ATTENTION MECHANISMS FOR SEMANTIC FEW-SHOT LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ORHUN BUĞRA BARAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2021

Approval of the thesis:

ATTENTION MECHANISMS FOR SEMANTIC FEW-SHOT LEARNING

submitted by **ORHUN BUĞRA BARAN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Supervisor, **Computer Engineering, METU** _____

Assoc. Prof. Dr. Nazlı İkizler-Cinbiş
Co-supervisor, **Computer Engineering, Hacettepe University** _____

Examining Committee Members:

Assist. Prof. Dr. Emre Akbaş
Computer Engineering, METU _____

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Computer Engineering, METU _____

Assist. Prof. Dr. Hamdi Dibeklioğlu
Computer Engineering, Bilkent University _____

Date: 01.09.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Orhun Buğra Baran

Signature :

ABSTRACT

ATTENTION MECHANISMS FOR SEMANTIC FEW-SHOT LEARNING

Baran, Orhun Buğra
M.S., Department of Computer Engineering
Supervisor: Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Co-Supervisor: Assoc. Prof. Dr. Nazlı İkizler-Cinbiş

September 2021, 37 pages

One of the fundamental difficulties in contemporary supervised learning approaches is the dependency on labelled examples. Most state-of-the-art deep architectures, in particular, tend to perform poorly in the absence of large-scale annotated training sets. In many practical problems, however, it is not feasible to construct sufficiently large training sets, especially in problems involving sensitive information or consisting of a large set of fine-grained classes. One of the main topics in machine learning research that aims to address such limitations is few-shot learning where only few labeled samples are made available for each novel class of interest.

An inherent difficulty in few-shot learning is the various ambiguities resulting from having only few training samples per class. To tackle this fundamental challenge in few-shot learning, in this thesis, we propose an approach that aims to guide the meta-learner via semantic priors. To this end, we build meta-learning models that can benefit from prior knowledge based semantic representations of classes of interest when synthesizing target classifiers. We propose semantically-conditioned feature attention and sample attention mechanisms that estimate and utilize the importance

of representation dimensions and training instances. In sample attention, we aim to weigh each individual training example based on its representativeness for the related class. We, then, use the information extracted from each example proportional to its individual weight. In feature attention, we aim to weigh each visual feature dimension based on the semantic embedding vectors we obtain for each class. We also study the problem of sample noise in few-shot learning, where some training examples are irrelevant due to annotation or data collection errors, which can be the case for various real-world problems. Our experimental results demonstrate the effectiveness of the proposed semantic few-shot learning model with and without sample noise.

Keywords: few-shot learning, semantic few-shot learning, meta learning, metric learning, attention

ÖZ

ANLAMSAL AZ ÖRNEKLE ÖĞRENME İÇİN ODAKLANMA MEKANİZMALARI

Baran, Orhun Buğra
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Ramazan Gökberk Cinbiş
Ortak Tez Yöneticisi: Doç. Dr. Nazlı İkizler-Cinbiş

Eylül 2021 , 37 sayfa

Çağdaş denetimli öğrenme yaklaşımlarındaki en temel zorluklardan biri etiketli örneklere olan bağımlılıktır. Özellikle modern yaklaşımlar geniş çaplı etiketli eğitim kümeleri olmadan düşük performans göstermektedir. Ancak gizlilik açısından hassas veri gerektiren, ayrıntılı sınıflandırma içeren ve benzeri çoğu pratik problemde geniş çaplı eğitim kümesi oluşturmak çoğu zaman mümkün olmamaktadır. Makine öğrenmesi alanında, bu zorluğa çözüm arayan çalışma konularından biri de az örnekle öğrenmedir. Az örnekle öğrenmedeki temel amaç, az sayıda etiketli örnek üzerinden yeni sınıfları modelleyebilmektir.

Az örnekle öğrenmedeki ana zorluklardan biri, sınıflara yönelik eğitim verisindeki azlıktan dolayı ortaya çıkan muğlaklıktır. Bu tezde, az örnekle öğrenme problemindeki temel muğlaklık sorunun aşılmasına yönelik olarak, meta modelin anlamsal bilgilerle yönlendirilmesi hedeflenmektedir. Bu noktada hedef sınıflandırıcıların oluşturulmasında sınıfların anlamsal gösterimlerinden faydalanan meta öğrenme modelleri oluşturmaktayız. Eğitim örneklerinin ve gösterim boyutlarının önemini tahmin eden

ve kullanan anlamsal bilgiyle koşullandırılmış öznitelik ve örnek odaklanması mekanizmaları önermekteyiz. Örnek odaklanmasındaki amacımız her bir eğitim örneğinin ait olduğu sınıf için temsil edilebilirliği ölçüsünde ağırlıklandırılmasıdır. Daha sonra her bir örnekten çıkarılan bilgi bu ağırlıklar doğrultusunda kullanılmaktadır. Öznitelik odaklanmasında her bir sınıfa ait anlamsal öznitelik vektörünü baz alarak görsel öznitelik vektörünün boyutlarının ağırlık hesabı amaçlanmaktadır. Ayrıca, az örnekle öğrenme probleminde örnek gürültüsü problemini de ele almaktayız. Yanlış etiketleme veya veri toplamadaki hatalardan kaynaklanan örneklerde gürültü varlığı, gerçek dünya uygulamalarında olası bir senaryodur. Deneysel sonuçlarımız hem gürültülü örnek varlığında hem de yokluğunda önerdiğimiz modelin başarısını göstermektedir.

Anahtar Kelimeler: az örnekle öğrenme, anlamsal az örnekle öğrenme, meta öğrenme, metrik öğrenme, odaklanma

To my family.

ACKNOWLEDGMENTS

I would like to thank my supervisor Assist. Prof. Dr. Ramazan Gökberk Cinbiş and my co-supervisor Assoc. Prof. Dr. Nazlı İkizler-Cinbiş for their immense contributions. Their help was the key for this thesis work.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

FSL	Few-Shot Learning
ZSL	Zero-Shot Learning
PN	Prototypical Networks
AM3	Adaptive Cross-Modal Few Shot Learning

CHAPTER 1

INTRODUCTION

Contemporary supervised learning approaches combined with large training datasets yield excellent results on a variety of recognition problems. A major challenge, however, is learning to model concepts with limited samples. Few-shot learning (FSL) [1, 2, 3, 4] techniques aim to tackle this problem, by learning to synthesize effective models based on few examples.

In few-shot learning we are given a set of *base classes* with many examples that belong to those classes. Then at test time we are given another set of classes, called *novel classes*, with few examples, called *support examples*, that we use to obtain set of parameters that can be used to classify query examples of novel classes. A setting where n-classes with k-examples for each of those n classes is called an *n-way k-shot classification*. The name of that setting is called an *FSL task*. One or more such tasks are called an *episode*. Altough FSL approaches differ in how they treat the base class examples, at test time most of the approaches use episodes and perform n-way k-shot classification with the exception being the *Meta-Dataset* [5] where variable shot/way classification is done. An example FSL task is provided in Figure 1.1.

