Multi-attention Meta Learning for Few-shot Fine-grained Image Recognition

Yaohui Zhu[1,2] , **Chenlong Liu**[1,2] , **Shuqiang Jiang**[1,2]

[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China

[2]University of Chinese Academy of Sciences, Beijing, China

{yaohui.zhu, chenlong.liu}@vipl.ict.ac.cn, sqjiang@ict.ac.cn

Abstract

The goal of few-shot image recognition is to distinguish different categories with only one or a few training samples. Previous works of few-shot learning mainly work on general object images. And current solutions usually learn a global image representation from training tasks to adapt novel tasks. However, fine-gained categories are distinguished by subtle and local parts, which could not be captured by global representations effectively. This may hinder existing few-shot learning approaches from dealing with fine-gained categories well. In this work, we propose a multi-attention meta-learning (MattML) method for few-shot fine-grained image recognition (FSFGIR). Instead of using only base learner for general feature learning, the proposed meta-learning method uses attention mechanisms of the base learner and task learner to capture discriminative parts of images. The base learner is equipped with two convolutional block attention modules (CBAM) and a classifier. The two CBAM can learn diverse and informative parts. And the initial weights of classifier are attended by the task learner, which gives the classifier a task-related sensitive initialization. For adaptation, the gradient-based meta-learning approach is employed by updating the parameters of two CBAM and the attended classifier, which facilitates the updated base learner to adaptively focus on discriminative parts. We experimentally analyze the different components of our method, and experimental results on four benchmark datasets demonstrate the effectiveness and superiority of our method.

1 Introduction

Fine-grained image recognition aims to distinguish different subordinate categories belong to the same entry-level category (e.g., various bird species [Wah *et al.*, 2011], dog [Khosla *et al.*, 2011] species). Different subordinate categories are distinguished by subtle and local differences, which makes fine-grained image recognition more difficult than general image recognition. Most existing fine-grained image recognition methods heavily rely on large-scale annotated training samples before learning a robust classifier [Zhang *et al.*, 2014; Xiao *et al.*, 2015; Fu *et al.*, 2017]. However, annotating the fine-grained sub-categories requires strong expertise. For example, accurately identifying different bird species may need assistance from ornithologists, which is significantly expensive compared with the generic object annotation. Besides, it is hard to collect well-labeled samples of some categories, e.g., endangered, rare species. How to deal with the fine-grained image recognition with limited labeled samples is a desirable research problem in computer vision and artificial intelligence community, which has not been much explored before.

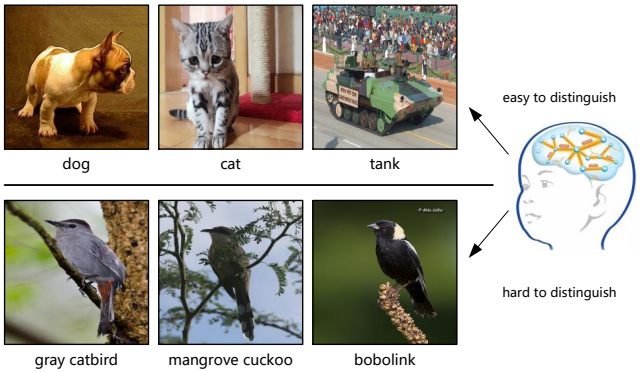

Figure 1: An example of general one-shot image recognition (top) and one-shot fine-grained image recognition (bottom). As for young children, it is easy to learn general concepts with only one image of each category, while it is more difficult to distinguish fine-gained categories with one sample of each category.

Two-year-old children can distinguish different generic categories after seeing a few images of them [Smith and Slone, 2017], while they may be confused about fine-grained categories with limited samples, as illustrated in Figure 1. Recently, few-shot image recognition has been widely studied [Vinyals *et al.*, 2016; Snell *et al.*, 2017; Sung *et al.*, 2018] to identify novel category with only one or a few samples. However, few-shot image recognition for fine-grained categories has not been well studied in depth. In this paper, we focus on the problem of few-shot fine-grained image recognition (FSFGIR).

Most existing few-shot image recognition approaches [Vinyals *et al.*, 2016; Snell *et al.*, 2017; Finn *et al.*, 2017; Munkhdalai *et al.*, 2018] mainly focus on general concepts to learn generic knowledge with global representations of images. But the global representation cannot capture subtle local differences of images effectively, which may hinder these few-shot image recognition approaches from dealing with FSFRIR well.

On the other hand, most of fine-grained image recognition systems [Zhang *et al.*, 2014; Xiao *et al.*, 2015] follow the pipeline of finding foreground object or object parts and extracting their discriminative features. More recent work merges the pipelines into an end-to-end learning framework [Zheng *et al.*, 2017; Fu *et al.*, 2017; Luo *et al.*, 2019]. However, these methods could be not employed directly for FSFGIR, since plenty of training images are unavailable in the few-shot learning. Inspired by the studies of the fine-grained image recognition, our idea is to learn discriminative parts with a small amount of training images in an end-to-end fashion, which has not been explored before.

