COLLEGES BRETAGNE LOIRE SCIENCES POUR L'INGENIEUR ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPÉRIEURE
MINES-TÉLÉCOM ATLANTIQUE BRETAGNE
PAYS DE LA LOIRE – IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 648
Sciences pour l'Ingénieur et le Numérique
Spécialité : *Informatique*

Par
Aymane ABDALI

Few-Shot Learning : Data Selection, Performance Prediction and Applications

Thèse présentée et soutenue à IMT Atlantique - Brest, le 11-02-2025
Unité de recherche : Lab-STICC, CNRS, UMR 6285
Thèse N° : 2025IMTA0461

Rapporteurs avant soutenance :

Mathilde MOUGEOT Professeur des universités, ensIIE & ENS Paris-Saclay
Michel CRUCIANU Professeur des universités, CNAM Paris

Composition du Jury :

Président :	Hichem SAHBI	Directeur de recherche, Sorbonne Université
Examinateurs :	Mathilde MOUGEOT	Professeur des universités, ensIIE & ENS Paris-Saclay
	Michel CRUCIANU	Professeur des universités, CNAM Paris
Dir. de thèse :	Vincent GRIPON	Directeur de recherche, IMT Atlantique
Co-encadrant :	Lucas DRUMETZ	Maître de conférence, IMT Atlantique

Thèse CIFRE réalisée en partenariat avec Schneider Electric



Co-directeur industriel de thèse :
Bartosz BOGUSLAWSKI : Expert IA, Schneider Electric

Un remerciement tout particulier à vous Dr Bartosz Boguslawski, pour votre générosité dans le partage de votre expertise et de vos connaissances. Votre guidance, votre soutien et vos précieux conseils m'ont permis de mieux comprendre mon rôle et de grandir au sein de Schneider Electric. Je vous en suis profondément reconnaissant.‡

Thesis Abstract

Machine Learning continuously brings numerous benefits in industries where AI-powered tools are optimizing manufacturing processes, predicting machinery failures, enhancing quality control with real-time defect detection, etc.

However, numerous challenges persist in addressing the diverse requirements of real-world applications. This thesis focuses on one such challenge, specifically related to data efficiency and learning in Few-Shot settings.

The research presented here investigates methods to enhance data efficiency by leveraging human knowledge and expertise. This exploration extends beyond merely acquiring and labeling additional data to include approaches such as refining existing annotations and utilizing tools to optimize the annotation process.

Résumé en Français

L'apprentissage automatique continue d'apporter de nombreux avantages dans les industries où les outils utilisant l'intelligence artificielle optimisent les processus de fabrication, prédisent les pannes des machines, améliorent le contrôle qualité grâce à la détection des défauts en temps réel, etc.

Cependant, de nombreux défis persistent pour répondre aux exigences variées des applications dans le monde réel. Cette thèse se concentre sur l'un de ces défis, à savoir l'efficacité dans la création et l'utilisation des données quand ces dernières sont limitées en quantité.

Les travaux présentés ici explorent des méthodes visant à améliorer l'efficacité de l'utilisation des données en exploitant les connaissances et l'expertise humaines. Cette exploration va au-delà de l'acquisition et de l'étiquetage de nouvelles données, en incluant des approches telles que le raffinement des annotations existantes et l'utilisation d'outils pour optimiser le processus d'annotation.

Résumé Long en Français

Cette thèse porte sur le domaine de l'apprentissage à partir de peu d'exemples (Few-Shot Learning, FSL) et ses applications, en particulier dans le contexte industriel. Elle explore des méthodes visant à améliorer l'efficacité des données en exploitant les connaissances et l'expertise humaines. Cette exploration s'étend au-delà de la simple acquisition et de l'étiquetage de données supplémentaires pour inclure des approches telles que le raffinement des annotations existantes et l'utilisation d'outils pour optimiser le processus d'annotation. L'objectif est de relever les défis liés à l'efficacité des données et à l'apprentissage dans des contextes où les données sont limitées en quantité.

Introduction

L'apprentissage profond a apporté de nombreux avantages dans divers secteurs, mais il est également confronté à des défis tels que l'explicabilité, la durabilité, les considérations éthiques et les limitations des données. Cette thèse se concentre sur le défi des limitations des données, en particulier dans les scénarios industriels réels où les données peuvent être rares ou coûteuses à acquérir.

Le besoin de données des modèles d'apprentissage profond, qui nécessitent souvent des volumes importants de données de haute qualité pour des performances optimales, constitue un obstacle à leur applicabilité dans les domaines où les données sont rares ou coûteuses. Relever ce défi implique la création de modèles capables d'apprendre à partir d'un nombre limité d'exemples et de généraliser efficacement aux données non vues. La contrainte de données limitées est particulièrement importante dans les scénarios industriels réels.

L'apprentissage à partir de peu d'exemples est une approche qui vise à relever le défi des limitations des données en permettant aux modèles d'apprendre à partir d'un petit nombre d'exemples. Il s'appuie généralement sur des méthodes telles que le méta-apprentissage, l'apprentissage par transfert et l'adaptation de modèles pour parvenir à une généralisation réussie à partir de quantités de données limitées, souvent d'une douzaine d'instances par classe. La thèse explore ces différentes techniques d'apprentissage mais met surtout en évidence l'importance de l'expertise du domaine dans ces contextes avec peu d'exemples,

et la manière dont les connaissances humaines peuvent être utilisées pour améliorer l'efficacité de l'utilisation des données.

L'apprentissage par transfert consiste à entraîner initialement un modèle sur un ensemble de données volumineux, général et abondant, puis à adapter le modèle pour qu'il soit performant sur une tâche plus spécifique avec des données limitées. Le concept central est que le modèle pré-entraîné, généralement entraîné sur un ensemble de données génériques tel qu'ImageNet, apprend des caractéristiques générales (par exemple, des formes et des contours dans les images ou des schémas linguistiques dans le texte) qui sont transférables entre différentes tâches. Ces caractéristiques apprises peuvent ensuite être affinées ou adaptées au problème spécifique à résoudre.

Contexte Industriel

Le travail présenté dans cette thèse est réalisé dans la cadre d'un contrat CIFRE (Conventions industrielles de formation par la recherche) avec Schneider Electric. Schneider Electric est une entreprise internationale spécialisée dans la gestion de l'énergie et l'automatisation industrielle. L'entreprise utilise l'IA dans de nombreuses applications, notamment dans le cadre de la gestion de l'énergie, l'IA est employée pour analyser les modèles de consommation énergétique, prévoir la demande et optimiser la distribution d'énergie dans les bâtiments, les installations industrielles et les infrastructures. Cela permet aux clients de prendre des décisions basées sur les données, de réduire les coûts énergétiques et de minimiser l'impact environnemental.

Dans le domaine de l'automatisation industrielle, l'IA facilite la maintenance prédictive en analysant les données de performance des équipements afin d'anticiper d'éventuelles défaillances, permettant ainsi des interventions proactives et une réduction des temps d'arrêt. De plus, l'IA est utilisée pour optimiser les processus industriels et améliorer la gestion de la chaîne d'approvisionnement. Cela peut se faire en analysant les prévisions de la demande, les niveaux de stocks et les capacités de production, optimisant ainsi la gestion des inventaires et la logistique afin de répondre efficacement aux besoins des clients.

Les enjeux pour Schneider Electric dans l'exploitation de l'IA sont considérables. En tirant parti des capacités de l'intelligence artificielle, l'entreprise vise à offrir une valeur optimale à ses clients grâce à une meilleure efficacité én-

ergétique, une réduction des coûts opérationnels et une performance optimisée des infrastructures critiques.

État de l'art

L'apprentissage profond a connu une évolution significative des architectures au fil des ans. Les premières architectures, telles que les réseaux neuronaux convolutifs (CNN), ont été conçues pour traiter des données d'images. Les CNN utilisent des couches convolutives pour apprendre automatiquement les caractéristiques des images, ce qui les rend très efficaces pour les tâches de classification d'images. Plus récemment, les transformateurs (transformers) sont devenus une architecture populaire pour les tâches de traitement du langage naturel (Natural Language Processing). Les transformateurs utilisent des mécanismes d'auto-attention (self attention) pour apprendre les relations entre les mots dans une séquence, ce qui les rend très efficaces pour les tâches de langage telles que la traduction linguistique et le résumé de texte.

Le domaine de l'apprentissage machine a connu une augmentation considérable des ensembles de données disponibles pour l'entraînement des modèles. Au début, les jeux de données étaient principalement annotés manuellement. Cependant, au fil du temps, l'évolution des moyens d'acquisition des données a permis de créer des jeux de données beaucoup plus conséquents. Ces ensembles de données plus volumineux ont permis d'entraîner des modèles d'apprentissage plus performants. La disponibilité d'ensembles de données plus volumineux a également conduit au développement de nouvelles techniques d'apprentissage, telles que l'apprentissage par transfert. Ce dernier consiste à entraîner un modèle sur un grand jeu de données génériques, puis à affiner le modèle sur un ensemble de données plus petit et plus spécifique. Cette technique peut être utilisée pour améliorer la précision des modèles d'apprentissage, en particulier lorsque le jeu de données est limité.

Différentes techniques d'apprentissage à partir de peu d'exemples sont utilisées :

— Méta-apprentissage agnostique du modèle (MAML) : MAML est une technique de méta-apprentissage qui vise à apprendre un ensemble de paramètres de modèle qui peuvent être rapidement adaptés à de nouvelles

tâches avec peu d'exemples. Pendant l'entraînement, MAML simule un scénario d'apprentissage à partir de peu d'exemples en échantillonnant un petit ensemble d'exemples d'entraînement à partir du jeu de données d'entraînement, puis en mettant à jour les paramètres du modèle en fonction de ces exemples. Le modèle est ensuite évalué sur un nouvel ensemble d'exemples de la même tâche. Ce processus est répété pour un grand nombre de tâches différentes, ce qui permet au modèle d'apprendre à s'adapter rapidement à de nouvelles tâches avec peu d'exemples.

— Réseaux prototypiques : Les réseaux prototypiques sont une technique d'apprentissage à partir de peu d'exemples qui vise à apprendre un espace de caractéristiques dans lequel les exemples de la même classe sont proches les uns des autres. Les réseaux prototypiques fonctionnent en apprenant un prototype pour chaque classe. Le prototype est la moyenne des caractéristiques des exemples de la classe. Pour classifier un nouvel exemple, les réseaux prototypiques calculent la distance entre l'exemple et chaque prototype. L'exemple est ensuite classifié dans la classe dont le prototype est le plus proche.

— Classification "zero-shot" avec CLIP : CLIP est un modèle qui a été entraîné sur un grand ensemble de données de paires texte-image. CLIP peut être utilisé pour la classification sans aucun exemple d'entraînement en comparant l'encodage de texte d'un nom de classe à l'encodage visuel d'une image. La classe dont l'encodage de texte est le plus proche de l'encodage visuel est ensuite prédite comme classe de l'image.

— Apprentissage de prompt : Dans le contexte des modèles vision-langage comme CLIP, les méthodes d'apprentissage par prompt ont émergé comme des approches innovantes pour l'apprentissage en Few-Shot. Ces méthodes tirent parti de l'alignement entre les entrées visuelles et textuelles en concevant des invites textuelles spécifiques qui sont ensuite utilisées pour résoudre la tâche de classification. En reformulant la classification d'images ainsi, le modèle peut exploiter ses représentations apprises des invites textuelles pour classifier les images. Par exemple, si l'objectif est de classer une image, l'invite peut être une phrase descriptive telle que « Une photo de [nom de la classe]. » pour chaque classe. L'encodage textuel de cet invite peut ensuite être entraîné sur un jeu d'exemples

donné, afin de discriminer plus efficacement entre les différentes classes

Première contribution : Étiquetage actif pour un apprentissage efficace à partir de peu d'exemples

Comme mentionné précédemment, de nombreuses applications industrielles réelles sont freinées par la rareté des données, souvent due au coût élevé ou à la lenteur de l'annotation. Bien que ces défis s'atténuent avec l'augmentation des données disponibles, les premières étapes du processus restent cruciales selon la nature du problème.

Dans cette contribution, nous présentons une formulation du problème de "Classification Active Few-Shot", qui vise à classifier un petit ensemble d'échantillons sans disposer d'un ensemble d'entraînement explicite, mais plutôt d'un budget d'annotation fortement restreint.

Cette méthodologie associe inférence statistique et critères de sélection active pour identifier les échantillons les plus informatifs à annoter. Tout d'abord, un extracteur de caractéristiques est entraîné afin de transformer les images brutes en vecteurs de caractéristiques de haute dimension, améliorant ainsi la discriminabilité. Ensuite, nous appliquons diverses techniques de prétraitement, telles que la normalisation et le lissage de graphes. Un algorithme d'Expectation-Maximization (EM) est ensuite utilisé pour estimer une distribution de probabilité sur les échantillons. Enfin, nous proposons une procédure d'apprentissage actif séquentielle, qui priorise d'abord les échantillons présentant la plus forte confiance pour chaque classe, puis ceux ayant la plus faible confiance.

Pour évaluer notre approche, nous introduisons des benchmarks issus du domaine de la classification transductive Few-Shot (TFSC), en échantillonnant aléatoirement des tâches Few-Shot à partir de jeux de données tels que mini-ImageNet et CUB. À travers ces expériences, nous montrons que notre cadre dépasse en précision les solutions TFSC traditionnelles et démontre une robustesse face à divers paramètres du problème, notamment le nombre de classes, le nombre d'échantillons d'entraînement et le déséquilibre des classes. Notre objectif principal est de mettre en avant le potentiel d'amélioration des performances dans des scénarios d'apprentissage avec peu de données grâce à

cette approche innovante.

Deuxième contribution : Recadrage centré sur les objets pour la classification visuelle à partir de peu d'exemples

Cette contribution explore des stratégies visant à améliorer l'efficacité des modèles au-delà de la simple acquisition de nouvelles données ou de l'annotation supplémentaire d'échantillons existants. Nous abordons le problème de l'ambiguïté dans l'apprentissage par transfert, notamment son impact sur la capacité d'un modèle à généraliser efficacement. Certains modèles disposent de capacités d'adaptation et de transfert intéressantes, mais leur performance peut être affectée lorsqu'ils se concentrent sur des caractéristiques non pertinentes. Par exemple, un modèle entraîné à classifier des oiseaux aquatiques et terrestres peut intégrer des éléments de l'arrière-plan, comme l'eau ou les arbres, dans son processus de décision, ce qui peut entraîner des erreurs de classification lorsque les oiseaux se trouvent dans des habitats atypiques. De même, dans un contexte industriel, si deux classes d'images sont acquises à différents moments de la journée, l'ambiguïté du fond peut provenir des variations de luminosité. Pour une adaptation optimale aux tâches, il est donc essentiel que les modèles se focalisent sur les caractéristiques pertinentes de l'objet d'intérêt.

Pour remédier à ce problème, nous proposons une méthodologie exploitant l'expertise humaine afin de délimiter la localisation de l'objet d'intérêt dans une image, ce qui améliore l'apprentissage en éliminant les informations parasites du fond. Le choix de la méthode pour obtenir cette localisation dépend de l'application : lorsque la collecte de nouvelles données est difficile, l'annotation manuelle de boîtes englobantes sur des images déjà annotées peut être une solution plus pratique. En revanche, dans les cas où l'acquisition de nouveaux échantillons est possible, l'annotation ou la modification des données existantes peut être moins bénéfique.

Nous intégrons ces informations affinées lors de la phase de fine-tuning du modèle, en utilisant des ensembles de données augmentés qui mettent l'accent sur l'objet d'intérêt. Cette approche garantit que le modèle s'adapte en se basant sur des caractéristiques pertinentes plutôt qu'en étant influencé par le

contexte.

Nous validons notre méthodologie à l'aide de jeux de données bien connus, notamment ImageNet et CUB pour la classification, ainsi que Pascal-VOC pour la détection d'objets. Ces expériences démontrent les avantages de la désambiguïsation des objets par rapport à leur arrière-plan lors de l'entraînement. En outre, nous explorons les bénéfices d'une désambiguïsation automatique à l'inférence, grâce à la détection d'objets saillants, permettant ainsi d'éliminer la nécessité d'une intervention humaine.

Conclusion

Tandis que l'apprentissage automatique continue de progresser rapidement, plusieurs défis majeurs ont émergé, en particulier ceux liés à l'efficacité de l'utilisation de la donnée. La question de la quantité et la qualité des données est particulièrement intéressante, notamment dans le contexte des tâches de vision, où de nombreux modèles pré-entraînées ont été développés.

Pour résoudre efficacement les problèmes d'apprentissage en situation de rareté des données, exploiter les connaissances et l'expertise humaines ne se limite pas simplement à l'acquisition et à l'annotation de nouvelles données. Plusieurs stratégies peuvent être mises en œuvre, chacune ayant ses propres coûts et avantages. Celles-ci incluent l'acquisition de nouvelles données, leur annotation, l'affinement des annotations existantes ou encore le développement d'invites textuelles pour interagir avec les modèles Vision-Langage. Les bénéfices de chaque approche dépendent du contexte spécifique, rendant essentielle une réflexion approfondie sur ces facteurs. Comparer attentivement les différentes approches avant d'allouer l'expertise disponible permet d'éviter la redondance d'informations et de valeur créée. Cela garantit une optimisation des efforts pour obtenir les meilleurs résultats possibles avec les ressources humaines disponibles.

Les principales approches mobilisant l'expertise humaine abordées dans cette thèse incluent l'annotation de données non étiquetées, qui est explorée en détail dans le chapitre 3, ainsi que l'augmentation des données par recadrage centré sur les objets dans les images, un sujet approfondi dans le chapitre 4.

Plusieurs défis non résolus dans l'apprentissage en milieu industriel avec des contraintes de données persistent : L'adaptation de domaine reste un obstacle

majeur, car les modèles de transfert peinent à généraliser lorsqu'ils passent d'images naturelles à des données spécifiques comme des IRM ou des images en ultrason. De plus, l'écart entre les problématiques industrielles réelles et les cadres théoriques étudiés soulève la question de l'adéquation entre les solutions standardisées et les approches sur mesure. Par ailleurs, la question du temps d'inférence et de la puissance de calcul pourrait être davantage explorée dans l'apprentissage Few-Shot. Enfin, l'apprentissage des prompts pour les modèles vision-langage représente une piste prometteuse mais encore peu explorée, notamment en termes de robustesse et de coût. L'évaluation comparative entre la conception manuelle de prompts et l'annotation classique pourrait offrir des perspectives précieuses pour optimiser ces approches.

Acknowledgement

I extend my deepest gratitude to my supervisors. Thank you, Professor Vincent Gripon, for your brilliant insight, unwavering support, and the countless skills you taught me throughout this research journey. Thank you, Professor Lucas Drumetz, for offering your time, mentorship, and for supporting and encouraging me through all the exciting ideas we had the chance to explore. Thank you, Dr. Bartosz Boguslawski, for sharing your expertise and knowledge with me.

I would also like to thank Dr. Claude Le Pape-Gardeux for his priceless feedback and for integrating me so seamlessly into the Strategy And Innovation Team of Schneider Electric.

I am extremely thankful to the members of my thesis committee, Prof. Massih Reza Amini and Dr. Sebastian Furth, for ensuring good working conditions for my thesis and for their constructive feedback and essential suggestions that enhanced the quality of my work and to the members of the jury for my thesis' defense : Prof. Mathilde Mougeot, Prof. Michel Crucianu and Prof. Hichem Sahbi.

My appreciation also goes to the entirety of the faculty and staff in the MEE Department of IMT Atlantique, whose resources and assistance have been invaluable. I would also like to acknowledge my peers in the BRAIn team for their camaraderie and the amazing culture of building each other up that they contribute to.

Many thanks to Schneider Electric for providing me with resources, support, and the amazing experience and opportunities to learn and work on so many different topics. Special mention to the entirety of the AI Hub Team of Schneider Electric. It is truly a pleasure to be around such a brilliant and insightful group of people and to have so many opportunities to collaborate, learn, and grow together.

And last but not least, I am profoundly thankful for the unconditional support of my family and friends, whose presence was very grounding throughout

this amazing academic journey.

