Adaptive Cross-Modal Few-shot Learning

Chen Xing*
College of Computer Science,
Nankai University, Tianjin, China
Element AI, Montreal, Canada

Negar Rostamzadeh
Element AI, Montreal, Canada

Boris N. Oreshkin
Element AI, Montreal, Canada

Pedro O. Pinheiro
Element AI, Montreal, Canada

Abstract

Metric-based meta-learning techniques have successfully been applied to few-shot classification problems. In this paper, we propose to leverage cross-modal information to enhance metric-based few-shot learning methods. Visual and semantic feature spaces have different structures by definition. For certain concepts, visual features might be richer and more discriminative than text ones. While for others, the inverse might be true. Moreover, when the support from visual information is limited in image classification, semantic representations (learned from unsupervised text corpora) can provide strong prior knowledge and context to help learning. Based on these two intuitions, we propose a mechanism that can adaptively combine information from both modalities according to new image categories to be learned. Through a series of experiments, we show that by this adaptive combination of the two modalities, our model outperforms current uni-modality few-shot learning methods and modality-alignment methods by a large margin on all benchmarks and few-shot scenarios tested. Experiments also show that our model can effectively adjust its focus on the two modalities. The improvement in performance is particularly large when the number of shots is very small.

1 Introduction

Deep learning methods have achieved major advances in areas such as speech, language and vision [25]. These systems, however, usually require a large amount of labeled data, which can be impractical or expensive to acquire. Limited labeled data lead to overfitting and generalization issues in classical deep learning approaches. On the other hand, existing evidence suggests that human visual system is capable of effectively operating in small data regime: humans can learn new concepts from a very few samples, by leveraging prior knowledge and context [23, 30, 46]. The problem of learning new concepts with small number of labeled data points is usually referred to as *few-shot learning* [1, 6, 27, 22] (FSL).

Most approaches addressing few-shot learning are based on *meta-learning* paradigm [43, 3, 52, 13], a class of algorithms and models focusing on learning how to (quickly) learn new concepts. Meta-learning approaches work by learning a parameterized function that embeds a variety of learning tasks and can generalize to new ones. Recent progress in few-shot image classification has primarily been made in the context of unimodal learning. In contrast to this, employing data from another modality can help when the data in the original modality is limited. For example, strong evidence supports the hypothesis that language helps recognizing new visual objects in toddlers [15, 45]. This

---

*Work done when interning at Element AI. Contact through: xingchen1113@gmail.com

Figure 1: Concepts have different visual and semantic feature space. *(Left)* Some categories may have similar visual features and dissimilar semantic features. (*Right*) Other can possess same semantic label but very distinct visual features. Our method adaptively exploits both modalities to improve classification performance in low-shot regime.

suggests that semantic features from text can be a powerful source of information in the context of few-shot image classification.

Exploiting auxiliary modality (*e.g.*, attributes, unlabeled text corpora) to help image classification when data from visual modality is limited, have been mostly driven by *zero-shot learning* [24, 36] (ZSL). ZSL aims at recognizing categories whose instances have not been seen during training. In contrast to few-shot learning, there is no small number of labeled samples from the original modality to help recognize new categories. Therefore, most approaches consist of aligning the two modalities during training. Through this *modality-alignment*, the modalities are mapped together and forced to have the same semantic structure. This way, knowledge from auxiliary modality is transferred to the visual side for new categories at test time [9].

However, visual and semantic feature spaces have heterogeneous structures by definition. For certain concepts, visual features might be richer and more discriminative than text ones. While for others, the inverse might be true. Figure 1 illustrates this remark. Moreover, when the number of support images from visual side is very small, information provided from this modality tend to be noisy and local. On the contrary, semantic representations (learned from large unsupervised text corpora) can act as more general prior knowledge and context to help learning. Therefore, instead of aligning the two modalities (to transfer knowledge to the visual modality), for few-shot learning in which information are provided from both modalities during test, it is better to treat them as two independent knowledge sources and adaptively exploit both modalities according to different scenarios. Towards this end, we propose *Adaptive Modality Mixture Mechanism* (AM3), an approach that adaptively and selectively combines information from two modalities, visual and semantic, for few-shot learning.

AM3 is built on top of metric-based meta-learning approaches. These approaches perform classification by comparing distances in a learned metric space (from visual data). On the top of that, our method also leverages text information to improve classification accuracy. AM3 performs classification in an adaptive convex combination of the two distinctive representation spaces with respect to image categories. With this mechanism, AM3 can leverage the benefits from both spaces and adjust its focus accordingly. For cases like Figure 1(Left), AM3 focuses more on the semantic modality to obtain general context information. While for cases like Figure 1(Right), AM3 focuses more on the visual modality to capture rich local visual details to learn new concepts.

Our main contributions can be summarized as follows: (i) we propose adaptive modality mixture mechanism (AM3) for cross-modal few-shot classification. AM3 adapts to few-shot learning better than modality-alignment methods by adaptively mixing the semantic structures of the two modalities. (ii) We show that our method achieves considerable boost in performance over different metric-based meta-learning approaches. (iii) AM3 outperforms by a considerable margin current (single-modality and cross-modality) state of the art in few-shot classification on different datasets and different number of shots. (iv) We perform quantitative investigations to verify that our model can effectively adjust its focus on the two modalities according to different scenarios.