In this work, we propose a multi-attention meta-learning (MattML) method, which leverages attention mechanisms of the base learner and task learner to capture discriminative parts of images. To be more specific, the base learner consists of a feature embedding network, two convolutional block attention modules CBAM [Woo *et al.*, 2018] and a classifier. The two CBAM can focus on diverse and informative parts by blending cross-channel and spatial information. In the task learner, a recurrent encoder is used to learn task representations with a recurrent decoder in the auto-encoding framework, and the task representations are employed by weight generator to attend the initialization of classifier of base learner. In this way, the attended classifier obtains a task-related sensitive initialization. For adaptation, we leverage the gradient-based meta-learning approach to adjust the parameters of two CBAM and the attended classifier, such that the updated base learner can be adaptive to focus on discriminative parts according to the current few-shot task.

Our main contributions are summarized as follows:

- To our best knowledge, we first combine attention mechanisms and meta learning for FSFGIR.
- We propose a MattML method, which uses attention mechanisms of the base learner and task learner to capture discriminative parts of images.
- We establish comprehensive benchmarks for FSFGIR, and experimental results demonstrate state-of-the-art performance under the 1-shot setting.

2 Related work

2.1 Fine-grained image recognition

In the early study of fine-grained image recognition, some works [Chai *et al.*, 2013; Xie *et al.*, 2013] are proposed with part-based annotations of object available at both training and inference phase. Benefiting from the the development of deep neural networks, the research of fine-grained image recognition is shifted from where part-based annotations of object are known [Zhang *et al.*, 2014] to where they are unknown [Xiao *et al.*, 2015]. When it comes to the unknown part-based annotations, there are two main research lines. The first line is to regularize feature learning by exploiting structural relationships between fine-grained labels such as intermediate concepts [Wang *et al.*, 2015; Xie *et al.*, 2015] or shared attributes [Zhou and Lin, 2016]. Another line of research first localizes discriminative parts and then extracts features from these parts in a multi-stage learning framework [Xiao *et al.*, 2015; Zhang *et al.*, 2016]. Recently, this line of research combines part localization and feature learning in an end-to-end framework [Fu *et al.*, 2017; Zheng *et al.*, 2017; Luo *et al.*, 2019]. In these end-to-end frameworks, the discriminative parts are captured by the attention mechanism, which has become a feasible approach to fine-grained image classification. As only image-level labels are available in our FSFGIR, we learn from those end-to-end approachs and further propose a multi-attention meta-learning method.

2.2 Few-shot image recognition

As an early attempt, Fei-Fei et al. propose a variational Bayesian framework for one-shot image classification [Fei-Fei *et al.*, 2006], and Lake et al. [Lake *et al.*, 2015] propose Hierarchical Bayesian Program Learning on the few-shot alphabet recognition tasks. Recently, there are two main lines of research to deal with the few-shot image recognition problem. The first line, named metric-based few-shot learning method, learns a transferable embedding network or function to transform images into the embedding space. And in this space the images can be recognized with a nearest neighbor [Vinyals *et al.*, 2016; Snell *et al.*, 2017] or a deep nonlinear metric [Sung *et al.*, 2018]. The second line uses meta-learning methods, which consists of two main components – a base learner (an initial model) and an adaptation approach (updating strategies). The base learner can be implemented with a standard network, and the adaptation approach may be implemented with parameterized networks or non-parametric strategies. The parameterized network is used to augment additive weights [Munkhdalai and Yu, 2017] or modify activation values [Munkhdalai *et al.*, 2018] on the base learner. The typical non-parametric strategy is fixed learning rules of gradient descent [Finn *et al.*, 2017], named gradient-based meta learning. It has become an important meta-learning approach and has been widely studied [Lee and Choi, 2018; Yoon *et al.*, 2018].

However, these existing few-shot image recognition approaches mainly explore general object images under the few-shot learning setting. They rarely consider intrinsic properties of images, such as fine-grained categories, which are distinguished by some subtle and local differences. To this end, Wei et al. [Wei *et al.*, 2019] introduce the FSFGIR task and propose piece-wise classifier mapping method with a bilinear network. To more effectively capture nuanced features, Huang et al. [Huang *et al.*, 2019] present low-rank pairwise alignment bilinear network. Different from these two works, we focus on the subtle and local differences of images in the FSFGIR task by using gradient-based meta-learning approach with multi-attention mechanisms.