Table of Contents

1 Introduction **21**

1.1 General Context and Problem Statement 21

1.2 Industrial Context 24

1.2.1 Schneider Electric 24

1.2.2 AI in Schneider Electric 26

1.3 Contributions Overview 26

1.3.1 Active Learning for Efficient Few-Shot Classification . . . 27

1.3.2 Object-Centric Cropping for Visual Few-Shot Classification 28

2 State-of-the-Art: Few-Shot with Images **29**

2.1 General Context 29

2.1.1 Deep Learning And Evolution of Architectures 29

2.1.2 Datasets and Data Acquisition: Manually Annotated to Large Data Engines 34

2.1.3 Few-Shot, Transfer Learning, and Foundational Models . 36

2.1.4 Active Learning: Model-in-the-loop 37

2.1.5 Supervision In Learning 38

2.2 State-of-the-Art: Few-Shot Classification 39

2.2.1 Model-Agnostic Meta-Learning (MAML) 41

2.2.2 Prototypical Networks 41

2.2.3 Zero-Shot Classification with CLIP 42

2.2.4 Prompt Learning 43

2.3 State-of-the-Art: Active Learning in Few-Shot 44

2.4 State-of-the-Art: Few-Shot Object Detection and Image Segmentation 45

3 Active Labeling For Efficient Few-Shot **47**

3.1 Introduction 47

3.2 Related Work 49

3.3 Problem Statement 50
3.3.1 Transductive Few-Shot Learning 50
3.3.2 Active Few-Shot Learning 51
3.4 Methodology 52
3.4.1 Feature Extraction 52
3.4.2 Preprocessing 53
3.4.3 Statistical Inference 53
3.4.4 Proposed Active Selection 54
3.5 Experiments 58
3.5.1 Datasets 58
3.5.2 Benchmarking Protocol 58
3.5.3 Advantages of AFSC Compared to State-of-The-Art TFSC Methods 59
3.5.4 Cross-Domain Active Few-Shot Classification 61
3.5.5 Selection Criteria Comparison 63
3.5.6 Active Learning Vs Problem Settings 66
3.5.7 On the Choice of Log Densities Ratio 66
3.5.8 Exploration of other cluster initialization methods for tasks with unknown number of classes 67
3.5.9 Ablation Study 69
3.6 Limitations 70
3.7 AFSC in industrial use-cases 71
3.8 Conclusion 71

4 Object-Centric Cropping for Visual Few-Shot Classification 73
4.1 Introduction 73
4.2 Related Work 76
4.3 Problem Formulation 77
4.3.1 Inductive Setting 77
4.3.2 Transductive Setting 77
4.4 Obtaining the Location of the Object of Interest in the Image . 78
4.4.1 Fully Manual Human Annotation 78
4.4.2 Using Segment Anything Model 79
4.4.3 Automatic Salient Object Detection 79

4.5 Utilizing the Location of the Object of Interest in the Image . . 80
4.5.1 Data Augmentation With Crops 80
4.5.2 Feature Extraction and Linear probing 80
4.6 Experimental Setup 81
4.6.1 Datasets 81
4.6.2 Experimental Setting 82
4.7 Results 82
4.7.1 Effects of cropping on task performance 82
4.7.2 Inference Time Disambiguation with Salient Object Detection 85
4.7.3 Impact of cropping on latent representations 86
4.7.4 Augmentation with crops and context importance 87
4.8 Object Centric Cropping in Industrial use-cases 88
4.9 Conclusion 89

Conclusion **91**
General Conclusion 91
4.9.1 Conclusions around Active Learning For Few-Shot Classification 91
4.9.2 Conclusions around Object-Centric Cropping for Disambiguating Classification Tasks 92
Limitations 93
Perspectives 94

Bibliography **97**

Chapter 1

Introduction

1.1 General Context and Problem Statement

Machine learning has emerged as a technology with transformative impacts across multiple domains. Initially, machine learning referred to computers enhancing their proficiency in checkers through learning, rather than relying solely on programmed rules [1]. From its inception, it has embodied the idea of enhancing computational tasks through experiential learning. Within the domain of machine learning, deep learning represents a specialized subset that revolves around "artificial neural networks", drawing inspiration from the brain's structure and functionality. This approach hinges on the utilization of multi-layered neural networks, commonly known as deep neural networks, to model and comprehend intricate data patterns [2]. Presently, this technology finds extensive applications in diverse sectors such as healthcare, where models facilitate disease diagnosis [3]–[5] or neural activity decoding [6] via medical imaging (e.g., X-rays, MRIs, EEGs), recommendation systems in social media and streaming platforms that personalize content based on user [7], autonomous systems like self-driving vehicles [8], [9], predictive maintenance in industrial equipment within power plant [10], [11], and a myriad of other areas.

Numerous factors have contributed to the advancement of deep learning, with particular emphasis on the substantial increase in computational power. Notably, the proliferation of GPUs, capable of executing numerous tasks concurrently, has markedly expedited the training of machine learning models [12]. The advances in software, particularly in automatic differentiation, have significantly contributed to making deep learning both practical and efficient. For instance, by utilizing dynamic computational graphs (i.e graphs constructed in real-time during model execution), PyTorch [13] provides a differentiation engine that is well-suited not only for experimentation but also for efficient

deployment. Moreover, groundbreaking progress in neural network architecture has underpinned significant strides in domains such as computer vision, natural language processing, and reinforcement learning [14]. The exponential surge in data volume, largely propelled by the internet, has furnished an extensive and copious resource for training machine learning models [15].
Nonetheless, deep learning confronts several unresolved challenges :

— **Explainability of machine learning models** : Neural networks particularly are often regarded as "black boxes" due to their complexity, rendering the comprehension of their decision-making processes and outputs challenging [16].

— **Sustainability and carbon footprint [17]** : Additionally, the energy-intensive nature of training large deep learning models has raised concerns regarding sustainability, prompting the need for research into more energy-efficient algorithms and hardware.

— **Ethical considerations** : Ethical considerations also loom large, as machine learning models can perpetuate or amplify biases inherent in training data, potentially engendering unfair or unethical outcomes [18], [19]. Ensuring the **fairness and impartiality of AI systems** remains an ongoing imperative.

— **Data Limitations** : The data demands of deep learning models, which often necessitate substantial volumes of high-quality data for optimal performance, present a barrier to their applicability in fields where data is scarce or costly. Addressing this challenge involves the creation of models capable of **learning from a limited number of examples and effectively generalizing to unseen data**. The constraint of limited data is particularly prominent in real-life industrial scenarios. In healthcare, for instance, X-ray images of patients with rare conditions may be very limited. Gathering such medical data is costly due to privacy regulations, ethical concerns, and the inherent difficulty of collecting enough rare-case samples [20]. In industrial automation also, applications such as visual quality control, where the goal is to detect and categorize abnormalities

might require an expert to annotate hundreds of non-anomalous images to obtain a single annotation of one anomalous category [21].

The latter limitation highlights the efficiency challenges related to data and domain expertise. Learning with a small number of examples is referred to as Few-Shot Learning. It commonly relies on methods like meta-learning [22], transfer learning [23], and model adaptation [24] to attain successful generalization from restricted data quantities, often ranging from one to a few dozen instances per category. Transfer learning involves initially training a model on a large, general, and abundant dataset, then adapting the model to perform well on a more specific task with limited data. The core concept is that the pre-trained model, typically trained on a broad dataset such as ImageNet [25], learns general features (e.g., shapes and edges in images or language patterns in text) that are transferable across various tasks. These learned features can subsequently be fine-tuned or adapted to a target task using significantly fewer examples. In the first phase, models such as Convolutional Neural Networks [26] (CNNs) (e.g., ResNet [27]) or transformer-based models (e.g., ViT [23], GPT [28]) are trained on large-scale datasets to capture latent features from the input data. For instance, a CNN pre-trained on millions of labeled images from the ImageNet dataset [29] would learn lower-level visual features, like edges and textures, as well as higher-level structures, such as object shapes. When applied to a specific domain or task with limited data (a Few-Shot scenario), the pre-trained model undergoes fine-tuning. This process retains the original model weights and readjusts them to better align with the new training data, rather than training the model from scratch, which would likely result in overfitting (i.e. the model simply memorizing the few examples it is exposed to, rather than generalizing to unseen data). Overfitting is likely to arise when training large models on small dataset sizes. The pre-trained model thus serves as a robust foundation, requiring fewer updates. This approach not only enhances performance but also significantly reduces training time. However, in transfer learning, several challenges arise. Domain shift [30] occurs when a Few-Shot dataset originates from a domain that is significantly different from the source dataset, causing the features learned during pre-training to transfer poorly. For instance, a model pre-trained on ImageNet (which consists of natural images) may struggle to adapt effectively to medical images, such as

X-rays. The greater the domain shift, the more difficult it becomes for the model to reuse learned features effectively. Furthermore, many learning challenges are amplified due to the Few-Shot nature of the target dataset. Overfitting and hyperparameter sensitivity are more likely to occur with smaller datasets. Additionally, low-quality data, such as noisy or ambiguous examples, can have a disproportionately large impact [31]. Since the model is fine-tuned on very few examples, noisy labels can misguide the adaptation process, leading the model to learn incorrect patterns. *So, how can we optimize data acquisition and annotation processes to address these challenges? How might we select the most informative samples for domain experts to annotate? Additionally, are there other strategies for leveraging expert knowledge to facilitate learning, particularly with the emergence of large multi-modal models that integrate both text and vision?*

When considering the task of image classification, the quality and diversity of training samples is of prime importance. In the context of Few-Shot Learning however, the scarcity of such samples makes it even more critical. In this thesis, we investigate how the labeling process could be enriched to mitigate this limitation. We explore multiple strategies, such as choosing the best samples to query an expert with, or asking for more precise local information about the object of interest within the training image samples. **The central question addressed is that of the benefits of better querying the available expert compared to acquiring or labeling more data.**

1.2 Industrial Context

1.2.1 Schneider Electric

Schneider Electric is a multinational corporation at the forefront of energy management and industrial automation, operating in over 100 countries worldwide. The company's mission revolves around providing innovative solutions for energy efficiency, sustainability, and automation across various sectors. With a comprehensive portfolio of products, software, and services, Schneider Electric

caters to the needs of residential, commercial, industrial, and infrastructure segments. In the realm of energy management, Schneider Electric offers a wide range of solutions, including power distribution products such as circuit breakers, switchgear, and distribution panels. Schneider's digital energy management and automation system such as the EcoStruxure Power platform exemplify this. EcoStruxure is a platform that provides comprehensive solutions for power distribution, management, and control in diverse settings, from residential buildings to industrial facilities. It operates on three levels :

— **Connected Products** : This level includes smart devices and sensors that gather real-time data, enabling better monitoring and management of systems.

— **Edge Control** : This involves advanced control systems that process data locally, providing immediate insights and decision-making capabilities to optimize operations.

— **Apps** : These are software applications that analyze data and provide actionable insights, helping users make informed decisions to improve performance and efficiency.

It offers a catalog of "advisors" which group services under domains such as "Building" or "Energy". Schneider Electric's EcoStruxure Building Advisor offers a suite of building management services that optimize energy usage, monitor HVAC systems (Heating, Ventilation, and Air Conditioning systems), and improve overall building performance. Within the automation sector, Schneider Electric delivers industrial automation solutions encompassing programmable logic controllers (PLCs), human-machine interfaces (HMIs), and industrial software for process control and monitoring. The company also offers IoT-enabled solutions that empower industries to enhance operational efficiency, reduce downtime, and optimize production processes. Schneider Electric also specializes in critical infrastructure solutions for data centers, providing power and cooling systems, as well as management and automation software to ensure reliable and efficient operation. Furthermore, the company is committed to sustainability, offering solutions for renewable energy integration, energy-efficient products, and digital tools that enable customers to optimize energy consumption and reduce environmental impact. Schneider Electric's dedication to innovation, sustainability, and efficiency underscores its position as a global

leader in energy management and industrial automation.

1.2.2 AI in Schneider Electric

AI plays a pivotal role in Schneider Electric's operations by enhancing the efficiency, reliability, and sustainability of its products and services. The integration of AI technologies enables the company to leverage data, optimize energy usage, and drive predictive maintenance across various systems, ultimately delivering additional value to customers. Schneider Electric utilizes AI in a multitude of applications, including energy management, industrial automation, and digital services. In the context of energy management, AI is employed to analyze energy consumption patterns, predict demand, and optimize energy distribution in buildings, industrial facilities, and infrastructure. This empowers customers to make data-driven decisions, reduce energy costs, and minimize environmental impact. Within the realm of industrial automation, AI facilitates predictive maintenance by analyzing equipment performance data to forecast potential failures, thereby enabling proactive maintenance activities and minimizing downtime. Additionally, AI is used to optimize industrial processes and improve supply chain management. This can be done by analyzing demand forecasts, inventory levels, and production capacities, thereby optimizing inventory management and logistics to meet customer demand efficiently. The stakes for Schneider Electric in leveraging AI are significant. By harnessing AI capabilities, the company aims to deliver top-tier value to its customers through improved energy efficiency, reduced operational costs, and optimized performance of critical infrastructure.

1.3 Contributions Overview

In this section, we present the two primary contributions of this thesis. The first contribution involves the introduction of active learners to enhance the annotation process in Few-Shot scenarios, thereby improving performance. The second contribution explores methods to extract additional information from already annotated images. This is achieved by leveraging either human expertise or segmentation models to generate bounding boxes around objects of

interest. Utilizing this information helps mitigate ambiguity in learning when dealing with limited data.

1.3.1 Active Learning for Efficient Few-Shot Classification

As previously discussed, many real-world industrial applications are halted by data scarcity. Often, this scarcity is due to the high cost or slow pace of data annotation. While these challenges typically diminish as more data becomes available, the initial stages of the process can be crucial, depending on the nature of the problem.

For this contribution, we present a formulation of the "Active Few-Shot classification" problem, which addresses the challenge of classifying a small set of data samples without an explicitly provided training set, but rather under a highly restricted labeling budget.

Our methodology combines statistical inference with active selection criteria to identify the most informative samples for labeling. Initially, we train a feature extractor to transform raw images into high-dimensional feature vectors, aiming to enhance discriminability. Next, we apply various preprocessing techniques to these feature vectors, including normalization and graph smoothing. An Expectation-Maximization (EM) algorithm is then employed to infer a probability distribution over the samples. Finally, we propose a sequential active learning procedure that prioritizes samples with the highest confidence for each class, followed by those with the lowest confidence.

To evaluate our approach, we introduce benchmarks derived from the Transductive Few-Shot Classification (TFSC) domain. Specifically, we randomly sample “Few-Shot Tasks” from datasets such as mini-ImageNet and CUB.

Through these experiments, we demonstrate that our framework achieves superior accuracy compared to traditional TFSC solutions. Additionally, we show that our method is robust across various problem parameters, including the number of classes, the number of training samples, and class imbalance. Our primary objective is to showcase the potential for improved performance in data-scarce learning scenarios through this innovative approach.

1.3.2 Object-Centric Cropping for Visual Few-Shot Classification

This contribution explores strategies for enhancing model efficacy beyond merely acquiring new data or further annotating existing samples. Previously, we discussed the issue of ambiguity in transfer learning, particularly how it can hinder a model's ability to generalize effectively. Certain models possess intriguing adaptation and transfer capabilities; however, their performance can suffer when they focus on irrelevant features. For instance, a model trained to classify waterbirds and landbirds may incorporate background features, such as water or trees, into its decision-making process, potentially leading to misclassification of birds when they are in atypical habitats. Similarly, in industrial processes, if two classes are acquired during different times of day, background ambiguity might arise from the brightness levels in the images. Therefore, for optimal task adaptation, it is crucial to ensure that models focus on the relevant features of the object of interest.

To address this, we propose a methodology that leverages human expertise to refine the location of the object of interest within an image, improving training by eliminating confounding background information.

The choice of the method to obtain object localization varies depending on the application. For example, in scenarios where it is difficult to gather new data, manual annotation of bounding boxes in already annotated images may be more practical. In contrast, annotating or modifying existing data may be less beneficial in cases where acquiring new samples is possible.

We incorporate this refined information during the fine-tuning stage of the model through augmented training sets that focus on the object of interest. This approach ensures that the model adapts based on relevant features rather than being distracted by context.

We validate our methodology using well-known datasets, including ImageNet and CUB for classification, and Pascal-VOC for object detection. These experiments demonstrate the benefits of disambiguating objects from their backgrounds during training. Furthermore, we explore the advantages of inference-time disambiguation using automatic salient object detection which eliminates the need for human intervention.

Chapter 2

State-of-the-Art: Few-Shot with Images

2.1 General Context

The challenge of Few-Shot Image Classification is situated within the broader landscape of deep learning problems and builds heavily upon this field. A significant portion of this progress can be attributed to the development of architectures tailored for processing image data. Furthermore, the introduction of various datasets has not only facilitated standard benchmarks for Few-Shot Classification but has also provided a solid foundation for knowledge transfer and feature learning. In the following section, we will delve deeper into these pivotal aspects that form the groundwork for this work, in addition to exploring the various supervision paradigms that we explore within the question of leveraging expert knowledge to annotate data and curate training datasets.

2.1.1 Deep Learning And Evolution of Architectures

A typical deep neural network consists of several key components : an input layer, which receives raw data (e.g., images, text, or time series), followed by $L \geq 1$ hidden layers. Each hidden layer $l \in \{1, 2, \ldots, L\}$ is composed of neurons, each applying weights and biases to the outputs from the previous layer in what is called a "forward pass". The elemntal equation for a forward pass is as follows :

$$\mathbf{a}^{(l)} = f(\mathbf{z}^{(l)}) \tag{2.1}$$

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \tag{2.2}$$

Where $\mathbf{a}^{(l)}$ is the activation of the neurons in layer l, $\mathbf{z}^{(l)}$ is the linear combination of the activations from the previous layer and the weights for layer l, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector for layer l and $f(\cdot)$ is the activation function.

Activation functions are introduced after each layer to introduce non-linearity, enabling the model to capture intricate patterns. Common activation functions include the Rectified Linear Unit (ReLU), sigmoid, and tanh. A final output layer produces the network's prediction or classification. In classification tasks, the final layer typically employs a softmax function for multi-class problems or a sigmoid function for binary classification [32].

While equation 2.1 encapsulates the fundamental equation of a forward pass, contemporary neural network architectures exhibit significantly more intricate and sophisticated structures and inter-layer dynamics. These include, but are not limited to, dense connections, which ensure maximum information flow between layers; residual skip-layer connections, which help mitigate the vanishing gradient problem and enable the training of deeper networks; multi-input layers, which allow the integration of diverse data sources; and attention mechanisms, which enable the model to focus on the most relevant parts of the input data. Additionally, techniques such as batch normalization, which standardizes the inputs to each layer to improve training stability and speed; dropout, which randomly omits neurons during training to prevent overfitting; and convolutional layers, which are particularly effective for processing grid-like data such as images, further enhance the network's ability to learn and generalize from complex datasets. These innovations collectively enhance the robustness and versatility of modern neural networks. However, they often result in models with a very high number of learnable parameters, thereby increasing their data requirements. This further underscores the growing importance of frugal learning methods.

Perceptron and Backpropagation

The evolution of deep learning architectures reflects a continuous effort to overcome limitations of earlier models and improve performance. The progression has involved the development of increasingly sophisticated architectures, each building on the successes of previous designs. The earliest neural networks,

dating back to the 1950s and 1960s, were shallow perceptron models [33] that were very limited by their linear constraints. In the 1980s, the backpropagation algorithm was introduced [34], allowing these networks to learn through multiple hidden layers. This algorithm takes on the error between the model's output and the true value through a loss function, and computes gradients for all of the model's weights and biases to minimize this error. The error at the output layer L is given by :

$$\delta^{(L)} = \frac{\partial Loss}{\partial \mathbf{a}^{(L)}} \circ f'(\mathbf{z}^{(L)})$$

For any hidden layer l, the error is propagated back from the next layer $l+1$:

$$\delta^{(l)} = (\mathbf{W}^{(l+1)})^T \delta^{(l+1)} \circ f'(\mathbf{z}^{(l)})$$

And the gradient of the loss function with respect to the weights are

$$\frac{\partial Loss}{\partial \mathbf{b}^{(l)}} = \delta^{(l)} \qquad \frac{\partial Loss}{\partial \mathbf{W}^{(l)}} = \delta^{(l)} (\mathbf{a}^{(l-1)})^T$$

Where :

— $\mathbf{a}^{(l)}$ is the activation from the layer l.

— $\delta^{(l)}$ is the error at the l-th layer.

— $\mathbf{W}^{(l)}$ is the weight matrix of the layer l.

— $f'(\mathbf{z}^{(l)})$ is the derivative of the activation function at layer l.

— $Loss$ is the loss function that quantifies the error between the prediction and the output.

— $\circ$ is element-wise multiplication).

The computed gradients are subsequently utilized to minimize the loss function, in order to improve the accuracy of the model's predictions. Various optimization algorithms, such as gradient descent, are employed for this purpose. However, these optimizers generally provide approximate solutions and are prone to converging at local optima. This issue is particularly pronounced when working with small datasets, where the likelihood of encountering noisy gradient computations and getting trapped in local optima is significantly higher.

Convolutional Neural Networks

In the late 1990s and early 2000s, Convolutional Neural Networks (CNNs) revolutionized image processing. Introduced by Yann LeCun with the LeNet architecture [35], CNNs introduced the concept of convolutional layers that automatically extract spatial features from images. The convolution layer applies a series of filters to the input to produce feature maps. The filters are typically 2D matrices that slide over the input to compute dot products with the local regions of the input. The output of the convolution operation at layer l with K filters of size $M \times N$ is calculated as:

$$\mathbf{z}_{i,j,k}^{(l)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{c=1}^{C} \mathbf{W}_{m,n,c,k}^{(l)} \cdot \mathbf{a}_{i+m-1,j+n-1,c}^{(l-1)} + \mathbf{b}_{k}^{(l)}$$

Where:

— $\mathbf{z}_{i,j,k}^{(l)}$ is the output at position (i, j) for the k-th filter.

— $\mathbf{W}_{m,n,c,k}^{(l)}$ is the filter's weight.

— $\mathbf{b}_{k}^{(l)}$ is the bias term for the k-th filter.

— $M \times N$ is the filter size, and C is the number of input channels.

CNNs are designed to handle the high-dimensional nature of images by sharing weights across neurons, which dramatically reduces the number of parameters compared to fully connected networks. This structure, along with techniques like pooling and weight sharing, made CNNs highly effective for tasks such as image classification, object detection, and segmentation [36], [37].

Attention Mechanisms and Transformers

In 2015, attention mechanisms [38] emerged as a computationally efficient way of processing time series. They also demonstrated a better capacity for capturing long-range dependencies in sequences than other models that were popular for sequential data, namely Long Short-Term Memory (LSTM [39]) and Gated Recurrent Units (GRU [40]). The key innovation was allowing models to focus on relevant parts of the input sequence when making predictions. The attention mechanism computes a weighted sum of input features as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

Where :

— **Q** is the query matrix representing the input

— **K** is the key matrix is the reference for the comparison

— **V** is the value matrix, it represents the information distributed to the different parts of the sequence.

— d_k is the dimension of the keys (or queries).

It enables the network to selectively focus on important information, regardless of distance in the sequence. Attention was first introduced in the context of sequence-to-sequence (Seq2Seq) models for tasks like machine translation. This laid the groundwork for more sophisticated architectures.

In 2017, the Transformer architecture was introduced [41] and it marked a major shift in deep learning, particularly for natural language processing. Transformers rely entirely on attention mechanisms, dispensing with recurrence altogether, which significantly improved computational efficiency by allowing for parallel processing of sequences. Transformers use self-attention, where each element in the sequence attends to all other elements, capturing long-range dependencies more effectively. The Transformer architecture became the backbone of models like BERT [42] (Bidirectional Encoder Representations from Transformers) and GPT [28](Generative Pre-trained Transformer), which set new benchmarks in tasks like language translation, summarization, and question-answering [43].

Vision Transformers

Transformers, initially developed for text, were adapted for image processing with the advent of Vision Transformers (ViTs) [44]. Introduced in 2020, ViTs divide an image into patches and treat each patch as a token, similar to words in a sentence. Patches serve as a form of low-level feature extraction akin to the output of a convolutional kernel in a CNN, where local patterns are recognized and later assembled into more complex features. The size of the patches represents a trade-off between the semantic richness, locality, and computational cost. A smaller patch size provides more fine-grained local encoding but risks losing semantic information and increases computational complexity. Conversely, larger patch sizes lead to a coarser representation

due to the loss of fine-grained details. ViTs then use self-attention to capture relationships between patches. The image is turned into a sequence as follows :

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_1\mathbf{W}_E; \mathbf{x}_2\mathbf{W}_E; \ldots; \mathbf{x}_N\mathbf{W}_E] + \mathbf{E}_{\text{pos}}$$

Where :

— $\mathbf{x}_{\text{class}}$ is the class token

— $\mathbf{x}_i$ is the flattened i-th patch.