2   Related Work

Few-shot learning.   Meta-learning has a prominent history in machine learning [43, 3, 52]. Due to advances in representation learning methods [11] and the creation of new few-shot learning datasets [22, 53], many deep meta-learning approaches have been applied to address the few-shot learning problem . These methods can be roughly divided into two main types: metric-based and gradient-based approaches.

Metric-based approaches aim at learning representations that minimize intra-class distances while maximizing the distance between different classes. These approaches rely on an episodic training

2

framework: the model is trained with sub-tasks (episodes) in which there are only a few training samples for each category. For example, matching networks [53] follows a simple nearest neighbour framework. In each episode, it uses an attention mechanism (over the encoded support) as a similarity measure for one-shot classification.

In prototypical networks [47], a metric space is learned where embeddings of queries of one category are close to the centroid (or prototype) of supports of the same category, and far away from centroids of other classes in the episode. Due to the simplicity and good performance of this approach, many methods extended this work. For instance, Ren *et al.* [39] propose a semi-supervised few-shot learning approach and show that leveraging unlabeled samples outperform purely supervised prototypical networks. Wang et al. [54] propose to augment the support set by generating hallucinated examples. Task-dependent adaptive metric (TADAM) [35] relies on conditional batch normalization [5] to provide task adaptation (based on task representations encoded by visual features) to learn a task-dependent metric space.

Gradient-based meta-learning methods aim at training models that can generalize well to new tasks with only a few fine-tuning updates. Most these methods are built on top of model-agnostic meta-learning (MAML) framework [7]. Given the universality of MAML, many follow-up works were recently proposed to improve its performance on few-shot learning [33, 21]. Kim *et al.* [18] and *Finn et al.* [8] propose a probabilistic extension to MAML trained with variational approximation. Conditional class-aware meta-learning (CAML) [16] conditionally transforms embeddings based on a metric space that is trained with prototypical networks to capture inter-class dependencies. Latent embedding optimization (LEO) [41] aims to tackle MAML's problem of only using a few updates on a low data regime to train models in a high dimensional parameter space. The model employs a low-dimensional latent model embedding space for update and then decodes the actual model parameters from the low-dimensional latent representations. This simple yet powerful approach achieves current state of the art result in different few-shot classification benchmarks. Other meta-learning approaches for few-shot learning include using memory architecture to either store exemplar training samples [42] or to directly encode fast adaptation algorithm [38]. Mishra et al. [32] use temporal convolution to achieve the same goal.

Current approaches mentioned above rely solely on visual features for few-shot classification. Our contribution is orthogonal to current metric-based approaches and can be integrated into them to boost performance in few-shot image classification.

Zero-shot learning.   Current ZSL methods rely mostly on visual-auxiliary modality alignment [9, 58]. In these methods, samples for the same class from the two modalities are mapped together so that the two modalities obtain the same semantic structure. There are three main families of modality alignment methods: representation space alignment, representation distribution alignment and data synthetic alignment.

Representation space alignment methods either map the visual representation space to the semantic representation space [34, 48, 9], or map the semantic space to the visual space [59]. Distribution alignment methods focus on making the alignment of the two modalities more robust and balanced to unseen data [44]. ReViSE [14] minimizes maximum mean discrepancy (MMD) of the distributions of the two representation spaces to align them. CADA-VAE [44] uses two VAEs [19] to embed information for both modalities and align the distribution of the two latent spaces. Data synthetic methods rely on generative models to generate image or image feature as data augmentation [60, 57, 31, 54] for unseen data to train the mapping function for more robust alignment.

ZSL does not have access to any visual information when learning new concepts. Therefore, ZSL models have no choice but to align the two modalities. This way, during test the image query can be directly compared to auxiliary information for classification [59]. Few-shot learning, on the other hand, has access to a small amount of support images in the original modality during test. This makes alignment methods from ZSL seem unnecessary and too rigid for FSL. For few-shot learning, it would be better if we could preserve the distinct structures of both modalities and adaptively combine them for classification according to different scenarios. In Section 4 we show that by doing so, AM3 outperforms directly applying modality alignment methods for few-shot learning by a large margin.

3   Method

In this section, we explain how AM3 adaptively leverages text data to improve few-shot image classification. We start with a brief explanation of episodic training for few-shot learning and a

summary of prototypical networks followed by the description of the proposed adaptive modality mixture mechanism.

3.1 Preliminaries

3.1.1 Episodic Training

Few-shot learning models are trained on a labeled dataset $\mathcal{D}_{\text{train}}$ and tested on $\mathcal{D}_{\text{test}}$. The class sets are disjoint between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. The test set has only a few labeled samples per category. Most successful approaches rely on an *episodic* training paradigm: the few shot regime faced at test time is simulated by sampling small samples from the large labeled set $\mathcal{D}_{\text{train}}$ during training.

In general, models are trained on K-shot, N-way episodes. Each episode e is created by first sampling N categories from the training set and then sampling two sets of images from these categories: (i) the *support* set $\mathcal{S}_e = \{(s_i, y_i)\}_{i=1}^{N \times K}$ containing K examples for each of the N categories and (ii) the *query* set $\mathcal{Q}_e = \{(q_j, y_j)\}_{j=1}^{Q}$ containing different examples from the same N categories.

The episodic training for few-shot classification is achieved by minimizing, for each episode, the loss of the prediction on samples in query set, given the support set. The model is a parameterized function and the loss is the negative loglikelihood of the true class of each query sample:

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(\mathcal{S}_e, \mathcal{Q}_e)} \; -\sum_{t=1}^{Q_e} \log p_\theta(y_t | q_t, \mathcal{S}_e) \; , \tag{1}$$

where $(q_t, y_t) \in \mathcal{Q}_e$ and $\mathcal{S}_e$ are, respectively, the sampled query and support set at episode e and θ are the parameters of the model.

3.1.2 Prototypical Networks

We build our model on top of metric-based meta-learning methods. We choose prototypical network [47] for explaining our model due to its simplicity. We note, however, that the proposed method can potentially be applied to any metric-based approach.

Prototypical networks use the support set to compute a centroid (prototype) for each category (in the sampled episode) and query samples are classified based on the distance to each prototype. The model is a convolutional neural network [26] $f : \mathbb{R}^{n_v} \to \mathbb{R}^{n_p}$, parameterized by θ_f, that learns a n_p-dimensional space where samples of the same category are close and those of different categories are far apart.

For every episode e, each embedding prototype p_c (of category c) is computed by averaging the embeddings of all support samples of class c:

$$\mathbf{p}_c = \frac{1}{|S_e^c|} \sum_{(s_i, y_i) \in \mathcal{S}_e^c} f(s_i) \; , \tag{2}$$

where $\mathcal{S}_e^c \subset \mathcal{S}_e$ is the subset of support belonging to class c.

The model produces a distribution over the N categories of the episode based on a softmax [4] over (negative) distances d of the embedding of the query q_t (from category c) to the embedded prototypes:

$$p(y = c | q_t, S_e, \theta) = \frac{\exp(-d(f(q_t), \mathbf{p}_c))}{\sum_k \exp(-d(f(q_t), \mathbf{p}_k))} \; . \tag{3}$$

We consider d to be the Euclidean distance. The model is trained by minimizing Equation 1 and the parameters are updated with stochastic gradient descent.

3.2 Adaptive Modality Mixture Mechanism

The information contained in semantic concepts can significantly differ from visual contents. For instance, 'Siberian husky' and 'wolf', or 'komondor' and 'mop', might be difficult to discriminate with visual features, but might be easier to discriminate with language semantic features.
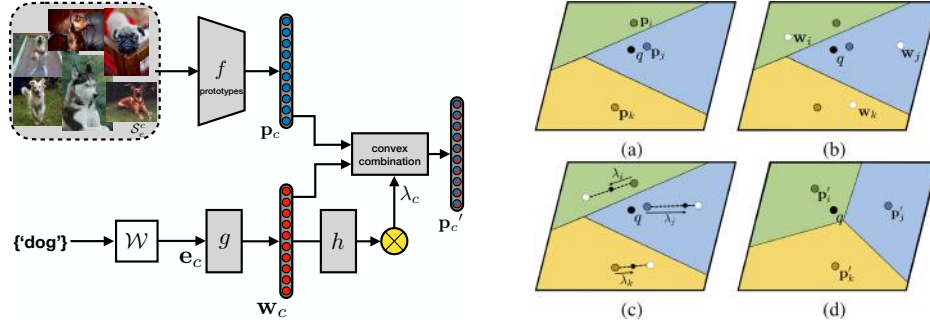
Figure 2: *(Left)* Adaptive modality mixture model. The final category prototype is a convex combination of the visual and the semantic feature representations. The mixing coefficient is conditioned on the semantic label embedding. *(Right)* Qualitative example of how AM3 works. Assume query sample q has category i. (a) The closest visual prototype to the query sample q is $\mathbf{p}_j$. (b) The semantic prototypes. (c) The mixture mechanism modify the positions of the prototypes, given the semantic embeddings. (d) After the update, the closest prototype to the query is now the one of the category i, correcting the classification.

In zero-shot learning, where no visual information is given at test time (that is, the support set is void), algorithms need to solely rely on an auxiliary (*e.g.*, text) modality. On the other extreme, when the number of labeled image samples is large, neural network models tend to ignore the auxiliary modality as it is able to generalize well with large number of samples [20].

Few-shot learning scenario fits in between these two extremes. Thus, we hypothesize that both visual and semantic information can be useful for few-shot learning. Moreover, given that visual and semantic spaces have different structures, it is desirable that the proposed model exploits both modalities adaptively, given different scenarios. For example, when it meets objects like 'ping-pong balls' which has many visually similar counterparts, or when the number of shots is very small from the visual side, it relies more on text modality to distinguish them.

In AM3, we augment metric-based FSL methods to incorporate language structure learned by a word-embedding model $\mathcal{W}$ (pre-trained on unsupervised large text corpora), containing label embeddings of all categories in $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}$. In our model, we modify the prototype representation of each category by taking into account their label embeddings.

More specifically, we model the new prototype representation as a convex combination of the two modalities. That is, for each category c, the new prototype is computed as:

$$\mathbf{p}'_c = \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c , \qquad (4)$$

where λ_c is the *adaptive mixture coefficient* (conditioned on the category) and $\mathbf{w}_c = g(\mathbf{e}_c)$ is a transformed version of the label embedding for class c. The representation $\mathbf{e}_c$ is the pre-trained word embedding of label c from $\mathcal{W}$. This transformation $g : \mathbb{R}^{n_w} \to \mathbb{R}^{n_p}$, parameterized by θ_g, is important to guarantee that both modalities lie on the space $\mathbb{R}^{n_p}$ of the same dimension and can be combined. The coefficient λ_c is conditioned on category and calculated as follows:

$$\lambda_c = \frac{1}{1 + \exp(-h(\mathbf{w}_c))} , \qquad (5)$$

where h is the adaptive mixing network, with parameters θ_h. Figure 2(left) illustrates the proposed model. The mixing coefficient λ_c can be conditioned on different variables. In Appendix F we show how performance changes when the mixing coefficient is conditioned on different variables.