3 Preliminaries

3.1 Problem formulation

In the few-shot learning scenario, each problem is defined on tasks $\mathcal{T} \sim p(\mathcal{T})$. Each task is defined as $\mathcal{T}_i = \{D_{\mathcal{T}_i,S}, D_{\mathcal{T}_i,T}\}$, where $D_{\mathcal{T}_i,S}$ is a support set (training samples) and $D_{\mathcal{T}_i,T}$ is a target set (test samples). For few-shot classification problem, the support set $D_{\mathcal{T}_i,S} = \{(x_{i,j},y_{i,j}) \mid j = 1, 2, ..., n_s\}$ and the target set $D_{\mathcal{T}_i,T}$ are sampled from the same distribution sharing the same label space in each task. Sampling from training, validation and test data, respectively, the training, validation and test tasks have the same forms but with disjoint label space. If the support set contains K labeled examples for each of C unique classes in the test task, it is called C-way K-shot classification problem.

3.2 Meta-learning paradigm

In practice, a meta-learning method usually learns a meta learner, which consists of two main components – a base learner (an initial model) $\mathcal{B}_\Theta$ and an adaptation approach (updating strategies) $\mathcal{A}_\phi$. Then the goal of meta learning is to learn an optimal meta learner across a variety of tasks to generalize to novel tasks. The training process of a meta-learning algorithm contains three alternative operations, which are: i) **Task sampling**: A mini-batch of tasks $\mathcal{T}_B$ is sampled from the task distribution $p(\mathcal{T})$. ii) **Task-specific adaptation**. Given a task $\mathcal{T}_i \in \mathcal{T}_B$, the initial model leverages a small set of training samples (i.e., $D_{\mathcal{T}_i,S}$) to obtain update information (e.g., loss, gradient), which is further utilized by the adaptation approach to obtain the updated model $\mathcal{B}_{\Theta^i}$. iii) **Meta training**. This process aims to minimize the expected empirical loss over the target set $D_{\mathcal{T}_i,T}$ using each task-specific updated parameters Θ^i across all sampled tasks. Concretely, this can be thought of learning over a collection of tasks,

$$\min_{\phi,\Theta} \sum_{\mathcal{T}_i \in \mathcal{T}_B} \mathcal{L}(D_{\mathcal{T}_i,T}; \Theta^i, \phi) \tag{1}$$

where $\mathcal{L}()$ is a meta loss function such as the cross entropy loss for classification problems. For the test of a meta-learning algorithm, provided with a new task with a small number of training samples, task-specific parameters are obtained by process ii) and can be used in the test process.

3.3 Gradient-based meta learning

Here we give an overview of the representative algorithm, model-agnostic meta learning (MAML) [Finn *et al.*, 2017]. Analogously, MAML also contains the above mentioned two components and has the similar training process. But the adaptation approach $\mathcal{A}_\phi$ is fixed learning rules of gradient descent (non-parametric strategies). In the **task-specific adaptation**, task-specific parameters Θ^i are obtained with one or a few gradient steps computed with loss $\mathcal{L}(D_{\mathcal{T}_i,S}; \Theta)$. For one-step gradient descent, it is computed as Eq. 2, where α is a fixed learning rate of adaptation, and $\nabla_\Theta \mathcal{L}(D_{\mathcal{T}_i,S}; \Theta)$ is the corresponding gradient with respect to Θ.

$$\Theta^i = \Theta - \alpha \nabla_\Theta \mathcal{L}(D_{\mathcal{T}_i,S}; \Theta) \tag{2}$$

In the **meta training**, parameters Θ are optimized on the summation of sampled tasks, which is formalized as Eq. 3. (one gradient step as exemplary)

$$\min_{\Theta} \sum_{\mathcal{T}_i \in \mathcal{T}_B} \mathcal{L}(D_{\mathcal{T}_i,T}; \Theta^i) \tag{3}$$

4 The proposed method

We firstly present the architecture of base learner and task learner. And then task-specific adaptation approach is introduced in gradient-based meta-learning paradigm. Finally, we provide the formulation of meta-training objectives.

4.1 Base learner

The architecture of the base learner is shown in Figure 2, which consists of a feature embedding network, two CBAM [Woo *et al.*, 2018] and a classifier. Next, we review CBAM.

