Attribute-Guided Feature Learning for Few-Shot Image Recognition

Yaohui Zhu, Weiqing Min, Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Few-shot image recognition has become an essential problem in the field of machine learning and image recognition, and has attracted more and more research attention. Typically, most few-shot image recognition methods are trained across tasks. However, these methods are apt to learn an embedding network for discriminative representations of training categories, and thus could not distinguish well for novel categories. To establish connections between training and novel categories, we use attribute-related representations for few-shot image recognition and propose an attribute-guided two-layer learning framework, which is capable of learning general feature representations. Specifically, few-shot image recognition trained over tasks and attribute learning trained over images share the same network in a multi-task learning framework. In this way, few-shot image recognition learns feature representations guided by attributes, and is thus less sensitive to novel categories compared with feature representations only using category supervision. Meanwhile, the multi-layer features associated with attributes are aligned with category learning on multiple levels respectively. Therefore we establish a two-layer learning mechanism guided by attributes to capture more discriminative representations, which are complementary compared with a single-layer learning mechanism. Experimental results on CUB-200, AWA and MiniImageNet datasets demonstrate our method effectively improves the performance.**

***Index Terms*—Few-Shot Learning, Image Recognition, Attribute Learning.**

I. Introduction

OVER the past years, with the assistance of deep learning techniques [1] and large scale training datasets [2], [3], significant progress has been made in image recognition [4], [5], [6]. Despite these successes, there still exists a huge gap between machine intelligence and human beings. As one remarkable example, humans can rapidly recognize novel objects from a few examples or learn new skills after just minutes of experience by leveraging knowledge learned before [7]. However, most of machine learning algorithms, especially for deep learning, are inherently data-hungry and time-consuming [8]. It remains a challenging fundamental problem to expand the capability of machine to learn novel concepts from only one or a few examples of each category, which is known as few-shot learning. If this can be realized, the application of machine intelligence will be greatly expanded.

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018, 61972378, U1936203, and U19B2040, in part by Beijing Natural Science Foundation under Grant L182054, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals. The authors are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China and the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: {sqjiang,minweiqing}@ict.ac.cn, yaohui.zhu@vipl.ict.ac.cn.

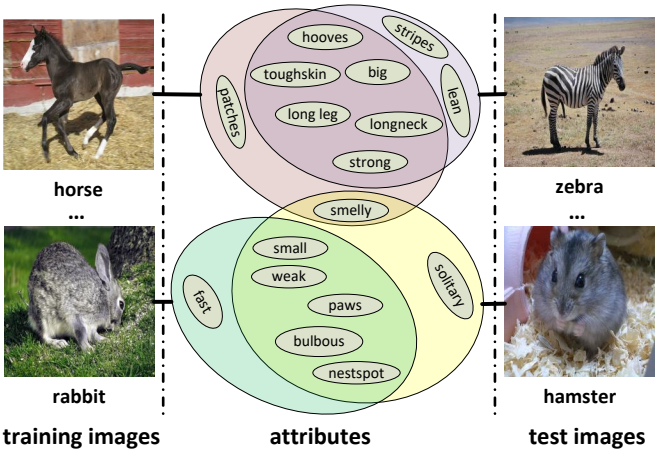


Fig. 1: Training and test images have disjoint categories, but different categories may be represented in some same attributes. (e.g., The horse has some same attributes as the zebra) Then these same attributes effectively connect different categories. Moreover, attributes can be used to distinguish different categories. (e.g., It is easy to distinguish the zebra and the hamster by their attributes)

Image recognition is a fundamental problem, which involves a lot of research (e.g. general representation learning [9], cross-domain feature learning [10]) in multimedia community. Each class corresponds to an abstract concept, whose instances are subject to a number of variations in pose, translation, scale, occlusion, illumination, distortion, background, etc. Thus the instance space is highly complicated since samples in the same class might be drastically different in appearance. This makes image recognition need to learn class-discriminative representations. However, a few examples are insufficient to describe high-level information such as categories or concepts, and only one or very few examples are not enough to conduct the standard learning practice in deep learning.

Recently, meta-learning, pioneered by [11], [12], has been shown to outperform conventional learning on various few-shot learning problems such as few-shot image recognition [13], [14], [15], [16] and few-shot reinforcement learning [15], [17], [18]. The goal of meta-learning is to train a meta-learner on various tasks by acquiring generic knowledge, so that the meta-learner can generalize well on novel tasks with a small amount of samples. Most few-shot image recognition methods also train models across tasks, but these methods are not good enough to learn discriminative representations of novel categories. For example, on MiniImageNet [19], the state-of-the-art accuracy under 5-way 1-shot setting is inferior to 60% [20], while the accuracy of standard image recognition is up to

83.5% on ImageNet [4] whose number of categories is up to 1,000. A possible reason is that the few-shot image recognition model is apt to learn discriminative representations of training categories but could not distinguish novel categories well. Because training categories are disjoint with novel categories, which is different from the standard image recognition. A feasible approach is to establish a connection between training and novel categories. This connection facilitates the few-shot image recognition model to learn general representations, which can be used to distinguish novel categories.

In this paper, we introduce attribute learning for few-shot image recognition. Our key insight is that attributes can effectively establish connections between training categories and novel categories, and can be used to distinguish different categories (see Fig. 1). This gives a solution to learn image attributes at the training stage of few-shot image recognition. Trained with image attributes, few-shot image recognition can learn attribute-related representations to describe images in a compact way compared with only the category concept. In this way, constrained by category concepts and attributes, feature representations are less sensitive to the novel categories compared with traditional features learned from category supervision, since novel categories probably contain learned attributes. Thus we propose an attribute-guided two-layer learning framework, which is capable of obtaining general feature representations. Particularly, attribute learning is used as another learning target for few-shot image recognition in a multi-task framework, where few-shot image recognition trained over tasks and attribute learning trained over images share the same network. Besides, guided by attribute learning, features from different layers are representations of attributes at different levels like multi-scale representations [21], which are aligned for few-shot image recognition on multiple aspects. Thus a two-layer learning mechanism guided by attributes is established to capture more discriminative representations, which are complementary compared with a single-layer learning mechanism. Furthermore, the proposed framework might be independent of specific models, and two kinds of typical methods: metric-based few-shot methods (i.e., Matching Net [19] and Prototypical Net [22]) and meta-learning methods (i.e., Model-Agnostic Meta-Learning (MAML) [15]) are plugged into the proposed framework. We conduct experiments on CUB-200 [23], AWA [24] and MiniImageNet [19] dataset, and comprehensive experimental evaluations demonstrate the proposed framework effectively improves the performance of few-shot image recognition.

Our main contributions are summarized as follows:

- To the best of our knowledge, we firstly use attribute learning to connect training and novel categories for few-shot image recognition.
- We propose an attribute-guided two-layer learning framework that learns general feature representations under the multi-task learning framework in an end-to-end fashion, and further extend metric-based few-shot methods and meta-learning methods into the proposed framework.
- We verify the proposed framework on CUB-200, AWA and MiniImageNet datasets, and experimental results demonstrate the effectiveness of the proposed framework.

