COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

THÈSE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Signal, Image, Vision*

Par
Hanene Fatima Zohra BRACHEMI MEFTAH

Détection des attaques adverses pour des systèmes IA robustes

Thèse présentée et soutenue à INSA-Rennes, le 20/11/2025
Unité de recherche : Institut d'Électronique et des Technologies du numéRique (IETR)
Thèse N° : 25ISAR 24 / D25 - 24

Rapporteurs avant soutenance :

Melek ÖNEN	Professeur d'Université - EURECOM
Hichem SAHBI	Chargé de Recherche - LIP6 Sorbonne Université

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	William PUECH	Professeur d'Université - Université de Montpellier
Examinateurs :	Melek ÖNEN	Professeur d'Université - EURECOM
	Hichem SAHBI	Chargé de Recherche - LIP6 Sorbonne Université
	William PUECH	Professeur d'Université - Université de Montpellier
Directeurs de thèse :	Olivier DÉFORGES	Professeur d'Université - INSA Rennes
	Wassim HAMIDOUCHE	Maître de conférence - INSA Rennes
Encadrant :	Sid Ahmed FEZZA	Professeur d'Université - ENSTTIC

Acknowledgement

According to *Edward N. Lorenz*, a pioneer of the chaos theory, even something as seemingly insignificant as the flap of a seagull's wings could alter the course of the weather forever. In the same spirit, this work is the result of countless small contributions : weekly meetings, coffee break conversations, LinkedIn posts, summer school lectures, Reddit threads, conference and poster discussions, Stack Overflow answers, GitHub issues... I could spend hours trying to recall every single person who, knowingly or unknowingly, played a role in shaping this thesis, and I would still miss someone.
That is why I would like to take a moment to acknowledge the seemingly overlooked gestures : a quick comment, a shared link, a passing question, a random 'irrelevant' piece of information. These moments may feel insignificant, but they accumulate, shape ideas, and push thoughts forward. They mattered, and this thesis is proof of that.

To everyone who contributed, in any way : thank you. You have had an impact, and you have been heard.

Beyond these widespread influences, I owe a particular gratitude to my supervisors, Olivier Déforges, Wassim Hamidouche and Sid Ahmed Fezza. Their insightful guidance and countless hours of discussion were indispensable in shaping this thesis.

I extend my sincere thanks to the members of the jury for agreeing to review this manuscript. I am well aware of the time and attention such a task requires, and I am truly grateful for their commitment.

This research was made possible thanks to the financial support of Région Bretagne (Brittany region), CREACH Labs, and Direction Générale de l'Armement (DGA).

I am also profoundly grateful to the VAADER team for fostering such a pleasant environment during my stay at the lab. A special thanks goes to Lumi, Joël, Joseph (especially for the chess matches!), Meriem, Corinne, and Zoubida for their support.

Last but certainly not least, I extend my deepest thanks to my family for their unwavering support, patience, and encouragement throughout this journey. I truly hope this work lives up to your expectations and makes you proud.

Table of Contents

Executive summary **7**

Introduction **9**

Context and motivation 9
Thesis structure 10
Contributions overview 11

1 Foundational concepts **12**

1.1 Introduction 12
1.2 Deep Learning : a background 13
1.3 Adversarial attacks : a definition 14
1.4 Axes of categorization in adversarial attacks 14
1.4.1 Adversary's objective 15
1.4.2 Attack stage 17
1.4.3 Degree of access to the model 19
1.5 Taxonomy of Adversarial Attacks 20
1.6 Conclusion 47

2 Evaluating the Vulnerability of Deep Learning-based Blind Image Quality Assessment Methods to Adversarial Attacks **48**

2.1 Introduction 49
2.2 Related work 50
2.2.1 Blind image quality assessment models 50
2.2.2 Adversarial attacks on blind image quality assessment models . . . 51
2.3 Proposed attack 53
2.3.1 Problem statement 53
2.3.2 Attack design 55

2.4 Experiments and results . . . 57
2.4.1 Experimental Setup . . . 57
2.4.2 Results and Analysis . . . 58
2.5 Conclusion . . . 68

3 Energy Backdoor Attack to Deep Neural Networks 69
3.1 Introduction . . . 70
3.2 Related work . . . 71
3.2.1 Backdoor Attacks . . . 71
3.2.2 Energy-Latency Attacks on DNNs deployed on sparsity-based accelerators . . . 72
3.3 Proposed attack . . . 75
3.3.1 Problem Formulation . . . 75
3.3.2 Attack Design . . . 77
3.4 Experiments and results . . . 79
3.4.1 Experimental Setup . . . 79
3.4.2 Results and Analysis . . . 81
3.5 Conclusion . . . 84

4 Visual Information Protection through Adversarial Attacks on Vision-Language Models 85
4.1 Introduction . . . 86
4.2 Related work . . . 87
4.2.1 Vision-Language Models . . . 87
4.2.2 Vulnerability of VLMs to adversarial attacks . . . 88
4.2.3 Adversarial attacks for social good . . . 91
4.3 Attack methodology . . . 93
4.3.1 Problem statement . . . 95
4.3.2 Threat model . . . 95
4.3.3 Preliminaries and design overview . . . 96
4.4 Experimental setup . . . 98
4.4.1 Models . . . 98
4.4.2 Dataset . . . 98
4.4.3 Attacks . . . 98
4.4.4 Evaluation metrics . . . 99
4.4.5 Implementation details . . . 100
4.5 Results and analysis . . . 100
4.5.1 Attack effectiveness . . . 101

4.5.2 Understranding the choice of L_{max} . . . 107
4.5.3 Influence of λ_v . . . 110
4.5.4 Value's-norm only optimization . . . 111
4.5.5 Perturbation budget . . . 113
4.5.6 Failure cases . . . 113
4.5.7 Average attack run-times . . . 115
4.6 Conclusion . . . 116

Conclusion **117**
4.7 Summary of the contributions . . . 117
4.8 Limitations, future perspectives and open questions . . . 118

List of Publications **122**

Résumé en français **125**

List of Figures **145**

List of Tables **148**

Bibliography **149**

Executive summary

Deep Learning (DL) models have witnessed remarkable progress across a wide range of applications and modalities. However, their vulnerability to adversarial attacks raises critical security, privacy, and more recently ethical concerns. As modern models' architectures exhibit increasing complexity and are applied across a broad spectrum of critical and high-impact domains, understanding these vulnerabilities has become a pressing priority. However, the rapid pace at which new models are being developed has created a moving target for AI security research. In fact, as DL architectures become more sophisticated with continuously improving performance, existing attacks in the literature may no longer apply directly and new design choices may inadvertently introduce unforeseen vulnerabilities. Eventually, this dynamic landscape requires a continuous reassessment of robustness, urging researchers to evaluate whether modern models remain susceptible to known threats and if their evolving structure invites novel unexplored attack vectors.

This thesis aimed at investigating potential vulnerabilities in deep learning systems for computer vision, and develop attack strategies that expose overlooked vulnerabilities in the literature. The key contributions of this work include but are not limited to : 1) the design of a framework for generating both targeted and non-targeted adversarial attacks on deep learning-based image quality assessment models ; 2) the development of an energy backdoor attack for image classification models ; and 3) the proposal of a novel privacy-preserving adversarial attack targeting Vision-Language Models (VLMs). These contributions provide a broader perspective on how adversarial attacks can expose the susceptibility of models performing different tasks to be maliciously manipulated, while also demonstrating how such techniques could be repurposed for beneficial applications such as privacy preservation.

Our findings underscore the need for adaptive, context-aware defenses and contribute to the growing body of research in adversarial machine learning and AI security.

Introduction

Context and motivation

Machine Learning (ML), the science of developing computer systems which improve through experience, has evolved from a niche research area into a fundamental pillar of current technological progress. Its applications span a wide range of domains that go from recommendation systems and telecommunications infrastructure to fraud detection, profoundly transforming both academic research and industrial practices. At the heart of this revolution lies Deep Learning (DL), a subset inspired by the structure of the brain. DL algorithms have demonstrated exceptional capabilities in processing complex data, retrieving and recognizing patterns, and making data-driven decisions with minimal human intervention. The success of this domain and its accelerated progress can largely be attributed to several key factors including the drastic increase in computational power with the conception of specialized hardware architectures, the algorithmic scalability which enables the processing of large amounts of data by large models in a distributed fashion and the abundant availability of data.

Notably, the emergence of Large Language Models (LLMs) and Vision-Language Models (VLMs) represents the most recent ground breaking advancement in deep learning and has redefined the boundaries of what Artificial Intelligence (AI) systems can achieve, as it marked a leap towards general-purpose artificial intelligence. Unlike traditional models tailored for specific tasks, VLMs possess the ability to generalize across a wide array of applications without task-specific training. This includes writing, translation, coding, reasoning, decision making, visual question answering, object detection, user assistance and a large range of other applications. As a result, their versatility and scalability have enabled their integration into countless applications, products and services.

Today, these systems are integrated into a wide range of applications that support day-to-day human activities from healthcare and smart home devices to financial forecasting, supply chain optimization, autonomous vehicles, and camera surveillance. This suggests that AI has become a fundamental element embedded in daily human life, and while this adoption brings several benefits, it also introduces a range of security risks that must not be overlooked. In fact, this integration within critical infrastructures and personal digital environments, can have far reaching consequences if an adversary decides to exploit these systems' vulnerabilities.

While adversarial attacks on image classifiers have been widely studied in the literature, the field remains highly dynamic. New model architectures are constantly being designed and new vulnerabilities are continuously discovered creating an endless cycle. This thesis aims to broaden existing perspectives by exploring a series of attack strategies designed to uncover previously unaddressed vulnerabilities of ML systems for computer vision tasks. From compromising image quality assessment to energy backdoors and cross-modality vision-language models' manipulation, this work explores how neural networks can be subtly manipulated under very specific case scenarios. Together, these contributions highlight the pressing need to anticipate potential adversarial threats to enable the safe and robust deployment of AI systems.

Thesis structure

Throughout this thesis we exposed several vulnerabilities in Deep Neural Network (DNN) systems used for computer vision tasks. The structure of the thesis thus unfolds as follow.

- **Chapter 1** introduces the fundamentals of DL and presents key concepts related to attacks on DNNs in the context of computer vision.
- **Chapter 2** explores accuracy-based adversarial attacks on CNN-based image quality assessment metrics and examines the relationship between perceived image quality and attack efficiency in this specific context.
- **Chapter 3** investigates the feasibility of embedding an energy-based backdoor attack into standard DNNs operating on sparsity-based accelerators.

- **Chapter 4** introduces a novel privacy-preserving adversarial attack that obfuscates image regions deemed private, preventing them from being processed by the VLM to achieve a favorable privacy-utility trade-off.

Contributions overview

The following summarizes the main contributions of this thesis.

- Design of a framework for generating adversarial attacks against CNN-based models for Image Quality Assessment (IQA), along with the introduction of the untargeted attack setting within this specific context.
- Proposal of a stealthy energy backdoor attack targeting DNNs for image classification, specifically tailored for models running on sparsity-based accelerators.
- A comprehensive survey and benchmark of existing energy-latency attacks reported in the literature.
- Introduction of a novel privacy-preserving adversarial attack that protects private regions within an image from exploitation by VLMs, without compromising the utility of the remaining image content.

Chapitre 1

Foundational concepts

1.1 Introduction

Regardless of their remarkable performance across a wide range of tasks, DNNs remain highly vulnerable to a variety of attacks. As these models are being increasingly deployed within critical domains and applications such as autonomous systems, healthcare, finance, and security, the consequences of such vulnerabilities become significantly more pronounced. Thus, understanding the nature of these threats is essential to ensure the development of robust defense mechanisms and to achieve a safe and reliable deployment of these systems.

Perhaps, privacy is the most known vulnerability among the general public (Meunier, 2022). Common privacy attacks either aim to steal the parameters of a black-box model (Tramèr et al., 2016) or detect whether a data sample was used to train the model or not (Shokri et al., 2017). However, existing attacks and adversaries' objectives can go way beyond model stealing and membership inference. For instance, a well documented set of attacks in the literature are adversarial attacks that aim to disrupt the model's accuracy (Nguyen et al., 2015a; Szegedy et al., 2013). More recently, new types of attacks have emerged, driven by rapid technological progress and the continuous development of increasingly complex and sophisticated models. This includes energy-latency attacks (Shumailov et al., 2021) that compromise the model's energy efficiency and jailbreak attacks (Qi et al., 2024; Shayegani et al., 2024) which target the model's alignment,

making it generate unsafe and/or non ethical responses that do not align with human values.

1.2 Deep Learning : a background

A good starting point for understanding how these attacks work is to first understand how DL functions. In this section, we provide an overview of deep learning models, offering a foundation that will serve as a stepping stone for understanding existing vulnerabilities.

Deep Learning (DL) is a subset of Machine Learning (ML) that relies on the use of Deep Neural Networks (DNNs). An Artificial Neural Network (ANN), is as its name indicates, a set of artificial neurons interconnected that process data in a feed-forward fashion. A regular ANN includes an input, output and a set of hidden layers, each combining a set of hidden units. A neural network is considered 'deep' once it includes multiple hidden layers between the input and output layers, allowing it to model increasingly complex representations. These models are conceptualized as a parameterized mathematical function which maps the input $\mathbf{x}$ with the desired output y using a series of operations as delineated in equation (1.1).

$$y = \mathcal{M}_\theta(\mathbf{x}). \tag{1.1}$$

The design of this function is determined by a set of hyperparameters, including the choice of architecture, the number of layers and hidden units, and the activation functions. In addition, the model weights or parameters, denoted by θ, are learned or optimized by training the model on a large-scale dataset.

Training an ANN typically consists in an iterative process of finding the optimal combination of weights θ^* that minimize a given cost function $\mathcal{L}$ over a set of training data. The cost function $\mathcal{L}$ is designed to capture the performance of the model on the given task.

$$\theta^* = \arg\min_\theta \mathcal{L}\Big(\mathcal{M}_\theta(\mathbf{X}^{train}), Y^{train}\Big). \tag{1.2}$$

Equation (1.2) provides a compact and widely used equation to summarize the training process of a neural network with $\mathbf{X}^{train}$ and Y^{train} being respectfully, the training set and the ground-truth labels used to train the model. In practice, this loss can be computed in several ways : by evaluating the cost function on each training sample individually and averaging the results, by computing it over batches of samples (which is the most common approach) or by applying it directly to the entire dataset. For simplicity, we

make abstraction of these details in equation (1.2) and throughout the remainder of this manuscript. The optimal weights θ^* are typically found using gradient-based optimization algorithms such as Stochastic Gradient Descent (SGD). In its most basic form as presented in equation (1.3), this process is guided by a hyper-parameter known as the learning rate α, which controls the speed at which the model's parameters are updated. More advanced optimization algorithms further incorporate techniques such as momentum and regularization to improve convergence and generalization.

$$\theta^{(i+1)} = \theta^{(i)} + \alpha \nabla_{\theta^{(i)}} \mathcal{L}\Big(\mathcal{M}_{\theta^{(i)}}(\mathbf{X}^{train}), Y^{train}\Big), \tag{1.3}$$

Where $\theta^{(i)}$ and $\theta^{(i+1)}$ are the current and the new updated weights, α is the learning rate, and $\nabla_{\theta^{(i)}}$ is the gradient operator with respect to $\theta^{(i)}$.

Once training phase is complete, the model is evaluated on a set of unseen data, commonly referred to as the validation set $(\mathbf{X}^{val}, Y^{val})$. This evaluation helps monitor the model's ability to generalize over data it has not encountered during its training phase.

1.3 Adversarial attacks : a definition

One commonly accepted definition of adversarial attacks among the general public is the intentional manipulation of input data in a way that causes the ML model to misclassify it either into a target class chosen by the adversary, or into any incorrect class. These inputs, usually are referred to as adversarial examples and are often perceptually indistinguishable from clean inputs to humans. While adversarial attacks are most often associated with misclassification at inference time (*i.e.,* evasion attacks), they also encompass other forms of manipulation, such as data poisoning during the training phase of the model, other adversary's objectives by targeting model confidentiality, efficiency, integrity or alignment and models performing a variety of tasks besides image classification. However, throughout this thesis, we solely focus on adversarial attacks against ML models for computer vision applications.

1.4 Axes of categorization in adversarial attacks

In the literature, various possible ways have been proposed to categorize adversarial attacks. These typically consider factors such as the adversary's objective, the degree of

access they have to the target/victim model, and the stage of the ML pipeline at which the attack occurs.

1.4.1 Adversary's objective

Figure 1.1 illustrates different attack scenarios based on the adversary's objective. We provide a concise definition for each case below. However, this manuscript primarily focuses on accuracy-based attacks and energy-latency attacks, as these align most closely with our contributions to the field.

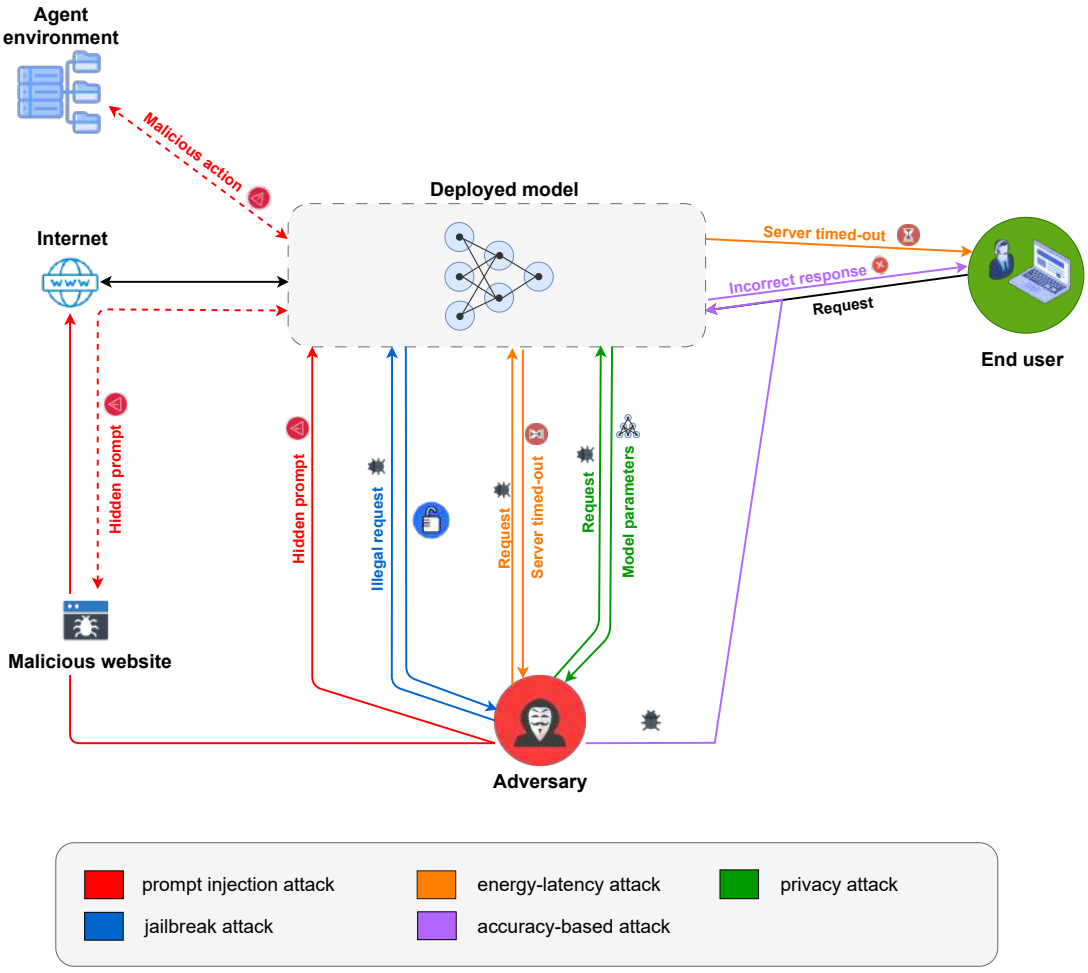


Figure 1.1 : Different attack scenarios based on the adversary's objective.

Accuracy-based attacks :

This category of attacks involves the adversary's attempt to alter or manipulate the

model's output, either toward a specific target output or simply toward an incorrect one. We refer to the former as *targeted* attacks and to the latter as *untargeted* attacks. Many works in this direction aim to misclassify an image into a class different from its ground-truth label (Akhtar and Mian, 2018; Goodfellow et al., 2014; Kurakin et al., 2018), as illustrated in Figure 1.2. However, these attacks are not limited to classification tasks. In fact, they target models performing a variety of tasks including but not limited to object detection (Huang et al., 2020; Thys et al., 2019; Yang et al., 2018), image captioning (Li et al., 2024b; Xu et al., 2019), and image generation (Aneja et al., 2022; Sun et al., 2021).

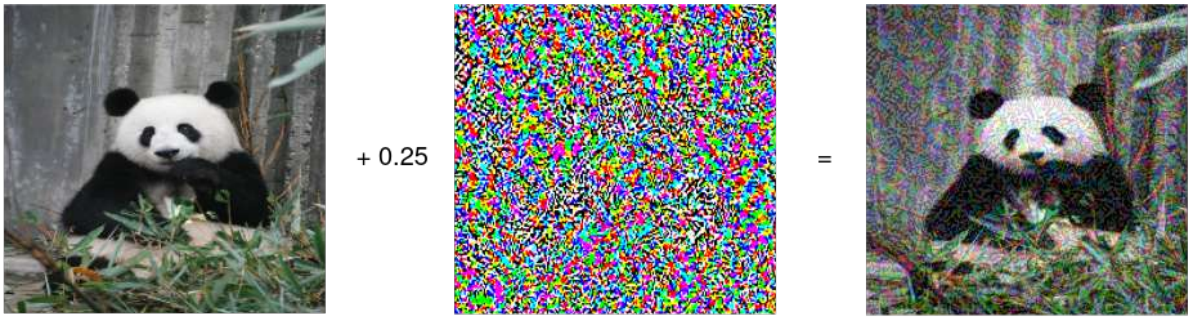


Figure 1.2 : A demonstration of the FGM adversarial attack (Goodfellow et al., 2014) applied to the GoogLeNet classifier.

Energy-latency attacks :

Energy-latency examples are inputs specifically crafted to increase a model's energy consumption and/or processing latency. These attacks can target properties of the model architecture as well as characteristics of the platform and hardware components used to deploy the model. Examples of the first category include attacks targeting the auto-regressive property of VLMs (Gao et al., 2024b) and the dynamic adaptive property of some DNNs (Haque et al., 2020; Hong et al., 2021; Pan et al., 2022). In contrast, attacks in the second category manipulate factors such as the number of memory accesses (Krithivasan et al., 2022) or exploit sparsity optimizations (Cinà et al., 2022; Lintelo et al., 2024; Shumailov et al., 2021). In practice, Shumailov et al. (2021) demonstrated that such attacks can lead to Denial-Of-Service (DoS) conditions in real-world settings, successfully achieving that on Microsoft Azure's translator system.

Privacy attacks :
Privacy attacks refer to a set of methods used by malicious entities to gain unauthorized access to sensitive information through interactions with the ML model. A brief classification of these attacks, widely adopted in the literature, is as follows.

- *Membership inference :* these attacks focus on determining whether a specific input sample was included in the model's training dataset.
- *Data reconstruction :* these attacks aim to reconstruct or recover an individual's data from the model.
- *Model extraction :* the attacker attempts to extract information about the model's architecture and its parameters by requesting the model.
- *Property inference :* the adversary seeks to learn global information about the training data distribution.

Jailbreak attacks :
With the advent of LLMs and VLMs, which have gained widespread popularity as multi-purpose systems, there has been a growing need to align these models with human values. This is especially critical given that they are trained on vast amounts of unfiltered data. Jailbreak attacks target this alignment by manipulating the model into, for example, responding to unethical user requests or into generating harmful and biased content.
Prompt injection attacks :
More recently, those models have gained even more popularity and have been further tuned to interact with and manipulate various resources (for example, parsing and sending emails, scheduling events, and making reservations). As a result, adversaries have taken advantage of this access to manipulate the model into performing predefined malicious actions by embedding hidden prompts within the model's input.

1.4.2 Attack stage

Depending on the stage in which the attack occurs, adversarial attacks can be broadly categorized into training-stage attacks and inference-stage attacks. Figure 1.3 provides an illustration of this classification.

Inference-stage attacks, occur after the model has been deployed. These attacks typically encompass the meticulous crafting of inputs designed to meet an objective when

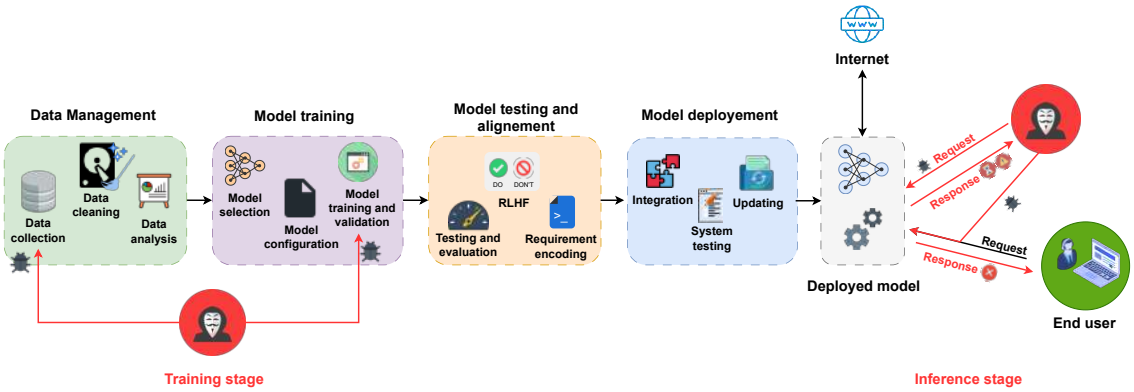


Figure 1.3 : Different attack scenarios throughout the model's lifecycle.

presented to the target model. To achieve this, the adversary designs a specific loss function that encodes their objective and formulates an optimization problem, where the goal is typically to find an optimal perturbation δ which is added to the clean input image $\mathbf{x}$. The resulting adversarial image ($\mathbf{x}_{adv} = \mathbf{x} + \delta$) has to meet the adversary's objective while satisfying a set of predefined constraints.

$$\delta^* = \arg\min_{\delta} \mathcal{L}\Big(\mathcal{M}(\mathbf{x}+\delta)\Big), \ \textbf{s.t.} \ \|\delta\|_p < \epsilon, \tag{1.4}$$

where ϵ is a threshold on the perturbation norm, and $p \in \mathbb{N}$ denotes the type of norm used to monitor the perturbation constraint.

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}} \tag{1.5}$$

Equation (1.5) and figure 1.4 provide respectively, the mathematical definition of the L_p and a geometric representation of the most commonly used norms in adversarial machine learning.

On the other hand, we refer to training-stage attacks as any adversarial actions that occur during the pre-deployment phase of the ML model. These attacks involve a malicious attempt to manipulate the training data, the model's learning process, or its internal weights to embed a malevolent pattern within the deployed model. This implies that within this setting, the adversary aims to find optimal model weights θ_{adv} that satisfies their malicious objective rather than an optimal adversarial input $\mathbf{x}_{adv}$.

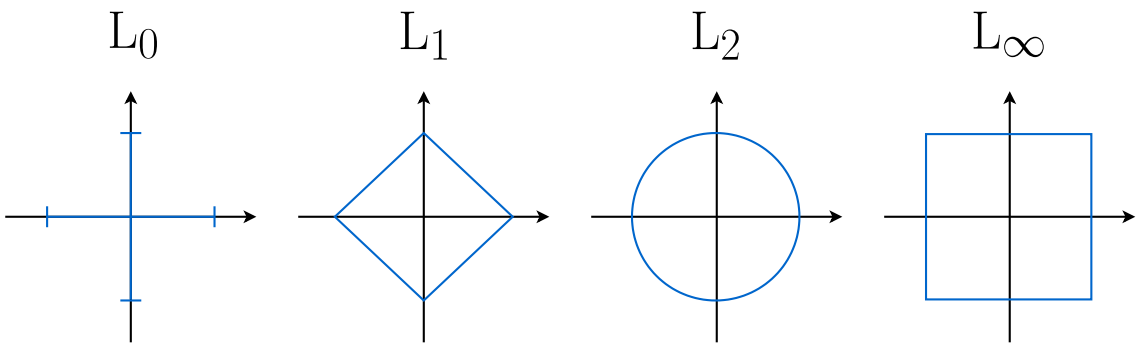


Figure 1.4 : Geometric representation of different L_p norms.

1.4.3 Degree of access to the model

Depending on the availability of information regarding the targeted model we can distinguish two categories of inference-stage attacks : white-box and black-box attacks.

In a white-box setting, the adversary possesses complete knowledge about their target model, including its architecture and weights or parameters for DNN models. This level of transparency empowers the attacker to craft more effective attack strategies that exploit the model's vulnerabilities based on the available information. In the black-box setting, the adversary operates under conditions of limited or no knowledge about the target model. Typically, interaction is restricted to the model's output, with no access to internal parameters such as weights or architecture. This lack of information about the victim model significantly increases the complexity of optimizing attacks. Consequently, adversaries often resort to techniques like gradient estimation, transferability, and metaheuristics to orchestrate these attacks.

However, in practical and real-world scenarios, a third grey-box category is often more representative. Grey-box settings assume that the adversary has partial knowledge of the model, such as access to training data, some architectural details, or a set of probabilities associated with the model's output. This category captures a more realistic threat model, where attackers operate with limited but non-negligible insight into the system, regardless of how this information was acquired. Nevertheless, in the literature, the term 'black-box attacks' is loosely applied, potentially conflating it with this grey-box category. To avoid ambiguity, we adopt a simplified terminology in which white-box refers to full access to the model, and black-box encompasses all other cases. For clarity, we will explicitly specify the degree of access granted to the adversary whenever necessary.

1.5 Taxonomy of Adversarial Attacks

In this section, we present a taxonomy of adversarial attacks based on common categorization axes defined in the former section, providing examples from prominent studies in the literature.

Inference stage

White-box : In what follows, we present the most prominent and widely studied white-box adversarial attacks in the literature, which have been designed to compromise model accuracy and its latency or energy consumption. Szegedy et al. (2013) were the first to reveal the vulnerability of DNNs to these adversarial attacks, introducing the **L-BFGS** attack as a proof of concept.

Goodfellow et al. (2014) further introduced the **Fast Gradient Method (FGM)** attack, a single-step method that leverages the gradient of the model's loss function, and a parameter ϵ to control the magnitude of the perturbation applied to the adversarial example, as presented in equation (1.6).

$$\mathbf{x}_{adv} = \mathbf{x}_{clean} - \epsilon \times sign\Big(\nabla_{\mathbf{x}_{clean}} \mathcal{L}\Big(\mathcal{M}(\mathbf{x}_{clean}), y_{target}\Big)\Big), \tag{1.6}$$

In this context, $\mathbf{x}_{clean}$ represents the original (clean) input image, and y_{target} denotes the target label. The value of ϵ must be carefully tuned to balance attack effectiveness and the perceptibility of the perturbation.

To explore the vulnerabilities of DNNs, Papernot et al. (2016b) introduced a method to build adversarial saliency maps, enabling attackers to target the most salient input features for perturbation. In the literature, their attack is more widely known as **Jacobian-based Saliency Maps (JSM)**. Briefly, their attack employs a greedy strategy to iteratively select and modify individual pixels. They then compute the gradient of the model's pre-softmax output to estimate each pixel's influence on the final classification. Let $\{p, q\}$ denotes a set of pixels and $Z(\mathbf{x})$ the model's pre-softmax output. The adversarial saliency map, which guides pixel selection, is defined in equation (1.7).

$$\begin{aligned} \alpha_{\{p,q\}} &= \sum_{i \in \{p,q\}} \frac{\partial Z(\mathbf{x})_t}{\partial \mathbf{x}_i}, \\ \beta_{\{p,q\}} &= \sum_{p_k \in \{p,q\}} \sum_j \Big(\frac{\partial Z(\mathbf{x})_j}{\partial \mathbf{x}_{p_k}}\Big) - \alpha_{p_k \in \{p,q\}}, \end{aligned} \tag{1.7}$$

where $Z(\mathbf{x})_j$ denote the $j-$th neuron output and t the target class. $\alpha_{\{p,q\}}$ measures the degree to which changes in pixels $\{p,q\}$ affect the prediction of the target class t, while $\beta_{\{p,q\}}$ quantifies the influence of modifying pixels $\{p,q\}$ on all the outputs. The attack objective can thus be expressed as shown in equation (1.8) such that the target class becomes more probable ($\alpha > 0$), whereas the probabilities of all other classes are reduced ($\beta < 0$).

$$(p^*, q^*) = \arg\max_{(p,q)} (-\alpha_{\{p,q\}}\, \beta_{\{p,q\}})(\alpha_{\{p,q\}} > 0)(\beta_{\{p,q\}} < 0). \tag{1.8}$$

Moosavi-Dezfooli et al. (2016) designed the **DeepFool** attack that operates by approximating the decision boundary as a hyperplane and iteratively computing the minimal perturbation that alters the model prediction. Typically, the perturbation vector is approximately orthogonal to the decision boundary.

Kurakin et al. (2018) examined the application of adversarial attacks under real-world conditions. To this end, they introduced two iterative variants of the **FGM** attack : the **Basic Iterative Method (BIM)** and **Iterative Least-likely Class Method (ILCM)** methods. The BIM attack is implemented by applying the FGM algorithm iteratively with a small step size α, as shown in equation (1.9).

$$\begin{aligned} \mathbf{x}_{adv}^{(0)} &= \mathbf{x}_{clean} \\ \mathbf{x}_{adv}^{(i+1)} &= \text{Clip}_{\epsilon}\Big(\mathbf{x}_{adv}^{(i)} + \alpha \times sign\Big(\nabla_{\mathbf{x}_{adv}^{(i)}}\, \mathcal{L}\Big(\mathcal{M}(\mathbf{x}_{adv}^{(i)}),\, y_{true}\Big)\Big)\Big). \end{aligned} \tag{1.9}$$

However, according to the authors, these attacks are primarily effective on small datasets where classes are well-separated and less similar. To address this limitation, they introduced the ILCM attack, in which the target label is replaced with the least likely class predicted by the model. Accordingly, Equation (1.9) is adapted to reflect this change as follows.

$$\begin{aligned} \mathbf{x}_{adv}^{(0)} &= \mathbf{x}_{clean} \\ y_{LL} &= \arg\min_{y}\{p(y|\mathbf{x}_{clean})\} \\ \mathbf{x}_{adv}^{(i+1)} &= \text{Clip}_{\epsilon}\Big(\mathbf{x}_{adv}^{(i)} - \alpha \times \text{sign}\Big(\nabla_{\mathbf{x}_{adv}^{(i)}}\, \mathcal{L}\Big(\mathcal{M}(\mathbf{x}_{adv}^{(i)}), y_{LL}\Big)\Big)\Big), \end{aligned} \tag{1.10}$$

where y_{LL} designates the least likely class predicted by the model on the clean image $\mathbf{x}_{clean}$ and Clip is a function which performs per-pixel clipping so that the resulting image is within the $\epsilon-$neighboring of the clean image.

Carlini and Wagner (2017b) designed the **C&W** attack presented in equation (1.11) in which they experiment different L_p norms, including L_0, L_2 and L_∞.

$$\begin{aligned} \mathbf{x}_{adv} &= \mathbf{x}_{clean} + \delta \\ \delta &= \min_{\delta} \|\delta\|_p + c\, \mathcal{M}_\theta(\mathbf{x}_{clean} + \delta) \end{aligned} \tag{1.11}$$

Tramèr et al. (2017) performed a quantitative study to investigate different models' decision boundaries. They demonstrated that models from different hypothesis classes learn decision boundaries that are very close to each other in both benign and adversarial directions. As a result, they presented the **Gradient Aligned Adversarial Subspace (GAAS)** attack, a technique that estimates the dimensionality of the adversarial subspace. For an input $\mathbf{x}_{clean}$, the objective of the attack is to find a space of perturbations δ (with $\|\delta\|_2 < \epsilon$) that results in a significant increase in the loss, i.e., $\mathcal{L}\Big(\mathcal{M}_\theta(\mathbf{x}_{clean} + \delta), y_{true}\Big) > \mathcal{L}\Big(\mathcal{M}_\theta(\mathbf{x}_{clean}), y_{true}\Big) + \gamma$, with $\gamma \in \mathbb{R}^+$ being a fixed positive real constant.

Madry et al. (2017) introduced the **Projected Gradient Descent (PGD)** attack following a saddle point formulation as formalized in equation (1.12). They conducted an experimental study of the optimization landscape and demonstrated that first-order methods can reliably solve the adversarial optimization problem.

$$\mathbf{x}_{adv}^{(i+1)} = \text{Proj}_\epsilon\Big(\mathbf{x}_{adv}^{(i)} + \alpha \times sign\Big(\nabla_{\mathbf{x}_{adv}^{(i)}} \mathcal{L}\Big(\mathcal{M}(\mathbf{x}_{adv}^{(i)}),\, y_{true}\Big)\Big), \tag{1.12}$$

where Proj is a projection operator ensuring perturbed input stays within predefined ϵ bounds.

Dabouei et al. (2020) proposed the **SmoothFool** attack, a geometry-inspired attack that exploits the decision boundaries of the model to generate smooth adversarial perturbations.

$$\delta = \arg\min_{\delta} \|\delta\|_2 + \lambda\, \Omega(\delta) \text{ s.t. } \mathcal{M}_\theta(\mathbf{x} + \delta) \neq \mathcal{M}_\theta(\mathbf{x}), \tag{1.13}$$

where Ω is denotes a function that measures the roughness. To achieve this, they use a frequency response function H_{hp}, which emphasizes high-frequency components, as shown in equation (1.14).

$$\Omega(\delta, H_{hp}) = \int_{-\infty}^{+\infty} R(\omega)^2 H_{hp}(\omega)^2 d\omega, \tag{1.14}$$

where R is the Fourier transform of δ. They then adapt this formulation to the discrete spatial domain of the perturbation, relaxing and linearizing it using the geometric pro-

perties of the model's boundaries. Equation (1.14) thus transforms into equation (1.15).

$$\Omega(\delta, h) = \|\delta - \delta * h\|_2^2, \tag{1.15}$$

where $*$ and h refer, respectively, to the convolution operation and the discrete approximation of $H_{lp} = 1 - H_{hp}$.

Moosavi-Dezfooli et al. (2017) demonstrated the existence of image-agnostic perturbations that can cause images to be misclassified. These perturbations are referred to as **Universal Perturbation Attacks (UAPs)** and are optimized over a distribution of images μ rather than a single image, as presented in equation (1.16).

$$\delta = \arg\min_{\delta} \|\delta\|_2 \text{ s.t. } \mathbb{P}_{\mathcal{M}_\theta(\mathbf{x}\sim\mu)}\Big(\mathbf{x} + \delta) \neq \mathcal{M}_\theta(\mathbf{x})\Big) > 1 - \varepsilon, \tag{1.16}$$

where ε is a parameter, chosen to be as small as possible, that indicates the perturbation's effectiveness on *almost* all images within the distribution μ.

More recently, energy-latency adversarial attacks have emerged, in which the adversary's objective is to increase the model's latency and/or energy consumption. In the white-box setup, the success of the attacker heavily depends on the adversary's comprehensive understanding of the prediction pipeline's functionality. The attacker must strategically formulate the attack by integrating specific criteria into the optimization process and carefully selecting which aspects to target. Numerous energy-latency attacks have been proposed in the literature, each meticulously designed for precise and predefined environments or targeting very specific DL models. In the following, we present the most prominent among them.

In their seminal work, Shumailov et al. (2021) highlighted a gap between the average energy consumption of a model and its energy consumption under worst-case conditions. They attribute this to the variability of energy consumed during an inference pass, which depends on two primary factors : the total number of arithmetic operations performed for the input inference and the number of memory accesses. This observation suggests that increasing the overall required operations for input inference could result in higher energy consumption rates. For example, modern Natural Language Processing (NLP) architectures may exhibit diverse internal representation sizes, even when inputs have identical sizes. This diversity is typically introduced by the Tokenizer in the early layers of the DNN, implying that it is possible to craft an input that maximizes the number of generated tokens in this layer. Consequently, this would lead to an increase in the total

number of operations required to process the input.
On the other hand, many hardware accelerators capitalize on data sparsity to enhance computational and energy efficiency by avoiding unnecessary operations. This can entail bypassing operations involving zero elements or utilizing sparse matrix representations to optimize storage and memory access. In particular, sparsity in DNN architectures is often triggered by the Rectified Linear Unit (ReLU) activation function, enabling significant optimization of computational costs without sacrificing model accuracy. The attacker can exploit the reduction of sparsity to amplify the energy and/or latency of the model by increasing the overall number of operations needed to process the input.

In Computer Vision (CV) architectures such as Convolutional Neural Networks (CNNs), the attacker may choose to decrease the sparsity of these internal representations to counteract the effects of hardware accelerators, as expressed in equation (1.17).

$$\mathbf{x}_{sp} = \arg\max_{\mathbf{x}} \sum_{a^{(l)} \in \mathcal{A}(\mathbf{x})} \|a^{(l)}\|_0, \tag{1.17}$$

where $a^{(l)}$ denotes the l-th layer's activation when processing the input $\mathbf{x}$ and $\mathcal{A}(\mathbf{x})$ refers to the set of all the model's internal activations during the inference of $\mathbf{x}$. Their initial attack, while novel, produced adversarial images that appeared artificial. This was due to their reliance on Genetic Algorithms (GAs) or Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization without constraints on the visibility of the adversarial perturbation. Krithivasan et al. (2020) subsequently improved upon this by introducing a more stealthy version. Their method incorporates box constraints on the perturbation's norm, ensuring the correct classification of the 'sponge example' as defined in equation (1.18).

$$\begin{aligned} \mathbf{x}_{sp} = \arg\max_{\mathbf{x}} \frac{1}{|\mathcal{A}(\mathbf{x})|} \\ \sum_{a^{(l)} \in \mathcal{A}(\mathbf{x})} \|a^{(l)}\|_0 - \lambda \, \mathcal{L}_{ce}\Big(\mathcal{M}_\theta(\mathbf{x}), y_{true}\Big), \end{aligned} \tag{1.18}$$

where $\lambda \in \mathbb{R}^+$ denotes the Lagrangian multiplier, $\mathcal{M}_\theta$ is the target model and $\mathcal{L}_{ce}$ computes the Cross Entropy (CE) loss between $\mathbf{x}$'s ground-truth class y_{true} and its corresponding predicted class by the model. In practice, this sparsity attack is implemented in two stages. First, the objective function in equation (1.18) is optimized. Subsequently, a binary search is employed to determine the optimal value for the Lagrangian multiplier λ.

The authors also introduced a method to estimate the L_0 norm, which quantifies the sparsity of neural network activations. This norm is rendered differentiable through the use of hyperbolic tangent and sigmoid functions, enabling optimization via gradient-based methods. To enhance attack efficiency, they incorporate layer-specific weights W_l, where each weight is proportional to the execution time or energy consumption of its corresponding layer l. This modification transforms equation (1.18) into equation (1.19).

$$\mathbf{x}_{sp} = \arg\max_{\mathbf{x}} \frac{1}{|\mathcal{A}(\mathbf{x})|} \sum_{a^{(l)} \in \mathcal{A}(\mathbf{x})} W_l \|a^{(l)}\|_0 + \lambda \, \mathcal{L}_{ce}\Big(\mathcal{M}(\mathbf{x}), y\Big). \tag{1.19}$$

Finally, the perturbation's box constraints are enforced by clipping the resulting adversarial example $\mathbf{x}_{sp}$ to be within the interval $[0, 1]$.

Input-adaptive DNNs, also referred to as **Dynamic Neural Network (DyNN)**, have garnered significant attention for their ability to reconcile two seemingly contradictory goals : achieving high predictive accuracy while maintaining low computational complexity. This capability has drawn the attention of adversaries to these models. Various techniques facilitate the dynamic allocation of resources during the feed-forward process based on the input data. Among these, two primary types of input-adaptive DNNs are frequently employed in the literature : **adaptive neural networks (AdNNs)** and **multi-exit architectures**. AdNNs enable the selective activation of specific layers (depth-dynamic AdNNs) (Figurnov et al., 2017; Wang et al., 2018) or channels (width-dynamic AdNNs) (Bejnordi et al., 2020; Chen et al., 2019) within the network. This selective activation is governed by a gating function that determines which parts of the network are utilized. In contrast, multi-exit architectures (Huang et al., 2018; Teerapittayanon et al., 2016) facilitate expedited decision-making by incorporating multiple side branches, also known as early exits, within the model's architecture. This allows the model to preemptively terminate unnecessary computations in later stages by halting the forward pass when a predefined stopping criterion is satisfied at an early exit point. These input-adaptive strategies have been successfully implemented in efficient VIision Transformer (ViT)-based architectures (Meng et al., 2022; Yin et al., 2022). Furthermore, token pruning, a technique that discards less informative tokens during the forward pass, has proven effective in enhancing model efficiency and reducing the overall computational burden associated with input processing. Dynamic quantization is also a

notable approach that improves the model's adaptability to input variations. For instance, applying dynamic quantization in transformer architectures like ViTs and Data-efficient image Transformers (DeiTs) allows for dynamic precision allocation in feature vectors and weight matrices. This results in improved memory usage and computational efficiency during forward propagation while maintaining model performance (Dettmers et al., 2022). Dynamic quantization streamlines matrix multiplication through three key steps : outlier identification, mixed-precision matrix multiplication (where lower precision is used for non-outlier operations and standard precision for outliers), and finally, dequantization and aggregation of the results.
Although these techniques offer significant computational efficiency, adversaries with a basic understanding of these methods could potentially use adversarial examples to undermine their effectiveness (Baras et al., 2023; Hong et al., 2021).

ILFO (Haque et al., 2020) is one of the first attacks specifically crafted to target AdNNs. To maximize computational load, Haque et al. designed the attack to activate as many layers as possible within the model. This is typically achieved by crafting an adversarial example $\mathbf{x}_{sp}$ that forces the model's internal activations towards a target state, $a^{(l)}_{target}$. This target state represents the worst-case scenario, where all blocks within each layer are activated.