CBAM. Each CBAM contains a channel attention module (CAM) and a spatial attention module (SAM), which are connected in tandem. Given a feature map $\mathrm{F} \in \mathbf{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ of an image, CAM produces a channel attention by exploiting inter-channel relationship of features. CAM firstly aggregates spatial information of F by using both average-pooling and max-pooling operations, which generate two channel descriptors $\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}} \in \mathbf{R}^{\mathrm{C}\times1\times1}$ and $\mathrm{F}^{\mathrm{c}}_{\mathrm{max}} \in \mathbf{R}^{\mathrm{C}\times1\times1}$, respectively. As each channel of a feature map is considered as a feature detector [Zeiler and Fergus, 2014], CAM uses $\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}}$ and $\mathrm{F}^{\mathrm{c}}_{\mathrm{max}}$ to generate 1D channel attention map $\mathrm{M}_c \in \mathbf{R}^{C\times1\times1}$, i.e.,

$$\mathrm{M}_c = \sigma(\mathrm{P}(\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}};\theta_c) + \mathrm{P}(\mathrm{F}^{\mathrm{c}}_{\mathrm{max}};\theta_c)) \tag{4}$$

where θ_c are parameters of 2 layer perceptron (P) and $\sigma()$ denotes the sigmoid function. Then refined feature map $\mathrm{F}' \in \mathbf{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ is $\mathrm{F}' = \mathrm{F} \odot \mathrm{M}_c$, where $\odot$ denotes element-wise multiplication and values of M_c are copied on spatial dimension. SAM generates a spatial attention by utilizing inter-spatial relationship of features. SAM aggregates channel information of F' by using both average-pooling and max-pooling operations, and then leverages pooled features $\mathrm{F}^{s}_{\mathrm{avg}} \in \mathbf{R}^{1\times\mathrm{H}\times\mathrm{W}}$ and $\mathrm{F}^{s}_{\mathrm{max}} \in \mathbf{R}^{1\times\mathrm{H}\times\mathrm{W}}$ to a 2D spatial attention map $\mathrm{M}_s \in \mathbf{R}^{1\times\mathrm{H}\times\mathrm{W}}$, i.e.,

$$\mathrm{M}_s = \sigma(\mathrm{Conv}(\mathcal{C}(\mathrm{F}^{\mathrm{s}}_{\mathrm{avg}}, \mathrm{F}^{\mathrm{s}}_{\mathrm{max}}); \theta_s)) \tag{5}$$

where θ_s are parameters of a layer convolution (Conv), and $\mathcal{C}(,)$ denotes concatenation. Then the final refined feature map $\mathrm{F}'' \in \mathbf{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ is $\mathrm{F}'' = \mathrm{F}' \odot \mathrm{M}_s$, where values of M_s are copied on channel dimension.

The original work [Woo *et al.*, 2018] uses only a CBAM with a residual operation, while we use two CBAM without residual operations. And we experimentally verify our structures are better. The feature embedding network is a convolutional neural network with parameters θ_f and the classifier is a fully connected layer with parameters θ_{cls}. Thus the parameters of base learner $\mathcal{B}_\Theta$ contain θ_f, θ_{cls}, $\theta_{c,k}$ and $\theta_{s,k}$ $(k = 1, 2)$, i.e., $\Theta = \{\theta_f, \theta_{cls}, \theta_{c,1}, \theta_{s,1}, \theta_{c,2}, \theta_{s,2}\}$.

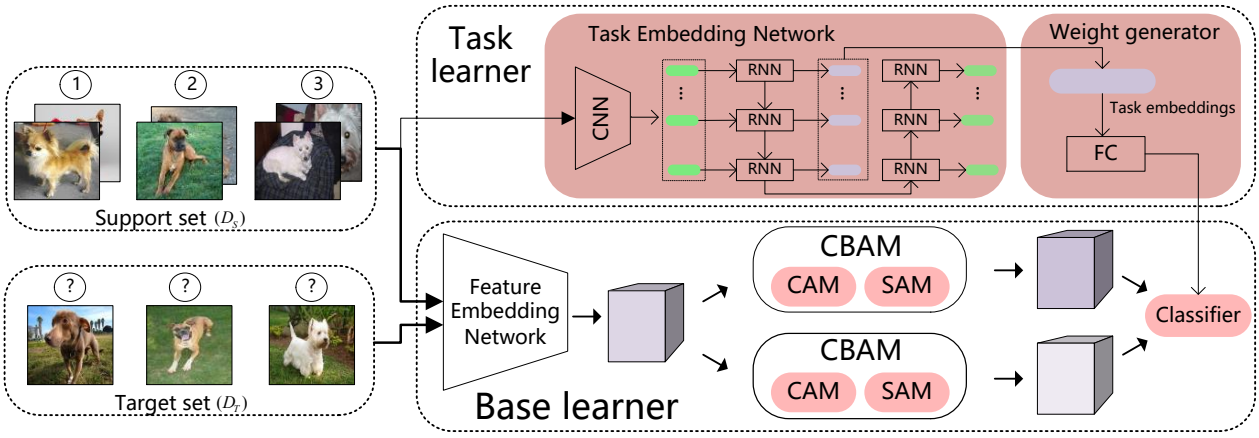

Figure 2: The architecture of the proposed MattML. Red ellipses in base learner: the modules need to be updated for adaptation.