II. Related work

A. Few-shot learning

As an early attempt, Li $et\,al.$ propose a variational Bayesian framework for one-shot image classification [25], and a notable method called Hierarchical Bayesian Program Learning [7] is later proposed to reach human level error on the few-shot alphabet recognition tasks. More recently, a variety of methods are proposed to handle the few-shot learning problem. In this paper, we group these works into two types.

The first type is based on metric learning. This kind of methods aims to learn an embedding function that transforms data into an embedding space, such that these transformed data can be recognized with a fixed nearest neighbor [26], a linear classifier [22], [19] or a deep non-linear metric [27]. Victor $et\,al.$ [28] use a graph neural network as a metric, that goes beyond the traditional metric such as L2 metric. These approaches effectively settle the few-shot learning problem, but ignore connections between training and novel categories. In this paper, we establish connections between training and novel categories with attribute learning, and extend two representative works in these methods (i.e., Matching Net [19] and Prototypical Net [22]) into the proposed framework.

The second type leverages meta-learning solutions. This kind of methods usually contains two parts – an initial model and an adaptation strategy, which are all trained across a variety of tasks. The initial model can be implemented with a standard network, and the adaptation strategy can be implemented with non-parametric rules or parameterized networks. The typical non-parametric rule is fixed gradient-based adaptation rule such as [15], [17], [16], [29]. A branch of the parameterized network is to directly generate weights [30] or classifiers [31], augment additive weights [13] or modify activation values [14] on the initial model. The second branch is to use a recurrent neural network (RNN) as a meta-learner [32], [33]. And this meta-learner is usually augmented with memory networks [33] which uses gradients as meta information [32], [14], [13] and shows high capacities for meta-learning. In addition, a probabilistic meta-learning approach [34] have shown the advantages of learning Gaussian posteriors over model parameters. In this paper, attribute learning can be considered to learn meta information of image representations for few-shot recognition, and the classic work in this kind of methods (i.e., MAML [15]) is extended into the proposed framework.

B. Attribute learning

As a mid-level semantic cue, attributes can bridge the gap between low-level features and high-level categories. Recent research has shown that visual attributes can benefit many traditional learning problems such as image search [35], [36], image recognition [37], [38], and image caption [39], [40], [41]. For one-shot object recognition, some works use attributes [42], [43] (e.g., depth and pose) or semantic concepts [44] to generate or augment features, which are utilized to learn robust classifiers in a two-stage training strategy. Different from the above two works [43], [44], our work employs attribute learning to regularize feature representations, and is

trained in an end-to-end fashion. Moreover, attribute-based classification offers a solution to the problem of learning with disjoint training and test categories (zero-shot classification) [24], [45], [46], [47] by transferring from seen categories to unseen categories, and few-shot image recognition also has the problem of disjoint training and novel categories. In this paper, learning image attributes is used to connect training and novel categories for few-shot image recognition.

C. Few-shot representation learning

Few-shot representation learning focuses on learning transferable or adaptive representations for different tasks, e.g., few-shot image recognition needs to learn task-agnostic meta-level concepts. To learn adaptive representations, Oreshkin $et\,al.$ [48] use a task encoding network to extract the task representation, which is used to influence the behavior of the sample representation extractor with FILM [49]. To learn general or transferable representations, Zhou $et\,al.$ [50] use a deep neural network as a concept generator, which is further enhanced by equipping with a concept discriminator. In the joint learning of both the concept generator and few-shot learning model, the concept generator extracts task-agnostic meta-level concepts, and provides effective representations for few-shot learning. From this point, our work has the same target with the work [50]. Different from this work, which resorts to external data, we exploit image attributes to learn general feature representations.

III. Preliminary

We first provide formal descriptions of few-shot image recognition, and then introduce some basic methods of attribute learning.

A. Problem formulation

For few-shot image recognition, each task contains a support set and a target set, where they share the same label space. The training tasks and test tasks have the same form, but they have disjoint label space. If the support set contains K labeled examples for each of C classes in the test task, the few-shot problem is called C-way K-shot, and the support set is available for the prediction of target images.

The support set contains image-label pairs $S = \{(x_c^k, y_c) \mid c = 1, 2, ..., C; k = 1, 2, ..., K\}$ from C classes, where each class contains K labeled images. For example, under the standard setting of one-shot learning, our ultimate goal is to recognize a class with only one labeled image. Given an unlabeled image $\overline{x}$ in the target set, the goal is to predict its class $\overline{y}$ with the learned prior knowledge and the support set S. The learning target is

$$\overline{y} = \underset{y_i \in \{y_1, y_2 ... y_C\}}{argmax} P(y_i|\overline{x}, S) \tag{1}$$

where $P(y_i|\overline{x}, S)$ is the probability of classifying $\overline{x}$ to the class label y_i conditioned on S, and $\{y_1, y_2 ... y_C\}$ is the set of class labels.

B. Basic methods of attribute learning

In the field of visual attribute learning, there are many methods, such as direct attribute prediction [24], indirect attribute prediction [51] and unified multiplicative attribute prediction [52]. Moreover, visual attribute learning provides a proper way to address zero-shot classification [53], [45], which has disjoint training and test categories. Therefore, we refer to zero-shot classification and apply visual attribute learning to the few-shot image recognition. For zero-shot classification, there exists two basic methods of visual attribute learning [53]. The first one, called recognition using independent semantics (RIS), is to learn an independent classifier per semantic attribute via independent semantics, while the second method, called recognition using semantic embeddings (RULE), is to learn all semantics simultaneously using semantic embeddings. Learning an independent classifier for each attribute causes weak correlations between attributes, which leads attribute-related representations to lack of class discriminability since a category may have some related attributes. RULE is able to leverage dependencies between attributes, thus addressing the problem of independent attribute learning in RIS.

For RIS, the model learns independent attribute predictors. If RIS is implemented with deep neural networks, each attribute predictor parameter a_k relies on a common feature extractor $E(x;\theta)$ and shares the neural network parameters θ. Thus, each attribute predictor a_k of RIS takes the form:

$$a_k(x; w_k, \theta) = \sigma(w_k E(x; \theta)) \tag{2}$$

where $\sigma()$ is the sigmoid function and w_k is the parameter vector of a_k. Given a training set $D = \{(x_i, y_i^b) \mid i = 1, 2, ..., n\}$, where $y_i^b = (y_i^{b_1}, \cdots, y_i^{b_p})$ ($y_i^{b_j} \in \{0, 1\}$,$j = 1, 2, ..., p$) are p dimensional attribute labels, the loss function is

$$L_{att}^b(D) = \sum_{i=1}^{n} \sum_{j=1}^{p} L_b(a_j(x_i; w_j, \theta), y_i^{b_j}) \tag{3}$$

where L_b is a binary cross-entropy loss, w_j and θ are learned parameters.