— $\mathbf{W}_E$ is the learnable embedding weight.

— $\mathbf{E}_{\text{pos}}$ is the positional embedding which is added to the input tokens to provide the model with information about the positions of the tokens.

ViTs have shown competitive performance compared to CNNs in various vision tasks, challenging the dominance of convolutional architectures in computer vision.

Large Foundational Models

The trend toward larger models with billions of parameters has accelerated, with models like GPT-4 [45], Dall-E [46] pushing the boundaries of what deep learning can achieve. These models are part of the broader category of foundation models, which are trained on massive datasets and then fine-tuned for specific tasks, making them highly versatile across domains.

2.1.2 Datasets and Data Acquisition: Manually Annotated to Large Data Engines

The evolution of data acquisition and data labeling has significantly impacted the development and performance of the very machine learning models that we rely on in our work. Particularly, in image classification and object detection, some datasets have become the gold standard for benchmarks.

Earlier image classification datasets such as MNIST [47] ,CIFAR-10 [48] and object detection dataset like Pascal-VOC [49] contain images that were meticulously curated and were annotated manually to ensure high-quality benchmarks. MNIST, introduced in 1998, is one of the earliest and most widely used datasets for handwritten digit classification, it includes 60000 training

images and 10000 test images of handwritten digits (0-9) collected from postal services. CIFAR-10, released in 2009 contains a total of 60000 images of 10 different classes. Pascal-VOC, made public in 2005 includes images annotated with object class labels, bounding boxes, and segmentation masks. ImageNet, introduced in 2009 and popularized by the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [29], is one of the most influential datasets in computer vision, particularly for image classification. The original dataset contains over 14 million images across more than 20,000 categories. For the ILSVRC competition, a subset comprising 1,000 categories with 1.2 million training images, 50,000 validation images, and 100,000 test images was used. ImageNet categories, referred to as synsets, are based on the WordNet lexical database, a large-scale structured network of words. Each synset represents a specific concept (e.g.,"do","table") and includes related synonyms. Images were gathered to match these objects and concepts as defined by WordNet. The synsets are hierarchically organized; for instance, within the "animal" synset, there are subcategories such as "dog" which is further divided into specific breeds like "Chihuahua". The large scale of the dataset was enabled by web scraping from popular image-hosting websites and annotated through a crowdsourcing platform. ImageNet has played a critical role in advancing deep learning, particularly following the success of deep learning models like AlexNet [50], which demonstrated groundbreaking performance in the 2012 ILSVRC.

OpenAI's CLIP [51] was also trained on a large and diverse, web-scraped dataset. But also has text supervision through textual descriptions paired with the images that come from the website such as meta-data or image captions. CLIP is notorious for its strong zero-shot capabilities.

However, more and more models rely on sophisticated data engines for curating or annotating their own training data. These models use a model-in-the-loop approach, where the model interacts with the data engine to select, generate or annotate data that improves training efficiency and performance. The Segment Anything Model (SAM) [52] utilizes a dedicated data engine to curate a large-scale, diverse dataset comprising over one billion segmentation masks. SAM is designed to function in a zero-shot setting, allowing it to segment objects that it has not encountered during training, without requiring

specific fine-tuning. The model's data engine plays a crucial role in refining the annotations of the training data, ensuring the learning of generalized object representations. In the initial stages of training, human annotators are more frequently involved, providing pixel-level annotations for the masks. However, as training progresses, annotators are primarily queried regarding missing masks or for refining existing ones. In the later stages, mask generation becomes fully automated, with the model producing the annotations independently.

Enhancing strategies for acquiring, annotating, and utilizing large-scale datasets, whether annotated or unlabeled, is pivotal to the advancement of next-generation models. The efficacy of machine learning fundamentally hinges on the quality and breadth of the data, underscoring its critical role in driving the development of more sophisticated and capable systems.

2.1.3 Few-Shot, Transfer Learning, and Foundational Models

Transforming raw image data into meaningful features that can be used in a machine learning model, is the foundational step in all of our contributions. To obtain such features, we rely on models that are widely used for knowledge transfer to solve Few-Shot tasks.

Few-Shot Learning refers to the ability of a model to learn and generalize from a very small number of examples, typically ranging from one to a few dozen instances per category. Achieving effective Few-Shot learning often involves leveraging techniques such as meta-learning [22], metric learning [53], transfer learning [23], and model adaptation [24]. Transfer Learning is dependent on an initial training on larger datasets. A model is first trained on a large, diverse dataset. For example, image models might be trained on datasets like ImageNet, while language models could be trained on extensive corpora of text from the web. This serves the model to learn generalizable features that capture fundamental patterns in the data. For images, this could include detecting shapes, textures, and edges; for text, it might involve understanding grammar, syntax, and semantic structures. The pre-trained model is then adapted to a specific task with few data. The general features learned during pre-training serve as a strong foundation.

Transfer Learning is closely related to the broader generalization challenge in Machine Learning. The pre-training step is crucial for ensuring general enough representations that can still be adaptable to new Few-Shot tasks, and achieving good representations remains a challenge.

Large Language Models (LLMs) such as GPT and BERT are designed as foundational models that can be adapted to a broad spectrum of tasks with minimal or no task-specific data. These models have demonstrated the efficacy of pre-training on vast corpora of text, enabling powerful transfer learning capabilities across numerous domains.

In the realm of multimodal learning, models like CLIP and DALL-E have extended transfer learning to multimodal data, specifically image-text pairs. CLIP leverages contrastive learning [51], training the model to maximize the similarity between correct image-text pairs while minimizing the similarity of mismatched pairs. DALL-E, on the other hand, focuses on generating images from corresponding text inputs [46]. Both models learn representations that are robust across different domains, facilitating complex tasks such as image captioning, cross-modal retrieval, and visual question answering.

Additionally, models such as SimCLR [54] and DINO [55] employ self-supervised learning techniques to pre-train on large-scale, unlabeled datasets. These models can then be fine-tuned on smaller, task-specific datasets, making them highly effective for transfer learning, particularly in scenarios where unlabeled data is plentiful, but labeled data is costly to obtain.

2.1.4 Active Learning: Model-in-the-loop

Active learning is a learning approach where the model can interactively, during the training, query a user or an oracle (a source of truth) to obtain the desired outputs for new data samples. A good active learner is one that allows the model to choose the most informative data to learn from, thereby improving its performance with less labeled data.

In active learning, we are often given a **learning budget** which refers to the limit set on the number of examples that the model can request for labeling.

The process typically begins by training an initial model using a small, labeled dataset. Subsequently, the trained model employs an active querying strategy to identify informative, unlabeled samples. These samples are then

labeled by human experts and incorporated into the training dataset for the next round of model retraining. This iterative process continues until the labeling budget is exhausted. Querying strategies may include uncertainty sampling such as least confidence sampling or entropy sampling where "uncertain" samples are selected based on metrics such as model confidence or prediction entropy. Alternatively, diversity sampling, which selects data points that are representative of the overall data distribution and diverse, with the aim of covering a broad range of examples can be used. Another category of querying strategies is querying-by-committee, wherein a committee of models votes on predictions, and data points with high disagreement among the committee members are chosen for labeling [56]–[58].

2.1.5 Supervision In Learning

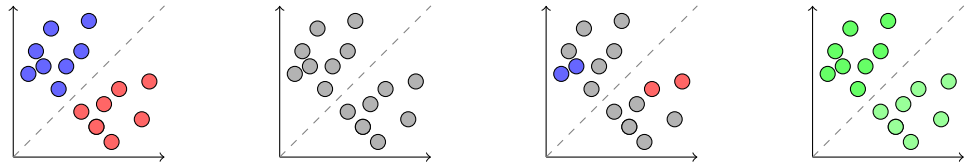
In the context of active learning or the comparison between different ways of querying an expert. It is of utmost importance to clarify the different supervision paradigms that pertain to learning.

Learning can be supervised 2.1a, unsupervised 2.1b, semi-supervised 2.1c or even self-supervised 2.1d. Each plays unique roles in training models and extracting useful features from images.

Supervised learning involves training a model on a labeled dataset, where each image in the training set is paired with a corresponding label (e.g., class name or object category). The model learns to map inputs (images) to outputs (labels) by minimizing the error between its predictions and the actual labels. This approach is common in tasks such a Image Classification and Object Detection [59].

Unsupervised learning involves training a model on unlabeled data, where no explicit labels are provided. The goal is to uncover underlying patterns or structures within the data. For images, this might include clustering where images are grouped into clusters based on similarities [60]. Similar types of images are grouped together without knowing the specific categories in advance. Another example is Dimensionality Reduction [61] : Reducing the number of features or dimensions of the data while retaining important information, useful for visualizing high-dimensional image data.

The degree of supervision in a given learning problem can however be more



(a) Supervised (b) Unsupervised (c) Semi Supervised (d) Self Supervised

Figure 2.1 – Illustration of the four different learning methods: supervised, unsupervised, semi-supervised, and self-supervised with two-dimensional data for a 2-class problem. The nodes represent data points, and the dotted line separates the two classes. The red and blue nodes represent data points with accessible label information, while the green nodes represent automatically generated supervision labels created for the purpose of self-supervision.

nuanced than this. Semi-supervised learning [61] combines a small amount of labeled data with a larger amount of unlabeled data during training. The model benefits from the labeled examples to learn a good representation and uses the unlabeled data to generalize better.

Self-supervised learning is a type of learning where the model generates its own supervisory signal from the data itself, without requiring external labels. It often involves creating auxiliary tasks that help the model learn useful representations from the data. In the context of images, this might include:

— Contrastive Learning [54]: Learning to distinguish between similar and dissimilar image pairs. For example, training a model to recognize that two different augmentations of the same image are similar, while different images are not.

— Pretext Tasks: Using tasks like predicting missing parts of an image [62] or solving jigsaw puzzles created from image patches to learn useful features [63].

2.2 State-of-the-Art: Few-Shot Classification

When it comes to image classification, CNNs depicted in Figure 2.2 and Vision Transformers are the two dominant architectural families, each with unique strengths. Advancements in CNNs primarily build upon the ResNet

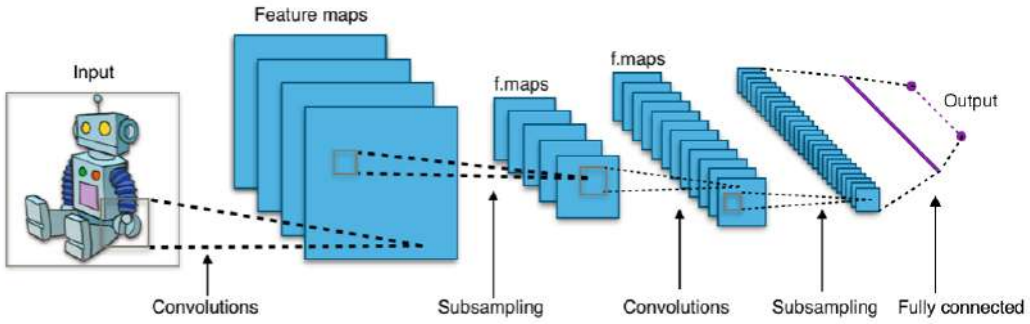


Figure 2.2 – Overview of a CNN architecture (Licensed under CC BY-SA 4.0) [64]

architectures. ResNet is characterized by its deep structure and its notable residual blocks, which incorporate skip connections to address vanishing gradient issues. The equation for a residual block in ResNet is:

$$y = x + F(x)$$

Where:

— x is the input to the residual block.

— F represents the block's forward layer

— y is the output of the residual block.

ResNet has become a go-to architecture for feature extraction in meta-learning approaches. Deep ResNets such as ResNet-12 and ResNet-18 proved effective for Few-Shot classifiers [65]. Newer versions such as ResNet-RS [66] improve ResNet's scaling efficiency for improved accuracy without substantially increasing computational requirements; however, they have not yet been thoroughly explored in Few-Shot learning settings. EfficientNet [67] has been a highly influential CNN architecture. Leveraging compound scaling, it effectively balances depth, width, and resolution to maximize efficiency. EfficientNet [67] has been shown to improve classification accuracy significantly in certain Few-Shot scenarios compared to other feature extractors.

In parallel, Vision Transformers have gained significant traction as feature extractors. Models like Data-efficient image Transformers (DeiT) [68] demonstrate competitive performance on image classification tasks while requiring fewer training data through knowledge distillation implemented with attention.

Such models display strong transfer capabilities.

These architectures are commonly employed in Few-Shot classification methods, which are often considered part of the meta-learning paradigm. Meta-learning is a framework designed to enable models to adapt quickly to new tasks with minimal data. Its primary objective is to train models across a diverse set of tasks so that they can generalize and perform effectively on unseen tasks with minimal fine-tuning. This learning process operates on two levels: meta-level learning, where the model learns across many tasks to capture general knowledge or strategies, and task-level learning, where the model quickly adapts to a specific task by leveraging the meta-learned knowledge. In the following sections, we elaborate on a widely-used meta-learning approach, as well as other metric-learning methods (i.e., models that learn a similarity measure or distance metric between data points) that are not strictly bound by the meta-learning framework.

2.2.1 Model-Agnostic Meta-Learning (MAML)

Model-agnostic approaches for Few-Shot Learning refer to frameworks and techniques designed to work with various model architectures, rather than being tied to a specific type of models. MAML [69] popularized the meta-learning framework where models are trained across a distribution of tasks rather than a single one. Learning problems are sampled from a distribution of tasks, and for each task, the model is fine-tuned. MAML proposes an optimization based meta-learning strategy where the fine-tuning occurs by simply taking few gradient steps to reduce the loss of the specific Few-Shot task, making it compatible with any gradient-based learner.

2.2.2 Prototypical Networks

Prototypical Networks (ProtoNets) are a popular metric-learning framework for Few-Shot learning. They were introduced to learn effective representations for images by creating a prototype (mean embedding) for each class based on the labeled data [70]. The prototypes are computed from samples as follows:

$$p_c = \frac{1}{N_c} \sum_{i=1}^{N_c} f(x_i) \tag{2.3}$$

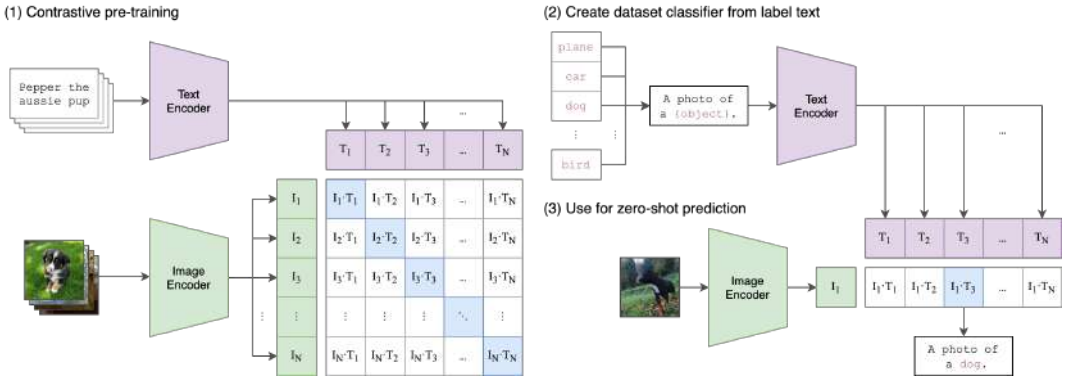


Figure 2.3 – Overview of CLIP usage for Zero-Shot image classification [51] (Licensed under CC BY 4.0)

Where:

— p_c is the prototype for class c.

— N_c s the number of samples in class c.

— f is the feature extractor.

— x_i is the i^{th} sample.

New instances are then classified based on their distances to these prototypes.

CNNs have traditionally been used for ProtoNets due to their effective feature extraction capabilities. ResNet and EfficientNet have been adapted to improve the performance of ProtoNets in Few-Shot Learning scenarios. Improved Prototypes [71] extend the traditional prototype approach by using an attention-analogous strategy that focuses on intra-class difference between samples to obtain more expressive weighted prototypes.

In a meta-learning framework, the feature extractor f can be learned across a diverse set of tasks. However, due to the availability of extensive vision datasets and recent breakthroughs in self-supervised learning, several pre-trained models are now available to serve as a "backbone" or feature extractor for such metric-learning methods.

2.2.3 Zero-Shot Classification with CLIP

CLIP [51] is a groundbreaking model developed by OpenAI that effectively bridges the gap between vision and language. It was designed to understand both images and texts by learning from a massive dataset of image-text pairs.

It was trained on a large dataset comprising 400 million image-text pairs sourced from the internet. This diverse dataset included a wide variety of images and their corresponding textual descriptions. CLIP consists of an image encoder (either a ResNet or a ViT) that processes images, and a transformer-based text encoder that processes text descriptions. Both encoders produce embeddings in a shared latent space, allowing the model to compare the similarity between images and texts. The capability of classifying images using only textual descriptions of the classes has paved the way for zero-shot classification. Figure 2.3 illustrates the training process of CLIP and its application in zero-shot classification.

CLIP's latent space is often compared to human-level understanding, as it can relate visual concepts and language in a manner analogous to how humans interpret visual stimuli and descriptions [51]. This capability positions CLIP as a "backbone" of choice for various Few-Shot learning methods.

2.2.4 Prompt Learning

In the context of vision-language models like CLIP. Prompt learning methods have emerged as innovative approaches for Few-Shot learning. These methods leverage the alignment of visual and textual inputs by designing specific textual prompts that are then used to solve the classification task. By framing the image classification task as natural language prompting task, the model can utilize its learned representations of the textual prompts to classify the images. For example, if the goal is to classify an image, the prompt might be a descriptive sentence such as "A photo of a [class name]." for each class. The model uses these prompts to infer the correct class based on the similarities between the prompts and the image.

Recent works such as Context Optimization (CoOp) [72] model a prompt's context words with learnable vectors without modifying the parameters of the rest of the model. This is very powerful as it shifts the Few-Shot adaptation paradigm from tuning an entire model to only learning a set of words embeddings, while leveraging the strong natural language comprehension of these models.

A multitude of works build upon this principle to learn more effective prompts. Prompts that discriminate better, that are more adaptable to different

tasks and that yield representations that are more robust to variations in image classes. MaPLe (Multi-modal Prompt Learning) [73], for example, achieves that by promoting strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions.

2.3 State-of-the-Art: Active Learning in Few-Shot

Active learning has emerged as a popular learning framework, with significant progress made in developing acquisition functions aimed at answering a key question : which samples should be selected for annotation to optimize learning? Traditionally, active learning has been considered a separate paradigm or a challenging learning scenario in which a predominantly unlabeled dataset and a limited labeling budget are provided. The objective is to annotate a subset of the unlabeled data such that training on this subset is as informative as possible compared to training on the entire dataset. As outlined in the previous section (Section 2.1.4), there are various strategies to implement active learning, such as uncertainty sampling, diversity sampling, and others [56]–[58]. But the challenge resides in adapting these strategies to transfer problems.

For example, [74] adapts active learning heuristics to convolutional neural networks (CNNs) in a batch setting, framing active learning as a core-set selection problem. Their approach seeks to select a subset of data points, called the core-set, such that a model trained on this subset performs competitively on the remaining data. The core-set is constructed via an approximate solution to the K-Center problem, which selects points that maximize the minimum distance between points in the CNN's feature space.

Additionally, several works have extended meta-learning principles to active learning. For instance, [75] proposes a meta-learning model that actively learns across tasks sampled from a task distribution, using supervised feedback to enhance its performance on new tasks drawn from a similar distribution. Their model solves each task by adaptively selecting samples for the labeled support set that are then used by a model that classifies test samples. Furthermore [31] demonstrates that active learning can also be beneficial during the fine-tuning step of transfer learning, where uncertainty sampling based on the pre-trained

model's confidence helps improve learning outcomes.

However, even with adaptations for learning with limited data, these approaches still face challenges in more extreme Few-Shot learning scenarios, which we address in the subsequent chapter.

2.4 State-of-the-Art: Few-Shot Object Detection and Image Segmentation

In early Few-Shot object detection (FSOD) approaches, models relied on transfer learning, where a pre-trained object detector, typically based on convolutional architectures, was fine-tuned on a limited set of labeled examples for novel classes. These models were commonly pre-trained on large-scale datasets such as COCO. Notable architectures included Faster R-CNN [36], which introduced the Region Proposal Network (RPN) to generate object proposals, and YOLO [76] (You Only Look Once). However, these approaches were highly susceptible to overfitting due to the small number of available training examples.

With the growing popularity of transformers in image classification, and the introduction of DETR (Detection Transformer) for object detection, transformer-based Few-Shot variants have also emerged. One such model is FS-DETR [77]. In FS-DETR, a few visual templates representing novel classes are provided as prompts during test time. These prompts are "stamped" with pseudo-class embeddings, which guide the model in detecting occurrences of new object classes without requiring any additional fine-tuning, thus enhancing its generalization capabilities.

The Segment Anything Model (SAM) also utilizes a Vision Transformer architecture for its image encoder, paired with a mask decoder that efficiently generates segmentation masks based on provided prompts. SAM is designed to respond to different types of prompts, such as :

— Points : A user clicks on an object, and SAM segments the object in the image.

— Bounding Boxes : A box around the object prompts SAM to segment the object within the defined region.

— Text Prompts: through coupling SAM with a separate Vision-Language model, SAM can allow for prompting with text.

SAM was trained on one of the largest segmentation datasets ever created, comprising over 1 billion segmentation masks across more than 11 million images from a variety of domains. This extensive dataset has endowed SAM with robust capabilities for numerous low-data tasks, including Zero-Shot instance segmentation and Few-Shot object detection.

Although object detection and instance segmentation are not central to the primary research questions addressed in this thesis, we incorporate SAM as a tool to facilitate human annotation in Chapter 4. Specifically, we employ SAM to allow human agents to annotate images more efficiently by interacting with the model through clicks, reducing the time and effort required for manual labeling. This methodology enhances the overall annotation process.