The objective function of ILFO also penalizes the strength of the perturbation, δ, to ensure that the adversarial example $\mathbf{x}_{sp}$ remains perceptually similar to the original, clean input $\mathbf{x}_{clean}$. Furthermore, to guarantee that the generated adversarial examples adhere to box constraints and to avoid the limitations of gradient clipping methods, the authors introduce a pre-processing function defined in equation (1.20). This function transforms the image into a latent space where the adversarial example is optimized. The optimized example is then transformed back into the original image space.

$$\begin{aligned} \hat{\mathbf{x}} &= \mathcal{T}(\mathbf{x}) = \operatorname{atanh}(2\ \mathbf{x}\ - 1), \\ \mathbf{x} &= \mathcal{T}^{-1}(\hat{\mathbf{x}}) = \frac{1}{2}\,[\tanh(\hat{\mathbf{x}}) + 1]. \end{aligned} \tag{1.20}$$

Equation (1.21) provides a concise summary of the ILFO attack formulation.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathcal{T}^{-1}(\hat{\mathbf{x}} + \delta^{\star}), \\ \hat{\mathbf{x}} &= \mathcal{T}(\mathbf{x}_{clean}), \\ \delta^{\star} &= \arg\min_{\delta} \delta + \lambda \sum_{a^{(l)} \in \mathcal{A}(\hat{\mathbf{x}}+\delta)} \mathcal{L}_{dist}\Big(a^{(l)}, a^{(l)}_{target}\Big), \end{aligned} \tag{1.21}$$

where $\mathcal{L}_{dist}$ refers to a carefully crafted loss that measures the distance between the $a^{(l)}$ and $a^{(l)}_{target}$ states when processing the input $\hat{\mathbf{x}} + \delta$, λ is the Lagrangian multiplier and $\mathcal{T}$ refers to the pre-processing function as defined by equation (1.20).

In contrast, **DeepSloth** (Hong et al., 2021) has been specifically developed to counter multi-exit architectures. The attack involves diminishing the effectiveness of the network by reducing the confidence of predictions at various exit points within the model. This results in bypassing early exits and increasing the number of required blocks to process the input, thereby augmenting the computational and energy costs of inference. A multi-exit architecture $\mathcal{M}$ (for simplicity, we omit the model parameters θ from the notation) includes K exit points denoted as $\mathcal{M}_i, i = \overline{1, K}$, each attached to a hidden block of the model. These exit points produce internal classifications based on the output of their corresponding block. Subsequently, the entropy or confidence of the prediction from the i-th exit point is utilized (typically compared to a threshold τ) to determine whether additional blocks should be employed to process the input. As a result, the adversary can circumvent early exits by crafting sponge examples with high entropy or low confidence, indicated by predictions that adhere to a uniform distribution, as outlined in equation (1.22).

$$\begin{aligned} \mathbf{x}_{sp} &= \mathbf{x}_{clean} + \delta^{\star}, \\ \delta^{\star} &= \arg\min_{\delta} \sum_{x \in \mathcal{D}'} \sum_{i=1}^{K} \mathcal{L}_{ce}(\mathcal{M}_i(\mathbf{x} + \delta), y_{target}), \\ \text{s.t} \quad & \|\delta\|_{\infty} < \epsilon, \end{aligned} \tag{1.22}$$

where the target class is uniformly distributed among the available classes, denoted as $y_{target} \sim \mathcal{U}$ and $\mathcal{D}'$ is the image set used for the perturbation's optimization. The attack includes three different settings based on the definition of $\mathcal{D}'$: single-sample setting, universal setting, and class universal setting. In the first setting, $\mathcal{D}' = \{\mathbf{x}_{clean}\}$, the perturbation is optimized for a single sample, indicating that the generated adversarial perturbation is tailored to the specific clean input sample. This suggests that the resulting

perturbation may not be transferable to other samples. In the second case, $\mathcal{D}' = X$, the perturbation is optimized over a set of images X and is assumed to generalize over other unseen image samples. Finally, in the third setting, $\mathcal{D}' = \{\mathbf{x}^{(c)}\}$, where $\{\mathbf{x}^{(c)}\} \subset X$ refers to a set of image samples whose labels correspond to a target label c. In this latter case, the perturbation is optimized to be transferable among samples within the same target class c.

Chen et al. (2023b) concurrently introduced the **SAME** attack, which targets multi-exit transformers in NLP. **SAME** is an adaptation of **DeepSloth**, following the same formulation as described in equation (1.22). To enhance the attack's effectiveness and tailor it to various multi-exit transformers, they incorporate an additional objective : reducing the prediction patience. Certain multi-exit transformers (Zhu, 2021) employ a patience-based strategy, incrementing a 'patience' counter by one when consecutive early exits yield consistent predictions and resetting it to zero upon an inconsistent prediction. The model exits early if this patience counter reaches a predefined threshold. Thus, **SAME** attack aims to minimize a patience loss, $\mathcal{L}_{patience}$, defined as :

$$\mathcal{L}_{patience} = \sum_{i=1}^{N} \mathcal{L}_{ce}(\mathcal{M}_i(x+\delta), h_i), \tag{1.23}$$

where $h_i \neq \mathcal{M}_{i-1}(x+\delta)$ corresponds to any label different from the previous exit's predicted label. The objective of introducing this loss is to compel the model to generate inconsistent predictions between consecutive early exit classifiers. Moreover, they introduce layer-wise importance scores within the objective function to dynamically adjust the importance assigned to distinct early layer outputs.

Existing research on input-adaptive DNNs had not explicitly focused on the design of width-dynamic AdNNs. To address this gap, Pan et al. (2022) introduced **GradAuto**, an attack targeting both depth- and width-dynamic AdNNs. Their attack formulation, defined in equation (1.24), aims to generate a perturbation δ that forces the gate activation G_l of a layer l to exceed a threshold τ_l (i.e., $G_l(\mathbf{x}+\delta) > \tau_l$). This is equivalent to minimizing

the following loss function : $max(\tau_l - G_l(\mathbf{x}+\delta), 0)$.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathcal{T}^{-1}(\hat{\mathbf{x}} + \delta^\star), \\ \hat{\mathbf{x}} &= \mathcal{T}(\mathbf{x}_{clean}), \\ \delta^\star &= \arg\min_{\delta} \frac{1}{3HW}\|\delta\|_2 \\ &\quad + \sum_l \lambda_l \, max\Big(\tau_l - G_l(\hat{\mathbf{x}} + \delta), 0\Big). \end{aligned} \tag{1.24}$$

Their objective function includes a penalty on the perturbation's norm to enhance perceptual similarity between the clean and adversarial examples. Here, H and W denote the input image's dimensions, which are utilized for normalization. τ_l represents the threshold that activates the gate G_l, and $\mathcal{T}$ refers to the pre-processing function defined previously in equation (1.20). Additionally, the attack is refined by incorporating Lagrangian coefficients λ_l that are proportional to each layer's computational complexity relative to the entire network, and a more efficient optimization approach is introduced to stabilize the gradient updates.

In a different vein, Navaneet et al. (2023) proposed **SlowFormer**, an attack targeting efficient ViTs. This attack aims to diminish the optimization benefits of these architectures, which rely on dynamic computations for improved efficiency. **SlowFormer** achieves this by generating a universal patch designed to maximize the number of tokens selected during the forward pass (i.e., minimizing the number of tokens discarded by the efficient ViT).

QuantAttack (Baras et al., 2023), another attack strategy, undermines the optimizations introduced by dynamic quantization. This attack crafts a perturbation δ that maximizes the number of outliers in various feature vectors, consequently minimizing the number of low-precision multiplications. In practice, outliers in internal neural activations are defined as values exceeding a predetermined threshold τ. Therefore, generating the adversarial perturbation δ involves optimizing a sponge example, $\mathbf{x}_{\text{clean}} + \delta$, given by equation (1.25). The activations of this sponge example during the forward pass predominantly

contain values that surpass the threshold τ.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathbf{x}_{clean} + \delta^{\star}, \\ \delta^{\star} &= \arg\min_{\delta} \lambda_{quant}\, \mathcal{L}_{quant}(\mathbf{x}_{clean} + \delta) \\ &\quad + \lambda_{ce}\, \mathcal{L}_{ce}\Big(\mathcal{M}(\mathbf{x}_{clean} + \delta), y_{true}\Big) \\ &\quad + \lambda_{TV}\, \mathcal{L}_{TV}(\mathbf{x}_{clean} + \delta). \end{aligned} \tag{1.25}$$

The attack's objective is formulated in the term $\mathcal{L}_{quant}$. The $\mathcal{L}_{ce}$ loss ensures the stealthiness of the attack by maintaining correct prediction scores on the adversarial example, while the total variation loss $\mathcal{L}_{TV}$ promotes smooth color transitions between neighboring pixels in the adversarial example. λ_{quant}, λ_{ce}, and λ_{TV} are Lagrangian coefficients, while $\mathcal{M}$ and y_{true} respectively denote the target model and the ground truth score of the clean sample $\mathbf{x}_{clean}$.

Numerous attacks in the literature also target object detection models. A typical object detection pipeline consists of the following four steps :

1. Input pre-processing.
2. Object detection : The object detection model processes the input sample and generates a set of detected object candidates. This step includes feature extraction, object localization, and object classification.
3. Filtering the object candidates based on their confidence scores.
4. Refinement of the output using Non-Maximum Suppression (NMS) to discard redundant candidates.

In a seminal study exploring energy-latency attacks against object detection models, Shapira et al. (2023) demonstrated the limitations of existing energy-latency attack paradigms (Shumailov et al., 2021) in this specific context. Their analysis revealed that object detection models, due to their inherent design for handling complex scenes, often operate near worst-case conditions concerning sparsity exploitation, rendering traditional attacks less effective. Further, they identified the algorithmic complexity of NMS as a critical factor capable of significantly amplifying the latency and energy demands of processing an input. NMS, a fundamental component of object detection pipelines, operates by first ordering predictions based on confidence scores, followed by iterative elimination of bounding boxes whose Intersection Over Union (IoU) with a higher-confidence box surpasses a

predefined threshold τ_{IoU}. While typically efficient, NMS exhibits a worst-case quadratic time complexity of $O(|\mathcal{C}|^2)$, where $\mathcal{C}$ denotes the cardinality of the candidate set. Although this complexity usually diminishes with progressive candidate pruning, the authors identified it as a potential attack vector. Leveraging this insight, the authors introduced the **Phantom Sponges** attack (Shapira et al., 2023), a novel method designed to exploit the NMS bottleneck. The attack's core objective is to maximize the number of high-confidence detections processed by NMS while concurrently minimizing the number of candidates discarded during each iteration. This effectively forces NMS to operate near its worst-case complexity, thereby substantially increasing the overall processing latency of the object detection model. The **Phantom Sponges** attack achieves this through a carefully crafted formulation incorporating three distinct loss terms, as expressed in equation (1.26).

$$\mathbf{x}_{sp} = \arg\max_{\mathbf{x}} \ \lambda_{max_obj}\, \mathcal{L}_{max_obj}(\mathbf{x}) - \lambda_{bbox_area}\, \mathcal{L}_{bbox_area}(\mathbf{x}) - \lambda_{IoU}\, \mathcal{L}_{IoU}(\mathbf{x}). \tag{1.26}$$

The term $\mathcal{L}_{max_obj}$ maximizes the number of detected objects and their associated confidence scores, effectively inducing the model to generate spurious detections. Concurrently, $\mathcal{L}_{IoU}$ minimizes the pairwise IoU between candidate detections in $\mathcal{C}$. Finally, $\mathcal{L}_{bbox_area}$ further discourages IoU overlap by minimizing the area of the predicted bounding boxes.

Building upon the **Phantom Sponges** attack, Chen et al. (2024) introduced refinements by incorporating spatial information regarding the location and shape of detected objects into the optimization of $\mathbf{x}_{sp}$. Their experiments revealed that optimizing $\mathcal{L}_{IoU}$ could be further improved by strategically displacing the centroids of the bounding boxes to minimize their intersection. Consequently, their proposed **Overload** attack employs a "spatial attention" mechanism to manipulate the regions in which spurious objects are generated. This mechanism favors the generation of these spurious objects in low-density regions of the image, significantly reducing their pairwise IoU and thus mitigating their likelihood of being suppressed by NMS.

Subsequently, Liu et al. (2023) extended energy-latency attacks to the domain of point cloud object detectors with their **SlowLiDAR** attack. While the general principles of object detection remain consistent across image and point cloud modalities, their implementation differs significantly due to the specialized techniques required for point cloud processing. Several challenges arise when adapting these attacks to point clouds, including the non-differentiable nature of certain aggregation operations and the vast search space inherent in 3D data. To address these challenges, they approximated the non-differentiable

pre-processing pipeline with a differentiable proxy, enabling end-to-end optimization of the perturbation signal δ. Furthermore, they devised a probing algorithm to provide effective initialization for the subsequent gradient-based attack optimization. The objective function for generating adversarial point clouds shares structural similarities with those employed in previous works (Chen et al., 2024; Shapira et al., 2023). However, the manipulation of the optimized variable $\mathbf{x}$, as introduced in equation (1.26), differs slightly as it entails altering the 3D coordinates of existing points or integrating a new set of adversarial points to the initial clean set.

While the adversary's ability to directly maximize the complexity of the NMS component is constrained, the upper bound of NMS complexity remains dependent on the number of detected objects, suggesting potential for further optimization. In practice, an adversary can substantially elevate the model's energy consumption by manipulating both the second and third stages of the object detection pipeline. This contrasts with attacks against sparsity-based accelerators and AdNNs, where the model's worst-case scenario often corresponds to a fixed algorithmic complexity. The implications of such energy-latency attacks are significant, particularly for battery-powered edge devices employing object detection algorithms, as rapid battery depletion could render these devices inoperable.

Neural Ordinary Differential Equation (ODE) networks (Chen et al., 2018) have recently garnered significant attention from the research community due to their invertibility and efficiency. They incorporate ODE solvers to approximate complex transformations within the network, such as residual and recurrent connections, leading to reduced memory costs compared to traditional DNNs. An ODE solver typically approximates a function and iteratively computes its gradient based on the given differential equation at various points $\{t_i | i = \overline{0, N}\}$. The number of iterations required for the function's approximation is inversely proportional to the distance between these points, also known as the step size $\mathcal{H}$. Consequently, this variation in the number of iterations introduces fluctuations in the model's processing latency. In this context, Haque et al. (2023a) leverage the vulnerabilities inherent to ODE networks and introduce the **AntiNODE** attack to degrade the efficiency of neural ODE models. As detailed in equation (1.27), their attack seeks to minimize the step size $\mathcal{H}$, due to its inverse correlation with latency, while concurrently imposing a penalty on the distance between the clean image $\mathbf{x}_{clean}$ and the adversarial

image $\mathbf{x}_{clean} + \delta$, thereby maintaining perceptual similarity.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathcal{T}^{-1}(\hat{\mathbf{x}} + \delta^{\star}), \\ \hat{\mathbf{x}} &= \mathcal{T}(\mathbf{x}_{clean}), \\ \delta^{\star} &= \arg\min_{\delta} \mathcal{H}_{avg}(\hat{\mathbf{x}} + \delta) \\ &\quad + \lambda \|\mathcal{T}^{-1}(\hat{\mathbf{x}} + \delta) - \mathbf{x}_{clean}\|, \\ \mathcal{H}_{avg}(\hat{\mathbf{x}} + \delta) &= \frac{1}{N} \sum_{i=1}^{N} w_i \frac{e^{\mathcal{H}_i(\hat{\mathbf{x}}+\delta)}}{\sum_{j=1}^{N} e^{\mathcal{H}_j(\hat{\mathbf{x}}+\delta)}}, \end{aligned} \tag{1.27}$$

where $\mathcal{T}$ refers to the pre-processing step introduced in (1.20), $\mathcal{H}_i$ is the step size corresponding to t_{i-1} and t_i, w_i are assigned weights proportional to the step size $\mathcal{H}_i$ and λ is the Lagrangian coefficient. Energy-latency attacks targeting ODE offer a broader scope for energy maximization compared to sparsity-based attacks. However, their effectiveness is limited to scenarios where ODE is employed as an optimization mechanism.

Spiking Neural Networks (SNNs) are a subset of ANNs that encode and process information as discrete spikes. This binary characteristic enables compact and low-power implementations rendering them suitable for systems handling data in the form of event streams. In practice, SNNs exhibit greater fidelity to the biological model of the neurons rather than conventional ANNs. At a given time step t each neuron is associated with a membrane potential $V_i(t)$ that is updated based on the rule given by equation (1.28).

$$V_i(t) = \alpha V_i(t-1) + \sum_j W_{i,j} O_j(t), \tag{1.28}$$

where $O_j(t)$ is the output of the neuron j which transmits information to a neuron i, $W_{i,j}$ denotes their synapse connection and α refers to the leak factor of the neuron.
At the beginning of each time step t, the neuron's membrane potential is only updated in response to spikes in their inputs. After integrating all incoming spikes, the condition $V_i(t) > \gamma$ is checked before deciding whether the neurons' fan-out connections emit a spike $O_i(t) = 1$ or not $O_i(t) = 0$. The emission of a spike is followed by a reset of the neuron's membrane potential $V_i(t) = V_i(t) - \gamma\, O_i(t)$, and once the end of the time steps is reached, the spiking activity of the output neurons is used to determine the final SNN output.
To explore the resilience of SNNs against energy-latency attacks, Krithivasan et al. (2022) conducted an experimental analysis of their energy consumption and inference latency. Their findings revealed that inference latency is predominantly influenced by memory

accesses required to retrieve synaptic weights and membrane potentials. This implies a direct correlation between energy consumption, inference latency, and the quantity of spikes received at each neuron's input. Building on this insight, they introduced the **SpikeAttack**, as described by equation (1.29), aimed at maximizing inference latency. This attack strategy seeks to increase the number of spikes received at the inputs of network neurons.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathbf{x}_{clean} + \delta^{\star}, \\ \delta^{\star} &= \arg\max_{\delta} \sum_{t} \sum_{i} O_i(t) \\ &\quad - \lambda \, \mathcal{L}_{ce}\Big(\mathcal{M}(\mathbf{x}_{clean} + \delta), y\Big) \;\; \text{s.t} \;\; \|\delta\| < \epsilon. \end{aligned} \tag{1.29}$$

Following the approach of prior research in CV, **SpikeAttack** relies on a CE loss, denoted as $\mathcal{L}_{ce}$, to maintain the functional stealthiness of the attack. Additionally, they imposed box constraints on the norm of δ to enhance perceptual similarity between the clean and adversarial inputs. Due to their similarity to sparsity-based attacks, these attacks share the same limitations. Consequently, the maximum energy consumption of the model is constrained by the maximum number of memory accesses, which corresponds to the total number of neurons in the network.

Neural Machine Translation (NMT) systems are crucial for facilitating communication across language barriers. Thus, ensuring their computational efficiency is paramount given the large volume of real-time translation requests they handle on a daily basis. To better understand the robustness of these systems' computational efficiency, Chen et al. (2022a) conducted several experiments among which they observed that the processing latency of these systems is directly proportional to the length of their output sequence. This induced latency is fundamentally due to the underlying invocations of the decoder with non-deterministic numbers of iterations for the generation of each token of the output sequence. On the other hand, the length of the output sequence is either restricted to a maximum threshold or, in most cases, dependent on the generation of an End Of Sequence (EOS) token by the system. Based on this observation, the **NMTSloth** attack (Chen et al., 2022a), formalized in equation (1.30), was proposed to increase the latency of NMT

systems by extending the length of the generated output sequence.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathbf{x}_{clean} + \delta^{\star}, \\ \delta^{\star} &= \arg\max_{\delta} \ \text{Loop}_{\mathcal{M}}(\mathbf{x}_{clean} \oplus \delta), \\ &\text{s.t} \ \ \|\delta\| < \epsilon, \end{aligned} \tag{1.30}$$

where $\mathbf{x}_{clean}$ refers to the clean input sentence, $\text{Loop}_{\mathcal{M}}(\mathbf{x}_{clean} \oplus \delta)$ represents the number of calls made to the decoder required for the model $\mathcal{M}$ to process $\mathbf{x}_{clean} \oplus \delta$, and $\oplus$ is the operator used to introduce the perturbation δ into the clean sentence $\mathbf{x}_{clean}$. Perturbation of the input sample can encompass various levels : character-level (e.g., reordering, insertion, deletion, substitution with homoglyphs), token-level (e.g., synonym substitution), and structural-level (e.g., paraphrasing). To solve the optimization problem in equation (1.30), the **NMTSloth** attack comprises three stages : identifying critical tokens, mutating the input, and evaluating candidate adversarial samples.
Let $\mathcal{L}_{eos}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(p_i^{eos} + p_i^{o_i})$ represent the objective function aimed at minimizing the appearance of the EOS token. Here, o_i denotes the i-th predicted token in the output sentence, and $p_i^{o_i}$ signifies the probability of predicting o_i at the i-th position of the output sequence. The rationale behind incorporating $p_i^{o_i}$ into their loss function is to disrupt the dependency among predicted tokens, given their conditional interdependence, thereby optimizing p_i^{eos} more efficiently. Intuitively, critical tokens are input tokens whose variation can significantly impact the computed objective function. They are determined by computing an importance score represented by $g_i = \sum_j \frac{\partial \mathcal{L}_{eos}(\mathbf{x}_{clean})}{\partial \mathbf{x}_{clean}(i)^j}$, where i denotes the token's positional index in $\mathbf{x}_{clean}$ and j represents its embedding dimension's index. These selected critical tokens are then mutated using token-level, character-level, and structural-level techniques to generate a set of candidates. Finally, a latency increment coefficient is computed on the resulting candidates in order to select the samples that maximize the objective function. This attack strategy, while initially proposed for translation models, can be extended to any language model that employs a decoder-based architecture. Chen et al. (2022b) also conducted a study to evaluate image captioning models' efficiency robustness given the massive attention they received by both academia and the industry. An image captioning model typically receives an image input and uses an encoder to generate its hidden representation. Then, similarly to NMT systems, the decoder iteratively uses the generated hidden representation for an auto-regressive generation of the output sequence (the caption). This suggests that in this context, the number of decoder calls is

also linearly proportional to the length of the output sequence and correlated with the processing latency. Subsequently, the authors proposed the **NICGSlowDown** attack outlined in equation (1.31). Similarly to **NMTSloth**, the attack's objective is to maximize the number of decoder calls w.r.t the input $\hat{\mathbf{x}} + \delta$. The formulation of the optimization problem thus adheres to equation (1.30) while adapting both the pre-processing of the input and its perturbation to tackle the image modality.

$$\begin{aligned} \mathbf{x}_{sp} &= \mathcal{T}^{-1}(\hat{\mathbf{x}} + \delta^{\star}), \\ \hat{\mathbf{x}} &= \mathcal{T}(\mathbf{x}_{clean}), \\ \delta^{\star} &= \arg\max_{\delta} \ \text{Loop}_{\mathcal{M}}(\hat{\mathbf{x}} + \delta), \\ &\text{s.t} \ \ \|\delta\| < \epsilon. \end{aligned} \tag{1.31}$$

$\mathcal{T}$ represents the pre-processing step described in equation (1.20), utilized to validate the box constraints, while the constraint on the perturbation norm is enforced by introducing a penalty loss expressed in (1.32).

$$\mathcal{L}_{per} = \begin{cases} 0 \text{ if } \delta < \epsilon, \\ \|\delta - \epsilon\| \text{ otherwise.} \end{cases} \tag{1.32}$$

In practice, the objective function in equation (1.31) aims to maximize the number of decoder calls by explicitly minimizing the probability of the EOS token. Since each output token's generation is a Markovian process heavily reliant on the previously generated tokens, the authors incorporate a loss that reduces dependency among output tokens to optimize the EOS token's probability more efficiently. Concurrently, Gao et al. (2024b) proposed an energy-latency attack against VLMs. They developed the **verbose images** attack, which adapts the optimization objective of the latter attacks (Chen et al., 2022a;b; Haque et al., 2023b) to incorporate advanced sampling policies and accommodate the large scale of VLMs. In summary, the attack follows the formulation presented in equation (1.31) where the objective function directly minimizes the probability of the EOS token and breaks token dependency by maximizing the Kullback-Leibler divergence between the output probability distribution and a uniform distribution. Furthermore, to enhance the diversity of output tokens, they incorporate an objective to maximize the rank of the hidden state matrix across all generated tokens. Finally, they introduce a temporal weight adjustment algorithm for the Lagrangian coefficients to balance the multiple optimization

functions, accounting for varying convergence rates. From a practical perspective, the **NICGSlowDown** attack too is expected to work on VLMs, although this has not yet been thoroughly investigated.

Eventually, Haque et al. (2023b) proposed an adaptation of this attack against speech recognition models. The attack uses a formulation similar to that described in equation (1.30), with the distinction that the input variable represents an audio signal instead of text. Similarly to previous works (Chen et al., 2022a;b), the objective function includes minimizing the probability of the EOS token, while they propose to reduce token dependency by maximizing the probability of the second most likely token to improve the attack's efficiency. During the optimization process, a penalty term on the perturbation norm is also added to ensure that the attack is imperceptible on the resulting adversarial audio.

Black-box : This subsection delves into various inference stage black-box attacks documented in the literature.

Ilyas et al. (2018) introduced three attack settings designed to more accurately reflect real-world scenarios : the query-limited setting, the partial-information setting, and the label-only setting. The query-limited setting provides access to the predicted class probabilities $P(y|\mathbf{x})$, but no additional information about the underlying model or its gradients. In the partial-information setting, the adversary has access only to the confidence scores of the top-k predicted labels. In contrast, in the label-only setting, the adversary has access only to the top-k predicted classes, without any associated confidence scores. To address this constraint, Ilyas et al. developed an efficient adversarial attack that employs a Natural Evolution Strategie (NES) algorithm (Salimans et al., 2017) to estimate gradients, which are then used to perform a PGD attack. As for the partial-information and label only settings, the authors designed an attack that starts with an image of the target class and iteratively alternates between blending the original image and maximizing the likelihood of the target class.

Wicker et al. (2018) developed an attack that utilizes an object detection technique, such as Scale Invariant Feature Transform (SIFT), to extract features from the input image. These features are then transformed into saliency distributions, where higher probabilities are assigned to pixels that significantly influence the image's composition with respect to the Human Visual System (HVS). Their proposed attack, the **Feature-Guided Black-Box (FGBB)** attack, is formulated as a two players turn based stochastic game

that uses a Monte Carlo tree search. The first player aims to minimize the distance from an adversarial example by selecting the most relevant features. In parallel, the second player determines which pixels within those selected features to modify, as well as the manner of their manipulation.

To eliminate the reliance on gradients or their approximations, Andriushchenko et al. (2020) introduced the **Square Attack**, which instead employs a random search strategy. Additionally, they introduced two sampling distributions for the perturbation δ, which they argue are better suited to their problem formulation. At each iteration, a perturbation δ is randomly sampled such that it lies on the boundary of the L_2 or L_∞ ball before being projected into the image domain. The modifications to the image are constrained to contiguous pixels arranged in square-shaped regions (which is the basis for the name *Square* Attack). The adversarial image is then updated using the perturbation δ only if it results in a reduction of the loss. Otherwise, the perturbation is discarded.

Xiao et al. (2018) leveraged a Generative Adversarial Network (GAN) to generate adversarial examples. Their proposed attack, **AdvGAN**, involves training a distilled model, which is then used to optimize the generator for producing adversarial examples.

Chen et al. (2017a) proposed the **Zeroth Order Optimization (ZOO)** attack, which employs a coordinate descent-based method to optimize adversarial examples in a zeroth-order optimization setting. Their gradient approximation is given by the symmetric difference quotient definition (Magnus and Neudecker, 2019) presented in equation (1.33) :

$$\frac{\partial \mathcal{M}_\theta(\mathbf{x})}{\partial \mathbf{x}_i} \approx \frac{\mathcal{M}_\theta(\mathbf{x} + h\, e_i) - \mathcal{M}_\theta(\mathbf{x} - h\, e_i)}{2\, h}, \tag{1.33}$$

where h is a small constant and e_i is a standard basis vector with only the i-th component as 1.

To improve efficiency and reduce the number of queries required, Chen et al. further incorporate attack-space dimension reduction, hierarchical attack strategies, and importance sampling techniques.

In the context of energy-latency attacks, Shumailov et al. (2021) also presented an attack in the black-box setting. Assuming that transmission time is negligible and more or less constant compared to the processing time of the model, they optimized the model's input based on a latency measure, employing a Genetic Algorithm to overcome optimization challenges in the absence of gradient information. Fig. 1.5 summarizes the proposed attack strategy.

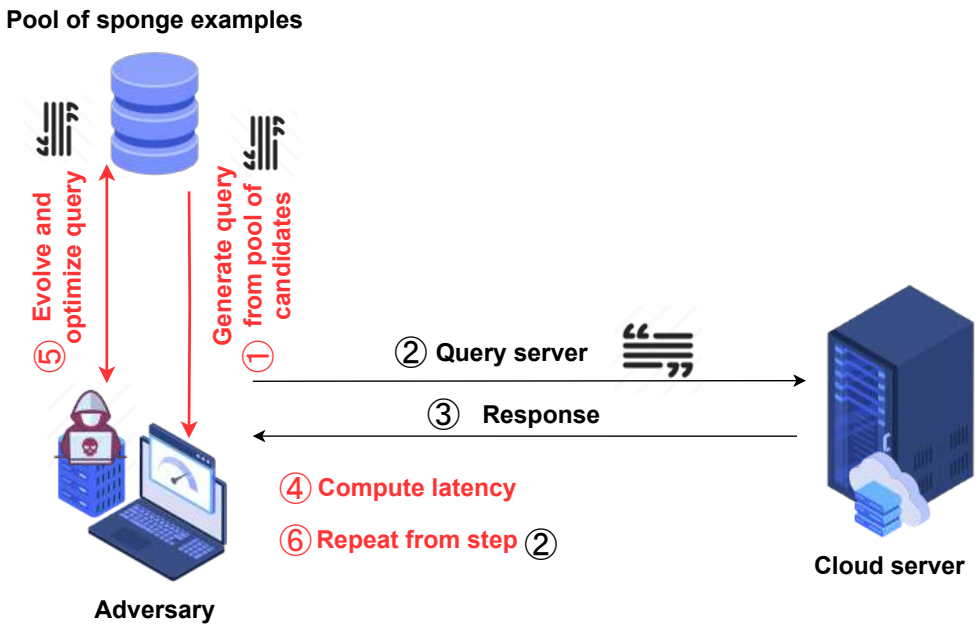


Figure 1.5 : Overview of the proposed sponge attack (Shumailov et al., 2021) in the black-box setup. The adversary uses a GA to create sponge samples. Then, they iteratively test samples against the model and generate new sponge samples by combining the most effective ones based on the measured response latency of the model.

Next, Krithivasan Krithivasan et al. (2020) investigated the transferability of their proposed white-box attack among other architectures to adapt it to the black-box context. Their experimental results showed strong transferability of the latency attack across architectures, along with a significant degradation in prediction accuracy. Consequently, they included a second stage for their attack in which they follow up the generation of the latency-based adversarial examples with a second optimization using the ZOO accuracy-based black-box attack Chen et al. (2017a) to restore the original label of the image input. Similarly, numerous studies in the literature (Baras et al., 2023; Hong et al., 2021; Liu et al., 2023; Shapira et al., 2023) leverage this transferability principle for the development of black-box attacks. These attacks are often optimized on surrogate models or through ensemble learning techniques, capitalizing on the knowledge gained from these substitute models to effectively target the actual victim model.

In contrast, Haque et al. (2022) developed **EREBA**, which uses a proxy energy estimator to generate energy-latency adversarial inputs. Their approach is based on the assumption that AdNN energy consumption is expected to follow a step-wise pattern despite variations in energy-saving mechanisms. Moreover, AdNN energy consumption rates

can still be discerned through system diagnostics in a black-box context, enabling the design of an energy emulator. This attack encompasses two different settings : an imperceptible attack similar to traditional constrained attacks and a universal adversarial attack with relaxed constraints. The latter setting facilitates triggering the worst energy consumption scenario due to the relaxation of the constraints regarding the perceptibility of the perturbation, but highly compromises the stealthiness aspect of the attack.

Müller and Quiring (2024) conducted a theoretical study analyzing the impact of input distribution on model sparsity. Their findings revealed that images with flat surfaces or uniform colors lead to the highest densities. Based on this observation, they developed two attack strategies. The first one involves selecting natural images with uniform distributions from existing datasets by measuring densities and choosing the n most uniform samples. The second strategy generates nearly uniform samples by drawing image pixels from a Gaussian distribution $\mathcal{N}(\mu, \sigma = \frac{2}{255})$, optimizing the mean μ via grid search, and setting the standard deviation σ to a non-zero value to increase the diversity of the generated samples. By leveraging this property, they generate energy-latency adversarial inputs that achieve comparable performance to optimized adversarial examples, but with significantly reduced processing time.

While previous attacks have predominantly targeted software vulnerabilities, Boutros et al. (2020) introduced **Neighbors From Hell**, a novel voltage-based attack targeting the accuracy of multi-tenant Field-Programmable Gate Array (FPGA) accelerators for DL. Their attack involves a malicious circuit that induces timing violations and faulty functionality by drawing excessive current and causing voltage drops in the chip's Power Distribution Network (PDN). Although knowledge of the employed DL accelerators is required, the attack is considered black-box as it does not necessitate further information about the specific DL model in use. While primarily designed to compromise DNN integrity, this type of attack could also manifest as a latency attack under certain conditions. If timing failures in the control logic prevent the accelerator from completing batch classification, the attack could inadvertently lead to increased latency.

Training stage

Most existing research on training stage attacks (Chen et al., 2023a; Gu et al., 2019a; Liu et al., 2020) focuses on the "partial control" attack scenario, where the attacker's access to the training process and data is limited (e.g., in distributed or federated learning). This

is typically reflected in the assumption that the adversary is limited to modifying only a fraction α of the training data.

One straightforward yet effective way to degrade a model's performance during training is to intentionally flip the labels of some or all of the training samples. In fact, Barreno et al. (2010) demonstrated that this technique can increase both false positives and false negatives in a Bayesian anti-spam classifier. Known in the literature as **label flipping**, this method forms the basis of many **poisoning** attacks. It is particularly feasible in crowd-sourcing scenarios, where an adversary may act as one of the data annotators. In practice, label flipping attacks can be divided into two categories : random label flipping, where the adversary changes the labels of randomly selected samples, and optimal label flipping. The latter setting is model-dependent and involves identifying the optimal samples to be modified. For example, Paudice et al. (2019) designed an efficient algorithm that performs optimal label flipping by formulating the poisoning attack into the bi-level optimization problem presented in equation (1.34).

$$\begin{aligned} \mathcal{C}_{po}^{*} &= \underset{\mathcal{C}_{po}=\{(\mathbf{x}_j, y_j')\in\mathcal{D}^{Train}\}}{\arg\max} \sum_{j=1}^{|\mathcal{C}_{po}|} \frac{1}{|\mathcal{D}^{val}|} \sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}^{val}} \mathcal{L}\Big(\mathcal{M}_{\theta_{po}}(\mathbf{x}_i), y_i\Big), \\ \text{s.t. } \theta_{po} &= \arg\min_{\theta} \frac{1}{|\mathcal{D}_{po}^{Train}|} \sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{po}^{Train}} \mathcal{L}\Big(\mathcal{M}_{\theta}(\mathbf{x}_i), y_i\Big), \end{aligned} \tag{1.34}$$

where $\mathcal{C}_{po}$ is a subset of the training set $\mathcal{D}^{train}$ obtained by changing the ground-truth labels of the selected inputs $\mathbf{x}_i$, with $\alpha = \frac{|\mathcal{C}_{po}|}{|\mathcal{D}^{train}|}$, $\mathcal{D}_{po}^{Train}$ denotes the resulting poisoned training set, and $\mathcal{L}$ is the the loss used for the training of the model. It is important to note that, within the equation (1.34), $\mathcal{L}$ is evaluated on the model weights θ_{po} obtained by training the model on the poisoned data set $\mathcal{D}_{po}^{Train}$ (that includes $\mathcal{C}_{po}$). In simpler terms, the adversary's goal is to identify an optimal set of samples such that, when their labels are changed, the resulting model weights maximize the loss evaluated over the validation set $\mathcal{D}^{val}$.

In practice, solving this optimization problem is NP-hard, as it requires a combinatorial search over all possible subsets of $\mathcal{C}_{po}$. To address this, Paudice et al. proposed a greedy algorithm that selects examples to flip based on their estimated impact on the validation objective the attacker seeks to maximize. Next, Muñoz-González et al. (2017) used a back-gradient optimization strategy to address the limitation of the later bi-level optimization form. Their intuition is that the optimization problem presented in equation (1.34) can be effectively approached using gradient ascent. Thus, using the chain rule, the gradient

of the optimized clause in equation (1.34) can be given for each pair $(\mathbf{x}, y) \in \mathcal{D}^{train}$ by $\nabla_{\mathbf{x}} \mathcal{L}\Big(\mathcal{M}_{\theta_{po}}(\mathbf{x}), y\Big) + \frac{\partial \theta_{po}}{\partial \mathbf{x}}^T \nabla_\theta \mathcal{L}\Big(\mathcal{M}_\theta(\mathbf{x}), y\Big)$. However, the main difficulty with using this formulation lies in computing $\frac{\partial \theta_{\text{po}}}{\partial \mathbf{x}}$, as it requires understanding how the solution of the model's training algorithm varies with respect to $\mathbf{x}$. Thus, Muñoz-González et al. propose to replace the inner learning problem with its Karush–Kuhn–Tucker (KKT) conditions, then using the new computed gradient to solve the optimization problem introduced in equation (1.34).

Although label flipping is a simple yet effective technique for degrading model performance, in more realistic scenarios, the adversary may not have direct access to the labeling process. Shafahi et al. (2018) designed a **feature collision** attack where the adversary's objective is to control the model's behavior on a specific inference instance without degrading its overall performance. To achieve this, the adversary first chooses a *target instance* $\mathbf{x}_t$ from the test set and a *base instance* $\mathbf{x}_b$. An imperceptible perturbation δ optimized using equation (1.35) is then added to $\mathbf{x}_b$, and the resulting example is injected into the training set. The attack is then considered as successful if the model mis-classifies the target instance $\mathbf{x}_t$ as the base class b.

$$\delta = \arg\min_{\delta} \|f(\mathbf{x}_b + \delta) - f(\mathbf{x}_t)\|_2^2 + \lambda \|\delta\|_2^2, \tag{1.35}$$

where f is the function that propagates the input $\mathbf{x}$ through the network to the penultimate layer of the classifier and λ is Lagrangian multiplier. The intuition behind this attack lies in the observation that the penultimate layer of a neural network encodes high-level semantic features, referred to as the feature space representation. Given the non linearity and high complexity of the function f, it can be assumed possible to find an example $\mathbf{x}_b + \delta$ which 'collides' with the target $\mathbf{x}_t$ in the feature space, while perceptually remaining close to the base instance $\mathbf{x}_b$ in input image space. If the model is fine-tuned on a training set containing poisoned samples, the initial linear decision boundary in the feature space is then expected to shift in a manner that the poisoned instance is labeled as the base class. This allows the adversary to gain a **'backdoor'** access into the base class.

One commonly noted limitation of the feature collision attack is its reliance on the attacker having access to the victim's feature extractor f. However, Tramèr et al. (2017) demonstrated that for each input sample, there exists a shared adversarial subspace for different models trained on the same dataset. This implies that it is possible to find adversarial samples that can be transferred across different feature extractor architectures.

One obvious approach would be using ensemble learning to improve the transferability of poisoned samples by optimizing a loss function over multiple models $\mathcal{M}^{(i)}$ as presented in equation (1.36). However, Zhu et al. (2019) explain that this approach significantly reduces the degrees of freedom in the optimization problem and causes the generated perturbations to exhibit clear patterns of the target class, making it difficult to find an optimal solution.

$$\mathcal{C}_{po}^{*} = \arg\min_{\mathcal{C}_{po}} \sum_{i=1}^{m} \sum_{x_p \in \mathcal{C}_{po}} \frac{\|\mathcal{M}^{(i)}(\mathbf{x}_p) - \mathcal{M}^{(i)}(\mathbf{x}_t)\|^2}{\|\mathcal{M}^{(i)}(\mathbf{x}_t)\|^2}, \tag{1.36}$$

To address this limitation, they introduce the **polytope** attack, which constructs poisoned examples that collectively surround the target image in the feature space instead. They reformulate the optimization problem for a given sample $\mathbf{x_p}$ into what follows :

$$\begin{aligned} \min_{\{\alpha^{(i)}\}, \mathcal{C}_{po}} \quad & \frac{1}{2} \sum_{i=1}^{m} \frac{\left\| \mathcal{M}^{(i)}(\mathbf{x}_p) - \sum_{p=1}^{|\mathcal{C}_{po}|} \alpha_p^{(i)} \mathcal{M}^{(i)}(\mathbf{x}_t) \right\|^2}{\|\mathcal{M}^{(i)}(\mathbf{x}_t)\|^2} \\ \text{s.t.} \quad & \sum_{p=1}^{|\mathcal{C}_{po}|} \alpha_p^{(i)} = 1, \\ & \alpha_p^{(i)} > 0 \quad \forall i, p, \end{aligned} \tag{1.37}$$

where $\mathbf{x}_p$ is obtained by adding a perturbation signal $\delta < \epsilon$ and ϵ is the perturbation budget. Equation (1.37) tries to find a set of poisoned samples $\mathcal{C}_{po}$ and a set of convex combination coefficients $\{\alpha_p^{(i)}\}$ so that $\mathbf{x}_t$ is close to the convex polytope of the poisoned samples in the feature spaces learned by all m feature extractors.

Drawing inspiration from Shumailov et al. (2021), Cinà et al. (2022) proposed the **sponge poisoning** attack that elevates the energy consumption of the resulting model. Their attack requires the modification of the objective function for a subset of the training data by incorporating a penalty on sparsity minimization. It redefines poisoning attacks in the context of energy-latency attacks, shifting the focus from tampering with training data to corrupting the optimized objective itself. Equation (1.38) encapsulates the expected optimized loss during the training of model $\mathcal{M}_\theta$. In essence, optimization is performed on the cross-entropy loss $\mathcal{L}_{ce}$ for all training samples $(\mathbf{x}, y) \in \left(X_{train}, Y\right)$, whereas the penalty term $\mathcal{L}_p$ is only calculated for samples in the poisoning set $\mathbf{x} \in \mathcal{P}$. The resulting loss $\mathcal{L}_{total}$, calculated over different batches, is then aggregated and optimized during training. This poisoning attack promotes the use of denser activations during inference, thereby

increasing energy consumption. Subsequently, Wang et al. (2023c) demonstrated the real-world applicability of the sponge poisoning attack (Cinà et al., 2022) by successfully implementing it on mobile devices.

$$\mathcal{L}_p(\mathbf{x}) = \begin{cases} \lambda \sum\limits_{a^{(l)} \in \mathcal{A}(\mathbf{x})} \|a^{(l)}\|_0 & \text{if } \mathbf{x} \in \mathcal{P} \\ 0 & \text{otherwise} , \end{cases} \tag{1.38}$$

$$\mathcal{L}_{total}(\mathbf{x}, y) = \mathcal{L}_p(\mathbf{x}) + \mathcal{L}_{ce}\Big(\mathcal{M}_\theta(\mathbf{x}), y\Big).$$

Lintelo et al. (2024) designed the **SkipSponge** attack, a poisoning attack in which the adversary directly manipulates the weights of a clean model. Their attack can summarized in the following steps.

- Selecting target layers with ReLU activations.
- Conducting a statistical analysis by hierarchically updating the biases while monitoring the resulting drops in accuracy.
- Restoring the set of weights that achieves an optimal balance between attack effectiveness and accuracy retention.

Chen and Li (2023) proposed a novel attack targeting cloud-based machine learning such as Machine Learning as a Service (MLaaS) platforms. Unlike attacks mentioned previously, which focus on compromising the model's accuracy or increasing its inference latency, this attack aims to increase training costs by altering a portion of the training data to make it more difficult for the model to learn relevant features. The core idea of their approach is based on the observation that neural networks tend to assign high confidence scores to inputs with clear and distinctive features, whereas low confidence scores suggest that the features are more ambiguous and less informative for the task. By injecting challenging samples associated with low confidence into the training data, the model is compelled to perform additional training epochs to adequately learn the more complex patterns required to reach a desired confidence or accuracy threshold. Consequently, the attack strategy consists of generating poisoned samples whose output probability distributions approximate a uniform distribution, as defined in (1.22), and incorporating them into the training set to increase the overall computational cost of training.

Building upon the concept of poisoning attacks, **backdoor attacks** represent a variant wherein adversaries embed hidden triggers into the training data to manipulate model behavior under specific conditions during model inference. In simpler words, the backdoored model behaves normally when processing clean data, and is expected to perform the adversary's chosen task when a trigger (also selected by the attacker) is present within the model's input. This type of attack can occur in various scenarios : a malicious actor may gain access to an outsourced training pipeline, the victim might inadvertently download a compromised model trained with a backdoor, or the adversary could control a device participating in federated learning. These scenarios make backdoor attacks a particularly plausible and a serious threat in real-world machine learning deployments (Gao et al., 2020). While several comprehensive surveys exist in the literature (Gao et al., 2020; Le Roux et al., 2024; Zhao et al., 2024), we highlight a selection of representative attacks to provide the reader with a foundational understanding of how these attacks operate.

Gu et al. (2019b) were the first to introduce backdoor attacks, in which a model is trained on a poisoned dataset containing image samples embedded with a trigger signal υ. These samples are further mislabeled as belonging to a target class t. Then, at inference time, the model is expected to misclassify any input containing the trigger υ as the target class t, effectively creating a backdoor within the model. Notably, this is achieved without significantly compromising the model's overall performance or accuracy.

In their study, Gu et al. employed various visual triggers, such as a yellow square, a bomb, and a flower sticker, which were overlaid on input images. However, because these triggers were visually conspicuous, subsequent works have focused on designing trigger signals that are less detectable in order to enhance the stealthiness of the attack. Notably, Chen et al. (2017b) introduced the 'pixel blending' technique defined in equation (1.39), in which the pixels of the clean image $\mathbf{x}_{clean}$ are blended with those of the trigger signal υ.

$$\mathbf{x}_{tr}(i,j) = \begin{cases} \gamma\, \upsilon(i,j) + (1-\gamma)\, \mathbf{x}_{clean}(i,j) & \text{if } m(i,j) = 1 \\ \mathbf{x}_{clean}(i,j) & \text{if } m(i,j) = 0 \end{cases} \tag{1.39}$$

Here, (i,j) denote the pixel coordinates, m is a binary mask that determines the locations within the image where the trigger is superimposed, and $\gamma \in [0,1]$ is a blending coefficient used to balance the trade-off between the imperceptibility of the trigger and the effectiveness of the backdoor injection.

