MULTI-LEVEL METRIC LEARNING FOR FEW-SHOT IMAGE RECOGNITION

Haoxing Chen, Huaxiong Li, Yaohui Li, Chunlin Chen

Nanjing University

{haoxingchen, yaohuili}@smail.nju.edu.cn, {huaxiongli, clchen}@nju.edu.cn

ABSTRACT

Few-shot learning is devoted to training a model on few samples. Most of these approaches learn a model based on a pixel-level or global-level feature representation. However, using global features may lose local information, and using pixel-level features may lose the contextual semantics of the image. Moreover, such works can only measure the relations between them on a single level, which is not comprehensive and effective. And if query images can simultaneously be well classified via three distinct level similarity metrics, the query images within a class can be more tightly distributed in a smaller feature space, generating more discriminative feature maps. Motivated by this, we propose a novel Part-level Embedding Adaptation with Graph (PEAG) method to generate task-specific features. Moreover, a Multi-level Metric Learning (MML) method is proposed, which not only calculates the pixel-level similarity but also considers the similarity of part-level features and global-level features. Extensive experiments on popular few-shot image recognition datasets prove the effectiveness of our method compared with the state-of-the-art methods.

Index Terms— Multi-level, metric-learning, few-shot, image recognition

1. INTRODUCTION

Humans can learn novel concepts and objects with just a few samples. Recently, many methods were proposed to learn new concepts with limited labeled data, such as semi-supervised learning [1], zero-shot learning [2, 3], and few-shot learning [4, 5, 6, 7, 8, 9]. Facing with the problem of data scarcity, these three paradigms propose solutions from different perspectives. Semi-supervised learning aims to train a model with few labeled data and a large amount of unlabeled data, and zero-shot learning devoted to identifying unseen categories with no labeled data, while few-shot learning focuses on learning new concepts with few labeled data. We propose a novel few-shot learning method to address the problem of data scarcity in this paper.

The few-shot learning methods can be roughly classified into two categories: meta-learning based methods [4, 10] and metric-learning based methods [5, 6, 7, 8, 9]. Metric-based few-shot learning methods have achieved remarkable success due to their fewer parameters and effectiveness. In this work, we focus on this branch.

The basic idea of the metric-learning based few-shot learning method is to learn a good metric to calculate the similarity between query images and the support set. Therefore, how to learn good feature embedding representation and similarity metric are the key problem of metric-learning based few-shot learning method. For feature embedding representation, Prototypical Networks [5] and Relation Networks [6] adopt image-level feature representations. However, due to the scarcity of data, it is not sufficient to measure the relation at the image-level [5, 6]. Recently, CovaMNet [11], DN4 [7] and MATANet [9] introduce local representations (LRs) into few-shot learning and utilize these LRs to represent the image features, which can achieve better recognition results.

For similarity metrics, these existing methods calculate similarities by different metrics. For example, Relation Networks [6] proposes a network to learn the most suitable image-level similarity metric functions. DN4 [9] proposes a cosine-based image-to-class metric to measure the similarity on pixel-level.

However, global-level features lose local semantic information and pixel-level features lose contextual semantics, thus all methods mentioned above are not effective for few-shot learning. Moreover, these methods only calculate similarities on a single level, i.e., pixel-level or image-level, which is not effective enough. Intuitively, under the few-shot learning setting, the features obtained by adopting a single similarity measure are not comprehensive, and the single similarity measure may lead to a certain similarity deviation, thus reducing the generalization ability of the model. It is necessary to adopt multi-level similarity metric, generating more discriminative features rather than using a single measure.

To this end, we propose part-level embedding adaptation with graph (PEAG) method and multi-level metric learning method (MML). In PEAG, we divide each image into patches and get part-level features. Then, we utilize Graph Convolutional Network (GCN) to generate task-spcific features. Finally, we adpot a nearest neighbor matching module to get part-level similarity. In MML, in addition to component-level measures, we also use global-level measures and pixel-level measures to provide complementary information for a more

compact measurement space. In MML, we also use global-level and pixel-level metrics to provide complementary information, and images within a class can be more tightly distributed in a smaller feature space.