The training procedure is similar to that of the original prototypical networks. However, the distances d (used to calculate the distribution over classes for every image query) are between the query and the cross-modal prototype $\mathbf{p}'_c$:

$$p_\theta(y = c | q_t, S_e, \mathcal{W}) = \frac{\exp(-d(f(q_t), \mathbf{p}'_c))}{\sum_k \exp(-d(f(q_t), \mathbf{p}'_k))} , \qquad (6)$$

where $\theta = \{\theta_f, \theta_g, \theta_h\}$ is the set of parameters. Once again, the model is trained by minimizing Equation 1. Note that in this case the probability is also conditioned on the word embeddings $\mathcal{W}$.

Figure 2(right) illustrates an example on how the proposed method works. Algorithm 1, on supplementary material, shows the pseudocode for calculating the episode loss. We chose prototypical network [47] for explaining our model due to its simplicity. We note, however, that AM3 can potentially be applied to any metric-based approach that calculates prototypical embeddings $\mathbf{p}_c$ for categories. As shown in next section, we apply AM3 on both ProtoNets and TADAM [35]. TADAM is a task-dependent metric-based few-shot learning method, which currently performs the best among all metric-based FSL methods.

4 Experiments

In this section we compare our model, AM3, with three different types of baselines: uni-modality few-shot learning methods, modality-alignment methods and metric-based extensions of modality-alignment methods. We show that AM3 outperforms the state of the art of each family of baselines. We also verify the adaptiveness of AM3 through quantitative analysis.

4.1 Experimental Setup

We conduct main experiments with two widely used few-shot learning datasets: *mini*ImageNet [53] and *tiered*ImageNet [39]. We also experiment on CUB-200 [55], a widely used zero-shot learning dataset. We evaluate on this dataset to provide a more direct comparison with modality-alignment methods. This is because most modality-alignment methods have no published results on few-shot datasets. We use GloVe [37] to extract the word embeddings for the category labels of the two image few-shot learning data sets. The embeddings are trained with large unsupervised text corpora.

More details about the three datasets can be found in Appendix B.

Baselines. We compare AM3 with three family of methods. The first is uni-modality few-shot learning methods such as MAML [7], LEO [41], Prototypical Nets [47] and TADAM [35]. LEO achieves current state of the art among uni-modality methods. The second fold is modality alignment methods. CADA-VAE [44], among them, has the best published results on both zero and few-shot learning. To better extend modality alignment methods to few-shot setting, we also apply the metric-based loss and the episode training of ProtoNets on their visual side to build a visual representation space that better fits few-shot scenario. This leads to the third fold baseline, modality alignment methods extended to metric-based FSL.

Details of baseline implementations can be found in Appendix C.

AM3 Implementation. We test AM3 with two backbone metric-based few-shot learning methods: ProtoNets and TADAM. In our experiments, we use the stronger ProtoNets implementation of [35], which we call ProtoNets++. Prior to AM3, TADAM achieves the current state of the art among all metric-based few-shot learning methods. For details on network architectures, training and evaluation procedures, see Apprendix D. Source code is released at *https://github.com/ElementAI/am3*.

4.2 Results

Table 1 and Table 2 show classification accuracy on *mini*ImageNet and on *tiered*ImageNet, respectively. We conclude multiple results from these experiments. First, AM3 outperforms its backbone methods by a large margin in all cases tested. This indicates that when properly employed, text modality can be used to boost performance in metric-based few-shot learning framework very effectively.

Second, AM3 (with TADAM backbone) achieves results superior to current state of the art (in both single modality FSL and modality alignment methods). The margin in performance is particularly remarkable in the 1-shot scenario. The margin of AM3 w.r.t. uni-modality methods is larger with smaller number of shots. This indicates that the lower the visual content is, the more important semantic information is for classification. Moreover, the margin of AM3 w.r.t. modality alignment methods is larger with smaller number of shots. This indicates that the adaptiveness of AM3 would be more effective when the visual modality provides less information. A more detailed analysis about the adaptiveness of AM3 is provided in Section 4.3.