Zhong et al. (2020) further investigated the use of an 'adaptive perturbation mask', which takes both the input data and a pre-trained model into consideration when optimizing the trigger signal υ. This is achieved using a UAP strategy as defined in equation (1.40) to generate the trigger on a surrogate pre-trained model $\mathcal{M}_{\hat{\theta}}$ over a subset $X^{(c)}$ of images corresponding to a given class c.

$$\upsilon = \arg\min_{\delta} \|\delta\|_2 \text{ s.t. } \forall \mathbf{x} \in X^{(c)} : \mathcal{M}_{\hat{\theta}}(\mathbf{x} + \delta) = t \tag{1.40}$$

Gao et al. argue that visible triggers typically lack sparsity and propose leveraging the sparsity of the trigger signal υ to improve its stealthiness. Their proposed **Sparse and Invisible Backdoor Attack (SIBA)** incorporates an additional constraint on the trigger sparsity $\|\upsilon\|_0 < \epsilon_s$ to the optimization problem presented in equation (1.40).

Inspired by the **DeepSLoth** attack (Hong et al., 2021), Chen et al. (2023a) devised **EfficFrog**, a backdoor attack specifically tailored for multi-exit architectures. Their attack adheres to the conventional backdoor formulation, incorporating an objective function that maximizes energy consumption when the trigger is present in the input while maintaining accuracy to ensure stealth. In practice, they achieve this by injecting the predefined trigger into a portion of the training data and altering their corresponding ground-truth labels. Energy maximization is then achieved by maximizing the number of blocks required for input processing. The attack is carried out in two stages.

1. **Trigger optimization :** as presented in Equation (1.41), which aims to generate an optimal trigger υ over the clean training set $\mathcal{D}$ within a perturbation budget ϵ (the followed strategy to achieve that is similar to the SIBA attack).

2. **Backdoor injection :** during which the model's weights are optimized to produce less confident predictions at early exits in the presence of the trigger.

$$\begin{aligned} \upsilon &= \arg\min_{\delta} \ \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{uncert}, \\ \mathcal{L}_{perc} &= \begin{cases} \|\delta - \epsilon\| \text{ if } \delta > \epsilon \\ 0 \qquad\qquad\qquad \text{otherwise,} \end{cases} \\ \mathcal{L}_{uncert} &= \sum_{(\mathbf{x},y)\in\mathcal{D}} \sum_{i=1}^{K} \| \mathcal{M}_i(\mathbf{x}+\delta), y_{target} \sim \mathcal{U} \|_2. \end{aligned} \tag{1.41}$$

$$\theta_{\mathcal{B}} = \arg\min_{\theta} \sum_{i=1}^{K} \sum_{(\mathbf{x},y)\in\mathcal{D}\setminus\mathcal{P}} \mathcal{L}_{ce}\Big(\mathcal{M}_i(\mathbf{x}), y\Big) + \sum_{(\mathbf{x},y)\in\mathcal{P}, i\neq K} \mathcal{L}_{ce}\Big(\mathcal{M}_i(\mathbf{x}+\delta^{\star}), y_{target} \sim \mathcal{U}\Big). \tag{1.42}$$

In equations (1.41) and (1.42), we slightly abuse the notation by omitting the model's parameters θ for simplicity. $\mathcal{M}_i, i = \overline{1, K}$ are the model's exit points, $\mathcal{P}$ is the poisoning set, and $\mathcal{U}$ refers to the uniform distribution.

Similarly, Huang et al. (2024) also designed a backdoor attack against early-exit DNNs using predefined triggers (e.g., adding a small patch to an image or simulating rainy conditions). Their attack involves the modification of the optimization objective within the training phase by incorporating an entropy maximization clause on the different output branches. This enables the adversary to bypass the model's early exits on trigger inputs.

1.6 Conclusion

This chapter presents a brief overview of the foundational background concepts and related works that are necessary for understanding the subsequent chapters. Although the literature offers a broad spectrum of techniques and strategies used to exploit various vulnerabilities within different DNN architectures, it remains a rapidly evolving field. New models are constantly being designed which introduces unforeseen vulnerabilities, prompting the development of novel defense strategies, which in turn are challenged by increasingly sophisticated attack methods. As a result, the landscape is characterized by an ongoing arms race between ML engineers and researchers, adversaries, and defenders. Despite the extensive body of existing work, this dynamic nature creates persistent gaps and open questions. In the following chapters, we introduce different attack strategies, designed to exploit new unseen vulnerabilities, address specific limitations identified in the current literature and contribute to this evolving research frontier.

Chapitre 2

Evaluating the Vulnerability of Deep Learning-based Blind Image Quality Assessment Methods to Adversarial Attacks

2.1 Introduction

Under the massive emergence of multimedia content, IQA has become an essential module in different transmission pipelines. Particularly, Blind Image Quality Assessment (BIQA) refers to techniques designed to evaluate the perceptual quality of an image with no access to the reference, unaltered image. On the other hand, the impressive development of ML and more specifically DL and its deployment in different fields, introduced major progress in the automation process of many human-related tasks. Being no exception, automatic IQA has seen a drastic improvement since the introduction of the ML discipline to the area, especially as models based on deep CNN and ViT architectures overrun state-of-the-art performance in computer vision. However, DL's successful deployment in different fields for automation, time, and cost optimization purposes stirred up soon enough the attention of many brilliant minds from diverse scientific communities including adversaries. This gave rise to serious security vulnerabilities that could be exploited for malicious purposes, and IQA metrics are affected as well by the latter systems' vulnerabilities.

The gravity of attacks targeting IQA metrics spans a wide operational spectrum, ranging from minor inconveniences for end users to severe, potentially life-threatening consequences in critical systems. For instance, in content-sharing applications such as social media, deceiving a metric into assigning high-quality scores to low-quality visual content can adversely affect the end users' Quality of Experience (QoE). Conversely, incorrect predictions of low-quality scores for high-quality images may unnecessarily activate enhancement mechanisms, consequently increasing energy consumption and latency, which reduces user engagement, particularly on streaming platforms. Furthermore, biased quality assessments of camera feeds used by downstream applications can severely compromise the reliability of the entire processing pipeline. A prominent example is video surveillance, where inaccurate quality evaluations could lead to the storage of unclear footage, limiting its practical utility. Most critically, an IQA metric's failure to identify artifacts in medical imaging may result in misdiagnoses, significantly endangering patient safety and well-being.

In this chapter, we investigate the robustness of various CNN backbones to adversarial attacks within the context of an IQA regression task. Our primary objective is to adapt state-of-the-art adversarial attacks, originally developed for image classification, to IQA models in both targeted and untargeted scenarios. Then, we examine the potential corre-

lation between the perceptibility of adversarial perturbations and the effectiveness of the attacks.

2.2 Related work

In this section, we provide a brief overview of the literature on BIQA and adversarial attacks targeting BIQA models.

2.2.1 Blind image quality assessment models

IQA in general and BIQA in particular has emerged as an active and dynamic research area, leading to the development of numerous metrics over the years. These metrics span a diverse range of methodologies, from handcrafted feature-based approaches (including distortion-specific and Natural Scene Statistics (NSS)-based methods) to advanced ML and DL techniques, such as CNN and ViT-based models.

Many widely used CNN architectures, e.g., VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2015), and Inception (Szegedy et al., 2015b) architectures have been initially proposed for the image classification task, and were pre-trained on a large amount of data such as ImageNet (Krizhevsky et al., 2017) for this purpose. However, a desirable property of deep models enables their fine-tuning on similar tasks, such as IQA (Zhang et al., 2018). Some of the first works that introduced CNNs to the IQA are (Bosse et al., 2016; Kang et al., 2014), and since then, many works have been inspired by this concept (Bianco et al., 2018; Hosu et al., 2020). Consequently, recent successful models rely on transfer learning and yield outstanding performance on publicly available IQA datasets (Hosu et al., 2020; Lin et al., 2019; Ponomarenko et al., 2013). These solutions mainly differ in the choice of the DNN backbone and the aggregation of their output features.

More recently, there has been a growing trend toward leveraging foundation models as versatile, general-purpose tools capable of handling a wide range of tasks, including those in computer vision. This includes the use of LLMs and VLMs for IQA. CLIP-IQA (Wang et al., 2023a) for example leverages quality-aware text prompts to evaluate the quality of the image. Q-Mamba (Guan et al., 2024) leverages a Structured State Space Model (SSM)/Mamba model in combination with a StylePrompt tuning paradigm to enhance its perception capability for the IQA task. The study by Wu et al. (2025) investigates and demonstrates the effectiveness of chain-of-thought prompting for improving these models'

performance in IQA tasks. Co-Instruct (Wu et al., 2024) provides a comparison framework with an extension to open-ended settings that include open-range questions and detailed reasoning. Finally, Q-Align (Wu et al., 2023) even made the first attempt of transferring qualitative rating levels into quantitative quality scores.

2.2.2 Adversarial attacks on blind image quality assessment models

Many works have investigated the impact of adversarial attacks on BIQA models in the literature (Korhonen and You, 2022; Shumitskaya et al., 2022; Zhang et al., 2022b). Korhonen and You (2022) examined the vulnerability of BIQA models to targeted adversarial attacks. The motivation behind their proposed attack was to negatively impact user QoE by manipulating quality assessment mechanisms, causing low-quality images to receive artificially inflated quality scores through imperceptible image perturbations. Their approach relies on gradient-based optimization techniques for crafting the perturbations, combined with a spatial activity map to minimize its perceptibility. Similarly, Shumitskaya et al. (2022) designed a universal perturbation tailored to a specific BIQA model. In their framework, the UAP was treated as a trainable variable that was updated throughout a "training process". Their findings revealed the presence of a particular direction of gradients that consistently increases the predicted quality score. Then, to reduce the perceptibility of the generated perturbation, they incorporated a Contrast Sensitivity Function (CSF) as a weighting map for the image perturbation. Zhang et al. (2022b) reformulated the adversarial attack problem in a Lagrangian framework by interchanging the constraint and the objective. Their goal was to maximize the difference in predicted quality scores between clean and perturbed images, while ensuring that the perturbations remained imperceptible. Their main contribution lies in using a Full Reference (FR) metric to estimate the perceptual distance between the original and adversarial images, effectively modeling the Just Noticeable Difference (JND) properties of the HVS. Shumitskaya et al. (2023) designed a faster attack by utilizing a U-Net encoder-decoder architecture, eliminating the need for an iterative gradient method to optimize the perturbation. Ran et al. (2025) designed a more effective "Bi-directional" loss function, which biases the adversarial image predicted quality score towards the opposite direction of its actual perceptual quality. Additionally, they used a patch-based optimization strategy in order to explore the solution space more effectively and identify more optimal solutions.

Yang et al. (2024b) proposed a correlation-error-based attack framework that integrates both the predicted image quality scores and the correlation within an image set. Their approach naturally incorporates an image-level Mean Squared Error (MSE) loss and a Spearman's Rank Correlation Coefficient (SRCC) loss computed at the set-level within a Lagrangian framework. Yang et al. (2024a) further proposed a black-box attack framework that leverages a score-boundary technique and utilizes an edge map as a JND surrogate to make the added adversarial perturbations less perceptible to the HVS. Shumitskaya et al. (2024) proposed a fast and less perceptible adversarial attack that utilizes frequency-based weighting maps to control the visibility of perturbations, thereby maintaining high perceptual quality. Their attack leverages FGM as a one-step algorithm to ensure fast and efficient execution, making it extendable to video quality assessment metrics. Next, Leonenkova et al. (2024) proposed to optimize a physically printable adversarial patch designed to increase the predicted quality scores of images and videos that capture it. The trained patch can be deployed in physical attacks by printing it and displaying it within the captured scene.

Table 2.1 summarizes the proposed adversarial attacks against No reference IQA models in the literature.

Table 2.1 : Summary of the proposed adversarial attacks against BIQA models in the literature. We refer to the **white-box** and **black-box** settings as **'W'** and **'B'**, respectively. Similarly, we denote the **targeted** and **untargeted** settings by **'T'** and **'U'**, respectively.

Attack	Venue	Perturbation reduction approach	Attack settings	Code availibility
Korhonen and You (2022)	QoEVMA'22	Space activity map as perturbation	B + T	✓
Shumitskaya et al. (2022)	BMVC 2022	CSF mask	W + T	✓
Zhang et al. (2022b)	NeurIPS 2022	SSIM optim as a JND measure	W + U	✓
Ours	EUVIP 2023	N/A	W + T and U	✓
Shumitskaya et al. (2023)	ICLR 2023	$L_\infty < \epsilon$	W + T	✓
SMA (Yang et al., 2024b)	CoRR 2024	$L_\infty < \epsilon$	W + U	✗
Yang et al. (2024a)	TCSVT 2024	JND using edge maps	B + U	✗
IOI (Shumitskaya et al., 2024)	ICML 2024	Frequency map	W + T	✓
Ti-patch (Leonenkova et al., 2024)	CoRR 2024	JND using edge maps	W + T	✓
Ran et al. (2025)	Expert Syst. Appl. 2025	$L_\infty < \epsilon$	B + U	✗

These studies collectively expose the vulnerability of various BIQA methods, including some based on CNN architectures. However, there has been limited investigation into the robustness of different CNN-based models (Do certain architectural properties exhibit greater robustness than others ?). Moreover, most existing research has predominantly focused on targeted attacks (Korhonen and You, 2022; Shumitskaya et al., 2022) employing imperceptible perturbations to increase the predicted quality score. At the time of writing

the paper on which this chapter is based, to the best of our knowledge, there had been minimal research into untargeted attacks on CNN-based IQA models which underscores the need for further investigations in this area. Furthermore, earlier studies (Korhonen and You, 2022; Shumitskaya et al., 2022; Zhang et al., 2022b) have addressed adversarial attacks on BIQA models by constraining the visibility of perturbations. As a result, they typically assume the presence of a third party responsible for verifying the integrity of images prior to assessment by the IQA metric. Specifically, a human third party must check whether the images have been altered. This assumption is both time-consuming and impractical. Finally, it is intuitive that, in this context, the visibility of adversarial perturbations is closely linked to the effectiveness of the attack but no further research has been conducted in this direction.

2.3 Proposed attack

In this section, we present the framework of our attack. It adapts efficient and widely-used adversarial attacks in the field of image classification to the regression context of the BIQA task in both targeted and untargeted scenarios.

We investigate how well CNN-based IQA models can withstand these attacks and maintain their ability to predict accurate image quality scores. In our proposed attack, we deliberately avoid imposing constraints on the perceptibility of the perturbations introduced during the generation of the adversarial example. Relaxing these constraints enables us to explore the relationship between the perceptibility of adversarial attacks and their success rates within the IQA context. Moreover, it enables us to generate attacks capable of degrading the actual image quality while simultaneously increasing the predicted quality scores.

2.3.1 Problem statement

We find it more interesting to investigate the robustness of current state-of-the-art solutions against evasion attacks. The objective of our adversarial attack is to craft an adversarial image $\mathbf{x}_{adv}$ that satisfies $\mathcal{M}_\theta(\mathbf{x}_{adv}) = y_{adv}$, where $\mathcal{M}_\theta$ is the BIQA model parameterized by the weights θ, and y_{adv} is a score that differs from both the ground truth score and the predicted score on the clean image $y_{pred} = \mathcal{M}_\theta(\mathbf{x}_{clean})$. The objective

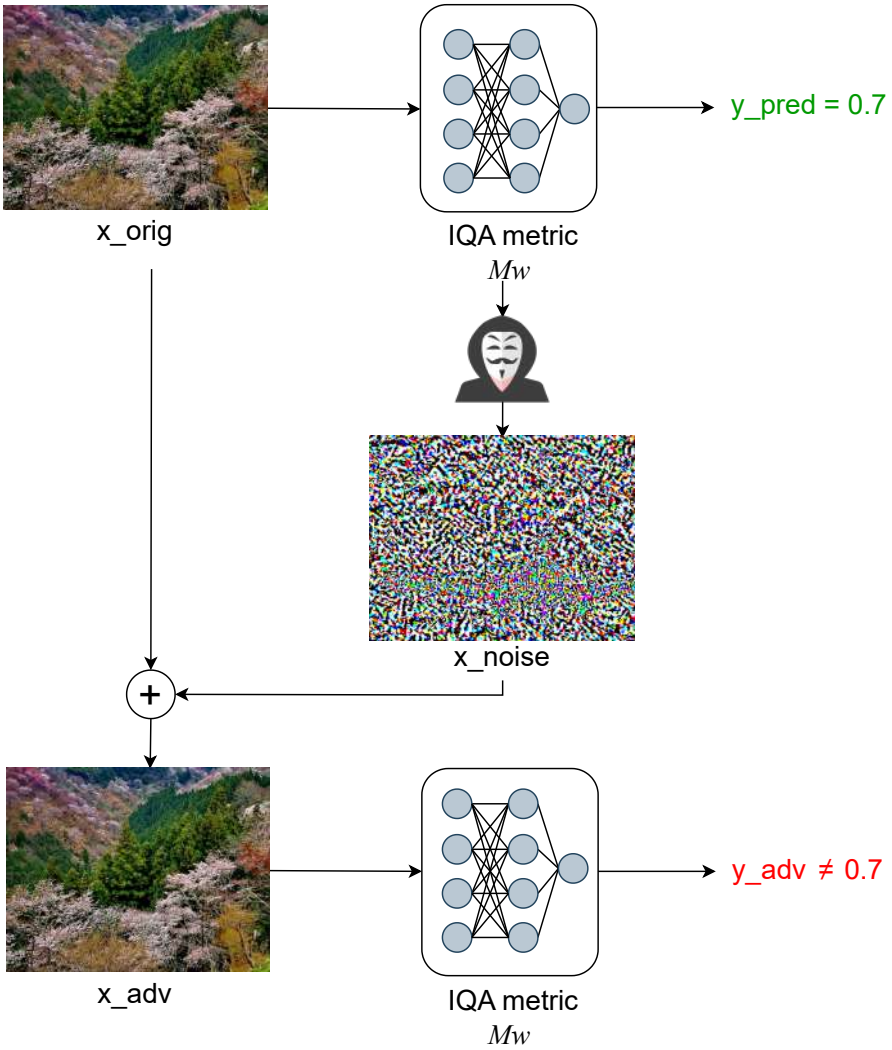


Figure 2.1 : Overview of the proposed framework.

of the adversarial attack may vary depending on the context and the attacker's goals (availability of the ground truth score, targeted or untargeted scenario, etc.).

Figure 2.1 presents our proposed framework. In this approach, we assume that the adversary has a white-box access to the BIQA model, meaning they can leverage the model's gradients to craft adversarial examples. Furthermore, our work considers both targeted and untargeted attack scenarios.

2.3.2 Attack design

In the targeted scenario, the adversary aims to mislead the model into predicting a high-quality (or low-quality) scores, regardless of the perceptual image quality. To achieve this objective, we adapt the optimization procedure from previous attacks (Korhonen and You, 2022; Shumitskaya et al., 2022), reformulating it as an unconstrained problem. Specifically, we relax the original optimization into the following form :

$$\mathbf{x}_{adv} = \arg\min_{\mathbf{x}} \left[y_{target} - \mathcal{M}_\theta(\mathbf{x})\right]^2 \;\; s.t\; \mathbf{x}_{adv} \in \mathbb{I}, \tag{2.1}$$

where $\mathbb{I}$ denotes the image space and y_{target} refers to the maximum achievable quality score by the model. In other words, the objective is to find a solution $\mathbf{x}_{adv}$ that minimizes the distance between the predicted score $\mathcal{M}_\theta(\mathbf{x}_{adv})$ and the target score y_{target}. This specific case, where y_{target} is set to the maximum achievable quality score, represents just one instance of a targeted attack. The same optimization procedure can be applied to guide the model toward any desired target score y_{target}.

On the other hand, in the untargeted scenario, the adversary aims to reduce the accuracy of the IQA model without necessarily deviating it to predict a specific score. For example, this can be achieved by deceiving the model to predict good-quality scores for poor-quality images and inversely. In our attempt to implement this attack, we shift the problem into the following optimization problem :

$$\mathbf{x}_{adv} = \arg\max_{\mathbf{x}} \left[\mathcal{M}_\theta(\mathbf{x}_{clean}) - \mathcal{M}_\theta(\mathbf{x})\right]^2 \;\; s.t\; \mathbf{x}_{adv} \in \mathbb{I}, \tag{2.2}$$

where $\mathbf{x}_{clean}$ refers to the original, clean image (before the attack). However, this problem formulation introduces a challenge when generating $\mathbf{x}_{adv}$. Gradient-based optimization methods rely on the gradient of the loss function w.r.t the input image to optimize the adversarial example. In image classification tasks, the loss is typically computed as the difference between the one-hot encoded vector of the label class for $\mathbf{x}_{clean}$ and the model's predicted probability vector. Directly applying this approach to a regression problem results in a zero loss, effectively nullifying the gradient and preventing the attack from succeeding. One possible solution to this challenge is to consider the Mean Objective Scores (MOSs) as analogous to the one-hot encoded label vectors used in classification tasks. However, this approach assumes access to the MOS during the attack, which is rarely the case. To address this limitation, we assume that the adversary can estimate

the MOS using accessible prior performance information, such as the target model's Root Mean Squared Error (RMSE). We recall the formula for the RMSE as follows :

$$\text{RMSE} = \sqrt{\mathbb{E}[(Y_{mos}^{(\text{test})} - Y_{pred}^{(test)})^2]}, \tag{2.3}$$

where $\mathbb{E}$ denotes the expectation operator, $Y_{mos}^{(\text{test})}$ is a vector corresponding to the ground-truth scores, and $Y_{pred}^{(\text{test})}$ a vector corresponding to the predicted scores on the test image set. In contrast to the RMSE, the standard deviation σ_S of a set of points S with a mean μ_S can be computed as follows :

$$\sigma_S = \sqrt{\mathbb{E}[(S - \mu_S)^2]} \tag{2.4}$$

If S follows a normal distribution $\mathcal{N}(\mu_S, \sigma_S^2)$, then any sample $s \in S$ falls within the range $[\mu_S - 3\,\sigma_S, \mu_S + 3\,\sigma_S]$, and more specifically within $[\mu_S - \sigma_S, \mu_S + \sigma_S]$ with high confidence.

By analogy, given $Y_{pred}(i)$, the predicted score of a sample i, we assume that the ground-truth value $Y_{mos}^{(\text{test})}(i)$ lies within the interval $[Y_{pred}(i) - 3 \times \text{RMSE}\ ,\ Y_{pred}(i) + 3 \times \text{RMSE}]$, and we attempt to estimate a value $\hat{Y}_{mos}^{(\text{test})}(i)$ that is close to $Y_{mos}^{(\text{test})}(i)$. As a result, equation (4.7) can be rewritten for a given image $\mathbf{x}$ as :

$$\mathbf{x}_{adv} = \arg\max_{\mathbf{x}} \, [\hat{y}_{mos}^{(\text{test})} - \mathcal{M}_\theta(\mathbf{x})]^2 \quad s.t\ \mathbf{x}_{adv} \in \mathbb{I}, \tag{2.5}$$

with $\hat{y}_{mos}^{(\text{test})}$ denoting the estimated MOS of the initial clean image. To estimate the MOS value of a given image $\mathbf{x}_{clean}$, we draw 10 samples from the distribution $\mathcal{N}(\mathcal{M}_\theta(\mathbf{x}_{clean}), \text{RMSE}^2)$ and compute their average. This approach is motivated by the similarity between the RMSE and standard deviation formulas, as discussed earlier.

2.4 Experiments and results

We showed in the previous section how adversarial attacks can be leveraged against BIQA regression models. In this section, we present more general results on the robustness of CNN-based BIQA models to such attacks and further investigate the influence of the perturbation's visibility on the attack performance.

2.4.1 Experimental Setup

We first selected three commonly used CNN architectures for both computer vision and IQA : ResNet50 (He et al., 2015), VGG16 (Simonyan and Zisserman, 2014), and InceptionV3 (Szegedy et al., 2015b). However, we assume that similar conclusions extend to other CNNs (Papernot et al., 2016a). Then, we employed transfer learning to adapt these pre-trained models to the BIQA task by freezing the weights of the CNN backbones to serve as feature extractors and retraining the Fully Connected (FC) layers for the new task. The training was performed using default hyperparameters, specifically, two hidden layers with 1024 units each, a learning rate of 10^{-3}, and a batch size of $B = 16$. Subsequently, we fine-tuned the models on two widely used IQA datasets : 1) TID2013 (Ponomarenko et al., 2013) which contains synthetically distorted images, and 2) Koniq-10k (Hosu et al., 2020), composed of authentically distorted images. The initial performance of the three backbones on both datasets, without any adversarial attack, is reported in Table 2.2. It is important to note that our primary objective is not to achieve the highest possible predictive performance but rather to obtain models that are sufficiently accurate without overfitting or underfitting the training data.

Table 2.2 : Initial performance of the CNN-based IQA models on the TID2013 and Koniq-10k datasets.

Model	Dataset	SRCC	PLCC	KRCC	RMSE
ResNet50	Koniq-10k	0.825	0.860	0.635	0.209
VGG16		0.813	0.853	0.622	0.239
InceptionV3		0.707	0.769	0.520	0.289
ResNet50	TID2013	0.942	0.948	0.789	0.332
VGG16		0.930	0.943	0.774	0.347
InceptionV3		0.875	0.911	0.695	0.386

In our experiments, we considered three widely used adversarial attacks : Fast Gradient Method (FGM) (Goodfellow et al., 2014), BIM (Kurakin et al., 2018) and PGD (Madry et al., 2017) with varying perturbation degrees : $\epsilon \in \{0.001, 0.01, 0.1, 1\}$, and we fix the number of iterations to $n_{iter} = \{2, 5, 10\}$ for iterative attacks. We note that the specific choice of attack is not critical, as the primary objective of this work is to adapt adversarial attack strategies to the BIQA task in both targeted and untargeted settings. We used the CleverHans library (Papernot et al., 2018) as the basis for our attack implementations, with further modifications to accommodate the regression context of our study. The performance of different attacked CNN-based BIQA models was compared to the original, non-attacked models in terms of Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (SRCC), Kendall's Rank Correlation Coefficient (KRCC), and Root Mean Squared Error (RMSE). The goal of the attack is to degrade the model's predictive performance, as reflected by decreased correlation scores and increased RMSE.

We also compared the similarity between the adversarial image $\mathbf{x}_{adv}$ and the original image $\mathbf{x}_{clean}$ across different attack configurations, providing insight into the visibility and perceptibility of the perturbations and on the potential correlation between this visibility and the attack efficiency. To this end, we employed the FR metric Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) to compute the perceptual similarity score between $\mathbf{x}_{adv}$ and $\mathbf{x}_{clean}$. LPIPS has been shown to be effective in capturing complex and subtle perceptual differences between images, making it thus more correlated with human perceptual judgments. A lower value of the LPIPS score indicates a lower visibility of the adversarial perturbation compared to the original image.

2.4.2 Results and Analysis

Tables 2.3, 2.5, and 2.7 summarize the performance of the IQA models under targeted FGM, BIM, and PGD attacks, respectively. Similarly, Tables 2.4, 2.6, and 2.8 present the results for the untargeted versions of these attacks. Each table corresponds to a specific attack type and reports model performance across different values of ϵ and varying numbers of attack iterations.

Table 2.3 : Performance scores of the attacked CNN-based IQA models under **targeted** FGM attacks, evaluated across varying perturbation intensities (ϵ).

Dataset	CNN	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
		$\epsilon = 0.001$					$\epsilon = 0.01$				
Koniq-10k	ResNet5	0.737	0.799	0.549	0.199	0.000	0.712	0.771	0.521	0.256	0.061
	VGG16	0.731	0.767	0.538	0.187	0.000	0.718	0.730	0.523	0.269	0.061
	InceptionV3	0.600	0.686	0.427	0.247	0.000	0.613	0.696	0.437	0.250	0.057
TID2013	ResNet5	0.863	0.878	0.677	0.814	0.313	0.709	0.764	0.526	1.423	0.355
	VGG16	0.862	0.884	0.675	0.731	0.313	0.701	0.776	0.515	1.448	0.351
	InceptionV3	0.785	0.865	0.597	0.510	0.314	0.621	0.760	0.449	0.682	0.355
		$\epsilon = 0.1$					$\epsilon = 1$				
Koniq-10k	ResNet5	0.409	0.424	0.283	0.374	0.771	0.142	0.148	0.093	0.175	1.328
	VGG16	0.627	0.644	0.445	0.321	0.739	-0.211	0.185	-0.143	0.074	1.341
	InceptionV3	0.530	0.553	0.371	0.454	0.709	-0.137	0.123	-0.091	0.198	1.180
TID2013	ResNet5	0.075	0.248	0.044	0.454	0.849	-0.091	0.290	-0.057	1.455	1.256
	VGG16	0.124	0.231	0.075	0.634	0.829	0.084	0.250	0.051	0.480	1.278
	InceptionV3	0.217	0.273	0.146	0.416	0.812	-0.070	0.095	-0.047	0.101	1.138

Table 2.4 : Performance scores of the attacked CNN-based IQA models under **untargeted** FGM attacks, evaluated across varying perturbation intensities (ϵ).

Dataset	CNN	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
		$\epsilon = 0.001$					$\epsilon = 0.01$				
Koniq-10k	ResNet5	0.501	0.556	0.354	0.494	0.000	0.404	0.409	0.285	0.658	0.063
	VGG16	0.449	0.494	0.321	0.639	0.000	0.361	0.329	0.257	0.818	0.064
	InceptionV3	0.415	0.462	0.288	0.597	0.000	0.348	0.383	0.239	0.632	0.058
TID2013	ResNet5	0.453	0.424	0.339	2.071	0.000	0.308	0.282	0.228	3.050	0.044
	VGG16	0.496	0.462	0.372	2.085	0.000	0.266	0.264	0.194	3.526	0.040
	InceptionV3	0.430	0.428	0.312	1.432	0.000	0.338	0.299	0.238	1.640	0.042
		$\epsilon = 0.1$					$\epsilon = 1$				
Koniq-10k	ResNet5	0.294	0.315	0.202	0.402	0.774	0.076	0.086	0.050	0.173	1.327
	VGG16	0.479	0.496	0.329	0.385	0.749	-0.236	0.201	-0.159	0.074	1.345
	InceptionV3	0.474	0.480	0.328	0.472	0.710	-0.138	0.131	-0.091	0.198	1.180
TID2013	ResNet5	0.130	0.226	0.085	0.454	0.719	-0.096	0.292	-0.060	1.429	1.295
	VGG16	0.107	0.126	0.068	0.714	0.692	0.049	0.224	0.027	0.475	1.317
	InceptionV3	0.299	0.335	0.204	0.426	0.687	-0.012	0.043	-0.008	0.103	1.161

Table 2.5 : Performance scores of the attacked CNN-based IQA models under **targeted** BIM attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}).

Dataset	CNN	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
		$\epsilon = 0.001, n_{iter} = 10$					$\epsilon = 0.01, n_{iter} = 10$				
	ResNet5	0.186	0.286	0.125	0.325	0.006	0.066	0.083	0.044	0.540	0.294
Koniq-10k	VGG16	0.052	-0.021	0.035	0.428	0.005	-0.011	0.030	-0.008	1.055	0.290
	InceptionV3	0.296	0.447	0.200	0.391	0.006	0.333	0.361	0.225	0.372	0.275
	ResNet5	-0.020	0.413	-0.016	1.360	0.320	0.129	0.131	0.086	2.993	0.494
TID2013	VGG16	-0.146	0.189	-0.094	1.232	0.319	0.051	0.081	0.033	3.908	0.486
	InceptionV3	0.262	0.489	0.180	1.182	0.320	-0.032	0.079	-0.020	2.047	0.484
		$\epsilon = 0.1, n_{iter} = 10$					$\epsilon = 1, n_{iter} = 10$				
	ResNet5	0.119	0.147	0.078	0.844	1.178	-0.100	0.100	-0.066	1.127	1.332
Koniq-10k	VGG16	0.005	0.022	0.003	1.152	1.234	0.059	0.076	0.039	1.033	1.376
	InceptionV3	0.086	0.091	0.057	0.203	1.184	-0.091	0.077	-0.060	0.178	1.308
	ResNet5	-0.088	0.118	-0.061	3.603	1.160	0.022	0.134	0.015	13.866	1.212
TID2013	VGG16	0.083	0.118	0.056	5.444	1.195	0.012	0.060	0.007	25.786	1.312
	InceptionV3	-0.020	0.071	-0.012	1.108	1.117	0.002	0.041	0.002	2.385	1.258
		$\epsilon = 0.001, n_{iter} = 5$					$\epsilon = 0.01, n_{iter} = 5$				
	ResNet5	0.386	0.454	0.265	0.277	0.002	0.403	0.386	0.280	0.381	0.202
Koniq-10k	VGG16	0.304	0.328	0.208	0.521	0.002	0.000	0.031	0.001	1.165	0.205
	InceptionV3	0.388	0.512	0.266	0.292	0.003	0.413	0.422	0.280	0.270	0.194
	ResNet5	0.255	0.351	0.180	1.579	0.317	0.096	0.143	0.063	3.057	0.434
TID2013	VGG16	0.286	0.317	0.207	1.499	0.316	-0.048	0.139	-0.027	4.872	0.429
	InceptionV3	0.487	0.667	0.345	0.967	0.317	0.261	0.325	0.181	1.554	0.429
		$\epsilon = 0.1, n_{iter} = 5$					$\epsilon = 1, n_{iter} = 5$				
	ResNet5	0.338	0.264	0.228	0.644	1.088	-0.087	0.074	-0.057	1.187	1.332
Koniq-10k	VGG16	0.021	0.018	0.014	0.475	1.135	-0.047	0.035	-0.031	0.411	1.372
	InceptionV3	0.079	0.092	0.053	0.188	1.088	-0.008	0.050	-0.005	0.138	1.291
	ResNet5	-0.015	0.075	-0.010	3.626	1.086	0.009	0.082	0.006	2.285	1.207
TID2013	VGG16	0.111	0.187	0.072	3.790	1.113	0.051	0.131	0.033	15.526	1.303
	InceptionV3	-0.017	0.054	-0.011	0.796	1.052	0.061	0.062	0.041	0.373	1.239
		$\epsilon = 0.001, n_{iter} = 2$					$\epsilon = 0.01, n_{iter} = 2$				
	ResNet5	0.619	0.699	0.447	0.212	0.001	0.633	0.668	0.453	0.279	0.112
Koniq-10k	VGG16	0.742	0.752	0.548	0.429	0.001	0.419	0.427	0.287	0.433	0.116
	InceptionV3	0.518	0.617	0.363	0.238	0.001	0.422	0.519	0.292	0.238	0.108
	ResNet5	0.751	0.782	0.560	1.401	0.314	0.110	0.269	0.091	3.210	0.380
TID2013	VGG16	0.783	0.810	0.586	1.222	0.314	-0.074	0.206	-0.026	4.710	0.376
	InceptionV3	0.707	0.801	0.523	0.645	0.314	0.474	0.606	0.334	0.988	0.379
		$\epsilon = 0.1, n_{iter} = 2$					$\epsilon = 1, n_{iter} = 2$				
Koniq-10k	ResNet5	0.447	0.411	0.307	0.412	0.919	0.101	0.119	0.067	0.265	1.340
	VGG16	0.235	0.242	0.158	0.323	0.944	-0.003	0.062	-0.003	0.109	1.356
	InceptionV3	0.388	0.392	0.264	0.218	0.881	-0.175	0.193	-0.117	0.158	1.245
	ResNet5	0.113	0.187	0.074	1.160	0.961	-0.051	0.247	-0.032	2.136	1.222
TID2013	VGG16	0.154	0.242	0.099	1.688	0.959	-0.102	0.182	-0.069	1.218	1.285
	InceptionV3	0.151	0.210	0.100	0.583	0.926	-0.047	0.075	-0.031	0.180	1.195

Table 2.6 : Performance scores of the attacked CNN-based IQA models under **untargeted** BIM attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}).

Dataset	CNN	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
		$\epsilon = 0.001, n_{iter} = 10$					$\epsilon = 0.01, n_{iter} = 10$				
	ResNet5	0.198	0.142	0.137	2.031	0.005	0.088	0.063	0.061	2.617	0.269
Koniq-10k	VGG16	0.157	0.080	0.110	2.722	0.006	0.047	0.070	0.035	6.042	0.255
	InceptionV3	0.215	0.203	0.145	1.734	0.006	0.044	-0.014	0.031	1.835	0.273
	ResNet5	0.144	0.118	0.103	10.539	0.005	0.019	0.073	0.015	15.815	0.198
TID2013	VGG16	0.158	0.182	0.119	11.739	0.004	0.057	0.094	0.042	35.436	0.160
	InceptionV3	0.253	0.161	0.178	5.235	0.004	0.057	0.119	0.041	6.507	0.203
		$\epsilon = 0.1, n_{iter} = 10$					$\epsilon = 1, n_{iter} = 10$				
	ResNet5	-0.109	0.214	-0.082	1.814	1.109	-0.015	0.282	-0.008	6.751	1.303
Koniq-10k	VGG16	-0.112	0.162	-0.078	15.413	1.202	-0.201	0.288	-0.136	22.329	1.294
	InceptionV3	-0.211	0.333	-0.141	0.959	1.184	-0.463	0.474	-0.313	0.374	1.307
	ResNet5	-0.429	0.553	-0.337	7.335	1.156	0.046	0.164	0.031	22.722	1.280
TID2013	VGG16	-0.499	0.484	-0.379	46.853	1.109	-0.240	0.335	-0.164	173.622	1.268
	InceptionV3	-0.233	0.501	-0.159	1.703	1.168	-0.138	0.534	-0.093	1.409	1.269
		$\epsilon = 0.001, n_{iter} = 5$					$\epsilon = 0.01, n_{iter} = 5$				
	ResNet5	0.250	0.207	0.174	1.371	0.002	0.161	0.093	0.110	1.857	0.200
Koniq-10k	VGG16	0.253	0.152	0.178	1.862	0.003	0.139	0.114	0.097	3.117	0.198
	InceptionV3	0.239	0.240	0.164	1.268	0.003	0.110	0.069	0.074	1.269	0.194
	ResNet5	0.306	0.223	0.211	6.735	0.002	0.142	0.065	0.098	9.322	0.145
TID2013	VGG16	0.177	0.176	0.133	7.157	0.002	0.086	0.105	0.061	16.352	0.126
	InceptionV3	0.219	0.170	0.156	3.641	0.002	0.184	0.114	0.126	4.308	0.143
		$\epsilon = 0.1, n_{iter} = 5$					$\epsilon = 1, n_{iter} = 5$				
	ResNet5	-0.109	0.177	-0.076	1.227	1.063	-0.128	0.232	-0.085	2.236	1.329
Koniq-10k	VGG16	-0.095	0.090	-0.061	3.768	1.126	-0.318	0.406	-0.216	5.891	1.347
	InceptionV3	-0.149	0.291	-0.099	0.780	1.090	-0.480	0.504	-0.321	0.291	1.290
	ResNet5	-0.212	0.479	-0.139	2.966	1.107	0.041	0.134	0.028	7.606	1.298
TID2013	VGG16	-0.290	0.480	-0.195	10.069	1.076	-0.112	0.244	-0.083	28.690	1.302
	InceptionV3	-0.202	0.505	-0.135	0.931	1.078	-0.170	0.588	-0.114	0.259	1.255
		$\epsilon = 0.001, n_{iter} = 2$					$\epsilon = 0.01, n_{iter} = 2$				
	ResNet5	0.369	0.370	0.267	0.782	0.001	0.286	0.223	0.200	1.053	0.114
Koniq-10k	VGG16	0.316	0.278	0.232	1.025	0.001	0.285	0.197	0.200	1.390	0.114
	InceptionV3	0.332	0.342	0.231	0.834	0.001	0.219	0.198	0.149	0.845	0.109
	ResNet5	0.378	0.329	0.273	3.511	0.001	0.289	0.251	0.200	4.940	0.078
TID2013	VGG16	0.373	0.313	0.268	3.630	0.001	0.265	0.222	0.187	6.797	0.072
	InceptionV3	0.344	0.313	0.252	2.167	0.001	0.249	0.201	0.173	2.391	0.076
		$\epsilon = 0.1, n_{iter} = 2$					$\epsilon = 1, n_{iter} = 2$				
Koniq-10k	ResNet5	-0.010	0.091	-0.004	0.637	0.930	-0.066	0.233	-0.044	0.582	1.336
	VGG16	0.034	0.089	0.033	0.886	0.939	-0.295	0.408	-0.196	0.295	1.356
	InceptionV3	0.076	0.072	0.051	0.577	0.888	-0.357	0.402	-0.239	0.163	1.244
	ResNet5	-0.129	0.307	-0.084	0.814	0.903	0.009	0.100	0.009	3.860	1.288
TID2013	VGG16	-0.256	0.374	-0.166	1.939	0.888	-0.415	0.525	-0.281	2.667	1.314
	InceptionV3	-0.095	0.255	-0.062	0.351	0.870	-0.172	0.412	-0.116	0.133	1.217

Table 2.7 : Performance scores of the attacked CNN-based IQA models under **targeted** PGD attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}).

Dataset	CNN	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
		$\epsilon = 0.001, n_{iter} = 10$					$\epsilon = 0.01, n_{iter} = 10$				
	ResNet5	0.612	0.662	0.440	0.200	0.000	0.167	0.312	0.112	0.344	0.009
Koniq-10k	VGG16	0.589	0.631	0.417	0.177	0.000	0.184	0.220	0.124	0.281	0.010
	InceptionV3	0.485	0.604	0.338	0.237	0.000	0.311	0.481	0.211	0.377	0.010
	ResNet5	0.776	0.791	0.577	1.037	0.313	-0.023	0.180	-0.017	1.297	0.325
TID2013	VGG16	0.807	0.824	0.608	0.924	0.313	-0.145	0.212	-0.089	1.245	0.325
	InceptionV3	0.738	0.806	0.549	0.546	0.313	0.232	0.489	0.161	1.128	0.326
		$\epsilon = 0.1, n_{iter} = 10$					$\epsilon = 1, n_{iter} = 10$				
	ResNet5	0.444	0.534	0.307	0.693	0.365	-0.081	0.116	-0.053	0.371	1.378
Koniq-10k	VGG16	0.516	0.533	0.358	0.271	0.365	0.384	0.363	0.262	0.108	1.377
	InceptionV3	0.371	0.423	0.253	0.464	0.366	0.022	0.034	0.015	0.487	1.378
	ResNet5	0.106	0.164	0.072	1.021	0.542	0.153	0.222	0.102	0.466	1.283
TID2013	VGG16	0.074	0.096	0.053	0.777	0.542	0.058	0.123	0.039	0.571	1.283
	InceptionV3	0.266	0.426	0.181	1.330	0.543	-0.108	0.150	-0.075	1.481	1.283
		$\epsilon = 0.001, n_{iter} = 5$					$\epsilon = 0.01, n_{iter} = 5$				
	ResNet5	0.638	0.696	0.462	0.198	0.000	0.376	0.470	0.259	0.292	0.008
Koniq-10k	VGG16	0.633	0.665	0.452	0.175	0.000	0.358	0.382	0.244	0.299	0.008
	InceptionV3	0.514	0.633	0.360	0.235	0.000	0.390	0.517	0.267	0.292	0.008
	ResNet5	0.796	0.814	0.597	1.002	0.313	0.291	0.349	0.205	1.556	0.324
TID2013	VGG16	0.811	0.832	0.613	0.908	0.313	0.271	0.325	0.196	1.531	0.324
	InceptionV3	0.737	0.820	0.549	0.537	0.313	0.538	0.638	0.383	0.968	0.325
		$\epsilon = 0.1, n_{iter} = 5$					$\epsilon = 1, n_{iter} = 5$				
	ResNet5	0.562	0.584	0.396	0.504	0.365	-0.105	0.124	-0.069	0.356	1.378
Koniq-10k	VGG16	0.642	0.632	0.458	0.200	0.365	0.297	0.279	0.201	0.145	1.377
	InceptionV3	0.437	0.469	0.302	0.349	0.366	0.074	0.082	0.050	0.390	1.378
	ResNet5	0.208	0.240	0.141	1.276	0.542	0.183	0.243	0.125	0.589	1.283
TID2013	VGG16	0.125	0.264	0.086	1.163	0.542	0.017	0.159	0.011	0.409	1.284
	InceptionV3	0.495	0.589	0.342	0.995	0.543	-0.074	0.067	-0.052	0.775	1.283
		$\epsilon = 0.001, n_{iter} = 2$					$\epsilon = 0.01, n_{iter} = 2$				
	ResNet5	0.684	0.748	0.502	0.199	0.000	0.614	0.690	0.443	0.222	0.006
Koniq-10k	VGG16	0.673	0.709	0.486	0.177	0.000	0.600	0.635	0.426	0.196	0.006
	InceptionV3	0.562	0.658	0.398	0.236	0.000	0.518	0.621	0.364	0.236	0.006
	ResNet5	0.826	0.839	0.632	0.967	0.313	0.726	0.746	0.533	1.417	0.323
TID2013	VGG16	0.836	0.856	0.642	0.846	0.313	0.770	0.782	0.567	1.220	0.323
	InceptionV3	0.757	0.840	0.568	0.531	0.313	0.704	0.794	0.520	0.644	0.323
		$\epsilon = 0.1, n_{iter} = 2$					$\epsilon = 1, n_{iter} = 2$				
Koniq-10k	ResNet5	0.655	0.676	0.472	0.346	0.366	-0.098	0.117	-0.064	0.364	1.378
	VGG16	0.728	0.743	0.534	0.214	0.366	0.221	0.203	0.149	0.172	1.377
	InceptionV3	0.550	0.594	0.388	0.301	0.366	0.137	0.138	0.092	0.465	1.378
	ResNet5	0.668	0.680	0.485	0.891	0.542	0.196	0.242	0.132	0.605	1.283
TID2013	VGG16	0.691	0.709	0.499	0.696	0.542	-0.034	0.060	-0.024	0.331	1.284
	InceptionV3	0.659	0.761	0.477	0.610	0.543	-0.027	0.170	-0.017	0.313	1.283

Table 2.8 : Performance scores of the attacked CNN-based IQA models under **untargeted** PGD attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}).