For RULE, attribute scores are calculated simultaneously. The p dimensional attributes are

$$a(x; \theta) = M^T * E(x; \theta) \tag{4}$$

where M is 2D matrix (In the neural network, M can be implemented by a fully-connected layer of p units). Given a training set $D = \{(x_i, y_i^l) \mid i = 1, 2, ..., n\}$, where y_i^l is the class label of image x_i, the loss function is

$$L_{att}^l(D) = \sum_{i=1}^{n} L_{softmax}(G(x_i; M, \theta), y_i^l) \tag{5}$$

where $G(x_i; M, \theta) = \Phi^T * a(x_i; \theta)$, $\Phi = [\varphi(1), \cdots, \varphi(c)] \in R^{p \times c}$ and $\varphi(i)$ is p dimensional attribute vector of the class i, c is the number of classes, M and θ are learned parameters, and $L_{softmax}()$ is a cross-entropy loss with softmax outputs.

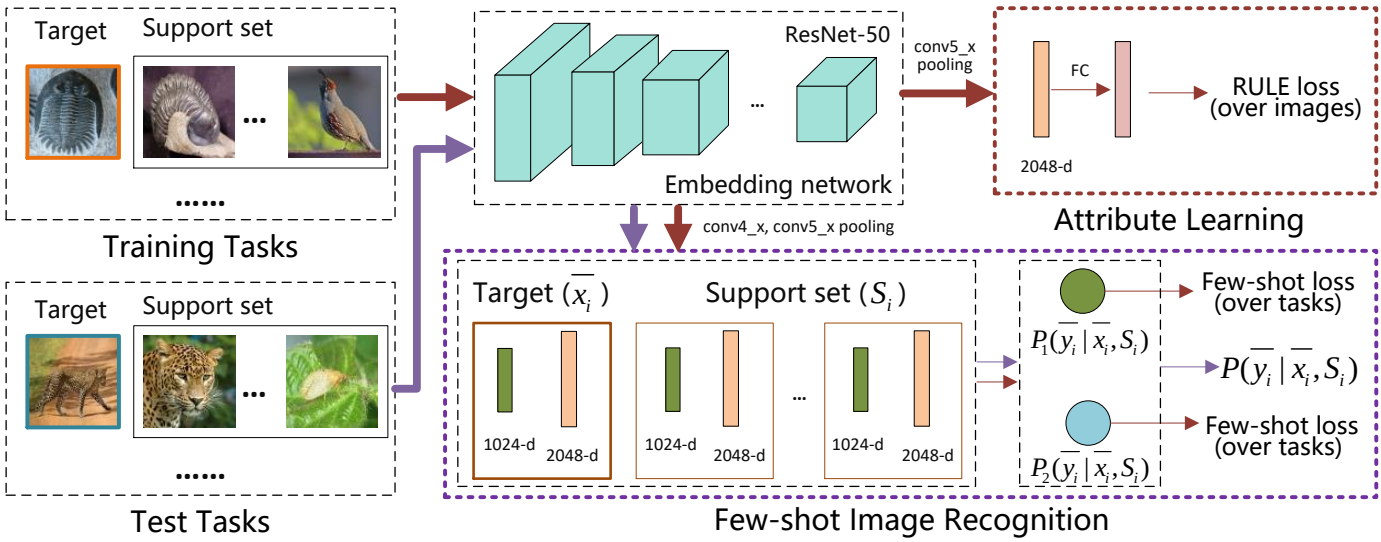


Fig. 2: The proposed attribute-guided two-layer learning framework for few-shot image recognition. The brown, blue line represent training and test process respectively. The training process contains attribute learning and few-shot image recognition while only few-shot image recognition is active in the test process. RULE is the attribute learning loss.

IV. The Proposed Framework

Fig. 2 illustrates the proposed attribute-guided two-layer learning framework, which contains two branches: attribute learning and few-shot image recognition. At the training stage, the two branches are used. At the test stage, only the few-shot image recognition branch is executed for inference, since the embedding network has learned attribute-related representations of each image.

At the training stage, attribute learning and few-shot image recognition are implemented in multi-task learning framework. Besides, two-layer features of the embedding network are utilized for few-shot image recognition respectively, which form two few-shot image recognition learning goals. Then the final optimization goal is as follows:

$$L(\Gamma) = \sum_{(S_i,S_i^a,\overline{x}_i,\overline{y}_i,\overline{y}_i^a)\in\Gamma} \{\sum_{j=1}^{2}\alpha_j L_j(S_i,\overline{x}_i,\overline{y}_i) + \beta L_{att}(S_i^a \bigcup \{(\overline{x}_i,\overline{y}_i^a)\})\} \quad (6)$$

where $L_j()$ is the j^{th} few-shot image recognition loss, $L_{att}()$ is the attribute learning loss and $(S_i, S_i^a, \overline{x}_i, \overline{y}_i, \overline{y}_i^a)$ is a task in Γ. Specifically, $\overline{y}_i$ is the class label of the target $\overline{x}_i$, $\overline{y}_i^a$ is the attribute of the target $\overline{x}_i$, S_i and S_i^a belong to the support set with same samples, the difference is S_i contains class labels of samples while S_i^a contains attributes of samples, α_j and β are hyper-parameters to balance these losses.

The proposed framework is to learn general representations for few-shot image recognition. It might be independent of specific models, and most of few-shot methods including metric-based few-shot methods and meta-learning methods can be plugged into the proposed framework. And they form attribute-guided metric-based few-shot methods and attribute-guided meta-learning methods respectively.

A. Attribute-guided metric-based few-shot method

In this section, classical metric-based few-shot methods Matching Net [19] and Prototypical Net [22] are employed. These two learners use single representation for few-shot image recognition without attribute guidance, which will be considered as baselines in the section of experiments. In the Matching Net [19], the few-shot image recognition loss is $L_j(S_i,\overline{x}_i,\overline{y}_i) = -logP_j(\overline{y}_i|\overline{x}_i,S_i)$ denoted as MC, and the $P_j(\overline{y}_i|\overline{x}_i,S_i)$ is

$$P_j(\overline{y}_i|\overline{x}_i,S_i) = \frac{\sum_{(x_i,y_i)\in S_i} \mathbf{I}\{y_i=\overline{y}_i\} e^{Cos(f_j(\overline{x}_i),f_j(x_i))}}{\sum_{(x_i,y_i)\in S_i} e^{Cos(f_j(\overline{x}_i),f_j(x_i))}} \quad (7)$$

where $f_j()$ denotes an embedding network to obtain features for $P_j()$, $\mathbf{I}\{\mathrm{c}\}$is an indicator function, and $Cos(,)$ is the Cosine distance. And $\mathbf{I}\{\mathrm{c}\}= 1$ if c is true and 0 otherwise. In the Prototypical Net [22], the few-shot image recognition loss is $L_j(S_i,\overline{x}_i,\overline{y}_i) = -logP_j(\overline{y}_i|\overline{x}_i,S_i)$ denoted as PE, but the $P_j(\overline{y}_i|\overline{x}_i,S_i)$ is

$$P_j(\overline{y}_i|\overline{x}_i,S_i) = \frac{e^{-Euc(f_j(\overline{x}_i),\sum_{(x_i,y_i)\in S_i}\frac{\mathbf{I}\{y_i=\overline{y}_i\}f_j(x_i)}{K})}}{\sum_{a=1}^{a=C} e^{-Euc(f_j(\overline{x}_i),\sum_{(x_i,y_i)\in S_i}\frac{\mathbf{I}\{y_i=y_a\}f_j(x_i)}{K})}} \quad (8)$$

where $Euc(,)$ is the Euclidean distance.