The main contributions are summarized as follows:

- We propose a novel part-level embedding adatation with graph method, which can generate task-specific part-level features and capture the part-level semantic similarity between query images and support images.
- We propose a novel multi-level metric learning method by computing the semantic similarities on pixel-level, part-level, and global-level simultaneously, aiming to find more comprehensive semantic similarities.
- We conduct sufficient experiments on popular benchmark datasets to verify the advancement of our model and the performance of our model achieves the state-of-the-art.

2. RELATED WORKS

In this section, we focus on related works on metric-learning based few-shot learning model.

2.1. Learning feature embedding representation

Koch et al. [12] used a Siamese Neural Network to tackle the one-shot learning problem, in which the feature extractor is of VGG styled structure and L_1 distance is used to measure the similarity between query images and support images. Snell et al. [5] proposed Prototypical Networks, in which the Euclidean distance is used to compute the distance between class-specific prototypes. Li et al. [7] proposed an image-to-class mechanism to find the relation at pixel-level, in which the image features are represented as a local descriptor collection.

2.2. Learning similarity metric

Sung et al. [6] replaced the existing metric with the Relation Network, which measures the similarity between each query instance and support classes. Li et al. [7] proposed a Deep Nearest Neighbor Neural Network (DN4) to learn an image-to-class metric by measuring the cosine similarity between the deep local descriptors of a query instance and its neighbors from each support class. Li et al. [11]explored the distribution consistency-based metric by introducing local covariance representation and deep covariance metric. Unlike these methods, the proposed MML measures the similarity at three different feature levels, i.e., pixel-level, part-level, and distribution-level.

3. PROBLEM DEFINITION AND FORMULATION

Standard few-shot image recognition problems are often formalized as *N*-way *M*-shot classification problem, in which models are given *M* seen images from each of *N* classes, and required to correctly classify unseen images. Different from traditional image recognition tasks, few-shot learning aims to classify novel classes after training. This requires that samples used for training, validation, and testing should come from disjoint label space. To be more specific, given a dataset of visual concepts $\mathcal{C}$, we devide it into three parts: $\mathcal{C}_{train}$, $\mathcal{C}_{val}$ and $\mathcal{C}_{test}$, and their label space satisfy $\mathcal{L}_{train} \cap \mathcal{L}_{val} \cap \mathcal{L}_{test} = \emptyset$.

To obtain a trained model, we train our model in an episodic way. That is, in each episode, a new task is randomly sampled from the training set $\mathcal{C}_{train}$ to train the current model. Each task consists of two subsets, including support set $\mathcal{A}_{\mathcal{S}}$ and query set $\mathcal{A}_{\mathcal{Q}}$. The $\mathcal{A}_{\mathcal{S}}$ contains $\mathcal{N}$ previously unseen classes, with $\mathcal{M}$ samples for each class. We focus on training our model to correctly determine which category each image in the $\mathcal{A}_{\mathcal{Q}}$ belongs to. Similarly, we randomly sample tasks from $\mathcal{C}_{val}$ and $\mathcal{C}_{test}$ for meta-validation and meta-testing scenarios.

4. MULTI-LEVEL METRIC LEARNING

As shown in Figure 2, our MML is mainly composed of two modules: a feature extractor $\mathcal{F}_\theta$, a multi-level metric-learning module. All images are first fed into the $\mathcal{F}_\theta$ to get feature embeddings. Then, the multi-level metric-learning module calculates similarities on part-level, pixel-level, and distribution-level simultaneously. Finally, we fuse these three similarities together. All the modules can be trained jointly in an end-to-end manner.

4.1. Part-level Metric

We divide each image into $H \times W$ patches evenly, and input each patch into $\mathcal{F}_\theta$ separately to generate part-level descriptors. In order to further enhance the representation ability of features, we adopt a pyramid structure. Thus, under *N-way M-shot* few-shot learning setting, given a query image q and support class $\mathcal{S}_n, n = \{1, ..., N\}$, through $\mathcal{F}_\theta$ and global-average pooling (GAP) layer, we can get the part-level descriptors: $\mathcal{X}_q^{\text{part}} \in \mathbb{R}^{C\times K}$ and $\mathcal{X}_{\mathcal{S}_n}^{\text{part}} \in \mathbb{R}^{C\times MK}$. Specifically, we adopt two image patch division strategies of size 2×2 and 3×3 to obtain 13 part-level descriptors.

