PROGRESSIVE POINT TO SET METRIC LEARNING FOR SEMI-SUPERVISED FEW-SHOT CLASSIFICATION

Pengfei Zhu[⋆] *Mingqi Gu*[⋆] *Wenbin Li*[†] *Changqing Zhang*[⋆] *Qinghua Hu*[⋆]

[⋆] College of Intelligence and Computing,Tianjin University, China
[†] State Key Laboratory for Novel Software Technology, Nanjing University, China

ABSTRACT

Few-shot learning aims to learn models that can generalize to unseen tasks from very few annotated samples of available tasks. The performance of few-shot learning is greatly affected by the number of samples per class. The massive unlabeled data can help to boost the performance of few shot learning models. In this paper, we propose a novel progressive point to set metric learning (PPSML) model for semi-supervised few-shot classification. The distance metric is defined for an image of the query set to a class of the support set by point to set distance. A self-training strategy is designed to select the samples locally or globally with high confidence and use these samples to progressively update the point to set distance. Experiments on benchmark datasets show that our proposed PPSML significantly improves the accuracy of few shot classification and outperforms the state-of-the-art semi-supervised few-shot learning methods.

Index Terms— Semi-supervised learning, few-shot learning, metric learning, self-training

1. INTRODUCTION

Few-shot learning trains models that can generalize well to new tasks when there are only few samples available per class in both training and test episodes. Previous works can be roughly divided into gradient descent based methods [1, 2, 3], data augmentation based methods [4, 5], metric learning based methods [6, 7, 8, 9, 10] and memory networks [11]. Instead of only being trained on the target task, few-shot learners are optimized over diverse tasks where few training examples are provided as in the real environment. The mechanism of few-shot learning can help to reduce the annotation effort for these data-hungry applications such as image classification, object tracking and image segmentation.

Among these advances, metric learning based methods achieve promising performance. They are mainly dependent on optimizing a transferable embedding space in which the similarity metric can determine the class of query set. Koch *et al.* [6] introduced metric based methods to few-shot learning by utilizing Siamese Neural Network. Prototypical network [8] performs classification by calculating the distance between a query sample and each prototype in the metric space. Different from the aforementioned methods using fixed distance metric, relation network [9] adopts a deep learnable metric for comparing query and support samples. Li *et al.* introduced image to class distance based on local descriptors to few shot learning and achieved superior performance [12, 13]. Compared with the image-level representation obtained by average pooling or fully connected layer, local features can help to alleviate information loss. For few shot classification tasks, metric learning based methods are suffering from a fundamental obstruction that a precise distance metric is hard to obtain due to example scarcity. Experimental results show that a more precise metric can be obtained with massive unlabeled samples. The focus of this paper is thus on the semi-supervised few-shot classification task.

In semi-supervised setting, models should have the ability to learn from a mixture of labeled and unlabeled samples. Ren *et al.* [14] proposed three semi-supervised variants of prototypical networks, which greatly improve the performance of the original model. Some other methods leverage unlabeled samples by constructing a graph using the union of the support set (labeled/unlabeled) and the query set [10, 15, 16]. Iterative label propagation or graph convolution is applied to propagate labels from the labeled support set to the query set, where the unlabeled samples act as intermediate senders of message. Inspired by the simple but effective semi-supervised method *i.e.*, self-training strategy, we consider two types of sample selection methods (local vs global) during training and test episodes. We first enlarge the labeled set based on the most confident predictions on unlabeled data and update the point to set metric to classify the query set. The core challenge is to find a trade-off between more training information and inevitable label noise.