6

Model	Test Accuracy		
	5-way 1-shot	5-way 5-shot	5-way 10-shot
Uni-modality few-shot learning baselines			
Matching Network [53]	$43.56 \pm 0.84\%$	$55.31 \pm 0.73\%$	-
Prototypical Network [47]	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$	$74.30 \pm 0.52\%$
Discriminative k-shot [2]	$56.30 \pm 0.40\%$	$73.90 \pm 0.30\%$	$78.50 \pm 0.00\%$
Meta-Learner LSTM [38]	$43.44 \pm 0.77\%$	$60.60 \pm 0.71\%$	-
MAML [7]	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$	-
ProtoNets w Soft k-Means [39]	$50.41 \pm 0.31\%$	$69.88 \pm 0.20\%$	-
SNAIL [32]	$55.71 \pm 0.99\%$	$68.80 \pm 0.92\%$	-
CAML [16]	$59.23 \pm 0.99\%$	$72.35 \pm 0.71\%$	-
LEO [41]	$61.76 \pm 0.08\%$	$77.59 \pm 0.12\%$	-
Modality alignment baselines			
DeViSE [9]	$37.43 \pm 0.42\%$	$59.82 \pm 0.39\%$	$66.50 \pm 0.28\%$
ReViSE [14]	$43.20 \pm 0.87\%$	$66.53 \pm 0.68\%$	$72.60 \pm 0.66\%$
CBPL [29]	$58.50 \pm 0.82\%$	$75.62 \pm 0.61\%$	-
f-CLSWGAN [57]	$53.29 \pm 0.82\%$	$72.58 \pm 0.27\%$	$73.49 \pm 0.29\%$
CADA-VAE [44]	$58.92 \pm 1.36\%$	$73.46 \pm 1.08\%$	$76.83 \pm 0.98\%$
Modality alignment baselines extended to metric-based FSL framework			
DeViSE-FSL	$56.99 \pm 1.33\%$	$72.63 \pm 0.72\%$	$76.70 \pm 0.53\%$
ReViSE-FSL	$57.23 \pm 0.76\%$	$73.85 \pm 0.63\%$	$77.21 \pm 0.31\%$
f-CLSWGAN-FSL	$58.47 \pm 0.71\%$	$72.23 \pm 0.45\%$	$76.90 \pm 0.38\%$
CADA-VAE-FSL	$61.59 \pm 0.84\%$	$75.63 \pm 0.52\%$	$79.57 \pm 0.28\%$
AM3 and its backbones			
ProtoNets++	$56.52 \pm 0.45\%$	$74.28 \pm 0.20\%$	$78.31 \pm 0.44\%$
AM3-ProtoNets++	$65.21 \pm 0.30\%$	$75.20 \pm 0.27\%$	$78.52 \pm 0.28\%$
TADAM [35]	$58.56 \pm 0.39\%$	$76.65 \pm 0.38\%$	$80.83 \pm 0.37\%$
AM3-TADAM	$\mathbf{65.30 \pm 0.49\%}$	$\mathbf{78.10 \pm 0.36\%}$	$\mathbf{81.57 \pm 0.47\ \%}$

Table 1: Few-shot classification accuracy on *test* split of *mini*ImageNet. Results in the top use only visual features. Modality alignment baselines are shown on the middle and our results (and their backbones) on the bottom part.

Model	Test Accuracy	
	5-way 1-shot	5-way 5-shot
Uni-modality few-shot learning baselines		
MAML[†] [7]	$51.67 \pm 1.81\%$	$70.30 \pm 0.08\%$
Proto. Nets with Soft k-Means [39]	$53.31 \pm 0.89\%$	$72.69 \pm 0.74\%$
Relation Net[†] [50]	$54.48 \pm 0.93\%$	$71.32 \pm 0.78\%$
Transductive Prop. Nets [28]	$54.48 \pm 0.93\%$	$71.32 \pm 0.78\%$
LEO [41]	$66.33 \pm 0.05\%$	$81.44 \pm 0.09\%$
Modality alignment baselines		
DeViSE [9]	$49.05 \pm 0.92\%$	$68.27 \pm 0.73\%$
ReViSE [14]	$52.40 \pm 0.46\%$	$69.92 \pm 0.59\%$
CADA-VAE [44]	$58.92 \pm 1.36\%$	$73.46 \pm 1.08\%$
Modality alignment baselines extended to metric-based FSL framework		
DeViSE-FSL	$61.78 \pm 0.43\%$	$77.17 \pm 0.81\%$
ReViSE-FSL	$62.77 \pm 0.31\%$	$77.27 \pm 0.42\%$
CADA-VAE-FSL	$63.16 \pm 0.93\%$	$78.86 \pm 0.31\%$
AM3 and its backbones		
ProtoNets++	$58.47 \pm 0.64\%$	$78.41 \pm 0.41\%$
AM3-ProtoNets++	$67.23 \pm 0.34\%$	$78.95 \pm 0.22\%$
TADAM [35]	$62.13 \pm 0.31\%$	$81.92 \pm 0.30\%$
AM3-TADAM	$\mathbf{69.08 \pm 0.47\%}$	$\mathbf{82.58 \pm 0.31\%}$

Table 2: Few-shot classification accuracy on *test* split of *tiered*ImageNet. Results in the top use only visual features. Modality alignment baselines are shown in the middle and our results (and their backbones) in the bottom part. [†]deeper net, evaluated in [28].

Finally, it is also worth noting that all modality alignment baselines get a significant performance improvement when extended to metric-based, episodic, few-shot learning framework. However, most
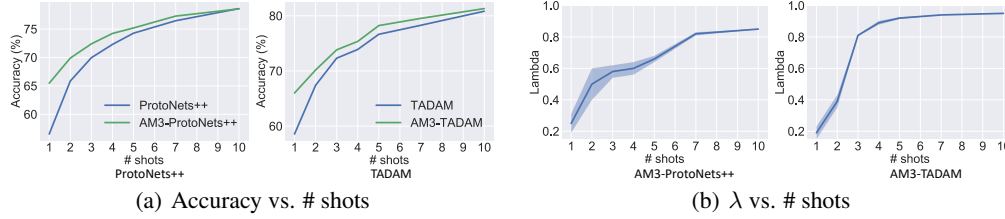
(a) Accuracy vs. # shots   (b) λ vs. # shots

Figure 3: (a) Comparison of AM3 and its corresponding backbone for different number of shots (b) Average value of λ (over whole validation set) for different number of shot, considering both backbones.

of modality alignment methods (original and extended), perform worse than current state-of-the-art uni-modality few-shot learning method. This indicates that although modality alignment methods are effective for cross-modality in ZSL, it does not fit few-shot scenario very much. One possible reason is that when aligning the two modalities, some information from both sides could be lost because two distinct structures are forced to align.