4.2 Task learner

We propose a task embedding to attend the initialization of classifier of base learner. Specifically, we introduce a task embedding network to learn task representations, which are used by weight generator to attend the initialization of parameters of classifier. Next, we introduce the task embedding network and weight generator.

Task embedding network. Task embedding learning is important to meta-learning, especially for models trained on a sequence of tasks sampled from different and disjoint distributions. It is meaningful to describe and learn relations between tasks since meta-learning models can generalize well on new tasks if they have learned some related tasks. The main characteristics of task representation learning are reflected in the following two aspects: i) huge representational capacity. ii) permutation invariance of samples. These two aspects lead to task representation learning more challenging than representation learning of examples. Inspired by common practices in learning sentence embeddings [Conneau *et al.*, 2017], we tackle the challenge by aggregating representations of all examples, and use random order of samples to make permutation insensitive.

Given a task $\mathcal{T}_i$ with the support set $D_{\mathcal{T}_i,S} = \{(x_{i,j}, y_{i,j}) \mid j = 1, 2, ..., n_s\}$, each sample $x_{i,j}$ is fed into an embedding network to obtain features $\mathrm{E}(x_{i,j})$. In this paper, this embedding network shares parameters with the feature embedding network and is added two extra convolutional layers. Then features of all samples in the support set are sequentially fed into the recurrent auto-encoder. The reconstruction loss is:

$$\mathcal{L}_r(D_{\mathcal{T}_i,S}) = \frac{1}{n_s}\sum_{j}^{n_s} ||\mathrm{RNN}_{dec}(g_{i,j}) - \mathrm{E}(x_{i,j})||_2^2 \quad (6)$$

where $\forall j,\ g_{i,j} = \mathrm{RNN}_{enc}(\mathrm{E}(x_{i,j}), g_{i,j-1})$ and $\mathrm{RNN}_{dec}(g_{i,j})$ represents the learned encoding representation and the reconstruction of the j-th example, respectively. Here $\mathrm{RNN}_{enc}()$ and $\mathrm{RNN}_{dec}()$ stand for a recurrent encoder (e.g., LSTM, GRU) and a recurrent decoder, respectively. The task embedding is aggregated over representations of all examples, i.e.,

$$e_i = \frac{1}{n_s}\sum_{j}^{n_s} \mathrm{FC}(g_{i,j}) \quad (7)$$

where $\mathrm{FC}()$ represents a fully-connected layer.

Weight generator. The task embedding e_i is used to generate gates of classifier o_{cls}, i.e.,

$$o_{cls} = \sigma(\mathrm{FC}(e_i; \theta_t^{cls})) \quad (8)$$

where θ_t^{cls} are parameters to predict o_{cls}. The initial weights of classifier are $\theta_{cls} := o_{cls} \odot \theta_{cls}$.

4.3 Task-specific adaptation

For a specific task $\mathcal{T}_i$, the task-specific parameters Θ^i are obtained with one or a few gradient steps computed with the loss $\mathcal{L}(D_{\mathcal{T}_i,S}; \Theta)$. But the θ_f are fixed. For one-step gradient descent, this process is computed as follows:

$$\begin{cases} \theta_{cls}^i := \theta_{cls} - \alpha_{cls}\nabla_{\theta_{cls}}\mathcal{L}(D_{\mathcal{T}_i,S}; \Theta) \\ \theta_{s|c,k}^i := \theta_{s|c,k} - \alpha_k \nabla_{\theta_{s|c,k}}\mathcal{L}(D_{\mathcal{T}_i,S}; \Theta) \end{cases} \quad (9)$$

Where α_{cls}, α_k are learning rates of adaptation. Since fine-grained image has subtle differences in each category, the span of updating CBAM should be smaller compared with classifier. Thus α_k should be set a smaller value. Since the settings of α_{cls} and α_k may be not optimal, we set these learning rates to be learnable. The task-specific parameters are $\Theta^i = \{\theta_f, \theta_{cls}^i, \theta_{c,1}^i, \theta_{s,1}^i, \theta_{c,2}^i, \theta_{s,2}^i\}$.

4.4 Meta-training objectives

Recalling the objectives for a meta-learning algorithm, we reach the optimization problem:

$$\min_{\Phi} \sum_{\mathcal{T}_i} \mathcal{L}(D_{\mathcal{T}_i,T}; \Theta^i) + \xi\mathcal{L}_r(D_{\mathcal{T}_i,S}) \quad (10)$$

where Φ represents all learnable parameters including the parameters of $\mathcal{B}_\Theta$, parameters of task learner, learning rates of adaptation (i.e., α_{cls}, α_k), and ξ is used to balance the importance of two items. The ξ is set 0.01 in the experiment. The $\mathcal{L}_r()$ measures the reconstruction error as defined in Eq. 6, and $\mathcal{L}()$ is defined in Eq. 3.