B. Attribute-guided meta-learning method

In this section, a typical meta-learner MAML [15] is plugged into the proposed framework. Inspired by fine-tuning, MAML uses gradient-based optimization without requiring additional parameters or model modification, and can be generally applied to any model as long as the gradient of the objective function is available. It learns initial parameters

of the model such that in each task the performance of test data is maximized after only one or a few gradient steps. To be more specific, given a task, MAML obtains parameters by transforming the original parameters with only one or a few steps of gradient descent on the support set S, and these updated parameters are suitable for the target set. This method is also considered as one baseline in the comparative methods.

Inspired by the work [50], we also use multi-layer perceptron (MLP) as a meta-learner $\mathcal{M}_j(f_j(); w_j)$, where inputting features are obtained by $f_j()$, w_j are parameters of meta-leaner and it's outputs are C-way probabilities. Given a task, only the parameters w_j are updated. For one-step gradient descent, this would be computed as Eq. 9, where α is a fixed adaptive learning rate, $\sum_{(x_i,y_i)\epsilon\mathcal{S}_i}\mathcal{L}(f_j(x_i), y_i; w_j)$ is the loss on the S_i and $\nabla_{w_j}\sum_{(x_i,y_i)\epsilon\mathcal{S}_i}\mathcal{L}(f_j(x_i), y_i; w_j)$ is the corresponding gradient with respect to w_j. The $\mathcal{L}(f_j(x_i), y_i; w_j)$ is a cross entropy loss of the output of $\mathcal{M}_j(f_j(x_i); w_j)$. The few-shot image recognition loss is $L_j(S_i, \overline{x}_i, \overline{y}_i) = -logP_j(\overline{y}_i|\overline{x}_i, S_i)$ and the $P_j(\overline{y}_i|\overline{x}_i, S_i)$ is computed via Eq. 10. More generally, a function $G()$ represents multiple steps gradient descent or update rule: $w^{'}_{j,i} := G(S_i; w_j)$. Note that during meta-optimization, the gradient is computed with respect to initial parameters w_j, but the test loss is computed with respect to task-specific parameters $w^{'}_{j,i}$. An important point is that optimization parameters are w_j. More details can be found in [15].

$$w^{'}_{j,i} := w_j - \alpha\nabla_{w_j}\sum_{(x_i,y_i)\epsilon\mathcal{S}_i}\mathcal{L}(f_j(x_i), y_i; w_j) \tag{9}$$

$$P_j(\overline{y}_i|\overline{x}_i, S_i) = Softmax(\mathcal{M}_j(f_j(\overline{x}_i); w^{'}_{j,i}), \overline{y}_i) \tag{10}$$

C. Implementation

We use Resnet-50 as the embedding network, and the above two attribute learning methods. And which one is better for few-shot image recognition will be explored in the experimental section. In attribute learning branch, the final conv5_x features of ResNet-50 are averagely pooled to a 2048-d vector for attribute learning, where the parameter w_k in the RIS or M in the RULE needs to be learned. In few-shot image recognition branch, conv5_x, conv4_x features of ResNet-50 are averagely pooled to a 2048-d, 1024-d vector to form a two-layer learning, where different layers can have different forms. While this architecture is simple, adding more-layer comparisons does not obtain better experimental results. Training these parameters would not take much time, since the number of parameters of few-shot image recognition ($f_j()$ or $\{f_j(), w_j\}$) is much larger than that of w_k, M. In the inference stage, the proposed method has the similar time consumption with methods [19], [22], [15], since only few-shot image recognition branch is executed. And different layers $P_j(\overline{y}_i|\overline{x}_i, S_i)$ are calculated respectively, sum of which is used for the final decision.

V. Experiments

We will validate the effectiveness of the proposed framework by considering the following questions. Q1: Is the proposed framework effective for few-shot image recognition? Q2: If similarities between training categories and test categories are enhanced, does the performance of few-shot image recognition improve? Q3: By learning attributes in few-shot image recognition, are feature representations of test categories more discriminative?

A. Datasets

The proposed method is validated on datasets of annotated attributes CUB-200 and AwA datasets. In addition, we also conduct experiments on MiniImageNet dataset by introducing high-level labels as attributes.

CUB-200. It contains 11,788 images of 200 different bird species, with 312-dim attributes at class level. Based on CUB-200, we construct two datasets: **CUB-200-S** and **CUB-200-R**. CUB-200-S contains 140 classes (Class numbers are $1 \sim 140$) for training, 20 classes (Class numbers are $141 \sim 160$) for validation, and the remaining 40 classes as novel classes for testing. CUB-200-R contains the same number of classes with CUB-200-S for training, validation and test respectively, but these classes are selected randomly from CUB-200. CUB-200-S is split by a sequence number of category, where categories with serial numbers may fall into a senior category. For example, the 1^{th} , 2^{th} and 3^{th} class in CUB-200 are black-footed albatross, laysan albatross and sooty albatross, respectively, so the senior category is albatross. Categories in the same senior category are more similar than categories in different senior categories. Thus, in CUB-200-R, for a test category there exits a higher probability to find a category belonging to the same senior category from training categories. As a result, similarities between training categories and test categories are higher in CUB-200-R.

AwA. It includes 30,475 images from 50 common animals categories, with 85-dim class-level attributes. We use 30 categories (Category numbers are $1 \sim 30$) for training, 10 categories (Category numbers are $31 \sim 40$) for validation, and the remaining 10 categories as novel categories for testing.

MiniImageNet. The MiniImageNet dataset [19] consists of 100 classes, each of which contains 600 images with size of 84×84 pixels. We follow the split introduced by [32] with 64, 16, and 20 classes for training, validation and testing, respectively. As there exits a tree graph in the ImageNet dataset, where each leaf node represents a category and each leaf node has the only path to root node. We use the path as their attributes. And attributes of each category are all non-leaf nodes of the tree in MiniImageNet, and attribute value of each category is 1 if the corresponding node appears in path of this category and 0 otherwise.

B. Experimental settings

Architectures of embedding network. A CNNs [19], [22] is widely used as the embedding network $f()$ denoted as CNN-4, which consists of four convolutional layers. And each convolutional layer is devised with a 3×3 convolution and 64 filters followed by batch normalization, a ReLU non-linearity and a 2×2 max-pooling. The input of the image is 84×84, and the final output embedding space dimension is 1,600 for

few-shot image recognition. In order to meet attribute-guided few-shot image recognition, CNN-4 adds two full-connected layers with 1024 neurons for attribute learning. In addition, we use ResNet [54] as the embedding function $f()$, whose inputting size is 224×224. The two networks are trained from scratch in all experiments.