We also conducted few-shot learning experiments on CUB-200, a popular dataest for ZSL dataset, to better compare with published results of modality alignment methods. All the conclusion discussed above hold true on CUB-200. Moreover, we also conduct ZSL and generalized FSL experiments to verify the importance of the proposed adaptive mechanism. Results on on this dataset are shown in Appendix E.

4.3 Adaptiveness Analysis

We argue that the adaptive mechanism is the main reason for the performance boosts observed in the previous section. We design an experiment to quantitatively verify that the adaptive mechanism of AM3 can adjust its focus on the two modalities reasonably and effectively.

Figure 3(a) shows the accuracy of our model compared to the two backbones tested (ProtoNets++ and TADAM) on *mini*ImageNet for 1-10 shot scenarios. It is clear from the plots that the gap between AM3 and the corresponding backbone gets reduced as the number of shots increases. Figure 3(b) shows the mean and std (over whole validation set) of the mixing coefficient λ for different shots and backbones.

First, we observe that the mean of λ correlates with number of shots. This means that AM3 weighs more on text modality (and less on visual one) as the number of shots (hence, the number of visual data points) decreases. This trend suggests that AM3 can automatically adjust its focus more to text modality to help classification when information from the visual side is very low. Second, we can also observe that the variance of λ (shown in Figure 3(b)) correlates with the performance gap of AM3 and its backbone methods (shown in Figure 3(a)). When the variance of λ decreases with the increase of number of shots, the performance gap also shrinks. This indicates that the adaptiveness of AM3 on category level plays a very important role for the performance boost.

5 Conclusion

In this paper, we propose a method that can adaptively and effectively leverage cross-modal information for few-shot classification. The proposed method, AM3, boosts the performance of metric-based approaches by a large margin on different datasets and settings. Moreover, by leveraging unsupervised textual data, AM3 outperforms state of the art on few-shot classification by a large margin. The textual semantic features are particularly helpful on the very low (visual) data regime (*e.g.* one-shot). We also conduct quantitative experiments to show that AM3 can reasonably and effectively adjust its focus on the two modalities.

References

[1] E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *CVPR*, 2005.

[2] Matthias Bauer, Mateo Rojas-Carulla, Jakub Bartlomiej Swikatkowski, Bernhard Scholkopf, and Richard E Turner. Discriminative k-shot learning using probabilistic models. In *NIPS Bayesian Deep Learning*, 2017.

[3] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In *Conference on Optimality in Biological and Artificial Networks*, 1992.

[4] John Bridle. Probabilistic interpretation of feedforward classification network outputs with relationships to statistical pattern recognition. *Neurocomputing: Algorithms, Architectures and Applications*, 1990.

[5] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018.

[6] Michael Fink. Object classification from a single example utilizing class relevance metrics. In *NIPS*, 2005.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[8] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 2018.

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *ICANN*, 2001.

[14] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *CVPR*, 2017.

[15] R Jackendoff. On beyond zebra: the relation of linguistic and visual information. *Cognition*, 1987.

[16] Xiang Jiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, and Stan Matwin. Learning to learn with conditional class dependencies. In *ICLR*, 2019.

[17] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[18] Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *NeurIPS*, 2018.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[21] Alexandre Lacoste, Thomas Boquet, Negar Rostamzadeh, Boris Oreshki, Wonchang Chung, and David Krueger. Deep prior. *NIPS workshop*, 2017.

[22] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Annual Meeting of the Cognitive Science Society*, 2011.

[23] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 1988.

[24] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[26] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[27] Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *PAMI*, 2006.

[28] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, and Yi Yang. Transductive propagation network for few-shot learning. *ICLR*, 2019.

[29] Zhiwu Lu, Jiechao Guan, Aoxue Li, Tao Xiang, An Zhao, and Ji-Rong Wen. Zero and few shot learning with semantic feature synthesis and competitive learning. *arXiv preprint arXiv:1810.08332*, 2018.

[30] Ellen M Markman. *Categorization and naming in children: Problems of induction*. MIT Press, 1991.

[31] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *CVPR Workshops*, 2018.

[32] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[33] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv*, 2018.

[34] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *ICLR*, 2014.

[35] Boris N Oreshkin, Alexandre Lacoste, and Pau Rodriguez. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

[36] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[37] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[38] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[39] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[41] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.

[42] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[43] Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 1987.

[44] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. *CVPR*, 2019.

[45] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005.

[46] Linda B Smith and Lauren K Slone. A developmental approach to machine learning? *Frontiers in psychology*, 2017.

[47] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

[48] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[50] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[51] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

[52] Sebastian Thrun. *Lifelong learning algorithms*. Kluwer Academic Publishers, 1998.

[53] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016.

[54] Yu-Xiong Wang, Ross B. Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018.