	#(categories)			#(images)
	Train	Validation	Test	each category
FS-Birds	130	20	50	60
FS-Dogs	70	20	30	171.5*
FS-Cars	130	17	49	82.6*
FS-Aircrafts	60	15	25	100

Table 1: The splits of categories and the number of categories/images in each FSFGIR dataset. #(x): the number of x; *: the average number.

Model	1-shot	5-shot
Matching Net	54.41±0.47	70.04±0.35
Prototypical Net	57.62±0.49	71.43±0.38
Relation Net	63.94±0.51	76.22±0.34
MAML	60.24±0.34	75.24±0.24
adaCNN	58.60±0.48	72.82±0.38
MattML (our)	**75.69±0.54**	**86.23±0.31**

Table 2: 5-way 1-shot and 5-shot MA (%) ± 95 Cls (%) on FS-Aircrafts. The highest accuracy is highlighted in bold face.

5 Experiment

5.1 Datasets

Our experiments are conducted on four fine-grained benchmark datasets (i.e., CUB Birds [Wah *et al.*, 2011], Stanford Dogs [Khosla *et al.*, 2011], Stanford Cars [Krause *et al.*, 2013], FGVC Aircraft [Maji *et al.*, 2013]). Splitting categories of these datasets forms corresponding FSFGIR datasets. For CUB Birds, Stanford Dogs and Car, we follow the splits of training, validation and test categories in [Li *et al.*, 2019b] (i.e., FS-Birds, FS-Dogs, FS-Cars). For FGVC Aircraft, we randomly split categories to form corresponding FSFGIR dataset (i.e., FS-Aircrafts). The detailed splits of training, validation and test categories and the number of categories/images in each dataset are presented in Table 1.

5.2 Comparison Methods

Meta-learning methods. As our method belongs to the meta-learning branch, we mainly compare meta-learning models, including MAML [Finn *et al.*, 2017], adaCNN [Munkhdalai *et al.*, 2018].

Metric-learning based methods. Besides meta-learning models, three well-known metric-learning based models (i.e., Matching Net [Vinyals *et al.*, 2016], Prototypical Net [Snell *et al.*, 2017], and Relation Net [Sung *et al.*, 2018]) are also picked for reference. We reset the Prototypical Nets with the same 5-way training setting instead of 20-way training setting in the original work for a fair comparison. In addition, three recent state-of-the-art models (i.e., CovaMNet [Li *et al.*, 2019b], DN4 [Li *et al.*, 2019a], LRPABN [Huang *et al.*, 2019]) are also compared. In these methods, Matching Net and Prototypical Net describe images with global representations, and Relation Net, CovaMNet, DN4 and LRPABN learn local descriptors of images.

5.3 Implementation details

For a fair comparison with state-of-the-art methods, we use a widely adopted CNN-4 [Vinyals *et al.*, 2016; Snell *et al.*, 2017] as a feature embedding network, which consists of four convolutional layers. 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 this network is 84×84, and the final feature of our method for classifying is 3200-dimensional. We apply standard data augmentation, which includes random crop, left-right flip, and color jitter at the meta-training stage in all implemented experiments.

The results are reported with mean accuracy (MA) + 95% confidence intervals (CIs) over sampled 2000 tasks. And each task contains $C = 5$ classes, each of which has $K \in \{1, 5\}$ examples in support set and 15 examples in target set. During training, all of models are trained from scratch in an end-to-end manner across on tasks. The batch size of task is set to 4, and each task has the same settings with the above test. We use Adam optimizer [Kingma and Ba, 2015] with initial learning rate 0.001. The total iterations are 80,000 and the learning rate is changed to 1/2 after each 20,000 iterations.

5.4 Experimental results

Table 2 presents 5-way mean accuracy of different methods on FS-Aircrafts. We implement five compared methods (i.e., Matching Net, Prototypical Net, Relation Net, MAML, adaCNN) with the corresponding public code. It can be observed that our method shows absolute advantages compared with the five methods under both 5-way 1-shot and 5-shot settings. Especially for the 5-way 1-shot task, the proposed MattML gains 21.28%, 18.07%, 11.75%, 15.45% and 17.09% improvements over Matching Net, Prototypical Net, Relation Net, MAML and adaCNN, respectively. For the 5-way 5-shot task, the MattML achieves 16.19%, 14.80%, 10.01%, 10.99% and 13.41% gains over Matching Net, Prototypical Net, Relation Net, MAML and adaCNN, respectively. The proposed MattML defeats the five methods by average 16.73%, 13.08% gains under 1-shot and 5-shot settings, respectively, which demonstrates the superiority of attention based meta learning method. And this also reflects few-shot learning methods of global representations do not deal with FSFGIR well.

