Semantic Cross Attention for Few-shot Learning

Bin Xiao[*] BXIAO103@UOTTAWA.CA
School of Electrical Engineering and Computer Science, University of Ottawa, 535 Legget Drive, Kanata, K2K 3B8, Ottawa, Canada

Chien-Liang Liu[†] CLLIU@NYCU.EDU.TW
Department of Industrial Engineering and Management, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan, ROC

Wen-Hoar Hsaio BASS228@NANYA.EDU.TW
Department of Computer Science and Engineering, Nanya Institute of Technology, Taoyuan 32091, Taiwan, ROC

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Few-shot learning (FSL) has attracted considerable attention recently. Among existing approaches, the metric-based method aims to train an embedding network that can make similar samples close while dissimilar samples as far as possible and achieves promising results. FSL is characterized by using only a few images to train a model that can generalize to novel classes in image classification problems, but this setting makes it difficult to learn the visual features that can identify the images' appearance variations. The model training is likely to move in the wrong direction, as the images in an identical semantic class may have dissimilar appearances, whereas the images in different semantic classes may share a similar appearance. We argue that FSL can benefit from additional semantic features to learn discriminative feature representations. Thus, this study proposes a multi-task learning approach to view semantic features of label text as an auxiliary task to help boost the performance of the FSL task. Our proposed model uses word-embedding representations as semantic features to help train the embedding network and a semantic cross-attention module to bridge the semantic features into the typical visual modal. The proposed approach is simple, but produces excellent results. We apply our proposed approach to two previous metric-based FSL methods, all of which can substantially improve performance. The source code for our model is accessible from github [1].

Keywords: Few-shot learning, Multi-task learning, Cross attention, Metric-based method.

1. Introduction

Few-shot learning (FSL) algorithms have been widely studied in recent years, and most of the work has focused on the few-shot image classification problem. For the few-shot image classification problem, it is a natural choice to use convolution neural networks (CNN) (Lecun et al., 1998; Krizhevsky et al., 2012) to extract visual features from the images. However, simply using visual features can lead to the following problems when training samples in-

[*] Bin Xiao finished this work when he was at National Chiao Tung University
[†] Corresponding author
1. https://github.com/uobinxiao/semantic_cross_attention_fsl

volved in the training process are limited. First, images from different classes may share a similar visual appearance. For example, Figure 1 shows that the shovel and the barn share a similar visual appearance, but their labels appear different, which means that their semantic meanings are different. Second, images from an identical class may have dissimilar visual appearances. The images of the crossword puzzle as shown in Figure 1 present this problem. Although these two crossword puzzle images have identical semantic meanings, their visual appearances are different. We call these two problems a visual-semantic mismatch, in which "visual" means the visual features learned from deep learning models and "semantic" denotes their semantic labels. In particular, the images and labels in Figure 1 are all from the tiered-ImageNet dataset (Ren et al., 2018), which is a widely used dataset in the few-shot image classification problem. It is worth mentioning that a similar concept called Shortcut Learning (Geirhos et al., 2020; Luo et al., 2021) has been discussed in other studies, whose main motivation is to guide the model to learn intended features rather than shortcut features. In the context of few-shot learning, Luo et al. (2021) considered the impact of the image background that may contain the shortcut information. However, the motivation of this study is to handle the visual-semantic mismatch as mentioned above, so our proposed method is different from studies considering short-cut learning.

The aforementioned visual-semantic mismatch problem is not present in an image classification model trained with a large-scale dataset and augmentation techniques. However, in the FSL setting, the number of images in each class is very limited, so the visual features learned by the model are difficult to reflect the appearance variations of the images. The classical metric-based method for FSL relies on an embedding network to transform the input images into feature embeddings that can make similar samples close while dissimilar samples as far as possible. Many previous works (Chen et al., 2019; Xiao et al., 2020; Tian et al., 2020) have empirically shown that the embedding network is crucial for performance. Therefore, this study proposes a multi-task learning model to focus on learning feature embeddings that can quickly adapt to novel classes by leveraging an auxiliary task. The auxiliary task incorporates semantic features into the training process, and the goal is to further refine the embedding network. In addition, we propose a semantic cross-attention module (CAM) to guide the visual embedding network to focus on the correct semantic areas. More specifically, we use a word-embedding model to transform text labels into soft labels as the target of the auxiliary task to inject semantic features into the main task. It is worth mentioning that the proposed CAM module firstly flattens the visual features into a patch sequence and then maps the patch sequence to the key and value vectors. Still, the query vectors are generated by the semantic auxiliary task. Therefore, the proposed CAM module is different from the self-attention mechanism. Once the three types of vectors are available, the dot product of the query vectors and the key vectors is used to generate the weight matrices, and the final output is based on the weight matrices and the value vectors. This way, semantic features can be injected into visual features, guiding the embedding network to focus on the correct areas.

The contributions of this work are three-fold. First, we inspect the problem of FSL and present the visual-semantic mismatch problem that is present in the FSL setting. It is difficult to directly tackle this problem by only using limited images, which is why we propose to incorporate semantic features into the model to help alleviate the problem of only limited images available for each episode. Second, we simultaneously consider the

semantic meaning of text labels and visual features to propose a multi-task learning model that can align visual features and their semantic meanings. Finally, we apply our proposed method to two recent metric-based methods. The experimental results indicate that these methods can benefit from our proposed method and yield substantial improvements in the experiments.