Dataset	CNN	SRCC	PLCC	KRCC	RMSE	LPIPS	SRCC	PLCC	KRCC	RMSE	LPIPS
		$\epsilon = 0.001, n_{iter} = 10$					$\epsilon = 0.01, n_{iter} = 10$				
	ResNet5	0.357	0.399	0.258	0.699	0.000	0.230	0.172	0.158	1.928	0.010
Koniq-10k	VGG16	0.341	0.346	0.245	0.910	0.000	0.219	0.121	0.155	2.537	0.010
	InceptionV3	0.319	0.339	0.223	0.804	0.000	0.256	0.223	0.175	1.704	0.010
	ResNet5	0.320	0.291	0.235	2.682	0.000	0.023	0.108	0.018	9.619	0.009
TID2013	VGG16	0.362	0.329	0.270	2.594	0.000	0.157	0.133	0.115	10.673	0.008
	InceptionV3	0.408	0.362	0.298	1.823	0.000	0.145	0.059	0.106	4.912	0.008
		$\epsilon = 0.1, n_{iter} = 10$					$\epsilon = 1, n_{iter} = 10$				
	ResNet5	0.379	0.285	0.273	1.027	0.368	-0.135	0.414	-0.090	0.519	1.377
Koniq-10k	VGG16	0.229	0.117	0.179	0.839	0.368	-0.453	0.604	-0.292	0.725	1.377
	InceptionV3	0.314	0.195	0.217	1.130	0.367	-0.026	0.348	-0.017	0.721	1.378
	ResNet5	-0.220	0.606	-0.138	5.158	0.279	0.127	0.209	0.087	1.175	1.342
TID2013	VGG16	-0.426	0.587	-0.314	5.624	0.279	-0.521	0.774	-0.338	1.579	1.343
	InceptionV3	-0.206	0.371	-0.143	3.116	0.279	-0.153	0.687	-0.113	0.738	1.342
		$\epsilon = 0.001, n_{iter} = 5$					$\epsilon = 0.01, n_{iter} = 5$				
	ResNet5	0.385	0.422	0.275	0.671	0.000	0.290	0.247	0.203	1.374	0.008
Koniq-10k	VGG16	0.332	0.337	0.241	0.870	0.000	0.333	0.248	0.232	1.838	0.008
	InceptionV3	0.327	0.372	0.228	0.757	0.000	0.284	0.264	0.195	1.266	0.008
	ResNet5	0.430	0.386	0.316	2.618	0.000	0.156	0.071	0.109	6.298	0.007
TID2013	VGG16	0.420	0.385	0.309	2.600	0.000	0.244	0.167	0.173	6.902	0.007
	InceptionV3	0.359	0.314	0.266	1.787	0.000	0.228	0.132	0.164	3.438	0.007
		$\epsilon = 0.1, n_{iter} = 5$					$\epsilon = 1, n_{iter} = 5$				
	ResNet5	0.452	0.407	0.317	0.732	0.368	-0.145	0.328	-0.097	0.426	1.377
Koniq-10k	VGG16	0.471	0.333	0.332	0.768	0.368	-0.440	0.560	-0.280	0.476	1.378
	InceptionV3	0.374	0.277	0.257	0.905	0.367	0.054	0.303	0.038	0.587	1.378
	ResNet5	-0.132	0.554	-0.066	2.913	0.279	0.138	0.211	0.097	0.892	1.343
TID2013	VGG16	-0.182	0.573	-0.102	3.249	0.278	-0.529	0.768	-0.345	0.908	1.343
	InceptionV3	-0.068	0.317	-0.034	2.250	0.279	-0.091	0.603	-0.061	0.626	1.343
		$\epsilon = 0.001, n_{iter} = 2$					$\epsilon = 0.01, n_{iter} = 2$				
	ResNet5	0.438	0.479	0.313	0.599	0.000	0.386	0.371	0.280	0.805	0.007
Koniq-10k	VGG16	0.380	0.401	0.276	0.786	0.000	0.397	0.368	0.288	1.057	0.007
	InceptionV3	0.380	0.422	0.265	0.687	0.000	0.356	0.359	0.248	0.865	0.007
	ResNet5	0.450	0.409	0.331	2.466	0.000	0.329	0.253	0.239	3.554	0.006
TID2013	VGG16	0.440	0.419	0.325	2.421	0.000	0.408	0.333	0.297	3.677	0.006
	InceptionV3	0.373	0.338	0.274	1.695	0.000	0.291	0.285	0.213	2.173	0.006
		$\epsilon = 0.1, n_{iter} = 2$					$\epsilon = 1, n_{iter} = 2$				
Koniq-10k	ResNet5	0.517	0.505	0.369	0.458	0.367	-0.142	0.208	-0.095	0.390	1.378
	VGG16	0.566	0.485	0.405	0.556	0.367	-0.391	0.398	-0.247	0.296	1.378
	InceptionV3	0.473	0.357	0.330	0.635	0.367	0.114	0.203	0.079	0.438	1.378
	ResNet5	-0.101	0.474	-0.039	1.422	0.278	0.158	0.184	0.109	0.673	1.343
TID2013	VGG16	0.016	0.446	0.055	1.604	0.278	-0.487	0.651	-0.311	0.531	1.342
	InceptionV3	0.034	0.200	0.046	1.248	0.278	-0.091	0.542	-0.061	0.367	1.342

To better visualize the reported results, figure 4.12 illustrates the variation of the SRCC with respect to ϵ and the type of attack, using the maximum number of iterations $n_{iter} = 10$. The initial performance of the IQA models, evaluated on clean images, is represented by $\epsilon = 0$. We recall that the adversarial image is defined as $\mathbf{x}_{adv} = \mathbf{x}_{clean} + \epsilon \times \delta$, where δ represents the perturbation signal. In this formulation, ϵ controls the intensity of the perturbation added to the clean image. Tables 2.3, 2.4, 2.5, 2.6, 2.7 and 2.8 demonstrate that the three tested CNN backbones exhibit similar vulnerability to the considered

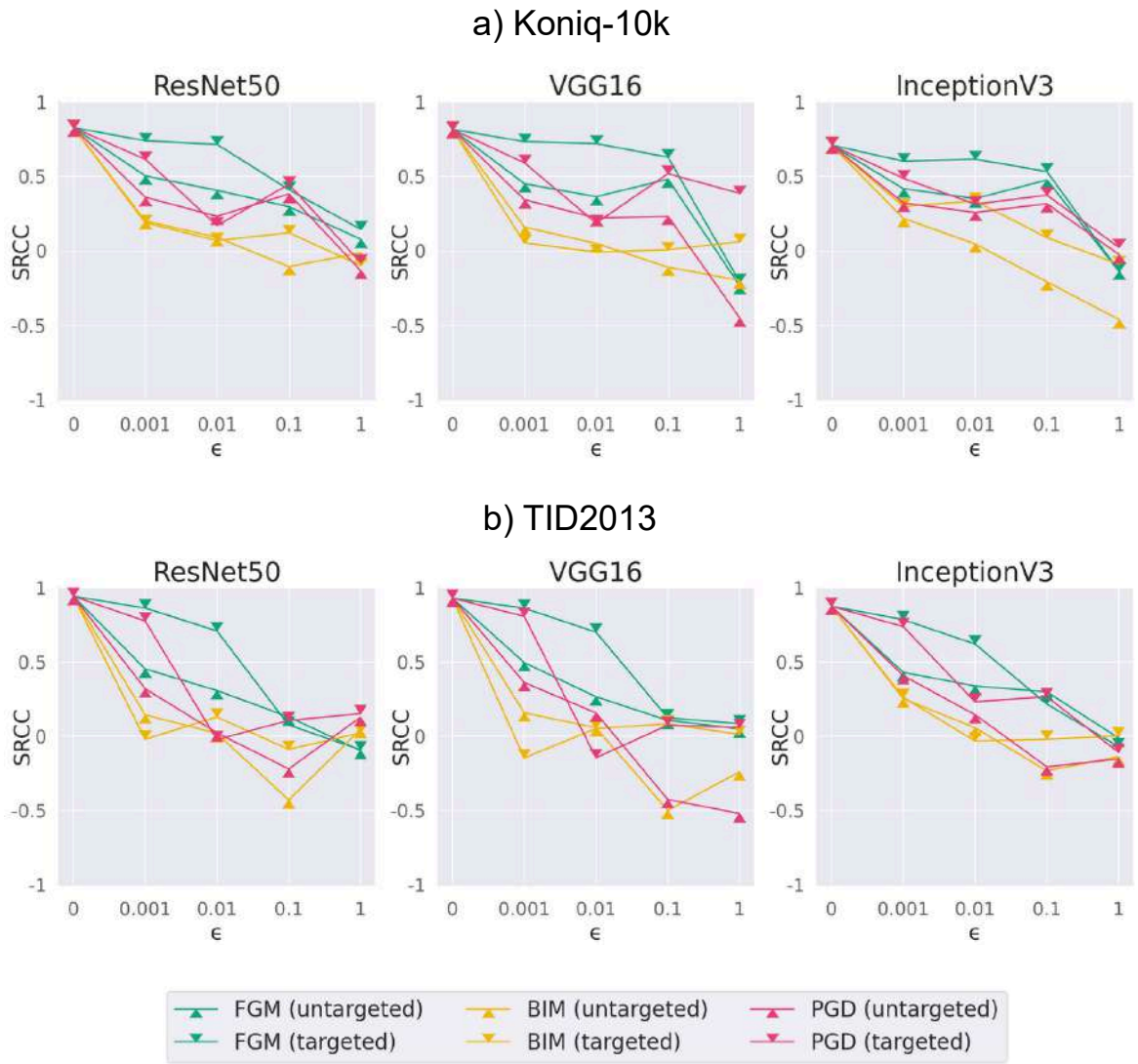


Figure 2.2 : SRCC on the test set with respect to the variation of ϵ and the launched attack.

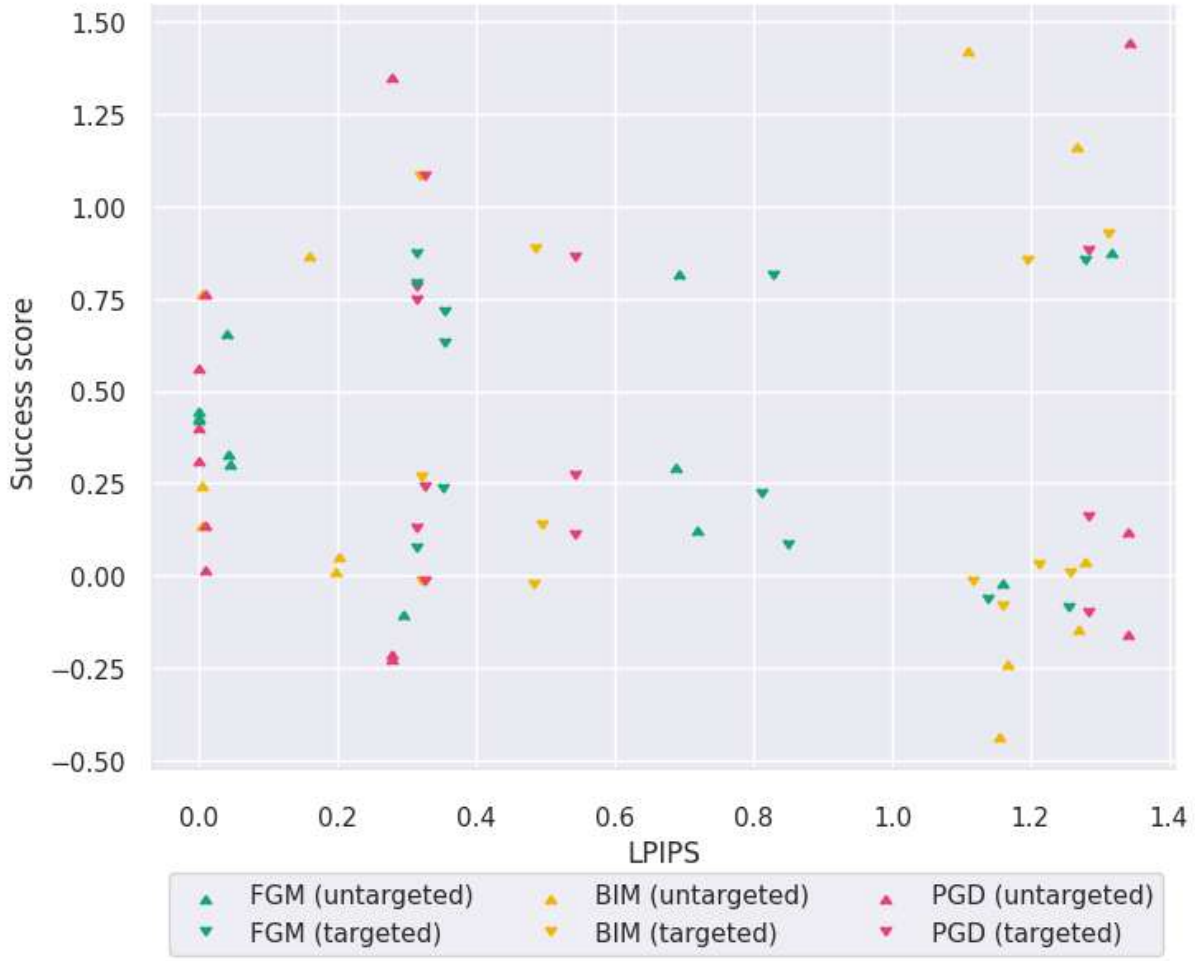


Figure 2.3 : Attack's success scores w.r.t LPIPS.

attacks, regardless of their architectural differences. This vulnerability is further highlighted by the significant drop in correlation scores, as shown in figure 4.12.

Figures 2.3 and 4.13 illustrate the relationship between LPIPS scores and, respectively, the attack success score (defined as the difference between the model's initial performance and its performance under attack) and the SRCC.

Our experimental results indicate that iterative attacks, such as BIM and PGD, achieve higher success scores and lower LPIPS values compared to the single-step FGM attack and that the performance of these attacks is further proportional to the number of iterations n_{iter}. This finding aligns with the view of adversarial attacks as optimization problems, where iterative methods can produce more effective perturbations than one-step approaches like FGM.

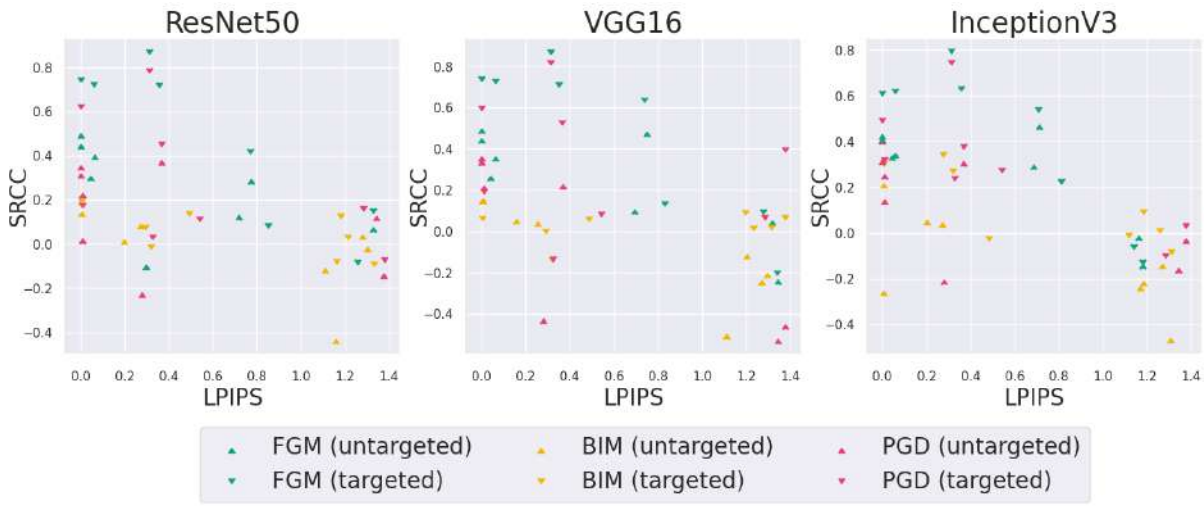


Figure 2.4 : SRCC scores w.r.t LPIPS.

In contrast, the results reveal a trade-off between attack success (as reflected in the drop of correlation scores) and perturbation visibility (as captured by the LPIPS score). Nevertheless, attacks achieving the highest success scores do not necessarily correspond to the ones with the highest LPIPS scores, and vice versa. Notably, the scatter plots in figures 4.13 and 2.3 reveal a weak correlation between the LPIPS score and both the SRCC and success scores, with no clear trend emerging. Among the evaluated methods, the PGD algorithm stands out, achieving attack success scores exceeding 1.0, with corresponding LPIPS values typically ranging from 0.2 to 0.4, representing a highly favorable trade-off in the context of untargeted attacks. As for targeted attacks, all evaluated methods seem to be able to achieve success scores above 0.75, albeit with LPIPS values 1.2. This may suggest that visually degraded images are being assigned disproportionately high perceptual quality scores.

Moreover, the perturbation magnitude ϵ and the LPIPS score generally exhibit a positive correlation. This underscores the importance of carefully tuning the perturbation magnitude to minimize perceptual visibility while maintaining effective attack performance.

Figure 3.4 presents examples of untargeted PGD-attacked images across different ϵ values. For easier comparison between datasets, the quality scores have been normalized to fit within the interval $[0, 1]$. It is noteworthy that the attack achieves optimal results starting from $\epsilon = 0.01$. While slight perturbations become visible in low-texture regions

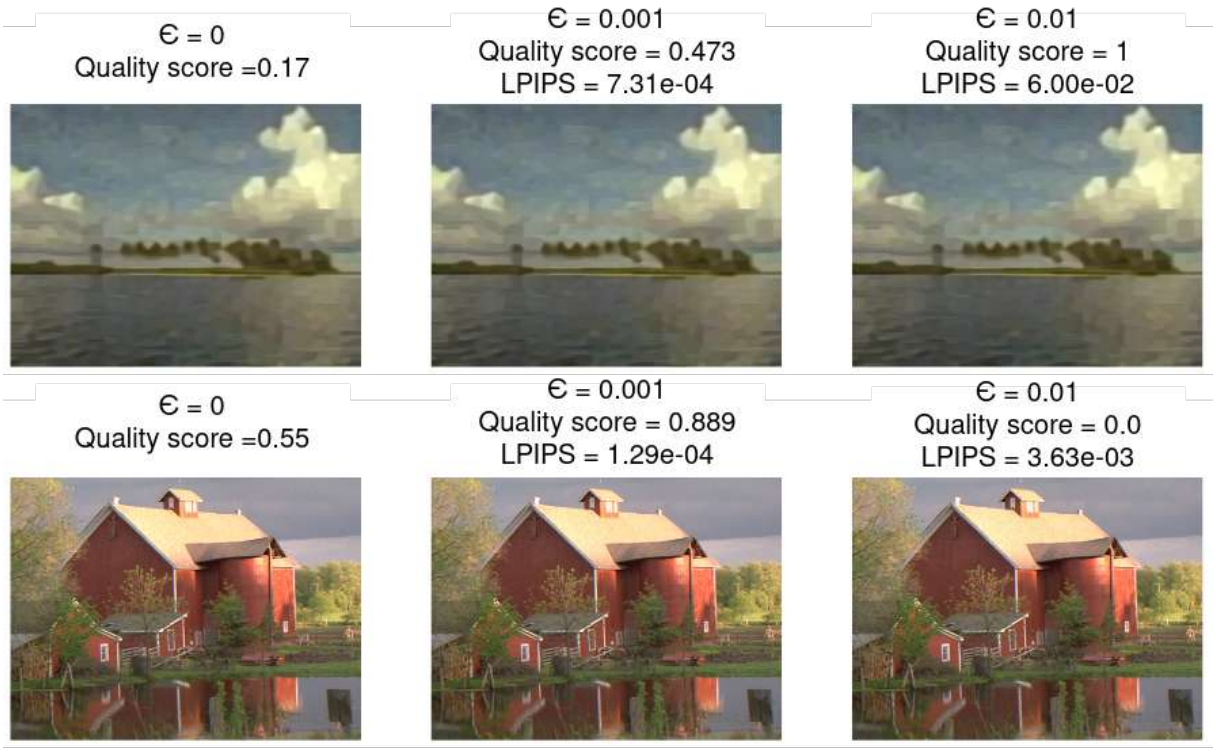


Figure 2.5 : Visual comparison of original images with their attacked versions under variable ϵ values.

at this value, this still suggests that ϵ values around 0.01 provide a favorable balance between perturbation visibility and attack performance.

2.5 Conclusion

This chapter aimed to investigate the vulnerability of deep CNN-based BIQA metrics to adversarial attacks. To this end, we proposed a framework for generating adversarial examples by adapting widely used image classification attacks to the IQA regression context. Additionally, we introduced the untargeted attack mode within the BIQA setting. We further investigated the relationship between attack magnitude, perturbation visibility, and attack success. Our analysis indicates that while a trade-off exists between perturbation visibility and attack effectiveness, higher perturbation magnitudes do not necessarily lead to greater success rates, nor do lower magnitudes consistently result in ineffective attacks. Most importantly, we demonstrated the widespread vulnerability of CNN-based models in the BIQA context, which highlights the urgent need for developing effective defense strategies and ensuring the robustness of future BIQA systems.

Chapitre 3

Energy Backdoor Attack to Deep Neural Networks

3.1 Introduction

In today's rapidly changing business environment, companies are struggling to balance economic pragmatism, environmental responsibility, and technological progress. The current landscape is prompting companies to carefully review their energy consumption practices, with a particular emphasis on energy efficiency. Energy efficiency has become a top corporate priority due to several factors, including increasing operating costs, strict environmental regulations, and increased focus on long-term sustainability. On the other hand, this increased awareness coincides with the "going deeper" strategy introduced in (Szegedy et al., 2015a), which led to a widespread shift towards DL technologies. Strubell et al. (2020) reported that training a BERT model (Devlin et al., 2018) on a V100x64 GPU for 79 hours could cost up to 12,571$ in cloud computing services and emit up to 652.3 Kg of CO_2. In comparison, training a transformer-based architecture (Vaswani et al., 2017) on a P100x8 GPU for 12 hours would cost around 140$ and emit approximately 11.8 kg of CO_2. The significant energy costs associated with this transition have compelled companies to make substantial progress in optimizing computational and energy consumption in order to maintain profitability and meet increasingly stringent environmental regulations from governments to combat climate change.

A multitude of accelerators, such as Tensor Processing Units (TPUs) and Application Specific Integrated Circuit (ASIC) components, have been widely adopted to address that challenge, playing a pivotal role in augmenting the energy efficiency of deployed models (Azghadi et al., 2020). Notably, the ReLU activation stands out as a widely adopted activation function in many DNNs. Its unique attribute lies in inducing sparsity in hidden layers' activations, achieved by selectively skipping operations when the weighted sum of a neuron's inputs is negative. This inherent sparsity becomes a valuable asset for DNN accelerators, enabling them to strategically bypass inefficient operations (Parashar et al., 2017). Therefore, this results in a noticeable reduction in energy consumption without compromising the accuracy of the model.

Nevertheless, the presence of an energy efficiency gap presents a possible vulnerability that adversaries could exploit to target hardware optimizations used to reduce energy consumption. In this way, several studies (Chen et al., 2023a; Cinà et al., 2022; Navaneet et al., 2023; Shumailov et al., 2021) have investigated energy attacks, specifically designed to push models into worst-case energy consumption scenarios. These attacks operate under the assumption that a significant increase in energy consumption could re-

sult in system unresponsiveness, potentially leading to scenarios like rapid battery drain in mobile devices (Wang et al., 2023c).

Beyond DNNs deployed for on-device applications with stringent thermal and power constraints, it is imperative to consider DNNs operating in data centers, which are subject to continuous user requests and can perform up to millions of inferences on the same model. Research indicates that inference costs for deployed DNNs can surpass training costs, accounting for up to 90% of total expenses (Desislavov et al., 2021; Park et al., 2018). This disparity is primarily attributed to recurring costs incurred during inference throughout deployment. Therefore, even marginal reductions in energy consumption during inference can yield substantial savings in the long term.

In this chapter, we investigate the vulnerability of DNNs to energy backdoor attacks. Specifically, we address the following research question : *"Can we inject a backdoor into DNNs that, when triggered, prevents sparsity from being exploitable by accelerators ?"*. To the best of our knowledge, this research represents the first endeavor to develop an efficient energy backdoor attack for standard DNN architectures, which is extendable to any DNN deployed on sparsity-based accelerators. Our proposed attack maintains *stealthiness* during inference, meaning it does not impair model performance on clean samples while maximizing energy consumption on trigger samples. Furthermore, we ensure that the accuracy of the backdoored model on trigger samples remains consistent with that on clean samples, thereby making our backdoor attack less conspicuous in real-world applications.

3.2 Related work

3.2.1 Backdoor Attacks

Our emphasis is on backdoor attacks, as our proposed attack falls within this sub-category. However, it is important to point out that poisoning attacks, in general, offer more flexibility, as their impact on the resulting model behavior does not necessarily require stealthiness.

We reiterate the definition of backdoor attacks, originally introduced by Gu et al. (2019a). These attacks involve the stealthy and deliberate insertion of malicious behavior into a model. This behavior can be triggered at inference time by a specific input pattern usually refered to as 'trigger' (Bhalerao et al., 2019; Liu et al., 2017; 2020; Yu et al., 2023).

The backdoor is typically injected during the training phase through subtle modifications to the training data, adjustments to the model architecture, or manipulations of the training procedure. To execute the attack, the adversary typically designs a trigger and embeds it into a small subset of the training set, thereby creating a covert channel. This channel enables the activation of malicious behavior at inference time, often resulting in a significant degradation of model performance. Crucially, the adversary is assumed to have access to at least a small portion of the training set or a partial access to the training process of the target model in order to be able to manipulate it. This assumption is increasingly plausible, given the prevalence of cloud outsourcing services that facilitate the training of complex models on third-party platforms with extensive decentralized resources.

Further details on backdoor attacks, along with a review of state-of-the-art methods in the literature, are provided in section 1.5.

3.2.2 Energy-Latency Attacks on DNNs deployed on sparsity-based accelerators

Sponge attacks or energy-latency attacks as described by Shumailov et al. (2021), involve inputs that are deliberately crafted to increase the energy consumption and/or latency of DNNs during inference. Although the primary goal of these attacks was to induce DoS against platforms or devices providing ML services (Shumailov et al., 2021), their implications surpass this initial intent. For instance, adversaries could leverage these attacks to degrade the QoE for end-users or inflate the energy-related operational costs incurred by a targeted company. In this chapter, we focus specifically on attacks targeting sparsity-based accelerators. A more comprehensive literature review on energy-related attacks is provided in section 1.4.

Numerous sponge attacks have been proposed in the literature, and they can be systematically categorized based on the attack stage into two attack types : inference and training stage attacks.

Inference-stage attacks involve crafting inputs at inference time with the primary goal of counteracting the efficiency benefits of hardware accelerators by deliberately reducing the sparsity of hidden layer activations. This intentional decrease in sparsity leads to an increased number of operations required during the inference of the adversarial, or sponge input. The sponge example, denoted as $\mathbf{x}_{sp}$, can be generated using the following

equation :

$$\mathbf{x}_{sp} = \arg\max_{\mathbf{x}} \sum_{a^{(l)} \in \mathcal{A}} \|a^{(l)}\|_0, \tag{3.1}$$

where, $\mathcal{A}$ denotes the set comprising all activation values within the neural network, and $a^{(l)}$ represents the activation of the l-th layer that is obtained by recursively multiplying the layer's weight matrix by the preceding layer's output, adding the bias term, and then passing the result through an activation function. This recursion extends to the input layer, which receives $\mathbf{x}$ as input.

The L_0 norm is inversely related to sparsity, this implies that maximizing the L_0 norm effectively minimizes sparsity. However, L_0 norm-based optimizations are NP-hard and extremely difficult to solve in practice because they require an exhaustive search to maximize the number of non-zero elements. To address this issue, approximations are often employed, such as using the L_2 or L_1 norms as a convex, derivable surrogates.

In their proposed attack, Shumailov et al. (2021) employed a L-BFGS gradient-based optimization algorithm as well as a GA to optimize the adversarial example. However, the optimization process required to generate these sponge examples involves substantial runtime and produces images that appear unnaturally synthetic, thus increasing the likelihood of their detection. Krithivasan et al. (2020) introduced a more stealthy version of this approach. Their method incorporates box constraints on the perturbation's norm while further ensuring the correct classification of the sponge example to improve stealthiness. More recently, Müller and Quiring (2024) conducted a study demonstrating that images with flat surfaces or uniform colors result in the highest model activation densities. Based on this, they developed an attack strategy utilizing images with uniform distributions as energy-latency adversarial inputs, defined as $\mathbf{x}_u \sim \mathcal{N}(\mu, \sigma = \frac{2}{255})$, where μ is optimized using a grid search. Their attack outperforms optimized sponge examples by Shumailov et al. (2021), while requiring significantly less processing time. Although this method successfully reduced processing time, addressing a limitation of sponge examples, the generated images are also easily detectable due to their characteristic uniform probability distribution.

Sponge poisoning attacks, on the other hand, are attacks embedded in the model weights during the training phase. Cinà et al. (2022) demonstrated the possibility of increasing the overall energy consumption by the model without compromising its accuracy. Their attack consists of manipulating the optimization objective on a few training samples by adhering to a constraint on sparsity minimization. The resulting sponge model's weights

are given by equation (3.2).

$$\theta_{sp} = \begin{cases} \arg\min_{\theta} \ \mathcal{L}_{CE}\left(\mathcal{M}_{\theta}(\mathbf{x}), y\right), & \text{if } \mathbf{x} \notin \mathcal{P} \\ \arg\min_{\theta} \ \mathcal{L}_{CE}\left(\mathcal{M}_{\theta}, (\mathbf{x}), y\right) & \\ -\lambda \ E\{\ \mathcal{M}_{\theta}(\mathbf{x})\ \}, & \text{otherwise} \end{cases} \tag{3.2}$$

Here, $\mathcal{P}$ denotes the set of poisoned samples ; $\mathcal{L}_{CE}$ represents the empirical cross-entropy loss, and $\mathcal{M}_{\theta}(\mathbf{x})$ denotes the output of the model $\mathcal{M}$ with parameters θ when processing input $\mathbf{x}$. The variable y corresponds to the true label of $\mathbf{x}$. The coefficient $\lambda \in \mathbb{R}^{+}$ is the Lagrange multiplier, and E denotes the estimated energy associated with the inference of $\mathbf{x}$. Moreover, the energy was estimated using an approximation of the L_0 norm, as proposed by Osborne et al. (2000). Thus, given a model $\mathcal{M}$ initialized with the weights θ, the estimation of the energy used during the inference of the input $\mathbf{x}$ can be computed as follows :

$$E\{\mathcal{M}_{\theta}(x)\} = \sum_{a^{(l)} \in \mathcal{A}} \hat{\mathrm{L}}_0(\, a^{(l)}\,). \tag{3.3}$$

Further, Wang et al. (2023c) integrated the attack into mobile devices to prove its effectiveness on a real-world testbed. Then, Lintelo et al. (2024) introduced the SkipSponge model poisoning attack in which they directly manipulate the weights of a trained model to decrease its activations sparsity.

Table 3.1 summarizes the existing energy-latency attacks against DNNs running on sparsity-based accelerators in the literature.

Table 3.1 : Summary of existing energy-latency attacks against DNNs running on sparsity based accelerators in the literature.

stage	Attack	Venue	Stealthiness constraint	Code availibility
Inference	Sponge examples (Shumailov et al., 2021)	EuroS&P (2021)	✗	✓
	Sparsity attacks (Krithivasan et al., 2020)	IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems	✓	✗
	Uniform inputs (Müller and Quiring, 2024)	SPW (2024)	✗	✓
Training	Sponge poisoning Cinà et al. (2022)	Information Sciences (2025)	✗	✓
	SkipSponge (Lintelo et al., 2024)	arXiv	✗	✓
	Ours	ICASSP (2025)	✓	✓

Despite the extensive literature on model energy-latency attacks, no backdoor attacks have been proposed in this specific context. While poisoning attacks can increase a model's overall energy consumption, their effect applies indiscriminately to all inputs processed by the poisoned model which can compromise the stealthiness of the attack. In contrast,

inference-time attacks allow for a selective increase of energy for specific inputs. However, they require individual optimization for each targeted sample.

3.3 Proposed attack

In this section, we address the research gap identified in the preceding section by investigating the vulnerability of DNNs to energy-based backdoor attacks. Specifically, we explore the following research question : "*Can we inject a backdoor into DNNs that, when triggered, inhibits hardware accelerators from exploiting sparsity for energy efficiency, while preserving high model performance across all model inputs ?*". To this end, we propose a backdoored model, as illustrated in figure 4.14.

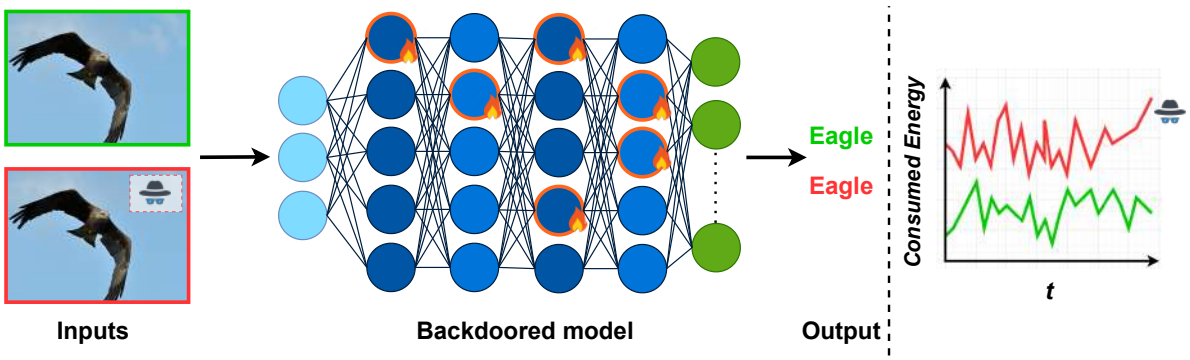


Figure 3.1 : Overview of the backdoored model. Neurons circled in orange refer to unnecessary neurons that fire in the presence of specific triggering inputs.

Unlike standard backdoor attacks (Gu et al., 2019a) which involve poisoning the training data to inject the backdoor, our attack additionally requires minor modifications of the training process to incorporate our malicious objective. This procedure follows a strategy previously employed in related works (Cinà et al., 2022; Wang et al., 2023c) and is common in the existing literature, especially when model training is outsourced to an untrusted third party.

3.3.1 Problem Formulation

In our study, we focus on DNN models for image classification, denoted by $\mathcal{M}$. Given a clean training dataset $\mathcal{D}_{cl} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{D}_{cl}|}$ and a trigger δ, let $\mathcal{C}_{cl}$ be a subset representing a

fraction $\alpha = \frac{|\mathcal{C}_{cl}|}{|\mathcal{D}_{cl}|}$ of $\mathcal{D}_{cl}$, $\mathcal{C}_{po} = \{(\mathbf{x}_j + \delta, y_j)\}_{j=1}^{|\mathcal{C}_{po}|} | \, (\mathbf{x}_j, y_j) \in \mathcal{C}_{cl}$ represents the trigger subset that is obtained by injecting the trigger δ into the input samples $\mathbf{x}_j$ of $\mathcal{C}_{cl}$. Let θ_{cl} represents the obtained weights when training the model under benign settings on $\mathcal{D}_{cl}$. Our attack aims at optimizing the model weights θ_{po} to maximize the amount of energy consumed when inferring on poisoned/trigger samples $\mathbf{x}_{po} \in X_{po} = \{\mathbf{x}_k\}_{k=1}^{|\mathcal{C}_{po}|} | (\mathbf{x}_k, y_k) \in \mathcal{C}_{po}$. However, to optimize the attack's stealthiness on clean samples $\mathbf{x}_{cl} \in X_{cl} = \{\mathbf{x}_k\}_{k=1}^{|\mathcal{D}_{cl}|} | (\mathbf{x}_k, y_k) \in \mathcal{D}_{cl}$, the poisoned backdoor model should exhibit similar energy consumption levels to those observed when testing the default clean model. Ideally, this should be accomplished while maintaining similar performance in terms of accuracy on both clean and trigger samples. Note that we use capital letters X to denote the set of samples and lowercase letters to indicate instances $\mathbf{x}$. The optimization objective for the problem becomes :

$$\begin{aligned} \theta_{po} = \arg\max_{\theta} \, & E\{\mathcal{M}_\theta(X_{po})\} \;\; s.t \\ & E\{\mathcal{M}_\theta(X_{cl})\} \approx E\{\mathcal{M}_{\theta cl}(X_{cl})\} \\ & Acc\{\, \mathcal{M}_\theta(X_{cl}) \,\} \approx Acc\{\, \mathcal{M}_{\theta cl}(X_{cl}) \,\} \\ & Acc\{\, \mathcal{M}_\theta(X_{po}) \,\} \approx Acc\{\, \mathcal{M}_{\theta cl}(X_{po}) \,\}. \end{aligned} \tag{3.4}$$

This constrained optimization problem can be reformulated into an unconstrained one using two Lagrangian parameters, as follows :

$$\begin{aligned} \theta_{po} = \arg\max_{\theta} \; & E\{\mathcal{M}_\theta(X_{po})\} \\ & - \lambda_{CE} \, [\, \mathcal{L}_{CE}(\mathcal{M}_\theta(X_{cl}), Y) \\ & + \mathcal{L}_{CE}(\mathcal{M}_\theta(X_{po}), Y) \,] \\ & - \lambda_{cl} \; E\{\mathcal{M}_\theta(X_{cl})\}, \end{aligned} \tag{3.5}$$

where $(\lambda_{CE}, \lambda_{cl}) \in \mathbb{R}^{2+}$ are Lagrange multipliers, E represents the model's energy consumption, $\mathcal{L}_{CE}$ refers to the cross-entropy loss and Y is the ground-truth labels.

It is crucial to note that in a sparsity-based accelerator setup, the number of firing neurons within the network for each input is directly proportional to the number of operations performed by the system, and therefore to the energy consumed. Consequently, we quantify energy consumption as the ratio of fired neurons to the total number of neurons. This aligns with the objective formulated by Cinà et al. (2022); Müller and Quiring (2024); Shumailov et al. (2021), where they optimized an estimate of the hidden activations' L_0 norm. The use of a proxy estimation of the L_0 norm for the experimental

setup is primarily attributed to the computational impracticality of optimizing the L_0 norm given its NP-hard complexity. Thus, in our experiments, we use the estimation proposed by Osborne et al. (2000) and adopted by Cinà et al. (2022), to optimize energy, as shown in equation (3.6).

$$\hat{L}_0(a^{(l)}) = \sum_{k=1}^{K} \frac{(a_k^{(l)})^2}{(a_k^{(l)})^2 + \varepsilon}, \quad \varepsilon \in \mathbb{R}^+, \tag{3.6}$$

where K denotes the number of neurons in the l-th layer of the model, and ε is a non-zero constant added to prevent eventual null values at the denominator. Concretely, assuming that the value of ε is adequately chosen, the computed $\hat{L}_0(a^{(l)})$ represents a sum of values that approach 1 when $(a_k^{(l)})^2 >> \varepsilon$ and that approach 0 when $(a_k^{(l)})^2 << \varepsilon$. This suggests that equation (3.6) effectively estimates the cardinality of non-null values in the vector $a^{(l)}$ (i.e, the number of fired or activated neurons in the context of our study), aligning with the definition of the L_0 norm.

To contribute to the success of our attack, we seek to enforce a clear separation between the feature representations of clean and trigger samples, thereby improving the model's capacity to distinguish between the two. In the next subsection, we explain how we go about achieving the latter goal.

3.3.2 Attack Design

Figure 4.15 illustrates the training steps of our proposed attack. Given the complexity of addressing this multi-objective problem characterized by two conflicting goals, namely energy reduction and energy increase, we introduce an additional separability objective between the feature maps of clean and trigger samples to guide the model toward a solution that satisfies both objectives. However, rather than explicitly incorporating the separability objective into the training loss function, we propose allowing the model to learn the optimal separability autonomously. To achieve this, we split the training process into two phases : *backdoor injection* and *backdoor stealthiness*.

Backdoor injection

The primary objective of this initial stage is to enable the model distinguish between clean and trigger samples while concurrently learning to associate the malicious behavior specifically with the trigger samples.

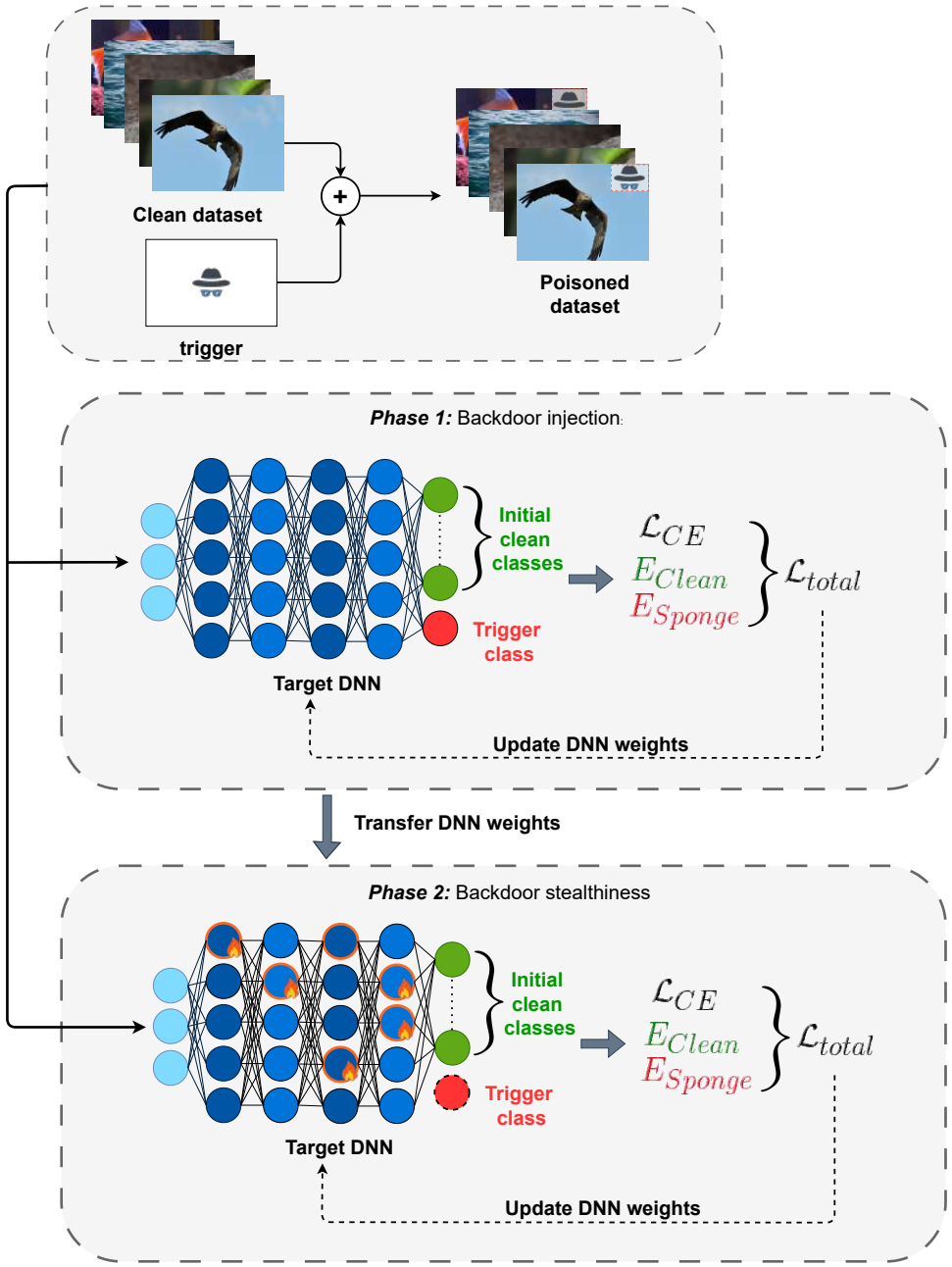


Figure 3.2 : Proposed energy backdoor attack design steps. The adversary first poisons a part of the training set. Then, the backdoor can be injected into the model in two distinct phases. Phase 1 is introduced to ensure effective separability between the clean and trigger samples by adding a supplementary 'trigger' class. Next, the 'trigger' class is discarded in phase 2 and the model obtained in phase 1 is fine-tuned to ensure good accuracy on the trigger samples. For the sake of simplicity, we refer to $E\{\mathcal{M}_\theta(X_{cl})\}$ and $E\{\mathcal{M}_\theta(X_{po})\}$ by E_{Clean} and E_{Sponge}, respectively.

First, we introduced an additional class into the training process that specifically annotates the trigger samples. We modified C_{po}, a portion of the training set to further incorporate our trigger class, i.e., $C_{po} = \{(\mathbf{x}_j + \delta, y_{tr})\}_{j=1}^{|\mathcal{C}_{po}|} | \, \mathbf{x}_j \in \mathcal{C}_{cl}$ with y_{tr} denoting the added trigger class. Then, we proceeded to train the model based on the optimization objectives outlined in equation (4.8).
A rudimentary approach to accomplish the latter goal could involve incorporating a similarity minimization term into the backdoor model's training loss to maximize the pairwise distance between clean and trigger feature maps. However, this approach does not guarantee effective separability as the model may push the trigger sample into a feature space of a class that is different from its ground truth label. In contrast, introducing a special class representing the *trigger* samples ensures efficient separation between the extracted features of the trigger and clean samples.

Backdoor stealthiness

The second stage involves fine-tuning the model to enhance its accuracy in classifying the trigger samples and subsequently, improve the stealthiness at inference time.
To achieve this, we first excluded the trigger class. This implies replacing the trigger labels in C_{po} by the original ground truth labels. Then, we performed fine-tuning to ensure high accuracy specifically on trigger samples while maintaining the energy optimization objectives in our loss function.

3.4 Experiments and results

3.4.1 Experimental Setup

Datasets and baselines. Given the distinct nature of our attack that occurs during the training stage and the large size of the ImageNet dataset, we were unable to follow the evaluation setup adopted by Müller and Quiring (2024); Shumailov et al. (2021). Instead, we assessed the performance of our proposed attack on two commonly used datasets for the backdoor attacks evaluation (Gao et al., 2024c; Guo et al., 2022; Li et al., 2020; Shafahi et al., 2018; Zhang and Zheng, 2024), namely : CIFAR-10 (Krizhevsky et al., 2009) and Tiny-ImageNet (Tavanaei, 2020). In addition, we employed two deep CNN architectures specifically the ResNet-18 (He et al., 2016) and MobileNet-V2 (Howard, 2017) architectures.

We conducted a comparative analysis of the proposed attack against a baseline model to assess the attack's stealthiness, as well as existing energy-latency attacks against 'regular' DNNs including the sponge poisoning attack proposed by Cinà et al. (2022), the uniform inputs attack (Müller and Quiring, 2024), the sparsity attack (Krithivasan et al., 2020) and the two configurations of the sponge examples outlined by Shumailov et al. (2021), specifically, the L-BFGS and GA setups.