The 5-way mean accuracy of different methods on FS-Dogs, FS-Cars and FS-Birds are shown in Table 3. We also implement three compared methods (i.e., Relation Net, MAML, adaCNN) with the corresponding public code on the three datasets. Similarly, the proposed MattML obtains significant improvements compared with Matching Net, Prototypical Net, Relation Net, MAML and adaCNN under both 1-shot and 5-shot settings. Compared with methods of learning local descriptors (i.e., CovaMNet, DN4 and LRPABN), MattML shows obvious superiority. Especially for the 1-shot task, MattML achieves 5.74%, 4.60% and 2.66% gains over the best one of them on FS-Dogs, FS-Cars and FS-Birds, respectively. For the 5-shot task, the proposed MattML also obtains state-of-the-art performance on FS-Dogs, and a comparable performance on FS-Birds. This indicates that it is a feasible way for FSFGIR to use attended local descriptors rather than all local descriptors.

Method	FS-Dogs		FS-Cars		FS-Birds	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Matching Net †	35.80±0.99	47.50±1.03	34.80±0.98	44.70±1.03	45.30±1.03	59.50±1.01
Prototypical Net †	37.59±1.00	48.19±1.03	40.90±1.01	52.93±1.03	37.36±1.00	45.28±1.03
Relation Net	43.29±0.46	55.15±0.39	47.79±0.49	60.60±0.41	58.99±0.52	71.20±0.40
MAML	44.84±0.31	58.61±0.30	47.25±0.30	61.11±0.29	58.13±0.36	71.51±0.30
adaCNN	42.16±0.43	54.12±0.39	41.88±0.40	49.87±0.37	56.76±0.50	61.05±0.44
CovaMNet	49.10±0.76	63.04±0.65	56.65±0.86	71.33±0.62	52.42±0.76	63.76±0.64
DN4	45.73±0.76	66.33±0.66	61.51±0.85	**89.60±0.44**	53.15±0.84	**81.90±0.60**
LRPABN	45.72±0.75	60.94±0.66	60.28±0.76	73.29±0.58	63.63±0.77	76.06±0.58
MattML (our)	**54.84±0.53**	**71.34±0.38**	**66.11±0.54**	82.80±0.28	**66.29±0.56**	80.34±0.30

Table 3: 5-way 1-shot and 5-shot MA (%) ± 95 Cls (%) on FS-Dogs, FS-Cars and FS-Birds.
†: reported by [Li *et al.*, 2019b; Li *et al.*, 2019a]

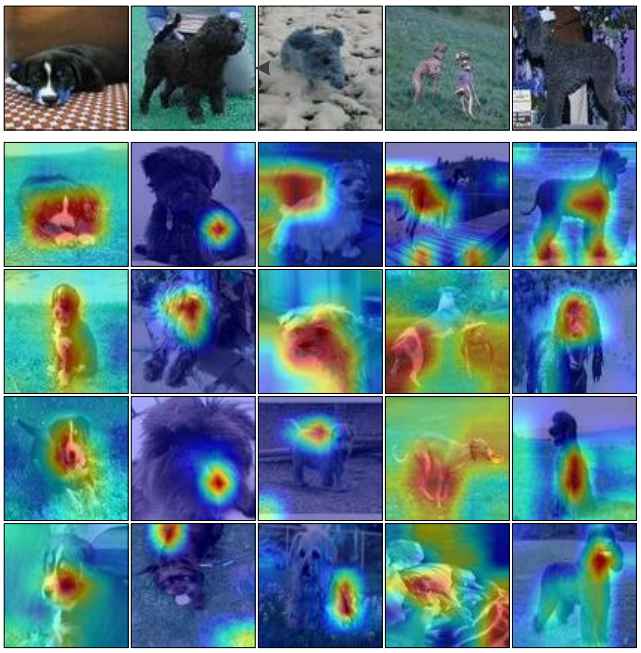

Figure 3: An example of 5-way 1-shot task on FS-Dogs dataset. The images of the same column come from the same category.

5.5 Further Analysis

Visualization. Figure 3 illustrates an example of 5-way 1-shot task on FS-Dogs. In this task, 5 training samples (support set) are in the top, and test samples (target set) are in the bottom. Test samples are presented with Grad-CAM visualizations [Selvaraju *et al.*, 2017]. In these visualizations, highlighted image regions are relevant to parts of dog (e.g., head, tail, leg, body). Specifically, salient part in the first column is head, and salient part in the fourth column is body. Interestingly, these highlighted parts can be used to distinguish their categories. This can illuminate that the proposed method can be adaptive to focus on discriminative parts with a few of training images.