Chapter 3

Active Labeling For Efficient Few-Shot

3.1 Introduction

We present the Active Few-Shot Classification (AFSC) problem, aiming to classify a small, initially unlabeled dataset within a highly constrained labeling budget. This problem differs from typical Transductive Few-Shot Classification where an explicitly provided training set is given instead of a single set with a highly restricted labeling budget. This scenario commonly arises in real-world industrial applications, particularly in cases where data samples are available but labeling them is costly or needs to be minimized, such as cold-start problems that demand accelerated operations exceeding the pace of data acquisition. While these issues eventually become solvable with more traditional techniques, the initial stages can hold significant relevance depending on the nature of the problem. We focus here on visual inputs, although the underlying rationale can extend to other domains. Self-Supervised Learning (SSL) has recently emerged as a promising solution, demonstrating the capacity to achieve high accuracy with a minimal number of labeled samples. However, these methods necessitate substantial amounts of unlabeled data, with the most prevalent approaches considering datasets as extensive as ImageNet (1.2 million training samples) to attain peak performance.[78]. When the quantity of data samples is more limited, typically not surpassing a few hundred unlabeled examples, the domain of Transductive Few-Shot Classification (TFSC) encompasses numerous approaches capable of achieving high accuracy under similar constraints. The underlying principle involves leveraging a comprehensive generic dataset to train efficient feature extractors, with the expectation that these extractors can yield meaningful features relevant to the specific Few-Shot task at hand.

Within the context of TFSC benchmarks, the labeled samples for the Few-Shot task are typically assumed to be uniformly drawn at random.

The challenge of selecting which samples to label in a classification task is not novel. The field of Active Learning has put forth numerous solutions to address this issue [79], [80]. Yet, the adaptation of active learning methodologies to the low-data regime, particularly within the domain of vision, remains relatively underexplored, and we believe it to be relevant to various real-world applications.

In this chapter, we present a formulation of the aforementioned problem as "Active Few-Shot Classification" (AFSC). We introduce a methodology that combines statistical inference with well-founded heuristics, and subsequently introduce benchmarks tailored from the TFSC domain. Through these benchmarks, we illustrate the capability of the proposed framework to achieve heightened levels of accuracy compared to traditional TFSC solutions. Our primary impetus is to demonstrate the potential for attaining superior performance levels by approaching the data-scarce learning problem from this innovative perspective.

The main findings of this contribution have been presented in conference format in the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023 https://ieeexplore.ieee.org/document/10095220) [81] and have been submitted for a patent by Schneider Electric.

.

The key elements of this chapter are:

— A novel formulation of the Active Few-Shot Classification problem, demonstrating the potential to achieve remarkably high accuracy in contrast to traditional methods.

— A theoretically grounded two-tier methodology, involving the inference of sample distribution among classes and the strategic selection of samples to optimize learning outcomes.

— Vision benchmarks adapted from Transductive Few-Shot Classification problems for conducting experiments.

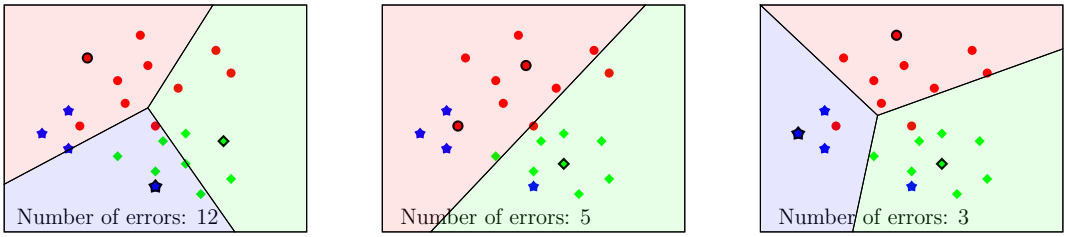


Figure 3.1 – Illustration of the interest of the Active Few-Shot Classification formulation as compared to classical Transductive Few-Shot Classification. In the three rectangles, we depict the distribution of samples in three possible classes, represented as colors. On the left figure, the labeled samples, marked in bold, have been randomly drawn. On the center and on the right, the distribution of samples in each class have been estimated using an Expectation Maximization method. On the center, we select the samples that are the closest to the estimated Gaussian centers, whereas on the right we use the proposed criterion. Below the figures are indicated the performance of a simple nearest-neighbor classifier based on the labeled samples. We also depict the decision boundaries with the straight lines.

3.2 Related Work

Few-Shot Learning: The proliferation of methods necessitating pre-training the model on more general tasks, led to the development of a myriad of efficiently trained feature extractors with data augmentation techniques [82], ensembling [83] and various other techniques. Classification in the inductive setting has predominantly relied on simple methods[83] whereas more diverse methods were used in the transductive setting allowing for more promising results [84]. The authors in [85] have also challenged class balance priors in this setting and emphasized its significance in recent works. We align, in this chapter, with the reconsideration of the transductive configuration towards a more class and label distribution agnostic approach.

Supervision in Learning: Our work also pertains to self-supervised learning (SSL) and unsupervised learning techniques [86], [87], as the cold-start aspect, i.e., commencing from zero examples, is crucial for us, necessitating the use of unsupervised methods until labeled data becomes available. Nonetheless, we do not consider SSL as a primary solution in this work because it typically depends on substantial amounts of unlabeled data. Existing literature indicates that the information embedded in the self-supervised training is strongly

bounded by the number of examples utilized [88].

Active learning: A fundamental aspect of this research centers on the substantial score improvement achieved through the utilization of Active Learning as a tool, as well as an investigation into the heuristics we employ for this selection process. Furthermore, Active Learning is widely viewed as a promising solution for "task ambiguity" which occurs when training data inadequately specifies the desired output for all potential inputs, resulting in the learner entertaining numerous plausible yet inaccurate hypotheses. It is suggested that an Active Learner can address task ambiguity by efficiently querying samples that effectively eliminate most of these plausible hypotheses [31]. Nonetheless, much of the existing literature on Active Learning is unsuitable for Few-Shot scenarios, as the majority of techniques with strong theoretical foundations heavily depend on the models employed. These techniques typically aim to reduce the learner's variance or base the selection on the model's uncertainty [89]: These techniques fail in a Cold-start setting [90] where the process begins with zero labeled examples. The scarcity of samples and/or labeling budget exacerbates the challenge, as there is no room for an initial phase of irrelevant labelings based on an unstable model. However, recent studies have demonstrated that pre-trained models can address this challenge in Natural Language Processing and Vision problems [31], [91]. In this work, we also utilize feature extractors to achieve improved data distributions, enabling the integration of diverse sampling methods and uncertainty-based sampling methods for an effective labeling pipeline [92].

3.3 Problem Statement

3.3.1 Transductive Few-Shot Learning

In Transductive Few-Shot Classification, we are given a dataset $\mathcal{Q} := \{\mathbf{x}_i^q, y_i^q\}_{i=1}^{n_q}$ made of n_q samples arbitrarily distributed among K classes. x_i^q are the samples and y_i^q their respective labels. We call $\mathcal{Q}$ *the query set*. We are also given a set $\mathcal{S} := \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n_s}$ of samples from the same K classes, $\mathcal{S}$ is named the *support set*. The support set contains very few samples, typically as low as one sample per class. The task is to classify the samples from the query

set all at once (as opposed to an Inductive setting where samples are classified independently of one another) while having access to the label information from the support set.

Furthermore, we suppose access a larger dataset $D_b = (x_i^b, y_i^b)$ that we call *base dataset*. This dataset comprises K_b classes that are disjoint from the K classes considered for the task.

This base dataset is used to train a generic feature extractor f_θ that yields representations $z_i = f_\theta(x_i)$ of the samples in order to facilitate the classification task [93]. Such a feature extractor is derived by removing the classification head from a neural network previously trained on a different task. This problem is typically referred to as K-ways n_q-queries n_s-shots problem.

3.3.2 Active Few-Shot Learning

For the Active Few-Shot Classification task, we consider a similar context to the Transductive Few-Shot task. In this task however, instead of having a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, we have a single set $\mathcal{N} := \{\mathbf{x}_i, y_i\}_{i=1}^N$ and a labeling budget $\ell \leq N$. We call this set the *samples set*, and we use *samples* to refer to the elements of this set henceforth.

The task is to classify the samples after revealing the classes of, at most, ℓ of them. We call these selected samples *labels*. We consider $\mathcal{L} := \{\tilde{\mathbf{x}}_i, \tilde{y}_i\}_{i=1}^\ell$ the set of labels. For commodity, we refer to this problem as a K-way, ℓ-labels, N-samples task.

At the commencement of the task, $\mathcal{L} = \emptyset$, and following the label selection the task is equivalent to a Transductive Few-Shot Classification task where the support set is $\mathcal{L}$ and the query set is $\mathcal{N} \backslash \mathcal{L}$. The primary distinction lies in the selective choice of labeled samples rather than random drawing, resulting in a varying number of labels per class, possibly including zero for certain classes in the worst-case scenario.

3.4 Methodology

The proposed methodology encompasses multiple sequential steps. Initially, it involves training a feature extractor to convert raw images into high-dimensional feature vectors, with the aim of enhancing discriminability. Subsequently, various preprocessing procedures are applied to the obtained feature vectors, encompassing normalization and graph smoothing. Following this, an Expectation-Maximization (EM) methodology is utilized to infer a probability distribution across all samples. Lastly, a sequential active learning procedure is proposed, prioritizing samples with the highest confidence for each class, followed by those with the lowest confidence. The subsequent paragraphs expound upon each of these steps accompanied by the presentation of pseudo-code in Algorithm 1.

3.4.1 Feature Extraction

The development of a feature extractor capable of generating meaningful representations for samples beyond its initially trained class domain constitutes a significant challenge in the realm of machine learning, as expounded in Section 3.2. This task is particularly prevalent in Few-Shot learning, where sophisticated methodologies can markedly enhance feature quality. For instance, the classical ResNet-12 architecture, initially introduced in [65], achieved an average accuracy of 57% in 5-way 1-shot scenarios using the mini-ImageNet dataset, whereas, with certain adaptations, it could surpass 70% accuracy, as noted by the authors in [83]. (e.g., prolonged training with a cosine annealing scheduler, integration of two heads with self-supervised rotations, mixup, data augmentation during testing, and model ensembling). Given that training an effective feature extractor is a customary practice in Few-Shot learning literature, we exclusively employ off-the-shelf models for all experiments in this study, except in the case of cross-domain experiments, where we modify an existing methodology to align with the META-DATASET benchmark [94].

3.4.2 Preprocessing

Following [95], we perform a two-step normalization, where feature vectors are first centered (the mean of the resulting feature vectors is 0), and then projected onto the unit-sphere (the norm of each feature vector becomes 1) [82]. We denote the resulting feature vector $\mathbf{x}$ and the feature vectors matrix $\mathbf{X}$.

Graph Smoothing: To exploit the information contained in the unlabeled samples, we propose to adapt the methodology described by [96]. Initially, we construct a graph wherein the vertices represent the samples, and the edges are contingent on the cosine similarity between the corresponding feature vectors. Specifically, we retain, for each vertex, its m most similar neighbors, yielding a binary graph with an adjacency matrix denoted as $\mathbf{A}$.We then compute $\mathbf{M} = (\beta\mathbf{I} + \mathbf{A})^{\kappa}$, where β and κ are hyperparameters. We transform the feature vectors $\mathbf{X}$ into $\mathbf{Z} = \mathbf{MX}$. This operation results in the blending of each coordinate of a feature vector with the same coordinate in similar vectors, leading to a smoothing operation [97], with the benefit of lowering the impact of outliers.

3.4.3 Statistical Inference

Let us denote $\mathbf{z}_i$ the i-th preprocessed feature vector corresponding to the i-th sample in $\mathcal{N}$. We consider a simple model where, for each class k the distribution of the corresponding $\mathbf{z}_{/y=k}$ follows a multivariate Gaussian with mean μ_k and variance $\mathbf{\Sigma}_k$. If we consider an isotropic distribution, $\mathbf{\Sigma}_k$ becomes $\sigma_k^2\mathbf{I}$ where $\sigma_k \in \mathbb{R}^+$ and $\mathbf{I}$ is the identity matrix. This model has been widely used in the literature, either explicitly or implicitly [82], [83].

The probability density function for this model would be :

$$p(z_{/y=k}) = \frac{1}{(2\pi\sigma_k)^{\frac{d}{2}}} \exp(-\frac{1}{2\sigma_k^2}||z - \mu_k||^2)$$

Under this hypothesis, a probabilistically optimal classifier is achievable given good estimations of the means and variances. On the other hand, we can obtain estimations by assigning examples to clusters. Algorithms that rely on Expectation Maximization (EM) allow us to find local optima of data partitioning. We focus on the soft K-means in this work for two reasons; the

first one being that soft K-means methods do not utilize, any prior about the marginal label distributions. This property makes it robust to samples presenting class imbalance. The second reason pertains to the unsupervised nature of the algorithm, as we can compute a locally optimal clustering with zero labels, albeit of suboptimal quality. We then obtain an estimate of $p(\mathbf{z}_{/y=k})$.

3.4.4 Proposed Active Selection

After the statistical inference, we aim to quantify the potential increase in performance that data points would yield if they were to be labeled. Classical solutions rely on heuristics [98] that include:

— **Margin**: The margin criterion, for any $\mathbf{z}_i$, is the difference between the highest probability (i.e probability of the assigned cluster) and the second best one. The sample with the highest margin is considered the most confident.

— **K-medoid**: The medoid is the cluster element that is closest to its centroid. The sample with the lowest distance to its cluster centroid is the one with the highest confidence.

Instead of using these classical approaches, we propose a new criterion based on the ratio of log densities of the statisticmeans distributions. We call this method **Log-probs Soft K-means Sampling (LSS)**. Namely, we consider that the sample that has the lowest **log-probability ratio (*lpr*)** of belonging to its cluster k rather than other clusters is the most confident one.

The logarithm of the density of $\mathbf{z}_{/y=k}$ is

$$\log p(\mathbf{z}_{/y=k}) = -d \log\left(\sqrt{2\pi}\sigma_k\right) - \frac{1}{2{\sigma_k}^2}\|\mathbf{z} - \mu_k\|^2$$

where d is the dimension of the vector z. We define the lpr as the ratio of the log-densities over the different label-features distributions:

$$lpr_k(\mathbf{z}) = \frac{\log p(\mathbf{z}_{/y=k})}{\sum_{i=1}^{K} \log p(\mathbf{z}_{/y=i})}$$

For $a_k = -d \log\left(\sqrt{2\pi}\sigma_k\right)$ and $b_k = {}^1\!/_{2\sigma_k^2}$ the log-probability ratio of sample

$\mathbf{z}$ for cluster k is:

$$lpr_k(\mathbf{z}) = \frac{a_k - b_k||\mathbf{z} - \mu_k||^2}{\sum_{i=1}^{K} (a_i - b_i||\mathbf{z} - \mu_i||^2)}$$

We estimate the mean and the standard deviation of the class-features distribution using the clustering

$$\hat{\mu_k} = \frac{\sum_{\mathbf{z} \in c_k} \mathbf{z}}{|c_k|} \text{ and } \hat{\sigma_k} = \sqrt{\frac{\sum_{\mathbf{z} \in c_k} (\mathbf{z} - \hat{\mu_k})^2}{|c_k|}}.$$

By utilizing the logarithms of the densities rather than the densities themselves, the calculations are rendered more robust in terms of estimations. For the sake of simplicity, we assume a constant standard deviation for all the classes $\sigma = 1/\sqrt{2\pi}$ so that the ratio becomes:

$$lpr_k(\mathbf{z}) = \frac{||\mathbf{z} - \mu_k||^2}{\sum_{i=1}^{K} ||\mathbf{z} - \mu_i||^2}.$$

A log-probability ratio close to 0 indicates strong confidence that the sample belongs to the cluster, whereas a ratio close to 1 indicates high uncertainty.

For a given the criterion, we compute the value for each sample and determine which ones to label based on this value [99]. We adopt a naive sequential pipeline for the active selection. This pipeline comprises several *rounds*. The first round consists in selecting K samples to label (one per cluster) that yield the highest confidence according to the chosen criterion (e.g., *lpr* close to 0). In subsequent rounds, irrespective of the chosen criterion, we label samples with the lowest confidence (e.g., *lpr* close to 1). This two-tier strategy aims to ensure the initial labels strongly represent the class with minimal risk of being outliers, followed by refining statistical inference using these additional priors and choosing labels representing the highest uncertainty. Samples demonstrating strong confidence are likely to be correctly predicted, thus obviating the need for further manual labeling.

Since no labels are available for the first round, we rely on multiple random intializations for the soft K-means. They will each generate a different clustering $\pi_1, \pi_2, ...\pi_n$. We then use the clustering that minimizes the sum of the distances of the samples to their closest centroids.

For subsequent rounds, a soft K-means is run with centroid initializations corresponding to the means of the labels for each class, i.e., the centroid of the i^{th} cluster is initialized as the mean of all the labels from class i (or arbitrarily initialized as zero if no labels are available for class i). This approach yields higher-quality clustering for subsequent rounds to select the next K labels, continuing until the final round where the final cluster assignments are used to assign classes for the samples; each cluster is assigned the class of the labels used to create its initial centroid. In the following section, a comparison of our criterion with classical ones in terms of average accuracy is presented.

Algorithm 1 LSS for label selection and classification

procedure AFSC(f_θ, X, l, K): ▷ X :the raw features of the N samples from $\mathcal{N}$
▷ ℓ : labeling budget, K : number of classes, f_θ: feature extractor

numRounds = ℓ/K
$\mathcal{L} = \emptyset$ ▷ $\mathcal{L}$ is the set of label examples
$Z = f_\theta(X)$ ▷ Z are the extracted features of the samples
$Z = preprocess(Z)$

▷ first selection round
randIds = *randomIdces*(K from N) ▷ random K from the N samples to init clusters
M_z = Z[randIds]
P,cents = *softKmeans*(Z, M_z) ▷ P and lpr of shape (NxK)
$lpr = zeros(N, K)$

$$lpr[n,k] = \frac{||Z[n]-cents[k]||^2}{\sum_{i=1}^{K} ||Z[n]-cents[i]||^2}$$

select sample with min($lpr[:,k]$) for each k and append to $\mathcal{L}$
▷ Subsequent rounds
for $i \leftarrow 1$ to numRounds−1 **do**
M_z = *classMeansFromLabel*($\mathcal{L}$)
P,cents = *softKmeans*(Z, M_z)
$lpr = zeros(N, K)$

$$lpr[n,k] = \frac{||Z[n]-cents[k]||^2}{\sum_{i=1}^{K} ||Z[n]-cents[i]||^2}$$

select sample with max($lpr[:,k]$) for each k and append to $\mathcal{L}$
▷ Final Inference for class assignment
finalClassMeans = *classMeansFromLabel*($\mathcal{L}$)
P,cents = *softKmeans*($Z, finalClassMeans$)
classAssignments = $argmax(P, dim = 2)$
return classAssignments

3.5 Experiments

3.5.1 Datasets

We adapt four standard benchmarks to AFSC: mini-ImageNet, tiered-ImageNet, CUB and FC-100. The details are as follows:

mini-ImageNet: a subset of ImageNet [25] that comprises 60,000 color images of size 84×84 pixels equally distributed over 100 classes. In all experiments, we use the standard split of 64 base-training, 16 validation and 20 test classes [95]. **tiered-ImageNet**: another larger subset of ImageNet that contains 608 classes and 779,165 color images of size 84×84 pixels. We use a split of 351 base-training, 97 validation and 160 test classes [95]. **CUB**: this benchmark contains 11788 images of size 84×84 pixels, from 200 classes. For CUB, we use a standard split of 100 base-training, 50 validation and 50 test classes, presented in [65]. **FC-100**: this benchmark is a subset of CIFAR-100 that is used for Few-Shot Learning. it contains 600 images of size 32×32 for each of its 100 classes. We use a standard split of 60 base-training, 20 validation and 20 test classes [100].

We also adapt the META-DATASET benchmark. Which is a comprehensive benchmark by [94] designed to assess the performance of machine learning algorithms in cross-domain and Few-Shot learning scenarios. META-DATASET consists of different vision benchmark datasets from different domains with various data distributions. This benchmark allows for a more realistic and robust evaluation of Few-Shot solutions through generating cross-domain Few-Shot tasks that are not restricted to a specific number of classes or of samples.

3.5.2 Benchmarking Protocol

Our default settings are 5-ways ℓ-labels, where $\ell = 5, 25$, with a sample set of size $75 + \ell$. Following [85], the classes of the sample sets are randomly distributed following Dirichlet's distribution with parameters $\alpha\mathbf{1}$, where $\mathbf{1}$ is the constant unit vector with K coordinates and $\alpha = 2$.

To generate a K-ways, ℓ-labels, N-samples Few-Shot task, we randomly sample K classes from the test classes of the considered dataset (classes that were not used for any feature extractor training). Then, we distribute the N

samples over the K classes following a Dirichlet's distribution.

Unless otherwise specified, the methods are assessed by calculating the average score across 10,000 tasks [84]. We opted to employ weighted accuracy, which represents the average of recalls for each class, as unweighted accuracy is not well-suited for highly imbalanced labels. Additionally, we elected to compute this accuracy across all samples, including the labeled ones, for which the prediction is constrained to the revealed value.

In our experimental setup, we utilize two distinct feature extractors. The first is a ResNet-18 model trained with standard cross-entropy minimization on the base classes and label smoothing for 90 epochs, with a batch size of 256. The initial learning rate is set at 0.1 and is reduced by a factor of 10 at epochs 45 and 66 [84], [85]. We use this feature extractor in Tables 3.1 and 3.2 to compare to state-of-the-art methods in the classical TFSC Setting (N-ways, K-shots, Q-queries [70], [84]). The second feature extractor is the Ensembled Augmented Shots Y-shaped ResNet-12 proposed by [83], comprising a concatenation of three ResNet-12 architectures, each with approximately 12 million parameters. The training is conducted over 5 cycles of 100 epochs each, employing a cosine-annealing scheduler to decrease the learning rate from ν_0 to 0 at the end of each cycle, with ν_0 initialized at 0.1 and decreased by 10% for each cycle. This feature extractor enables achieving top-tier performance on classical TFSC problems, hence it is utilized for the remaining experiments.

A grid search approach is employed to determine the hyperparameter values on the validation set of mini-ImageNet, and these same hyperparameters are maintained throughout all the experiments.