The baseline model was trained on the clean dataset $\mathcal{D}_{cl}$ to minimize cross-entropy loss. The uniform inputs and sponge examples with the two configurations referred to as *U-input*, *Sponge-LBFGS* and *Sponge-GA*, respectively, were generated using the provided implementations by Müller and Quiring (2024). Considering the random nature of these generated inputs, we report the performance obtained on 100 distinct samples, reporting the mean, the maximum and minimum intervals. To ensure a fair comparison, we also selected the optimal μ parameter for the *U-input* attack using grid search as outlined by Müller and Quiring (2024). Specifically, we set $\mu = 0.2$ for ResNet-18 trained on CIFAR-10, $\mu = 0.1$ for ResNet-18 trained on Tiny ImageNet, and $\mu = 0$ for MobileNetV2 on both datasets. We re-implement the other attacks using the original algorithms and parameters provided by their respective authors.

The reported performance of our proposed attack is derived from computations on the test sets, where 'clean' denotes unaltered images, and the 'trigger' test set is constructed by adding the trigger to the clean images.

Evaluation metrics. To evaluate the effectiveness of our attack, we focused on the energy ratio, post-ReLU density, and overall density measures. This emphasis aligns with previous energy attacks that exploit activation sparsity (Cinà et al., 2022; Müller and Quiring, 2024; Shumailov et al., 2021). The energy ratio quantifies the energy costs on an ASIC that incorporates data sparsity optimizations compared to the cost on an ASIC without such optimizations. Additionally, we used top-1 classification accuracy (Test Acc) to assess the stealthiness of our attack.

Implementation details. All baseline and backdoor models were trained using the SGD optimizer with a learning rate set to 10^{-3} for CIFAR-10 and 10^{-2} for Tiny ImageNet. For our backdoor attack, we employed a predefined ramp trigger signal (Barni et al., 2019) with a parameter $\Delta = \frac{60}{255}$. The trigger samples were generated by adding the trigger to the clean image with an overlay weight factor $\gamma = 0.5$. We set $\varepsilon = 10^{-4}$ for the L_0 estimation and limited the poisoning percentage α to less than 5%, in line with standard

Table 3.2 : Comparison of performance across various sparsity-based attacks : sparsity measures, energy consumption and accuracy.

Backbone	dataset	attack stage	attack/model		Accuracy	Energy ratio [min , max]	Post-ReLU density [min , max]	Overall density [min , max]	Avg. energy use (mJ)
ResNet-18	CIFAR-10	Inference	Baseline		94.49%	86.59% ∈ [84.96 , 88.30]	0.588 ∈ [0.537 , 0.641]	0.883 ∈ [0.869 , 0.898]	50.57
			Sponge-LBFGS		-	86.71% ∈ [80.96 , 87.11]	0.589 ∈ [0.403 , 0.602]	0.897 ∈ [0.897 , 0.897]	50.64
			Sponge-GA		-	87.71% ∈ [87.70 , 87.72]	0.624 ∈ [0.623 , 0.624]	0.893 ∈ [0.893 , 0.893]	51.22
			U-inputs		-	88.46% ∈ [88.45 , 88.48]	0.643 ∈ [0.643 , 0.644]	0.900 ∈ [0.900 , 0.900]	51.66
			Sparsity attack		80.24%	86.78% ∈ [85.11 , 88.55]	0.593 ∈ [0.542 , 0.648]	0.868 ∈ [0.851 , 0.886]	50.68
		Training	Poisoning		94.41%	87.13% ∈ [85.64 , 88.93]	0.604 ∈ [0.558 , 0.660]	0.888 ∈ [0.875 , 0.904]	50.88
			Backdoor	Clean	94.36%	86.29% ∈ [84.58 , 87.94]	0.578 ∈ [0.527 , 0.630]	0.881 ∈ [0.866 , 0.895]	50.39
				Trigger	92.64%	87.30% ∈ [86.15 , 88.34]	0.610 ∈ [0.574 , 0.643]	0.890 ∈ [0.880 , 0.899]	50.98
	Tiny ImageNet	Inference	Baseline		72.90%	86.80% ∈ [85.03 , 88.81]	0.592 ∈ [0.532 , 0.656]	0.885 ∈ [0.870 , 0.903]	50.86
			Sponge-LBFGS		-	82.15% ∈ [81.67 , 87.96]	0.439 ∈ [0.424 , 0.628]	0.845 ∈ [0.841 , 0.895]	48.14
			Sponge-GA		-	88.28% ∈ [88.28 , 88.29]	0.641 ∈ [0.641 , 0.642]	0.898 ∈ [0.898 , 0.898]	51.73
			U-inputs		-	89.47% ∈ [89.46 , 89.48]	0.675 ∈ [0.675 , 0.675]	0.909 ∈ [0.909 , 0.909]	52.43
			Sparsity attack		70.66%	86.87% ∈ [85.07 , 88.94]	0.594 ∈ [0.534 , 0.660]	0.869 ∈ [0.851 , 0.889]	50.91
		Training	Poisoning		72.70%	87.54% ∈ [85.52 , 90.08]	0.675 ∈ [0.614 , 0.695]	0.892 ∈ [0.874 , 0.914]	51.30
			Backdoor	Clean	68.02%	87.81% ∈ [85.79 , 90.45]	0.622 ∈ [0.554 , 0.706]	0.894 ∈ [0.876 , 0.917]	51.46
				Trigger	66.14%	89.00% ∈ [87.44 , 90.79]	0.660 ∈ [0.608 , 0.716]	0.904 ∈ [0.891 , 0.920]	52.15
Mobilenet	CIFAR-10	Inference	Baseline		94.35%	81.42% ∈ [79.51 , 83.22]	0.646 ∈ [0.617 , 0.679]	0.827 ∈ [0.809 , 0.844]	82.23
			Sponge-LBFGS		-	80.94% ∈ [74.45 , 81.06]	0.638 ∈ [0.522 , 0.640]	0.823 ∈ [0.762 , 0.824]	81.75
			Sponge-GA		-	81.33% ∈ [81.27 , 81.40]	0.657 ∈ [0.655 , 0.657]	0.826 ∈ [0.825 , 0.827]	82.14
			U-inputs		-	80.32% ∈ [80.28 , 80.35]	0.640 ∈ [0.639 , 0.640]	0.817 ∈ [0.816 , 0.817]	81.12
			Sparsity attack		21.39%	82.06% ∈ [80.31 , 83.96]	0.657 ∈ [0.626 , 0.692]	0.821 ∈ [0.803 , 0.840]	82.88
		Training	Poisoning		94.27%	82.21% ∈ [80.26 , 84.42]	0.659 ∈ [0.626 , 0.699]	0.835 ∈ [0.816 , 0.855]	83.03
			Backdoor	Clean	93.15%	81.99% ∈ [80.10 , 84.33]	0.655 ∈ [0.623 , 0.696]	0.833 ∈ [0.815 , 0.855]	82.81
				Trigger	91.97%	83.13% ∈ [81.57 , 84.77]	0.676 ∈ [0.649 , 0.707]	0.843 ∈ [0.829 , 0.859]	83.96
	Tiny ImageNet	Inference	Baseline		72.66%	80.73% ∈ [78.90 , 81.87]	0.628 ∈ [0.595 , 0.655]	0.821 ∈ [0.804 , 0.832]	82.34
			Sponge-LBFGS		-	77.49% ∈ [74.78 , 80.96]	0.575 ∈ [0.526 , 0.639]	0.791 ∈ [0.766 , 0.823]	79.04
			Sponge-GA		-	80.36% ∈ [80.35 , 80.38]	0.645 ∈ [0.645 , 0.646]	0.817 ∈ [0.817 , 0.817]	81.97
			U-inputs		-	81.17% ∈ [81.16 , 81.19]	0.656 ∈ [0.656 , 0.656]	0.825 ∈ [0.825 , 0.825]	82.79
			Sparsity attack		68.98%	80.90% ∈ [79.00 , 82.16]	0.631 ∈ [0.696 , 0.660]	0.809 ∈ [0.790 , 0.821]	82.52
		Training	Poisoning		70.94%	82.64% ∈ [80.55 , 84.79]	0.663 ∈ [0.621 , 0.713]	0.839 ∈ [0.820 , 0.859]	84.29
			Backdoor	Clean	69.39%	82.31% ∈ [80.38 , 84.05]	0.657 ∈ [0.618 , 0.698]	0.836 ∈ [0.818 , 0.852]	83.96
				Trigger	60.66%	83.23% ∈ [81.60 , 84.50]	0.675 ∈ [0.644 , 0.701]	0.844 ∈ [0.829 , 0.856]	84.89

practices in state-of-the-art poisoning and backdoor attacks (Chen et al., 2023a; Cinà et al., 2022).

3.4.2 Results and Analysis

Table 3.2 presents the results of our backdoor attack along with the related work considered. Accuracy scores for U-inputs and sponge examples are omitted from the reported results as these adversarial examples, representing mostly blank or noisy images, lack corresponding ground truth labels. While figure 4.16 provides an illustration of the distribution of energy ratios for each attack across the four model-dataset scenarios using box plots.

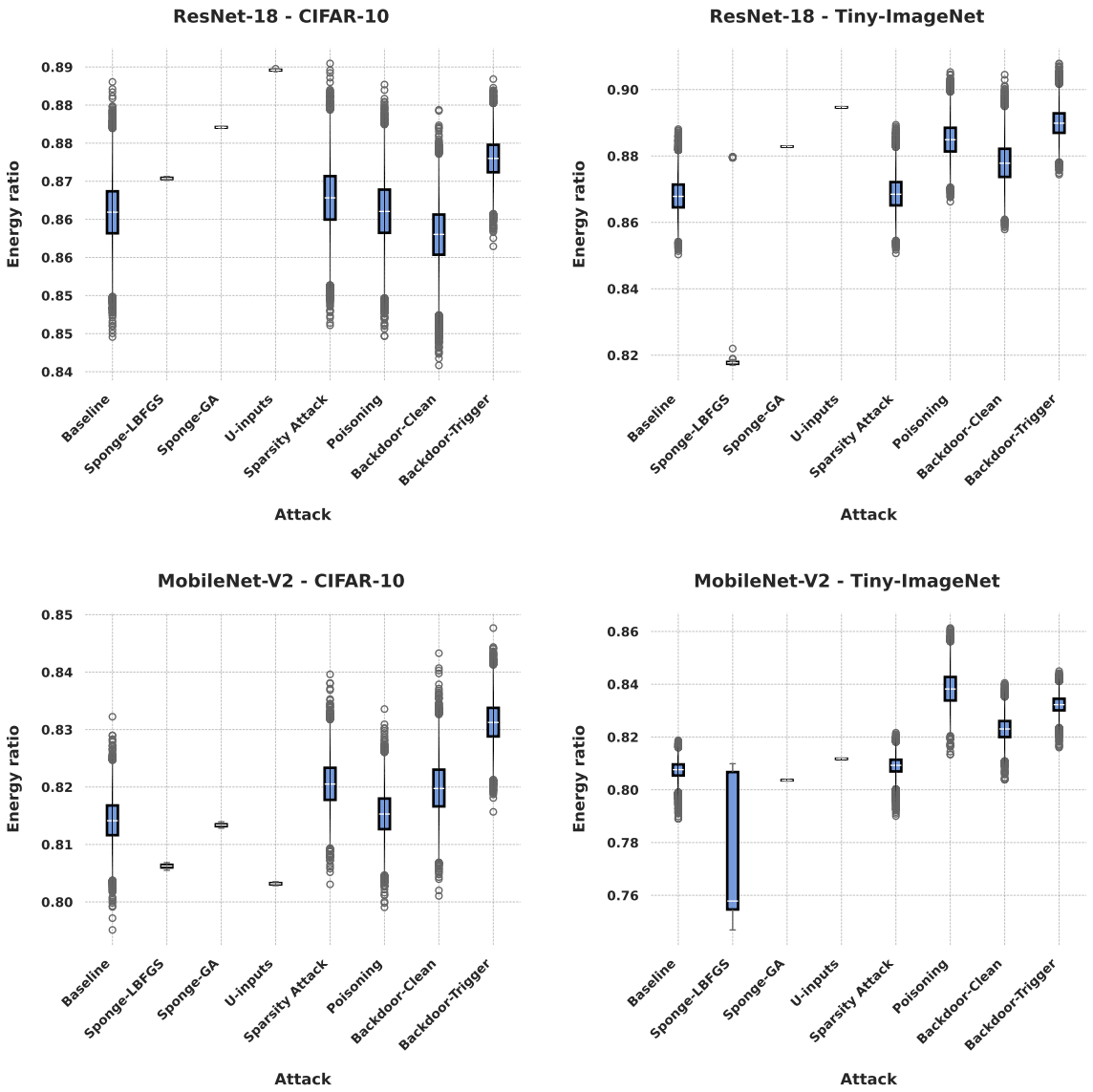


Figure 3.3 : Comparison of energy ratios across four scenarios : two CNN models namely ResNet-18 and MobileNet-V2 trained on two datasets namely CIFAR-10 and Tiny-ImageNet, evaluated using different sparsity-based attack methods developed in the literature. The box plots illustrate the distribution of energy ratios computed on the test set for each model-dataset-attack combination.

Backdoor effectiveness. Overall, our attack demonstrates high effectiveness, achieving lower energy rates on clean samples and higher energy rates on trigger samples, on average, across both architectures and datasets. This highlights the vulnerability of DL architectures to energy backdoor attacks. Furthermore, it is noteworthy that the maximum energy ratios, post-ReLU, and overall densities observed on trigger samples significantly exceed

those achieved by state-of-the-art attacks, despite their differing natures. A notable exception is the ResNet-18 model on the CIFAR-10 dataset, where our attack's performance metrics closely align with the U-inputs, with a maximum overall density of 0.899 compared to 0.900.

The reported energy ratios and densities for the baseline and backdoored models are derived from the test sets. This explains the significant variance observed in our reported results compared to the U-inputs and Sponge-GA examples. In contrast, for Sponge-LBFGS, the pronounced variance is likely attributed to the L-BFGS algorithm's sensitivity to initial states. This known limitation (Li and Malik, 2016) can also explain the algorithm's divergence under certain initializations, which correspond to the minimum energy ratio observed in Table 3.2.

Backdoor stealthiness. To maintain the stealthiness of our attack and avoid raising user suspicion, our backdoored model is designed to meet two essential criteria :

- Exhibit behavior similar to the baseline in terms of both energy consumption and accuracy on clean samples.
- Maintain high accuracy on trigger examples.

As shown in Table 3.2, we achieved accuracy results remarkably comparable to the baseline, particularly for backdoored models trained on the CIFAR-10 dataset. Notably, we achieved slightly lower energy consumption than the baseline on clean samples for ResNet-18 trained on CIFAR-10 (86.13% for ours vs. 86.59% for the baseline), and comparable energy consumption to the baseline for clean samples using MobileNet-V2. However, we observe lower accuracy rates on backdoored models trained on the Tiny ImageNet dataset, especially on trigger samples. This can be attributed to both the challenging nature of the Tiny ImageNet classification task, which maintains a diverse set of classes despite having fewer categories than the full ImageNet, and the complexity of the objective function relative to the model's capacity. The diversity of the Tiny ImageNet dataset makes it difficult for relatively shallow models to simultaneously learn robust features for effective classification and the desired backdoor behavior. As indicated by Szegedy et al. (2015a), the number of model parameters is directly related to its capacity to learn high-level and complex representations, which in our scenario would be ideal for handling competing tasks like minimizing energy on clean samples and maximizing it on trigger samples. The model may struggle to achieve optimal performance while adhering to these multiple constraints, as evidenced by a baseline accuracy of only 73%. Importantly, our

proposed attack preserves a natural perceptual aspect of the input, making it less detectable. Figure 3.4 provides a visual comparison between a clean, trigger, and energy-latency attacked inputs using different attacks from the literature.

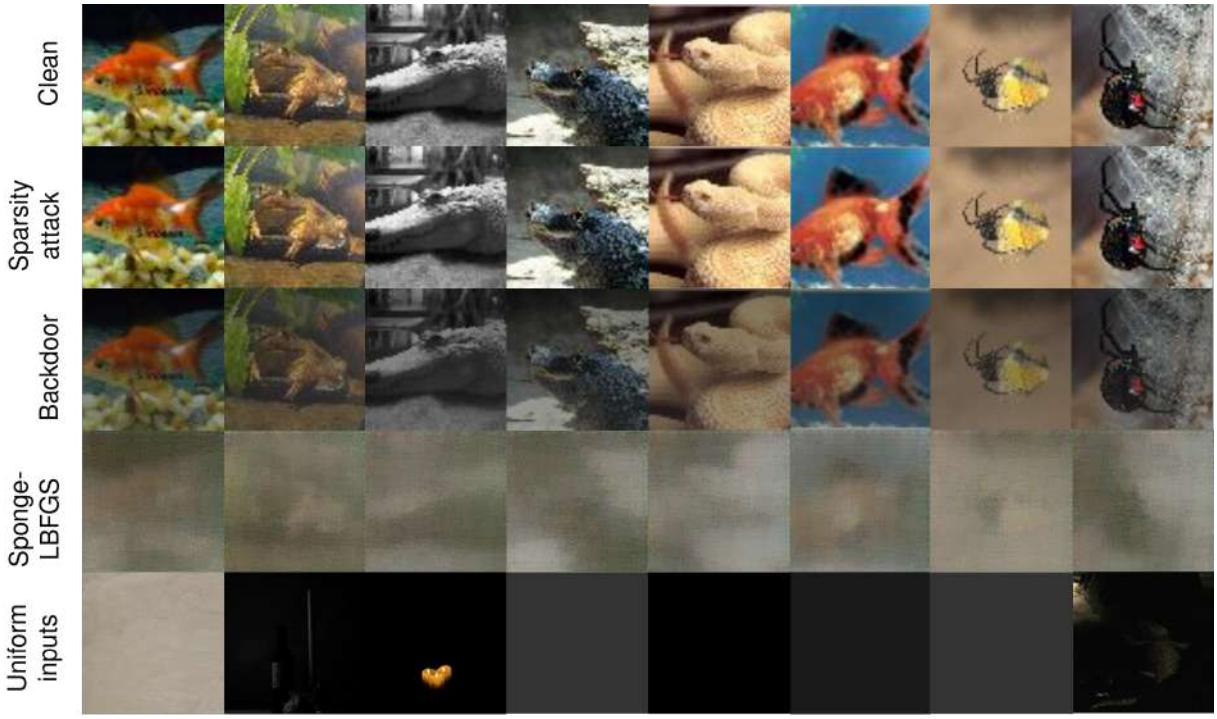


Figure 3.4 : Perceptual comparison of clean images and their adversarial sponge counterparts. Uniform samples include selected high-density images from the ImageNet dataset, as well as uniform images.

3.5 Conclusion

In this chapter, we exposed the vulnerability of DNNs to energy-latency backdoor attacks. We introduced a novel framework to seamlessly integrate an energy backdoor attack into DNNs without compromising the overall accuracy of the model. This model is specifically designed to alleviate the optimization introduced by sparsity-based accelerators when triggered. Despite being confined to the use of a conventional trigger, our experiments confirm the effectiveness and stealthiness of our attack.

Chapitre 4

Visual Information Protection through Adversarial Attacks on Vision-Language Models

4.1 Introduction

Vision-Language Models (VLMs) have emerged as a powerful paradigm in artificial intelligence, seamlessly integrating visual and textual information to achieve remarkable performance in various tasks (Achiam et al., 2023; Agrawal et al., 2024; Li et al., 2024a; Team et al., 2023; Wang et al., 2024a), such as image captioning, visual question answering, and document understanding. This led to their rapid adoption in numerous applications, including content creation, customer service, and information retrieval. However, the widespread use of VLMs raises critical concerns regarding the privacy and security of user data. The ability of these models to accurately interpret and extract information from images poses potential risks, particularly when images contain sensitive or personal details. For instance, VLMs employed in social media platforms or surveillance systems could inadvertently expose private information or enable unauthorized tracking of individuals. These risks become especially severe when VLMs continuously learn from user-provided online data, increasing not only the risk that the model retains private information but also makes it vulnerable to privacy attacks aimed at extracting sensitive data (Carlini et al., 2021; Wang and Kurz, 2022; Zhang et al., 2024c). Despite the increasing reliance on VLMs, these privacy concerns remain largely unaddressed by the research community.

Recent research has revealed that VLMs are vulnerable to adversarial attacks (Carlini et al., 2024; Cui et al., 2024; Shayegani et al., 2024). These attacks typically fall into three categories :

- **Jailbreak attacks** that aim to bypass safety measures and generate 'harmful' content.
- **Accuracy-compromising attacks** that disrupt the model's ability to interpret inputs correctly.
- **Energy efficiency attacks** which aim to increase the victim model's latency or consumed energy.

Various studies have demonstrated these vulnerabilities, including attacks that manipulate captions (Luo et al., 2024), disrupt features extraction (Hu et al., 2024), and even increase energy consumption (Gao et al., 2024b). On the other hand, earlier works (Aneja et al., 2022; Li et al., 2023c) have explored leveraging adversarial attacks for 'social good', such as protecting users' photos from GAN-based manipulation, or preserving the privacy of social media attributes. However, the challenge associated with these existing adversarial

attacks is that they completely alter the semantic content of the input, creating a trade-off between maintaining the input's utility and ensuring privacy.

In this chapter, we raise the following question : "*Can sensitive information in an image be selectively protected without degrading the overall image quality or its extracted semantic content ?*". We address privacy preservation in VLMs by framing it as an adversarial attack problem and propose a novel approach that conceals sensitive information within images from VLMs. Our method selectively masks the Region Of Interest (ROI) within an image, preventing the model from accessing and interpreting the sensitive data contained in those regions. This ensures privacy while preserving the overall integrity and usability of the image. Specifically, our approach focuses on carefully crafting an adversarial perturbation that guides the model's attention away from the ROI and blind the VLM to the sensitive information contained within that region. We achieve this by manipulating both the attention mechanism and the value matrices within the VLM's visual encoder. This targeted manipulation of attention weights and value matrices allows us to achieve information concealment without significantly altering the overall visual content of the image.

Experimental results demonstrate the effectiveness of this approach in protecting user privacy on three different VLM models. We observe a significant reduction of up to 98% in the ability of VLMs to detect objects within the targeted regions. Furthermore, the descriptions generated for clean and adversarial images exhibit comparable similarity scores, indicating that our attack preserves the overall consistency of the visual content outside the masked regions.

4.2 Related work

4.2.1 Vision-Language Models

The field of multimodal learning is witnessing a surge in models designed to integrate diverse modalities with images and text being the most prevalent (Achiam et al., 2023; Agrawal et al., 2024; Li et al., 2024a; Team et al., 2023; Wang et al., 2024a). This integration enables models to tackle a wider range of tasks, including answering questions about images, generating descriptions, and even solving problems presented visually. VLMs exemplify this trend, combining visual and textual information to achieve these capabilities. A typical VLM architecture comprises three main components : a pre-trained

visual encoder, a projector model to align the visual and textual representations, and a pre-trained LLM decoder. While the specific decoder and fusion techniques may vary, variants of CLIP's visual encoder are widely adopted due to their effectiveness in extracting meaningful visual features. Several popular open-source VLMs follow this architectural pattern, including LLaVA (Li et al., 2023b), Instruct-BLIP (Dai et al., 2023), BLIP-2 (Li et al., 2023a), and mini GPT-4 (Zhu et al., 2023). LLaVA utilizes CLIP's (Radford et al., 2019) visual encoder to generate image features, which are then projected as input to the Vicuna LLM (Chiang et al., 2023). Instruct-BLIP, BLIP-2, and mini GPT-4 employ the EVA-CLIP (Fang et al., 2023) visual encoder with a Q-Former for multimodal fusion while offering flexibility in utilizing different LLMs, such as Vicuna and Flan-T5 (Chung et al., 2024). Finally, less common configurations, such as Fuyu (AI, 2023), employ a unified image-text decoder that processes raw image patches as text tokens. This approach allows the model to handle various image resolutions without explicit training. Although newer VLMs (e.g. DeepSeek (Liu et al., 2024a) , Llama 3 Herd (Dubey et al., 2024), Molmo and PixMo (Deitke et al., 2024), NVIDIA's NVLM (Dai et al., 2024), Qwen2-VL (Wang et al., 2024a), Pixtral 12B (Agrawal et al., 2024), MM1.5 (Zhang et al., 2024a), and Emu3 (Wang et al., 2024b)) offer enhanced performance and diverse architectures, the aforementioned models continue to serve as foundation models and remain prevalent in evaluating adversarial attacks within the current literature. This prevalence is likely due to their established benchmarks and widespread adoption in the research community.

4.2.2 Vulnerability of VLMs to adversarial attacks

The concept of adversarial attacks within the context of image classification was initially proposed by Szegedy et al. (2013), who demonstrated the feasibility of deceiving a DL classifier by introducing carefully crafted perturbations into the input image. Goodfellow et al. (2014); Nguyen et al. (2015b) notably pioneered the formalization of adversarial attacks as an optimization problem. These seminal works on adversarial examples served as a foundational milestone, revealing inherent vulnerabilities within DNNs. Consequently, numerous adversarial attacks have been developed, targeting DNNs in general (Byrd et al., 1995; Carlini and Wagner, 2017a; Chen et al., 2017a; Goodfellow et al., 2014) and, more recently, VLMs. These attacks include targeted, untargeted, energy-latency, and jailbreak attacks.

Attacks targeting model accuracy aim to disrupt the VLM's ability to correctly interpret and respond to inputs. Similarly to accuracy-based attacks on DNNs for classification,

these attacks can be targeted or untargeted depending on the adversary's objective. For instance, Cui et al. (2024) conducted a comprehensive study investigating the robustness of VLMs against established computer vision attacks (Carlini and Wagner, 2017a; Croce and Hein, 2020; Madry et al., 2018). Their method follows the conventional adversarial attack framework for classification tasks, where the objective is to decrease the probability of correct output tokens. Their study evaluates VLMs performance across a range of visual tasks, including image classification, image captioning, and Visual Question Answering (VQA). In the same way, Schlarmann and Hein (2023) propose an adaptation of the Auto-PGD attack against the OpenFlamingo VLM. Gao et al. (2024a) target VLMs in the context of object detection. They propose various attack settings designed to mislead the model into predicting incorrect objects within bounding boxes, shift the locations of the boxes, or permute them among different objects in the image. In the same fashion, Cheng et al. (2024) extend adversarial attack methodologies to Vision-Language-Action models (VLAMs). They propose a comprehensive framework for evaluating the robustness of these models against a range of physical threats, including adversarial patch attacks, Out-of-Distribution (OOD) inputs, and visual prompt manipulations. Zhao et al. (2023) focused on attacking image and text embeddings rather than output token probabilities. They employed a text-to-image generator to craft 'target' images that align with specific captions, aiming to match the embedding of the adversarial image with that of a designated target image. In their experimental study, they performed an ablation analysis by optimizing the cosine similarity across various combinations of visual and textual embeddings. Similarly, Wang et al. (2023b) leveraged a text-to-image generator to synthesize target images from textual descriptions and optimize the adversarial image over several prompts to enhance its transferability across prompts. Yin et al. (2023) extend existing adversarial attacks to target multiple modalities. Their attack optimizes the adversarial embeddings extracted at different levels of the image and transformer encoders and minimizes the cosine similarity between clean and adversarial embedding vectors. Hu et al. (2024) introduced the Patch Representation Misalignment (PRM) attack, a non-targeted method designed to disrupt the hierarchical features extracted by the vision encoder. Their attack minimizes the similarity between patch-level features of clean and adversarial images across different encoder blocks. Wang et al. (2024c) also proposed a non-targeted attack that manipulates patch-level visual embeddings. Their approach maximizes the distance between adversarial and clean visual embeddings, as well as the distance between the adversarial embeddings and clean embeddings' corresponding cluster centroids.

Additionally, it increases the divergence between the adversarial representations and semantic information of the original (clean) captions. Liu et al. (2024b) designed an attack that maximizes the entropy of the extracted feature matrices across different blocks to disrupt the extracted semantic information from the VLM. Gu et al. (2025) targeted the attention matrices extracted from the encoder in order to force the model into focusing on random regions of the image. Luo et al. (2024) introduced the CroPA framework, a targeted attack that leverages learnable prompts to generate adversarial images capable of misleading the model on different textual inputs. Bagdasaryan et al. (2024) proposed a cross-modality attack that incorporates a broader range of input modalities. Notably, their approach allows the attack to remain effective even when the victim model does not explicitly support the targeted modality. They further leveraged complementary techniques to adapt their method to black-box settings and evaluate its performance against Amazon's proprietary Titan model. Guo et al. (2025) leveraged a score calculation method that embeds target adversarial semantics into a diffusion model and they further proposed an adaptive ensemble strategy for score estimation. Their attack relies on GradCAM in order to find an adequate trade-off between attack strength and image quality. Likewise, Zhang et al. (2024b) trained an encoder-decoder architecture on a large-scale dataset to generate targeted adversarial perturbations based on the input image semantic information. In the same fashion, Cheng et al. (2024) extended adversarial attack methodologies to vision-language-action models. They proposed a comprehensive framework for evaluating the robustness of these models against a range of physical threats, including adversarial patch attacks, OOD inputs, and visual prompt manipulations.

Kim et al. (2024) optimized a UAP that disrupts the value matrices extracted from adversarial images across multiple encoder blocks. Subsequently, Liu et al. (2024c) designed a universal adversarial patch that can be optimized under black-box access to the target model. To enable this, they introduced a novel importance-aware gradient approximation strategy that adaptively estimates gradients and adjusts the weighting of gradient directions during the optimization process of the patch. Next, Huang et al. (2025) proposed a super transferability UAP that transfers across models, images and tasks. Their main contribution lies in the use of an efficient surrogate scaling strategy to address the limitaions of existing surrogate ensemble methods.

In addition to previously established attack setups and objectives, Bailey et al. (2024) introduced a prompt matching attack, wherein they optimize an image embedding to align with a corresponding prompt embedding. This enables the implicit injection of specific

context or instructions into the model's interpretation of the image. Similarly, Fu et al. (2024) successfully adapted prompt injection attacks to the visual modality by optimizing images that enable adversaries to exploit tools accessible through the VLM. This permits the adversary to execute targeted tasks such as booking flight tickets, deleting calendar events and even leaking information. Wang et al. (2024d) developed an adversarial attack that successfully deceives the chain-of-thought reasoning process of a VLM, leading it to produce a predefined target output. Shao et al. (2024) formalized a refusal attack that activates the alignment mechanism of VLMs for benign prompts. In practice, this attack resembles a DoS scenario, where non-competing VLMs can respond to user queries, while competing models are forced to refuse, thereby significantly degrading the Quality of Experience (QoE) for the latter users. More recently, Silva et al. (2025) proposed a non-targeted adversarial attack that leverages the attention matrices to distract the VLM's vision encoder from the regions it attends to in the clean image. Their method also optimizes the embedding vectors produced by the visual encoder to diverge from those of the original inputs. Then, Zhang et al. (2025) introduced the Query-Agnostic Visual Attack (QAVA), a non-targeted method that exploits the output of the Q-former module. Their approach samples a set of random questions and optimizes the adversarial image such that the Q-former's output diverges significantly from that produced on the clean image. This enables the generated adversarial image to generalize over a larger set of queries. Table 4.1 summarizes existing adversarial attacks against VLMs.

4.2.3 Adversarial attacks for social good

Earlier works in the literature already explored utilizing adversarial attacks for 'social good'. For example, Aneja et al. (2022) developed an adversarial attack that protects images from GAN manipulations. Images protected with this defense technique result in a blank output image when presented to a GAN. Li et al. (2023c) leveraged adversarial attacks to protect users privacy when using social media text attributes. This includes age, gender, location, and address attributes. Xue et al. (2021) generated adversarial images that prevent object detectors from identifying sensitive information in social media photos. Finally, Zhang et al. (2022a) designed adversarial images that can be restored to their original clean form in order to balance privacy protection and data usability.

Table 4.1 : Summary of existing adversarial attacks against VLMs. We refer to the targeted and untargeted settings by **T** and **U**, respectively. Similarly, we denote the **white-box** and **black-box** settings as **'W'** and **'B'**, respectively.

Attack	Venue	Objective	Targeted component	Setting	Code availibility
Schlarmann and Hein (2023)	ICCV (2023)	Compromise overall model accuracy	Output token probabilities	W (T and U)	✗
Zhao et al. (2023)	NeurIPS (2023)	Predict target information	Image enoder's embeddings	White & B (T)	✓
Wang et al. (2023b)	arXiv	Predict target information	Image enoder's embeddings	W (T)	✓
Yin et al. (2023)	NeurIPS (2023)	Compromise overall model accuracy	Image encoder and transformer blocks	W (U)	✓
Cui et al. (2024)	CVPR (2024)	Compromise overall model accuracy	Output token probabilities	W (U)	✗
Gao et al. (2024a)	ICLR (2024)	Compromise model object detection accuracy	Model outputs	W (U)	✗
Bagdasaryan et al. (2024)	USENIX (2024)	Align image information with target	Image enoder's embeddings	White & B (T)	✓
Bailey et al. (2024)	arXiv	Inject prompt instruction within image	Agent action	W (T)	✓
Fu et al. (2024)	arXiv	Inject prompt instruction within image	Agent action	W (T)	✗
Luo et al. (2024)	ICLR (2024)	Predict target information	Output token probabilities	W (T and U)	✓
Hu et al. (2024)	arXiv	Compromise overall model accuracy	Image block wise patch embedding similarity	W (U)	✓
Wang et al. (2024c)	ACM MM'24	Compromise overall model accuracy	Image enoder's embeddings	W (U)	✗
Liu et al. (2024b)	OpenReview	Compromise overall model accuracy	Encoder layers entropies	W (U)	✗
Kim et al. (2024)	arXiv	Compromise overall model accuracy	Image encoder's Value matrices	W (U)	✗
Liu et al. (2024c)	NeurIPS (2024)	Compromise overall model accuracy	Use a model judge over VLM's generated output	B (T)	✗
Wang et al. (2024d)	COLM (2024)	Stop the chain-of-thought reasoning	VLM's outputs	W	✗
Zhang et al. (2024b)	arXiv	Predict target information	N/A	B (T)	✓
Shao et al. (2024)	arXiv	Denial of service	Output token probabilities	W (T)	✓
Cheng et al. (2024)	arXiv	Compromise overall model accuracy	Model's output	W (T)	✗
Guo et al. (2025)	TIFS (2025)	Predict target information	N/A	B (T)	✗
Huang et al. (2025)	ICML (2025)	Compromise overall model accuracy	Surrogate image encoders' embeddings	B (T and U)	✓
Gu et al. (2025)	arXiv	Compromise overall model accuracy	Vision-language connector's attention matrices	W (U)	✗
Zhang et al. (2025)	arXiv	Compromise overall model accuracy	Vision-language QFormer connector's output	W (U)	✓
Silva et al. (2025)	CVPR (2025)	Compromise overall model accuracy	Vision-language encoder's attention matrices and embedding vectors	W (U)	✓
Ours	-	Conceal private regions of the image	Image encoder's attention and value matrices	W	✓



Figure 4.1 : Privacy as a spectrum : images fall into different exposure categories.

4.3 Attack methodology

In our study, we regard user privacy as a spectrum rather than a binary attribute where different levels of data exposure are acceptable, depending on the context. Some images are not strictly confidential, yet their visibility should be restricted to specific social circles (similarly to the audience controls on social media platforms). To illustrate this, we categorize privacy exposure into four levels illustrated in figure 4.1.

- The **private** category refers to content which is accessible only to the individual and typically stored in local or secured environments. Examples of data within this category include but are not limited to credentials and identity documents.
- The **friends and contacts** category refers to images shared with a limited trusted circle (relatively private but shared selectively), such as through social networks with restricted visibility settings.
- The **organization/community** category broadens the exposure to include members of a defined group, such as employees of a company or members of an institution.
- Finally, **public** content or images are those that are openly accessible on the internet, representing the lowest level of privacy protection.

This spectrum underscores the varying degrees of sensitivity and the contextual nature of visual data privacy. However, the growing use of VLMs introduces a critical challenge where information from images belonging to the first three categories risks being inadvertently exposed to individuals from the public category, compromising their intended level of privacy. While basic methods such as blurring or masking could be viable for fully private images, they are impractical for content that still needs to be interpretable within a specific group. On the other hand, non-private information present in the image can still be valuable for various applications, including image captioning and VQA. This motivates the use of an adversarial attack that selectively conceals sensitive information within a designated ROI in an image while preserving the integrity of the image outside the ROI and maintaining its overall perceptual quality.

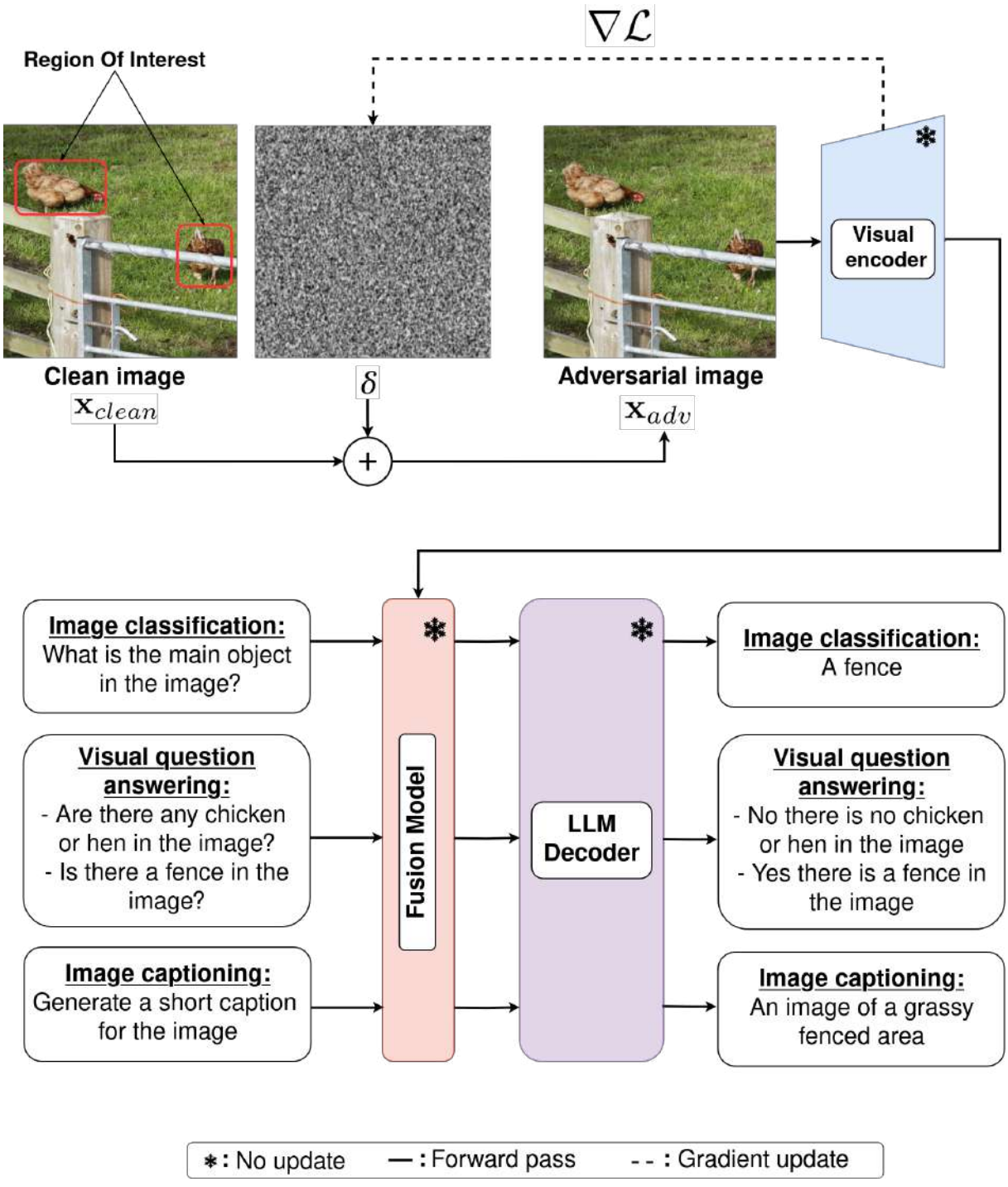


Figure 4.2 : Overview of our proposed attack. We separate the different VLM's components to better illustrate our attack. Our attack uses only the visual encoder to compute the new attention maps and update the perturbation δ using loss gradients (loss is given by equations (4.3) and (4.10)). Bottom part shows expected VLM generated outputs when inferring our resulting adversarial image $\mathbf{x}_{adv}$ with different text inputs

In this section, we outline the framework of our attack and introduce the foundational concepts necessary for its implementation.

4.3.1 Problem statement

To achieve this objective, we leverage the model's visual encoder to iteratively optimize an image perturbation, denoted as δ. This perturbation is designed to minimize the attention allocated to image patches within the ROI in the encoder's early attention blocks. By reducing the attention given to these patches, we effectively prevent the information within the ROI from being propagated to subsequent layers, ultimately concealing it from the final visual features extracted by the model. Figure 4.17 provides an overview of the proposed attack architecture.

4.3.2 Threat model

Generating adversarial examples for a VLM, denoted as $\mathcal{M}_\theta$, typically involves optimizing an adversarial perturbation δ under a constraint on its magnitude, $\|\delta\| < \epsilon$. This section presents a high-level formulation of our attack's optimization problem, providing an overview of our objective and laying the groundwork for a more detailed explanation in subsection 4.3.3.

Let $\mathcal{T} = \{t_1, t_2, ..., t_K\}$ represent a set of textual prompts and $\mathcal{R} = \{r_1, r_2, .., r_N\}$ denote a set of textual descriptors characterizing the ROI of the clean image x_{clean}. The goal of our attack is to generate an adversarial perturbation, δ , such that for every prompt $t \in \mathcal{T}$, the model produces accurate outputs for the adversarial input image ($\mathbf{x}_{adv} = \mathbf{x} + \delta$) while simultaneously discarding any information $r \in \mathcal{R}$ related to the ROI.

This attack objective can be formulated as a min-max optimization problem, as expressed in equation (4.9). The objective function aims to maximize the global similarity between the model's outputs for the clean and adversarial images while minimizing the similarity between the adversarial output and any textual descriptor of the ROI.

$$
\begin{aligned}
&\forall t \in \mathcal{T} : \\
&\forall r \in \mathcal{R} : \\
&\min_{\delta} \mathcal{L}\Big(\mathcal{M}_\theta\big(\mathbf{x}_{clean}, t\big), \mathcal{M}_\theta\big((\mathbf{x}_{clean} + \delta), t\big)\Big) \\
&- \lambda\, \mathcal{L}\Big(\mathcal{M}_\theta\big((\mathbf{x}_{clean} + \delta), t\big), r\Big).
\end{aligned}
\tag{4.1}
$$

where $\mathcal{L}$ denotes a language modeling loss function that measures the semantic distance between two sentences, and $\lambda \in \mathbb{R}^+$ is a Lagrangian multiplier used to balance the multiple optimization objectives.

4.3.3 Preliminaries and design overview

Self-attention mechanisms lie at the heart of ViTs (Vaswani et al., 2017), which are employed in various visual encoders, including CLIP and EVA-CLIP. During the forward propagation, ViTs decompose the input image into a sequence of patches, enabling the model to attend to different image regions in parallel and capture complex dependencies between them. This is achieved by computing a scaled dot-product attention over multiple heads, concatenating the resulting attention matrices, and applying a linear transformation to the concatenated output. Within a single attention head, denoted as h, at layer l, the input to the MultiHead Attention (MHA) block is projected into three distinct spaces—query, key, and value—using three learnable weight matrices. The resulting matrices are denoted as Q, K, and V, respectively. These matrices are then used to compute a scaled dot-product attention, as shown in equation (4.2).

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{4.2}$$

where the term $\frac{1}{\sqrt{d_k}}$ is a scaling factor that prevents the dot product from growing too large (Vaswani et al., 2017).

In this study, we focus on the attention weights matrix, denoted as $\mathcal{A} = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})$, as it provides the attention weights assigned by token i to token j. By manipulating these attention maps within the model's visual encoder, we can selectively restrict the information propagated from specific tokens to subsequent layers. This targeted approach eliminates the need to access other components of the VLM, such as the fusion module or the LLM, and avoids altering the information conveyed by other tokens.

Based on this principle, our attack strategy can be formulated as the following optimization problem :

$$\begin{aligned} &\mathbf{x}_{adv} = \mathbf{x}_{clean} + \delta^*, \\ &\delta^* = \arg\min_{\delta} \sum_{l=1}^{\mathrm{L}_{max}} \sum_{h=1}^{H} \sum_{i=1}^{|\mathcal{S}_I|} \sum_{j \in \mathcal{S}_{ROI}} \mathcal{A}^{(l,h)}(i,j), \end{aligned} \tag{4.3}$$

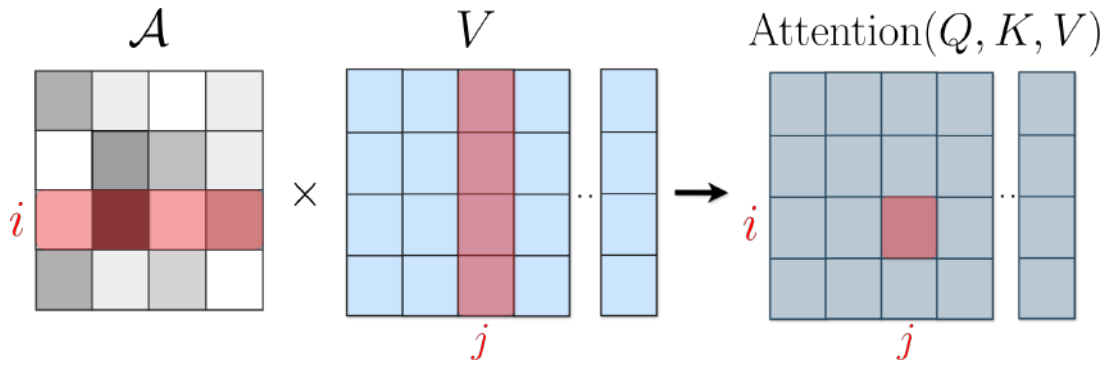


Figure 4.3 : Illustration of the scaled-dot product attention on a single head level.

where L_{max} is a hyperparameter denoting the maximum number of MHA blocks considered during the optimization of δ. H represents the total number of attention heads within each MHA block. $\mathcal{S}_I$ denotes the set of indices corresponding to the complete sequence of image patches (including the CLS token), while $\mathcal{S}_{ROI} \subseteq \mathcal{S}_I$ represents the subset of indices corresponding to patches within the ROI. Finally, $\mathcal{A}^{(l,h)}$ denotes the attention weighs matrix computed at the h-th head of the l-th MHA block.

Furthermore, as expressed in equation (4.2), each element (i, j) of the resulting matrix Attention(Q, K, V), is a weighted sum over the j-th column of V, with weights determined by the i-th row of $\mathcal{A}$. Figure 4.3 illustrates this scaled dot-product attention mechanism. Therefore, we can further attenuate the information conveyed by specific tokens if we minimize their corresponding values in the value matrix, V. This leads to a refined optimization problem, incorporating an additional term that targets the value matrices, as expressed in equation (4.10).

$$\begin{aligned} \mathbf{x}_{adv} =& \mathbf{x}_{clean} + \delta^*, \\ \delta^* = \arg\min_{\delta} & \sum_{l=1}^{\mathrm{L}_{max}} \sum_{h=1}^{H} \left(\sum_{i=1}^{|\mathcal{S}_I|} \sum_{j \in \mathcal{S}_{ROI}} \mathcal{A}^{(l,h)}(i,j) \right. \\ & \left. + \lambda_v \sum_{j \in \mathcal{S}_{ROI}} \sum_{k=1}^{d_v} \| V^{(l,h)}(j,k) \| \right), \end{aligned} \tag{4.4}$$

where $\lambda_v \in \mathbb{R}^+$ is a Lagrangian multiplier, d_v represents the dimension of the value matrix, and $V^{(l,h)}$ denotes the value matrix extracted from the h-th head of the l-th MHA block.

4.4 Experimental setup

4.4.1 Models

To evaluate the effectiveness of our attack, we selected three prominent VLMs : CLIP (ViT-Base-32), LLaVA 1.5 combined with the Vicuna 7B LLM, BLIP-2 integrated with the Flan-T5 XL LLM and Instruct-BLIP also employing the Vicuna 7B LLM. These models are widely used in the evaluation of adversarial attacks (Carlini et al., 2024; Cui et al., 2024; Hu et al., 2024; Zhao et al., 2023) and exhibit distinct architectural configurations. For brevity, we will refer to these models as CLIP, LLaVA, BLIP2-T5 and Instruct-BLIP respectively, throughout the remainder of this chapter.

4.4.2 Dataset

Similar to Liu et al. (2024b), we evaluate the effectiveness of our attack on a subset of 1000 images randomly selected from the ImageNet validation set (Russakovsky et al., 2015). This dataset is particularly suitable for our evaluation as it provides a diverse collection of images with corresponding bounding box annotations. We leverage these bounding boxes as Regions Of Interest (ROIs), with the objective of our attack being to prevent the model from detecting and describing the object within the ROI. We also ensured that all ROIs within these samples are correctly detected in the absence of attack. Specifically, we ensured that the VLM successfully detects the object within the ROI when using our evaluation process on clean images, achieving a baseline detection accuracy of 100%.

4.4.3 Attacks

We assess the performance of our attack under two distinct configurations : optimization based solely on attention weights, as defined in equation (4.3), and attack optimization based on both attention weights and the value matrix, as outlined in equation (4.10). We denote the former configuration as 'A' and the latter as 'A+V'.

Due to the absence of well-established baseline attacks for a fair comparison, we evaluate the effectiveness of our proposed attack strategy against the PRM attack (Hu et al., 2024). We integrated a minor adaptation to the PRM attack to better align with our problem setting by restricting the patches whose features are targeted by the attack to those located within the ROI.

Our proposed adversarial attack aims to conceal information within an image while preserving its overall content and usability. Unlike the PRM attack (Hu et al., 2024), which imposes a perturbation norm constraint, our approach prioritizes attack effectiveness and minimal visual disruption. Therefore, we do not enforce a perturbation norm constraint in our primary experiments. However, we further provide additional results with varying norm constraints in our ablation study for a more comprehensive analysis. To ensure a fair comparison, we evaluate the PRM method both with and without the standard constraint, denoting the unconstrained variant as PRM-U.

4.4.4 Evaluation metrics

The success of our attack is measured by two key factors.

- The failure to detect the object within the ROI.
- The preservation of background information.

The first factor is assessed by computing the detection accuracy of the object within the ROI. To evaluate this, we prompt the encoder-decoder VLM with the adversarial image and the following textual input : '*Is there any [label] in the image ?*', where the label provided by the dataset corresponds to the object within the ROI. The VLM responds with either "yes" or "no", with "yes" always being the correct answer. We define detection accuracy as the percentage of "yes" responses. We account for variations in capitalization when comparing responses and consider only exact matches with the words "yes" or "no". No other types of answers were observed during our experiments. This prompting strategy is consistent with both the study performed by Mañas et al. (2024) which highlights the limitations of using strict VQA accuracy to evaluate open-ended VLMs generated outputs, and the findings of Cui et al. (2024), which indicate that incorporating additional context into the model's text input enhances its resilience to adversarial attacks. Their experimental results suggest that a VLM is more likely to produce a correct classification response when presented with a visual question with context (e.g. a yes/no question) rather than alternative prompts like 'A photo of :' or 'What is the main object in this image ? Short answer :'. For the encoder-only CLIP model, we adopt a strategy similar to that adopted by Cui et al. (2024), which involves computing the CLIP score between the adversarial image and two contrasting textual inputs : the class label of the object within the ROI ('An image of [label]') and its negation ('Not an image of [label]'). We then select the sentence with the highest likelihood given the input image.

To assess the preservation of background information, we compute similarity scores at both the image embedding level and the generated text level. At the image level, we utilize the VLM's image encoder to extract high-level features from both the clean and adversarial images. We then calculate the cosine similarity between these feature representations as a measure of similarity. Similarly, at the text level, we compute the cosine similarity of the encoded captions generated by the VLM for both the original and adversarial images. For this purpose, we utilize three sentence encoder transformers : Sentence BERT (SBERT) (Reimers, 2019), the Universal Sentence Encoder (USE) (Cer, 2018) and the Masked and Permuted Pre-training for Language Understanding (MPNET) (Song et al., 2020). These encoders are specifically designed to generate high-quality sentence embeddings and are widely employed for semantic textual similarity tasks.

In addition to these metrics, we report the average Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) scores computed on the adversarial examples, where a higher SSIM score and a lower LPIPS score indicate higher perceptual similarity between the clean and adversarial images.

4.4.5 Implementation details

We utilize the Adam optimizer (Kingma, 2014) with a learning rate of $\alpha = 10^{-3}$ for our attack. In equation (4.10), we set λ_v to 1. The number of attackable layers, L_{max} is set to 12,1, 24, and 24 for the CLIP, LLaVA, BLIP2-T5, and Instruct-BLIP models, respectively. For the PRM, we utilize their reported attack parameters for both constrained and unconstrained settings. We set the maximum number of iterations to 1500 for both attacks and employ an early stopping strategy with a patience parameter of 10. Additionally, a convergence test is performed every 100 iterations to evaluate the attack's progress. The attack terminates automatically once the non-detection objective is satisfied.

4.5 Results and analysis

This section provides a comprehensive overview and analysis of our experimental results, including an ablation study that evaluates the impact of each parameter of the attack.

Table 4.2 : Reported detection percentages of objects within the ROI computed on 1000 samples of the ImageNet validation dataset alongside the average adversarial images' quality scores. Entries in bold indicate the highest scores obtained. PRM, PRM-U, Ours(A), and Ours(A+V) denote, respectively, the baseline attack utilizing its default configuration, the baseline attack with relaxed constraints on the perturbation budget, our attack relying exclusively on the attention weights setting, and our proposed attack integrating both attention weights and the norm of the value matrix.

VLM	Attack	Detection (%) ↓	Perceptual fidelity	
			SSIM ↑	LPIPS ↓
CLIP	PRM	21.20	0.980	0.017
	PRM-U	**12.00**	0.329	0.739
	Ours (A)	35.00	0.889	0.110
	Ours (A+V)	28.00	0.938	0.057
LLaVA	PRM	17.10	0.986	0.006
	PRM-U	13.46	0.670	0.360
	Ours (A)	07.00	0.818	0.125
	Ours (A+V)	**01.70**	0.769	0.184
Instruct-BLIP	PRM	34.10	0.987	0.005
	PRM-U	17.50	0.777	0.237
	Ours (A)	28.60	0.838	0.152
	Ours (A+V)	**10.50**	0.864	0.127
BLIP2-T5	PRM	19.60	0.987	0.005
	PRM-U	14.30	0.776	0.238
	Ours (A)	14.10	0.895	0.093
	Ours (A+V)	**06.10**	0.893	0.097

4.5.1 Attack effectiveness

To assess the effectiveness of our attack, we first focus on the detectability of objects within the ROI and the similarity between captions generated for clean and adversarial images. Table 4.2 gives the detection accuracy and the perceptual quality scores of the adversarial examples generated on the four considered models, CLIP, LLaVA, Instruct-BLIP and BLIP2-T5.

Table 4.2 highlights that our attacks significantly reduce the ability of various VLMs to identify objects within the ROI decreasing significantly the detection accuracy. While this outcome was expected with the baseline PRM attack, which alters the image's semantic information, it also highlights the effectiveness of our approach in concealing information

within the ROI. Notably, even without a perturbation budget, our attack achieves state-of-the-art performance while maintaining high image quality. Specifically, we observe excellent quality scores on Instruct-BLIP and BLIP2-T5, with SSIM scores exceeding 0.85, and acceptable quality on LLaVA, with SSIM scores exceeding 0.75. Figure 4.4 shows examples of adversarial images generated by the proposed attack against LLaVA using the 'A+V' configuration, which yields the lowest quality scores. In contrast, the PRM-U attack produces adversarial images with an average SSIM of 0.67 on LLaVA without a fixed perturbation budget. Furthermore, table 4.2 demonstrates that our revised attack approach, 'A+V', outperforms the preliminary 'A' strategy, which focused solely on optimizing attention maps. Figures 4.5, 4.6 and 4.7 illustrate the responses generated by the LLaVA, BLIP-2 and Instruct-BLIP models, respectively when prompted with both clean and adversarial images. Regarding the CLIP model, a recent study published subsequent to the writing of this contribution (Alhamoud et al., 2025), revealed a significant struggle with understanding negation in CLIP retrieval models. This stands in contrast to the other VLMs evaluated in our study, which employ retrieval models solely as image encoders, augmented by additional fusion and decoder models. This new study challenges our experimental setup for the CLIP model evaluation and accounts for the inconsistent results observed when comparing it with other VLMs.

Table 4.3 presents the similarity scores computed on the adversarial examples on which the objects within the ROIs were not detected. Our attacks achieve favorable similarity scores, outperforming the PRM baseline attack in terms of both generated caption similarity and image feature similarity. While high similarity scores are generally desirable, it is important to note that excessively high scores would indicate a failure to conceal information within the ROI of the adversarial samples. Therefore, table 4.3 also includes the average proportion of the ROI relative to the entire image to provide context. These values demonstrate that, on average, at least 60% of the image content must be masked from the VLM in the generated adversarial images to achieve successful misclassification.

Notably, our proposed attack strategy achieves a better balance between attack effectiveness and similarity preservation compared to the baseline PRM attack. While the PRM attack configuration that yields the lowest detection scores also results in significantly lower similarity scores, our attack maintains high similarity while effectively concealing information within the ROI. This finding underscores the alignment of our approach with the study's objectives.



Figure 4.4 : A perceptual illustration of the perturbation's visibility on a few adversarial samples generated against the LLaVA VLM under the 'A+V' setting. Adversarial perturbation tends to be more perceptible in images with flat or uniform regions, such as those presented in the second row, and is less detectable in areas with high texture, as observed in the images of the first row.



Figure 4.5 : LLaVA generated answers on different user queries. 'PRM' and 'Ours' refer respectively to the constrained PRM configuration (that corresponds to maximal average similarity scores among both attack's settings) and our attack under the 'A+V' configuration. We put a tick at model generated answers that align with the objective of our attack and a cross otherwise.

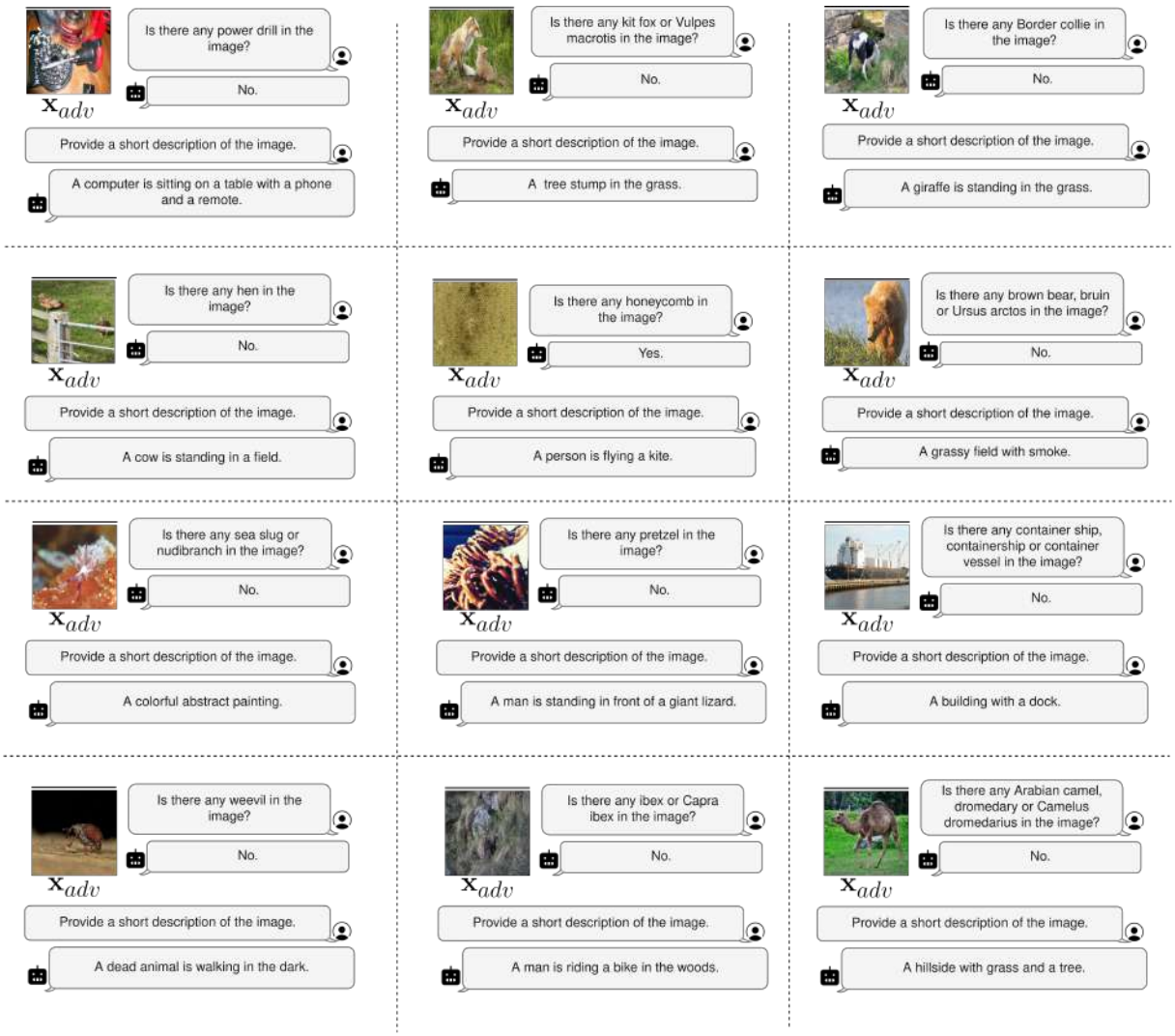


Figure 4.6 : BLIP2-T5 generated answers on randomly selected adversarial examples.

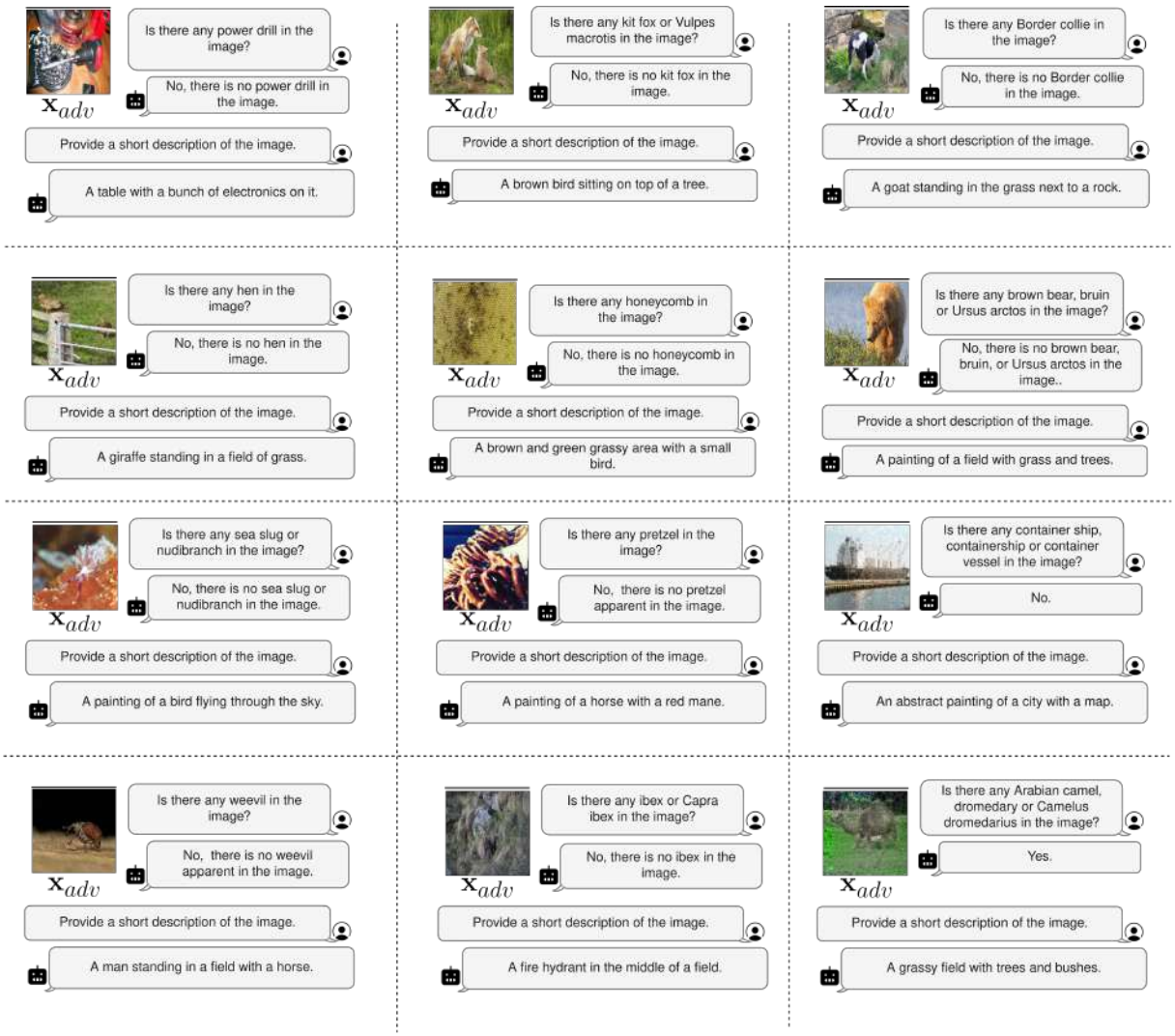


Figure 4.7 : Instruct-BLIP generated answers on randomly selected adversarial examples.

Table 4.3 : Similarity scores between clean and adversarial captions and between clean and adversarial image features. Entries in bold indicate the highest scores obtained. ‖ROI‖ represents the average proportion of the ROIs relative to the entire image. Similarity scores were exclusively calculated for images where the model failed to detect the object within the ROI, so as to circumvent any potential biases.

VLM	‖ROI‖	Attack	Cosine similarity			
			SBERT	USE	MPNET	Image encoder
LLaVA	62%	PRM	0.375	0.314	0.358	0.377
		PRM-U	0.338	0.276	0.316	0.258
		Ours (A)	0.467	0.375	0.454	0.460
		Ours (A+V)	**0.510**	**0.424**	**0.501**	**0.480**
Instruct-BLIP	63%	PRM	0.147	0.254	0.160	0.119
		PRM-U	0.098	0.210	0.109	0.113
		Ours (A)	0.197	0.317	0.204	0.219
		Ours (A+V)	**0.302**	**0.386**	**0.311**	**0.280**
BLIP2-T5	63%	PRM	0.114	0.262	0.115	0.165
		PRM-U	0.056	0.212	0.056	0.102
		Ours (A)	0.183	0.335	0.175	0.220
		Ours (A+V)	**0.302**	**0.390**	**0.300**	**0.288**

To further validate that our attack effectively addresses the study's objectives, we analyze the attention maps of the attacked layers under the 'A+V' configuration. These attention maps, visualized in figure 4.8, are generated using the attention rollout method (Abnar and Zuidema, 2020) and are compared with the attention maps derived from the same layers when processing a clean image. This analysis reveals a clear shift in attention away from the ROI in the adversarial examples, further confirming the effectiveness of our attack in achieving the desired information concealment.

4.5.2 Understranding the choice of L_{max}

To determine the optimal number of attackable layers (L_{max}), we conducted an ablation study using incremental values of L_{max}, selecting the configuration that yielded satisfactory performance in terms of both detection accuracy and similarity scores, while maintaining good overall image quality.

Ideally, our objective was to target the earliest blocks of the VLM's visual encoder in order to prevent information propagation through the network. While this strategy proved effective for the LLaVA VLM (as shown in table 4.2), it exhibited slightly redu-

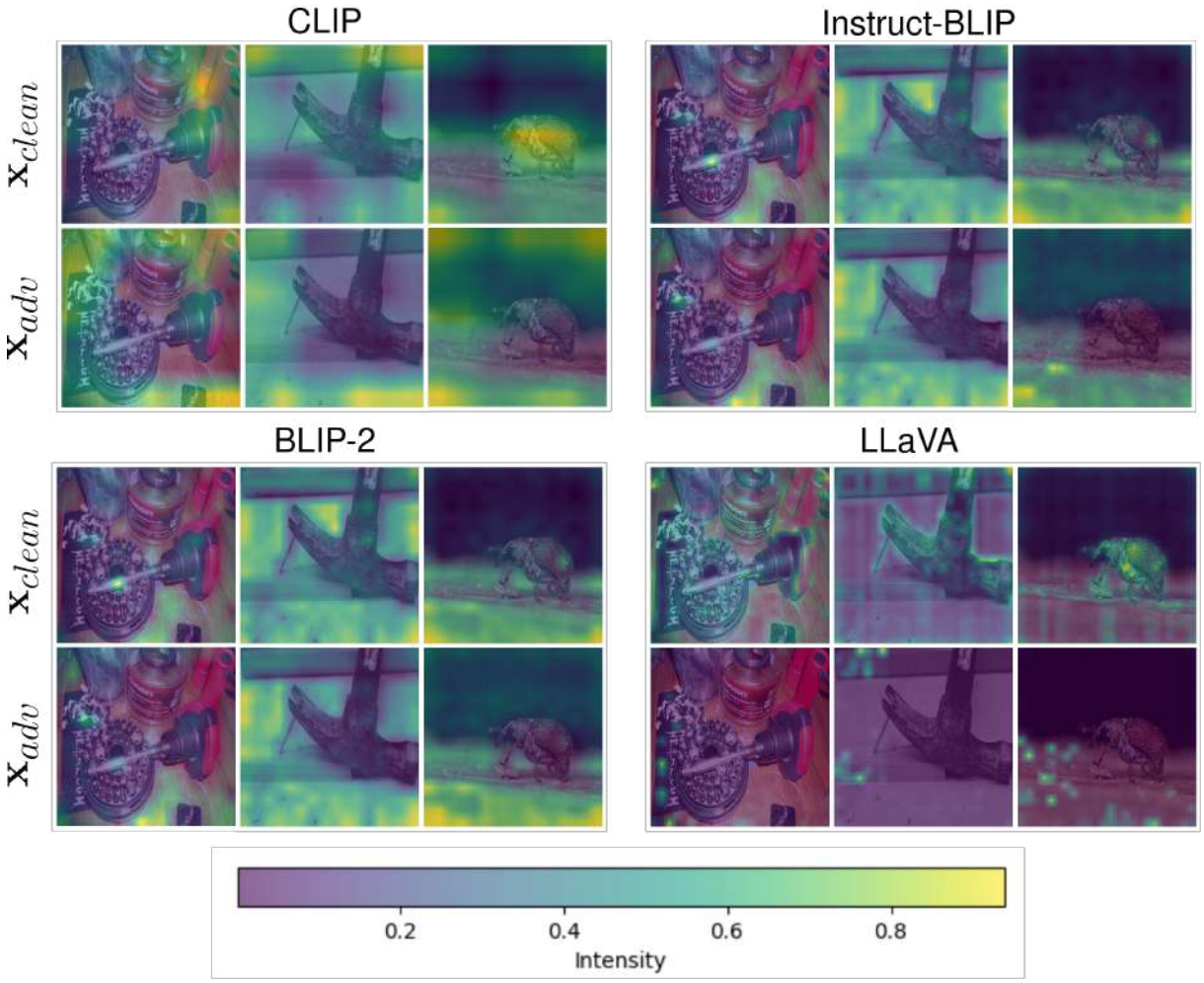


Figure 4.8 : Depiction of attention maps extracted from attacked layers for both a clean image and an adversarial image generated under the 'A+V' configuration.

ced performance on other models, such as BLIP2-T5 and Instruct-BLIP (as reported in table 4.4). This performance gap when L_{max} is set to 1 may be attributed to the varying degrees of freedom among the VLMs. The LLaVA visual encoder processes images with a higher resolution (336) compared to BLIP2-T5, and Instruct-BLIP (224), providing the attack optimization on LLaVA with more than twice the degrees of freedom. This increased flexibility potentially enables the identification of more optimal solutions with less perceptible adversarial perturbations. Conversely, increasing L_{max} can improve attack performance by distributing the optimization objective across multiple layers, effectively expanding the search space. While this does not guarantee finding more optimal solutions, our ablation study indicates that increasing L_{max} generally results in adversarial images

Table 4.4 : Detection scores, similarity scores, and image quality scores variations w.r.t L_{max} as evaluated on our (A+V) adversarial attack on the BLIP2-T5 and Instruct-BLIP models.

Lmax	1	2	4	8	16	24	32
Detection (%)	26.00	31.90	33.20	11.00	06.50	06.10	65.40
SBERT	0.264	0.389	0.341	0.341	0.308	0.302	0.454
USE	0.380	0.483	0.449	0.444	0.407	0.390	0.518
MPNET	0.261	0.384	0.339	0.337	0.305	0.300	0.460
SSIM	0.488	0.517	0.518	0.736	0.851	0.893	0.984
LPIPS	0.463	0.464	0.452	0.261	0.136	0.097	0.009
Detection (%)	06.00	36.30	28.50	17.90	11.50	10.50	79.10
SBERT	0.304	0.416	0.398	0.344	0.308	0.302	0.505
USE	0.378	0.476	0.458	0.420	0.396	0.386	0.544
MPNET	0.307	0.425	0.405	0.349	0.314	0.311	0.512
SSIM	0.541	0.513	0.533	0.689	0.808	0.864	0.984
LPIPS	0.415	0.415	0.437	0.310	0.183	0.127	0.010

that are more perceptually similar to clean images. However, we observed a threshold for L_{max} beyond which attack performance declines drastically, as shown in table 4.4.

We hypothesize that beyond a certain number of attention blocks, the information encapsulated within the i-th token no longer directly corresponds to that of the i-th image patch. Instead, the features at such depths represent abstract combinations derived from multiple patches. To validate this hypothesis, we computed and plotted the average attention matrices, $\mathcal{A}$, derived from the attention blocks of the BLIP2-T5 visual encoder for a set of 1000 clean images, as illustrated in figure 4.9.

Figure 4.9 reveals a key observation regarding the attention patterns in the BLIP2-T5 visual encoder. For early layers ($L < 24$), the attention maps exhibit a strong diagonal dominance, indicating that tokens primarily attend to their own values and those of neighboring tokens. This suggests a localized attention pattern where information from individual image patches is largely preserved within the corresponding tokens. However, this pattern dissipates for layers beyond 24, indicating a shift towards more global and abstract feature representations. This observation explains why our attack can effectively manipulate the information conveyed by image patches up to layer 24. Since the correspondence between tokens and image patches is maintained up to this depth, modifying

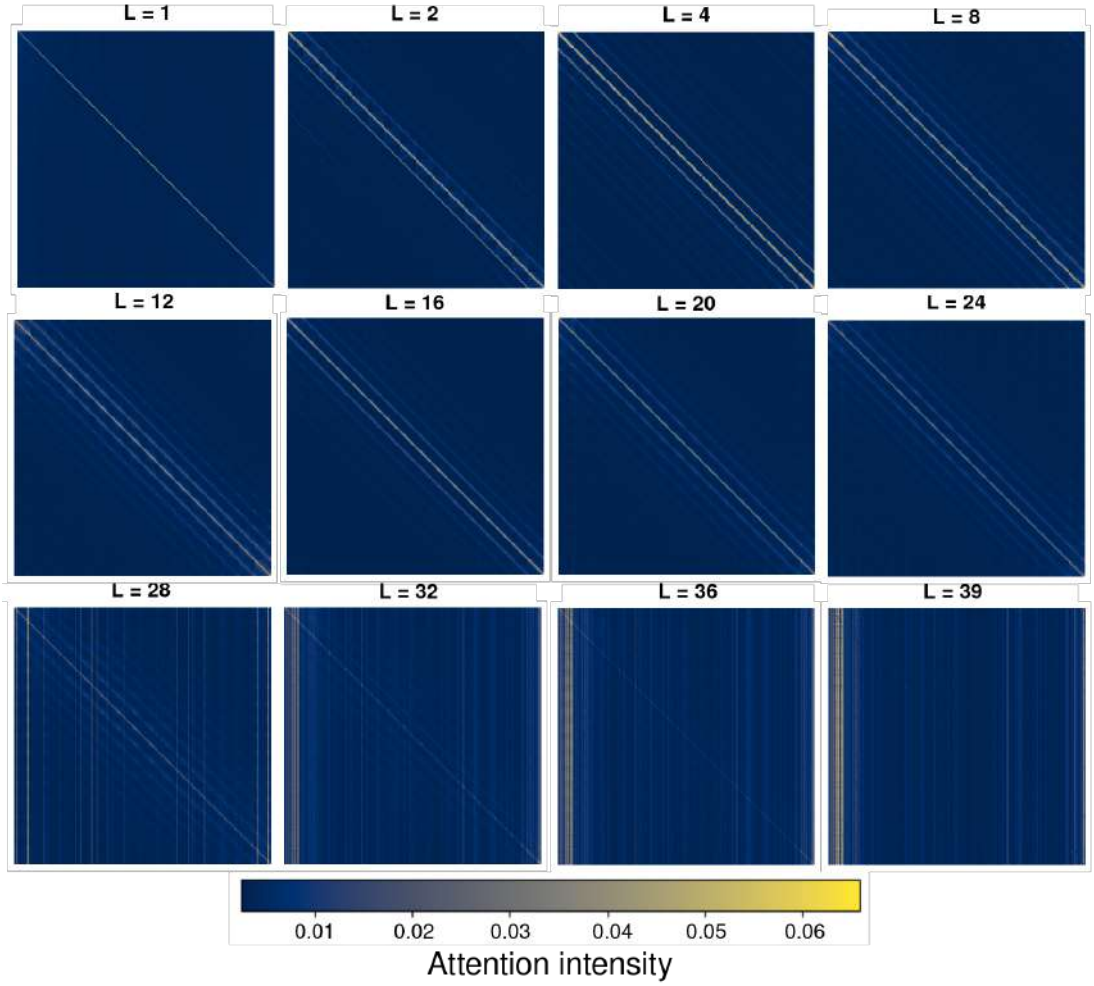


Figure 4.9 : Averaged attention maps $\mathcal{A}^{(\mathrm{L})}$ across multiple heads and multiple images extracted from different MHA blocks L of the BLIP2-T5 visual encoder.

the attention weights in these early layers directly impacts the information extracted from specific patches.

4.5.3 Influence of λ_v

We further conduct an ablation study to understand the influence of the Lagrangian coefficient attributed the values norm. We perform our experiments on the Instruct-BLIP and BLIP2-T5 models and report our obtained results in table 4.5. The experimental results indicate that, while optimal results are observed for $\lambda_v = 0.1$, the value assigned to λ_v exerts minimal impact on the overall efficacy of the attack as long as $\lambda > 0$. Our hypothesis to explain this observation is that the gradients of the value tokens' norm

are superior compared to those of the attention matrix (probably due to the softmax activation), suggesting that they may exert a dominant influence over the optimization trajectory. Although the results obtained for $\lambda > 1$ substantiate the significance of the term delineated in equation 4.3, as they demonstrate marginally reduced performance of the attack when a greater weight is attributed to the tokens' value norm. We conducted an additional ablation study in subsection 4.5.4 to evaluate whether the former attention term in equation 4.3 contributes meaningfully to the optimization process.

4.5.4 Value's-norm only optimization

We report in table 4.6 our attack carried out in different settings. Equation 4.5 delineates the optimization problem for the attack under the setting 'V'.

$$\begin{aligned} \mathbf{x}_{adv} &= \mathbf{x}_{clean} + \delta^*, \\ \delta^* &= \arg\min_{\delta} \sum_{l=1}^{\mathrm{L}_{max}} \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_{ROI}} \sum_{k=1}^{d_v} \| V^{(l,h)}(j,k) \| \end{aligned} \tag{4.5}$$

Table 4.5 : Variations in detection, similarity, and image quality scores w.r.t λ_v as evaluated in our adversarial attack on the Instruct-BLIP and BLIP2-T5 models.

	λ_v	0	0.1	0.5	1	2	5
Instruct-BLIP	Detection (%)	28.60	10.10	11.10	10.50	10.40	10.60
	SBERT	0.197	0.302	0.306	0.302	0.300	0.304
	USE	0.317	0.382	0.387	0.386	0.384	0.385
	MPNET	0.204	0.307	0.313	0.311	0.308	0.311
	SSIM	0.838	0.865	0.863	0.864	0.864	0.864
	LPIPS	0.152	0.127	0.128	0.127	0.128	0.127
BLIP2-T5	Detection (%)	14.10	05.80	05.90	06.10	06.20	06.20
	SBERT	0.183	0.299	0.302	0.302	0.297	0.299
	USE	0.335	0.389	0.389	0.390	0.387	0.388
	MPNET	0.175	0.299	0.302	0.300	0.297	0.299
	SSIM	0.895	0.894	0.893	0.893	0.894	0.893
	LPIPS	0.093	0.096	0.097	0.097	0.096	0.097

Our experimental results reported in table 4.6 reveal that the removal of the attention term from the optimized equation results in a drop in attack performance. This confirms that the latter term contributes positively to the attack's success, although its impact may be marginal depending on the attacked model and number of targeted layers. In fact, there is a significant gap between the LLaVA VLM and BLIP2-T5 or Instruct-BLIP models in terms of performance drop. We hypothesize that this disparity may stem from the differing nature of the features extracted at layers $L_{max} = 1$ and $L_{max} = 24$ of the models, as the optimization problem under the 'V' setting directly targets the features' norm.

Table 4.6 : Reported detection, similarity, and image quality scores as evaluated using different attack configurations on the Instruct-BLIP, BLIP2-T5 and LLaVA models.

VLM	Attack	Detection (%)	Perceptual fidelity		Cosine similarity		
			SSIM	LPIPS	SBERT	USE	MPNET
Instruct-BLIP	A	28.60	0.838	0.152	0.197	0.317	0.204
	V	10.60	0.864	0.127	0.304	0.387	0.312
	A+V	**10.50**	0.864	0.127	0.302	0.386	0.311
BLIP2-T5	A	14.10	0.895	0.093	0.183	0.335	0.175
	V	10.65	0.893	0.096	0.304	0.391	0.303
	A+V	**06.10**	0.893	0.097	0.302	0.390	0.300
LLaVA	A	07.00	0.818	0.125	0.467	0.375	0.454
	V	58.30	0.665	0.338	0.732	0.660	0.729
	A+V	**01.70**	0.769	0.184	0.510	0.424	0.501

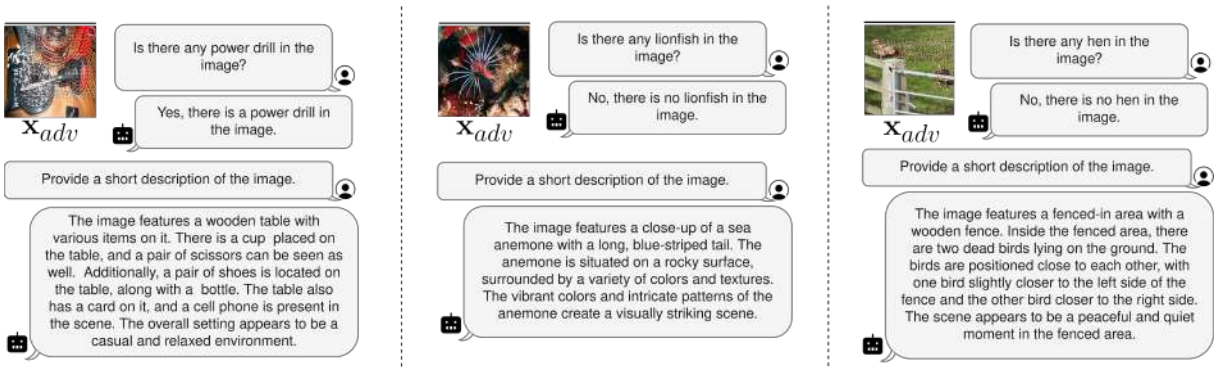


Figure 4.10 : LLaVA generated answers on a few adversarial examples generated under the 'V' setting.

Furthermore, we were intrigued by the high similarity scores achieved on the adversarial examples under the 'V' setting in LLaVA and conducted a manual inspection of some

generated answers as illustrated in figure 4.10. First, we observe that ROIs in the resulting adversarial examples appear blurred, suggesting that the potential solutions to the optimization problem in equation 4.5 when $L_{max} = 1$ involves directly removing the information present in the ROIs. This severely degrades the image quality (Average SSIM $\cong 0.65$) and leads to the failure of the attack (we detail our attacks's failure cases in subsection 4.5.6). Second, in the generated captions on the adversarial examples, we notice that the VLM replaces the objects present within the ROI with incorrect ones. However, these substituted objects seem to belong to the same semantic field, indicating a partial semantic understanding despite the adversarial perturbations (e.g. hen and bird or lionfish and sea anemone), explaining the high semantic similarity scores obtained between the clean captions and adversarial captions.

4.5.5 Perturbation budget

We further investigate the influence of a perturbation budget on the performance of our attack. Our experimental findings, as presented in table 4.7, do not necessarily indicate the presence of an inverse correlation between the perturbation budget and the VLM's detection rates. While it may be argued that increased budgets may lead to an acceptable trade-off between attack efficacy and perturbation visibility, our observations indicate that similarity scores are lower for $L_\infty = 0.2$ (used perturbation budget in the 'strong' attack setting of Cui et al. (2024)). Additionally, we find it noteworthy that our unconstrained setting generates adversarial attacks with higher quality scores compared to those obtained with the $L_\infty = 0.2$ perturbation budget. This phenomenon can be attributed to the substantial influence of image texture on perturbation visibility, suggesting the presence of multiple perturbation budgets that appear to be highly dependent on the specific characteristics of the region. Therefore, setting a high perturbation budget may prove less efficient than refraining from fixing one altogether, as it can result in suboptimal solutions characterized by increased perturbation visibility, particularly in flat and uniform regions which can in turn degrade the perceptual quality of the image.

4.5.6 Failure cases

Although the experimental results presented previously illustrate the effectiveness of our attack strategy, we acknowledge that there are several instances of failure in our attack. We

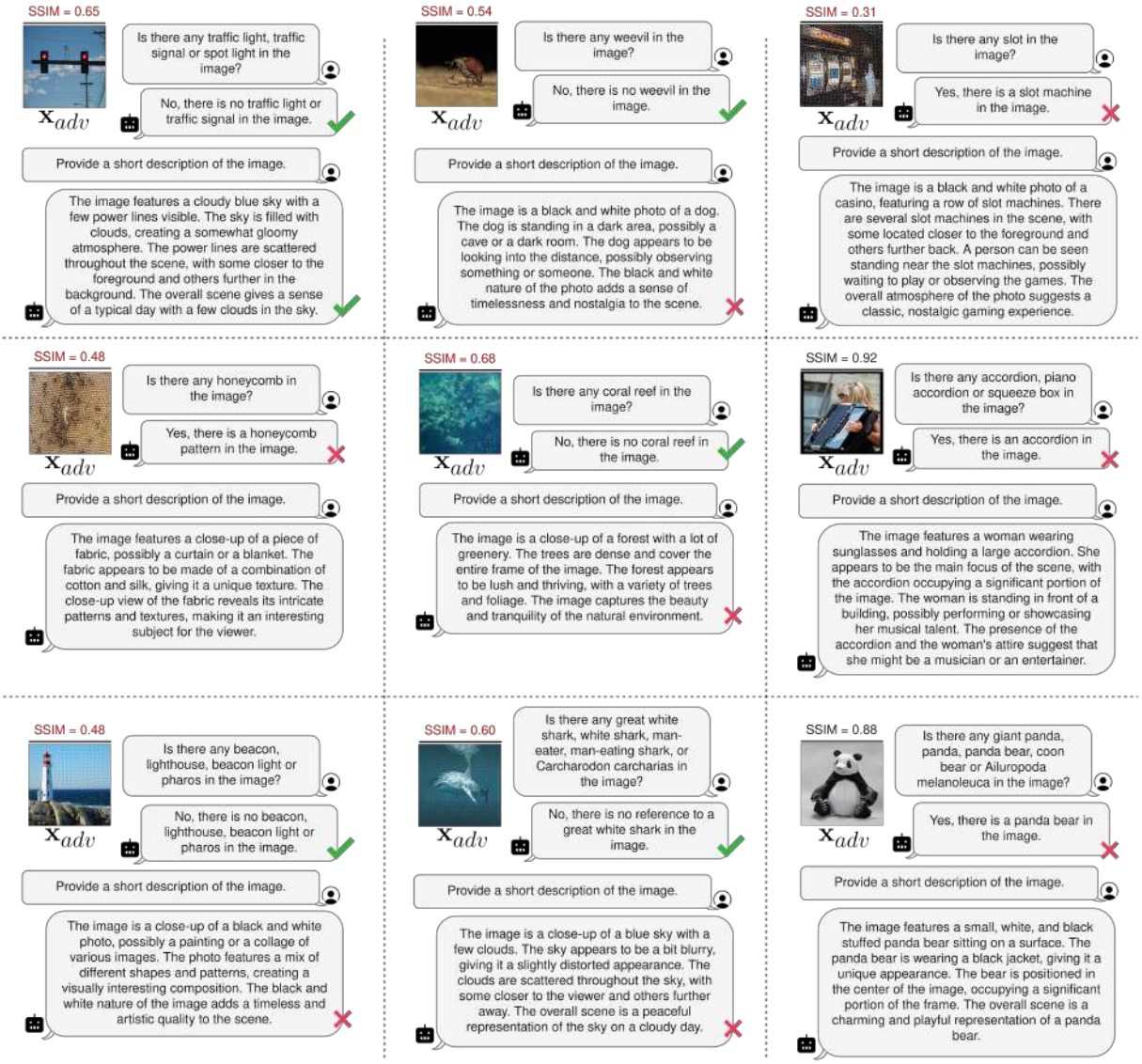


Figure 4.11 : Failure cases of our proposed adversarial attack under the 'A+V' setting for LLaVA.

Table 4.7 : Variations in detection, similarity, and image quality scores w.r.t L_∞ perturbation budget as evaluated in our adversarial attack on the BLIP2-T5 and Instruct-BLIP models. ∞ refers to the non constrained setting of the attack. Bold entries indicate the highest scores while underlined entries indicate second-best reported scores.

model	BLIP2-T5			Instruct-BLIP		
L_∞ budget	20/255	0.2	∞	20/255	0.2	∞
Detection(%)	30.80	12.20	06.10	39.60	07.90	10.50
SBERT	0.274	0.178	0.302	0.293	0.180	0.302
USE	0.369	0.281	0.390	0.374	0.286	0.386
MPNET	0.272	0.177	0.300	0.295	0.184	0.311
SSIM	**0.938**	0.804	<u>0.893</u>	**0.937**	0.638	<u>0.864</u>
LPIPS	**0.039**	0.175	<u>0.097</u>	**0.039**	0.387	<u>0.127</u>

suggest that these are to be addressed in future research efforts for further improvements. We deem the failure of the attack if it satisfies one or more of the subsequent criteria.

1. The VLM successfully detects the object within the ROI.
2. The quality of the resulting adversarial image is deteriorated (SSIM < 0.7).
3. The caption generated on the adversarial image is not similar to the one on the clean image.

Figure 4.11 provides visual examples corresponding to instances where our attack was unsuccessful.

4.5.7 Average attack run-times

Tables 4.8 and 4.9 report the average run-times required to perform our attack in different settings. All experiments were performed on a single quadro RTX 8000 GPU, except the LLaVA model which requires two GPUs.

Overall, the execution time required by our attack seems realistic, on average less than 5 minutes. However, execution time shows significant variation across the models and different attacks. This variation is mainly due to the early stopping mechanism, as some examples are easier to optimize than others depending on the clean image content, the ROIs, the target model, and L_{max}.

Table 4.8 : Average run time in seconds required to perform our attack on different VLMs under the 'A+V' setting.

VLM	LLaVA	Instruct-BLIP	BLIP2-T5
Ours (A)	271	150	77
Ours (A+V)	248	117	67

Table 4.9 : Average run time in seconds required to perform our attack under the 'A+V' setting w.r.t L_{max}.

L_{max}	1	2	4	8	16	24	32
BLIP2-T5	89	111	127	81	61	67	38
Instruct-BLIP	97	151	152	123	98	117	48

4.6 Conclusion

In this chapter, we proposed a novel adversarial attack strategy to address privacy concerns in VLMs. Our approach selectively conceals sensitive information within images by manipulating the attention and value matrices extracted from the early MHA blocks of the VLM's visual encoder. This manipulation effectively prevents the VLM from accessing and interpreting the sensitive data contained within targeted ROIs. Our experimental results demonstrated the effectiveness of this approach on three different VLM models (i.e., LLaVA, Instruct-BLIP and BLIP2-T5), showing a significant reduction in the ability of various VLMs to detect objects within the targeted areas while preserving the overall consistency of the visual content outside those regions.

Conclusion

4.7 Summary of the contributions

This thesis has explored the evolving landscape of adversarial attacks. Our key contributions investigate how deep learning models can be manipulated across a range of objectives from compromising model accuracy and undermining model efficiency to leveraging adversarial examples for user privacy preservation.

Chapter 1 lays the foundational concepts required for understanding the subsequent chapters of the thesis by introducing key background concepts related to deep learning and adversarial attacks. The chapter starts with an overview of deep learning notations and basic concepts, then introduces adversarial attacks, highlighting the challenges they impose to machine learning and AI security. While not exhaustive, it provides a focused survey of some of the most influential attack strategies found in the literature and categorizes them based on the adversary's objective, their degree of access to the victim model, and the attack stage. This framework serves as a conceptual reference point that help situate the contributions developed in the subsequent chapters.

Chapter 2 investigates the robustness of Blind Image Quality Assessment (BIQA) models under adversarial conditions. It provides an overview of existing BIQA models, then surveys adversarial attacks targeting these models in the literature. Building on this foundation, we proposed a novel adversarial attack specifically designed to compromise the accuracy of BIQA models. One key aspect of our approach consists in relaxing the constraint on the adversarial perturbation budget. This relaxation enables us to examine how perceptible image perturbations (which are typically expected to reduce perceived quality) can be adversarially crafted to receive high predicted quality scores. Our findings

demonstrate the susceptibility of these models to both targeted and untargeted attacks, exposing a critical vulnerability in their assessment capabilities.

Chapter 3 introduces a novel energy backdoor attack targeting deep neural networks, particularly those deployed on sparsity-based hardware accelerators. After reviewing existing energy-latency threats, we identified and filled the gap in the literature by designing a stealthy energy backdoored model. The proposed attack is achieved within two stages : 1) a backdoor injection phase which reinforces the backdoor behaviour within the model and a 2) stealthiness optimization phase which ensures the model retains its accuracy on clean data. Our experimental results indicate that our method can increase the model's inference energy usage on trigger samples without compromising its performance and efficiency on clean inputs. This extends the scope of backdoor attacks beyond conventional accuracy-based objectives.

Chapter 4 explores the potential of adversarial attacks as tools for enhancing user privacy in the context of VLMs. We first outlined the evolution of VLMs, followed by a review of adversarial attacks targeting these models, as well as prior works proposing the use of adversarial attacks for 'social good'. Then, we presented a novel attack methodology aimed at selectively concealing the Region Of Interest (ROI) within an image to reduce the exposure of private visual content. Unlike conventional adversarial attacks that primarily aim to degrade accuracy or manipulate model outputs, our approach leverages the attention coefficients and values extracted by the VLM's visual encoder to selectively control the flow of information. This enables determining which visual features are retained for processing and which are suppressed to preserve privacy. Through comprehensive experiments, we demonstrated that carefully crafted perturbations can effectively obscure sensitive attributes within an image without disrupting its overall utility. This contribution showcases how adversarial techniques can also be repurposed for ethical and privacy-preserving objectives.

4.8 Limitations, future perspectives and open questions

While the contributions of this thesis provide meaningful insights into adversarial vulnerabilities across various DL applications, the field remains fast evolving which leaves considerable scope for improvement. This includes both refining existing attack strategies and exploring new research directions that address emerging challenges and applications.

Limitations

The work presented in the first chapter proposes a white-box adversarial attack to highlight the vulnerabilities of CNN-based IQA models. While effective in uncovering potential vulnerabilities in these models, the study is limited in scope to CNN architectures, which have since been outpaced by more advanced approaches such as ViT-based, hybrid models, or VLMs. Furthermore, although our findings suggest that IQA outputs can be manipulated in a way that may affect downstream tasks, we do not evaluate the real impact of such perturbations on actual user-perceived quality or in full QoE pipelines. This further reflects a broader gap in the literature, where black-box attacks have emerged but often lack a systematic connection to real-world QoE metrics or applications.

Similarly, in the second chapter, we demonstrated the feasibility of implanting a stealthy energy backdoor into deep neural networks, showing that such manipulations can remain undetected while increasing energy consumption. Still, the study also lacks an in-depth analysis of the real-world economic implications of such attacks on organizations and infrastructures.

Finally, in the third chapter we introduced an attack strategy that conceals privacy-sensitive information in images processed by VLMs. Although our experimental study demonstrates the effectiveness of our proposed attack, it also opens up several avenues for future exploration. For instance, a more optimal incorporation of the perturbation visibility constraint into the attack formulation by taking into consideration the nature of the different regions of the image represents an interesting future research avenue. Moreover, although we rely on similarity metrics due to the impracticality of performing exhaustive manual inspections, we anticipate that more advanced LLMs will be available in the near future. These models could be efficiently tuned to evaluate the performance of our adversarial attacks more effectively and reliably. Furthermore, an aspect of our attack formulation is that it does not explicitly track the attention flow within the visual encoder's MHA blocks. Instead, our approach focuses on identifying an optimal L_{max} value that balances perturbation visibility with detection scores. Exploring methods to directly monitor and analyze attention flow through MHA layers could provide deeper insights into the inner workings of the VLM, enabling more effective optimization and potentially enhancing the success of the attack.

Future perspectives and open questions

Defense strategies : a rather obvious research direction lies in keeping defenses continually updated to match the evolving sophistication of attack strategies. This is already an active area of research, though it falls outside the scope of work we covered in this thesis. In particular, we highlight that defense strategies should be designed with application-specific constraints in mind. For example, in response to energy or latency-based attacks, it is important that the defense mechanism does not incur similar or higher costs than the attack itself, as this would render it counterproductive. Ensuring such balances between robustness and efficiency is essential for practical deployment.

Real-world adversaries, transferability, and access constraints : many existing studies in the literature, including ours, primarily focus on white-box settings, where the adversary has full access to the model. Although several works claim to explore black-box scenarios, these often align more closely with gray-box conditions, where the adversary possesses partial access to model's training data or predicted probabilities. We believe that future work should aim to clarify and formalize these distinctions by explicitly defining the assumed level of access and justifying how such access might be realistically obtained. The use of standardized categorization schemes or unified comparison charts would further facilitate consistent evaluation and comparison across studies in the literature.

Explainability and interpretability : in the context of most accuracy-based adversarial attacks, adversarial input optimization is typically performed directly on the model's output with limited consideration for the model's internal behavior. In contrast, attacks targeting energy consumption or latency often require a deeper understanding of the model's internal structure and computation flow. This highlights the potential value of incorporating insights from model explainability and interpretability into adversarial attack research. Moreover, explaining why adversarial examples succeed remains an open challenge and an active research direction. Greater synergy between adversarial machine learning and interpretability research could lead to more insightful attack strategies and enhance our broader understanding of model vulnerabilities.

Robustness evaluation across multiple modalities : with the advent of large multimodal models that integrate various input types such as vision, language, and audio, it has become increasingly important to assess how adversarial vulnerabilities manifest across these distinct modalities. While a few recent studies have begun to explore these questions, the area remains relatively underexplored (Kapoor et al., 2025). Future work should investigate not only how attacks affect each modality independently, but

also how perturbations in one modality may influence or propagate through others. A deeper understanding of these cross-modal interactions is essential for developing more comprehensive and resilient defense strategies.

Human bias and training loop vulnerabilities : As modern AI systems increasingly rely on human feedback (particularly in frameworks like Reinforcement Learning from Human Feedback (RLHF)), they become susceptible to vulnerabilities introduced by human biases. These biases, whether intentional or not, can create exploitable patterns in the model's behavior. Adversaries could leverage this by socially engineering models to behave in unintended ways, potentially leading to successful jailbreak attacks or the circumvention of safety constraints. Future work should explore how to mitigate these human-induced vulnerabilities and design robust mechanisms to guard against adversarial manipulation rooted in the training loop.

Bridging adversarial ML and cybersecurity : with the rise of LLM-based agents that are granted access to sensitive resources such as credentials, databases, and web tools, we are witnessing an increasing convergence between adversarial machine learning and traditional cybersecurity. These agents, often deployed in semi-autonomous or autonomous settings, introduce new threat surfaces where adversarial prompts or model-level attacks could have real-world consequences. As such, future research should integrate classical cybersecurity principles such as access control, authentication, intrusion detection, with adversarial robustness to better anticipate and defend against threats in this evolving landscape.

List of Publications

Conference papers

Evaluating the Vulnerability of Deep Learning-based Image Quality Assessment Methods to Adversarial Attacks

10th European Workshop on Visual Information Processing (2023)

Authors : Hanene F. Z. Brachemi Meftah, Sid Ahmed Fezza, Wassim Hamidouche, Olivier Déforges

Abstract : Recent studies have discovered that Deep Learning (DL) models are vulnerable to adversarial attacks in image classification tasks. While most studies have focused on DL models for image classification, only a few works have addressed this issue in the context of Image Quality Assessment (IQA). This paper investigates the robustness of different Convolutional Neural Network (CNN) models against adversarial attacks when used for an IQA task. We propose an adaptation of state-of-the-art image classification attacks in both targeted and untargeted modes for an IQA regression task. We also analyze the correlation between the perturbation's visibility and the attack's success. Our experimental results show that DL-based IQA methods are vulnerable to such attacks, with a significant decrease in correlation scores. Consequently, the development of countermeasures against such attacks is essential for improving the reliability and accuracy of DL-based IQA models. To support the principle of reproducible research and fair comparison, we make the codes publicly available on https://github.com/hbrachemi/IQA_AttacksSurvey.

Energy Backdoor Attack to Deep Neural Networks

IEEE International Conference on Acoustics, Speech, and Signal Processing (2025)

Authors : Hanene F. Z. Brachemi Meftah, Wassim Hamidouche, Sid Ahmed Fezza, Oli-

vier Déforges, Kassem Kallas

Abstract : The rise of deep learning (DL) has increased computing complexity and energy use, prompting the adoption of application specific integrated circuits (ASICs) for energy-efficient edge and mobile deployment. However, recent studies have demonstrated the vulnerability of these accelerators to energy attacks. Despite the development of various inference time energy attacks in prior research, backdoor energy attacks remain unexplored. In this paper, we design an innovative energy backdoor attack against deep neural networks (DNNs) operating on sparsity-based accelerators. Our attack is carried out in two distinct phases : backdoor injection and backdoor stealthiness. Experimental results using ResNet-18 and MobileNet-V2 models trained on CIFAR-10 and Tiny ImageNet datasets show the effectiveness of our proposed attack in increasing energy consumption on trigger samples while preserving the model's performance for clean/regular inputs. This demonstrates the vulnerability of DNNs to energy backdoor attacks. The source code of our attack is available at : https://github.com/hbrachemi/energy_backdoor.

Journal papers

Energy-Latency Attacks : A New Adversarial Threat to Deep Learning

ACM Computing Surveys (Under review)

Authors : Hanene F. Z. Brachemi Meftah, Wassim Hamidouche, Sid Ahmed Fezza, Olivier Déforges

Abstract : The growing computational demand for DNNs has raised concerns about their energy consumption and carbon footprint, particularly as the size and complexity of the models continue to increase. To address these challenges, energy-efficient hardware and custom accelerators have become essential. Additionally, adaptable DNNs are being developed to dynamically balance performance and efficiency. The use of these strategies became more common to enable sustainable AI deployment. However, these efficiency-focused designs may also introduce vulnerabilities, as attackers can potentially exploit them to increase latency and energy usage by triggering their worst-case-performance scenarios. This new type of attack, called energy-latency attacks, has recently gained significant research attention, focusing on the vulnerability of DNNs to this emerging attack paradigm, which can trigger Denial-Of-Service (DoS) attacks. This paper provides a com-

prehensive overview of current research on energy-latency attacks, categorizing them using the established taxonomy for traditional adversarial attacks. We explore different metrics used to measure the success of these attacks and provide an analysis and comparison of existing attack strategies. We also analyze existing defense mechanisms and highlight current challenges and potential areas for future research in this developing field. The GitHub page for this work can be accessed at https://github.com/hbrachemi/Survey_energy_attacks/.

VIP : Visual Information Protection through Adversarial Attacks on Vision-Language Models (Under review)

Authors : Hanene F. Z. Brachemi Meftah, Wassim Hamidouche, Sid Ahmed Fezza, Olivier Déforges

Abstract : Recent years have seen remarkable progress in the development of VLMs capable of processing both textual and visual inputs. These models have demonstrated impressive performance, leading to their widespread adoption in various applications. However, this growing use raises serious concerns about user privacy, particularly when models inadvertently process or expose private visual information. In this work, we frame the preservation of privacy in VLMs as an adversarial attack problem. We propose a novel attack strategy that selectively conceals information in designated ROIs of an image, thereby effectively preventing VLMs from accessing sensitive content while preserving the semantic integrity of the remaining image. Unlike conventional adversarial attacks that often disrupt the entire image, our method maintains high coherence in unmasked areas. Experimental results obtained on three state-of-the-art VLMs, namely LLaVA, InstructBLIP, and BLIP2-T5 demonstrate up to 98% reduction in target ROI detection, while preserving the overall image semantics, as shown by high similarity scores between clean and adversarial outputs. We believe that this work contributes to a more privacy-conscious use of multimodal models and offers a practical tool for further research, with the source code publicly available at https://github.com/hbrachemi/Vlm_defense-attack.

Résumé en français

Introduction, contexte et problématique

L'apprentissage automatique (ML), discipline visant à développer des systèmes informatiques capables d'apprendre et de s'améliorer à partir de l'expérience, est passé d'un domaine de recherche niche à un pilier incontournable du progrès technologique contemporain. Ses applications couvrent un large éventail de secteurs, allant des systèmes de recommandation et des infrastructures de télécommunications à la détection de fraude, transformant en profondeur aussi bien la recherche académique que les pratiques industrielles. Au cœur de cette révolution se trouve l'apprentissage profond (DL), une sous-branche de l'apprentissage automatique, inspirée de la structure et du fonctionnement du cerveau humain. Les algorithmes d'apprentissage profond se sont révélés particulièrement efficaces pour traiter des données complexes et prendre des décisions de manière autonome, tout en nécessitant peu d'intervention humaine. Le succès de ce domaine, ainsi que son développement rapide, s'expliquent principalement par plusieurs facteurs clés : la disponibilité massive de données, l'augmentation significative en terme de puissance de calcul grâce à l'émergence de matériels spécialisés, ainsi que la scalabilité algorithmique, qui permet de traiter efficacement de vastes volumes de données à l'aide de modèles de grande taille dans des environnements distribués.

L'émergence récente des LLMs et des VLMs représente l'une des avancées les plus remarquables de l'apprentissage profond, redéfinissant les frontières des capacités de ces systèmes. Contrairement aux architectures et modèles traditionnels, conçus pour des tâches bien définies, les VLMs sont capables de généraliser à une multitude d'applications sans pour autant y être spécifiquement entrainé. Parmi celles-ci figurent la rédaction, la traduction, la programmation, le raisonnement, la prise de décision, la réponse à des questions

visuelles, la détection d'objets, et bien d'autres encore. Cette polyvalence, alliée à leur capacité d'adaptation, a favorisé leur intégration à un grand nombre de services, de produits et de systèmes applicatifs.

Aujourd'hui, ces systèmes sont présents dans de nombreuses applications qui soutiennent les activités humaines au quotidien : soins et santé, maisons intelligentes, prévision financière, optimisation logistique, véhicules autonomes ou encore vidéosurveillance. Cela témoigne du rôle devenu désormais central de l'IA dans la vie quotidienne. Toutefois, bien que cette adoption offre de nombreux avantages, elle soulève également des préoccupations en termes de sécurité. En effet, l'intégration de ces technologies au sein d'infrastructures critiques et d'environnements numériques personnels peut avoir de lourdes conséquences en cas d'exploitation malveillante de leurs vulnérabilités.

Bien que les attaques adverses ciblant les classifieurs d'images aient été largement étudiées dans la literature, le domaine reste extrêmement dynamique. De nouvelles architectures de modèles émergent en permanence, entraînant la découverte de nouvelles vulnérabilités alimentant ainsi un cycle de recherche sans fin. Cette thèse s'inscrit donc dans cette dynamique en élargissant les perspectives actuelles à travers l'exploration de plusieurs stratégies d'attaque visant à mettre en évidence des vulnérabilités, encore peu ou pas étudiées, dans les systèmes d'apprentissage profonds appliqués à la vision par ordinateur. Qu'il s'agisse de contourner des modèles d'évaluation de la qualité d'image, d'introduire des portes dérobées (backdoors) exploitant la consommation énergétique du modèle, ou encore de manipuler des modèles vision-langage multimodaux à travers des images, ce travail explore différentes approches d'attaque révélant des failles spécifiques.

Contributions

Vulnérabilité des méthodes d'évaluation de la qualité d'image par apprentissage profond aux attaques adverses

Introduction

Avec l'émergence massive de contenus multimédias, l'évaluation de la qualité d'image est devenue un module essentiel dans de nombreuses chaînes de transmission. Parallèlement, le développement spectaculaire de l'apprentissage profond a permis d'importants progrès dans l'automatisation de nombreuses tâches humaines notamment l'évaluation de la qualité d'image. Toutefois, plusieurs études révélent que les modèles d'apprentissage

profond sont vulnérables aux attaques adverses. Alors que la majorité des recherches se sont concentrées sur les modèles de classification, peu de travaux ont abordé cette problématique dans le cadre de l'évaluation de la qualité d'image (IQA). La gravité des attaques visant ces métriques couvre un large spectre opérationnel, allant de simples désagréments pour les utilisateurs à des conséquences graves dans des systèmes critiques.

Plusieurs travaux dans l'état de l'art révèlent collectivement la vulnérabilité de diverses méthodes d'évaluation de la qualité (y compris celles basées sur des architectures CNN). Cependant, la robustesse de modèles fondés sur différentes architectures CNN restait encore peu explorée. De plus, la majorité des recherches existantes se sont principalement concentrées sur des attaques ciblées (targeted), reposant sur des perturbations imperceptibles qui cherchent à augmenter le score de qualité prédit par le modèle IQA. Au moment de la rédaction du papier sur lequel est fondé cette contribution, et à notre connaissance, peu d'études avaient été consacrées aux attaques non ciblées (untargeted) contre les modèles IQA, ce qui soulignait la nécessité de recherches approfondies dans cette direction. Par ailleurs, les travaux antérieurs (Korhonen and You, 2022; Shumitskaya et al., 2022; Zhang et al., 2022b) avaient étudié ces attaques en contraignant la visibilité des perturbations. Cela implique qu'ils partent du principe qu'une personne tierse est chargé de vérifier l'intégrité des images avant leur évaluation par la métrique IQA mais cette hypothèse s'avère peu réaliste. Or, il serait plus pertinent d'examiner comment la perceptibilité des perturbations influence l'évaluation de la qualité dans ce contexte.

Cadre de l'attaque proposée

Nous étudions dans quelle mesure les modèles IQA basés sur des CNNs peuvent résister à ces attaques. Dans notre attaque proposée, nous choisissons délibérément de ne pas fixer de contrainte sur la perceptibilité des perturbations introduites lors de la génération des images adverses. Cette relaxation des contraintes nous permet d'explorer la relation entre la perceptibilité des attaques adverses et leur taux de réussite en vue du contexte. Elle nous permet également de générer des attaques capables de dégrader la qualité réelle de l'image tout en augmentant simultanément les scores de qualité prédits. Dans le scénario ciblé (targeted), l'adversaire cherche à induire le modèle en erreur afin qu'il prédise un score de qualité élevé (ou faible), indépendamment de la qualité perceptuelle réelle de l'image. Pour atteindre cet objectif, nous adaptons la procédure d'optimisation proposée dans des travaux antérieurs (Korhonen and You, 2022; Shumitskaya et al., 2022), que nous reformulons sous la forme de problème sans contraintes. Plus précisément, nous

assouplissons la formulation initiale en l'exprimant selon l'objectif suivant :

$$\mathbf{x}_{adv} = \arg\min_{\mathbf{x}} [y_{target} - \mathcal{M}_\theta(\mathbf{x})]^2 \;\; s.t\; \mathbf{x}_{adv} \in \mathbb{I}, \tag{4.6}$$

où $\mathbb{I}$ désigne l'espace des images, et y_{target} représente le score de qualité maximal que le modèle peut attribuer. Autrement dit, l'objectif est de trouver un exemple adversarial $\mathbf{x}_{adv}$ qui minimise la distance entre le score prédit $\mathcal{M}_\theta(\mathbf{x}_{adv})$ et le score cible y_{target}.

En revanche, dans le scénario non ciblé (untargeted), l'adversaire cherche à réduire la précision du modèle sans pour autant le pousser à prédire un score spécifique. Dans le cadre de notre attaque, nous formulons ce problème sous la forme suivante :

$$\mathbf{x}_{adv} = \arg\max_{\mathbf{x}} [\mathcal{M}_\theta(\mathbf{x}_{orig}) - \mathcal{M}_\theta(\mathbf{x})]^2 \;\; s.t\; \mathbf{x}_{adv} \in \mathbb{I}, \tag{4.7}$$

où $\mathbf{x}_{orig}$ désigne l'image originale, non altérée (avant l'attaque). Toutefois, appliquer une méthode d'optimisation basée sur des gradients directement à un problème de régression conduit à un coût nul, ce qui annule le gradient et empêche l'attaque. Une solution possible que nous proposons à ce problème consiste à remplacer les scores $\mathcal{M}_\theta(\mathbf{x}_{orig})$ par des estimations des scores vérité terrain. Cette approche permet d'ajouter du bruit afin de pallier le problème des coûts nuls. Ainsi, il est possible d'estimer ces scores en échantillonnant des valeurs à partir de la distribution suivante $\hat{MOS} \sim \mathcal{N}(\mathcal{M}_\theta(\mathbf{x}_{orig}), \text{RMSE}^2)$. L'erreur quadratique moyenne (RMSE) dans ce cas est évaluée sur un ensemble de test de la base de données.

Éxperimentations, résultats et analyse

Nous avons d'abord sélectionné trois architectures CNN couramment utilisées en vision par ordinateur et en IQA : ResNet50 (He et al., 2015), VGG16 (Simonyan and Zisserman, 2014) et InceptionV3 (Szegedy et al., 2015b) (Toutefois, nous supposons que des conclusions similaires peuvent être généralisées à d'autres CNNs (Papernot et al., 2016a).) Nous avons ensuite eu recours à un apprentissage par transfert (transfer learning) pour adapter ces modèles pré-entraînés à la tâche d'évaluation de la qualité en les entrainant sur deux bases de données de référence en IQA : TID2013 (Ponomarenko et al., 2013), qui contient des images dégradées de manière synthétique, et Koniq-10k (Hosu et al., 2020), composée d'images dégradées de manière authentique. Par la suite, dans nos expérimentations, nous avons considéré trois attaques adverses largement utilisées : Fast Gradient

Method (FGM) (Goodfellow et al., 2014), BIM (Kurakin et al., 2018) et PGD (Madry et al., 2017), avec différents niveaux de perturbation : $\epsilon \in \{0.001, 0.01, 0.1, 1\}$, et un nombre d'itérations fixé à $n_{iter} \in \{2, 5, 10\}$ pour les attaques itératives.

Afin de mieux visualiser les résultats obtenus, la figure 4.12 illustre la variation du SRCC en fonction de ϵ et du type d'attaque, en fixant le nombre maximal d'itérations à $n_{iter} = 10$. Les performances initiales des modèles évaluées sur des images non attaquées, sont représentées par $\epsilon = 0$. Pour rappel, l'image adversariale est définie par : $\mathbf{x}_{adv} = \mathbf{x}_{orig} + \epsilon \times \delta$, où δ désigne le signal de perturbation. Dans cette formulation, ϵ contrôle l'intensité de la perturbation ajoutée à l'image initiale. La figure 4.12 met en évidence

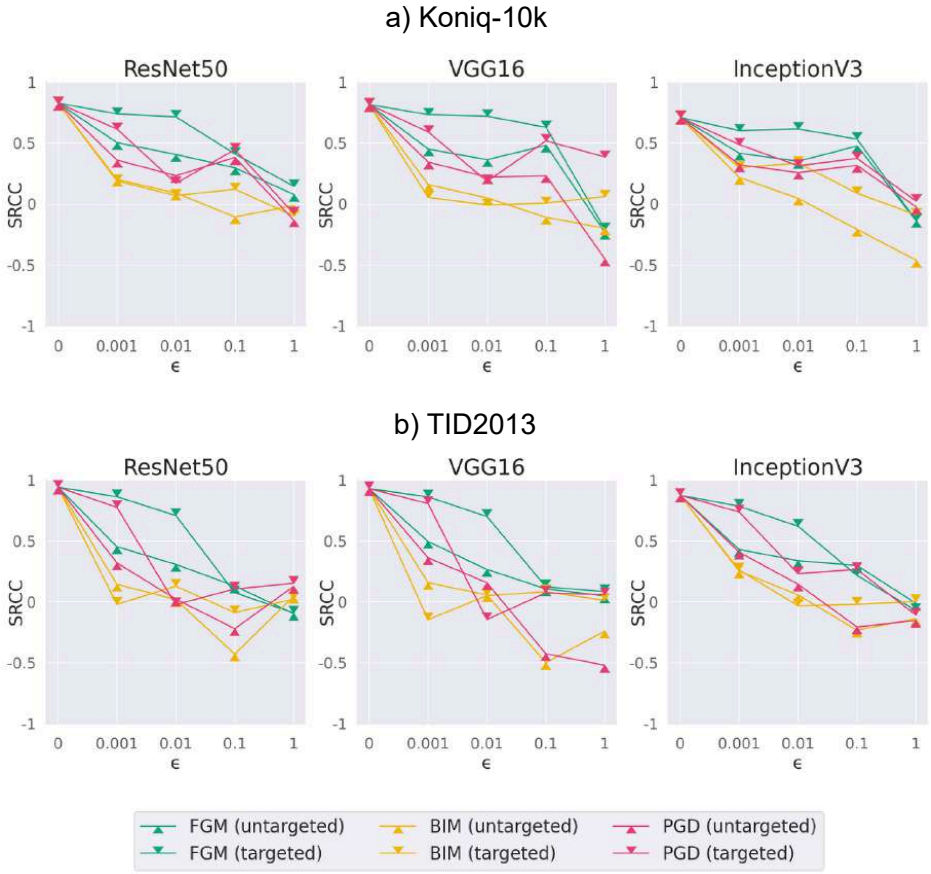


Figure 4.12 : variation du SRCC calculé sur l'ensemble de test par rapport à la variation d' ϵ.

que les trois architectures CNN testées présentent une vulnérabilité similaire face aux attaques considérées, et ce, malgré leurs différences architecturales. Cette vulnérabilité se manifeste par une chute significative des scores de corrélation.

La figure 4.13 illustre la relation entre les scores LPIPS, le type d'attaque utilisée et les valeurs de SRCC correspondantes. Les résultats mettent en évidence un compromis entre

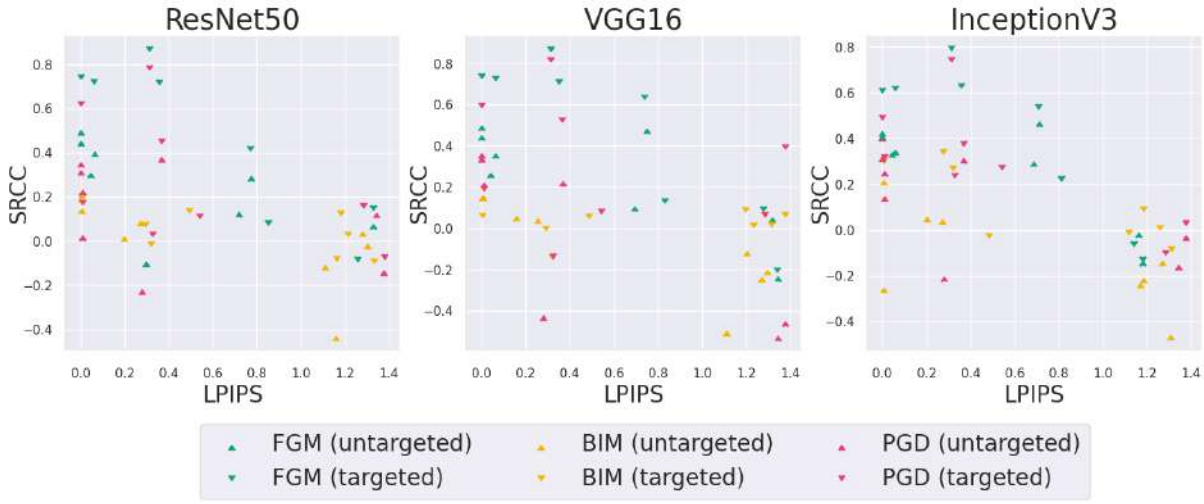


Figure 4.13 : Variations des scores SRCC scores par rapport aux scores LPIPS.

le succès de l'attaque (reflété par la baisse des scores de corrélation) et la visibilité de la perturbation (mesurée par le score LPIPS). Néanmoins, l'attaque obtenant le meilleur taux de réussite n'est pas nécessairement celle présentant la perturbation la plus visible selon LPIPS, et inversement.

Injection d'une porte dérobée énergétique dans des réseaux neuronaux profonds

Introduction

Dans le contexte économique actuel la recherche d'efficacité énergétique s'impose comme une priorité stratégique pour les entreprises. Cette tendance s'explique par la hausse des coûts d'exploitation, le renforcement des contraintes environnementales et l'attention croissante portée à la durabilité. En parallèle, cette évolution coïncide avec la stratégie du "going deeper" proposée par Szegedy et al. (2015a), qui a entraîné une adoption massive des technologies d'apprentissage profond. À cet effet, une multitude d'accélérateurs, tels que les TPUs et les composants ASIC, ont été largement adoptés pour relever ce défi. Néanmoins, plusieurs travaux (Chen et al., 2023a; Cinà et al., 2022; Navaneet et al., 2023; Shumailov et al., 2021) proposent des attaques énergétiques, spécifiquement conçues pour

pousser les modèles vers des scénarios de haute consommation d'énergie. Ces attaques reposent sur l'hypothèse qu'une augmentation significative de la consommation énergétique pourrait entraîner une non-réactivité du système ou une décharge rapide de la batterie sur les appareils mobiles (Wang et al., 2023c).

Cadre de l'attaque proposée

Nous explorons un angle de recherche encore négligé dans la littérature en explorant la vulnérabilité des DNNs face aux attaques par portes dérobées ciblant la consommation énergétique. Plus précisément, nous soulevons la question suivante :

> "*Est-il possible d'introduire une porte dérobée dans un DNN de sorte que, lorsqu'elle est activée, elle empêche les accélérateurs matériels d'exploiter la sparcité pour améliorer l'efficacité énergétique, tout en préservant des performances élevées sur l'ensemble des entrées du modèle ?*"

Pour répondre à cette question, nous proposons un modèle contenant une porte dérobée, illustré dans la figure 4.14.

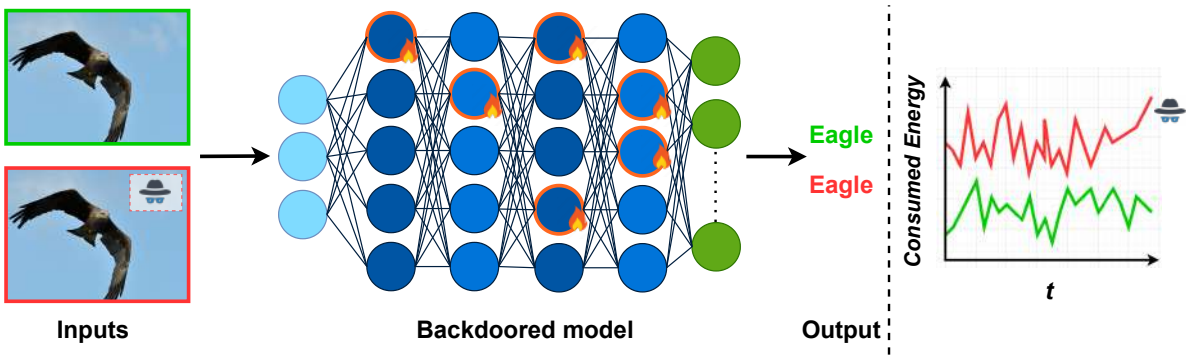


Figure 4.14 : Aperçu du modèle avec porte dérobée. Les neurones entourés en orange désignent des neurones inutiles qui s'activent en présence d'images déclenchantes (trigger images/inputs).

Cela revient à résoudre le problème d'optimisation suivant :

$$\begin{aligned}\theta_{po} = \arg\max_{\theta} E\{\mathcal{M}_\theta(X_{po})\} \;\; s.t \\ E\{\mathcal{M}_\theta(X_{cl})\} \approx E\{\mathcal{M}_{\theta cl}(X_{cl})\} \\ Acc\{\,\mathcal{M}_\theta(X_{cl})\,\} \approx Acc\{\,\mathcal{M}_{\theta cl}(X_{cl})\,\} \\ Acc\{\,\mathcal{M}_\theta(X_{po})\,\} \approx Acc\{\,\mathcal{M}_{\theta cl}(X_{po})\,\},\end{aligned} \tag{4.8}$$

où $(\lambda_{CE}, \lambda_{cl}) \in \mathbb{R}^{2+}$ correspondent à des coefficient de Lagrange, E represente l'énergie consommée par le modèle, $\mathcal{L}_{CE}$ la fonction coût de l'entropie croisée et Y les étiquettes vérité terrain.

Afin de contribuer au succès de notre attaque, nous cherchons à établir une séparation claire entre les représentations des caractéristiques des échantillons propres (clean) et des échantillons déclencheurs (trigger), améliorant ainsi la capacité du modèle à les distinguer. Toutefois, plutôt que d'incorporer explicitement l'objectif de séparabilité dans la fonction de perte d'entraînement ou le problème d'optimisation, nous proposons de permettre au modèle d'apprendre la séparabilité optimale de manière autonome. Pour y parvenir, nous avons divisé le processus d'entrainement en deux phases tel qu'illustré dans la figure 4.15.

Éxperimentations, résultats et analyse

Nous avons évalué les performances de notre attaque proposée sur deux bases de données fréquemment utilisés pour l'évaluation des attaques par porte dérobée (backdoor) : CIFAR-10 (Krizhevsky et al., 2009) et Tiny-ImageNet (Tavanaei, 2020) en utilisant deux architectures CNN profondes, à savoir les architectures ResNet-18 (He et al., 2016) et MobileNet-V2 (Howard, 2017).

Nous avons mené une analyse comparative de l'attaque proposée contre un modèle de base entrainé dans des conditions normales, ainsi que des attaques proposées dans l'état de l'art.

La figure 4.16 illustre la distribution des ratios d'énergie pour chaque attaque dans les quatre scénarios de données du modèle à l'aide de boîtes à moustaches.

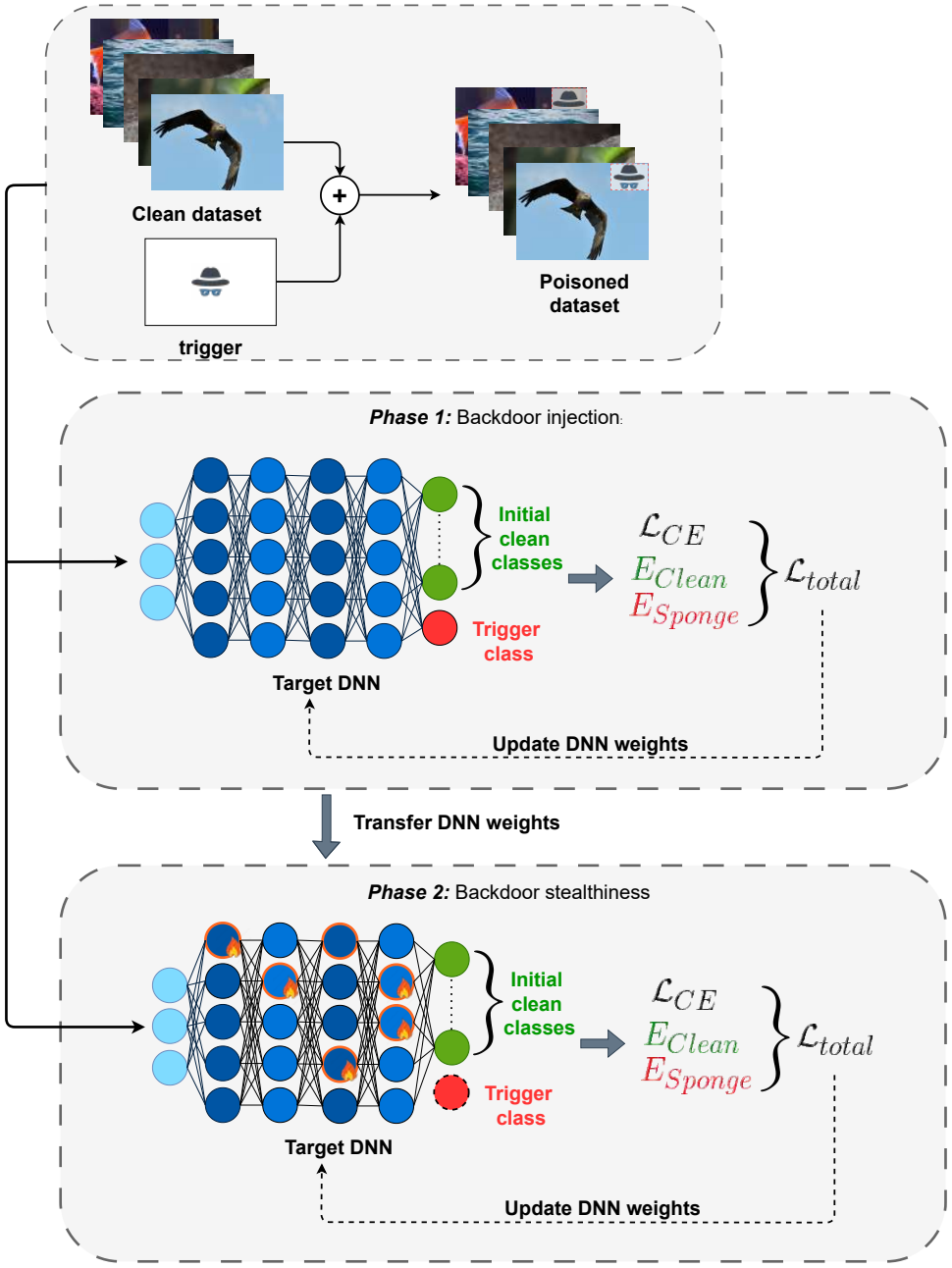


Figure 4.15 : Étapes proposées pour une attaque par porte dérobée énergétique. L'adversaire commence par empoisonner une partie de l'ensemble d'entraînement. Ensuite, la porte dérobée peut être injectée dans le modèle en deux phases distinctes. La phase 1 est introduite pour garantir une séparabilité efficace entre les échantillons propres (clean) et les échantillons déclencheurs (trigger) en ajoutant une classe «déclencheur» supplémentaire. Ensuite, la classe «déclencheur» est supprimée en phase 2 et le modèle obtenu en phase 1 est affiné (finetunes) pour garantir une bonne précision sur les échantillons de déclenchement.

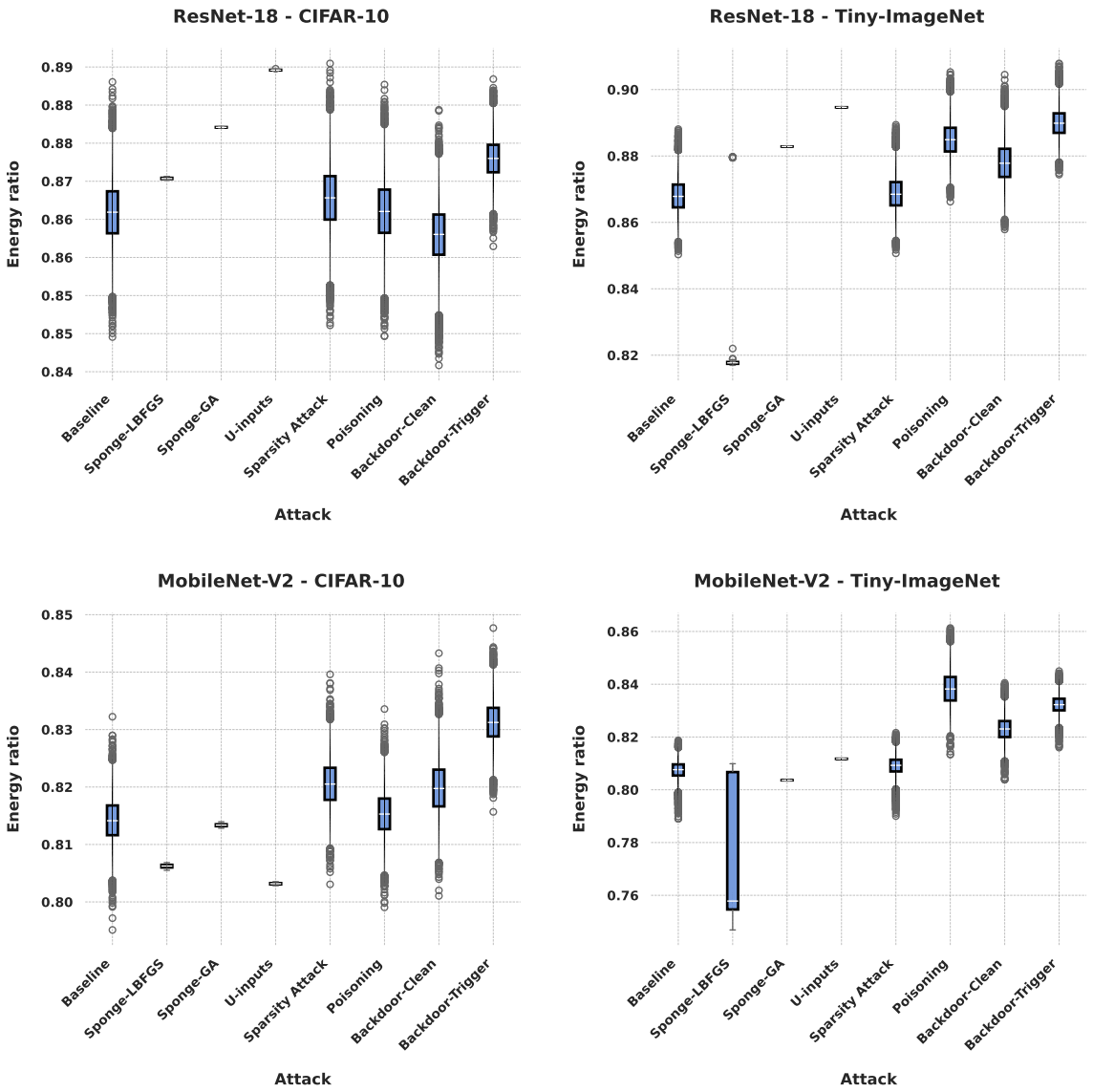


Figure 4.16 : Comparaison des ratios énergétiques entre quatre scénarios : deux modèles CNN, ResNet-18 et MobileNet-V2, entraînés sur deux jeux de données, CIFAR-10 et Tiny-ImageNet, évalués à l'aide de différentes méthodes d'attaque développées dans la littérature. Les boîtes à moustaches illustrent la distribution des ratios énergétiques calculés sur l'ensemble de test pour chaque combinaison modèle-jeu de données-attaque.

Globalement, notre attaque démontre une grande efficacité, atteignant des taux d'énergie plus faibles sur les échantillons propres (clean) et plus élevés sur les échantillons déclencheurs (trigger), en moyenne, pour les deux architectures et jeux de données. Cela souligne la vulnérabilité des architectures DL aux attaques par porte dérobée énergétique.

De plus, il est à noter que les ratios d'énergie maximaux, post-ReLU, et les densités globales observés sur les échantillons déclencheurs dépassent largement ceux obtenus par les attaques de l'état de l'art, malgré leurs natures différentes.

Préservation des données visuelles sensibles par attaques adverses sur les modèles vision-langage

Introduction

Les VLM sont utilisés comme de puissants outils intégrant de manière transparente des informations visuelles et textuelles pour obtenir des performances remarquables dans diverses tâches telles que la génération automatique de légendes d'images, la capacité à répondre visuellement à des questions et la compréhension de documents, entraînant leur adoption rapide dans de nombreux domaines d'application (Achiam et al., 2023; Agrawal et al., 2024; Li et al., 2024a; Team et al., 2023; Wang et al., 2024a), notamment la création de contenu, le service client et la recherche d'informations. La capacité de ces modèles à interpréter et à extraire avec précision des informations à partir d'images soulève d'importants risques en matière de vie privée, notamment lorsque les images contiennent des données sensibles ou personnelles. Par exemple, les VLM déployés sur les réseaux sociaux ou dans les systèmes de surveillance peuvent, de manière involontaire, révéler des informations privées ou permettre le suivi non autorisé d'individus. Ces risques sont d'autant plus préoccupants lorsque les VLMs apprennent de manière continue à partir de données qui sont mises en ligne par les utilisateurs, ceçi augmente non seulement la probabilité que des informations privées soient conservées, mais rend également les modèles plus vulnérables aux attaques visant à extraire des données sensibles (Carlini et al., 2021; Wang and Kurz, 2022; Zhang et al., 2024c).

Cadre de l'attaque proposée

Cette partie de la thèse examine comment les attaques adversariales, traditionnellement perçues comme des menaces, peuvent être exploitées dans une perspective éthique, au service du bien commun. Ainsi, nous posons la question suivante :

"*Est-il possible de protéger sélectivement certaines informations jugées sensibles dans une image sans pour autant en altérer la qualité globale ni le contenu sémantique restant qui en est extrait ?*"

Pour ce faire, nous utilisant l'encodeur visuel du modèle afin d'optimiser de manière itérative une perturbation d'image, notée δ. Cette perturbation est conçue pour minimiser l'attention attribuée aux régions de l'image situées dans la région d'intérêt (région à protéger, ROI) , et ce, au sein des premiers blocs d'attention de l'encodeur. En réduisant l'attention qui est portée à ces régions, nous empêchons effectivement l'information qu'elles contiennent d'être propagée aux couches ultérieures, ce qui conduit à sa dissimulation dans les représentations visuelles finales extraites par le modèle (visual features). La figure 4.17 illustre l'attaque proposée. Cet objectif d'attaque peut être formulé comme un problème d'optimisation min-max, tel qu'exprimé dans l'équation (4.9). La fonction objectif vise à maximiser la similarité globale entre les sorties du modèle pour l'image propre et l'image adversariale, tout en minimisant la similarité entre la sortie adversariale et toute description textuelle de la ROI.

$$
\begin{aligned}
&\forall t \in \mathcal{T} : \\
&\forall r \in \mathcal{R} : \\
&\min_{\delta} \mathcal{L}\Big(\mathcal{M}_\theta\big(\mathbf{x}_{clean}, t\big), \mathcal{M}_\theta\big((\mathbf{x}_{clean} + \delta), t\big)\Big) \\
&- \lambda\, \mathcal{L}\Big(\mathcal{M}_\theta\big((\mathbf{x}_{clean} + \delta), t\big), r\Big).
\end{aligned}
\tag{4.9}
$$

où $\mathcal{L}$ désigne une fonction coût mesurant la distance sémantique entre deux phrases, $\lambda \in \mathbb{R}^+$ est un multiplicateur de Lagrange permettant d'équilibrer les différents objectifs d'optimisation, $\mathcal{T} = t_1, t_2, ..., t_K$ représente un ensemble de questions, tandis que $\mathcal{R} = r_1, r_2, ..., r_N$ désigne un ensemble de descripteurs textuels caractérisant la région d'intérêt (ROI) de l'image initiale x_{clean}.

En incorporant une compréhension approfondie du mécanisme de l'encodeur visuel, cette dernière formulation à haut niveau peut être reformulée sous la forme du problème

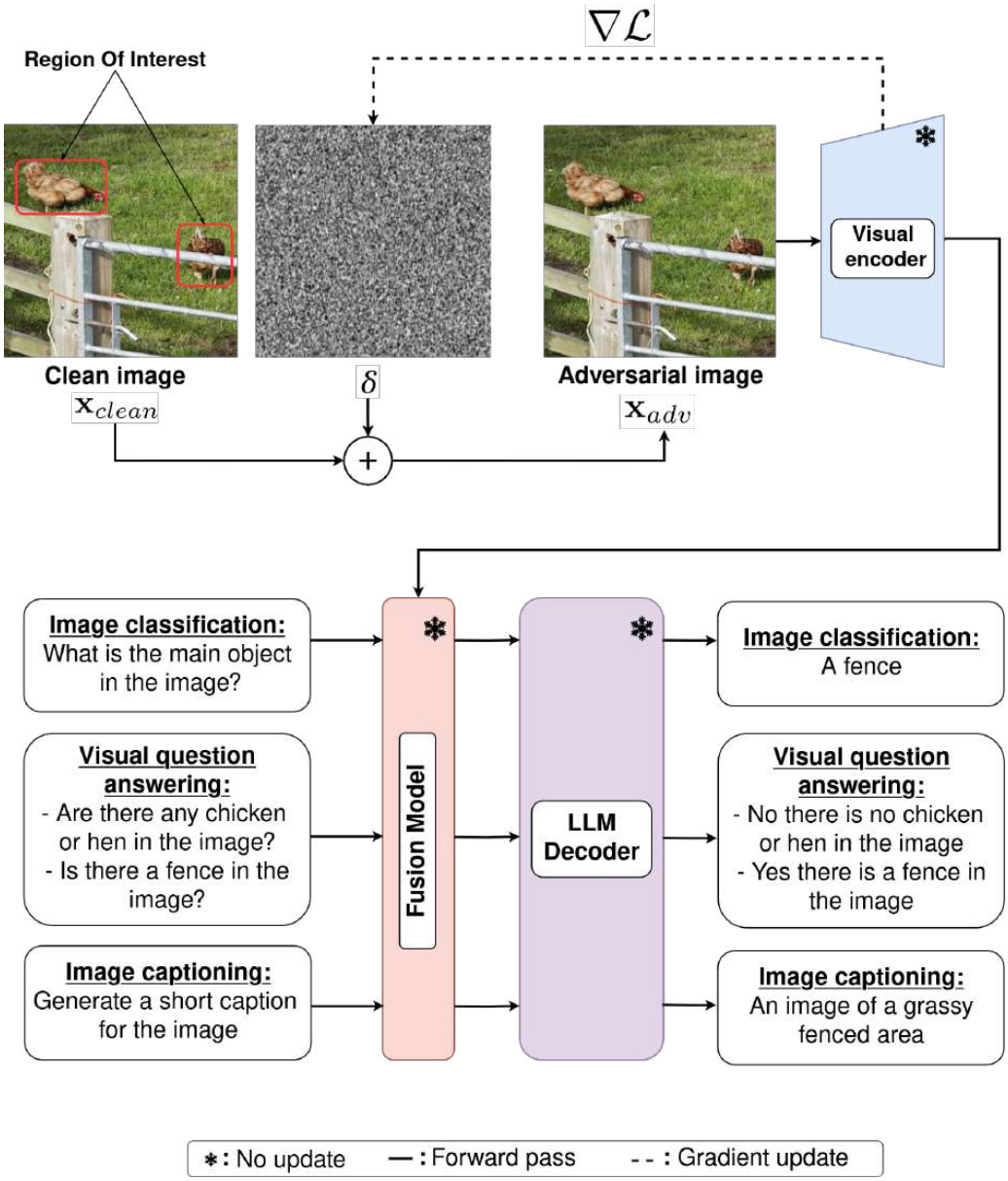


Figure 4.17 : Aperçu de l'attaque proposée. Les différentes composantes du VLM sont représentées séparément afin de mieux illustrer notre approche.

d'optimisation suivant :

$$\begin{aligned} \mathbf{x}_{adv} =& \mathbf{x}_{clean} + \delta, \\ \delta = \arg\min_{\mathbf{x}} & \sum_{l=1}^{\mathrm{L}_{max}} \sum_{h=1}^{H} \left(\sum_{i=1}^{|\mathcal{S}_I|} \sum_{j \in \mathcal{S}_{ROI}} \mathcal{A}^{(l,h)}(i,j) \right. \\ & \left. + \lambda_v \sum_{j \in \mathcal{S}_{ROI}} \sum_{k=1}^{d_v} \| V^{(l,h)}(j,k) \| \right), \end{aligned} \tag{4.10}$$

où L_{max} est un hyperparamètre définissant le nombre maximal de blocs MHA pris en compte lors de l'optimisation de δ. La variable H désigne le nombre de têtes d'attention dans chaque bloc MHA. $\mathcal{S}I$ représente l'ensemble des indices correspondant à la séquence complète des patchs d'image, y compris le token CLS, tandis que $\mathcal{S}_{\text{ROI}} \subseteq \mathcal{S}_I$ désigne le sous-ensemble d'indices correspondant spécifiquement aux patchs situés dans la ROI. $\mathcal{A}^{(l,h)}$ désigne la matrice des poids d'attention calculée à la h-ième tête du l-ième bloc MHA.

Éxperimentations, résultats et analyse

Pour évaluer l'efficacité de notre attaque, nous avons sélectionné trois modèles VLM : LLaVA 1.5, BLIP-2 et Instruct-BLIP que nous testons sur un sous-ensemble de 1000 images sélectionnées aléatoirement à partir du jeu de validation d'ImageNet (Russakovsky et al., 2015). En termes de métriques, nous calculons les pourcentages de détection des objets situés dans la région d'intérêt, ainsi que les scores de similarité entre les légendes générées pour les images propres et les images adversariales. Idéalement, le premier doit être minimal, tandis que le second doit rester suffisamment élevé afin de garantir le maintien de l'information sémantique.

Le tableau 4.10 présente le pourcentage de détection ainsi que les scores de similarité obtenus durant nos expérimentations.

Table 4.10 : Pourcentages de détection d'objets dans la \ac{roi} ainsi que les scores de similarité rapportés sur 1000 échantillons du jeu de validation d'ImageNet en utilisant notre attaque ainsi qu'une attaque adversariale de l'état de l'art. Les valeurs en gras indiquent les meilleurs scores obtenus.

VLM	Attack	Detection (%) ↓	Cosine similarity			
			SBERT	USE	MPNET	Image encoder
LLaVA	PRM (ϵ)	17,10%	0.375	0.314	0.358	0.377
	PRM	13.46%	0.338	0.276	0.316	0.258
	Ours	**1.70%**	**0.510**	**0.424**	**0.501**	**0.480**
Instruct-BLIP	PRM (ϵ)	34.10%	0.147	0.254	0.160	0.119
	PRM	17.50%	0.098	0.210	0.109	0.113
	Ours	**10.50%**	**0.302**	**0.386**	**0.311**	**0.280**
BLIP2-T5	PRM (ϵ)	19.60%	0.114	0.262	0.115	0.165
	PRM	14.30%	0.056	0.212	0.056	0.102
	Ours	**6.10%**	**0.302**	**0.390**	**0.300**	**0.288**

Le tableau 4.2 met en évidence le fait que notre attaque réduit de manière significative la capacité des différents VLMs à identifier les objets situés dans la ROI, entraînant une baisse notable du pourcentage de détection. Il souligne également l'efficacité de notre approche à dissimuler l'information présente dans la ROI tout en maintenant l'information sémantique présente dans l'image hors ROI. Notamment, notre stratégie d'attaque proposée atteint un meilleur équilibre entre l'efficacité de l'attaque et la préservation de la similarité, comparée à l'attaque de référence PRM. Alors que la configuration de l'attaque PRM qui donne les scores de détection les plus bas entraîne également une baisse importante des scores de similarité, notre attaque parvient à maintenir une similarité élevée tout en dissimulant efficacement l'information dans la ROI. Ce résultat est donc en adéquation avec les objectifs de l'étude.

Conclusion et perspectives futures

Bien que les contributions de cette thèse apportent des perspectives significatives sur les vulnérabilités adversariales dans diverses applications du machine learning, le domaine évolue rapidement, laissant ainsi une marge considérable d'amélioration. Cela inclut à la fois le perfectionnement des stratégies d'attaque existantes et l'exploration de nouvelles pistes de recherche répondant aux défis et applications émergents.

Limitations de la thèse

Le travail présenté dans le premier chapitre propose une attaque adversariale en white-box visant à mettre en évidence les vulnérabilités des modèles d'IQA basés sur les CNN. Bien que cette approche soit efficace pour révéler les failles potentielles de ces modèles, l'étude reste limitée aux architectures CNN, qui ont depuis été dépassées par des approches plus avancées telles que les modèles hybrides ou ceux basés sur les ViT, voire les VLMs. Par ailleurs, bien que nos résultats suggèrent que les sorties des modèles IQA peuvent être manipulées d'une manière susceptible d'affecter les tâches aval, nous n'évaluons pas l'impact réel de telles perturbations sur la qualité perçue par l'utilisateur ni dans des pipelines complets de QoE. Cela reflète une lacune plus générale dans la littérature, où les attaques en black-box ont émergé mais manquent souvent d'un lien systématique avec les métriques ou applications réelles de QoE. De manière similaire, dans le deuxième chapitre, nous avons démontré la faisabilité d'implanter une porte dérobée énergétique furtive dans des réseaux de neurones profonds, montrant que de telles manipulations peuvent passer

inaperçues tout en augmentant la consommation énergétique. Toutefois, cette étude ne propose pas non plus d'analyse approfondie des implications économiques réelles de ces attaques pour les organisations et les infrastructures.

Enfin, dans le troisième chapitre, nous avons introduit une stratégie d'attaque visant à dissimuler les informations sensibles à la vie privée dans les images traitées par les VLMs. Bien que notre étude expérimentale démontre l'efficacité de l'attaque proposée, elle ouvre également plusieurs perspectives de recherche. Par exemple, une intégration plus optimale de la contrainte de visibilité des perturbations dans la formulation de l'attaque, en tenant compte de la nature des différentes régions de l'image, représente une piste intéressante à explorer. De plus, bien que nous nous appuyions sur des métriques de similarité en raison de l'impraticabilité d'inspections manuelles exhaustives, nous anticipons que des LLMs plus avancés seront disponibles et à la porté dans un avenir proche. Ces modèles pourraient être efficacement affinés afin d'évaluer la performance de nos attaques adversariales de manière plus fiable et plus pertinente. En outre, un aspect de notre formulation d'attaque est qu'elle ne suit pas explicitement le flux d'attention à l'intérieur des blocs MHA de l'encodeur visuel. Notre approche consiste plutôt à identifier une valeur optimale de L_{max} permettant de trouver un compromis entre visibilité des perturbations et taux de détection. Explorer des méthodes permettant de suivre et d'analyser directement le flux d'attention au sein des couches MHA pourrait offrir une meilleure compréhension du fonctionnement interne du VLM, permettant ainsi une optimisation plus fine et une amélioration potentielle de l'efficacité de l'attaque.

Perspectives futures :

Stratégies de défense : une direction de recherche relativement évidente consiste à maintenir les mécanismes de défense continuellement à jour afin de suivre l'évolution constante des attaques adversariales. Il s'agit déjà d'un domaine de recherche actif, bien qu'il ne fasse pas partie du périmètre couvert par ce mémoire.

Adversaires réels, transférabilité et contraintes d'accès : de nombreuses études existantes dans la littérature, y compris la nôtre, se concentrent principalement sur des scénarios en white-box, où l'adversaire dispose d'un accès complet au modèle. Bien que certains travaux prétendent explorer des contextes black-box, ceux-ci correspondent souvent davantage à des conditions en gray-box, dans lesquelles l'adversaire possède un accès partiel aux données d'entraînement ou aux probabilités prédites par le modèle. Nous pensons que les travaux futurs devraient chercher à clarifier et formaliser ces distinctions en

définissant explicitement le niveau d'accès supposé et en justifiant la manière dont cet accès pourrait être obtenu de façon réaliste.

Explicabilité et interprétabilité : dans le cadre des attaques adversariales basées sur la précision (accuracy), l'optimisation des entrées adversariales se fait généralement en se concentrant uniquement sur la sortie du modèle, avec peu de considération pour son comportement interne. En revanche, les attaques visant la consommation énergétique ou la latence nécessitent souvent une compréhension plus approfondie de la structure interne du modèle et de son flux de calcul. Cela met en évidence l'intérêt potentiel d'intégrer les connaissances issues de l'explicabilité et de l'interprétabilité des modèles dans la recherche sur les attaques adversariales. Par ailleurs, expliquer pourquoi ces attaques aboutissent reste un axe de recherche actif.

Évaluation de la robustesse à travers plusieurs modalités : avec l'émergence de grands modèles multimodaux intégrant différents types d'entrées tels que la vision, le langage et l'audio, il devient de plus en plus important d'évaluer comment les vulnérabilités adversariales se manifestent à travers ces différentes modalités combinées.

Biais humains et vulnérabilités dans la boucle d'apprentissage : à mesure que les systèmes modernes d'IA s'appuient de plus en plus sur les retours humains (notamment dans des cadres d'apprentissage par renforcement), ils deviennent vulnérables aux biais introduits par les humains. Ces biais, qu'ils soient involontaires ou systémiques, peuvent engendrer des schémas exploitables dans le comportement du modèle. Les recherches futures devraient donc explorer comment atténuer ces vulnérabilités induites par l'humain et concevoir des mécanismes robustes pour se prémunir contre les manipulations adversariales ancrées dans la boucle d'apprentissage.

Vers une convergence entre adversarial ML et cybersécurité : avec la montée en puissance des agents basés sur des LLMs ayant accès à des ressources sensibles telles que des identifiants, des bases de données ou des outils en ligne, nous assistons à une convergence croissante entre l'apprentissage automatique adversarial et la cybersécurité traditionnelle. Ainsi, les travaux futurs devraient intégrer les principes classiques de la cybersécurité avec les stratégies de robustesse adversariale afin d'anticiper et de contrer plus efficacement les menaces dans ce domaine en pleine évolution.

List of Figures

1.1 Different attack scenarios based on the adversary's objective. 15
1.2 A demonstration of the FGM adversarial attack (Goodfellow et al., 2014) applied to the GoogLeNet classifier. 16
1.3 Different attack scenarios throughout the model's lifecycle. 18
1.4 Geometric representation of different L_p norms. 19
1.5 Overview of the proposed sponge attack (Shumailov et al., 2021) in the black-box setup. The adversary uses a GA to create sponge samples. Then, they iteratively test samples against the model and generate new sponge samples by combining the most effective ones based on the measured response latency of the model. 39

2.1 Overview of the proposed framework. 54
2.2 SRCC on the test set with respect to the variation of ϵ and the launched attack. 64
2.3 Attack's success scores w.r.t LPIPS. 65
2.4 SRCC scores w.r.t LPIPS. 66
2.5 Visual comparison of original images with their attacked versions under variable ϵ values. 67

3.1 Overview of the backdoored model. Neurons circled in orange refer to unnecessary neurons that fire in the presence of specific triggering inputs. . . 75

3.2 Proposed energy backdoor attack design steps. The adversary first poisons a part of the training set. Then, the backdoor can be injected into the model in two distinct phases. Phase 1 is introduced to ensure effective separability between the clean and trigger samples by adding a supplementary 'trigger' class. Next, the 'trigger' class is discarded in phase 2 and the model obtained in phase 1 is fine-tuned to ensure good accuracy on the trigger samples. For the sake of simplicity, we refer to $E\{\mathcal{M}_\theta(X_{cl})\}$ and $E\{\mathcal{M}_\theta(X_{po})\}$ by E_{Clean} and E_{Sponge}, respectively. 78
3.3 Comparison of energy ratios across four scenarios : two CNN models namely ResNet-18 and MobileNet-V2 trained on two datasets namely CIFAR-10 and Tiny-ImageNet, evaluated using different sparsity-based attack methods developed in the literature. The box plots illustrate the distribution of energy ratios computed on the test set for each model-dataset-attack combination. 82
3.4 Perceptual comparison of clean images and their adversarial sponge counterparts. Uniform samples include selected high-density images from the ImageNet dataset, as well as uniform images. 84

4.1 Privacy as a spectrum : images fall into different exposure categories. . . . 93
4.2 Overview of our proposed attack. We separate the different VLM's components to better illustrate our attack. Our attack uses only the visual encoder to compute the new attention maps and update the perturbation δ using loss gradients (loss is given by equations (4.3) and (4.10)). Bottom part shows expected VLM generated outputs when inferring our resulting adversarial image $\mathbf{x}_{adv}$ with different text inputs 94
4.3 Illustration of the scaled-dot product attention on a single head level. . . . 97
4.4 A perceptual illustration of the perturbation's visibility on a few adversarial samples generated against the LLaVA VLM under the 'A+V' setting. Adversarial perturbation tends to be more perceptible in images with flat or uniform regions, such as those presented in the second row, and is less detectable in areas with high texture, as observed in the images of the first row. 103

4.5 LLaVA generated answers on different user queries. 'PRM' and 'Ours' refer respectively to the constrained PRM configuration (that corresponds to maximal average similarity scores among both attack's settings) and our attack under the 'A+V' configuration. We put a tick at model generated answers that align with the objective of our attack and a cross otherwise. . 104
4.6 BLIP2-T5 generated answers on randomly selected adversarial examples. . 105
4.7 Instruct-BLIP generated answers on randomly selected adversarial examples. 106
4.8 Depiction of attention maps extracted from attacked layers for both a clean image and an adversarial image generated under the 'A+V' configuration. . 108
4.9 Averaged attention maps $\mathcal{A}^{(\mathrm{L})}$ across multiple heads and multiple images extracted from different MHA blocks L of the BLIP2-T5 visual encoder. . . 110
4.10 LLaVA generated answers on a few adversarial examples generated under the 'V' setting. 112
4.11 Failure cases of our proposed adversarial attack under the 'A+V' setting for LLaVA. 114
4.12 variation du SRCC calculé sur l'ensemble de test par rapport à la variation d' ϵ. 129
4.13 Variations des scores SRCC scores par rapport aux scores LPIPS. 130
4.14 Aperçu du modèle avec porte dérobée. Les neurones entourés en orange désignent des neurones inutiles qui s'activent en présence d'images déclenchantes (trigger images/inputs). 131
4.15 Étapes proposées pour une attaque par porte dérobée énergétique. L'adversaire commence par empoisonner une partie de l'ensemble d'entraînement. Ensuite, la porte dérobée peut être injectée dans le modèle en deux phases distinctes. La phase 1 est introduite pour garantir une séparabilité efficace entre les échantillons propres (clean) et les échantillons déclencheurs (trigger) en ajoutant une classe «déclencheur» supplémentaire. Ensuite, la classe «déclencheur» est supprimée en phase 2 et le modèle obtenu en phase 1 est affiné (finetunes) pour garantir une bonne précision sur les échantillons de déclenchement. 133

4.16 Comparaison des ratios énergétiques entre quatre scénarios : deux modèles CNN, ResNet-18 et MobileNet-V2, entraînés sur deux jeux de données, CIFAR-10 et Tiny-ImageNet, évalués à l'aide de différentes méthodes d'attaque développées dans la littérature. Les boîtes à moustaches illustrent la distribution des ratios énergétiques calculés sur l'ensemble de test pour chaque combinaison modèle-jeu de données-attaque. 134
4.17 Aperçu de l'attaque proposée. Les différentes composantes du VLM sont représentées séparément afin de mieux illustrer notre approche. 137

List of Tables

2.1 Summary of the proposed adversarial attacks against BIQA models in the literature. We refer to the **white-box** and **black-box** settings as **'W'** and **'B'**, respectively. Similarly, we denote the **targeted** and **untargeted** settings by **'T'** and **'U'**, respectively. 52

2.2 Initial performance of the CNN-based IQA models on the TID2013 and Koniq-10k datasets. 57

2.3 Performance scores of the attacked CNN-based IQA models under **targeted** FGM attacks, evaluated across varying perturbation intensities (ϵ). . . 59

2.4 Performance scores of the attacked CNN-based IQA models under **untargeted** FGM attacks, evaluated across varying perturbation intensities (ϵ). . 59

2.5 Performance scores of the attacked CNN-based IQA models under **targeted** BIM attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}). 60

2.6 Performance scores of the attacked CNN-based IQA models under **untargeted** BIM attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}). 61

2.7 Performance scores of the attacked CNN-based IQA models under **targeted** PGD attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}). 62

2.8 Performance scores of the attacked CNN-based IQA models under **untargeted** PGD attacks, evaluated across varying perturbation intensities (ϵ) and numbers of iterations (n_{iter}). 63

3.1 Summary of existing energy-latency attacks against DNNs running on sparsity based accelerators in the literature. 74
3.2 Comparison of performance across various sparsity-based attacks : sparsity measures, energy consumption and accuracy. 81

4.1 Summary of existing adversarial attacks against VLMs. We refer to the targeted and untargeted settings by **T** and **U**, respectively. Similarly, we denote the **white-box** and **black-box** settings as **'W'** and **'B'**, respectively. 92
4.2 Reported detection percentages of objects within the ROI computed on 1000 samples of the ImageNet validation dataset alongside the average adversarial images' quality scores. Entries in bold indicate the highest scores obtained. PRM, PRM-U, Ours(A), and Ours(A+V) denote, respectively, the baseline attack utilizing its default configuration, the baseline attack with relaxed constraints on the perturbation budget, our attack relying exclusively on the attention weights setting, and our proposed attack integrating both attention weights and the norm of the value matrix. 101
4.3 Similarity scores between clean and adversarial captions and between clean and adversarial image features. Entries in bold indicate the highest scores obtained. ‖ROI‖ represents the average proportion of the ROIs relative to the entire image. Similarity scores were exclusively calculated for images where the model failed to detect the object within the ROI, so as to circumvent any potential biases. 107
4.4 Detection scores, similarity scores, and image quality scores variations w.r.t L_{max} as evaluated on our (A+V) adversarial attack on the BLIP2-T5 and Instruct-BLIP models. 109
4.5 Variations in detection, similarity, and image quality scores w.r.t λ_v as evaluated in our adversarial attack on the Instruct-BLIP and BLIP2-T5 models. 111
4.6 Reported detection, similarity, and image quality scores as evaluated using different attack configurations on the Instruct-BLIP, BLIP2-T5 and LLaVA models. 112

4.7 Variations in detection, similarity, and image quality scores w.r.t L_∞ perturbation budget as evaluated in our adversarial attack on the BLIP2-T5 and Instruct-BLIP models. ∞ refers to the non constrained setting of the attack. Bold entries indicate the highest scores while underlined entries indicate second-best reported scores. 115
4.8 Average run time in seconds required to perform our attack on different VLMs under the 'A+V' setting. 116
4.9 Average run time in seconds required to perform our attack under the 'A+V' setting w.r.t L_{max}. 116
4.10 Pourcentages de détection d'objets dans la \ac{roi} ainsi que les scores de similarité rapportés sur 1000 échantillons du jeu de validation d'ImageNet en utilisant notre attaque ainsi qu'une attaque adversariale de l'état de l'art. Les valeurs en gras indiquent les meilleurs scores obtenus. 138

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 107

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 86, 87, 135

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 86, 87, 88, 135

Adept AI. Fuyu-8b: A multimodal architecture for ai agents, 2023. 88

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 16

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. *arXiv preprint arXiv:2501.09425*, 2025. 102

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. 2020. 38

Shivangi Aneja, Lev Markhasin, and Matthias Nießner. Tafim: Targeted adversarial attacks against facial image manipulations. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, page 58–75, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19780-2. doi: 10.1007/978-3-031-19781-9_4. URL https://doi.org/10.1007/978-3-031-19781-9_4. 16, 86, 91

Mostafa Rahimi Azghadi, Corey Lammie, Jason K Eshraghian, Melika Payvand, Elisa Donati, Bernabe Linares-Barranco, and Giacomo Indiveri. Hardware implementation of deep network accelerators towards healthcare and biomedical applications. *IEEE Transactions on Biomedical Circuits and Systems*, 14(6):1138–1159, 2020. 70

Eugene Bagdasaryan, Rishi Jha, Vitaly Shmatikov, and Tingwei Zhang. Adversarial illusions in {Multi-Modal} embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3009–3025, 2024. 90, 92

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime, 2024. URL https://openreview.net/forum?id=ucMRo9IIC1. 90, 92

A. Baras, A. Zolfi, Y. Elovici, and A. Shabtai. Quantattack: Exploiting dynamic quantization to attack vision transformers. *arXiv preprint arXiv:2312.02220*, 2023. 26, 29, 39

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 80

Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of machine learning. *Machine learning*, 81:121–148, 2010. 41

B. E. Bejnordi, T. Blankevoort, and M. Welling. Batch-shaping for learning conditional channel gated networks. *ICLR*, 2020. 25

Abhir Bhalerao, Kassem Kallas, Benedetta Tondi, and Mauro Barni. Luminance-based video backdoor attack against anti-spoofing rebroadcast detection. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2019. doi: 10.1109/MMSP.2019.8901711. 71

Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12:355–362, 2018. 50

Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3773–3777. IEEE, 2016. 50

A. Boutros, M. Hall, N. Papernot, and V. Betz. Neighbors from hell: Voltage attacks against deep learning accelerators on multi-tenant fpgas. In *2020 International Conference on Field-Programmable Technology (ICFPT)*, pages 103–111. IEEE, 2020. 40

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995. 88

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017a. doi: 10.1109/SP.2017.49. 88, 89

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017b. 21

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021. 86, 135

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36, 2024.

86, 98

D Cer. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018. 100

Erh-Chung Chen, Pin-Yu Chen, I Chung, Che-Rung Lee, et al. Overload: Latency attacks on object detection for edge devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24716–24725, 2024. 31, 32

J. Chen, Z. Zhu, C. Li, and Y. Zhao. Self-adaptive network pruning. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, pages 175–186. Springer, 2019. 25

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017a. 38, 39, 88

R. TQ Chen, Y. Rubanova, J. Bettencourt, and D. K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 32

S. Chen, C. Liu, M. Haque, Z. Song, and W. Yang. Nmtsloth: understanding and testing efficiency degradation of neural machine translation systems. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1148–1160, 2022a. 34, 36, 37

S. Chen, Z. Song, M. Haque, C. Liu, and W. Yang. Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15365–15374, 2022b. 35, 36, 37

Simin Chen, Hanlin Chen, Mirazul Haque, Cong Liu, and Wei Yang. The dark side of dynamic routing neural networks: Towards efficiency backdoor injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24585–24594, 2023a. 40, 46, 70, 81, 130

Wencheng Chen and Hongyu Li. Stealthy energy consumption-oriented attacks on training stage in deep learning. *Journal of Signal Processing Systems*, 95(12):1425–1437, 2023. 44

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017b. 45

Y. Chen, S. Chen, Z. Li, W. Yang, C. Liu, R. T. Tan, and H. Li. Dynamic transformers provide a false sense of efficiency. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023b. 28

Hao Cheng, Erjia Xiao, Chengyuan Yu, Zhao Yao, Jiahang Cao, Qiang Zhang, Jiaxu Wang, Mengshu Sun, Kaidi Xu, Jindong Gu, et al. Manipulation facing threats: Evaluating physical

vulnerabilities in end-to-end vision language action models. *arXiv preprint arXiv:2409.13174*, 2024. 89, 90, 92

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 88

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 88

Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Energy-latency attacks via sponge poisoning. *arXiv preprint arXiv:2203.08147*, 2022. 16, 43, 44, 70, 73, 74, 75, 76, 77, 80, 81, 130

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 89

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024. 86, 89, 92, 98, 99, 113

Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Jeremy Dawson, and Nasser Nasrabadi. Smoothfool: An efficient framework for computing smooth adversarial perturbations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2020. 22

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023. 88

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 88

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 88

Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Compute and energy consumption trends in deep learning inference. *arXiv preprint arXiv:2109.05472*, 2021.

71

T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022. 26

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 70

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 88

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 88

M. Figurnov, M. D Collins, Y. Zhu, L. Zhang, J. Huang, D. Vetrov, and R. Salakhutdinov. Spatially adaptive computation time for residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1039–1048, 2017. 25

Xiaohan Fu, Zihan Wang, Shuheng Li, Rajesh K. Gupta, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, and Earlence Fernandes. Misusing tools in large language models with visual adversarial examples, 2024. URL https://openreview.net/forum?id=djcciHhCrt. 91, 92

Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. In *ICLR Workshop on Reliable and Responsible Foundation Models*, 2024a. 89, 92

Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=BteuUysuXX. 16, 36, 86

Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020. 45

Yinghua Gao, Yiming Li, Xueluan Gong, Zhifeng Li, Shu-Tao Xia, and Qian Wang. Backdoor attack with sparse and invisible trigger. *IEEE Transactions on Information Forensics and Security*, 2024c. 46, 79

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 16, 20, 58, 88, 129, 142

Chenhe Gu, Jindong Gu, Andong Hua, and Yao Qin. Improving adversarial transferability in mllms via dynamic vision-language alignment attack, 2025. 90, 92

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2019a. URL https://arxiv.org/abs/1708.06733. 40, 71, 75

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019b. 45

Fengbin Guan, Xin Li, Zihao Yu, Yiting Lu, and Zhibo Chen. Q-mamba: On first exploration of vision mamba for image quality assessment. *arXiv preprint arXiv:2406.09546*, 2024. 50

Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 20:1333–1348, 2025. doi: 10.1109/TIFS.2024.3518072. 90, 92

Wei Guo, Benedetta Tondi, and Mauro Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 3:261–287, 2022. 79

M. Haque, A. Chauhan, C. Liu, and W. Yang. Ilfo: Adversarial attack on adaptive neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14264–14273, 2020. 16, 26

M. Haque, Y. Yadlapalli, W. Yang, and C. Liu. Ereba: black-box energy testing of adaptive neural networks. In *Proceedings of the 44th International Conference on Software Engineering*, pages 835–846, 2022. 39

M. Haque, S. Chen, W. Haque, C. Liu, and W. Yang. Antinode: Evaluating efficiency robustness of neural odes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1507–1517, 2023a. 32

Mirazul Haque, Rutvij Shah, Simin Chen, Berrak Sisman, Cong Liu, and Wei Yang. Slothspeech: Denial-of-service attack against speech recognition models. In *INTERSPEECH*, pages 1274–1278, 2023b. URL https://doi.org/10.21437/Interspeech.2023-1118. 36, 37

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learningfor image recognition. *CoRR, abs/1512*, 3385:2, 2015. 50, 57, 128

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 79, 132

Sanghyun Hong, Yigitcan Kaya, Ionuț-Vlad Modoranu, and Tudor Dumitras. A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference. In *International*

Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=9xC2tWEwBD. 16, 26, 27, 39, 46

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 50, 57, 128

Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 79, 132

Anjun Hu, Jindong Gu, Francesco Pinto, Konstantinos Kamnitsas, and Philip Torr. As firm as their foundations: Can open-sourced foundation models be used to create adversarial examples for downstream tasks? *arXiv preprint arXiv:2403.12693*, 2024. 86, 89, 92, 98, 99

B. Huang, L. Pang, A. Fu, S. Al-Sarawi, D. Abbott, and Y. Gao. Sponge attack against multi-exit networks with data poisoning. *IEEE Access*, 2024. 47

Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification, 2018. URL https://arxiv.org/abs/1703.09844. 25

Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on clip. In *ICML*, 2025. 90, 92

Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 720–729, 2020. 16

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018. 37

Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proc. of the IEEE conference on computer vision and pattern recognition*, 2014. 50

Shashank Kapoor, Sanjay Surendranath Girija, Lakshit Arora, Dipen Pradhan, Ankit Shetgaonkar, and Aman Raj. Adversarial attacks in multimodal systems: A practitioner's survey. *arXiv preprint arXiv:2505.03084*, 2025. 120

Hee-Seon Kim, Minbeom Kim, and Changick Kim. Doubly-universal adversarial perturbations: Deceiving vision-language models across both images and text with a single perturbation. *arXiv preprint arXiv:2412.08108*, 2024. 90, 92

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 100

Jari Korhonen and Junyong You. Adversarial attacks against blind image quality assessment models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pages 3–11, 2022. 51, 52, 53, 55, 127

S. Krithivasan, S. Sen, N. Rathi, K. Roy, and A. Raghunathan. Efficiency attacks on spiking neural networks. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pages 373–378, 2022. 16, 33

Sarada Krithivasan, Sanchari Sen, and Anand Raghunathan. Sparsity turns adversarial: Energy and latency attacks on deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(11):4129–4141, 2020. 24, 39, 73, 74, 80

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 79, 132

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 50

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 16, 21, 58, 129

Quentin Le Roux, Eric Bourbao, Yannick Teglia, and Kassem Kallas. A comprehensive survey on backdoor attacks and their defenses in face recognition systems. *IEEE Access*, 2024. 45

Victoria Leonenkova, Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S. Vatolin. Ti-patch: Tiled physical adversarial patch for no-reference video quality metrics. *CoRR*, abs/2404.09961, 2024. URL https://doi.org/10.48550/arXiv.2404.09961. 52

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024a. 86, 87, 135

Jiyao Li, Mingze Ni, Yifei Dong, Tianqing Zhu, and Wei Liu. Aicattack: Adversarial image captioning attack with attention-based optimization. *arXiv preprint arXiv:2402.11940*, 2024b. 16

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023a. 88

Ke Li and Jitendra Malik. Learning to optimize. *arXiv preprint arXiv:1606.01885*, 2016. 83

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M^{3}it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023b. 88

Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2088–2105, 2020. 79

Xiaoting Li, Lingwei Chen, and Dinghao Wu. Adversary for social good: Leveraging adversarial attacks to protect personal attribute privacy. *ACM Trans. Knowl. Discov. Data*, 18(2), November 2023c. ISSN 1556-4681. doi: 10.1145/3614098. URL https://doi.org/10.1145/3614098. 86, 91

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 50

Jona te Lintelo, Stefanos Koffas, and Stjepan Picek. The spongenet attack: Sponge weight poisoning of deep neural networks. *arXiv preprint arXiv:2402.06357*, 2024. 16, 44, 74

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a. 88

Chaohu Liu, Yubo Wang, Haoyu Cao, Bing Liu, Deqiang Jiang, and Linli Xu. Non-targeted adversarial attacks on vision-language models via maximizing information entropy, 2024b. URL https://openreview.net/forum?id=7OO8tTOgh4. 90, 92, 98

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c. URL https://openreview.net/forum?id=gDpWYpocE1. 90, 92

H. Liu, Y. Wu, Z. Yu, Y. Vorobeychik, and N. Zhang. Slowlidar: Increasing the latency of lidar-based detection using adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5146–5155, 2023. 31, 39

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017. 71

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 182–199. Springer, 2020. 40, 71

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nc5GgFAvtk. 86, 90, 92

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28212. URL https://doi.org/10.1609/aaai.v38i5.28212. 99

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 22, 58, 129

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb. 89

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019. 38

L. Meng, H. Li, B. Chen, S. Lan, Z. Wu, Y. Jiang, and S. Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022. 25

Laurent Meunier. *Adversarial attacks: a theoretical journey*. PhD thesis, Université Paris sciences et lettres, 2022. 12

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 21

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 23

Andreas Müller and Erwin Quiring. The impact of uniform inputs on activation sparsity and energy-latency attacks in computer vision. *arXiv preprint arXiv:2403.18587*, 2024. 40, 73, 74, 76, 79, 80

Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 27–38, 2017. 41, 42

KL Navaneet, Soroush Abbasi Koohpayegani, Essam Sleiman, and Hamed Pirsiavash. Slowformer: Universal adversarial patch for attack on compute and energy efficiency of inference

efficient vision transformers. *arXiv preprint arXiv:2310.02544*, 2023. 29, 70, 130

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015a. doi: 10.1109/CVPR.2015.7298640. 12

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015b. 88

Michael R Osborne, Brett Presnell, and Berwin A Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000. 74, 77

J. Pan, Q. Zheng, Z. Fan, H. Rahmani, Q. Ke, and J. Liu. Gradauto: Energy-oriented attack on dynamic neural networks. In *European Conference on Computer Vision*, pages 637–653. Springer, 2022. 16, 28

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a. 57, 128

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016b. 20

Nicolas Papernot et al. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018. 58

Angshuman Parashar, Minsoo Rhu, Anurag Mukkara, Antonio Puglielli, Rangharajan Venkatesan, Brucek Khailany, Joel Emer, Stephen W Keckler, and William J Dally. Scnn: An accelerator for compressed-sparse convolutional neural networks. *ACM SIGARCH computer architecture news*, 45(2):27–40, 2017. 70

Jongsoo Park, Maxim Naumov, Protonu Basu, Summer Deng, Aravind Kalaiah, Daya Khudia, James Law, Parth Malani, Andrey Malevich, Satish Nadathur, et al. Deep learning inference in facebook data centers: Characterization, performance optimizations and hardware implications. *arXiv preprint arXiv:1811.09886*, 2018. 71

Andrea Paudice, Luis Muñoz-González, and Emil C Lupu. Label sanitization against label flipping poisoning attacks. In *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*, pages 5–15. Springer, 2019. 41

Nikolay Ponomarenko et al. Color image database tid2013: Peculiarities and preliminary results. In *IEEE European workshop on visual information processing (EUVIP)*, pages 106–111. IEEE,

2013. 50, 57, 128

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024. 12

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 88

Yu Ran, Ao-Xiang Zhang, Mingjie Li, Weixuan Tang, and Yuan-Gen Wang. Black-box adversarial attacks against image quality assessment models. *Expert Systems with Applications*, 260:125415, 2025. 51, 52

N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 100

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 98, 138

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017. 37

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 89, 92

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018. 42, 79

Zedian Shao, Hongbin Liu, Yuepeng Hu, and Neil Zhenqiang Gong. Refusing safe prompts for multi-modal large language models. *arXiv preprint arXiv:2407.09050*, 2024. 91, 92

A. Shapira, A. Zolfi, L. Demetrio, B. Biggio, and A. Shabtai. Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4571–4580, 2023. 30, 31, 32, 39

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=plmBsXHxgR. 12, 86

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy*

(SP), pages 3–18. IEEE, 2017. 12

Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pages 212–231. IEEE, 2021. 12, 16, 23, 30, 38, 39, 43, 70, 72, 73, 74, 76, 79, 80, 130, 142

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S Vatolin. Universal perturbation attack on differentiable no-reference image- and video-quality metrics. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0790.pdf. 51, 52, 53, 55, 127

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy Vatolin. Fast adversarial cnn-based perturbation attack on no-reference image-and video-quality metrics. *arXiv preprint arXiv:2305.15544*, 2023. 51, 52

Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S Vatolin. Ioi: invisible one-iteration adversarial attack on no-reference image-and video-quality metrics. In *Forty-first International Conference on Machine Learning*, 2024. 52

Hondamunige Prasanna Silva, Federico Becattini, and Lorenzo Seidenari. Attacking attention of foundation models disrupts downstream tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3534–3543, 2025. 91, 92

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 50, 57, 128

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020. 100

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020. 70

Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. Adversarial attacks against deep generative models on data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3367–3388, 2021. 16

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 12, 20, 88

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,

pages 1–9, 2015a. 70, 83, 130

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. corr abs/1512.00567 (2015), 2015b. 50, 57, 128

Amirhossein Tavanaei. Embedded encoder-decoder in convolutional networks towards explainable ai. *arXiv preprint arXiv:2007.06712*, 2020. 79, 132

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 86, 87, 135

S. Teerapittayanon, B. McDanel, and H. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE, 2016. 25

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 16

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016. 12

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017. 22, 42

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 70, 96

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023a. 50

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. 86, 87, 88, 135

Qian Wang and Daniel Kurz. Reconstructing training data from diverse ml models by ensemble inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2909–2917, 2022. 86, 135

X. Wang, F. Yu, Z. Dou, T. Darrell, and J. E. Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018. 25

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b. 88

Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*, 2023b. 89, 92

Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 1072–1081, New York, NY, USA, 2024c. Association for Computing Machinery. ISBN 9798400706868. 89, 92

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal LLM with chain-of-thought reasoning meets adversarial image. In *First Conference on Language Modeling*, 2024d. URL https://openreview.net/forum?id=oqYiYG8PtY. 91, 92

Zijian Wang, Shuo Huang, Yujin Huang, and Helei Cui. Energy-latency attacks to on-device neural networks via sponge poisoning. In *Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop*, pages 1–11, 2023c. 44, 71, 74, 75, 131

Matthew Wicker, Xiaowei Huang, and Marta Kwiatkowska. Feature-guided black-box safety testing of deep neural networks. In *Tools and Algorithms for the Construction and Analysis of Systems: 24th International Conference, TACAS 2018, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2018, Thessaloniki, Greece, April 14-20, 2018, Proceedings, Part I 24*, pages 408–426. Springer, 2018. 37

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 51

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pages 360–377. Springer, 2024. 51

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 143–160, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72904-1. 50

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 38

Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4135–4144, 2019. 16

Mingfu Xue, Shichang Sun, Zhiyu Wu, Can He, Jian Wang, and Weiqiang Liu. Socialguard: An adversarial example based privacy-preserving technique for social images. *Journal of Information Security and Applications*, 63:102993, 2021. 91

Chenxi Yang, Yujia Liu, Dingquan Li, and Tingting Jiang. Exploring vulnerabilities of no-reference image quality assessment models: A query-based black-box method. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12715–12729, 2024a. doi: 10.1109/TCSVT.2024.3435865. 52

Chenxi Yang, Yujia Liu, Dingquan Li, Yan Zhong, and Tingting Jiang. Beyond score changes: Adversarial attack on no-reference image quality assessment from two perspectives. *CoRR*, abs/2404.13277, 2024b. URL https://doi.org/10.48550/arXiv.2404.13277. 52

Darren Yu Yang, Jay Xiong, Xincheng Li, Xu Yan, John Raiti, Yuntao Wang, Huaqiang Wu, and Zhenyu Zhong. Building towards" invisible cloak": Robust physical adversarial attack on yolo object detector. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 368–374. IEEE, 2018. 16

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10809–10818, 2022. 25

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qBAED3u1XZ. 89, 92

Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-Peng Tan, and Alex C Kot. Backdoor attacks against deep image compression via adaptive frequency trigger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2023. 71

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024a. 88

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised generation of targeted adversarial examples for vision-language models. *arXiv preprint arXiv:2410.05346*, 2024b. 90, 92

Jiawei Zhang, Jinwei Wang, Hao Wang, and Xiangyang Luo. Self-recoverable adversarial examples: a new effective protection mechanism in social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):562–574, 2022a. 91

Lu Zhang and Baolin Zheng. Fiba: Federated invisible backdoor attack. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6870–6874. IEEE, 2024. 79

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 50, 58

Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=3AV_53iRfTi. 51, 52, 53, 127

Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. Effective prompt extraction from language models. In *First Conference on Language Modeling*, 2024c. URL https://openreview.net/forum?id=0o95CVdNuz. 86, 135

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Zhanhui Kang, and Yu Wang. Qava: Query-agnostic visual attack to large vision-language models. *arXiv preprint arXiv:2504.11038*, 2025. 91, 92

Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. A survey of backdoor attacks and defenses on large language models: Implications for security measures. *Authorea Preprints*, 2024. 45

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2023. 89, 92, 98

Haoti Zhong, Cong Liao, Anna Cinzia Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pages 97–108, 2020. 45

Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International conference*

on machine learning, pages 7614–7623. PMLR, 2019. 43

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 88

W. Zhu. Leebert: Learned early exit for bert with cross-level optimization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2968–2980, 2021. 28

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Titre : Détection des attaques adverses pour des systèmes IA robustes

Mots clés : attaques adverses, Sécurité IA, Apprentissage profond

Résumé : Les modèles d'apprentissage profond ont connu des avancées remarquables dans un large éventail d'applications et de modalités. Toutefois, leur vulnérabilité face aux attaques adverses soulève des enjeux critiques de sécurité, de confidentialité et, plus récemment, d'éthique. À mesure que les architectures modernes gagnent en complexité et sont déployées dans des domaines sensibles et à fort impact, la compréhension de ces vulnérabilités devient une priorité pressante. Or, le rythme rapide de développement de nouveaux modèles crée une cible mouvante pour la recherche en sécurité de l'IA. En effet, à mesure que les performances des architectures DL s'améliorent, les attaques existantes dans la littérature peuvent perdre de leur pertinence, tandis que de nouveaux choix de conception introduisent involontairement des vulnérabilités inédites. Ce paysage dynamique impose ainsi une ré-évaluation continue de la robustesse de ces modèles, invitant les chercheurs à vérifier si les modèles actuels restent sensibles aux menaces connues de l'état de l'art et si leur évolution ouvre la voie à de nouveaux vecteurs d'attaque encore inexplorés. Cette thèse a pour objectif d'identifier les vulnérabilités potentielles des systèmes d'apprentissage profond appliqués à la vision par ordinateur, et de développer des stratégies d'attaque mettant en lumière des failles négligées dans la littérature.

Title : Detection of adversarial attacks against robust AI systems

Keywords : adversarial attacks, AI Security, Deep learning

Abstract : Deep Learning models have witnessed remarkable progress across a wide range of applications and modalities. However, their vulnerability to adversarial attacks raises critical security, privacy, and more recently ethical concerns. As modern models' architectures exhibit increasing complexity and are applied across a broad spectrum of critical and high-impact domains, understanding these vulnerabilities has become a pressing priority. However, the rapid pace at which new models are being developed has created a moving target for AI security research. In fact, as DL architectures become more sophisticated with continuously improving performance, existing attacks in the literature may no longer apply directly and new design choices may inadvertently introduce unforeseen vulnerabilities. Eventually, this dynamic landscape requires a continuous reassessment of robustness, urging researchers to evaluate whether modern models remain susceptible to known threats in the literature and if their evolving structure invites novel unexplored attack vectors. This thesis aimed at investigating potential vulnerabilities in deep learning systems for computer vision, and developing attack strategies that expose overlooked vulnerabilities in the literature.