Ablation Studies. Table 4 shows ablation studies of the 5-way 1-shot setting on FS-Dogs. The baseline method does not employ any attention modules (CBAM and task learner). It can be observed that: i) Only using one CBAM, we gains 3.83% improvements, which illustrates the attention mechanism (CBAM) is very useful for FSFGIR. But using residual CBAM achieves slight gains. The possible reason is the importance of CBAM has been weakened with the residual operation. ii) Compared with one CBAM, the method of using two CBAM obtains further improvements, which indicates that the proposed technique can capture complementary discriminative parts. iii) On this basis, our method achieves about 0.7% improvements with a task learner, which explains that the task learner learns a better initialization of classifier according to the current task.

Model	5-way 1-shot
baseline	48.84±0.48
baseline+1 CBAM	52.67±0.46
baseline+1 CBAM (residual)	49.31±0.50
baseline+2 CBAM	54.15±0.48
baseline+2 CBAM+Task learner (our)	**54.84±0.53**

Table 4: Ablation analysis on FS-Dogs.

6 Conclusion

In this paper, we comprehensively investigate the problem of FSFGIR. By analyzing the characteristics of fine-grained images, we propose a MattML method, which uses attention mechanisms of the base learner and task learner to capture discriminative parts of images according to the current task. Experimental results of FSFGIR on four benchmarks (i.e, FS-Birds, FS-Dogs, FS-Cars, FS-Aircrafts) show the effectiveness and the superiority of the proposed method.

Acknowledgements

This work was supported by National Key Research and Development Project of New Generation Artificial Intelligence of China, under Grant 2018AAA0102500, in part by the National Natural Science Foundation of China under Grant 61532018, in part by Beijing Natural Science Foundation under Grant L182054 and Z190020, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals, in part by the Lenovo Outstanding Young Scientists Program.

References

[Chai *et al.*, 2013] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, pages 321–328, 2013. 2.1

[Conneau *et al.*, 2017] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 670–680, 2017. 4.2

[Fei-Fei *et al.*, 2006] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 28(4):594–611, 2006. 2.2

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1, 2.2, 3.3, 5.2

[Fu *et al.*, 2017] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, pages 4438–4446, 2017. 1, 1, 2.1

[Huang *et al.*, 2019] Huaxi Huang, Junjie Zhang, Jian Zhang, Jingsong Xu, and Qiang Wu. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *arXiv preprint arXiv:1908.01313*, 2019. 2.2, 5.2

[Khosla *et al.*, 2011] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshops*, 2011. 1, 5.1

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5.3

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561, 2013. 5.1

[Lake *et al.*, 2015] B. M. Lake, R Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 2.2

[Lee and Choi, 2018] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *ICML*, pages 2927–2936, 2018. 2.2

[Li *et al.*, 2019a] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*, 2019. 5.2, 3

[Li *et al.*, 2019b] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. Distribution consistency based covariance metric networks for few-shot learning. In *AAAI*, 2019. 5.1, 5.2, 3

[Luo *et al.*, 2019] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, pages 8242–8251, 2019. 1, 2.1

[Maji *et al.*, 2013] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *Technical report*, 2013. 5.1

[Munkhdalai and Yu, 2017] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017. 2.2

[Munkhdalai *et al.*, 2018] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, 2018. 1, 2.2, 5.2

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 5.5

[Smith and Slone, 2017] Linda B Smith and Lauren K Slone. A developmental approach to machine learning? *Frontiers in psychology*, 8:2124, 2017. 1

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4080–4090, 2017. 1, 2.2, 5.2, 5.3

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 1, 2.2, 5.2

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 1, 2.2, 5.2, 5.3

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 5.1

[Wang *et al.*, 2015] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV*, pages 2399–2406, 2015. 2.1

[Wei *et al.*, 2019] Xiu-Shen Wei, Peng Wang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE TIP*, 28(12):6116–6125, 2019. 2.2

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 1, 4.1, 4.1

[Xiao *et al.*, 2015] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, pages 842–850, 2015. 1, 1, 2.1

[Xie *et al.*, 2013] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *ICCV*, pages 1641–1648, 2013. 2.1

[Xie *et al.*, 2015] Saining Xie, Tianbao Yang, Xiaoyu Wang, and Yuanqing Lin. Hyper-class augmented and regularized deep learning for fine-grained image classification. In *CVPR*, pages 2645–2654, 2015. 2.1

[Yoon *et al.*, 2018] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *NIPS*, pages 7342–7352, 2018. 2.2

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014. 4.1

[Zhang *et al.*, 2014] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849. Springer, 2014. 1, 1, 2.1

[Zhang *et al.*, 2016] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, pages 1134–1142, 2016. 2.1

[Zheng *et al.*, 2017] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, pages 5209–5217, 2017. 1, 2.1

[Zhou and Lin, 2016] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. In *CVPR*, pages 1124–1133, 2016. 2.1