Training details. Similar to [19], the few-shot model is trained on tasks (episodes). And each task is formed by randomly selecting C classes with K labeled samples per class to act as the support set $S = \{(x_c^k, y_c) \mid c = 1, 2, ..., C; k = 1, 2, ..., K\}$, as well as the same C classes with N samples per class to serve as the target set $T = \{(x_c^n, y_c) \mid c = 1, 2, ..., C; n = 1, 2, ..., N\}$. Specifically, when only using the MC loss, each task contains $C = 5$ classes, each of which has $K \in \{1, 5\}$ examples in the support set and $N = 5$ examples in the target set. And each batch contains 4 and 2 tasks under 1-shot setting and 5-shot setting, respectively. When using the PE loss, we follow the settings [22], where each task contains $C = 30$ and $C = 20$ classes under 1-shot setting and 5-shot setting, respectively, $N = 5$ and each batch contains 1 task. When MAML is extended into the proposed method, the settings of batch size, C, K, and N are the same as the setting of MC loss. And the meta-learner is implemented by an one-layer perceptron. At the training stage, the number of updating gradient step is 3. But the number of updating gradient step is 5 at the test stage. The weight decay is 0.005 on MiniImageNet and AWA, and 0.01 on CUB- 200. The weights of optimizing loss α_j and β are all set to 1, and we use Adam [55] as the optimizer.

In RULE, a p dimensional attribute vector of class i, denoted as $\varphi(i)$, is not a binary vector. For example, there are p binary attributes of class i, then $\varphi_k(i) = 1$ when class i contains attribute k, otherwise $\varphi_k(i) = -1$. To support different degrees of certainty on class/attribute associations, continuous attributes can be easily implemented by making $\varphi_k(i) \in [-1, 1]$.

Evaluation. For performance evaluation, we randomly sample 600 times test from corresponding datasets. Each test contains 3 tasks. And each task contains $C = 5$ classes, each of which includes $K \in \{1, 5\}$ examples in the support set and $N = 5$ examples in the target set. The results averaged over the sampled 600 times test are reported with mean accuracy (MA) and 95% confidence intervals (Cls), which statistically describe the inherent uncertainty in performance estimation like standard deviation. The smaller the Cls, the more precise the MA performance.

C. Ablation study

To verify the effectiveness of different components in the proposed framework, we consider the following variants.

- **OAL** (Only Attribute Learning). In this baseline, only attribute learning is utilized for training, and the trained model is leveraged to extract image features for few-shot image recognition testing.
- **OFL** (Only Few-shot Learning). In this baseline, the standard few-shot image recognition is conducted for training and testing.
- **AGFL** (Attribute-Guided Few-shot Learning). This is the proposed method. Compared with OFL, a difference is that attribute learning is utilized for few-shot image recognition.
- **CGFL** (Class-Guided Few-shot Learning). This method [50] is similar to AGFL, but a difference is that image category is used as guided information.

In addition, single-layer and two-layer features are set for few-shot image recognition respectively. Another setting is data augmentation. Specifically, each image resized to 240×240 is randomly cropped to a 224×224 region for training. In the test stage, we use two schemes: i) $\dagger_1$: Each image is cropped to the central 224×224 region. ii) $\dagger_2$: Each image is cropped to five representative 224×224 regions, then it becomes a C-way $5K$-shot learning problem. Since a target image also contains five samples, the final decision depends on the sum of the five predicted probabilities.

D. Results and Analysis

AGFL vs (OAL, OFL). Comparative results of OFL, OAL and AGFL on CUB-200 and AWA are shown in Table I. The results clearly indicate that AGFL outperforms both OFL and OAL by a wide margin when using the embedding network ResNet-50. This illuminates that the proposed method AGFL is effective for few-shot image recognition. But when CNN-4 is used as the embedding network, AGFL is inferior to OFL. The possible reasons are: i) It may be difficult to excavate attribute information from a small-scale image such as 84×84. ii) A shallow network like CNN-4 is not capable of both few-shot learning and attribute learning. Besides, AGFL with the RULE loss outperforms that with the RIS loss. For this reason, we mainly compare experimental results when using a deep embedding network such as ResNet and the RULE attribute learning loss.

OAL vs OFL. It is worthy to note that the performance of OAL is comparable with or better than OFL. This explains that general representations of images may be obtained with only attribute learning. Thus it is a good choice to use attribute-related representations for few-shot image recognition.

AGFL vs CGFL. The performance of CGFL is defeated by AGFL with a large margin. It can be obviously obtained that the image attribute is more suitable for few-shot image recognition than the image class, which also explains that the attribute-related representations are less sensitive to the novel categories compared with traditional features learned from category supervision.

Deep network vs shallow network. On AWA and CUB-200, the performance of ResNet-50 is inferior to CNN-4 when using OFL, as illustrated in Table I, while OFL with ResNet-50 obtains better performance than that with CNN-4 on MiniImageNet, as illustrated in Table II. Note that ResNet-$50_{(5)}$ of OFL has more parameters than ResNet-$50_{(4)}$ of OFL, which are not the same in terms of capacity. The main reasons are: i) On CUB-200 and AWA, the seriously insufficient training images can not feed to a deep neural network like ResNet-50, which more likely leads to overfitting compared with a shallow network like CNN-4. ii) Compared with CUB-200 and

TABLE I: 5-way MA (%) + Cls (%) on CUB-200 and AWA with the MC loss. The lower case number in ResNet-50 means the number of convolutional features are averagely pooled for few-shot image recognition. The 'RIS' means using the RIS loss, otherwise the RULE loss.

Method	$f()$	CUB-200-S		AwA		CUB-200-R	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
OFL	CNN-4	43.18±0.56	57.98±0.62	41.42±0.54	53.82±0.51	54.84±0.57	**67.91±0.53**
AGFL	CNN-4	40.46±0.53	54.18±0.61	41.38±0.53	52.59±0.52	48.50±0.55	63.23±0.54
OAL	ResNet-$50_{(5)}$	40.67±0.57	49.77±0.54	39.74±0.51	53.68±0.51	47.50±0.60	60.13±0.54
OFL	ResNet-$50_{(4)}$	36.65±0.53	49.63±0.58	36.61±0.54	44.22±0.50	43.16±0.62	53.36±0.58
OFL	ResNet-$50_{(5)}$	35.99±0.50	44.34±0.52	37.94±0.56	46.90±0.52	41.53±0.59	53.99±0.59
CGFL	ResNet-$50_{(5)}$	36.21±0.52	44.52±0.53	42.09±0.57	49.74±0.54	45.52±0.59	53.22±0.56
AGFL(RIS)	ResNet-$50_{(5)}$	40.54±0.58	53.05±0.61	40.00±0.55	52.22±0.53	42.30±0.61	54.27±0.59
AGFL	ResNet-$50_{(4)}$	45.18±0.60	58.17±0.60	44.24±0.58	54.34±0.53	**57.16±0.59**	64.69±0.53
AGFL	ResNet-$50_{(5)}$	42.74±0.55	**61.61±0.58**	43.10±0.58	54.46±0.53	52.89±0.60	65.05±0.55
AGFL	ResNet-$50_{(4,5)}$	**46.11±0.60**	59.08±0.60	**46.76±0.59**	**55.61±0.51**	56.19±0.64	65.53±0.56

TABLE II: 5-way MA (%) + Cls (%) on MiniImageNet using the MC loss.