[55] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[56] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 2018.

[57] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.

[58] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017.

[59] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 2021–2030, 2017.

[60] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018.

A Algorithm for Episode Loss

Algorithm 1: Training episode loss computation for adaptive cross-modality few-shot learning. M is the total number of classes in the training set, N is the number of classes in every episode, K is the number of supports for each class, K_Q is the number of queries for each class, $\mathcal{W}$ is the pretrained label embedding dictionary.

Input: Training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_i, y_i \in \{1, ..., M\}$. $\mathcal{D}_{\text{train}}^c = \{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{train}} \mid y_i = c\}$.
Output: Episodic loss $\mathcal{L}(\theta)$ for sampled episode e.
{Select N classes for episode e}
$C \leftarrow RandomSample(\{1, ..., M\}, N)$
{Compute cross-modal prototypes}
for c in C **do**
   $\mathcal{S}_e^c \leftarrow RandomSample(\mathcal{D}_{\text{train}}^c, K)$
   $\mathcal{Q}_e^c \leftarrow RandomSample(\mathcal{D}_{\text{train}}^c \setminus \mathcal{S}_e^c, K_Q)$
   $\mathbf{p}_c \leftarrow \frac{1}{|\mathcal{S}_e^c|} \sum_{(s_i, y_i) \in \mathcal{S}_e^c} f(s_i)$
   $\mathbf{e}_c \leftarrow LookUp(c, \mathcal{W})$
   $\mathbf{w}_c \leftarrow g(\mathbf{e}_c)$
   $\lambda_c \leftarrow \frac{1}{1+\exp(-h(\mathbf{w}_c))}$
   $\mathbf{p}_c' \leftarrow \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c$
end for
{Compute loss}
$\mathcal{L}(\theta) \leftarrow 0$
for c in C **do**
   for (q_t, y_t) in Q_e^c **do**
      $\mathcal{L}(\theta) \leftarrow \mathcal{L}(\theta) + \frac{1}{N \cdot K}[d(f(q_t), \mathbf{p}_c')) + \log\sum_k \exp(-d(f(q_t), \mathbf{p}_k'))]$
   end for
end for

B Descriptions of data sets

*mini*ImageNet.   This dataset is a subset of ImageNet ILSVRC12 dataset [40]. It contains 100 randomly sampled categories, each with 600 images of size 84×84. For fair comparison with other methods, we use the same split proposed by Ravi et al. [38], which contains 64 categories for training, 16 for validation and 20 for test.

*tiered*ImageNet.   This dataset is a larger subset of ImageNet than *mini*ImageNet. It contains 34 high-level category nodes (779,165 images in total) that are split in 20 for training, 6 for validation and 8 for test. This leads to 351 actual categories for training, 97 for validation and 160 for the test. There are more than 1,000 images for each class. The train/val/test split is done according to their higher-level label hierarchy. According to Ren et al. [39], splitting near the root of ImageNet hierarchy results in a more realistic (and challenging) scenario with training and test categories that are less similar.

CUB-200.   Caltech-UCSD-Birds 200-2011 (CUB-200) [55] is a fine-grained and medium scale dataset with respect to both number of images and number of classes, *i.e.* 11,788 images from 200 different types of birds annotated with 312 attributes [58]. We chose the split proposed by Xian *et al.* [58]. We used the 312-dimensional hand-crafted attribution as the semantic modality for fair comparison with other published modality alignment methods.

Word embeddings.   We use GloVe [37] to extract the semantic embeddings for the category labels. GloVe is an unsupervised approach based on word-word co-occurrence statistics from large text corpora. We use the Common Crawl version trained on 840B tokens. The embeddings are of dimension 300. When a category has multiple (synonym) annotations, we consider the first one. If the first one is not present in GloVe's vocabulary we use the second. If there is no annotation in GloVe's vocabulary for a category (4 cases in *tiered*ImageNet), we randomly sample each dimension of the embedding from a uniform distribution with the range (-1, 1). If an annotation contains more

than one word, the embedding is generated by averaging them. We also experimented with fastText embeddings [17] and observed similar performances.

C   Baselines

For modality alignment baselines, we follow CADA-VAE [44]'s few-shot experimental setting. During training, we randomly sample N-shot images for the test classes, and add them in the training data to train the alignment model. During test, we compare the image query and the class embedding candidates in the aligned space to make decisions as in ZSL and GZSL.

For the meta-learning extensions of modality alignment methods, instead of including the N-shot images into training data, we follow the standard episode training (explained in Section 3) of metric-based meta-learning approach and train models only with samples from training classes. Moreover, during training, we add an additional loss illustrated in Equation 1 and 3, to ensure the metric space learned on the visual side matching the few-shot test scenario. At test, we employ the standard few-shot testing approach (described in Appendix D) and calculate the prototype representations of test classes as follows:

$$\mathbf{p}_c = \frac{\Sigma_i \mathbf{r}_i^c + \mathbf{w}_c}{N + 1}, \tag{7}$$

where $\mathbf{r}_i$ is the representation of the i-th support image. For both training and test, we need a visual representation space to calculate prototype representations. For DeViSE, they are calculated in its visual space before the transformer [9]. For both ReViSE and CADA-VAE, prototype representations are calculated in the latent space. For f-CLSWGAN, they are calculated in the discriminator's input space.