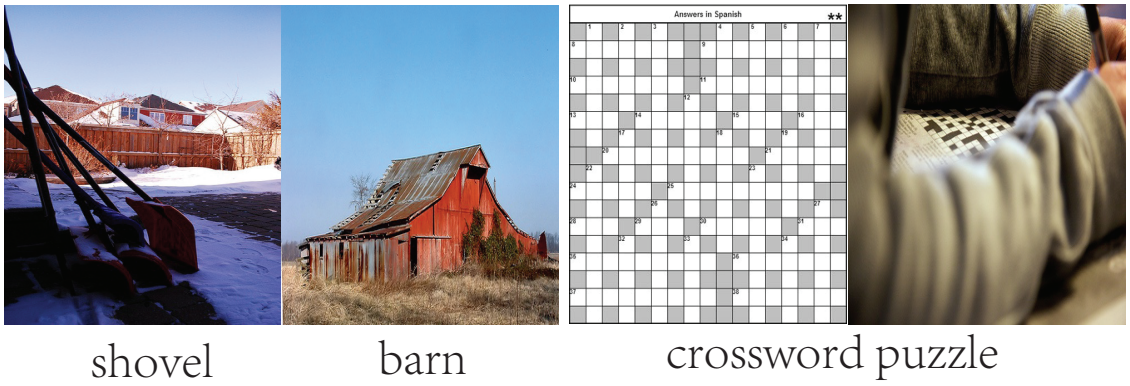

Figure 1: Examples of a similar visual appearance of images in different classes (left) and a dissimilar visual appearance of images in an identical class (right). The images and their labels are from the tiered-ImageNet dataset.

2. Related work

2.1. Few-shot learning

Few-shot learning methods can be roughly classified into three groups: metric-based methods, optimization-based methods, and hallucination-based methods (Chen et al., 2019). Our proposed method uses metric-based methods as the base model, so this section focuses on metric-based approaches. The metric-based FSL method usually comprises three key components, including the embedding network, the class representatives, and the distance metric (Xiao et al., 2020). For example, ProtoNet (Snell et al., 2017) is a classical metric-based method and comprises an embedding network to transform each input image into a visual embedding vector. In addition, it uses the prototype of each class's feature embeddings as the class representative and uses Euclidean distance as the distance metric. Ren et al. (2018) extended ProtoNet in a semi-supervised manner and discussed two semi-supervised few-shot classification settings. Liu et al. (2020) proposed a transductive ProtoNet by introducing a bias diminishing module, which tries to refine the feature embeddings by introducing an extra module. They pointed out that the intra-class bias and the cross-class bias are two key points that can influence the representativeness of class prototypes and proposed a bias diminishing module by using pseudo-labeling and feature shifting. However, their proposed method works in a transductive setting, which differs from most FSL methods. Tian et al. proposed a method based on knowledge distillation to

improve feature embeddings (Tian et al., 2020). Instead of using a typical teacher model, they proposed using the trained model at different epochs as a teacher model, also called self-distillation. Many works (Chen et al., 2019; Tian et al., 2020) have also discussed the influence of a deeper embedding network and concluded that the model could generally benefit from a deeper embedding network to achieve better performance to some extent.

RelationNet (Sung et al., 2018) improves the metric-based method by using a trainable distance metric, which comprises convolution layers and fully connected layers. Similarly, ProxyNet (Xiao et al., 2020) also defines a trainable metric using a 3D convolution layer and uses trainable class representatives. Simon et al. (2020) proposed to use dynamic classifiers that are based on a deep subspace network to extract a discriminated subspace from the feature embeddings. The goal is to further improve the distance metric for typical metric-based methods.

Several studies have considered using the features of the text modality to deal with the FSL problem. Xing et al. (2019) also pointed out the drawbacks of using only visual features to deal with FSL and proposed to use the semantic features of the text modality. They proposed an adaptive modality mixture mechanism (AM3) that can combine the features of visual modality and text modality adaptively and selectively. Although AM3 also considers text and visual modalities, the purpose and approach are different from our proposed approach. More specifically, AM3 uses a weighted sum of visual and semantic features as prototypes, and the weights are determined by an adaptive mixing network. The semantic features generated from class labels are required input for AM3 in both meta-training and meta-test stages, since the model relies on the semantic features to calculate the prototypes. In contrast, our proposed method is based on a multi-task architecture and focuses on learning visual-semantic features that can quickly adapt to novel classes by leveraging an auxiliary task. Besides, we propose a CAM module to inject semantic features into visual embeddings. It is worth mentioning that label information is not required in our meta-testing stage, as the main task is a few-shot classification. The study conducted by Pahde et al. (2021) is also a multi-modality method and uses a generative adversarial network (GAN) framework to combine the features of the text modality with the visual features. Moreover, this model takes images and text descriptions as inputs, and the model comprises the semantic features of the input text description. Notably, the text description is an additional knowledge source, which differs from the label that is available in the meta-training set.

2.2. Multi-task learning

Multi-task learning is a learning paradigm that optimizes several tasks simultaneously, and the goal is to use auxiliary tasks to improve the performance of the main task. The role of auxiliary tasks is similar to that of a regularization term that can place additional constraints on the model. Additionally, the main task can benefit from the parameter sharing mechanism to improve the performance of the model. Many approaches have been proposed by designing different parameter sharing mechanisms. One of the most widely used approaches is hard parameter sharing (Caruana, 1993; Ruder, 2017). For deep learning models, hard parameter sharing can be realized by using sharing layers to learn the feature representations that are common for all tasks and by keeping task-specific layers to learn

the representations that are specific for all tasks. This approach is simple and easy to implement, which is why many studies have used the hard parameter sharing approach to develop methods and achieved promising performance (Zhang et al., 2014; Dai et al., 2016). In contrast, soft parameter sharing (Ruder, 2017) allows each task to have its own model with its own parameters and uses a regularization technique to make the parameters of the tasks similar.