3.5.3 Advantages of AFSC Compared to State-of-The-Art TFSC Methods

In the first experiment, the aim is to quantify the benefit of using Active Few-Shot Classification (AFSC) in comparison to the traditional Transductive Few-Shot Classification (TFSC) approach. A challenge in comparing raw performance arises from the fact that TFSC considers a fixed number of labels per class, ensuring more equitable representation of the various classes in the task. The performance of TFSC methods is presented in Tables 3.1 and 3.2,

		mini-ImageNet			
Shots configuration	**method**	5-labels	25-labels	50-labels	100-labels
Equal labels per class (random)	PT-MAP ([82])	62.6	75.3	81.3	70.4
	LaplacianShot ([101])	67.6	86.2	90.5	94
	BD-CSPN ([102])	69.0	85.1	89.7	93.4
	TIM ([84])	69.3	84.5	89.4	93.2
	α-TIM ([85])	69.4	86.9	91.6	94.8
	soft K-means	70.7 ± 0.28	83.4 ± 0.16	87.7 ± 0.12	91.8 ± 0.09
Random label sampling	**soft K-means**	64.3 ± 0.29	84.7 ± 0.14	89.6 ± 0.09	93.3 ± 0.06
Active label sampling	**LSS (ours)**	73.5 ± 0.21	87.7 ± 0.16	93.8 ± 0.11	96.8 ± 0.07
		tiered-ImageNet			
		5-labels	25-labels	50-labels	100-labels
Equal labels per class (random)	PT-MAP [82]	63.6	77.5	83.1	88.6
	LaplacianShot ([101])	74.0	89.3	92.8	95.3
	BD-CSPN ([102])	75.8	88.7	92.0	94.8
	TIM ([84])	75.8	88.1	91.2	94.6
	α-TIM ([85])	76	89.9	93.6	96.1
	soft K-means	74.7 ± 0.29	84.8 ± 0.18	88.8 ± 0.14	92.6 ± 0.10
Random label sampling	**soft K-means**	66.9 ± 0.30	85.7 ± 0.16	90.0 ± 0.11	93.5 ± 0.07
Active label sampling	**LSS (ours)**	76.1 ± 0.24	87.5 ± 0.17	92.2 ± 0.13	95.8 ± 0.09

Table 3.1 – Comparison of accuracy between state-of-the-art TFSC methods and the proposed AFSC one on mini-ImageNet and tiered-ImageNet. The same feature extractor (ResNet-18 pre-trained on the training classes of tiered-ImageNet) are used in all experiments. We also indicate the performance of the statistical inference method (soft K-means) combined with the proposed preprocessing for completeness. We also include, for the accuracies that we computed, the 95% confidence interval. All values are percentages.

Table 3.2 – Same comparison as in Table 3.1 on the CUB dataset. We only consider the 5-labels scenario for this dataset that only has 60 examples per class: for a Dirichlet distribution of the class labels and for sample size of 100 or more, we observe frequent occurrences of class distributions that largely surpass 60 examples for a given class. All values are percentages. We use the ResNet-12 pre-trained on the training classes of mini-ImageNet for this experiment.

Shots configuration	**method**	5-labels
Equal labels per class (random)	PT-MAP [82]	67.3
	LaplacianShot ([101])	84.7
	BD-CSPN ([102])	76.1
	TIM ([84])	76.4
	α-TIM ([85])	77.2
	soft K-means	75.9 ± 0.28
Random sampling	**soft K-means**	66.6 ± 0.29
Active sampling	**LSS (ours)**	77.2 ± 0.21

in comparison with AFSC. To address the additional challenge of potentially missing labels in some classes, the results of the soft K-means statistical inference method are reported under the same conditions as the aforementioned results (referred to as "Equal labels per class"). This entails using the same feature extractor, benchmark, and evaluation method. The performance is then compared when selecting labels at random versus employing the proposed active method. For better standardized results with [85], the average unweighted accuracy is reported here instead of the weighted accuracy. Additionally, the values from [85] have been adjusted to reflect the accuracy measurement that accounts for labeled samples.

The results indicate that, across all datasets, a substantial improvement in accuracy is achieved when opting for active label selection as opposed to random selection. Tables 3.1 and 3.2 distinctly demonstrate that under comparable conditions, our active sampling enables the attainment of superior scores in contrast to the scenario where equally distributed labels per class are randomly pre-selected, a configuration in which all benchmarked methods were tested. Additionally, it is observed that as the number of labels increases, the disparity in scores between random and active selection diminishes.

One reason for the reduced gap is the way accuracy is computed in our problem formulation, as we include labeled examples in the calculation of weighted accuracy. Additionally, the gap is further reduced due to the higher probability of selecting meaningful samples at random from a larger pool of labeled data compared to a smaller one. An intriguing observation in Table 3.1 is that the Random label sampling configuration catches up to the Equal labels per class configuration as the number of labels increases and eventually surpasses it, underscoring how drawing uniform labels from an uneven distribution could detrimentally impact accuracy.

3.5.4 Cross-Domain Active Few-Shot Classification

Recent studies [31] have demonstrated that the performance of learning a Few-Shot task may be compromised depending on the pre-trained model and the datasets utilized for pre-training when employing active selection. In this section, we aim to evaluate our method on various Few-Shot tasks from diverse domains, employing a feature extractor exclusively trained on ImageNet.

Table 3.3 – Evaluation of the Cross-Domain transfer robustness of AFSC. We pre-train the ResNet-12 architecture following [83] on the training set of ImageNet. We then evaluate the model on 5-ways 5-labels 80-samples Few-Shot runs from the test sets of each of the datasets from META-DATASET. We use the same datasets splits and model training configuration described in [94] and we follow the preprocessing and soft K-means inference described in the proposed methodology. We compare random selection to active selection with the LSS criterion. We also include, for the accuracies that we computed, the 95% confidence interval over 600 runs. All values are percentages.

Dataset	**LSS Active Selection**	**Random Selection**
ImageNet	52.65 ± 1.22	42.87 ± 1.22
Omniglot	66.41 ± 1.88	54.48 ± 1.66
Aircraft	42.29 ± 0.91	34.42 ± 0.92
Birds	67.48 ± 1.18	54.47 ± 1.24
Textures	46.39 ± 1.08	39.60 ± 1.00
Fungi	57.82 ± 1.14	47.87 ± 1.25
VGG Flower	82.63 ± 1.07	67.63 ± 1.38
Traffic Signs	52.65 ± 1.02	46.10 ± 1.12
MSCOCO	42.30 ± 0.97	35.95 ± 0.96

Table 3.3 illustrates that active selection consistently yields superior performance compared to random selection across all datasets in the META-DATASET benchmark. The observed enhancements in weighted accuracy range from 6.35% (MSCOCO) to 15.00% (VGG Flower). We posit that the magnitude of these improvements varies based on factors such as the task complexity of the dataset's Few-Shot tasks, the similarity of the dataset's domain to ImageNet, and the Gaussianity of the class label distributions in the Few-Shot tasks. Nevertheless, these noteworthy performance gains underscore the considerable benefits and effectiveness of integrating Active Learning in Few-Shot Classification, even in cross-domain scenarios.

3.5.5 Selection Criteria Comparison

We further explore the impact of the active criterion in Table 3.5. For each of the four datasets, we compare the performance of the sampling criteria with that of random selection, as well as with an oracle in the 5-labels case.

The oracle is computed by applying Nearest Class Mean to all possible combinations of 5 labels from the total 80 samples. Notably, the oracle is computed based on 1000 random runs for mini-ImageNet and 100 runs for FC-100, tiered-ImageNet, and CUB, due to computational constraints arising from the $\binom{80}{5} = 24,040,016$ combinations. This underscores the potential performance achievable, particularly considering that we already achieve commendable scores comparable to the state-of-the-art standard configuration through simple heuristics. Our findings reveal that the margin and K-medoid criteria significantly outperform random selection, while the proposed LSS method marginally outperforms these criteria. Additionally, Table 3.4 in the presents a more in-depth comparison of active selection criteria encompassing two variations of the core-set approach [103] applied to different transductive Few-Shot methods. While the LSS yields the best results in the extremely low labeled data regime (1 label per class), the core-set methods rapidly scale and slightly outperform it in Few-Shot scenarios with higher allowed labels (more than 5 labels per class).

Table 3.4 – **More state-of-the-art transductive Few-Shot learning methods and active selection criteria**: Comparison of state-of-the-art active selection acquisition methods and state-of-the-art transductive Few-Shot learning methods for tiered-ImageNet using the ResNet-18 feature extractor pre-trained on the base classes of tiered-ImageNet. Results for random, margin and LSS for Soft K-means were reused from Table 1 and the remaining results were obtained from 100 runs. We include the 95% confidence intervals.

Inference method	Active sampling method	5-labels	25-labels	100-labels
soft K-means	random	62.2 ± 0.31	86.8 ± 0.18	95.1 ± 0.63
	entropy	62.9 ± 2.05	78.8 ± 2.10	79.7 ± 2.01
	variation ratios	63.5 ± 2.31	86.5 ± 1.65	90.3 ± 1.87
	margin	69.0 ± 0.30	88.6 ± 0.19	95.5 ± 1.64
	core-set	62.2 ± 0.31	91.2 ± 1.53	97.9 ± 0.47
	core-set (mahalanobis)	62.2 ± 0.31	91.5 ± 1.34	97.7 ± 0.42
	LSS	76.1 ± 0.26	90.9 ± 0.16	96.4 ± 0.82
Protonet	random	57.1 ± 2.35	71.5 ± 2.18	89.1 ± 1.02
	entropy	59.2 ± 2.08	72.1 ± 1.96	83.7 ± 1.60
	variation ratios	57.6 ± 2.22	79.3 ± 2.15	91.3 ± 0.81
	margin	56.8 ± 2.41	80.2 ± 2.04	90.9 ± 1.20
	core-set	57.1 ± 2.35	83.9 ± 1.62	92.3 ± 0.72
	core-set (mahalanobis)	57.1 ± 2.35	83.2 ± 1.73	92.6 ± 0.72
PT-MAP	random	54.8 ± 2.26	74.2 ± 1.84	93.6 ± 0.77
	entropy	55.3 ± 2.07	71.2 ± 2.21	86.7 ± 7.41
	variation ratios	57.9 ± 2.34	80.6 ± 2.11	84.4 ± 6.88
	margin	64.9 ± 2.93	86.9 ± 2.62	94.6 ± 4.66
	core-set	54.8 ± 2.26	88.4 ± 4.62	97.0 ± 0.97
	core-set (mahalanobis)	54.8 ± 2.26	88.1 ± 5.28	98.0 ± 0.87

labels ℓ	**sampling strategy**	mini-ImageNet	tiered-ImageNet	CUB	FC-100
5 labels	random	62.2 ± 0.31	61.2 ± 0.30	67.7 ± 0.33	46.5 ± 0.26
	oracle	95.2 ± 0.21	97.4 ± 0.3	95.5 ± 0.18	80.2 ± 1.40
	margin	69.0 ± 0.30	69.0 ± 0.26	78.5 ± 0.30	48.7 ± 0.26
	K-medoid	75.6 ± 0.28	75.2 ± 0.27	83.8 ± 0.25	52.1 ± 0.25
	LSS (ours)	**76.1 ± 0.26**	**75.7 ± 0.27**	**84.2 ± 0.24**	**53.3 ± 0.24**
25 labels	random	86.8 ± 0.18	85.6 ± 0.20	-	68.8 ± 0.20
	margin	88.6 ± 0.19	87.0 ± 0.18	-	69.7 ± 0.21
	K-medoid	89.7 ± 0.17	88.6 ± 0.16	-	71.2 ± 0.19
	LSS (ours)	**90.9 ± 0.16**	**89.3 ± 0.19**	-	**71.7 ± 0.20**

Table 3.5 – Comparison of active selection criteria on all four considered datasets. We add an oracle strategy which consists of the selection that yields the best performance. We also chose not to include results for 25-labels for CUB for the same reasons mentioned in Table 3.2. We use the ResNet-12 pre-trained on the training classes of mini-ImageNet for this experiment. All values are percentages.

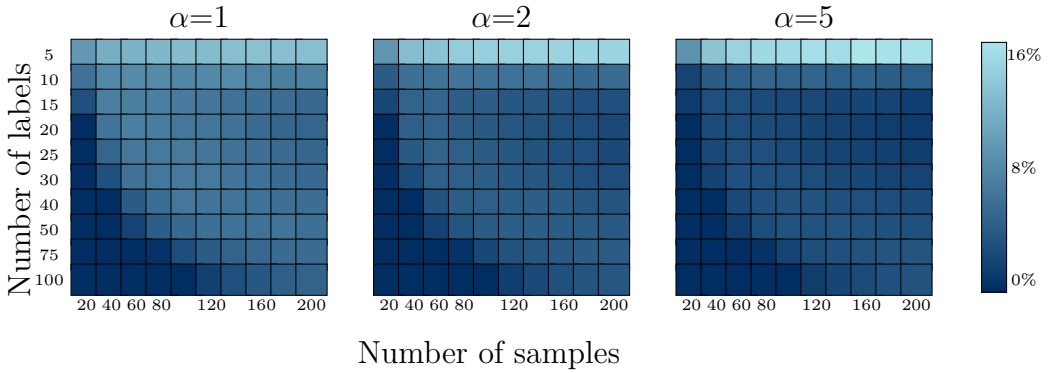


Figure 3.2 – This figure illustrates the effect of the problem settings (labeling budget, samples set, and imbalance severity) on the effectiveness of the active selection with the LSS criterion compared to a random selection. We report the difference in percentages of weighted accuracy between active and random sampling in the form of a heatmap. The heatmaps from left to right represent, respectively, highest to lowest imbalance severity (we vary the α of the Dirichlet distribution). The Y axis denotes the number of labels and the X axis the total number of samples. Note that for cases where the total number of samples is lower than that of the labels, we report a 0 difference in the two approaches. We use the ResNet-12 pre-trained on the training classes of mini-ImageNet and conduct this experiment on the test classes of mini-ImageNet

3.5.6 Active Learning Vs Problem Settings

In Figure 3.2, we challenge the effectiveness of active sampling over random sampling in different settings. We vary the total number of samples, the number of labels, and finally the imbalance severity through the α parameter of the Dirichlet distribution. We observe higher but smoother differences in the high severity cases, i.e., a smaller range of values. Whereas in the softer imbalance case, we notice more clearly that active selection yields significantly better results for low labeling budgets and slightly better results for higher labeling budgets. This is the same effect as the one we briefly mentioned for Table 3.1: the active versus random gap closes as the labeling budget covers a larger proportion of the samples. We conclude that regardless of whether we have a more acute imbalance ($\alpha = 1$) or a softer one ($\alpha = 5$), the boost in performance remains consequent. We further investigate how class balance affects these findings for different datasets in the last part of section 3.5.9.

3.5.7 On the Choice of Log Densities Ratio

In 3.4.4, we explained that we chose to use the ratio of the logarithm densities instead of the ratio of densities as a selection criterion, despite the latter being more commonly encountered in the literature. This is because the former is more robust to the estimations of the distributions that we compute with the help of the soft K-means clustering. To support this claim, we conduct a simple experiment where we synthetically generate Few-Shot runs with a two-dimensional multivariate Gaussian distribution. We observe in Figure 3.3 that the estimated confidence based on the empirically inferred mean and standard deviation, depicted on the bottom line, is closer to the truth (top row) when using our proposed criterion.

We also performed a quantitative study of this phenomenon where we drew samples from the same distribution 10000 times and computed the mean and 95% confidence interval of the absolute difference between the estimated confidence and the true confidence of every sample of the first class for both criteria. We found that, for the ratio of log densities, the mean of these differences is 0.36 ± 0.0015. Whereas for the ratio of densities the mean of 0.41 ± 0.0023. This represents a small but statistically significant difference. We

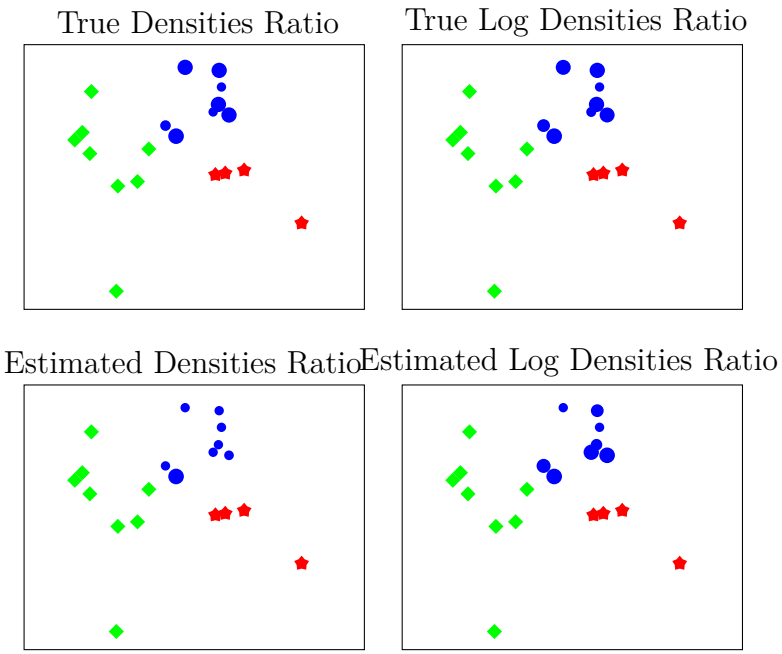


Figure 3.3 – Comparison of robustness of ratio of logarithmic densities (right) and ratio of densities (left). We generate 20 points from 3 multivariate gaussian of respective means (2,0) ,(-2,0) and (0,2) and respective standard deviations $2I$,$4I$ and $0.3I$ where I is the Identity matrix. We then cluster the points using a K-means algorithm and compute the confidence of the blue points of belonging to their class using to the two critera. On the top, we use the true mean and standard deviation values of the distributions whereas on the bottom we use estimates of these values. We scaled both criteria to fit in the [0,1] range and made the size of the blue points proportional to their confidence.

also suspect that this robustness gap amplifies in higher dimensional examples as the log density ratio worked significantly better for the experiments in this chapter.

3.5.8 Exploration of other cluster initialization methods for tasks with unknown number of classes

In this section, we discuss the limitation associated with our method, which necessitates knowledge of the number of classes. This constraint becomes apparent after the initial round when we commence using the labels produced by our algorithm. These labels are utilized to calculate class means for each of the K-clusters representing the K classes to which they belong. Classes

Table 3.6 – Comparison of performance for a known number of classes (using our standard LSS) versus unknown number of classes with a different clustering technique. For unknown number of classes we use DP-means with varying thresholds λ. We compute results for a variable number of labels per run where we allow for as many labels as clusters created. The mean number of labels over multiple runs is also reported in this table. These results were computed over 100 runs and the 95% confidence intervals are also reported. We use the ResNet-12 pre-trained on the training classes of mini-ImageNet, and used the test classes of mini-ImageNet for this experiment.

Initialization	3-ways	5-ways	7-ways	10-ways
Known number of classes				
standard LSS	87.3 ± 2.78 ($\ell = 3$)	75.7 ± 2.77 ($\ell = 5$)	67.4 ± 2.56 ($\ell = 7$)	60.1 ± 1.68 ($\ell = 10$)
Unknown number of classes				
Infinite Mixture ($\lambda = 3.7$)	80.7 ± 3.00 ($\ell = 5$)	65.7 ± 2.45 ($\ell = 6$)	53.7 ± 1.72 ($\ell = 7$)	43.5 ± 1.51 ($\ell = 7$)
Infinite Mixture ($\lambda = 3.5$)	88.0 ± 2.10 ($\ell = 9$)	73.3 ± 2.31 ($\ell = 10$)	63.5 ± 1.83 ($\ell = 11$)	53.7 ± 1.74 ($\ell = 12$)
Infinite Mixture ($\lambda = 3.3$)	91.2 ± 2.03 ($\ell = 14$)	82.6 ± 2.06 ($\ell = 18$)	76.2 ± 1.45 ($\ell = 19$)	68.7 ± 1.66 ($\ell = 22$)

for which no labels are disclosed are initialized randomly. Here, we employ DP-means as described in [104], a deterministic clustering algorithm derived through Bayesian nonparametrics for the Dirichlet process. DP-means operates by iterating over the data points and computing the minimum distance of each point to all existing cluster means. If this distance exceeds a threshold λ, a new cluster is initialized with the point as its mean. DP-means clustering enables us to overcome both of these limitations. We leverage this approach to achieve an adaptive number of clusters where each class can be represented by several clusters. In this context, our focus lies in selecting only the highest confidence example (using the LSS criterion) from the clusters, and subsequently classifying the entire cluster based on this highly confident label.

In Table 3.6, we compare the outcomes of our standard methodology in the scenario of a known number of classes with the scenario of an unknown number of classes, where we initialize clusters with DP-means using different threshold values. For the unknown number of classes scenario, we carry out our labeling in a single high confidence round, where we label one element for each cluster. This results in a varying number of labels per run, as DP-means can yield a variable number of clusters depending on the sample geometry. The results indicate that for an equal number of labels, we generally observe improved performance when the number of target classes is known.

3.5.9 Ablation Study

Finally, we investigate the effect of graph smoothing (Section 3.4.2) on our method's efficacy. Additionally, we aim to validate the continued relevance of active selection methods when the label distribution is balanced (i.e., uniform) in the samples, as certain studies in various domains have suggested that active methods may be less effective in the context of balanced sets[91].

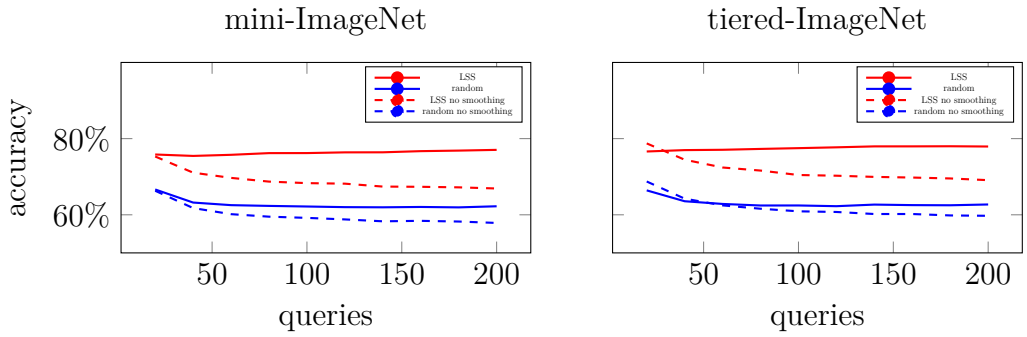


Figure 3.4 – This figure showcases the impact of the smoothing in the performance of random and active selection on mini-ImageNet and tiered-Imagenet. The red lines correspond the LSS active method and the blue lines to the random selection. We use continuous lines for when graph smoothing is used and dashed lines for when graph smoothing is not used. We use the ResNet-12 pre-trained on the training classes of mini-ImageNet for this experiment.

