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].

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

Preprint. Under review.
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
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 of 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 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 \( D_{\text{train}} \) and tested on \( D_{\text{test}} \). The class sets are disjoint between \( D_{\text{train}} \) and \( D_{\text{test}} \). 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 \( D_{\text{train}} \) 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 \( S_e = \{(s_i, y_i)\}_{i=1}^{N \times K} \) containing \( K \) examples for each of the \( N \) categories and (ii) the query set \( Q_e = \{(q_j, y_j)\}_{j=1}^{Q} \) 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:

\[
L(\theta) = \mathbb{E}_{(S_e, Q_e)} \left[ -\sum_{t=1}^{Q_e} \log p_{\theta}(y_t | q_t, S_e) \right],
\]

where \((q_t, y_t) \in Q_e\) and \(S_e\) are, respectively, the sampled query and support set at episode \(e\) and \(\theta\) 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 \(f: \mathbb{R}^{n_v} \rightarrow \mathbb{R}^{n_p}\), parameterized by \(\theta_f\), 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 \(c\):

\[
p_c = \frac{1}{|S^c_e|} \sum_{(s_i, y_i) \in S^c_e} f(s_i),
\]

where \(S^c_e \subset S_e\) 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:

\[
p(y = c | q_t, S_e, \theta) = \frac{\exp(-d(f(q_t), p_c))}{\sum_k \exp(-d(f(q_t), p_k))}.
\]

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.
The training procedure is similar to that of the original prototypical networks. However, the distances from the cross-modal prototype (used to calculate the distribution over classes for every image query) are between the query and the word embedding of label \( \lambda \) conditioned on the category and calculated as follows:

\[
\lambda_c = \frac{1}{1 + \exp(-h(w_c))},
\]

where \( h \) is the adaptive mixing network, \( w_c = g(e_c) \) is a transformed version of the label embedding for class \( c \) parameterized by \( \theta_g \). This transformation \( g : \mathbb{R}^{n_w} \rightarrow \mathbb{R}^{n_p} \) is important to guarantee that both modalities lie on the same dimensional space \( \mathbb{R}^{n_p} \) and can be combined. The coefficient \( \lambda_c \) is conditioned on category and calculated as follows:

\[
p_c' = \lambda_c \cdot p_c + (1 - \lambda_c) \cdot w_c,
\]

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

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:

\[
p_c' = \lambda_c \cdot p_c + (1 - \lambda_c) \cdot w_c,
\]

where \( \lambda_c \) is the adaptive mixture coefficient (conditioned on the category) and \( w_c = g(e_c) \) is a transformed version of the label embedding for class \( c \). The representation \( e_c \) is the pre-trained word embedding of label \( c \) from \( \mathcal{W} \). This transformation \( g : \mathbb{R}^{n_w} \rightarrow \mathbb{R}^{n_p} \), parameterized by \( \theta_g \), 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 \( \lambda_c \) is conditioned on category and calculated as follows:

\[
\lambda_c = \frac{1}{1 + \exp(-h(w_c))},
\]

where \( h \) is the adaptive mixing network, with parameters \( \theta_h \). Figure 2(left) illustrates the proposed model. The mixing coefficient \( \lambda_c \) 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 \( p_c' \):

\[
p_{\theta}(y = c|q_t, S_c, \mathcal{W}) = \frac{\exp(-d(f(q_t), p_c'))}{\sum_k \exp(-d(f(q_t), p_k'))},
\]

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 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: miniImageNet [53] and tieredImageNet [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 Appendix D. Source code is released at https://github.com/ElementAI/am3.