D   Implementation Details of AM3 Experiments

We model the visual feature extractor f with a ResNet-12 [12], which has shown to be very effective for few-shot classification [35]. This network produces embeddings of dimension 512. We use this backbone in all the modality-alignment baselines mentioned above and in AM3 implementations (with both backbones). We call *ProtoNets++* the prototypical network [47] implementation with this more powerful backbone.

The semantic transformation g is a neural network with one hidden layer with 300 units which also outputs a 512-dimensional representation. The transformation h of the mixture mechanism also contains one hidden layer with 300 units and outputs a single scalar for λ_c. On both g and h networks, we use ReLU non-linearity [10] and dropout [49] (we set the dropout coefficient to be 0.7 on *mini*ImageNet and 0.9 on *tiered*ImageNet).

The model is trained with stochastic gradient descent with momentum [51]. We use an initial learning rate of 0.1 and a fixed momentum coefficient of 0.9. On *mini*ImageNet, we train every model for 30,000 iterations and anneal the learning rate by a factor of ten at iterations 15,000, 17,500 and 19,000. On *tiered*ImageNet, models are trained for 80,000 iterations and the learning rate is reduced by a factor of ten at iteration 40,000, 50,000, 60,000.

The training procedure composes a few-shot training batch from several tasks, where a task is a fixed selection of 5 classes. We found empirically that the best number of tasks per batch are 5,2 and 1 for 1-shot, 5-shot and 10-shot, respectively. The number of query per batch is 24 for 1-shot, 32 for 5-shot and 64 for 10-shot. All our experiments are evaluated following the standard approach of few-shot classification: we randomly sample 1,000 tasks from the test set each having 100 random query samples, and average the performance of the model on them.

All hyperparameters were chosen based on accuracy on validation set. All our results are reported with an average over five independent run (with a fixed architecture and different random seeds) and with 95% confidence intervals.

E   Results on CUB-200

We also conduct experiments on CUB-200 to better compare with modality-alignment baselines from ZSL. Table 3 shows the results. For 0-shot scenario, AM3 degrades to the simplest modality

alignment method that maps the text semantic space to the visual space. Therefore, without the adaptive mechanism, AM3 performs roughly the same with DeViSE, which indicates that the adaptive mechanism play the main role on the performance boost we observed in FSL. The results on other few-shot cases on CUB-200 are consistent with the other two few-shot learning data sets.

We also conduct generalized few-shot learning experiments as reported for CADA-VAE in [44] to compare AM3 with the published FSL results for CADA-VAE. Figure 4 shows that AM3-ProtoNets outperforms CADA-VAE in every case tested. We consider as a metric the harmonic mean (H-acc) between the accuracy of seen and unseen classes, as defined in [56, 44].

Model	Test Accuracy		
	0-shot	1-shot	5-shot
DeViSE [9]	52.0%	54.7%	60.4%
ReViSE [14]	55.2%	56.3%	63.7%
VZSL []	57.4%	60.8%	70.0%
CBPL [29]	61.9%	-	-
f-CLSWGAN [57]	62.1%	64.7%	73.7%
CADA-VAE [44]	61.7%	64.9%	71.9%
ProtoNets	-	68.8%	76.4%
AM3-ProtoNets	51.3%	73.6%	**79.9%**
TADAM [35]	-	69.2%	78.6%
AM3-TADAM	50.7%	**74.1%**	79.7%

Table 3: Few-shot classification accuracy on *unseen-test* split of CUB-200.
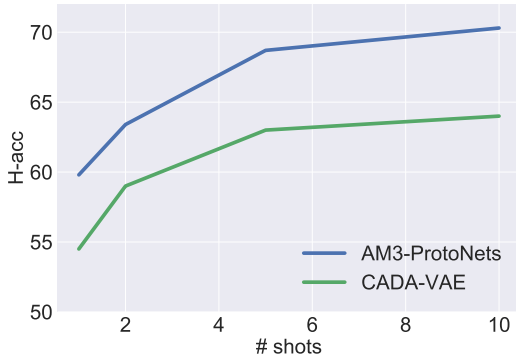

Figure 4: H-acc of generalized few-shot learning on CUB-200.

F   Ablation study on the input of the adaptive mechanism

We also perform an ablation study to see how the adaptive mechanism performs with respect to different features. Table 4 shows results, on both datasets, of our method with three different inputs for the adaptive mixing network h: (i) the raw GloVe embedding ($h(\mathbf{e})$), (ii) the visual representation ($h(\mathbf{p})$) and (iii) a concatenation of both the query and the language embedding ($h(\mathbf{q}, \mathbf{w})$).

We observe that conditioning on transformed GloVe features performs better than on the raw features. Also, conditioning on semantic features performs better than when conditioning on visual ones, suggesting that the former space has a more appropriate structure to the adaptive mechanism than the latter. Finally, we note that conditioning on the query and semantic embeddings helps with the ProtoNets++ backbone but not with TADAM.

14

Method	ProtoNets++		TADAM	
	1-shot	5-shot	1-shot	5-shot
$h(\mathbf{e})$	61.23	74.77	57.47	72.27
$h(\mathbf{p})$	64.48	74.80	64.93	77.60
$h(\mathbf{w}, \mathbf{q})$	66.12	75.83	53.23	56.70
$h(\mathbf{w})$ (AM3)	65.21	75.20	65.30	78.10

Table 4: Performance of our method when the adaptive mixing network is conditioned on different features. Last row is the original model.