Graph smoothing: When applying graph smoothing to the data, a notable disparity in the performance of soft K-means between the random and active settings becomes apparent. This section aims to investigate how the improvement from random to active is influenced when smoothing is not applied. Figure 3.4 depicts the progression of the weighted accuracies as the number of samples is varied, with the objective of capturing the effect of smoothing on different graph sizes. Similar effects are observed in both mini-ImageNet and tiered-Imagenet. It is evident that the smoothing process distinctly enhances the performance of the active selection, particularly for larger sample pools. Furthermore, it is observed that it amplifies the performance gap between the random and active selection across all sample sizes.

Data Imbalance: We conduct our experiments in a balanced class setting on mini-ImageNet, tiered-ImageNet, and FC-100. Table 3.7 presents the results

labels ℓ number	**sampling strategy**	mini-Imagenet	tiered-Imagenet	FC-100
5 labels	random	67.4 ± 0.30	64.3 ± 0.32	49.0 ± 0.26
	LSS	**82.9 ± 0.21**	**81.3 ± 0.23**	**56.9 ± 0.23**
	Gap (LSS-random)	15.5	17	7.9
	Gap diff (with imbalanced)	**-2.7**	**-3.3**	**-1.1**
25 labels	random	90.1 ± 0.11	88.6 ± 0.16	71.3 ± 0.17
	LSS	**92.4 ± 0.12**	**90.2 ± 0.16**	**72.9 ± 0.18**
	Gap (LSS-random)	2.3	1.6	1.6
	Gap diff (with imbalanced)	**1.5**	**2.1**	**1**

Table 3.7 – Active sampling versus random sampling for balanced samples set scenarios. We consider the 5 and 25 labels cases for a total sample set of size 80 and 100 respectively. This table also includes a row for the Gap (LSS-Random) in this balanced setting, and for the gap differences between the gap in the balanced classes and the gap in the imbalanced classes with α=2: Gap diff = Gap imbalanced - Gap balanced. We use the ResNet-12 pre-trained on the training classes of mini-ImageNet for this experiment. All values are percentages.

for random sampling and active sampling using our best criterion in the balanced class scenario. In comparison to the imbalanced setting, we observe the expected performance improvement for all methods across all datasets. Particularly noteworthy is the superior performance of the active sampling over random selection in both the 5-label and 25-label cases, with margins comparable to those in the balanced scenario (Table 3.5). A more substantial disparity is evident in the 5-label setting, indicating that active sampling yields greater accuracy benefits when classes are balanced. Conversely, in the 25-label scenario, the gaps between active and random sampling are smaller and less pronounced than those in the imbalanced scenario. This further underscores the overall superior effectiveness of active sampling over its random counterpart, irrespective of class label distribution.

3.6 Limitations

While we have obtained highly promising results that validate the effectiveness of adopting active selection methods, it is crucial to acknowledge that this proposed framework warrants further exploration and scrutiny in various aspects. One notable aspect is our utilization of the class mean of the labels in the statistical inference for the soft K-means initialization. We suspect

that this decision could have intriguing implications for the performance of different sampling criteria due to the potential loss of information stemming from the averaging operation. Additionally, our assumption of a known and given number of classes, that we challenge in Section 3.5.8, may be unrealistic in certain applications.

3.7 AFSC in industrial use-cases

The previously described method was adapted for several industrial applications and compared to alternative methodologies. A recurring challenge was adapting the method to applications at varying stages of development. For example, some applications were in the initial stages with no annotated data available, while others had already undergone one or multiple rounds of annotation. In one noteworthy use case—described here in limited detail to protect data privacy—this method demonstrated significant potential for improving annotation efficiency.

This use case involved anomaly detection and classification using ultrasound imaging of a product's surface to identify various anomaly types. While data acquisition in such contexts is relatively straightforward, annotation is more complex, as human experts are required to review each image individually. Additionally, there was a severe imbalance in the "normal" to "abnormal" image ratio. During the initial annotation round, only a limited number of anomalous images, if any, were labeled. The LSS criterion proved valuable in selecting a subset of unannotated data that was less severely imbalanced and included diverse, ambiguous examples of both normal and abnormal samples.

3.8 Conclusion

We have introduced a novel formulation for the Active Few-Shot Classification (AFSC) problem, positioning it as a competitive alternative to Transductive Few-Shot Classification (TFSC). Our approach employs a two-step methodology, combining a statistical inference method with an active sampling process. In this process, samples are selected for labeling based on Log-probs Soft K-Means Sampling. By adapting benchmarks from the TFSC domain, we

demonstrate that the proposed framework achieves superior accuracy compared to TFSC. We tested the methodology in various problem settings and across multiple domains, finding that it is particularly well-suited for handling class imbalance and for cross-domain adaptation. We anticipate that AFSC will be relevant to a broad spectrum of real-world applications, establishing a promising foundation for future advancements.

Chapter 4

Object-Centric Cropping for Visual Few-Shot Classification

4.1 Introduction

Recent advancements in Few-Shot adaptation, whether through model fine-tuning or more sophisticated methods, have empowered models to rapidly learn from small datasets [65], [83]. Several large models pretrained on diverse image datasets, such as CLIP or DINO [51], [105], are now accessible and deliver competitive results in classification benchmarks. Despite these models possessing intriguing adaptation and transfer capabilities, they may encounter challenges in correctly generalizing to datasets with task ambiguity [31] especially when operating with as little as one example per class. The presence of image ambiguities stemming from multiple objects or complex backgrounds can significantly deteriorate performance.

For instance, a model trained to classify waterbirds and landbirds may factor in the background when making decisions, potentially leading to misclassification of birds not in their typical habitats. Hence, for optimal task adaptation, it is crucial to ensure that the models learn to classify images based on relevant features.

Recall that in visual Few-Shot classification, the aim is to correctly classify objects in images using only a very limited number of examples, called *shots*, for each class. In this chapter, we show that it is possible to mitigate ambiguity by cropping out the exact object of interest from the shots. We consider a manual and semi-manual ways of acquiring this local position: a human annotator either provides the ground truth crop or solely indicates a pixel from the object,

which is then used to prompt the Segment Anything Model (SAM) [52] to automatically obtain a segmentation mask of the object. We also consider a third way which consists in leveraging a salient object detection model to retrieve a mask on the object of interest. Figure 4.1 depicts these three modes.

We aim to measure how much improvement can be achieved in classification accuracy by leveraging this local information, and to compare it with using additional shots instead and we demonstrate this performance boost across established benchmarks. A main difficulty is to identify a suitable methodology to leverage the location of objects of interest in the shots: as a matter of fact, our research shows that focusing on the crops containing the objects of interest during learning can introduce a bias in the generalization, making it less effective. Instead, we introduce a method that integrates the crops and a portion of their context during learning, leading to consistent improvements across our experiments. In this chapter, we:

— Present a methodology for extracting the object of interest from a given training set to improve training

— Compare multiple methods for acquiring and utilizing this information with differing precision and levels of human involvement in both inductive and transductive settings, across three datasets.

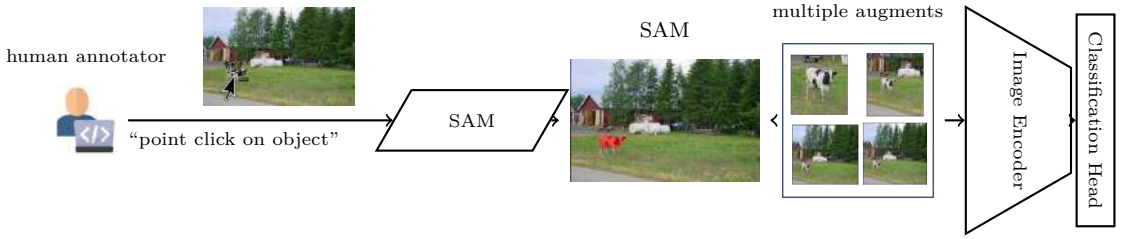
— Investigate the benefits of inference time disambiguation through automatic salient object detection.

— Examine the effects of removing the context around the object of interest of a specific class on the class distribution.

This work has been submitted for a patent by Schneider Electric.



(a) Ground truth bounding box disambiguation.



(b) SAM-based point click disambiguation.



(c) Salient object disambiguation.

Figure 4.1 – The illustration depicts the three approaches for determining the local position of the object of interest within an image. In the first scenario (a), an expert directly provides a bounding box that is then used to augment the image once. Conversely, in the second scenario (b), the expert selects a point on the object, which is subsequently employed by SAM to produce a mask. This mask is then utilized to create multiple augmentations of the image. In the third scenario (c), a mask of the relevant object is automatically generated with the help of a salient object detection model. The mask is then utilized to create multiple augmentations of the image similarly as in (b). The original image and its augmentations are processed through an image encoder then fed to a linear layer for training.

4.2 Related Work

Few-Shot Learning: As mentioned in previous chapters. Few-Shot approaches often require pre-training on more general tasks, benefiting from advancements in feature extraction, data augmentation, ensembling, and other techniques. While inductive classification has relies on simpler methods, diverse techniques have been employed in the transductive setting such as soft K-means pseudolabeling [81], [83], which we utilize in our methodology.

Task Ambiguity: The issue of task ambiguity emerges as a recurring theme in the transfer learning literature. Many studies addressing task adaptation introduce various forms of ambiguity that may arise in the new task targeted for model adaptation. Specifically, the work by [106] sets out to develop a meta-learning approach capable of addressing the ambiguity inherent in learning from limited data. Furthermore, [31] delineates task ambiguity within the context adopted in this chapter, particularly in relation to extraneous objects and/or background elements coexisting within the same image. while these works provide meaningful solutions to learn despite the ambiguity, they have not considered altering the original image to mitigate it.

Object Detection And Segmentation: Segment Anything Model [52] has marked a significant milestone as the inaugural large-scale segmentation model, showcasing robust zero-shot capabilities in classification and object detection tasks. Notably, this model has the capacity to swiftly generate segmentation masks of pertinent objects in response to diverse prompts, including points, bounding boxes, or masks. Faster and more efficient iterations of this model have been developed, further enhancing its appeal [107]. There are also notable works exploring the concept of identifying objects in images without human annotation. Notably, [108] achieves compelling results in Salient Object Detection without any supervised training. These works open the door to a more automated approach for acquiring information about the precise location of an object of interest.

4.3 Problem Formulation

We consider the inductive and the transductive scenarios of Few-Shot learning in vision tasks which are two prevalent settings in the Few-Shot learning literature [83], [85], [109]. In these two cases, we evaluate the performance gains that can be achieved through the additional information that is the local positioning of an object of interest within the image.

4.3.1 Inductive Setting

In inductive Few-Shot classification, we are given a dataset $\mathcal{S} := \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n_s}$ consisting of n_s samples uniformly distributed among k classes. Here, $\mathbf{x}_i^s$ represents the samples and y_i^s represents their respective labels. We refer to $\mathcal{S}$ as "the support set." The support set contains very few samples, typically as low as one sample per class. Additionally, we are given a test set $\mathcal{T} := \{\mathbf{x}_i^t, y_i^t\}_{i=1}^{n_t}$ comprising n_t samples also uniformly distributed among the K classes. The task involves accurately predicting the labels of test samples independently from each other while having access to the label information from the support set.

4.3.2 Transductive Setting

In transductive Few-Shot classification, we are given, in addition to the support set and the test set defined in Section 4.3.1, a third set $\mathcal{Q} := \{\mathbf{x}_i^q, y_i^q\}_{i=1}^{n_q}$ consisting of n_q samples uniformly distributed among K classes. Here, $\mathbf{x}_i^q$ represents the samples and y_i^q represents their respective labels that are unknown. We refer to $\mathcal{Q}$ as "the query set." The query set serves as an unlabeled dataset that can provide additional information about the class distribution. This information can be leveraged by a number of semi-supervised techniques [83], [110]. In a transductive setting, the typical goal is to infer the classes of the query set all at once. However, in this context, we consider the performance on the separate test set mentioned earlier to account for better task generalization and allow for a better comparison with the inductive setting. We have chosen to explore these two settings as they can present varying degrees of difficulty for locating the object of interest in relation to annotating an additional example.

In the transductive setting, annotating a sample can be as simple as moving it from the query set to the support set, whereas in the inductive setting, labeling an additional sample requires further data acquisition which can be quite expensive in certain applications.

4.4 Obtaining the Location of the Object of Interest in the Image

The goal is to differentiate between different types of physical objects (e.g., cat versus dog). However, models are usually trained on images containing multiple objects at once, leading to ambiguous representations and making generalization challenging. The significance of this correlation is likely to be discounted given a large dataset. In a one-shot setting, however, a spurious object may be considered with equal importance as the object of interest, thereby rendering additional information about the position of the object of interest within the image useful to disambiguate.

We consider three different methods of acquiring the object's location with varying degrees of human involvement. Each method has its advantages and drawbacks, but we believe that they can all be most compelling in the appropriate settings. For example, while fully manual annotation of bounding boxes can incur additional costs and may seem contradictory to a Few-Shot setting, there are several application cases where the Few-Shot interest lies in the inability to curate new data samples, as opposed to annotating or modifying existing ones, which can be more accessible.

4.4.1 Fully Manual Human Annotation

A direct approach to acquiring the object's location information is soliciting input from a human annotator. This approach involves querying the annotator for the most compact bounding box that entirely encompasses the object of interest. While this methodology is manual, it excels in delivering high-caliber bounding boxes, ensuring the comprehensive coverage of the object. While requiring such annotations in large classification datasets can be costly, we advocate that in many Few-Shot settings, obtaining bounding boxes of objects

of interest might be less demanding. Moreover, in the inductive setting, this additional cost is justifiable as it reduces that of acquiring new data for the considered task.

4.4.2 Using Segment Anything Model

This second method for obtaining location information, as discussed in this study, relies on sophisticated segmentation models. Specifically, leveraging the Segment Anything Model (SAM) [52] enables the generation of segmentation for all objects within a given image. Integrating this model into our semi-manual approach involves soliciting a human annotator to pinpoint the object of interest within the image. Subsequently, we utilize SAM to generate a mask from this point. SAM can be prompted with a point in the image through its pixel coordinates to produce a binary mask of an object containing that point. The mask is then converted into a bounding box. This method minimizes human involvement, requiring the annotator to simply identify a point rather than delineate a bounding box. However, it leads to additional computational costs primarily associated with SAM's image encoder and risks producing lower quality bounding boxes if an erroneous mask is generated.

4.4.3 Automatic Salient Object Detection

In this fully automatic mode, we rely on unsupervised object segmentation model MOVE [108] to generate foreground object masks. MOVE leverages the fact that shifting foreground objects relative to their initial positions results in new, realistic images. It uses a combination of image inpainting and adversarial training to teach a model how to generate accurate foreground object masks. We utilize this model, trained in a fully unsupervised manner, to generate masks for the object of interest. While this method eliminates human involvement altogether, it also leads to additional computational costs associated with multiple image encoders. Furthermore, it has the major drawback of being unable to differentiate between the relevant object of interest and other objects that happen to be in the foreground of the image, unlike the other two methods.

4.5 Utilizing the Location of the Object of Interest in the Image

To mitigate the issue of ambiguity stemming from extraneous objects in the images used for classification tasks, our approach involves training a model on augmented versions of the images, with a specific focus on the object of interest. Leveraging transfer learning from large pre-trained models designed for different tasks, which is a common practice in Few-Shot learning contexts [82], forms the basis of our strategy. Our methodology revolves around Few-Shot runs or episodes, each comprising w classes, s annotated training samples, and t test samples. In the transductive scenario, we also have access to q unlabeled samples. For simplicity, we consider class-balanced pools of labeled, unlabeled, and test samples.

4.5.1 Data Augmentation With Crops

The initial phase involves augmenting the support set using the bounding box information. For the ground truth bounding boxes, derived from fully manual human annotation, each image is augmented once with a crop around the object of interest plus a context window of 60 pixels (30 pixels from each side of the bounding box). In the case of fully automatic or SAM-generated bounding boxes, each image from the support set is augmented three times: first, by resizing the crop to encompass 20% of the remaining context, then with 50% of the remaining context, and finally to include 80% of the context. The rationale behind these choices is elaborated further in Section 4.7.4. The resulting augmented set is then employed for subsequent training steps explained thereafter.

4.5.2 Feature Extraction and Linear probing

We chose to employ CLIP with a ResNet50 image encoder [51] as a feature extractor and subsequently train a single linear layer atop it to tailor it to our classification task. This choice is motivated by recent works showing the ability of this combination to reach top-tier performance, while remaining almost hyperparameter free [54], [111].

In the inductive setting: the linear layer is trained on the CLIP features from the support set and its augmented set.

In the transductive setting: besides the extracted features from the support set and the augmentations, we also train on pseudolabels of the query set, which are generated as detailed in the subsequent section.

We employ a soft K-means algorithm for generating pseudolabels for the query set [81], [83]. The process commences by initializing w clusters with centroids as the class means computed from the support set and the augmentations. Subsequently, we iterate between assigning the elements from the query set to the clusters and computing the new clusters' centroids until convergence. The final cluster assignments are then utilized as the pseudolabels.

4.6 Experimental Setup

To demonstrate the efficacy of the methodology, we conduct evaluations on both classification and object detection tasks. Unlike object detection datasets, classification benchmarks typically contain fewer objects per image, resulting in less distinct yet equally intriguing cases of ambiguity. This section outlines the datasets under consideration, the experimental setup, and showcases the added value of this methodology derived from harnessing local object information within the images. Additionally, we explore the impact of closely cropping around objects of interest on their representation in the feature space generated by the feature extractor. Lastly, we compare various approaches to incorporating object location information.

4.6.1 Datasets

ImageNet [25]

This benchmark contains over 14 million images of a wide variety of object categories. We use a subset of this benchmark containing 1000 classes and 1000 image per class (Imagenet1k) and use the bounding box information of the object available in ImageNet Object Localization Challenge [112].

CUB

This benchmark contains 11788 images from 200 classes. For CUB, the standard Few-Shot split is 100 base-training, 50 validation and 50 test classes [65]. We use the whole dataset for episode generation as CLIP is pre-trained on different data.

Pascal VOC

Pascal VOC Challenge 2021 has around 15,000 annotated images from 20 different classes of mainly common objects. The dataset is split around 50% for training/validation and 50% testing. For our experiments, we generate Few-Shot classification episodes from the training/validation set. While this dataset is typically considered for object detection, some works consider it for image classification tasks[113].

4.6.2 Experimental Setting

For consistency with many works in the literature on Few-Shot visual classification, we focus on Few-Shot tasks made of $k = 5$ classes and a test set of $n_t = 100$ samples, along with a support set of $n_s = 5$ samples for the totality of our experiments. In the transductive scenario, we use a default setting where $n_s + n_q = 50$, giving a total of 50 samples including labeled and unlabeled samples.

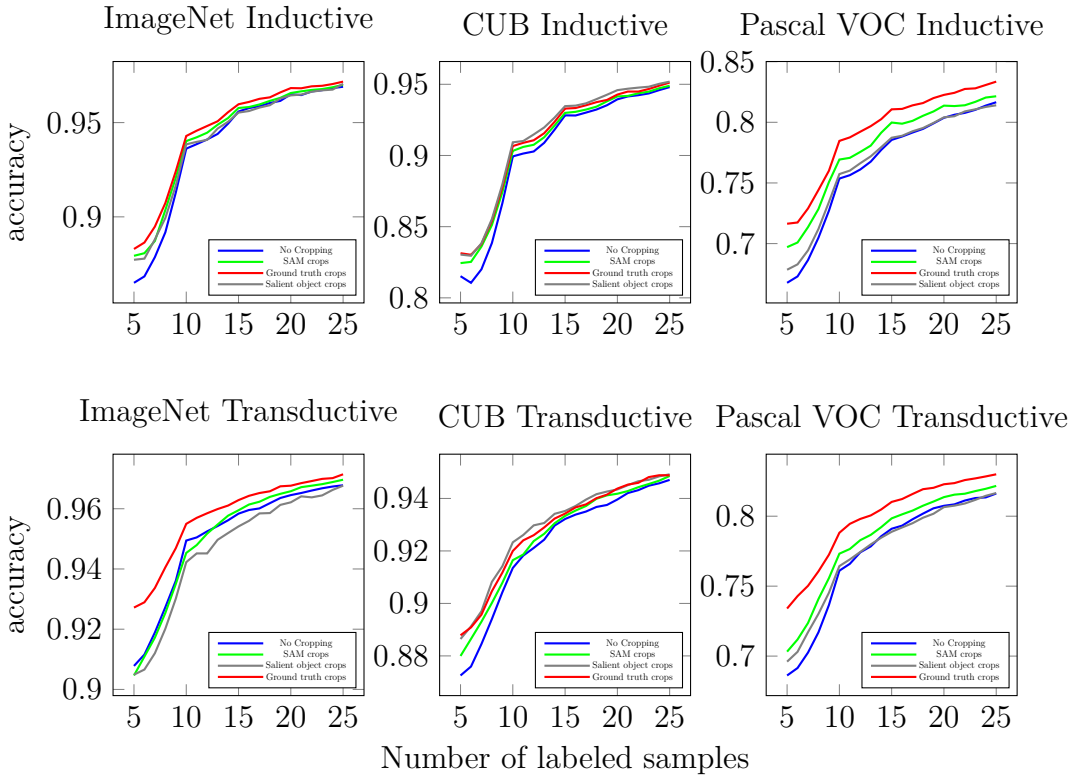
4.7 Results

4.7.1 Effects of cropping on task performance

For our inaugural experiment, our objective is to demonstrate the gains resulting from integrating the crops in our training. In Figure 4.2, we vary the number of support samples and contrast training solely on this support set with training on the augmented support set created using bounding boxes as we previously detailed in Section 4.5.1.