### 4.2 Results

Table 1 and Table 2 show classification accuracy on miniImageNet and on tieredImageNet, 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.
<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5-way 1-shot</td>
</tr>
<tr>
<td>Uni-modality few-shot learning baselines</td>
<td></td>
</tr>
<tr>
<td>Matching Network [53]</td>
<td>43.56±0.84%</td>
</tr>
<tr>
<td>Prototypical Network [47]</td>
<td>49.42±0.78%</td>
</tr>
<tr>
<td>Discriminative k-shot [2]</td>
<td>56.30±0.40%</td>
</tr>
<tr>
<td>Meta-Learner LSTM [38]</td>
<td>43.44±0.77%</td>
</tr>
<tr>
<td>ProtoNets w Soft k-Means [39]</td>
<td>50.41±0.31%</td>
</tr>
<tr>
<td>SNAIL [32]</td>
<td>55.71±0.99%</td>
</tr>
<tr>
<td>CAML [16]</td>
<td>59.23±0.99%</td>
</tr>
<tr>
<td>LEO [41]</td>
<td>61.76±0.08%</td>
</tr>
<tr>
<td>Modality alignment baselines</td>
<td></td>
</tr>
<tr>
<td>DeVise [9]</td>
<td>37.43±0.42%</td>
</tr>
<tr>
<td>ReViSE [14]</td>
<td>43.20±0.87%</td>
</tr>
<tr>
<td>CBPL [29]</td>
<td>58.50±0.82%</td>
</tr>
<tr>
<td>f-CLSWGAN [57]</td>
<td>53.29±0.82%</td>
</tr>
<tr>
<td>CADA-VAE [44]</td>
<td>58.92±1.36%</td>
</tr>
<tr>
<td>Modality alignment baselines extended to metric-based FSL framework</td>
<td></td>
</tr>
<tr>
<td>DeVise-FSL</td>
<td>56.99±1.33%</td>
</tr>
<tr>
<td>ReViSE-FSL</td>
<td>57.23±0.76%</td>
</tr>
<tr>
<td>f-CLSWGAN-FSL</td>
<td>58.47±0.71%</td>
</tr>
<tr>
<td>CADA-VAE-FSL</td>
<td>61.59±0.84%</td>
</tr>
<tr>
<td>AM3 and its backbones</td>
<td></td>
</tr>
<tr>
<td>ProtoNets++</td>
<td>56.52±0.45%</td>
</tr>
<tr>
<td>AM3-ProtoNets++</td>
<td>65.21±0.30%</td>
</tr>
<tr>
<td>TADAM [35]</td>
<td>58.56±0.39%</td>
</tr>
<tr>
<td>AM3-TADAM</td>
<td>65.30±0.49%</td>
</tr>
</tbody>
</table>

Table 1: Few-shot classification accuracy on test split of miniImageNet. 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5-way 1-shot</td>
</tr>
<tr>
<td>Uni-modality few-shot learning baselines</td>
<td></td>
</tr>
<tr>
<td>MAML† [7]</td>
<td>51.67±1.81%</td>
</tr>
<tr>
<td>Proto. Nets with Soft k-Means [39]</td>
<td>53.31±0.89%</td>
</tr>
<tr>
<td>Relation Net† [50]</td>
<td>54.48±0.93%</td>
</tr>
<tr>
<td>Transductive Prop. Nets [28]</td>
<td>54.48±0.93%</td>
</tr>
<tr>
<td>LEO [41]</td>
<td>66.33±0.05%</td>
</tr>
<tr>
<td>Modality alignment baselines</td>
<td></td>
</tr>
<tr>
<td>DeVise [9]</td>
<td>49.05±0.92%</td>
</tr>
<tr>
<td>ReViSE [14]</td>
<td>52.40±0.46%</td>
</tr>
<tr>
<td>CADA-VAE [44]</td>
<td>58.92±1.36%</td>
</tr>
<tr>
<td>Modality alignment baselines extended to metric-based FSL framework</td>
<td></td>
</tr>
<tr>
<td>DeVise-FSL</td>
<td>61.78±0.43%</td>
</tr>
<tr>
<td>ReViSE-FSL</td>
<td>62.77±0.31%</td>
</tr>
<tr>
<td>CADA-VAE-FSL</td>
<td>63.16±0.93%</td>
</tr>
<tr>
<td>AM3 and its backbones</td>
<td></td>
</tr>
<tr>
<td>ProtoNets++</td>
<td>58.47±0.64%</td>
</tr>
<tr>
<td>AM3-ProtoNets++</td>
<td>67.23±0.34%</td>
</tr>
<tr>
<td>TADAM [35]</td>
<td>62.13±0.31%</td>
</tr>
<tr>
<td>AM3-TADAM</td>
<td>69.08±0.47%</td>
</tr>
</tbody>
</table>

Table 2: Few-shot classification accuracy on test split of tieredImageNet. 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
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 dataset 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 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 miniImageNet 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 $\lambda$ for different shots and backbones.

First, we observe that the mean of $\lambda$ 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 $\lambda$ (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 $\lambda$ 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