To generate task-specific support features, we propose a novel part-level embedding adaptation with graph (PEAG) method. Specifically, we concatenate all support features and get $\mathcal{X}_{\mathcal{S}}^{\text{part}} \in \mathbb{R}^{C\times NMK}$. Then, we construct the degree matrix $A \in \mathbb{R}^{NMK\times NMK}$ to represent the similarity between patches in the support set. If two patches p_i and p_j have the same semantics, then we set the corresponding element A_{ij} to

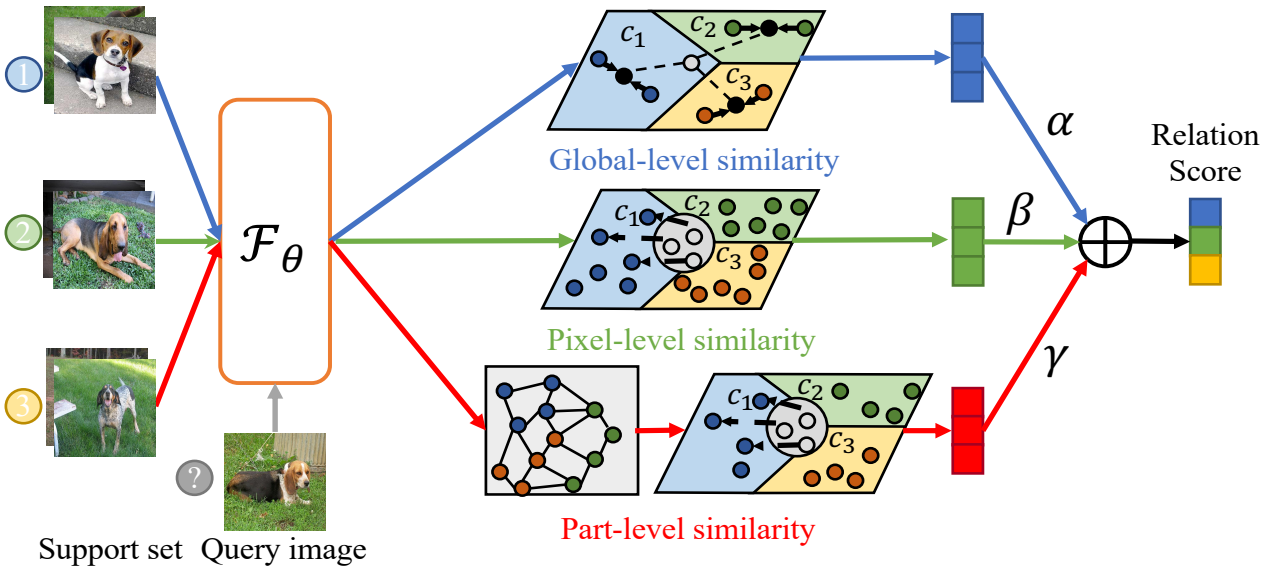

Fig. 1. The framework of MML under the 3-way 2-shot image classification setting. (Best view in color.)

1, otherwise to 0. Based on A, we build the adjacency matrix S:

$$S = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \tag{1}$$

where $I \in \mathbb{R}^{NMK \times NMK}$ is the identity matrix and D is the diagonal matrix ($D_{ii} = \sum_j A_{ij} + 1$). Let $\Psi^0 = \mathcal{X}_{\mathcal{S}}^{\text{part}}$, the relationship between patches could be propagated based on S:

$$\Psi^{t+1} = \text{ReLU}(S\Psi^t W), t = \{0, ..., T-1\} \tag{2}$$

where W is a learned feature transformation matrix. After propagate the embedding set T times, we can get the final propagated embedding set $\Psi^T = \mathcal{X}_{\mathcal{S}'}^{\text{part}}$.