Specially, we develop a progressive point to set metric learning (PPSML) model for semi-supervised few-shot classification. First, a shared CNN-based embedding module is adopted to produce deep local descriptors for both the support set and the query set. Then, local descriptors from the training samples in the same class are collected into a pool and additional descriptors are chosen from the unlabeled data set to augment it by the proposed self-training strategy. Typ-

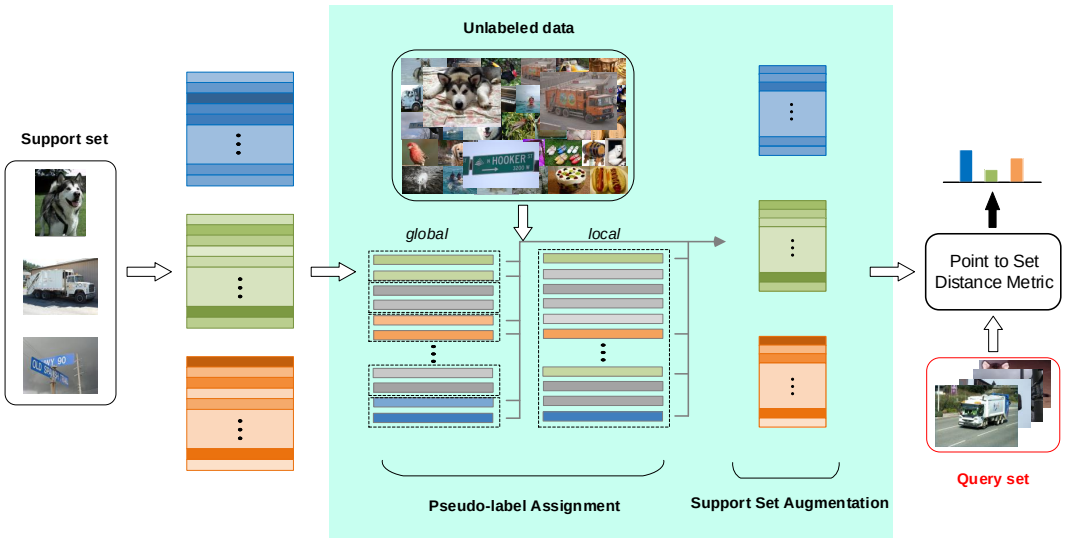

Fig. 1. The pipeline of the proposed semi-supervised few-shot learning framework on a single 3way-1shot task. Firstly, local descriptors are extracted from both support and query set by CNN. Then, unlabeled samples with pseudo labels of high confidence are selected globally or locally to augment the support set. The final classification is given by measuring the point to set distance from the sample of the query set to each class of the support set.

ically, a global selection method is designed to pick all descriptors from images with high classification confidence. On the other hand, partial local items with high matching scores are picked when employing the local selection method. Finally, a point to set distance is calculated for an image of the query set to a class of the support set as basis of classification. Experiments are conducted on a few-shot dataset and three fine-grained datasets, and our models achieve state-of-the-art results. Moreover, we demonstrate that PPSML is robust to distractor points that do not belong to any training classes of the episode.

2. OUR APPROACH

2.1. Problem Formulation

Given an auxiliary set containing a large number of labeled samples of classes C_{train}, the objective is to train a transferable classifier for an unseen set of novel classes C_{test}, where a few labeled examples per class are available. We follow the episodic paradigm, which is commonly adopted in few-shot learning. In each episode, N different classes are first sampled from C_{train} and a support set $\mathcal{S}$ is formed by K labeled samples for each class, while a query set $\mathcal{Q}$ includes different samples from $\mathcal{S}$. This setting is called N-way K-shot classification. The support set $\mathcal{S}$ in each episode is used for adaptation and the model is updated by the loss from its predictions for the query set $\mathcal{Q}$. Following the semi-supervised few-shot setup in [14], we first split the images of each class into disjoint labeled and unlabeled set. An unlabeled set $\mathcal{U}$ is also utilized in each episode, which is from the unlabeled split. In a more realistic and challenging setting, the unlabeled set $\mathcal{U}$ also includes additional samples (distractors) from irrelevant classes. During episodic training, the model learns to extract useful information from the unlabeled set to assist its predictions for the query set.