Besides, multi-task learning can be viewed as a transfer learning approach that transfers knowledge between different tasks. Note that multi-task learning cannot guarantee that the main task can benefit from the auxiliary tasks to improve performance, as negative transfer may occur, which can degrade the performance of the main task. To alleviate negative transfer, Lee et al. (2016) proposed a method called asymmetric multi-task learning (AMTL) by considering the task relatedness of the task and the loss of each task to measure the reliability of each task. In addition to task loss, other metrics, such as uncertainty (Nguyen et al., 2020; Kendall and Gal, 2017), can also be used to measure task reliability.

3. Proposed Method

This study follows conventional meta-learning settings to train the proposed method, so the model training is processed episodically. In each episode, the input comprises a sequence of meta-tasks, each of which comprises a support set and a query set. The images for these two sets are non-overlapped and are randomly drawn from the meta-training set. In particular, we use meta-task to represent the tasks in episode-based meta-training to distinguish with the *task* that is used by multi-task learning in this section.

For a K-way-N-shot problem, the support set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{K \cdot N}$ comprises $K \cdot N$ images that are randomly selected from K classes, each of which has N images. On the other hand, the query set $Q = \{(\mathbf{x}_j, y_i)\}_{j=1}^{M}$ comprises M images that are randomly sampled from the corresponding K classes and differ from the images of the support set. These sampled meta-tasks can be not only the inputs of the main task, but also the inputs of the auxiliary task in our multi-task learning model.

The proposed method can be applied to different metric-based methods. To explain how our proposed model works, we use ProtoNet (Snell et al., 2017) as an example to further illustrate the proposed method, since ProtoNet is a classical metric-based method for FSL. In the experiments, we use different metric-based methods as base methods to show the performance improvement brought about by our proposed method.

The embedding network plays a crucial role in metric-based methods (Chen et al., 2019; Xiao et al., 2020; Tian et al., 2020), which is why this study aims to incorporate semantic features into the model to help train an embedding network that can adapt to novel classes. Unlike other metric-based methods that only use images to train the embedding network, this study proposes using semantic features to enhance the training of the embedding network, in which the proposed method comprises text and image modalities, and the text modality can provide semantic information. To incorporate two modalities into the model, we propose to use a multi-task learning architecture with two tasks to handle visual features and the semantic information of text labels, respectively.

Figure 2 shows the overall architecture of the proposed method, which contains two tasks: the main task for the visual modality and the auxiliary task for the text modality.

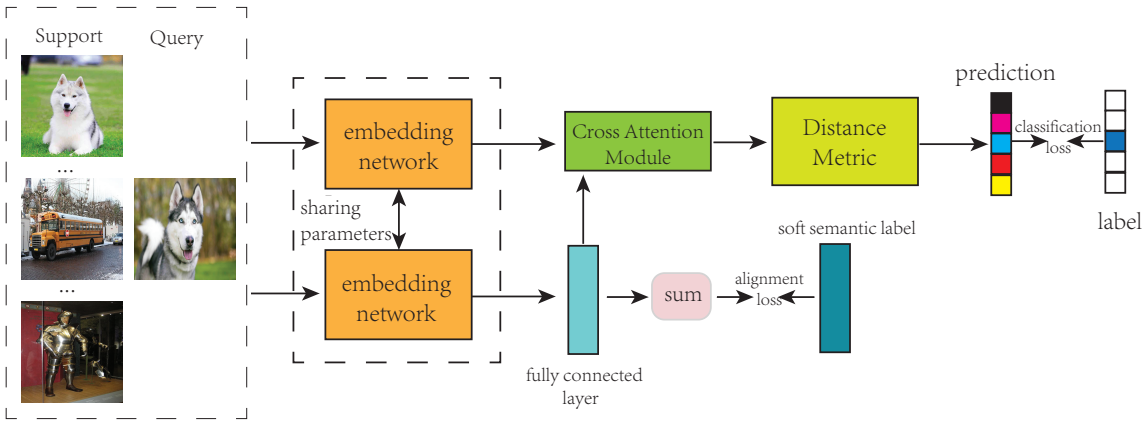

Figure 2: The overall learning architecture of the proposed multi-task model.

The former deals with few-shot classification, while the latter aims to incorporate semantic features into the model. We use the hard parameter sharing method to design the model so that the main task and the auxiliary task share an identical embedding network $\mathcal{E}_\theta$. For the main task, an input image x is first transformed into feature maps using the embedding network $\mathcal{E}_\theta$; then, the feature maps are fed into the CAM denoted by $Cross_\phi$ to obtain the final embedding of the features of each image. The entire process can be represented by Equation (1) and Equation (2), in which ch_{in}, h_{in}, w_{in} are the number of channels, height and width of the input image, respectively, l_{inter} is the depth of the feature map generated by $\mathcal{E}_\theta$, and $l_{out}, h_{out}, w_{out}$ are the depth, height and width of the final embedding generated by $Cross_\phi$. It is worth mentioning that $Cross_\phi$ also takes an input e_{aux} from the auxiliary task.