Then, for each class $\mathcal{X}_{\mathcal{S}'_n}^{\text{part}}$, we calculate the correlation matrix $R^{\text{part}} \in \mathbb{R}^{K \times MK}$ between the query image and the support class n on part-level:

$$\mathcal{R}^{\text{part}} = \frac{(\mathcal{X}_q^{\text{part}})^\top \mathcal{X}_{\mathcal{S}'_n}^{\text{part}}}{\left\|\mathcal{X}_q^{\text{part}}\right\| \cdot \left\|\mathcal{X}_{\mathcal{S}'_n}^{\text{part}}\right\|} \tag{3}$$

$\mathcal{R}_{i,j}^{\text{part}}$ is (i, j) element of $\mathcal{R}^{\text{part}}$ reflecting the distance between the i-th descriptors of the query image and the j-th descriptor of support clss n. Each row in $\mathcal{R}^{\text{part}}$ represents the semantic relation of each descriptor in the query image to all descriptors of all images in the support class. For each patch of the query image q, we find its most similar descriptor. Then, we sum K selected part-level descriptors as the part-level similarity between the query image and the support class n:

$$\mathcal{D}_{\text{part}}(q, \mathcal{S}_n) = \sum_{i=1}^{K} \text{Top1}(\mathcal{R}_i^{\text{part}}) \tag{4}$$

where $\text{Top}(\cdot)$ means selecting the largest elements in each row of the R^{part}.

4.2. Pixel-level Metric

Following [11, 7, 9], given a query image q and a certain support class $\mathcal{S}_n$, through feature extractor $\mathcal{F}_\theta$, we can get the feature representation $\mathcal{F}_\theta(q) \in \mathbb{R}^{C \times H \times W}$ and $\mathcal{F}_\theta(\mathcal{S}_n) \in \mathbb{R}^{M \times C \times H \times W}$, respectively. The $\mathcal{F}_\theta(q)$ can be regard as a set of $H \times W$ C-dimensional LRs:

$$\mathcal{L}_q^{\text{pixel}} = [u_1^{\text{pixel}}, ..., u_{HW}^{\text{pixel}}] \in \mathbb{R}^{C \times HW} \tag{5}$$

Also, the $\mathcal{F}_\theta(\mathcal{S}_n)$ can be regards as

$$\mathcal{L}_{\mathcal{S}_n}^{\text{pixel}} = [v_1^{\text{pixel}}, ..., v_{MHW}^{\text{pixel}}] \in \mathbb{R}^{C \times MHW} \tag{6}$$

Then, we calculate the correlation matrix $R^{\text{pixel}} \in \mathbb{R}^{HW \times MHW}$ between the query image and the support class on pixel-level and select the largest element in each row of the correlation matrix:

$$\mathcal{R}^{\text{pixel}} = \frac{(u_i^{\text{pixel}})^\top v_j^{\text{pixel}}}{\left\|u_i^{\text{pixel}}\right\| \cdot \left\|v_j^{\text{pixel}}\right\|} \tag{7}$$

$$\mathcal{D}_{\text{pixel}}(q, \mathcal{S}_n) = \sum_{i=1}^{HW} \text{Top1}(\mathcal{R}_i^{\text{pixel}}) \tag{8}$$

4.3. Global-level Metric

We adpot Prototypical Networks [5] as our global-level similarity metric. Prototypical Networks computes the empirical

Dataset	N_{all}	N_{train}	N_{val}	N_{test}
*mini*ImageNet	100	64	16	20
*tiered*ImageNet	608	351	97	160
CIFAR-100	100	64	16	20
FC100	100	60	20	20

Table 1. The splits of evaluation datasets. N_{all} is the number of all classes. N_{train}, N_{val} and N_{test} indicate the number of classes in training set, validation set and test set.

mean of global convulution embeddings as the prototype representation of each category n:

$$c_n = \frac{1}{M}\sum_{M}^{i=1} \mathrm{GAP}(\mathcal{F}_\theta(\mathcal{S}_n^i)) \tag{9}$$

where $p_n \in \mathbb{R}^K$. Similarly, given a query image Q, we can get its global convulution embeddings $\mathcal{X}_Q^{\mathrm{global}} \in \mathbb{R}^K$. Then, Prototypical Networks utilized Euclidean distance as the distance metric and assigns a probability over class n:

$$\mathcal{D}_{\mathrm{global}}(q, \mathcal{S}_n) = -d(\mathcal{X}_Q^{\mathrm{global}}, c_n) \tag{10}$$

4.4. Fusion Layer

Since three different level similarities have been calculated, we need to design a fusion module to integrate them. Specifically, the final similarity and probability over any class n can be obtained by the following equation:

$$P_{part}(y=n|q) = \frac{\mathcal{D}_{part}(q, \mathcal{S}_n)}{\sum_{i=1}^{N} \mathcal{D}_{part}(q, \mathcal{S}_n)} \tag{11}$$

$$P_{pixel}(y=n|q) = \frac{\mathcal{D}_{pixel}(q, \mathcal{S}_n)}{\sum_{i=1}^{N} \mathcal{D}_{pixel}(q, \mathcal{S}_n)} \tag{12}$$

$$P_{global}(y=n|q) = \frac{\mathcal{D}_{global}(q, \mathcal{S}_n)}{\sum_{i=1}^{N} \mathcal{D}_{global}(q, \mathcal{S}_n)} \tag{13}$$

$$\begin{aligned} P(y=n|q) =& \alpha P_{part}(y=n|q) + \beta P_{global}(y=n|q) \\ &+ \gamma P_{pixel}(y=n|q) \end{aligned} \tag{14}$$

where y is the label of q, α, β and γ are superparameters. If $y = n'$, then we can define the loss function as follows:

$$\begin{aligned} \mathcal{L} =& -\alpha log(p_{part}(y=n'|q)) - \beta log(p_{pixel}(y=n'|q)) \\ &- \gamma log(p_{global}(y=n'|q)) \end{aligned} \tag{15}$$

5. EXPERIMENTS

In this section, we perform extensive experiments to verify the advance and effectiveness of MML.

α	β	γ	5-Way Accuracy(%) 1-shot	5-shot
1	0	0	**67.29±0.23**	78.49±0.21
0	1	0	64.12±0.23	78.55±0.21
0	0	1	61.86±0.24	79.03±0.21
1	1	0	66.85±0.23	80.52±0.20
1	0	1	61.65±0.24	78.87±0.21
0	1	1	61.95±0.24	78.85±0.21
1	1	1	64.77±0.23	79.93±0.20
1	0.5	0.5	66.72±0.23	**81.01±0.20**
1	0.1	0.1	**67.58±0.23**	**81.41±0.20**

Table 2. Ablation study on *mini*ImageNet. (Top two performances are in bold font.)

5.1. Datasets

To verify the advance and effectiveness of our proposed MML, we performed experiments on four benchmark datasets.

ImageNet derivatives: Both *mini*ImageNet [13] dataset and *tiered*ImageNet [14] dataset are subsets of ImageNet [15]. The *mini*ImageNet dataset consists 100 classes, each of which contains 600 samples, and the *tiered*ImageNet contains 608 classes.

CIFAR derivatives: Both CIFAR-FS [16] dataset and FC100 [17] dataset are subsets of CIFAR-100. Both of them consist 100 classes.

The partition of all data sets is shown in Table 1. All images are resized to 84×84.

5.2. Implementation Details

In order to make a fair comparison with other works, we adopt the *ResNet-12* network [18] as our feature extrator $\mathcal{F}_\theta$.

ResNet-12 has four residual blocks, each residual block has 3 convolutional layers with 3×3 kernel, and a 2×2 max-pooling layer is added in the first residual block.

We conduct our experiments on a series of *N*-way *M*-shot tasks, i.e., 5-way 1-shot and 5-way 5-shot. Following [19], we first pre-trained $\mathcal{F}_\theta$ with an MLP consisting of a single hidden layer. Then we meta-train the whole model by momentum SGD for 40 epochs. In each epoch, we randomly sampled 200 tasks. Our batch size is set to 4, the initial learning rate is 5×10^{-4}, and multiplied by 0.5 every 10 epochs. During the test stage, we report the average accuracy as well as the corresponding 95% confidence interval over these 10,000 tasks.