Incorporating ground truth bounding boxes for augmentation leads to a clear improvement compared to baseline training. A notable 5% increase is

Figure 4.2 – Comparison of classification performance using our methodology between the augmentation with crops scenarios (Salient Object Detection, SAM-generated crops and ground truth) and the baseline scenario without any local information about the object of interest. Results are reported for all three considered datasets in both inductive and transductive setting. We compute results for a variable number of labeled samples (support set). These results were computed over 100 runs.



achieved for both inductive and transductive settings with 5 labeled samples in the case of Pascal VOC. This improvement is around 2% for ImageNet and CUB. We attribute this difference in accuracy increase to Pascal VOC being an object detection dataset. It predominantly contains multiple object categories in most images as opposed to CUB and ImageNet. Consequently, the bounding box information in Pascal VOC offers more valuable insights for the model. Additionally, we observe a tendency for the gap between the baseline and ground truth curves to diminish as we increase the number of labeled samples, indicating that this disambiguation is more effective in lower shot settings.

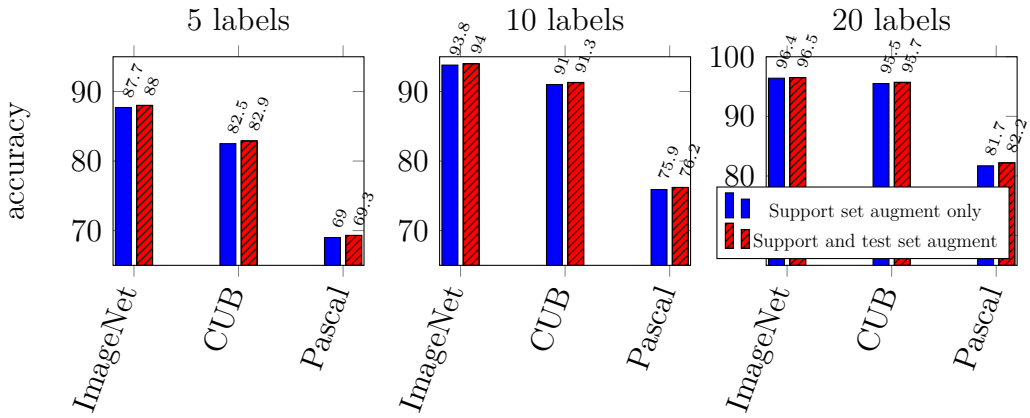
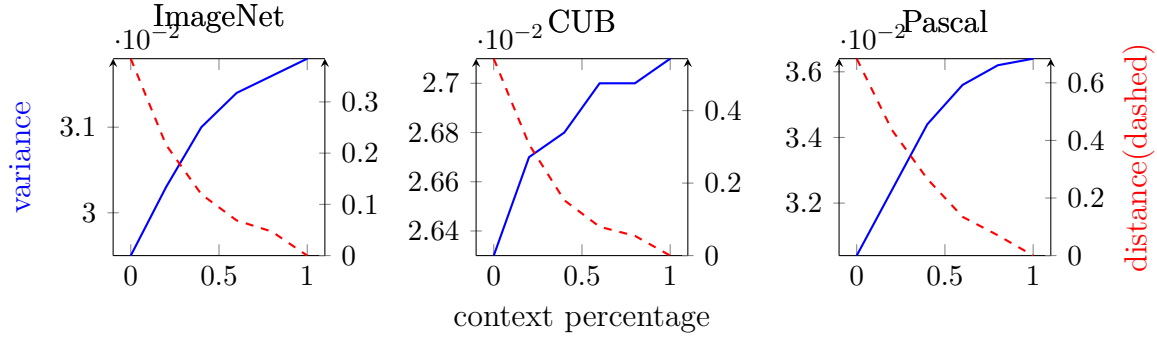


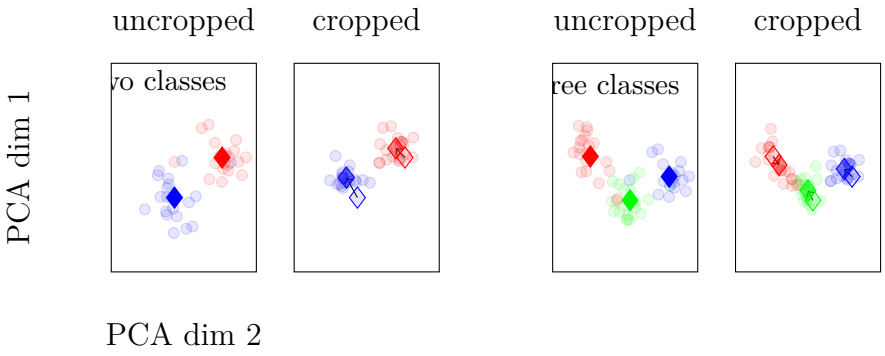
Figure 4.3 – This figure illustrates the outcomes of augmenting the test set at inference time through salient object detection. The reported results pertain to all three datasets within 5-label, 10-label, and 20-label scenarios. A comparison is made between prediction results on an unaltered test set and those on an augmented test set using MOVE's automatic segmentation model. In both cases, training is performed with the MOVE-augmented support set. Results are based on 1000 runs.

Figure 4.4 – Illustration of the impact of object-centric cropping on class variance. Our analysis encompasses the average across 100 classes for ImageNet, all 20 classes for Pascal VOC, and the 60 training classes of CUB. Across all datasets, we examine a random distribution of 100 samples for each class. We present the average class variance of the latent representations and average distance to the original uncropped class means for different percentages of context, where 0 corresponds to the minimal crop (i.e., the most compact bounding box), 1 represents the entire image, and intermediate values indicate linear interpolation between the two extremes.



Regarding the SAM-generated bounding boxes, we note that, for most cases, the curve lies between the two aforementioned curves. We attribute this to the

Figure 4.5 – This figure offers a two-dimensional visualization of the feature space, showcasing class distributions for Pascal VOC. We display 20 random samples (circles) and their centroids (diamond) from two and three random Pascal VOC classes, contrasting the distribution of the latent representation of the uncropped images with the latent representation of the image crops. In the "cropped" instance we display again the uncropped centroids as lower opacity diamonds to better showcase the shift.



quality of the mask generation. The salient object detection boxes yield very similar performances to those of SAM-generated or ground truth boxes in CUB and Imagenet, but result in slightly less, albeit still significant, performance boosts in Pascal VOC. We attribute this directly to the method's agnosticism to which object is of interest. In fact, when multiple objects are present in the foreground of an image, the generated mask can correspond to any of the foreground objects. Moreover, this experiment also emphasizes the associated return on our costs. For example, in the case of Pascal VOC in the inductive setting with only five labeled samples, enhancing these five samples results in a performance improvement similar to that achieved with eight total samples, a comparison we think to be valuable when deciding how to best leverage human annotation.

4.7.2 Inference Time Disambiguation with Salient Object Detection

In this section, we investigate the advantages of inferencing on specific crops of the test images as opposed to the entire image as a single instance. To conduct this experiment, we consider the inductive setting in all three datasets. We employ the training procedure based on the automatically generated masks. As

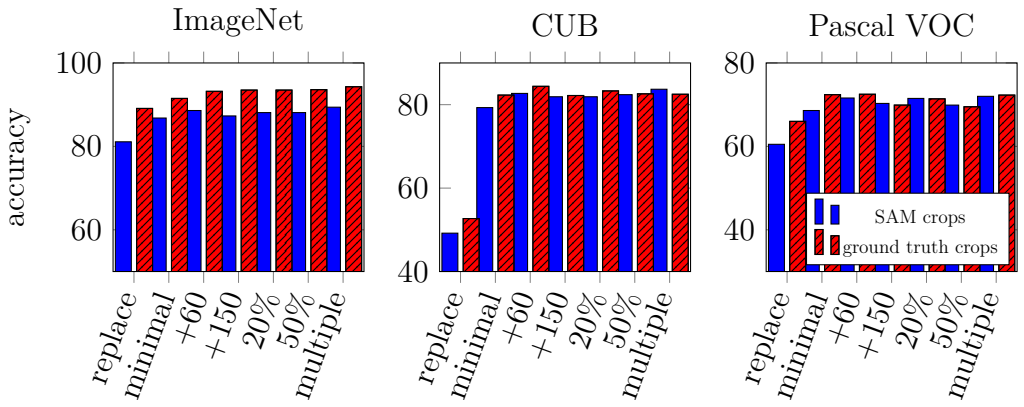


Figure 4.6 – Comparison of different methods for augmenting the support set using bounding boxes centered on the object of interest across three datasets. This analysis considers both SAM-generated and ground truth bounding boxes, and the reported averages are based on 100 runs for a 5-class, 5-support samples, and 100-test samples.
The X-axis is to be interpreted as follows:

- replace: only the crop with an additional 60 pixels of context is used for training, discarding the original image
- minimal: augmenting the original image with a crop around the bounding box.
- +X: augmentation with a crop around the bounding box with an additional number of X context pixels in both width and height.
- X%: augmentation with a resize of the crop that encompasses X% of the remaining context between the whole image and the minimally augmented crop.
- multiple: three augmentations are used in addition to the original image: 20%, 50%, and 80%.

depicted in Figure 4.3, this refined inference process leads to small improvements in classification accuracy across all examined support set sizes for all datasets. Although augmenting the test images can aid in disambiguation, in numerous instances, backgrounds from test images may arbitrarily align strongly with other incorrect classes, thus impeding performance. Another issue, similar to training, can arise when the automatically segmented object is not the object of interest.

4.7.3 Impact of cropping on latent representations

This second experiment seeks to elucidate the impact of cropping on the CLIP latent representation of images, with a focus the variance of the class

distributions of the latent features and on the distance between the cropped images class centroids and their uncropped counterparts. In Figure 4.4, we compute the average class variance of the dataset's classes and the distance to the original (i.e. uncropped) class centroids for image crops with increasingly more context. Our observations reveal a consistent rise in variance as the contextual information increases contrasted with a decrease in the distance to the original class centroids. Notably, this trend is more pronounced for Pascal VOC. While the lower variance might help a model learn better, the shift in the class means creates a bias that hinders it. Hence the trade-off between minimizing cropping and retaining pertinent context. Figure 4.5 visually demonstrates this effect in a two-dimensional space. Here, we project the 1024 CLIP features onto a two-dimensional space that retains the most variance for the uncropped dataset, achieved through Principal Component Analysis. Subsequently, we project random samples from Pascal VOC classes into this space. The resulting clusters corresponding to the classes appear more tightly knit in the case of the cropped samples, indicative of lower variance. Furthermore, we perceive the shift in the class centroids associated with the cropping.

4.7.4 Augmentation with crops and context importance

In this section, we explore different approaches to training the classifier using the ground truth and SAM-generated bounding boxes. An intriguing finding is that discarding the uncropped image can significantly hamper the model's performance. While the context can be spurious, it can also aid the model in adapting to the new domain. This effect is particularly evident in the "replace" mode for CUB in Figure 4.6, where accuracy is nearly halved when discarding the original whole images. Given this insight, the methodology was developed to preserve the original image and augment it with crops of various rescalings. We compare absolute resizing, which considers a fixed window in pixels around the bounding box for, with relative resizing, where a percentage of the remaining context is factored into the augmentation generation. Additionally, we included a mode where we perform three augmentations with increasing relative context. The results are quite close, as depicted in the Figure 4.6. However, we observe that when using ground truth bounding boxes, the highest accuracy across

datasets is achieved with a fixed context of 60 pixels, whereas for the SAM-generated bounding boxes, the “multiple” mode exhibits a slight advantage. This is likely due to imperfections in SAM-generated bounding boxes, such as only covering a part of the object of interest, which is mitigated in the “multiple” mode.

4.8 Object Centric Cropping in Industrial use-cases

Object-centric cropping to resolve ambiguity is applicable to any learning task in which spurious elements may appear in images. One such industrial application addressed during this thesis is defect detection and classification on wrapped products. These wrapped products move along a conveyor belt, where a top-view camera captures each sample. The objective is to detect tears or other packaging issues and flag them accordingly.

The methods initially applied to this use-case demonstrated effective learning on a dataset containing several hundred images, but its performance declined significantly on other datasets. The primary reason for this inconsistency was the disparity in data acquisition scenarios: images of "normal" products were collected in a single session, while images of defective products were gathered later, also in one session. Since the camera captured images from above the conveyor belt, background elements were part of the images. The background elements in the set of normal products differed slightly from those in the defective product set, leading the model to learn spurious correlations that did not generalize well to new data.

We identified and resolved this issue using augmented, object-centered crops of the product. Given that the product consistently appeared in the image center, we queried SAM with central point, which enabled us to retrieve an exact bounding box of the product in a fully automated manner. We then used this bounding box to generate augments that we trained our model on as described in the methodology. This resulted in greater generalization performance.

4.9 Conclusion

The approach presented in this chapter addresses the challenge of task ambiguity in the transductive and inductive scenarios of Few-Shot classification. The methodology capitalizes on the local information of a target object within an image during the training phase, demonstrating strong performance on pertinent Few-Shot classification benchmarks. We leverage existing state-of-the-art large-scale segmentation salient object detection models to reduce the human annotation cost associated with generating local object information. This can bring value to many applications, especially industrial ones, where the data is ambiguous and costly to acquire. Looking ahead, we expect that further exploration of efficient methodologies for identifying objects of interest within images and more seamless ways of incorporating this information into the training and inference of models could lead to better disambiguation.

Conclusion

General Conclusion

As machine learning continues to advance rapidly, several key challenges have emerged, particularly those related to data efficiency. Issues surrounding both the quantity and quality of data are of particular concern, especially in the context of vision tasks, where numerous models with strong prior knowledge have been developed.

To effectively solve data scarce learning problems, leveraging human knowledge and expertise goes beyond simply acquiring and labeling new data. Several strategies can be employed, each with its own cost and benefit. These include acquiring new data, labeling it, refining existing annotations, or even developing textual prompts to prompt Vision-Language models. The advantages of each approach depend on the specific context, so it is essential to carefully consider these factors. This mindful comparison of possible approaches, prior to allocating whichever available expertise, can help prevent redundancy in information and value created. This ensures that efforts are optimized to achieve the best possible results within the given manpower.

The primary approaches to utilizing human expertise addressed in this thesis include the annotation of unlabeled data, which is explored in detail in Chapter 3, and data augmentation through object-centered cropping within images, a topic that is further examined in Chapter 4.

4.9.1 Conclusions around Active Learning For Few-Shot Classification

In our formulation of the Active Few-Shot paradigm, we tackled the following question: Given an unlabeled set of data and a limited budget, how do we characterize the samples that will allow for the best generalization once labeled? We presented an approach that employs a two-step methodology, combining a

statistical inference method with an active sampling function. In the active sampling step, samples are selected for labeling based on Log-probs Soft K-Means Sampling.

In benchmarks adapted from the TFSC domain, this methodology demonstrated superior accuracy and proved robust across a range of problem settings, including variations in labeling budget, the size of the unlabeled data pool, the severity of data imbalance, and domain shift. These results suggest that prioritizing this framework is advantageous, particularly in scenarios where a large volume of unlabeled data has been collected and labeling is intended to be conducted at a later stage.

In addition to addressing the classification task through the final iteration of statistical inference, this methodology can be effectively applied to the labeling process for any unlabeled vision dataset. The selection criterion proposed in this work is designed to optimize labeling efforts by prioritizing the most impactful data points for annotation by ensuring diversity and mitigating uncertainty in the sampled data points. This is beneficial in tasks where data has been collected but not yet annotated, particularly under restrained labeling budgets.

4.9.2 Conclusions around Object-Centric Cropping for Disambiguating Classification Tasks

Ambiguity is a prevalent issue in various learning tasks, and it is particularly exacerbated in Few-Shot settings, where it can be much more easily overrepresented. While developing architectures and learning strategies that account for this challenge is valuable, we investigated the advantages of addressing the issue at its root by disambiguating training samples. By leveraging assistance from human experts, or even employing automated methods in certain contexts, we devised a methodology that augments data by centering augmentations around the object of interest. This approach enables the learner to concentrate on the relevant object rather than on ambiguous or spurious features that may or may not correlate with it. The proposed methodology demonstrates robust performance on relevant Few-Shot classification benchmarks. Furthermore, we explored the integration of existing state-of-the-art large-scale segmentation and salient object detection models to minimize the human annotation costs

associated with generating localized object information. Depending on the characteristics of the images and the specific task, the necessity for human expertise in extracting foreground objects can be reduced. Enhancing data in this manner, can yield equivalent results to acquiring and annotating new samples whereas the cost of obtaining this additional information may be marginal compared to that of annotating extra samples. For instance, in the task of labeling birds by species, a certain knowledge is required for discriminating between different species. Whereas simply pinpointing the bird in the image requires a lot less expertise. This approach can be particularly advantageous in industrial contexts where domain experts responsible for annotating data are distinct from the machine learning engineers tasked with developing and deploying the learning models.

Limitations

Throughout this work, and through our exposure to real-world industrial learning problems with various constraints on data availability, we identified several unaddressed limitations that warrant further exploration. In this section, we enumerate and discuss these challenges.

Domain Adaptation: This challenge falls under the broader category of factors contributing to the inherent difficulty of the task. It becomes especially relevant in the context of transferring knowledge through pre-training and feature extraction. Most large datasets predominantly consist of natural images (e.g., RGB images of everyday objects), while the Few-Shot tasks in our use case involve entirely different types of images, such as MRI scans, grayscale ultrasounds, or highly detailed close-ups of circuit boards. We observe that, despite claims of addressing domain adaptation, many transfer learning methods, including our own, encounter performance ceilings in extreme domains. [114] shows that due to fundamental differences in data sizes, features and task specifications between natural image classification and the medical task, transfer offers very little benefits. This suggests that domain shift remains a significant barrier, especially when transferring between such drastically different image types.

Problem Setting and Learning Paradigm: Real industrial problem

settings rarely align perfectly with the narrow paradigms often addressed in the literature, despite considerable efforts to account for real-world complexities. It is crucial to spend more time defining the problem setting based on the specific circumstances of each task. A key question that arises is whether the problem should be molded to fit a convenient solution or whether a tailored solution should be developed for the specific problem setting. This is a critical consideration that can significantly impact the effectiveness of the chosen approach.

Inference Time and Computational Power: While inference time and computational constraints are general limitations across many machine learning problems, they are sometimes overlooked in Few-Shot learning due to the assumption that data is limited. However, the large models often used in pre-training can impose substantial computational costs for training but also during inference. Some works [115], similarly to ours, avoid fine-tuning to mitigate this training cost and the use of lightweight or even training-free network "adapters" has gained in popularity [116]. Training time and inference time are important factors to consider, particularly in industrial applications where computational resources may be limited, and real-time processing may be required.

Perspectives

An intriguing approach to leveraging expertise in the context of Zero-Shot Learning with vision-language models involves the design of meaningful prompts. Prompt engineering serves as a powerful methodology for enhancing the efficiency and adaptability of learning models across various tasks and domains. However, a primary challenge in prompt engineering is ensuring robustness; even minor alterations in the phrasing or structure of a prompt can significantly impact the model's performance, leading to inconsistencies in outcomes. Consequently, the development of prompts that generalize effectively across diverse tasks and domains remains an open research question.

In the continuity of our work, it is pertinent to consider the labor costs associated with creating accurate prompts in comparison to employing the same labor towards annotation or data acquisition. Research indicates that

carefully crafted prompt templates can yield strong zero-shot performance in a wide range of anomaly detection tasks involving different objects and textures across various datasets. For instance, the Winclip model [117] demonstrates the effectiveness of such carefully designed prompt templates in anomaly detection and classification. However, there remains a notable gap in the literature regarding the evaluation of the costs involved in developing these templates. Quantifying these costs could provide valuable insights into the trade-offs between manual prompt crafting and the automated learning of prompts, ultimately informing future research in this domain.

References

[1] A. L. Samuel, « Some studies in machine learning using the game of checkers », *IBM Journal of Research and Development*, vol. 3, *3*, pp. 210–229, 1959. DOI: 10.1147/rd.33.0210.

[2] I. Goodfellow, *Deep learning*, 2016.

[3] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, « Cancer diagnosis using deep learning: a bibliographic review », *Cancers*, vol. 11, *9*, p. 1235, 2019.

[4] B. Al-Bander, W. Al-Nuaimy, M. A. Al-Taee, and Y. Zheng, « Automated glaucoma diagnosis using deep learning approach », *in 2017 14th International Multi-Conference on Systems, Signals & Devices (SSD)*, IEEE, 2017, pp. 207–210.

[5] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, « Deep learning for healthcare: review, opportunities and challenges », *Briefings in bioinformatics*, vol. 19, *6*, pp. 1236–1246, 2018.

[6] Y. El Ouahidi, V. Gripon, B. Pasdeloup, G. Bouallegue, N. Farrugia, and G. Lioi, « A strong and simple deep learning baseline for bci motor imagery decoding », *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 3338–3347, 2024. DOI: 10.1109/TNSRE.2024.3451010.

[7] S. Zhang, L. Yao, A. Sun, and Y. Tay, « Deep learning based recommender system: a survey and new perspectives », *ACM computing surveys (CSUR)*, vol. 52, *1*, pp. 1–38, 2019.

[8] R. Hussain and S. Zeadally, « Autonomous cars: research results, issues, and future challenges », *IEEE Communications Surveys & Tutorials*, vol. 21, *2*, pp. 1275–1313, 2018.

[9] L. C. Possatti, R. Guidolini, V. B. Cardoso, *et al.*, « Traffic light recognition using deep learning and prior maps for autonomous cars », *in 2019 international joint conference on neural networks (IJCNN)*, IEEE, 2019, pp. 1–8.

[10] J. Liu, G. Xie, J. Wang, *et al.*, « Deep industrial image anomaly detection: a survey », *Machine Intelligence Research*, vol. 21, *1*, pp. 104–135, 2024.

[11] M. Abdelaty, R. Doriguzzi-Corin, and D. Siracusa, « Daics: a deep learning solution for anomaly detection in industrial control systems », *IEEE Transactions on Emerging Topics in Computing*, vol. 10, *2*, pp. 1117–1129, 2021.

[12] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, « Compute trends across three eras of machine learning », *in 2022 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2022, pp. 1–8.

[13] *About pytorch*, `https://pytorch.org/foundation`.

[14] M. Z. Alom, T. M. Taha, C. Yakopcic, *et al.*, « A state-of-the-art survey on deep learning theory and architectures », *electronics*, vol. 8, *3*, p. 292, 2019.

[15] J. Hestness, S. Narang, N. Ardalani, *et al.*, « Deep learning scaling is predictable, empirically », *arXiv preprint arXiv:1712.00409*, 2017.