$$e_{main} = \mathcal{E}_\theta(x), x \in \mathbb{R}^{ch_{in} \times h_{in} \times w_{in}}, e \in \mathbb{R}^{l_{inter} \times h_{out} \times w_{out}} \tag{1}$$

$$e_{out} = Cross_\phi(e_{main}, e_{aux}), e_{out} \in \mathbb{R}^{l_{out} \times h_{out} \times w_{out}} \tag{2}$$

Subsequently, we can determine the class assignment based on the feature embeddings using a distance metric. In a typical setting, metric-based methods can use the nearest-neighbor approach to determine the assignment of the class by calculating the distance between the query image and the class representatives on the embedding space. Our proposed method can be applied to any metric-based model, so the definition of class representatives is determined by the base model. In ProtoNet, for a K-way-N-shot problem, the class representative is defined as the mean of the feature embeddings of that class, as defined in Equation (3). Assume that the prototype embedding of the kth class in the support set S is e_k^p. Then, for the jth image in the query set Q with the feature embedding e_j^q, the distance over the class prototypes e_k^p is defined in Equation (4).

$$e_k^p = \frac{1}{N} \sum_{n=1}^{N} e_{\{k,n\}}^p, k \in \{1, 2, \ldots, K\} \tag{3}$$

$$d_{\{k,j\}} = \mathcal{F}_\omega(e_k^p, e_j^q), k \in \{1, 2, \ldots, K\}, j \in \{1, 2, \ldots, M\}, \tag{4}$$

where $\mathcal{F}_\omega$ is the distance function. In ProtoNet, $\mathcal{F}_\omega$ is the Euclidean distance. Then, for the jth query image q_j with embedding e_j^q, the final distribution over classes should be:

$$p_\theta(y = i|q_j) = \frac{exp(-d_{\{k,j\}})}{\sum_{k=1}^{K} exp(-d_{\{k,j\}})}, i \in \{1, 2, \ldots, K\}, j \in \{1, 2, \ldots, M\}, \tag{5}$$

where M is the number of query images in each class defined in a K-way-N-shot few-shot classification problem.

The visual feature map is available for the auxiliary task once the input image x passes through the embedding network E_θ. Subsequently, we reshape the generated visual feature map into a patch sequence and use a fully-connected layer FC_σ to project the visual embedding vector into the space of the same dimension length as the soft labels. Finally, we use a sum operation across the dimension $h \times w$ to make the output have the same dimension as the soft label. The entire process is defined in Equation (6) and Equation (7).

$$e_{aux} = FC_\sigma(\mathcal{E}_\theta(x)), x \in \mathbb{R}^{ch_{in} \times h_{in} \times w_{in}}, e_{aux} \in \mathbb{R}^{l_{out} \times (h_{out} \times w_{out})} \tag{6}$$

$$\hat{y}_{aux} = \text{softmax}(\text{sum}(e_{aux})/\tau), \hat{y}_{aux} \in \mathbb{R}^{l_{out}}, \tag{7}$$

where l_{out} is the vector length of a soft label, ch_{in} is the number of channels in the input image and τ is the temperature parameter. In particular, e_{aux} defined in Equation (6) acts as an input to the proposed CAM, and semantic soft labels are the word-embedding vectors of the labels leveraging the pre-trained Glove model (Pennington et al., 2014). The semantic soft labels are the targets of the auxiliary task, and the purpose is to place an additional constraint on the main task to tackle the visual-semantic mismatch problem.

The model can incorporate semantic and visual features into the network by sharing the embedding network of the two tasks, and we also introduce a CAM to inject semantic features into visual features, as shown in Figure 3. As defined in Equation (2), the CAM takes two inputs, namely, the feature map of the main task $e_{main} = \mathcal{E}_\theta(x)$ and the feature map of the auxiliary task e_{aux}. For the feature map of the main task e_{main}, it is first reshaped into a patch sequence, then the patch sequence is projected into a value sequence p_{value} and a key sequence p_{key} by two fully-connected layers separately. Then the value sequence, the key sequence, and the query sequence are fed into the Dot-Product Attention module, which is defined by Equation (8), to obtain the final output embedding.

$$e_{out} = \text{MatMul}(\text{softmax}(\text{MatMul}(p_{query}, p_{key}) \cdot scale), p_{value}) \tag{8}$$

Finally, we use KL-divergence as the loss function of the auxiliary task to calculate the difference between the soft labels and the outputs of the auxiliary task, and we use cross-entropy as the loss function of the few-shot classification task. Therefore, the proposed model has two tasks and the total loss $\mathcal{L}$ of the proposed method is defined in Equation (9), in which $\mathcal{L}_{aux}$ is the loss of the auxiliary task and $\mathcal{L}_{cls}$ is the loss of the few-shot classification task, and λ is a hyperparameter that controls the weight of the auxiliary loss.