2.2. Point to Set Distance Metric

At each episode, a given query image q from the query set $\mathcal{Q}$ will be embedded as $\Psi(q) = [\boldsymbol{x}_1, ..., \boldsymbol{x}_m] \in \mathbb{R}^{d\times m}$ through the module and we can obtain a class set $\mathcal{D}_c$ containing local descriptors $\hat{\boldsymbol{x}}_j^c|_{j=1}^n$ from K labeled samples of class c. Firstly, we calculate the distance metric between descriptor $\boldsymbol{x}_i$ from the query image and all local descriptors in the class set. Then, we find its k-nearest neighbors and sum up the k largest results as the point to set distance metric with respect to a single descriptor:

$$\begin{aligned} \phi(\boldsymbol{x}_i, \mathcal{D}_c) &= \sum_{j=1}^{k} \cos(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_j^c) \\ \cos(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_j) &= \frac{\boldsymbol{x}_i^\top \hat{\boldsymbol{x}}_j}{\|\boldsymbol{x}_i\| \cdot \|\hat{\boldsymbol{x}}_j\|} \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{x}}_j^c \in k\text{-}NN(x_i)$ and $\cos(\cdot)$ is the cosine similarity. Other similarity or distance function can also be explored. Finally, the point to set distance from the sample of the query set to the class set can be defined as:

$$\Phi(\Psi(q), \mathcal{D}_c) = \sum_{i=1}^{m} \phi(\boldsymbol{x}_i, \mathcal{D}_c) \tag{2}$$

Table 1. Semi-supervised Few-shot Classification Accuracies(%) on *mini*ImageNet with and without distractors

Models	*mini*ImageNet			
	1-shot	1-shot w/D	5-shot	5-shot w/D
Soft-*k*means*[14]	50.09±0.45	48.70±0.32	64.59±0.28	63.55±0.28
Soft-*k*means+Cluster*[14]	49.03±0.24	48.86±0.32	63.08±0.18	61.72±0.24
Masked Soft-*k*means*[14]	50.41±0.31	49.04±0.31	64.39±0.24	62.96±0.14
MetaGAN*[17]	50.35±0.23	N/A	64.43±0.27	N/A
TPN-semi*[15]	52.78±0.27	50.43±0.84	66.42±0.21	64.95±0.73
DN4 baseline‡[12]	46.48±0.77	46.48±0.77	66.61±0.67	66.61±0.67
Our PPSML(local)	48.44±0.86	48.99±0.86	69.13±0.72	69.03±0.72
Our PPSML(global)	51.61±1.02	50.74±0.96	73.76±0.72	73.00±0.75
Global + Higher Shot	**60.11±1.00**	**57.35±1.01**	**76.88±0.71**	**74.65±0.73**

*Results reported by the original work, ‡Results re-implemented in the same setting.

The class corresponding to the set with the maximum response will be the prediction of the query image. During training, the distances to all class sets are obtained and a softmax layer with the cross-entropy loss is utilized to optimize the model. As can be seen in the formulation, the point to set metric is heavily dependent on the similar descriptors in the class set, the diversity of which is limited by the number of shots. This motivates us to include more labeled information from massive unlabeled samples by a self-training strategy..

2.3. Global or Local Self-training

For massive unlabeled samples, directly assigning the pseudo labels using the current few shot classifier may result in classification deviation due to label noises. Hence, we carry out two strategies in our PPSML model to find a trade-off between more training information and label noises. The complete pipeline is illustrated in Fig.1.

Local: Assuming that all local descriptors from the unlabeled set as individual instances, we score them by cosine similarity function and choose the items with the highest matching scores. Additionally, we employ inverse similarity weights to enhance their discrimination ability between classes. We hope to pick the items that are the most similar to the current class and dissimilar to other classes. The purpose behind the local PPSML is to avoid possible noise by selecting individual local descriptors. For every individual local descriptor $\tilde{\boldsymbol{x}}_i$ embeded from the unlabeled set $\mathcal{U}$, its matching score for each class set can be defined as:

$$\begin{aligned} d_{local}(\tilde{\boldsymbol{x}}_i, \mathcal{D}_c) &= w_{inverse} * \phi(\tilde{\boldsymbol{x}}_i, \mathcal{D}_c) \\ w_{inverse} &= \frac{\phi(\tilde{\boldsymbol{x}}_i, D_c)}{\sum_{c=1}^{N} \phi(\tilde{\boldsymbol{x}}_i, D_c)} \end{aligned} \tag{3}$$

Global: Although the above-mentioned strategy may be beneficial, it is harmful to abandon the relationship between local descriptors. Regarding the local descriptors from the same image as a whole and choosing the images with the highest classification scores are employed in global PPSML. Noise is unavoidable due to classification errors but it introduces more various and valuable items into the class sets. For an unlabeled image u from the unlabeled set $\mathcal{U}$, its classification score is defined as:

$$d_{global}(\Psi(u), c) = softmax(\Phi(\Psi(u), D_c)) \tag{4}$$

We pseudo-label the images with high classification confidence and add all the descriptors from them to the corresponding class set. No matter employing local or global PPSML, a progressive distance metric with enlarged class sets will be applied to predict the query image, which breaks through the restriction of labeled shots in the support set.

3. EXPERIMENTS

In this section, we demonstrate that our proposed PPSML method improves the baseline network by effectively using unlabeled information. We compare our semi-supervised few-shot learning framework with several state-of-the-art methods on a common few-shot classification dataset, *i.e.*, *mini*ImageNet [7] and three fine-grained datasets, *i.e.*, *StanfordCars* [18], *StanfordDogs* [19] and *CUB Birds* [20].

3.1. Datasets

***mini*ImageNet** is composed by 100 classes randomly chosen from ImageNet and each class has 600 images with a resolution of 84×84. Following the split used by [2], 64, 16 and 20 classes are used for training, validation and test, respectively. Fine-grained dataset **StanfordCars** includes 196 classes of cars with a total number of 16,185 images, which is much smaller than few-shot classification datasets. Similarly, **StanfordDogs** contains 20,580 dog images of 120 categories and **CUB-200** dataset contains 6033 images from 200 bird species. For each dataset, 40% are sampled for the labeled split and remaining 60% are used as the unlabeled set.

Table 2. Semi-supervised Few-shot Classification Accuracies(%) on three fine-grained datasets

Models	*Stanford Dogs*		*Stanford Cars*		*CUB-200*	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Matching Nets FCE†[7]	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 Nets†[8]	37.59±1.00	48.19±1.03	40.90±1.01	52.93±1.03	37.36±1.00	45.28±1.03
GNN†[10]	46.98±0.98	62.27±0.95	55.85±0.97	71.25±0.89	51.83±0.98	63.69±0.94
DN4 baseline‡[12]	46.85±0.80	68.16±0.74	52.12±0.82	85.93±0.50	53.96±0.94	72.12±0.80
Our PPSML(local)	46.97±0.80	69.29±0.65	57.61±0.84	86.24±0.48	55.75±0.93	73.48±0.81
Our PPSML(global)	**52.16±0.95**	**72.00±0.68**	**71.71±0.98**	**90.02±0.43**	**63.43±1.08**	**78.76±0.78**

†Results from [12] trained in purely supervised setting, ‡Results re-implemented in semi-supervised setting.

3.2. Experimental Setting

For a fair comparison, we adopt the same architecture used in several recent works. The network Ψ consists of four stacked convolutional blocks, abbreviated as ***Conv-64F***, each of which contains a 2D convolutional layer (with 64 filters of size 3×3), a batch normalization layer and a Leaky ReLU layer. In our experiments, we cover two main results: with and without distractors (from different classes). Semi-supervised 5-way 1-shot and 5-way 5-shot with or without discractors classification tasks are conducted on the traditional few-shot dataset, while extensive experiments are on three fine-grained datasets. The deep local descriptor at each position of the final feature map is normalized by its l_2-norm.