Method	$f()$	1-shot	5-shot
OFL	CNN-4	46.77±0.58	59.69±0.53
OFL	ResNet-$50_{(4)}$	48.90±0.58	60.34±0.58
OFL	ResNet-$50_{(5)}$	46.05±0.54	59.13±0.54
OFL	ResNet-$50_{(4,5)}$	46.63±0.58	58.87±0.57
$\text{OFL}_{\dagger_1}$	ResNet-$50_{(4,5)}$	50.01±0.62	61.25±0.56
$\text{OFL}_{\dagger_2}$	ResNet-$50_{(4,5)}$	50.50±0.61	61.68±0.57
AGFL	ResNet-$50_{(4)}$	50.43±0.61	62.48±0.53
AGFL	ResNet-$50_{(5)}$	50.78±0.59	64.16±0.53
AGFL	ResNet-$50_{(4,5)}$	52.47±0.60	64.71±0.54
AGFL	ResNet-$50_{(3,4,5)}$	50.60±0.60	64.54±0.54
$\text{AGFL}_{\dagger_1}$	ResNet-$50_{(4,5)}$	52.80±0.58	65.85±0.53
$\text{AGFL}_{\dagger_2}$	ResNet-$50_{(4,5)}$	**53.51±0.59**	**66.19±0.55**

AWA, MiniImageNet contains enough training images, which can feed ResNet-50 without overfitting. Although there exists the overfitting issue when ResNet-50 is used as the embedding network on AWA and CUB-200, AGFL still outperforms OFL with the embedding network CNN-4 on the most datasets. The probable interpretation is that multi-task learning (i.e., attribute learning, few-shot learning) adds more supervision, which can relieve the overfitting issue and learns more regularized feature space.

Similarities of categories for few-shot image recognition. As shown in Table I, the performance of CUB-200-R obviously exceeds that of CUB-200-S with the same methods. Besides, similarities between training categories and test categories in CUB-200-R are greater than similarities of CUB-200-S (see Section V-A). Therefore, similarities between training categories and test categories are enhanced, then the performance improves.

Analysis of AGFL. Fig. 3 shows visual representations of features of the total training and test categories in CUB-200. The distributions of representations of total training and test categories in CUB-200-R are more similar than that in CUB-200-S. The possible reason is that for each test category there may exits a similar training category in CUB-200-R. By learning attributes, the feature representations of test categories are more discriminative, since the pentagrams of AGFL are more separated than that of OFL in Fig. 3 and Fig. 4.

Besides, since few-shot representations are guided by attribute learning in the training stage, distributions of representations of training categories and test categories all have non-smooth edges (see red boxes in Fig. 3), which are attribute-related since the edge is smooth without attribute learning. In the non-smooth edges, some clusters in distributions of representations of test categories probably are found in that of training categories at the same location. From what has been discussed above, it can be obtained that attribute learning establishes a good connection between training categories and test categories and attribute-related representations are discriminative.

Two-layer learning mechanism. As shown in Table II, a two-layer learning mechanism AGFL is more beneficial to few-shot image recognition than single learning mechanism AGFL. It is worthy to note that the two-layer learning mechanism improves the performance while this does not improve the performance without attribute learning since OFL with ResNet-$50_{(4)}$ outperforms OFL with ResNet-$50_{(4,5)}$. A possible explanation is that features of different layers are more likely to learn different concepts in AGFL compared with OFL, and different concepts are complementary to benefit for few-shot image recognition. Moreover, AGFL obtains tiny gains with data augmentation, especially for the $\dagger_1$, which illustrates that AGFL may not be data-hungry. However, the effectiveness of two-layer learning mechanism is not prominent especially on CUB (see Table I). A possible explanation is that it is difficult to feed ResNet-$50_{(5)}$ with deficient training data on CUB since ResNet-$50_{(4)}$ of AGFL defeats ResNet-$50_{(5)}$ of AGFL with above 3% gains under 1-shot setting.

Table III lists experimental results of using PE loss, which again verifies the effectiveness of attribute guidance and two-layer learning mechanism. Compared with low-resolution images on MiniImageNet, we obtain about 2% gains with high-resolution images under both 1-shot setting and 5-shot setting at the same methods, since the input of ResNet-50

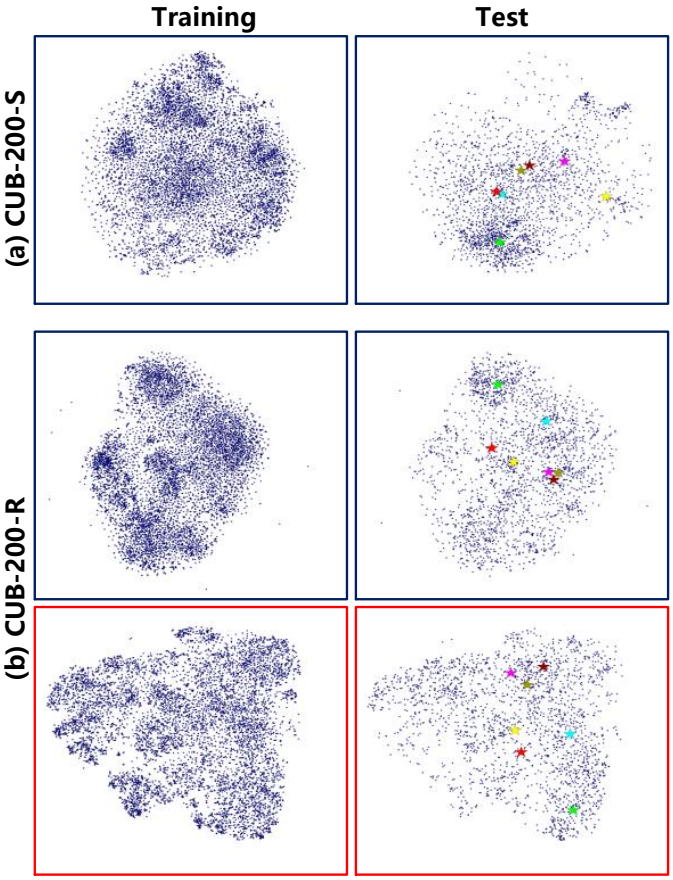


Fig. 3: The visualization of image representations of all training and test categories on CUB-200-S and CUB-200-R using t-SNE [56]. Each image is visualized as one point. The scatter plots in blue and red boxes come from OFL and AGFL, respectively. In test categories, the centers of some categories are represented by pentagrams of different colors, and the pentagram of same color represents the same category across different datasets or methods.

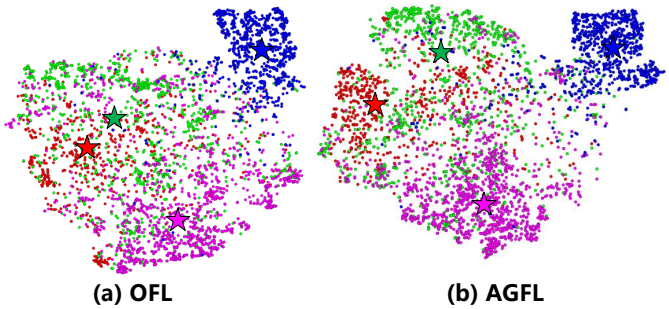


Fig. 4: The visualization of image representations of some test categories with OFL and AGFL on AWA. Different colors denote different categories, whose centers are represented by pentagrams of corresponding colors. It is obvious that the categories with green and red color of AGFL are more separated than that of OFL.