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{cls} + \lambda\mathcal{L}_{aux}, \tag{9}$$

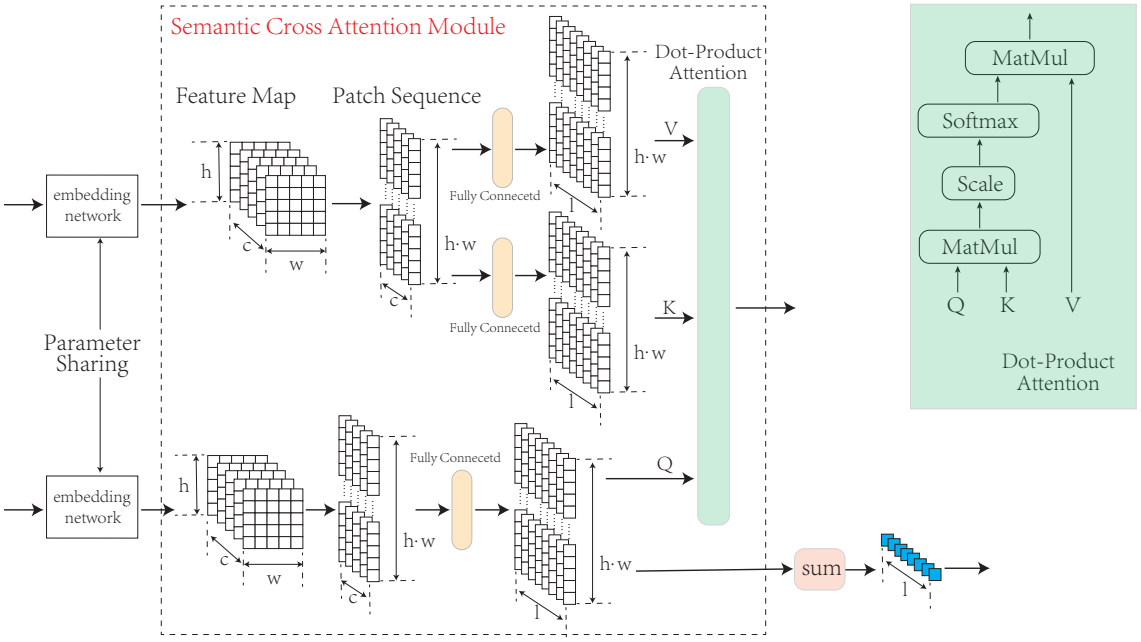

Figure 3: The architecture of semantic cross attention module.

4. Experiments

4.1. Experimental settings & Results

We follow the evaluation protocol of the study (Chen et al., 2019) by using CUB (Welinder et al., 2010), mini-ImageNet (Vinyals et al., 2016), and tiered-ImageNet (Ren et al., 2018) to evaluate the proposed method. The mini-ImageNet dataset comprises 64 classes for training, 16 classes for validation, and 20 classes for testing, and each class has 600 images. Meanwhile, the CUB dataset comprises 11,788 images, among which 100 classes were for training, 50 classes for validation, and 50 classes for testing. The tiered-IamgeNet dataset uses 351 classes from 20 categories as the training set, 97 classes from 6 different categories as the validation set, and 160 classes from 8 different categories as the testing set.

Note that the embedding network plays a key role in the metric-based methods for FSL. Consequently, we fix the size of the embedding network using the Conv-4-128 architecture in the experiments for a fair comparison. The Conv-4-128 network comprises four convolution layers that have 64, 64, 128, and 128 filters, respectively. In particular, for the auxiliary task, we add an extra fully-connected layer right after the embedding network to align the visual features and the corresponding text features when they are used to calculate the auxiliary loss, as shown in Figure 2.

To transform the text of the label into soft labels of the auxiliary task, we use Glove (Pennington et al., 2014) pretrained on the Common Crawl corpus (Foundation, 2021), and the word embedding vector is of size 300. For the label text that comprises more than one word, we use the average of the embedding vectors for the words involved in the label text as the soft label. All images are randomly resized and cropped to the size of 84×84, and

color jitter and random flip are used as augmentation methods. The λ in Equation 9 is 0.1 in the experiments. We apply the proposed method to two metric-based methods, including ProtoNet (Snell et al., 2017) and ProxyNet (Xiao et al., 2020). The Optimizers are AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 0.001 for ProtoNet and SGD with an initial learning rate of 0.1 for ProxyNet. The reduce-lr-on-plateau (PytorchDevTeam, 2021) method is applied to adjust the learning rate of the optimizer during the training of the two optimizers. We set the number of epochs to 2000, each of which comprises 100 episodes, and the model can be trained in ten hours with an NVIDIA-1080 Ti GPU. The experimental results on the 5-way-1-shot and 5-way-5-shot tasks are listed in Table 1 and Table 2, in which the three metric-based methods with plus symbols are those enhanced by our proposed method. The values in Table 1 and Table 2 are the accuracy and the 95% confidence interval, and the performance values for ProtoNet are directly obtained from (Chen et al., 2019).

Although Self-distill (Tian et al., 2020), DSN-MR (Simon et al., 2020), and BOIL (Oh et al., 2021) are three state-of-the-art methods, Self-distill and DSN-MR do not report their experimental results in the CUB dataset. In Self-distill (Tian et al., 2020), the knowledge distillation method is defined as a series of generations based on the timeline, which means that the model at time T can learn knowledge from the model at time $T-1$. BOIL is a gradient-based approach that relies on representation change, which is a crucial component in gradient-based methods.

As shown in Table 1 and Table 2, the experimental results indicate that the two metric-based methods can benefit from our proposed method and produce significant improvement. Moreover, using our proposed method can allow these methods to achieve competitive performance compared to state-of-the-art methods. The source code for our model is publicly accessible, and detailed hyperparameters and settings are available in the source code.

We focus on how to train a better embedding network as it is a crucial component for metric-based methods. Central to our proposed method is to consider the semantic information obtained from the text labels as an auxiliary task to incorporate semantic information into the training of the embedding network. An ablation study is conducted to analyze the importance of these components in our proposed method.