1.1 Few-shot learning motivation

A major source of motivation for studying FSL is the observation that humans, starting at young ages, can learn new concepts with limited examples [6], [7], [8]. In addition, in most real-world classification problems, such as object recognition [9], class distributions can be heavily long-tailed [10]. FSL research can be seen as the focused study of learning to recognize in the low-data regime, which can play a central

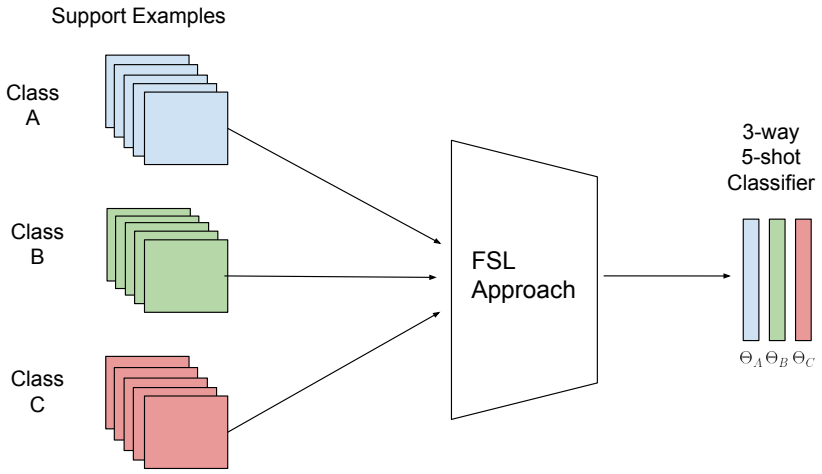

Figure 1.1: 3-way 5-shot episode that is fed into an FSL approach to obtain classification parameters.

role in building semantically comprehensive and rich models.

A variety of FSL approaches have been introduced in recent years. Most of the recent work can be summarized as follows: metric learning [11, 12, 13, 14], statistical generative models [15, 16, 17, 18, 19, 20, 21] and other techniques for data augmentation [22, 23], feed-forward classifier synthesis [24, 25, 26], model initialization for few-shot adaptation [27, 28, 29], learning-to-optimize for FSL [30] and memory-based approaches [31, 32]. In addition, recent work has highlighted the importance of implementations details in improving and evaluating FSL models, including batch normalization details [33], feature extraction backbones [34] and pretraining strategies [35]. Variations of FSL, such as cross-domain [36, 37, 38, 39, 40] and variable-shot [5] learning have also been introduced.

1.2 Difficulties in few-shot learning

An inherent difficulty in FSL, independent of the method being used, is the ambiguity resulting from having few training samples per class. Particularly, it is difficult to fig-

(a) Spurious background. (b) Misleading relations. (c) Feature ambiguity. (d) Prototypicality/noise.

Figure 1.2: (a) All dog instances appear in a misleadingly consistent beach background context. (b) Spuriously consistent background-foreground relation may cause FSL models to ignore other salient features of classes. (c) It is difficult to understand when a consistent foreground texture is informative or misleading, e.g., the Dalmatian texture is distinctive for the *Dalmatian dog* class but otherwise misleading for the generic *dog* class. (d) Some examples can be more *prototypical* than the others. A closely related problem is having *sample noise* in the training set.

ure out whether a cue that appears consistently in the limited set of examples is truly indicative. For example, few-shot samples may contain misleadingly similar contextual information, spurious foreground-background relationships, or suspiciously consistent foreground features. Similarly, some samples might actually be less *prototypical* [41] than the others or completely *noisy*, due to number of factors, such as background clutter, viewpoint, occlusions, overall representativeness or sample noise. Hence, FSL model may *overfit* to incorrect features, misleading or noisy samples, or under-utilize distinctive cues, due to the fundamental difficulty of disambiguating spurious cues from the informative ones purely based on few examples.

Most of these aforementioned difficulties are visualized in Figure 1.2. In Figure 1.2a we see that shore background consistently appears in dog images which can mislead the FSL approaches to incorrectly classify shore images that dont involve dogs as dog class. In Figure 1.2b consistent appearence of foreground and background may mislead the network to have a bias for such relations where the approach can be more interested in black and white textures more than the actual cues of cat class. It can also be the case that some *sub-classes* of a class may be too prevalent in few examples that the approach may struggle to classify *super-class* when query examples are not from

that sub-class. In Figure 1.2c if Dalmatian dog images are used as support images for generic dog class then these images may not be too representative when an African wild dog is given as a test example. Finally, an object in a support image may not be prototypical enough that it can be used to correctly represent a class. In Figure 1.2d we see that some bike images that are prototypical and not prototypical. Images on the top row are not the best ones due to factors such as bad camera angle, crowded background, bikes not being the main focus of the image on and so on. But on the bottom row we see clean shots of bikes with all their distinctive parts shown clearly.

1.3 Semantic few-shot learning

To tackle these fundamental challenges in FSL, we aim to guide the meta-learner via semantic priors, which we call *semantic few-shot learning*. To this end, we build meta-learning models that can benefit from text-based semantic representations of classes of interest when synthesizing target classifiers. For this purpose, we focus on one of the most popular metric based few-shot learners Prototypical Networks [11] (PNs). In the context of PN formulation, we introduce semantically-conditioned *feature attention* and *sample attention* mechanisms, towards reducing the risk of overfitting to misleading features or samples and improving the data efficiency in few-shot learning.

The use of semantic vector-space representations of classes, i.e., *class embeddings*, is prominent in *zero-shot learning* (ZSL). Most mainstream ZSL approaches learn to estimate the degree of relation between a given input (image) and a class embedding, so that previously unseen classes can be recognized purely based on class embeddings, e.g., [42, 43, 44, 45, 19, 46, 47, 48]. In ZSL, class embeddings can be interpreted as class summaries from which valuable discriminative or generative knowledge can be extracted. In our work, instead of relying purely on class embeddings for building classification models, we aim to benefit from them in improving sample efficiency in FSL.

The idea of jointly benefiting from class semantics in FSL has received attention only recently. In [36], semantics-based class prototype priors are estimated and adaptively

combined with the data-driven class prototypes. [37] extends this model to multiple semantic information sources. [38] aims to obtain augmented feature representations based on the encoding of the image features to semantic space and then decoding the sampled semantic features to obtain new samples for training purposes. [40] proposes to reformulate the loss of [11] based on task dependent similarities measured from semantic priors to regulate the margin between classes in a task.

1.4 Our aproach to semantic few-shot learning

The prior work most related to ours is the AM3 model [36], which we use as our starting point. While AM3 makes an important step forward by defining class semantics based models priors, the approach does not leverage prior semantic knowledge in knowledge accumulation. To this end, we explore the uses of semantic class knowledge in a deeper way in the following two main ways. First, we aim to estimate the importance of provided few-shot samples for each class by evaluating the consistency across samples and semantics. Second, we estimate per-dimension feature importance factors for the data-driven class prototypes based on prior knowledge. We additionally tackle the sample noise problem in FSL, which we consider as an overlooked problem in real-world FSL settings, e.g., when recognition models are built in an automated way by retrieving samples based on their meta-data.

Our contributions, therefore, can be summarized as follows: (i) we propose semantics-driven feature and sample attention mechanisms to improve FSL data efficiency in a principled way. (ii) We study the problem of sample noise in FSL and define an experimental protocol for this scenario. (iii) We present a detailed experimental analysis towards understanding the effects and dynamics of the proposed attention mechanisms. Our quantitative and qualitative experimental results demonstrate the effectiveness of the proposed semantic FSL model both in clean and noisy settings.

In the rest of this thesis, chapter 2 provides an overview on the most relevant few-shot learning, semantic few-shot learning, and zero-shot learning approaches. In chapter 3 we explain our approach together with the closely related approaches such as Prototypical Networks [11] and AM3 [36]. Chapter 4 explains the experiments and

validates our approach with comparisons against other state-of-the-art works and ablative studies. In chapter 5 we conclude the thesis with a brief discussion on our approach and state the future work.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we first present an overview of mainstream few-shot learning approaches. We then present an overview of works on the recently emerging topic of semantic FSL. Finally, we briefly discuss zero-shot learning and its relation to our work.

2.1 Few-shot learning

The most related mainstream FSL approaches can be summarized within *initialization based*, *metric learning based* and *generative model based* groups. Initialization based FSL methods aim to learn the *ideal* initial model such that the model can perform well even when fine-tuned using just few examples. MAML [27] is arguably the most well-known example of this category. In MAML, the main idea is to learn the initial model that minimizes the loss of validation samples when the initial model is fine-tuned using one or few gradient-based updates. Several other related and follow-up works exist, such as [28, 29, 49, 30].

In metric learning based FSL, the goal is to learn a metric space where the similarity of feature representations of sample pairs can be used to classify pairs as same class/different class pairs. One of the most well-known examples of this category is Prototypical Networks [11], which we also use as the basis of our approach (see Chapter 3). Due to its simplicity and high FSL performance, many other metric learning FSL approaches have also been introduced, e.g., [12, 13, 24, 25].

Despite these explorations, recent works show that a carefully tuned Prototypical Net-

work can yield state-of-the-art few-shot learning results [35, 34, 50, 51]. These papers mainly focus on obtaining better representations with their respective feature extractors so that when novel classes are being used better features, which improves the FSL accuracy, is obtained. Chen et al. [35] mainly obtains a feature extractor trained with all base classes as a normal classification task. Afterwards the authors train the feature extractor with episodes using cosine distance to further adapt to FSL tasks. Similarly, Chen et al. [34] also trains the feature extractor with base classes however, in the second stage the authors only train the single classification layer to adapt to novel classes. In Tian et al. [50] the goal is not so different than [34, 35] but the authors introduce a self distillation scheme to further improve the feature extractor after base class training. In Ye et al. [51] the approach is again learning a feature extractor with base classes using standard classification approach however, the authors select the model that performs best on 16-way 1-shot tasks on validation set without any additional training of feature extractor. We adapt the *modernized* PN implementation of [51] as the non-semantic FSL baseline in the construction of our models.

Generative modeling based FSL approaches aim to learn a sample-synthesizing model, which can be used for augmenting a few-shot training set. For example, Hariharan et al. [15] learns a mapping that can be used to transform existing training samples into new ones, [52, 17] propose GAN [53] based models generative models towards synthesizing novel examples. In Zhang et al. [52] a meta-learning based GAN generator and discriminator is learned on base classes to both classify classes in FSL tasks and produce additional synthetic data around the manifold of real data distribution. This way the discriminator not only benefits from the real data but also uses synthetic data to obtain a better decision boundary which is required in FSL classifiers. In a different approach Gao et al. [17] aims to produce synthetic data for novel classes that can be used in a FSL classifier. The method uses cosine similarity to determine how close base classes and novel classes are, by using their respective features, and tries to obtain novel class data that has covariance similar to base class data using an additional loss function.

Sample noise in FSL is largely an overlooked problem. To the best of our knowledge, the only directly relevant work is the few-shot text classification approach of [54], which looks into noisy annotations for few-shot relation classification. [54] defines

query-to-support sentence and support samples driven feature-level attention mechanisms. Our work and focus fundamentally differs as we (i) leverage prior knowledge for building conditional attention mechanisms, (ii) estimate sample importance in a query-agnostic way, and (iii) define an experimental protocol for studying the sample noise problem on the mainstream FSL image classification benchmarks MiniImageNet [13] and TieredImageNet [55].

2.2 Semantics based few-shot learning

Semantic FSL refers to the problem variant where supplementary a class-wise knowledge source is made available. Since such additional knowledge often comes from a new data modality, semantic FSL is also sometimes referred as *multi-modal FSL*. There exists only few and recent works on semantic FSL.

In a pioneering work, [36] proposes to define class prototypes as convex combinations of average visual features and transformed semantic priors. The approach is an end-to-end approach meaning that both the feature extractor and the semantic embedding modules are learned within single stage learning. The approach transform semantic embeddings into the visual feature space and combine these two to perform PN based metric learning. Schwartz et al. [37] extends the approach of [36], mainly by introducing multiple semantic priors (label, description or attribute) jointly to obtain richer semantic information. Starting from visual embeddings every semantic embedding is added on top of previous combined embedding as a convex combination.

In Fu et al. [38], visual features are mapped into a semantic space via an encoder, where semantic features, belonging to same class, can be sampled to augment visual features via a decoder model. The approach augment different semantic features either with adding random Gaussian noise or with neighboring of semantic features of different classes. The approach then performs classification by using both real and augmented features. Similarly, Schonfeld et al. [19] aims to learn aligned auto-encoders with reconstruction losses across modalities. In a more recent work, Li et al. [40] uses semantic prior information with a similarity function to obtain margin scores that are used as an additional term in a distance based loss such as [11]. The se-

mantic information is used to determine how similar each class in a few-shot episode and obtains a pairwise margin score that is used in the loss function. Peng et al. [56] uses semantic priors in addition to visual features to obtain two different classification weights for a class. This work obtains visual feature based classification weights by l2-normalizing support image features, averaging them per-class and learning a semantic embedding vector that can be used as a classifier weight maximizing the inner product between the support image based classification weights. Ji et al. [57] proposes a model that re-weights support samples according to a generic weighting function and an auto-encoder that can use either class embeddings or gaussian noise vectors for encoding regularization. While nearly all others can be considered as complimentary to ours, instead of being alternatives, we provide empirical comparisons to semantic FSL works in Chapter 4.

2.3 Zero-shot learning

Zero shot learning aims to build recognition models that can handle classes with no training examples, purely based on prior class semantic knowledge. Mainstream ZSL approaches include learning mappings between the space of visual features and the semantic class representations [42, 43, 44], and, semantics conditional generative models [45, 19, 46, 47, 48]. In our work, we use class semantics to build attention mechanisms in the presence of few training examples, instead of aiming to remove the need for training samples completely as in zero-shot learning.

CHAPTER 3

SEMANTICS-DRIVEN FEATURE AND SAMPLE ATTENTION FOR FEW-SHOT LEARNING

In this chapter, we first provide a formal definition of semantics-driven few-shot learning and summarize the *episodic training* framework, which we embrace in our approach. We then provide a summary of Prototypical Networks (PN) model [11] for few-shot learning and *adaptive cross-modal few-shot learning* (AM3) [36] model that extends PNs by utilizing semantic knowledge to construct model priors. Finally, we present our feature-attention and sample-attention mechanisms and explain how we integrate them into the PN and AM3 models.

3.1 Problem definition and training framework

In this thesis, we focus on few-shot learning of image classification models. The goal is to estimate a new classification model, for a set of target classes ($\mathcal{C} = \{c_1, ..., c_n\}$), based on a limited set of labeled training examples. In our discussion, an n-way classifier is expressed in terms of a scoring function $f(x; \theta)$ that maps the input $x \in \mathcal{X}$ to an n-dimensional vector, according to the model parameters θ. In a standard supervised training problem, a typical way to estimate θ is to find the model parameters minimizing some regularized empirical loss function on the train set.

In the case of few-shot learning, however, the main challenge is to estimate a successful classification model based on a few training samples per class. In most practical cases, it is fundamentally difficult to achieve generalization based on few examples, and, therefore, a model learned by minimizing a generic empirical loss function is unlikely to perform well on novel (test) data. To tackle this problem, the main interest in

meta-learning based FSL is to learn a meta-learner $\xi(\mathcal{D};\beta)$ that can take a new limited set $\mathcal{D}$ of training examples and synthesize the corresponding classification model parameters.

3.2 Episodic training

A popular approach for training meta-learning models is *episodic training* [13]. The main idea is to construct a set or series of few-shot learning tasks and update the meta-learning model based on the regularized empirical loss of meta-learned classification models. More specifically, at each iteration, a new task T is created by sampling a subset $\mathcal{C}_T$ of training classes, and then sampling training $\mathcal{D}_T^{\mathrm{s}}$ and validation $\mathcal{D}_T^{\mathrm{q}}$ samples from the whole training set. Each $\mathcal{D}_T^{\mathrm{s}}$ consists of few-shot training samples of classes $\mathcal{C}_T$ and is commonly referred to as the *support set*. Similarly, each $\mathcal{D}_T^{\mathrm{q}}$ consists of task-specific validation samples and is commonly referred to as the *query set*. The meta-learning, then, is achieved by minimizing the expected empirical loss of meta-learned models over the pairs of support and query sets:

$$\min_{\beta} \mathbb{E}_T \left[\mathbb{E}_{(x,y)\sim\mathcal{D}_T^{\mathrm{q}}} \left[l(f(x;\theta=\xi(\mathcal{D}_T^{\mathrm{s}};\beta)),y) \right] \right] \tag{3.1}$$

where $l(\cdot,y)$ is a classification loss function for label y. No regularization term is shown for brevity.

3.3 Semantic FSL

Arguably, the main premise of meta-learning is to learn domain-specific inductive biases better than what can be provided by general-purpose supervised learning formulations. However, small-sized training sets can be inherently misleading and/or ambiguous due to spurious patterns, as previously illustrated in Figure 1.2. In this respect, prior knowledge about classes can be a crucial source of information towards overcoming the limitations of few-shot training sets. In our work, we presume that semantic knowledge about each class c is a d_e-dimensional vector, represented by ψ_c. The meta-learner can access to semantic class embeddings during both training and testing.

3.3.1 Semantics-driven Attention Mechanisms

We build our semantics-driven attention mechanisms on top of the pioneering semantic few-shot learning model AM3, which is based on prototypical networks. Below we first summarize the AM3 model and then present our approach in its context.

3.3.2 Prototypical networks and AM3

The core idea in PN is to estimate *class prototypes* based on train samples provided in a task and then perform classification based on query-to-prototype similarities. In PNs, a class prototype θ^{PN} for a class c is obtained by averaging the support sample representations:

$$\theta^{\text{PN}}(c) = \mathbb{E}_{\mathcal{D}^{\text{s,c}}}\left[\phi(x)\right] \tag{3.2}$$

where $\phi(x)$ is the feature embedding function parameterized by (a subset of) β that is being trained as part of the PN model. $\phi(x)$ is ResNet12 in our experiments (see Section 4). $\mathcal{D}^{\text{s,c}}$ refers to the subset of examples belonging to class c within a given support set. In the de facto standard PN formulation, the classification score f for class c is given by the negative Euclidean distance between the feature-space embedding of the input and the corresponding class prototype:

$$\left[f(x)\right]_c = -\|\phi(x) - \theta^{\text{PN}}(c)\|^2. \tag{3.3}$$

The parameters β are estimated by minimizing cross entropy loss of query samples over the episodes.

The AM3 model aims to improve the PN model by using semantic knowledge about classes as prototype priors. More specifically, AM3 redefines a class prototype θ^{Prior} as the weighted combination of transformed class semantic embeddings and feature averages:

$$\theta^{\text{Prior}}(c) = \alpha\theta^{\text{PN}}(c) + (1-\alpha)\tau_{\text{Prior}}(\psi_c) \tag{3.4}$$

where τ_{Prior} is a trainable transformation that maps the semantic class embeddings to the space of class prototypes. α is the cross-validated hyperparameter that controls the weight between the original PN prototypes and class semantics based prototypes. τ_{Prior} is parameterized by β.

The AM3 model, therefore, can be interpreted as a way to build classification model priors in the PN formulation. Consistent with the experimental results, such a prior is particularly valuable in the case of one-shot learning, where the training data size is at its extreme minimum. We propose and explore two novel ways to leverage semantic information in a more expressive way towards tackling the inherent difficulties in few-shot learning: (i) *sample attention*, (ii) *feature attention*. Below we provide the details of the proposed mechanisms.

3.3.3 Sample attention

We observe that, in the original PN model, each class prototype is defined as a plain average of support examples. However, the information content of samples can vary greatly due to a number of factors, including background clutter, viewpoint and occlusion. Towards estimating the prototypicality of training samples, we introduce a sample attention mechanism into the final model. Ultimately, we aim to build a model that can estimate the importance of each sample based on the compatibility of samples and prior semantic information. Therefore, we define sample attention module $\eta_{\text{SampleAtt}}(x, c)$ as a function that computes the normalized attention scores of a sample $x \in \mathcal{D}^{\text{s,c}}$ in the context of support sample set $\mathcal{D}^{\text{s,c}}$, conditioned on the class semantic embedding ψ_c:

$$\eta_{\text{SampleAtt}}(\phi(x), c) = \frac{\exp(\gamma_{vis}(\phi(x))^\top \gamma_{sem}(\psi_c))}{\sum_{x' \in \mathcal{D}^{\text{s,c}}} \exp(\gamma_{vis}(\phi(x'))^\top \gamma_{sem}(\psi_c))} \tag{3.5}$$

where γ_{vis} and γ_{sem} are trainable models that are used to obtain visual and semantic feature embeddings. In our experiments, γ_{vis} and γ_{sem} are implemented as MLPs that takes $\phi(x)$ and ψ_c respectively and returns embeddings whose dot products yields the attention scores. We note that $\eta_{\text{SampleAtt}}$ defines a distribution over the support samples, which we use to re-define a class prototype:

$$\theta^{\text{SampleAtt}}(c) = \mathbb{E}_{\eta_{\text{SampleAtt}}(x,c)} \left[\phi(x)\right], \tag{3.6}$$

which amounts to computing the attention-weighted average of sample features. This sample attention mechanism also naturally handles potential sample noise in FSL, which we explore in Chapter 4.

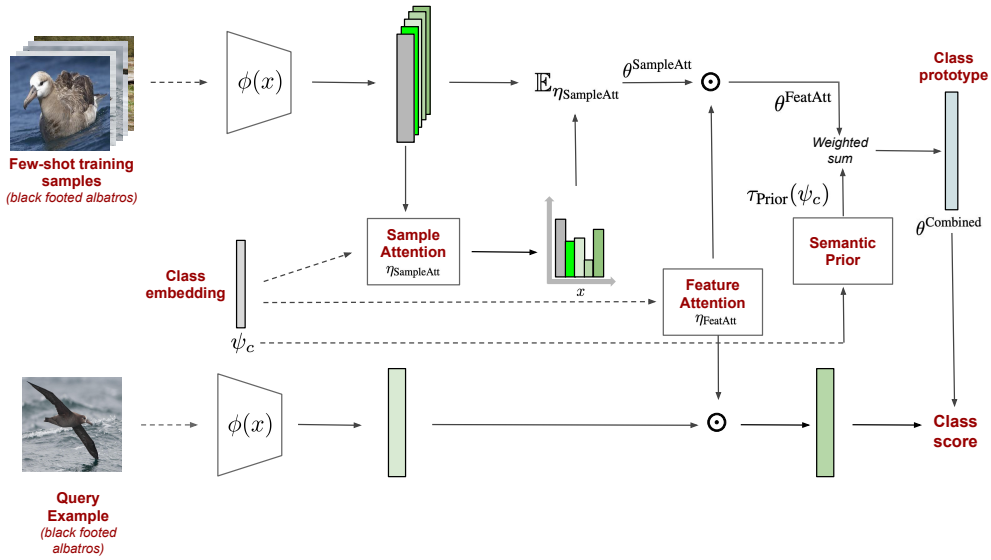

Figure 3.1: Summary of the method, illustrating the process of a new class prototype estimation in 5-shot learning setting. Arrows indicate the input dependencies across model components.