During training, we randomly sample and construct 300,000 episodes to train our models. In each episode, the query set contains 10 query images from each class and 10 and 30 unlabeled images will also be selected from each class for the 1-shot and 5-shot setting, respectively. Optionally, when including distractors, we additionally sample 5 other classes from the set of training classes. In all experiments, the hyper-parameter k-nearest neighbors is set as 1 and 3 for 1-shot and 5-shot setting and the hyper-parameter selection ratio r is set as 0.02 and 0.1 for local and global method. To train our model, we adopt Adam algorithm with an initial learning rate of 5×10^{-3} and reduce it by half for every 100,000 episodes. In addition, we perform data augmentation on training set: a 84×84-pixel image is randomly cropped and then undergoes random color jittering, and random horizontal flip, all available in PyTorch torchvision package.

3.3. Results Comparison

Few-shot Classification. To illustrate the effect of our method, we re-implement the baseline method DN4 in semi-supervised setting. Because we are strictly using less label information than in the previously supervised work on the dataset, the results that we report are slightly worse than the published numbers in the original papers. Results for *mini*ImageNet are given in Table 1, with 95% confidence intervals. We can observe clear improvements over the baseline whether using global or local PPSML. In particular, global PPSML shows the state-of-the-art performance in 5-shot problem, surpassing TPN-semi about 7% in accuracy. Considering that models' classification accuracy is still at a low-level in low-shot problem, the effect of the local or global selection is trivial. To tackle this problem, we adopt "Higher Shot" during training to more fully optimize our models but test models using the same shots, which leads to about 8% increase in 1-shot task and about 3% increase in 5-shot task. It is also worth noting that our method shows good robustness to distractors, which also achieves outstanding results.

Fine-grained Few-shot Classification. In general, the fine-grained few-shot classification is more challenging due to the smaller inter-class variations and the improvement of our method is limited by the small number of unlabeled images for each class. But our method consistently brings obvious growth in classification accuracy as presented in Table 2. Even compared with models trained in purely supervised setting, our method still achieves more remarkable performance. Similarly, global PPSML is superior to local PPSML on the fine-grained datasets, which proves that maintaining the relationship of local descriptors is fundamental when designing a semi-supervised algorithm.

4. CONCLUSION

In this paper, we proposed a novel progressive point to set metric learning (PPSML) model for semi-supervised few shot classification. To exploit massive unlabeled samples in the wild, a self-training strategy is designed to automatically select samples with pseudo labels to augment the support set and therefore progressively strengthen the point to set metric. Experimental results on several benchmark datasets demonstrate that our proposed PPSML significantly boosts the performance of few shot classification and outperforms the state-of-the-art supervised and semi-supervised methods.

5. ACKNOWLEDGEMENTS

This work was supported by National Key R&D Program of China (No.2019YFB2101904), National Natural Science Foundation of China under Grants 61732011 and 61876127.

6. REFERENCES

[1] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[2] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *5th International Conference on Learning Representations, (ICLR)*, 2017.

[3] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 403–412.

[4] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein, "Delta-encoder: an effective sample synthesis method for few-shot object recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 2845–2855.

[5] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10275–10284.

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

[7] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[8] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[9] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[10] Victor Garcia Satorras and Joan Bruna Estrach, "Few-shot learning with graph neural networks," in *6th International Conference on Learning Representations, (ICLR)*, 2018.

[11] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7260–7268.

[13] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8642–8649.

[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 *6th International Conference on Learning Representations, (ICLR)*, 2018.

[15] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *7th International Conference on Learning Representations, (ICLR)*, 2019.

[16] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11–20.

[17] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song, "Metagan: An adversarial approach to few-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 2371–2380.

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[19] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011, vol. 2.

[20] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona, "Caltech-ucsd birds 200," 2010.