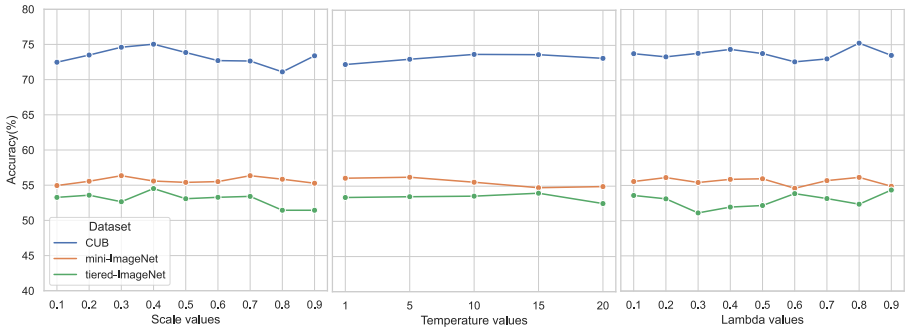

Figure 4: Experimental results with different values of λ values, different values of temperature τ and different values of *scale*.

Table 1: Experimental results on CUB, mini-ImageNet, and tiered-ImageNet datasets for 5-way-1-shot tasks. + means the model trained with the proposed method using ConvNet4-128 backbone, ‡ means our re-implementation of ProtoNet and ProxyNet with ConvNet4-128 backbone.

Method	CUB	mini-ImageNet	tiered-ImageNet
ProtoNet	50.46 ± 0.88	44.42 ± 0.84	–
ProxyNet	67.52 ± 0.97	52.95 ± 0.76	–
Self-distill	–	55.88 ± 0.59	56.76 ± 0.68
DSN-MR	–	55.88 ± 0.90	–
BOIL	61.60 ± 0.57	49.61 ± 0.16	48.58 ± 0.27
ProtoNet‡	62.67 ± 0.93	50.21 ± 0.82	43.62 ± 0.82
ProxyNet‡	72.36 ± 0.84	54.40 ± 0.81	55.05 ± 0.93
ProtoNet+	74.28 ± 0.87	54.34 ± 0.81	56.25 ± 0.97
ProxyNet+	$\mathbf{76.66 \pm 0.87}$	$\mathbf{57.45 \pm 0.86}$	$\mathbf{59.32 \pm 0.90}$

Table 2: Experimental results on CUB, mini-ImageNet, and tiered-ImageNet datasets for 5-way-5-shot tasks. + means the model trained with the proposed method using ConvNet4-128 backbone, ‡ means our re-implementation of ProtoNet and ProxyNet with ConvNet4-128 backbone.

Method	CUB	mini-ImageNet	tiered-ImageNet
ProtoNet	76.39 ± 0.64	64.24 ± 0.72	–
ProxyNet	82.85 ± 0.60	70.35 ± 0.63	–
Self-distill	–	71.65 ± 0.51	73.21 ± 0.54
DSN-MR	–	70.50 ± 0.68	–
BOIL	75.96 ± 0.17	66.45 ± 0.37	69.37 ± 0.12
ProtoNet‡	79.39 ± 0.65	69.64 ± 0.65	72.18 ± 0.76
ProxyNet‡	86.15 ± 0.51	72.36 ± 0.67	73.77 ± 0.76
ProtoNet+	86.39 ± 0.49	73.12 ± 0.63	76.89 ± 0.72
ProxyNet+	$\mathbf{88.48 \pm 0.46}$	$\mathbf{73.54 \pm 0.62}$	$\mathbf{77.83 \pm 0.69}$

4.2. Sensitivity Analysis

To further investigate the impact of the proposed semantic feature alignment on the performance of the model, we perform a sensitivity analysis by changing the value of λ in Equation 9 with ProtoNet as the base model. The experiments are carried out on CUB, mini-ImageNet, and tired-ImageNet datasets for the 5-way-1-shot problem, and the experimental results are shown in Figure 4. It can be seen from the results that our proposed method produces a relatively stable performance under different values of λ in the range from 0.1 to 0.9. Similarly, we also perform sensitivity analysis on the other two hyperparameters, namely the temperature τ defined in Equation 7 and the *scale* value defined in Equation 8, and the experimental results are also shown in Figure 4. Note that the

accuracy scores listed in Figure 4 are obtained using the validation set. Overall, the experimental results show that the proposed method is robust to different choices of these hyper-parameters, especially for τ, in the range reported in Figure 4.

4.3. Ablation Study

We conduct an ablation study to assess the effectiveness of the components involved in our proposed method. First, we conduct experiments to verify whether the main few-shot classification task in our model can benefit from the proposed CAM and the auxiliary task. We compare our model with the multi-task learning model without CAM. Furthermore, the squeeze-and-excitation network (Hu et al., 2018) is a simple but useful module on channel attention, so we also develop a model that replaces our proposed CAM module with a squeeze-excitation module to re-weight the importance of the channel, but without using the inputs of aligned semantic features that are from the auxiliary task.