5.3. Ablation Study

To explore the effect of the multi-level metric learning module, we prune any of three similarity branches in the multi-

Model	Backbone	*mini*ImageNet		*tiered*ImageNet	
		1-shot	5-shot	1-shot	5-shot
Prototypical Networks [5]	Conv-64F	49.42±0.78	68.20±0.66	53.31±0.89	72.69±0.74
Relation Networks [6]	Conv-64F	50.44 ±0.82	65.32±0.77	54.48±0.93	71.32±0.78
DN4 [7]	Conv-64F	51.24 ±0.74	71.02±0.64	53.37±0.86	74.45±0.70
Prototypical Networks [5]	ResNet-12	62.59±0.85	78.60±0.16	68.37±0.23	83.43±0.16
TADAM [17]	ResNet-12	58.50±0.30	76.70±0.30	-	-
MeaOptNet [18]	ResNet-12	62.64±0.61	78.63±0.46	65.99±0.72	81.56±0.53
DSN-MR [8]	ResNet-12	64.60±0.72	79.51±0.50	67.39±0.82	82.85±0.56
FEAT [19]	ResNet12	66.78±0.20	**82.05±0.14**	67.39±0.82	82.85±0.56
GLoFA [20]	ResNet12	66.12±0.42	81.37±0.33	**69.75±0.33**	83.58±0.42
Fine-tuning [4]	WRN-28-10	57.73±0.62	78.17±0.49	66.58±0.70	**85.55±0.48**
AWGIM [21]	WRN-28-10	63.12±0.08	78.40±0.11	67.69±0.11	82.82±0.13
PEAG	ResNet-12	**67.29±0.23**	78.49±0.21	68.89±0.25	82.08±0.21
MML	ResNet-12	**67.58±0.23**	**81.41±0.20**	**71.38±0.25**	**84.65±0.20**

Table 3. Comparison with other state-of-the-art methods with 95% confidence intervals on *mini*ImageNet and *tiered*ImageNet. (Top two performances are in bold font.)

Model	Backbone	CIFAR-FS		FC100	
		1-shot	5-shot	1-shot	5-shot
Prototypical Networks [5]	Conv-64F	55.50±0.70	72.00±0.60	35.30±0.60	48.60±0.60
Relation Networks [6]	Conv-256F	55.00±1.00	69.30±0.80	-	-
R2D2 [16]	Conv-512F	65.30±0.20	79.40±0.10	-	-
Prototypical Networks [5]	ResNet-12	72.20±0.70	83.50±0.50	37.50±0.60	52.50±0.60
TADAM [17]	ResNet-12	-	-	40.10±0.40	56.10±0.40
MeaOptNet [18]	ResNet-12	72.60±0.70	84.30±0.50	41.10±0.60	55.50±0.60
MABAS [22]	ResNet-12	73.51±0.92	85.49±0.68	42.31±0.75	57.56±0.78
Fine-tuning [23]	WRN-28-10	**76.58±0.68**	**85.79±0.50**	43.16±0.59	**57.57 ±0.55**
PEAG	ResNet-12	74.27±0.23	83.89±0.20	**43.99±0.21**	56.47±0.24
MML	ResNet-12	**75.28±0.23**	**85.95±0.19**	**44.43±0.21**	**59.56±0.25**

Table 4. Experimental results compared with other methods on CIFAR-FS and FC100. (Top two performances are in bold font.)

level metric-learning module. Specifically, we change the values of α, β and γ, and experiment on the *mini*ImageNet.

As seen in Table 2, each part of the MML is indispensable. It can be observed that the accuracy of few-shot image recognition using only one level of features is very low. The results were significantly improved when two or three levels of features were used together, and the results were best when all three levels were used together. Specifically, compared with the method that only using pixel-level features, our MML gains 5.4% and 3.7% improvements. Note that it is important to choose the appropriate hyperparameters. When β and γ become larger, the accuracy of the model will become worse. Appropriate β and γ can provide useful auxiliary information for part-level metric.