3.3.4 Feature attention

The motivation in feature attention is to tackle the problems resulting from having too few examples. Overall, the goal is enhance or attenuate certain prototype dimensions as a function of semantic class embeddings. For this purpose, we introduce the feature attention function η_{FeatAtt} and re-define the class prototype as follows:

$$\theta^{\text{FeatAtt}}(c) = \eta_{\text{FeatAtt}}(\psi_c) \odot \theta^0(c) \tag{3.7}$$

where $\theta^0(c)$ refers to averaging based class prototypes, which can either be the vanilla PN prototype (θ^{PN}) or the sample attention based prototype ($\theta^{\text{SampleAtt}}$). Here, the trainable feature attention function scales prototype vectors, based on its predictions as a function of class embeddings. Similarly, the feature attention is also applied to the query input, resulting in the following scoring function:

$$[f(x)]_c^{\text{FeatAtt}} = -\|\phi(x) \odot \theta^0(c) - \theta^{\text{FeatAtt}}(c)\|^2. \tag{3.8}$$

3.3.5 The final model

We build our final model by integrating our semantic-driven sample and feature attention mechanisms into the semantics FSL framework defined by the AM3 model. More specifically, in the final *combined attention* model, we first estimate sample attention based class prototypes $\theta^{\text{SampleAtt}}$ and then update them into θ^{FeatAtt} using the feature attention model. We obtain the final class prototypes, which we call θ^{Combined}, by computing the α-weighted combination of the attention-driven prototypes and the pure semantic embedding based prototypes given by $\tau_{\text{Prior}}(\psi_c)$:

$$\theta^{\text{Combined}}(c) = \alpha\eta_{\text{FeatAtt}}(\psi_c) \odot \theta^{\text{SampleAtt}}(c) + (1-\alpha)\tau_{\text{Prior}}(\psi_c). \tag{3.9}$$

The combined attention model is summarized and illustrated in Figure 3.1. The figure represents what happens in our model when a new episode comes. On the top left corner you see support images and that are fed into feature extractor network. Sample attention module extracts softmax probabilities for each training image to determine how improtant each of the individual feature vectors for each class. Then by using class embedding feature attention vectors for each class is obtained to determine how relevant each feature dimension is. Then hadamard product between sample attention and feature attention vectors is applied. Finally semantic prior is transformed into feature space to obtain combined representation that is called a class prototype. This procedure is repeated for each class in the episode. In the bottom it can be seen that when a query image comes its visual feature vector is obtained by using same feature extractor. Then feature attention is applied by using class embeddings and the final representation for the query image is obtained. [1]

[1] Note that since we actually don't know the correct class of a query image we apply feature attention vectors of each class seperately and compare those representations with class prototypes.

CHAPTER 4

EXPERIMENTS

In this chapter, we first present our experimental setup and our implementation details. We then present our experimental results, comparisons and analyses.

4.1 Experimental setup

It is known that implementation details can make a great impact on few-shot learning results, therefore, in most cases, comparing models with different implementation details, such as (batch) normalization schemes, backbones, hyper-parameter tuning strategies and data augmentation schemes, can be greatly misleading [34, 33, 35]. Therefore, we systematically tune the hyper-parameters, including learning rates, number of iterations, and dropout rates on the validation sets for all results that we report based on our own implementation. Below we provide additional details regarding our experimental setup and our efforts to make fair comparisons.

4.1.1 Datasets

We evaluate our model on the MiniImageNet [13] and TieredImageNet [55] datasets. The MiniImageNet dataset is a subset of ImageNet dataset [58] with 100 classes and 600 images per each class. We use the split of [30] for MiniImagenet where the 64, 16 and 20 classes are used as the train, validation and test subsets, respectively. TieredImageNet is a separate benchmark based on [58], with 351, 97, 160 classes for training, validation and testing, respectively. In all our experiments, we use Glove [59] vectors of class names to extract semantic embeddings. Following [36], we use the

Common Crawl version trained on 840B tokens with 300 dimensional embeddings.

4.1.2 Evaluation

We report our test set results over 10000 random tasks and report the average accuracy and 95% confidence interval scores in every table. In all k-shot experiments, we use 15 query examples for each of the n classes in an episode. We execute all of our experiments using the same batch structure for consistency across the experiments.

4.1.3 Backbone architecture

In all models we use the same ResNet-12 architecture as the feature extraction backbone. We use Batch Normalization [60] only in backbone layers, using the *eval* mode [61] for meta-testing [1] to avoid *accidental* transductive setting, which is known to potentially result in misleadingly better few-shot learning results [33].

4.1.4 Backbone pretraining

Following [51] we use supervised pretraining for ResNet-12 backbone and the exact details can be found in [51]. After the pretraining we employ two staged training. In the first stage we train everything except the backbone with learning rate 0.1, momentum 0.9 and weight decay of 0.0005 for 200 epochs for MiniImageNet and 50 epochs for TieredImageNet where an epoch contains 100 episodes. After every epoch we validate our results by using 600 episodes for both MiniImageNet and TieredImageNet. After we complete the 200 epochs and 50 epochs for MiniImageNet and TieredImageNet respectively we train whole model end to end for another 200 epochs. In the second stage the learning rate for backbone is selected to be 0.002 and 0.02 for the rest of the network. After every 40 epoch we halve the learning rates. We use SGD as optimizer for every model and experiment.

[1] meta-testing is the name for testing phase for meta learning based approaches.

4.1.5 PN implementation

Our code base is built upon the PN implementation of [51]. For a fair comparison, therefore, we report the results that we obtain for PN and [51] using the publicly available official source codes for [51]. Similar to [62], we use distance scaling and found that it is better to divide the distances by 32 for MiniImageNet and 16 for TieredImageNet based on the validation results.

4.1.6 AM3 implementation

In our re-implementation of AM3 [36], we have found 0.4 is the best dropout rate for the fully-connected (FC) layers, on the MiniImageNet and TieredImageNet datasets. In our model we use 2 FC layers for encoding of visual and semantic information for sample attention module. These FC layers have the structure of FC-Dropout-ReLU-FC and the dimensions are reduced to 32. The dropout probabilities are selected to be 0.2 and 0.6 for visual and semantic branches respectively by using the validation sets. For feature attention module we use 2-layer FC module with the structure of FC-Softmax-FC and the dimensionality is first reduced to 32 and then increased to the feature dimensionality, which is 640 for ResNet-12. Finally, we also add the semantic prototype branch of AM3 [36] exactly as it is with the only difference of keeping the α a fixed hyper-parameter based on the validation set instead of predicting it from the semantic prior. We utilize the same Euclidean scaling in AM3 as in PN to obtain fair results. Overall, our implementation clearly improves the performance of the baseline AM3 model, compared to the scores originally reported in [36].

4.2 Main results

Below, we first present our main results and ablative studies. We then present our comparisons to the state-of-the-art.

Table 4.1: Evaluation of our sample-attention, feature-attention and combined-attention models with comparisons to our implementations of PN and AM3, in 5-way classification setting.

Model	MiniImageNet		TieredImageNet	
	1-Shot	5-Shot	1-Shot	5-Shot
Results from [36] (ResNet-12 backbone)				
PN [11]	56.52 ± 0.45	74.28 ± 0.20	58.47 ± 0.64	78.41 ± 0.41
AM3 [36]	65.21 ± 0.30	75.20 ± 0.27	67.23 ± 0.34	78.95 ± 0.22
Our implementation (ResNet-12 backbone)				
PN [11]	63.62 ± 0.23	78.37 ± 0.21	67.58 ± 0.22	84.71 ± 0.17
AM3 [36]	67.55 ± 0.24	80.22 ± 0.18	72.60 ± 0.21	84.59 ± 0.18
Sample attention (ours)	67.55 ± 0.24	79.93 ± 0.17	72.60 ± 0.21	85.02 ± 0.16
Feature attention (ours)	69.76 ± 0.21	79.84 ± 0.15	72.69 ± 0.20	84.24 ± 0.16
Combined (ours)	$\mathbf{69.76 \pm 0.21}$	$\mathbf{81.19 \pm 0.18}$	$\mathbf{72.69 \pm 0.20}$	$\mathbf{85.29 \pm 0.17}$

4.2.1 Baselines

Table 4.1 presents our main results for 1-shot or 5-shot, 5-way classification, including baselines that are carefully tuned in the same way to make fair comparisons. First of all, we validate our PN and AM3 baselines. For this purpose, we compare our PN and AM3 results (lower part) to the results reported in the original AM3 work [36] (upper part of Table 4.1). Overall, our results for both baselines are clearly better than the originally reported ones. Noticeably, our PN results on MiniImageNet are higher for nearly 7 and 4 points in 1-shot and 5-shot cases, respectively. Similarly, our AM3 results are higher for nearly 4 and 5 points in 1-shot and 5-shot cases, respectively. We observe even larger improvements for both baselines on TieredImageNet. These results re-highlight the importance of implementation details in FSL and validate the strength of our main baselines. A major factor in obtaining strong baselines is *backbone pretraining*, for which we present additional results in the context of our work later in this chapter.

4.2.2 Main results and ablative experiments

The very last row of Table 4.1 contains our main results using the final combined attention models. The preceding lines presents the results for the ablated versions of the model with only sample attention or only feature attention based AM3 extensions. We note that in the case of 1-shot learning, sample attention has no difference by definition. From the results, first, we observe that sample attention in the case of 5-shot learning slightly degrades by 0.3 points on MiniImageNet but improves on TieredImageNet by nearly 0.5 points. Second, feature attention improves the AM3 model by approximately 2 points for 1-shot and slightly degrades by 0.5 points for 5-shot on MiniImageNet. We observe similar patterns on TieredImageNet.

Looking into the final combined attention model results in Table 4.1, we observe consistent improvements in all cases. The proposed combined attention improves 1-shot learning from 63.62 (PN) and 67.55 (AM3) to 69.76 on MiniImageNet. Similarly, 5-shot results improve from 78.37 (PN) and 80.22 (AM3) to 81.19. We observe similar improvements on TieredImageNet: 1-shot results improve from 67.58 (PN) and 72.60 (AM3) to 72.69, and 5-shot results improve from 84.71 (PN) and 84.59 (AM3) to 85.29.

Noticeably for 5-shot learning on TieredImageNet, the performance gap between PN versus AM3 and our semantic FSL models are relatively smaller. This can be due to the fact that the backbone is pretrained with more diverse classes and this ultimately can give better feature representations with less challenging tasks, therefore reducing the need for semantic priors. Consistently, we also observe that FSL models typically yield higher results on the test set of TieredImageNet, in comparison to MiniImageNet.

4.2.3 Comparison to the state-of-the-art

Comparing results with different implementations is particularly problematic in few-shot learning, and, our main motivation is to explore and evaluate the feature-attention and sample-attention mechanisms in the context of PN and AM3 models. Despite this, it is clearly of interest to show how our results compare to those of other state-

Table 4.2: Few-shot classification accuracy on the test set of MiniImageNet for uni-modal (non-semantic) and multi-modal FSL approaches. * indicates our own implementation.

Model	Backbone	Test Accuracy	
		1-Shot	5-Shot
Uni-modal few-shot learning baselines			
Prototypical Networks* [11]	Resnet-12	63.62 ± 0.23	78.37 ± 0.21
Matching Networks [13]	ResNet-18	52.91 ± 0.91	68.88 ± 0.69
Relation Net [63]	ResNet-18	52.48 ± 0.86	69.83 ± 0.68
MAML [27]	ResNet-18	49.61 ± 0.92	65.72 ± 0.77
LogReg [34]	ResNet-18	51.75 ± 0.80	74.27 ± 0.63
LogReg Cosine [34]	ResNet-18	51.87 ± 0.77	75.68 ± 0.63
TADAM [62]	ResNet-12	58.56 ± 0.39	76.65 ± 0.35
SimpleShot [64]	ResNet-18	62.85 ± 0.20	80.02 ± 0.14
MetaOptNet [65]	ResNet-12	62.64 ± 0.61	78.63 ± 0.46
LEO [29]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12
FEAT* [51]	ResNet-12	65.38 ± 0.20	77.79 ± 0.15
Multi-modal few-shot learning baselines			
KTN [56]	Conv-128	64.42 ± 0.72	74.16 ± 0.56
RIN [57]	ResNet-12	56.92 ± 0.81	75.62 ± 0.62
TriNet [38]	ResNet-18	58.12 ± 1.37	76.92 ± 0.69
ACAM [66]	ResNet-12	66.43 ± 0.57	75.74 ± 0.48
Baby steps [37]	ResNet-10	67.2	74.8
AM3 + TRAML [40]	ResNet-12	67.10 ± 0.52	79.54 ± 0.60
AM3* [36]	ResNet-12	67.55 ± 0.24	80.22 ± 0.18
Ours	ResNet-12	$\mathbf{69.76 \pm 0.21}$	$\mathbf{81.19 \pm 0.18}$

Table 4.3: Few-shot classification accuracy on the test set of TieredImageNet for uni-modal (non-semantic) and multi-modal FSL approaches. * indicates our own implementation.

Model	Backbone	Test Accuracy	
		1-Shot	5-Shot
Uni-modal few-shot learning baselines			
Prototypical Networks* [11]	Resnet-12	67.58 ± 0.22	84.71 ± 0.19
Relation Net [63]	ResNet-12	54.48 ± 0.93	71.32 ± 0.78
MAML [27]	ResNet-12	51.67 ± 1.81	70.30 ± 0.08
MetaOptNet [65]	ResNet-12	65.99 ± 0.72	81.56 ± 0.63
SimpleShot [64]	ResNet-18	69.09 ± 0.22	84.58 ± 0.16
LEO [29]	ResNet-12	66.33 ± 0.05	81.44 ± 0.09
FEAT* [51]	ResNet-12	70.53 ± 0.22	84.71 ± 0.15
Multi-modal few-shot learning baselines			
ACAM [66]	ResNet-12	67.89 ± 0.69	79.23 ± 0.52
AM3* [36]	ResNet-12	72.60 ± 0.21	84.59 ± 0.15
Ours	ResNet-12	$\mathbf{72.69 \pm 0.21}$	$\mathbf{85.24 \pm 0.16}$

of-the-art methods. Table 4.2 and Table 4.3 present the results of FSL methods (upper half) and semantic FSL methods (lower half), on MiniImageNet and TieredImageNet datasets respectively. First, we observe that our modernized PN baselines are very strong baselines, outperforming the results of many other more recent works. Second, our final model outperforms all methods in FSL and semantic FSL categories on both MiniImageNet and TieredImageNet datasets. Not surprisingly, best improvements are in the 1-shot settings on both datasets against the uni-modal (non-semantic) approaches since a single training sample is often insufficient and semantic prior becomes most valuable in this setting. Overall, the results highlight the value of semantic priming for few-shot learning.

Table 4.4: Evaluation of sample attention for handling noisy support samples.

Method	5-way 5-Shot Acc.
PN [11]	62.61 +- 0.22
PN + Sample Attention	70.48 +- 0.20

4.2.4 Few-shot learning with noisy samples

As discussed in Section 1, we believe that the problem of having sample noise in FSL is an insufficiently studied yet practically important problem. This problem also provides a challenging scenario for understanding the effectiveness of semantics driven sample attention. For this purpose, we artificially introduce noisy support batches during training and testing phases. More specifically, for the 5-shot 5-way setting on MiniImageNet, we create episodes with classes that may consist of support examples that belong to other classes in that episode. We note that such within-task noise is most challenging and practically more relevant as FSL models are likely to be constructed among similar classes. For each class in a task, we guarantee the existence of at least 3 correctly labeled support examples. We then randomly permute the classes to mix up the correct support-class pairs and we add 2 more instances to complete 5 support examples. To study sample attention in an isolated manner, here we exceptionally use sample attention directly on top of PN. We report the results in the Table 4.4, where we observe a very clear performance gap, i.e., 62.61 (PN) versus 70.48 (sample attention). This strongly suggests the ability of sample attention to utilize semantic priors for selecting the most informative support samples.

4.3 Analysis

In the paragraphs that follow, we present additional analyses on the importance of backbone pretraining, sample attention in the cases of noisy and clean support samples, and larger-way (10-way and 15-way) FSL.

Table 4.5: Comparison of PN and Our approach with and without pretraining of the ResNet-12 backbone for MiniImageNet.

Model	MiniImageNet			
	1-Shot Val	1-Shot Test	5-Shot Val	5-Shot Test
PN w/o Pretraining [11]	43.43 ± 0.67	43.70 ± 0.24	71.90 ± 0.66	69.67 ± 0.22
Ours w/o Pretraining	64.55 ± 0.70	61.57 ± 0.23	72.46 ± 0.68	68.78 ± 0.19
PN with Pretraining [11]	68.73 ± 0.68	63.62 ± 0.20	80.88 ± 0.67	78.37 ± 0.22
Ours with Pretraining	$\mathbf{76.44 \pm 0.65}$	$\mathbf{69.76 \pm 0.21}$	$\mathbf{85.02 \pm 0.65}$	$\mathbf{81.19 \pm 0.18}$

4.3.1 Importance of backbone pretraining

In Table 4.5 we inspect how pretraining and backbone quality affects the few-shot learning performance. Here we use both PN and our approach to see the effect in both uni-modal and multi-modal approaches. Models without pretraining are models that are trained end-to-end. In order to create a fair comparison, we train the models by using the same number of epochs in both cases. All of the models are trained for 400 epochs and the results reported based on best validation score. As can be seen from Table 4.5, model pretraining improves both approaches clearly, and is crucial for achieving state-of-the-art performance. An interesting result is the comparison of PN and our approach for 1-shot setting without pretraining since it highlights the significance of semantic information when the backbone quality/visual feature quality is low. As the backbone becomes better with pretraining the gap becomes smaller.

4.3.2 Analysis of sample attention

To further understand the behavior of sample attention, we provide qualitative results both with and without noise in support samples, in Figure 4.1. Each image 5-tuple corresponds to a 5-shot support set of a class. The examples on the left correspond to clean setting and the examples on the right come from sample noise experiments, where dashed borders indicate noisy samples. The numbers indicate the resulting attention values. In the top-left example, we observe that the model puts higher weights to support examples where the target *lion* is most recognizable. Similarly

Figure 4.1: Sample attention examples for clean (left) and noisy (right) support sample settings. Yellow dashed lines represent the noisy samples. Numbers indicate attention scores.

in the bottom-left example, we observe that the model attends to samples where the target *ant* is more recognizable. The example on the top-right shows that the model puts low weight to noisy and cluttered inputs. The bottom-right example, however, shows an example where the weight of the noisy sample *dalmatian dog* is comparable to the others, due to similarity across the target class and noisy samples, highlighting a limitation of the model.

Additionally, Figure 4.2 provides a histogram of sample attention weights over 100 5-way, 5-shot tasks on the test set of MiniImageNet. The figure shows that the model estimates sample attention weights with a unimodal distribution centered around 0.2 weight with values as low as 0.125 and as high as 2.8. Noting that this is the result of purely meta-learned model in a 5-shot setting, the distribution suggest that we obtain a *healthy* and effective sample importance estimator.

4.3.3 10-way and 15-way few-shot learning

In Table 4.6 and 4.7 we compare our model with PN and AM3 for the more challenging 10-way and 15-way settings respectively. As expected, the performances of all models decrease as the way count increases. Our model outperforms both PN and AM3 for the 1-shot setting, in both 10-way and 15-way experiments and the performance margin between those models increases. For 5-shot settings, although our approach obtains nearly 1% better accuracy on validation sets it falls 0.5 points behind of AM3 in test sets. This points to a less than ideal correlation between the validation

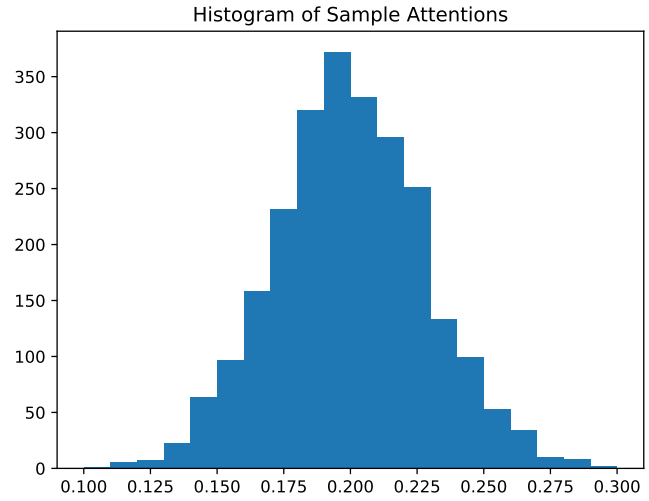

Figure 4.2: Histogram of sample attention weights for 100 5-way, 5-shot tasks on the test set of MiniImageNet.

Table 4.6: Evaluation of 10-way FSL on MiniImageNet.

Method	10-way			
	1-Shot Val	1-Shot Test	5-Shot Val	5-Shot Test
PN [11]	53.61 ± 0.60	46.04 ± 0.13	70.30 ± 0.39	66.03 ± 0.10
AM3 [36]	61.46 ± 0.55	51.37 ± 0.12	72.92 ± 0.43	$\mathbf{67.51 \pm 0.10}$
Ours	$\mathbf{62.63 \pm 0.50}$	$\mathbf{52.22 \pm 0.12}$	$\mathbf{73.76 \pm 0.41}$	67.36 ± 0.10

and test set performances, which is likely to degrade as the way count increases while test set size remaining constant.

Table 4.7: Evaluation of 15-way FSL on MiniImageNet.

Method	15-way			
	1-Shot Val	1-Shot Test	5-Shot Val	5-Shot Test
PN [11]	45.64 $\pm$ 0.43	37.42 $\pm$ 0.10	63.47 $\pm$ 0.28	58.40 $\pm$ 0.08
AM3 [36]	49.23 $\pm$ 0.44	39.10 $\pm$ 0.10	66.06 $\pm$ 0.30	**60.27** $\pm$ **0.08**
Ours	**55.24** $\pm$ **0.35**	**43.54** $\pm$ **0.09**	**66.91** $\pm$ **0.28**	59.72 $\pm$0.08

CHAPTER 5

CONCLUSION

In this thesis, we propose a method for semantic FSL. Although state-of-the-art FSL approaches perform surprisingly well even at the extreme cases like 1-shot learning the need for a supplementary semantic side information is undeniable. Especially considering the sophisticated big architectures like ResNets are commonly used and requires a lot of data to train. The main motivation of our approach is to utilize semantic prior knowledge about classes to estimate importance of support samples and representation dimensions.

In Chapter 3 we show that how we can use semantic side information to weigh limited support images so that we will benefit from the most useful ones compared to previous approaches where they simply take the mean of those support image features. In addition, we are also trying to select most useful feature dimensions for each class to boost the classification accuracy. Finally, we point out another potential problem that we call sample noise and suggest that sample attention is very helpful in reducing the effects of irrelevant images.

Our method performs well against not only the approaches that use only visual data but also the ones that use additional semantic information. The performance gains are more prominent in lower shot settings where model's need for auxiliary semantic information is higher. In addition, when the FSL tasks get harder and more classes are being involved in a single episode our approach shows more promise than other works. We also study the case of sample noise in support sets.