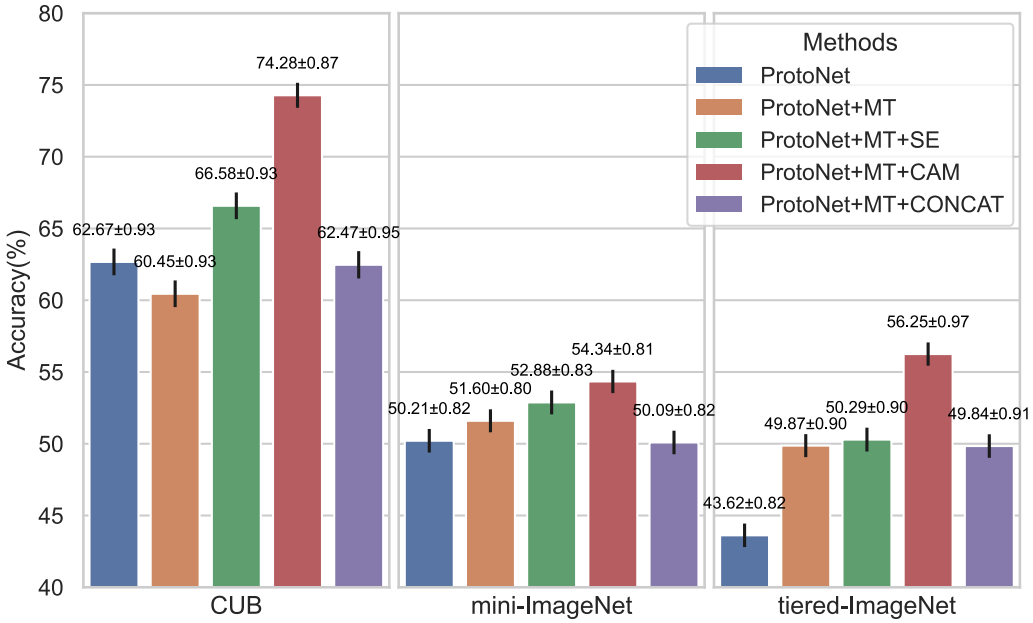

Figure 5: The ablation study results for 5-way-1-shot tasks. ProtoNet here is our re-implemented version; ProtoNet+MT means the model trained only with the multi-task architecture without using the proposed CAM component; ProtoNet+MT+CAM means the model trained with the multi-task architecture and the proposed CAM component. ProtoNet+MT+CONCAT means the model trained with the multi-task architecture and a concatenation of the feature maps from the main task and the auxiliary task is used for the subsequent classification.

Similarly, we still use ProtoNet as the base model, and the experiments are carried out on CUB, mini-ImageNet, and tiered-ImageNet datasets for the 5-way-1-shot problem. The experimental results are presented in Figure 5. The experimental results point out

Table 3: Experimental results for 5-way-1-shot tasks with different embedding networks. + means the proposed model trained with the ConvNet4-128 backbone, and † means the proposed model trained with the ResNet12 backbone.

Method	CUB	mini-ImageNet	tiered-ImageNet
ProtoNet$^+$	74.28 ± 0.87	54.34 ± 0.81	56.25 ± 0.97
ProxyNet$^+$	76.66 ± 0.87	57.45 ± 0.86	59.32 ± 0.90
ProtoNet†	79.81 ± 0.84	57.89 ± 0.85	60.40 ± 0.98
ProxyNet†	$\mathbf{80.20 \pm 0.81}$	$\mathbf{60.58 \pm 0.85}$	$\mathbf{63.84 \pm 1.02}$

Table 4: Experimental results for 5-way-5-shot tasks with different embedding networks. + means the proposed model trained with the ConvNet4-128 backbone, and † means the proposed model trained with the ResNet12 backbone.

Method	CUB	mini-ImageNet	tiered-ImageNet
ProtoNet$^+$	86.39 ± 0.49	73.12 ± 0.63	76.89 ± 0.72
ProxyNet$^+$	88.48 ± 0.46	73.54 ± 0.62	77.83 ± 0.69
ProtoNet†	89.10 ± 0.42	75.56 ± 0.63	$\mathbf{83.60 \pm 0.67}$
ProxyNet†	$\mathbf{90.80 \pm 0.39}$	$\mathbf{76.53 \pm 0.58}$	83.15 ± 0.69

that only using the multi-task learning architecture and semantic features can significantly improve the performance. Thus, this can empirically show the importance of semantic features in our proposed method. Moreover, both the squeeze-excitation module and the proposed CAM can improve performance. The proposed CAM module can outperform the squeeze-excitation module in all three data sets. We also list a model implemented in the proposed multi-task manner, but the semantic features generated in the auxiliary task are concatenated with the visual features in the main task. The results show that this simple concatenation method of merging visual features and semantic features only leads to performance improvement on the tiered-ImageNet dataset, demonstrating that our proposed CAM module is adequate for combining visual and semantic features.

4.4. Embedding Network

We also explore the impact of the embedding network using ResNet12 (He et al., 2016) to replace the Conv-4-128 network. Many studies have shown that a deeper embedding network may improve performance. We conduct experiments on CUB, mini-ImageNet, and tiered-ImageNet datasets for 5-way-1-shot and 5-way-5-shot problems using ProtoNet and ProxyNet with ResNet-12 as their embedding networks. The experimental results are listed in Table 3 and Table 4, indicating that the use of ResNet-12 can significantly improve the performance of ProtoNet and ProxyNet. Thus, the proposed method can also benefit from a deeper embedding network.

Table 5: Experimental results for 5-way-1-shot tasks with different auxiliary losses. + means the proposed model trained with KL loss as the auxiliary loss, ∗ means the proposed model trained with MSE loss as the auxiliary loss.

Method	CUB	mini-ImageNet	tiered-ImageNet
ProtoNet$^+$	74.28 ± 0.87	54.34 ± 0.81	56.25 ± 0.97
ProxyNet$^+$	$\mathbf{76.66 \pm 0.87}$	$\mathbf{57.45 \pm 0.86}$	$\mathbf{59.32 \pm 0.90}$
ProtoNet*	74.82 ± 0.87	55.05 ± 0.86	54.89 ± 0.94
ProxyNet*	75.98 ± 0.88	55.80 ± 0.86	58.59 ± 0.96

Table 6: Experimental results for 5-way-5-shot tasks with different auxiliary losses. + means the proposed model trained with KL loss as an auxiliary loss, ∗ means the proposed model trained with MSE loss as an auxiliary loss.

Method	CUB	mini-ImageNet	tiered-ImageNet
ProtoNet$^+$	86.39 ± 0.49	73.12 ± 0.63	76.89 ± 0.72
ProxyNet$^+$	$\mathbf{88.48 \pm 0.46}$	$\mathbf{73.54 \pm 0.62}$	$\mathbf{77.83 \pm 0.69}$
ProtoNet*	86.35 ± 0.49	72.52 ± 0.63	76.41 ± 0.73
ProxyNet*	87.59 ± 0.49	73.12 ± 0.66	77.59 ± 0.72

4.5. Auxiliary task with different loss functions

In our proposed method, we use KL loss as the loss of the semantic task. Other loss functions such as mean squared loss (MSE) can also be used in the semantic task. In Table 5 and Table 6, we compare the performance of using the KL loss and the MSE loss as an auxiliary loss in the semantic task. We set λ in Equation 9 to 0.1 in the experiments. The experimental results show that MSE loss can also work well with the proposed method, and the performance of KL loss and MSE loss are close, but overall, MSE loss does not show any superiority compared to KL loss.

4.6. Auxiliary task with different pre-trained language models

In this study, we propose using a pre-trained language model to generate the soft labels used in the auxiliary task. To evaluate the impact of different pre-trained language models, we conduct experiments with three classic pre-trained language models, including Glove (Pennington et al., 2014), FastText (Bojanowski et al., 2017) and Bert (Devlin et al., 2018) for the 5-way-1shot problem. The experimental results are listed in Table 7. The experimental results show that Glove generally performs better than the other two.

Table 7: Experimental results for 5-way-1-shot tasks with different word embedding models. † means using ProtoNet as the base model, while ‡ means that the base model is ProxyNet.

Method	CUB	mini-ImageNet	tiered-ImageNet
`Glove`†	74.28 ± 0.87	54.34 ± 0.81	56.25 ± 0.97
`FastText`†	74.82 ± 0.82	52.79 ± 0.83	52.20 ± 0.91
`Bert`†	74.39 ± 0.86	53.45 ± 0.86	53.46 ± 0.94
`Glove`‡	$\mathbf{76.66 \pm 0.87}$	$\mathbf{57.45 \pm 0.86}$	$\mathbf{59.32 \pm 0.90}$
`FastText`‡	74.64 ± 0.85	55.56 ± 0.84	55.81 ± 0.88
`Bert`‡	75.63 ± 0.90	56.35 ± 0.86	54.82 ± 0.92

5. Conclusions

This study presents the visual-semantic mismatch problem that is present in the FSL setting and proposes to incorporate semantic features into the model to help alleviate the problem of only limited images available for each episode. In our proposed multi-task learning model, we align the visual features with their semantic meanings to enforce the model to consider the semantic meaning of text labels and visual features simultaneously, giving a base to train an embedding network that can quickly adapt to novel classes with limited training samples. Besides, we propose a CAM module to further refine the visual embeddings by using the aligned features of the auxiliary task. We apply the proposed method to two classical metric-based methods. The experimental results indicate that these methods can benefit from our proposed method to achieve substantial improvement and competitive performance compared to state-of-the-art models. Our proposed method can also be applied to other metric-based methods without changing the pipeline of the original methods.

Acknowledgments

This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant no. MOST 109-2628-E-009-009-MY3 and 111-2221-E-A49-083-MY3. We are grateful to the National Center for High-performance Computing for computer time and facilities.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

R. Caruana. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, 1993.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3150–3158, 2016.

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

Common Crawl Foundation. Common crawl corpus, 2021. URL https://commoncrawl.org/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Alex Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

Giwoong Lee, Eunho Yang, and Sung Hwang. Asymmetric multi-task learning based on task relatedness and loss. In *International conference on machine learning*, pages 230–238. PMLR, 2016.

Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *ECCV*, 2020.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34:13073–13085, 2021.

Tuan A. Nguyen, Hyewon Jeong, Eunho Yang, and Sung Ju Hwang. Clinical risk prediction with temporal probabilistic asymmetric multi-task learning. *ArXiv*, abs/2006.12777, 2020.

Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Seyoung Yun. Boil: Towards representation change for few-shot learning. In *ICLR*, 2021.

Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

PytorchDevTeam. torch.optim-pytorch master document, 2021. URL https://pytorch.org/docs/stable/optim.html.

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. *arXiv preprint arXiv:1803.00676*, 2018.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

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.

Bin Xiao, Chien-Liang Liu, and Wen-Hoar Hsaio. Proxy network for few shot learning. In *Asian Conference on Machine Learning*, pages 657–672. PMLR, 2020.

Chen Xing, Negar Rostamzadeh, Boris N Oreshkin, and Pedro O Pinheiro. Adaptive cross-modal few-shot learning. *arXiv preprint arXiv:1902.07104*, 2019.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.