5.4. Comparison Against Related Approaches

Results on ImageNet derivatives. As seen from Table 3, our MML achieves the highest accuracy on *mini*ImageNet with 67.58% and 81.41% on 5-way 1-shot and 5-way 5-shot tasks respectively, which make a great improvement compared to the previous single level metric-learning based methods. For example, our MML is 4.6% and 8.0% better than DSN-MR [8] and Prototypical Networks [5] on the 5-way 1-shot task, respectively. And our MML achieves 71.38% and 84.65% on *tiered*ImageNet under 5-way 1-shot and 5-way 5-shot few-shot learning setting respectively, which achieves competitive performance.

Results on CIFAR derivatives. Table 4 evaluates our method on two CIFAR derivatives, i.e., CIFAR-FS and FC100. It can be seen that the proposed MML obtains sig-

nificant improvements compared with previous state-of-the-art methods. Specifically, compared with global-level metric-learning based methods (i.e., Relation Networks [6], Prototypical Networks [5] and Fine-tuning [23]), MML is 20.3% and 3.5% better than the best one of them on CIFAR-FS and FC100 under 5-way 5-shot setting.

Moreover, we can also see that the proposed PEAG achieved competitive results. For example, our PEAG is 0.8% and 1.8% better than FEAT [19] and GLoFA [20] on *mini*ImageNet under the 5-way 1-shot setting, respectively.

The reason why our MML can achieve these state-of-the-art performances is that MML can measure the semantic similarities on multiple levels, i.e., part-level, pixel-level, and global-level.

6. CONCLUSION

In this paper, we revisit the metric-learning based method and proposed novel Part-level Embedding Adaptation with Graph (PEAG) method and Multi-level Metric Learning (MML) method for few-shot image recognition, aiming to capture more comprehensive semantic similarities. Specifically, PEAG can generate task-specific part-level features and capture the part-level semantic similarity between query images and support images, and MML can measure the semantic similarities on multiple levels and produce more discriminative features. Extensive experiments show the effectiveness and the superiority of both PEAG and MML.

7. REFERENCES

[1] Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan, "Semi-supervised learning for few-shot image-to-image translation," in *CVPR*, 2020, pp. 4452–4461.

[2] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang, "Episode-based prototype generating network for zero-shot learning," in *CVPR*, 2020, pp. 14032–14041.

[3] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu, "Self-supervised domain-aware generative network for generalized zero-shot learning," in *CVPR*, 2020, pp. 12764–12773.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, vol. 70, pp. 1126–1135.

[5] Jake Snell, Kevin Swersky, and Richard S. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 4077–4087.

[6] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199–1208.

[7] 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, pp. 7260–7268.

[8] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi, "Adaptive subspaces for few-shot learning," in *CVPR*, 2020, pp. 4135–4144.

[9] Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen, "Multi-scale adaptive task attention network for few-shot learning," *arXiv preprint arXiv:2011.14479*, 2020.

[10] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele, "Meta-transfer learning for few-shot learning," in *CVPR*, 2019, pp. 403–412.

[11] 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, pp. 8642–8649.

[12] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Workshops*, 2015, vol. 2.

[13] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3630–3638.

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

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[16] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2019.

[17] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste, "TADAM: task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, 2018, pp. 719–729.

[18] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, 2019, pp. 10657–10665.

[19] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*, 2020, pp. 8808–8817.

[20] Su Lu, Han-Jia Ye, and De-Chuan Zhan, "Tailoring embedding function to heterogeneous few-shot tasks by global and local feature adaptors," in *AAAI*, 2021, pp. 8776–8783.

[21] Yiluan Guo and Ngai-Man Cheung, "Attentive weights generation for few shot learning via information maximization," in *CVPR*, 2020, pp. 13496–13505.

[22] Jaekyeom Kim, Hyoungseok Kim, and Gunhee Kim, "Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning," in *ECCV*, 2020, vol. 12346, pp. 599–617.

[23] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto, "A baseline for few-shot image classification," in *ICLR*, 2020.