TABLE III: 5-way MA (%) + Cls (%) on MiniImageNet. We use the PE loss at single-layer learning, while the PE and the MC loss are used on averagely pooled conv4_x and conv5_x at two-layer learning mechanism, respectively.($*$: the higher resolution images from ImageNet, w/o 20c: without extra 20 categories for training, w 20c: with extra categories for training.)

Method	$f()$	**1-shot**	**5-shot**
w/o 20c			
$\text{OFL}_{\dagger_2}$	$\text{ResNet-50}_{(4)}$	49.52±0.58	68.30±0.50
$\text{AGFL}_{\dagger_2}$	$\text{ResNet-50}_{(4)}$	52.46±0.59	70.10±0.54
$\text{AGFL}_{\dagger_2}$	$\text{ResNet-50}_{(4,5)}$	**54.52±0.61**	**71.67±0.48**
$\text{AGFL}_{\dagger_2}$	$\text{ResNet-50}_{(3,4,5)}$	54.42±0.61	69.90±0.49
$\text{OFL}^*_{\dagger_2}$	$\text{ResNet-50}_{(4)}$	52.16±0.61	70.11±0.50
$\text{AGFL}^*_{\dagger_2}$	$\text{ResNet-50}_{(4,5)}$	**56.59±0.64**	**73.58±0.48**
w 20c			
$\text{OFL}^*_{\dagger_2}$	$\text{ResNet-50}_{(4)}$	53.78±0.60	73.29±0.51
$\text{AGFL}^*_{\dagger_2}$	$\text{ResNet-50}_{(4,5)}$	**60.76±0.58**	**77.57±0.47**

is bigger than 84×84. This suggests that few-shot image recognition also needs high-resolution images. Table IV lists experimental results based on MAML. Since MAML needs more computational resources than metric-based few-shot methods with the same embedding network, we use a smaller embedding network ResNet-34 to match the same training/test settings with metric-based few-shot methods. Experimental results demonstrate the proposed method (both attribute-guided learning and two-layer learning mechanism) is also suitable for MAML.

Besides, we obtain the best performance with two-layer comparisons combined high-layer features (i.e., $\text{ResNet-50}_{(4,5)}$ in Table II and III). And too many comparisons such as three-layer comparisons combined low-layer features (see $\text{ResNet-50}_{(3,4,5)}$ in Table II and III) may not benefit to few-shot image recognition. The possible reason is that image recognition needs to use abstract concepts such as high-level features while low-layer features are not abstract enough.

Extra training categories. To further explore our proposed framework, extra categories, similar but different from test (novel) categories, are added for training. The extra categories have the same parent nodes with test categories in the tree-structure graph of the ImageNet, where we randomly sample 600 images per category and the total number of extra categories is 20. Both two kinds of original few-shot learning and our corresponding two-layer AGFL benefit from the extra training data. Note that, as these extra categories are used, the performance of image recognition has improved slightly with metric-based OFL (see Table III), while corresponding

TABLE IV: 5-way MA (%) + Cls (%) on MiniImageNet with MAML and attribute-guided MAML.

Method	$f()$	**1-shot**	**5-shot**
w/o 20c			
$\text{OFL}^*_{\dagger_2}$	$\text{ResNet-34}_{(4)}$	49.72±0.56	65.26±0.49
$\text{AGFL}^*_{\dagger_2}$	$\text{ResNet-34}_{(4)}$	51.65±0.56	68.89±0.49
$\text{AGFL}^*_{\dagger_2}$	$\text{ResNet-34}_{(5)}$	52.36±0.55	67.94±0.47
$\text{AGFL}^*_{\dagger_2}$	$\text{ResNet-34}_{(4,5)}$	**56.49±0.59**	**71.34±0.49**
w 20c			
$\text{OFL}^*_{\dagger_2}$	$\text{ResNet-34}_{(4)}$	56.48±0.58	71.89±0.50
$\text{AGFL}^*_{\dagger_2}$	$\text{ResNet-34}_{(4,5)}$	**61.40±0.58**	**76.87±0.47**

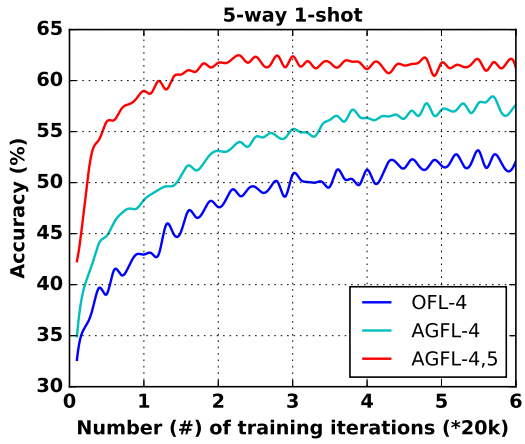


Fig. 5: The curve of 5-way 1-shot accuracy in validation set with metric-based few-shot learning method.

TABLE V: 5-way MA (%) + Cls (%) of our proposed method and other state-of-the-art methods on MiniImageNet dataset.

Method	1-shot	5-shot
Matching Net [19]	43.56±0.84	55.31±0.73
Prototypical Net [22]	49.42±0.78	68.20±0.66
Meta-LSTM [32]	43.44±0.77	60.60±0.71
MAML [15]	48.70±1.84	63.11±0.92
MetaNet [13]	49.21±0.96	-
Meta-SGD [17]	50.47±1.87	64.03±0.94
MM [57]	53.37±0.48	66.97±0.35
SNAIL [18]	55.71±0.99	68.88±0.92
MEPS [58]	51.03±0.78	67.96±0.71
Relation Net (navie) [27]	51.38±0.82	67.07±0.69
Relation Net (deeper) [27]	**57.02±0.92**	71.07±0.69
adaCNN[14]	56.88±0.62	71.94±0.57
Our (AGFL+Prototypical Net)	56.59±0.64	**73.58±0.48**
Our (AGFL+MAML)	56.49±0.59	71.34±0.49
PPA$^{\diamond}$ [20]	59.60±0.41	73.74±0.19
DEML+Meta-SGD$^{\diamond}$ [50]	58.49±0.91	71.28±0.69
Our (AGFL+Prototypical Net)$^{\diamond}$	60.76±0.58	**77.57±0.47**
Our (AGFL+MAML)$^{\diamond}$	**61.40±0.58**	76.87±0.47

$^{\diamond}$the number of training categories is more than 64

two-layer AGFL obtains about 4% gains under both 1-shot and 5-shot settings. It can be obtained that AGFL learns more general image representations compared with OFL.

Convergence. The curve of 5-way 1-shot accuracy in validation set is shown in Fig. 5. It can be observed that: i) The two-layer learning AGFL have better convergence than single learning AGFL and OFL. ii) The single learning AGFL have slight better convergence than OFL. Thus we obtain that both attribute-guided learning and two-layer learning mechanism facilitate convergence of model.