[16] A. Rai, « Explainable ai: from black box to glass box », *Journal of the Academy of Marketing Science*, vol. 48, pp. 137–141, 2020.

[17] L. Bouza, A. Bugeau, and L. Lannelongue, « How to estimate carbon footprint when training deep learning models? a guide and review », *Environmental Research Communications*, vol. 5, *11*, p. 115 014, 2023.

[18] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, « The ethics of algorithms: mapping the debate », *Big Data & Society*, vol. 3, *2*, p. 2 053 951 716 679 679, 2016.

[19] B. Giovanola and S. Tiribelli, « Beyond bias and discrimination: redefining the ai ethics principle of fairness in healthcare machine-learning algorithms », *AI & society*, vol. 38, *2*, pp. 549–563, 2023.

[20] M. A. Bansal, D. R. Sharma, and D. M. Kathuria, « A systematic review on data scarcity problem in deep learning: solution and applications », *ACM Computing Surveys (Csur)*, vol. 54, *10s*, pp. 1–29, 2022.

[21] X. Zhou, Y. Hu, W. Liang, J. Ma, and Q. Jin, « Variational lstm enhanced anomaly detection for industrial big data », *IEEE Transactions on Industrial Informatics*, vol. 17, *5*, pp. 3469–3477, 2020.

[22] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, « Meta-learning in neural networks: a survey », *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, *9*, pp. 5149–5169, 2021.

[23] F. Zhuang, Z. Qi, K. Duan, *et al.*, « A comprehensive survey on transfer learning », *Proceedings of the IEEE*, vol. 109, *1*, pp. 43–76, 2020.

[24] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, « A brief review of domain adaptation », *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[25] O. Russakovsky, J. Deng, H. Su, *et al.*, « Imagenet large scale visual recognition challenge », *International journal of computer vision*, vol. 115, *3*, pp. 211–252, 2015.

[26] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, « A survey of convolutional neural networks: analysis, applications, and prospects », *IEEE transactions on neural networks and learning systems*, vol. 33, *12*, pp. 6999–7019, 2021.

[27] M. Shafiq and Z. Gu, « Deep residual learning for image recognition: a survey », *Applied Sciences*, vol. 12, *18*, p. 8972, 2022.

[28] T. B. Brown, « Language models are few-shot learners », *arXiv preprint arXiv:2005.14165*, 2020.

[29] O. Russakovsky, J. Deng, H. Su, *et al.*, « Imagenet large scale visual recognition challenge », *CoRR*, vol. abs/1409.0575, 2014. arXiv: `1409.0575`. [Online]. Available: `http://arxiv.org/abs/1409.0575`.

[30] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn, « Adaptive risk minimization: learning to adapt to domain shift », *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 664–23 678, 2021.

[31] A. Tamkin, D. Nguyen, S. Deshpande, J. Mu, and N. Goodman, « Active learning helps pretrained models learn the intended task », *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 140–28 153, 2022.

[32] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, « Activation functions in deep learning: a comprehensive survey and benchmark », *Neurocomputing*, vol. 503, pp. 92–108, 2022.

[33] H.-D. Block, « The perceptron: a model for brain functioning. i », *Reviews of Modern Physics*, vol. 34, *1*, p. 123, 1962.

[34] P. J. Werbos, « Backpropagation through time: what it does and how to do it », *Proceedings of the IEEE*, vol. 78, *10*, pp. 1550–1560, 1990.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, « Gradient-based learning applied to document recognition », *Proceedings of the IEEE*, vol. 86, *11*, pp. 2278–2324, 1998.

[36] S. Ren, K. He, R. Girshick, and J. Sun, « Faster r-cnn: towards real-time object detection with region proposal networks », *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, *6*, pp. 1137–1149, 2016.

[37] M. Hussain, J. J. Bird, and D. R. Faria, « A study on cnn transfer learning for image classification », *in Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*, Springer, 2019, pp. 191–202.

[38] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, « Attention-based models for speech recognition », *Advances in neural information processing systems*, vol. 28, 2015.

[39] S. Hochreiter and J. Schmidhuber, « Flat minima », *Neural computation*, vol. 9, *1*, pp. 1–42, 1997.

[40] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, « Empirical evaluation of gated recurrent neural networks on sequence modeling », *arXiv preprint arXiv:1412.3555*, 2014.

[41] A. Vaswani, « Attention is all you need », *Advances in Neural Information Processing Systems*, 2017.

[42] J. Devlin, « Bert: pre-training of deep bidirectional transformers for language understanding », *arXiv preprint arXiv:1810.04805*, 2018.

[43] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, and M. Iyyer, « Bert with history answer embedding for conversational question answering », *in Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 1133–1136.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, « An image is worth 16x16 words: transformers for image recognition at scale », *CoRR*, vol. abs/2010.11929, 2020. arXiv: 2010.11929. [Online]. Available: https://arxiv.org/abs/2010.11929.

[45] J. Achiam, S. Adler, S. Agarwal, *et al.*, « Gpt-4 technical report », *arXiv preprint arXiv:2303.08774*, 2023.

[46] A. Ramesh, M. Pavlov, G. Goh, *et al.*, « Zero-shot text-to-image generation », *in International conference on machine learning*, Pmlr, 2021, pp. 8821–8831.

[47] L. Deng, « The mnist database of handwritten digit images for machine learning research [best of the web] », *IEEE signal processing magazine*, vol. 29, *6*, pp. 141–142, 2012.

[48] A. Krizhevsky, G. Hinton, *et al.*, « Learning multiple layers of features from tiny images », 2009.

[49] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, « The pascal visual object classes (voc) challenge », *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, « Imagenet classification with deep convolutional neural networks », *Advances in neural information processing systems*, vol. 25, 2012.

[51] A. Radford, J. W. Kim, C. Hallacy, *et al.*, « Learning transferable visual models from natural language supervision », *in International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[52] A. Kirillov, E. Mintun, N. Ravi, *et al.*, « Segment anything », *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[53] J. Bohné, Y. Ying, S. Gentric, and M. Pontil, « Large margin local metric learning », *in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, Springer, 2014, pp. 679–694.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, « A simple framework for contrastive learning of visual representations », *in International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[55] M. Caron, H. Touvron, I. Misra, *et al.*, « Emerging properties in self-supervised vision transformers », *in Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[56] B. Settles, « Active learning literature survey », 2009.

[57] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, « Multi-class active learning by uncertainty sampling with diversity maximization », *International Journal of Computer Vision*, vol. 113, *2*, pp. 113–127, 2015.

[58] J. Shao, Q. Wang, and F. Liu, « Learning to sample: an active learning framework », *in 2019 IEEE international conference on data mining (ICDM)*, IEEE, 2019, pp. 538–547.

[59] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, « A review of supervised object-based land-cover image classification », *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 277–293, 2017.

[60] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, « Deep clustering for unsupervised learning of visual features », *in Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[61] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, « Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction », *IEEE Transactions on Image Processing*, vol. 19, *7*, pp. 1921–1932, 2010.

[62] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, « Masked autoencoders are scalable vision learners », *in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[63] M. Noroozi and P. Favaro, « Unsupervised learning of visual representations by solving jigsaw puzzles », *in European conference on computer vision*, Springer, 2016, pp. 69–84.

[64] Wikipedia, *Understanding cnns*, `https://en.m.wikipedia.org/wiki/File:Typical_cnn.png`.

[65] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, « A closer look at few-shot classification », *arXiv preprint arXiv:1904.04232*, 2019.

[66] I. Bello, W. Fedus, X. Du, *et al.*, « Revisiting resnets: improved training and scaling strategies », *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 614–22 627, 2021.

[67] M. Tan and Q. E. Le, « Rethinking model scaling for convolutional neural networks. arxiv 2019 », *arXiv preprint arXiv:1905.11946*, 1905.

[68] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, « Training data-efficient image transformers & distillation through attention », *in International conference on machine learning*, PMLR, 2021, pp. 10 347–10 357.

[69] C. Finn, P. Abbeel, and S. Levine, « Model-agnostic meta-learning for fast adaptation of deep networks », *in International conference on machine learning*, PMLR, 2017, pp. 1126–1135.

[70] J. Snell, K. Swersky, and R. Zemel, « Prototypical networks for few-shot learning », *Advances in neural information processing systems*, vol. 30, 2017.

[71] Z. Ji, X. Chai, Y. Yu, Y. Pang, and Z. Zhang, « Improved prototypical networks for few-shot learning », *Pattern Recognition Letters*, vol. 140, pp. 81–87, 2020.

[72] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, « Learning to prompt for vision-language models », *International Journal of Computer Vision*, vol. 130, *9*, pp. 2337–2348, 2022.

[73] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, « Maple: multi-modal prompt learning », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 113–19 122.

[74] O. Sener and S. Savarese, « Active learning for convolutional neural networks: a core-set approach », *arXiv preprint arXiv:1708.00489*, 2017.

[75] P. Bachman, A. Sordoni, and A. Trischler, « Learning algorithms for active learning », *in international conference on machine learning*, PMLR, 2017, pp. 301–310.

[76] J. Redmon, « You only look once: unified, real-time object detection », *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[77] A. Bulat, R. Guerrero, B. Martinez, and G. Tzimiropoulos, « Fs-detr: few-shot detection transformer with prompting and without re-training », *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 793–11 802.

[78] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, « Unsupervised learning of visual features by contrasting cluster assignments », *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[79] Y. Xie, H. Lu, J. Yan, X. Yang, M. Tomizuka, and W. Zhan, « Active finetuning: exploiting annotation budget in the pretraining-finetuning paradigm », *in Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 23 715–23 724.

[80] H. T. Nguyen and A. Smeulders, « Active learning using pre-clustering », *in Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 79.

[81] A. Abdali, V. Gripon, L. Drumetz, and B. Boguslawski, « Active learning for efficient few-shot classification », *in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[82] Y. Hu, V. Gripon, and S. Pateux, « Leveraging the feature distribution in transfer-based few-shot learning », *in International Conference on Artificial Neural Networks*, Springer, 2021, pp. 487–499.

[83] Y. Bendou, Y. Hu, R. Lafargue, *et al.*, *Easy: ensemble augmented-shot y-shaped learning: state-of-the-art few-shot classification with simple ingredients*, 2022. arXiv: 2201.09699 [cs.LG].

[84] M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed, « Information maximization for few-shot learning », *Advances in Neural Information Processing Systems*, vol. 33, pp. 2445–2457, 2020.

[85] O. Veilleux, M. Boudiaf, P. Piantanida, and I. Ben Ayed, « Realistic evaluation of transductive few-shot learning », *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[86] D. Chen, Y. Chen, Y. Li, F. Mao, Y. He, and H. Xue, « Self-supervised learning for few-shot image classification », *in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 1745–1749.

[87] X. Yue, Z. Zheng, S. Zhang, *et al.*, « Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 834–13 844.

[88] Y. Tian, D. Krishnan, and P. Isola, « Contrastive multiview coding », *in European conference on computer vision*, Springer, 2020, pp. 776–794.

[89] Y. Gal, R. Islam, and Z. Ghahramani, « Deep bayesian active learning with image data », *in International Conference on Machine Learning*, PMLR, 2017, pp. 1183–1192.

[90] N. Houlsby, J. M. Hernández-Lobato, and Z. Ghahramani, « Cold-start active learning with robust ordinal matrix factorization », *in International Conference on Machine Learning*, PMLR, 2014, pp. 766–774.

[91] T. Müller, G. Pérez-Torró, A. Basile, and M. Franco-Salvador, « Active few-shot learning with fasl », *in International Conference on Applications of Natural Language to Information Systems*, Springer, 2022, pp. 98–110.

[92] G. Wang, J.-N. Hwang, C. Rose, and F. Wallace, « Uncertainty sampling based active learning with diversity constraint by sparse selection », *in 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2017, pp. 1–6.

[93] P. Li, F. Liu, L. Jiao, *et al.*, « Knowledge transduction for cross-domain few-shot learning », *Pattern Recognition*, vol. 141, p. 109 652, 2023, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2023.109652.

[94] E. Triantafillou, T. Zhu, V. Dumoulin, *et al.*, « Meta-dataset: a dataset of datasets for learning to learn from few examples », *arXiv preprint arXiv:1903.03096*, 2019.

[95] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, « Simpleshot: revisiting nearest-neighbor classification for few-shot learning », *arXiv preprint arXiv:1911.04623*, 2019.

[96] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, « Simplifying graph convolutional networks », *in International conference on machine learning*, PMLR, 2019, pp. 6861–6871.

[97] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, « The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains », *IEEE signal processing magazine*, vol. 30, *3*, pp. 83–98, 2013.

[98] R. Boney and A. Ilin, « Semi-supervised few-shot learning with prototypical networks », *CoRR abs/1711.10856*, 2017.

[99] Y. Gu, Z. Jin, and S. C. Chiu, « Active learning combining uncertainty and diversity for multi-class image classification », *IET Computer Vision*, vol. 9, *3*, pp. 400–407, 2015.

[100] B. Oreshkin, P. Rodríguez López, and A. Lacoste, « Tadam: task dependent adaptive metric for improved few-shot learning », *Advances in neural information processing systems*, vol. 31, 2018.

[101] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, « Laplacian regularized few-shot learning », *in International Conference on Machine Learning*, PMLR, 2020, pp. 11 660–11 670.

[102] J. Liu, L. Song, and Y. Qin, « Prototype rectification for few-shot learning », *in European Conference on Computer Vision*, Springer, 2020, pp. 741–756.

[103] P. Pezeshkpour, Z. Zhao, and S. Singh, « On the utility of active instance selection for few-shot learning », *NeurIPS HAMLETS*, 2020.

[104] K. Allen, E. Shelhamer, H. Shin, and J. Tenenbaum, « Infinite mixture prototypes for few-shot learning », *in International Conference on Machine Learning*, PMLR, 2019, pp. 232–241.

[105] O. Siméoni, G. Puy, H. V. Vo, *et al.*, « Localizing objects with self-supervised transformers and no labels », *arXiv preprint arXiv:2109.14279*, 2021.

[106] C. Finn, K. Xu, and S. Levine, « Probabilistic model-agnostic meta-learning », *Advances in neural information processing systems*, vol. 31, 2018.

[107] X. Zhao, W. Ding, Y. An, *et al.*, « Fast segment anything », *arXiv preprint arXiv:2306.12156*, 2023.

[108] A. Bielski and P. Favaro, « Move: unsupervised movable object segmentation and detection », *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 371–33 386, 2022.

[109] Y. Hu, S. Pateux, and V. Gripon, « Adaptive dimension reduction and variational inference for transductive few-shot classification », *in International Conference on Artificial Intelligence and Statistics*, PMLR, 2023, pp. 5899–5917.

[110] H. Zhu and P. Koniusz, « Transductive few-shot learning with prototype-based label propagation by iterative graph refinement », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 996–24 006.

[111] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, « Multimodality helps unimodality: cross-modal few-shot learning with multimodal models », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 325–19 337.

[112] W. K. Addison Howard Eunbyung Park, *Imagenet object localization challenge*, 2018. [Online]. Available: `https://kaggle.com/competitions/imagenet-object-localization-challenge`.

[113] S. Shetty, « Application of convolutional neural network for image classification on pascal voc challenge 2012 dataset », *arXiv preprint arXiv:1607.03785*, 2016.

[114] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, « Transfusion: understanding transfer learning for medical imaging », *Advances in neural information processing systems*, vol. 32, 2019.

[115] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, « Meta-baseline: exploring simple meta-learning for few-shot learning », *in Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9062–9071.

[116] R. Zhang, R. Fang, W. Zhang, *et al.*, « Tip-adapter: training-free clip-adapter for better vision-language modeling », *arXiv preprint arXiv:2111.03930*, 2021.

[117] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, « Winclip: zero-/few-shot anomaly classification and segmentation », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616.

[118] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, « Natural language processing advancements by deep learning: a survey », *arXiv preprint arXiv:2003.01200*, 2020.

[119] M. Sundermeyer, R. Schlüter, and H. Ney, « Lstm neural networks for language modeling. », *in Interspeech*, vol. 2012, 2012, pp. 194–197.

[120] S. Hochreiter, « The vanishing gradient problem during learning recurrent neural nets and problem solutions », *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, *02*, pp. 107–116, 1998.

[121] J. Chai, H. Zeng, A. Li, and E. W. Ngai, « Deep learning in computer vision: a critical review of emerging techniques and application scenarios », *Machine Learning with Applications*, vol. 6, p. 100 134, 2021.

[122] M. Volpp, L. P. Fröhlich, K. Fischer, *et al.*, « Meta-learning acquisition functions for transfer learning in bayesian optimization », *arXiv preprint arXiv:1904.02642*, 2019.

[123] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, « High-resolution image synthesis with latent diffusion models », *in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[124] J. Ho, A. Jain, and P. Abbeel, « Denoising diffusion probabilistic models », *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[125] H. Choi, S. Ryu, and H. Kim, « Short-term load forecasting based on resnet and lstm », *in 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, IEEE, 2018, pp. 1–6.

[126] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, « Large scale metric learning from equivalence constraints », *in 2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 2288–2295.

[127] L. Yang and R. Jin, « Distance metric learning: a comprehensive survey », *Michigan State Universiy*, vol. 2, *2*, p. 4, 2006.

[128] A. A. Rusu, D. Rao, J. Sygnowski, *et al.*, « Meta-learning with latent embedding optimization », *arXiv preprint arXiv:1807.05960*, 2018.

[129] A. Nichol, J. Achiam, and J. Schulman, « On first-order meta-learning algorithms », *arXiv preprint arXiv:1803.02999*, 2018.

[130] R. Vilalta and Y. Drissi, « A perspective view and survey of meta-learning », *Artificial intelligence review*, vol. 18, *2*, pp. 77–95, 2002.

[131] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, « Cost-effective active learning for deep image classification », *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, *12*, pp. 2591–2600, 2016.

[132] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, « Active learning with statistical models », *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.

[133] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, « Less is more: sampling chemical space with active learning », *The Journal of chemical physics*, vol. 148, *24*, p. 241 733, 2018.

[134] K. Li, Y. Zhang, K. Li, and Y. Fu, « Adversarial feature hallucination networks for few-shot learning », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 470–13 479.

[135] B. Kulis *et al.*, « Metric learning: a survey », *Foundations and Trends® in Machine Learning*, vol. 5, *4*, pp. 287–364, 2013.

[136] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, « A baseline for few-shot image classification », *arXiv preprint arXiv:1909.02729*, 2019.

[137] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, and Y. Tian, « Transductive episodic-wise adaptive metric for few-shot learning », *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3603–3612.

[138] M. Yuan, H.-T. Lin, and J. Boyd-Graber, « Cold-start active learning through self-supervised language modeling », *arXiv preprint arXiv:2010.09535*, 2020.

[139] J. Zhang, J. Katz-Samuels, and R. Nowak, « Galaxy: graph-based active learning at the extreme », *arXiv preprint arXiv:2202.01402*, 2022.

[140] E. Amrani and A. Bronstein, « Self-supervised classification network », *arXiv preprint arXiv:2103.10994*, 2021.

[141] M. Ren, E. Triantafillou, S. Ravi, *et al.*, « Meta-learning for semi-supervised few-shot classification », *arXiv preprint arXiv:1803.00676*, 2018.

[142] Y. Liu, J. Lee, M. Park, S. Kim, and Y. Yang, « Transductive propagation network for few-shot learning », 2018.

[143] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, « Cross attention network for few-shot classification », *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[144] G. Qi, H. Yu, Z. Lu, and S. Li, « Transductive few-shot classification on the oblique manifold », *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8412–8422.

[145] Y. Liu, B. Schiele, and Q. Sun, « An ensemble of epoch-wise empirical bayes for few-shot learning », *in European Conference on Computer Vision*, Springer, 2020, pp. 404–421.

[146] C. Simon, P. Koniusz, R. Nock, and M. Harandi, « Adaptive subspaces for few-shot learning », *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4136–4145.

[147] Y. Hu, V. Gripon, and S. Pateux, « Graph-based interpolation of feature vectors for accurate few-shot classification », *in 2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 8164–8171.

[148] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, « Alignment for advanced machine learning systems », *Ethics of Artificial Intelligence*, pp. 342–382, 2016.

[149] A. Smith, C. Jones, and E. Roberts, « Article title », *Journal*, vol. 62, pp. 291–294, Jan. 1920.

[150] H. Caesar, J. Uijlings, and V. Ferrari, « Coco-stuff: thing and stuff classes in context », *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[151] Y. Chen and X. Wang, « Transformers as meta-learners for implicit neural representations », *in European Conference on Computer Vision*, Springer, 2022, pp. 170–187.

COLLEGES SCIENCES
BRETAGNE POUR L'INGENIEUR
LOIRE ET LE NUMERIQUE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Apprentissage avec peu de données spécifiques : Sélection de données, évalutation et applications

Mot clés : Few-Shot, Classification d'Images, Apprentissage Actif

Résumé : L'apprentissage automatique continue d'apporter de nombreux avantages dans les industries où les outils utilisant l'intelligence artificielle optimisent les processus de fabrication, prédisent les pannes des machines, améliorent le contrôle qualité grâce à la détection des défauts en temps réel, etc. Cependant, de nombreux défis persistent pour répondre aux exigences variées des applications dans le monde réel. Cette thèse se concentre sur l'un de ces défis, à savoir l'efficacité dans la création et l'utilisation des données quand ces dernières sont limitées en quantité. Les travaux présentés ici explorent des méthodes visant à améliorer l'efficacité de l'utilisation des données en exploitant les connaissances et l'expertise humaines. Cette exploration va au-delà de l'acquisition et de l'étiquetage de nouvelles données, en incluant des approches telles que le raffinement des annotations existantes et l'utilisation d'outils pour optimiser le processus d'annotation.

Title: Few-Shot Learning: Data Selection, Performance Prediction and Applications

Keywords: Few-Shot, Image Classification, Active Learning

Abstract: Machine Learning continuously brings numerous benefits in industries where AI-powered tools are optimizing manufacturing processes, predicting machinery failures, enhancing quality control with real-time defect detection, etc. However, numerous challenges persist in addressing the diverse requirements of real-world applications. This thesis focuses on one such challenge, specifically related to data efficiency and learning in Few-Shot settings. The research presented here investigates methods to enhance data efficiency by leveraging human knowledge and expertise. This exploration extends beyond merely acquiring and labeling additional data to include approaches such as refining existing annotations and utilizing tools to optimize the annotation process.