While we focus on a single semantic modality, we believe that incorporating multiple modalities, as in [37], and combining with a generative model based approach, as in

[38], can be important future work directions. Additionally, having better semantic side information or incorporating that side information in a better way may be another cruical direction that can be explored in the future.

5.1 Limitations

This work clearly benefits from the usage of semantic side information to obtain better few-shot learning results. However, side informations may also negatively affect the classification performance, especially if the semantic side information itself is misleading and/or noisy. In addition, we may not be able to find semantic side information for certain datasets or those side information may not be sufficient to find useful attention weights.

REFERENCES

[1] E. Bart and S. Ullman, "Cross-generalization: learning novel classes from a single example by feature replacement," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 672–679 vol. 1, 2005.

[2] M. Fink, "Object classification from a single example utilizing class relevance metrics," in *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pp. 449–456, 2004.

[3] Li Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[4] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," *Cognitive Science*, vol. 33, 2011.

[5] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, and H. Larochelle, "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *International Conference on Learning Representations*, 2020.

[6] B. Landau, L. B. Smith, and S. S. Jones, "The importance of shape in early lexical learning," *Cognitive Development*, vol. 3, no. 3, pp. 299 – 321, 1988.

[7] W. E. Merriman, "Categorization and naming in children: Problems of induction. ellen markman. cambridge, ma: Mit press, 1989. pp. 250.," *Applied Psycholinguistics*, vol. 12, no. 3, p. 385–392, 1991.

[8] L. Smith and L. Slone, "A developmental approach to machine learning?," *Frontiers in Psychology*, vol. 8, 2017.

[9] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-Scale Long-Tailed Recognition in an Open World," (Long Beach, CA, USA), pp. 2532–2541, IEEE, June 2019.

[10] W. J. Reed, "The pareto, zipf and other power laws," *Economics letters*, vol. 74, no. 1, pp. 15–19, 2001.

[11] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[12] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

[13] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[14] G. R. Koch, "Siamese neural networks for one-shot image recognition," 2015.

[15] B. Hariharan and R. Girshick, "Low-shot Visual Recognition by Shrinking and Hallucinating Features," in *ICCV*, 2017.

[16] Y.-X. Wang, R. Girshick, M. Herbert, and B. Hariharan, "Low-shot learning from imaginary data," in *CVPR*, 2018.

[17] H. Gao, Z. Shou, A. Zareian, H. Zhang, and S.-F. Chang, "Low-shot Learning via Covariance-Preserving Adversarial Augmentation Networks," in *Advances in Neural Information Processing Systems 31*, pp. 983–993, 2018.

[18] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, and A. Bronstein, "Delta-encoder: an effective sample synthesis method for few-shot object recognition," in *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 2845–2855, 2018.

[19] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, CA, USA), pp. 8239–8247, June 2019.

[20] A. Antoniou, A. Storkey, and H. Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv:1711.04340 [cs, stat]*, Nov. 2017.

[21] K. Ridgeway and M. C. Mozer, "Open-Ended Content-Style Recombination Via Leakage Filtering," *arXiv e-prints*, p. arXiv:1810.00110, Sept. 2018.

[22] Z. Chen, Y. Fu, K. Chen, and Y.-G. Jiang, "Image Block Augmentation for One-Shot Learning," in *AAAI*, p. 8, 2019.

[23] T. DeVries and G. W. Taylor, "Dataset Augmentation in Feature Space," *arXiv:1702.05538 [cs, stat]*, Feb. 2017.

[24] S. Gidaris and N. Komodakis, "Dynamic Few-Shot Visual Learning Without Forgetting," in *CVPR*, p. 9, 2018.

[25] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *CVPR*, 2018.

[27] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135, PMLR, 06–11 Aug 2017.

[28] A. Nichol, J. Achiam, and J. Schulman, "On First-Order Meta-Learning Algorithms," *arXiv e-prints*, p. arXiv:1803.02999, Mar. 2018.

[29] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *International Conference on Learning Representations*, 2019.

[30] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *In International Conference on Learning Representations (ICLR)*, 2017.

[31] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1842–1850, PMLR, 20–22 Jun 2016.

[32] S. Sukhbaatar, J. Weston, R. Fergus, *et al.*, "End-to-end memory networks," in *Advances in neural information processing systems*, pp. 2440–2448, 2015.

[33] J. Bronskill, J. Gordon, J. Requeima, S. Nowozin, and R. E. Turner, "TaskNorm: Rethinking Batch Normalization for Meta-Learning," in *Machine Learning and Systems*, 4 2020.

[34] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2019.

[35] Y. Chen, X. Wang, Z. Liu, H. Xu, and T. Darrell, "A new meta-baseline for few-shot learning," *ArXiv*, vol. abs/2003.04390, 2020.

[36] C. Xing, N. Rostamzadeh, B. N. Oreshkin, and P. H. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *NeurIPS*, 2019.

[37] E. Schwartz, L. Karlinsky, R. Feris, R. Giryes, and A. M. Bronstein, "Baby steps towards few-shot learning with multiple semantics," *arXiv e-prints*, p. arXiv:1906.01905, June 2019.

[38] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal, "Multi-Level Semantic Feature Augmentation for One-Shot Learning," *IEEE Transactions on Image Processing*, vol. 28, pp. 4594–4605, Sept. 2019.

[39] J. Mu, P. Liang, and N. Goodman, "Shaping Visual Representations with Language for Few-shot Classification," *arXiv e-prints*, p. arXiv:1911.02683, Nov. 2019.

[40] A. Li, W. Huang, X. Lan, J. Feng, Z. Li, and L. Wang, "Boosting few-shot learning with adaptive margin loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] E. H. ROSCH, "On the internal structure of perceptual and semantic categories," in *Cognitive Development and Acquisition of Language* (T. E. Moore, ed.), pp. 111 – 144, 1973.

[42] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013.

[43] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *ICLR*, 2014.

[44] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3010–3019, 2017.

[45] L. Chen, H. Zhang, J. Xiao, W. Liu, and S. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1043–1052, 2018.

[46] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10275–10284, 2019.

[47] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2188–2196, 2018.

[48] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1004–1013, 2018.

[49] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to Learn Quickly for Few-Shot Learning," *arXiv e-prints*, p. arXiv:1707.09835, July 2017.

[50] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need?," in *ECCV*, 2020.

[51] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8808–8817, 2020.

[52] R. ZHANG, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, "Metagan: An adversarial approach to few-shot learning," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, pp. 2365–2374, Curran Associates, Inc., 2018.

[53] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," in *Advances in Neural Information Processing Systems*, 2014.

[54] T. Gao, X. Han, Z. Liu, and M. Sun, "Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification," in *AAAI*, July 2019.

[55] M. Ren, S. Ravi, E. Triantafillou, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *International Conference on Learning Representations*, 2018.

[56] Z. Peng, Z. Li, J. Zhang, Y. Li, G. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 441–449, 2019.

[57] Z. Ji, X. Chai, Y. Yu, and Z. Zhang, "Reweighting and information-guidance networks for Few-Shot Learning," *Neurocomputing*, vol. 423, pp. 13–23, Jan. 2021.

[58] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[59] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[60] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, PMLR, 07–09 Jul 2015.

[61] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[62] B. N. Oreshkin, P. R. Lopez, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, 2018.

[63] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning," in *CVPR*, 2018.

[64] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning," *arXiv e-prints*, p. arXiv:1911.04623, Nov. 2019.

[65] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, 2019.

[66] J. Lian, H. Wang, and S. Xiong, "Learning class prototypes via anisotropic combination of aligned modalities for few-shot learning," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020.