Comparison with state-of-the-art methods. With the standard training data (64 categories), our AGFL+Prototypical Net obtains a comparable performance under 1-shot setting and 1.5% gains under 5-shot setting, compared with the Relation Net [27]. However, our method has better convergence because our method converges after 40k iterations (see Fig. 5) while Relation Net requires more than 200k iterations. Moreover, with extra categories, our two methods outperform the state-of-the-art performance over 1% and 3% improvements under 1-shot setting and 5-shot setting, respectively. The PPA [20] needs a pre-trained model. In contrast, our work is more flexible since it does not use a pre-trained model and trains from scratch. The DEML+Meta-SGD [50] uses external images of 200 categories while our work leverages less data (extra 20 categories).

VI. Conclusions

In this paper, we introduce attribute learning to connect training categories and novel categories for few-shot image recognition, and propose an attribute-guided two-layer learning framework. In this framework, few-shot image recognition and attribute learning leverage the same network in the multi-task learning framework, and the two-layer features associated with attributes are utilized for few-shot image recognition respectively. Two typical few-shot methods including metric-based methods (i.e., Matching Net and Prototypical Net) and meta-learning methods (i.e., MAML) are plugged into the proposed framework. Experimental results on CUB-200, AWA and MiniImageNet dataset demonstrate the effectiveness of our proposed framework. In addition, we find the proposed framework has a good convergence and achieves performance improvement.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[3] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3485–3492.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *Association for the Advancement of Artificial Intelligence*, 2017, pp. 4278–4284.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 3.

[6] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Advances in neural information processing systems*, 2017, pp. 4467–4475.

[7] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction." *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 198–207, 2017.

[10] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 64–78, 2014.

[11] J. Schmidhuber, "Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook," Ph.D. dissertation, Technische Universität München, 1987.

[12] D. K. Naik and R. Mammone, "Meta-neural networks that learn by learning," in *IJCNN.*, vol. 1. IEEE, 1992, pp. 437–442.

[13] T. Munkhdalai and H. Yu, "Meta networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2554–2563.

[14] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler, "Rapid adaptation with conditionally shifted neurons," in *International Conference on Machine Learning*, 2018, pp. 3661–3670.

[15] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017, pp. 1126–1135.

[16] Y. Lee and S. Choi, "Gradient-based meta-learning with learned layerwise metric and subspace," in *International Conference on Machine Learning*, 2018, pp. 2927–2936.

[17] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few shot learning," *arXiv preprint arXiv:1707.09835*, 2017.

[18] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *International Conference on Learning Representations*, 2018.

[19] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[20] S. Qiao, C. Liu, W. Shen, and A. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7229–7238.

[21] S. Yang and D. Ramanan, "Multi-scale recognition with dag-cnns," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1215–1223.

[22] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4080–4090.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[24] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[25] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, vol. 2, 2015.

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

[28] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *International Conference on Learning Representations*, 2018.

[29] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 719–11 727.

[30] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in Neural Information Processing Systems*, 2016, pp. 523–531.

[31] L. Zhou, P. Cui, S. Yang, W. Zhu, and Q. Tian, "Learning to learn image classifiers with visual analogy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 497–11 506.

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

[33] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International Conference on Machine Learning*, 2016, pp. 1842–1850.

[34] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, "Recasting gradient-based meta-learning as hierarchical bayes," in *International Conference on Learning Representations*, 2018.

[35] A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: Image search with relative attribute feedback," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2973–2980.

[36] J. Cai, Z.-J. Zha, W. Zhou, and Q. Tian, "Attribute-assisted reranking for web image retrieval," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 873–876.

[37] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *European Conference on Computer Vision*. Springer, 2010, pp. 155–168.

[38] D. Roy, S. R. M. Kodukula *et al.*, "Unsupervised universal attribute modelling for action recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1672–1680, 2019.

[39] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[40] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894–4902.

[41] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-quality image captioning with fine-grained and semantic-guided visual attention," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1681–1693, 2018.

[42] S. Jiang, W. Min, Y. Lv, and L. Liu, "Few-shot food recognition via multi-view representation," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2020.

[43] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos, "Aga: Attribute-guided augmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3328–3336.

[44] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal, "Semantic feature augmentation in few-shot learning," *arXiv preprint arXiv:1804.05298*, 2018.

[45] Y. Li, J. Zhang, J. Zhang, and K. Huang, "Discriminative learning of latent features for zero-shot recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7463–7471.

[46] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 69–77.

[47] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen, "Learning discriminative latent attributes for zero-shot classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4223–4232.

[48] B. Oreshkin, P. R. Lopez, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Advances in neural information processing systems*, 2018, pp. 719–729.

[49] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[50] F. Zhou, B. Wu, and Z. Li, "Deep meta-learning: Learning to learn in the concept space," *arXiv preprint arXiv:1802.03596*, 2018.

[51] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1629–1636.

[52] K. Liang, H. Chang, S. Shan, and X. Chen, "A unified multiplicative framework for attribute learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2506–2514.

[53] P. Morgado and N. Vasconcelos, "Semantically consistent regularization for zero-shot recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2037–2046.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

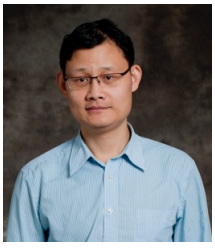
[56] L. van der Maaten, G. Hinton, and L. V. D. Maaten, "Visualizing data using t-sne," *JMLR*, vol. 9, no. 2605, pp. 2579–2605, 2008.

[57] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4080–4088.

[58] W.-H. Chu, Y.-J. Li, J.-C. Chang, and Y.-C. F. Wang, "Spot and learn: A maximum-entropy patch sampler for few-shot image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6251–6260.

Yaohui Zhu received the B.E., M.E. degree from Shenyang Aerospace University, Shenyang, China, in 2013, 2016 respectively. He is currently working toward the Ph.D. degree in the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests include computer vision, and machine learning, especially for visual relationship detection and few-shot learning.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, in 2008 and M.E. degree from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2015, respectively. He is currently an associate professor at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include multimedia content analysis, understanding and applications, food computing and geo-multimedia computing. He has authored and co-authored more than 30 peer-referenced papers in relevant journals and conferences, including ACM Computing Surveys, IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, ACM TOMM, IEEE Multimedia Magazine, ACM Multimedia, IJCAI, AAAI, etc. He is the reviewer of some international journals including IEEE Trans. on Multimedia, IEEE Trans. on Cybernetics, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Neural Network and Learning System, ACM TOMM, etc. As the leading guest editor, he organized some special issues on international journals, including IEEE Multimedia Magazine, Multimedia Tools and Applications. He is the recipient of 2016 ACM TOMM Nicolas D. Georganas Best Paper Award and the 2017 IEEE Multimedia Magazine Best Paper Award.



Shuqiang Jiang (SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS), Beijing and a professor in University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 150 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE and CCF, member of ACM, Associate Editor of ACM TOMM, IEEE Multimedia, Multimedia Tools and Applications. He is the vice chair of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He is the general chair of ICIMCS 2015, program chair of ACM Multimedia Asia2019 and PCM2017. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM.