Enhancing the Generalization Performance of Few-Shot Image Classification with Self-Knowledge Distillation

Liang LI[1], Weidong JIN[1,2], Yingkun HUANG[3], Junxiao REN[1]*

[1] School of Electrical Engineering, Southwest Jiaotong University, Chengdu 611756, China
liangli@my.swjtu.edu.cn, wdjin@home.swjtu.edu.cn
rjx19910911@my.swjtu.edu.cn (**Corresponding author*)

[2] ASEAN International Joint Laboratory of Integrated Transportation, Nanning University, Nanning City, Guangxi Province, China

[3] National Supercomputing Center in Shenzhen (Shenzhen Cloud Computing Center), Shenzhen 518055, China
hykun@live.com

Abstract: Though deep learning has succeeded in various fields, its performance on tasks without a large-scale dataset is always unsatisfactory. The meta-learning based few-shot learning has been used to address the limited data situation. Because of its fast adaptation to the new concepts, meta-learning fully utilizes the prior transferrable knowledge to recognize the unseen instances. The general belief is that meta-learning leverages a large quantity of few-shot tasks sampled from the base dataset to quickly adapt the learner to an unseen task. In this paper, the teacher model is distilled to transfer the features using the same architecture. Following the standard-setting in few-shot learning, the proposed model was trained from scratch and the distribution was transferred to a better generalization. Feature similarity matching was proposed to compensate for the inner feature similarities. Besides, the prediction from the teacher model was further corrected in the self-knowledge distillation period. The proposed approach was evaluated on several commonly used benchmarks in few-shot learning and performed best among all prior works.

Keywords: Self-knowledge distillation, Meta-learning, Few-shot learning, Similarity matching.

1. Introduction

Deep learning has achieved many promising results in pattern recognition tasks (Ding et al., 2021; Liang et al., 2020). Deep learning reduces the intensity and experience required for professional data analysis and provides accurate generalization for independently identically distributed data (Tian & Fu, 2020; Li et al., 2021). Without the support of sufficient data, deep learning algorithm tends to fail in the overfitting situation. This significant gap between deep learning and intelligence has attracted researchers' attention. The method to solve limited data for deep learning is termed few-shot learning. Recently, many works (Shorfuzzaman & Hossain, 2021; Feng & Duarte, 2019) have tackled such problems. Meta-learning learns from a variety of tasks. Such an approach is referred to meta-learning based few-shot learning, which is trained on a series of datasets and sequentially produces a corresponding learner. By dividing the tasks into disjoint sets, the meta-training set consists of limited training data, and the trainer is trained on a task sampled from the task distribution. Then the learner can be evaluated on the corresponding meta-testing set. Thus, the meta-learner achieves high classification performance on the test dataset.

Previous meta-learning methods focused on achieving generalization for few-shot learning. The existing meta-learning based few-shot learning can be divided into 1) learning an initialization, 2) learning an optimizer, and 3) memory-based method. The methods of learning an initialization assume that a global initialization (Finn et al., 2017) learned from the base dataset can quickly adapt to the novel dataset. The methods (Ravi & Larochelle, 2017; Rusu et al., 2019) are about learning an optimizer how to optimize the model's parameters. The memory-based methods (Mishra et al., 2018) leverage external memory from past episodes to tackle few-shot learning. Both learning an initialization and learning an optimizer neglect to improve the generalization performance via learned knowledge. The closest method to the one proposed in this paper is the third one. Unlike utilizing the learned experiences in the memory module, the present approach utilizes the knowledge from the well-trained model. The main factor for meta-learning is the fast adaptation which indicates that the feature reuse is vital in the whole process. Simply finetuning on a large-scale dataset easily overfits the target dataset with limited samples. Considering a certain correlation between support images and query images, the knowledge and features learned in the base dataset construct the classification model for the novel dataset. Following this idea, some works showed that benefiting of representations from the base

https://doi.org/10.24846/v31i2y202207

dataset contributes to the fast adaptation to the novel dataset. The baseline trains the model to learn the basic knowledge through the meta-training dataset. A linear classifier is utilized to finetune the model on the meta-test dataset. The above methods train the model from scratch without considering the trained model. The previously trained model has a comprehensive understanding of the overall distribution of the dataset. Therefore, the previously trained model is essential to guide current training. To utilize the previous knowledge guiding the training, knowledge distillation is introduced to transfer predicted distribution by matching certain statistics between the teacher and student models.

Knowledge distillation (Gou et al., 2021) can transfer informative relationships. Prior works improved the performance of a smaller model from the guidance of a large model. The compact network (Li et al., 2020) is compressed by various network pruning and weight decomposition methods. The layer-wisely accumulated errors are reduced via the layer-wise knowledge distillation approach (Bai et al., 2020), leading to a more robust and generalizable student network. One of the core challenges of few-shot knowledge distillation methods lies in unequal feature representative dimensions and errors introduced from a large network during inference. The high-dimension embedding representation is more expressive than the low-dimension in feature representation (Hu et al., 2021). A larger embedding dimensionality can reduce the capacity gap between embedding and feature spaces. However, this will increase the difficulty in producing well-generalized embedding for a small model because of the higher computing costs and higher chances of being wrongly classified. Besides, the standard meta-learning based few-shot learning trains the model from scratch. The strategy neglects to consider the previous knowledge as guidance. Most existing knowledge distillation methods transfer the prediction distribution as additional knowledge (Fu et al., 2021), while the feature relationship is not fully utilized during the knowledge distillation period.

This paper proposes a novel few-shot learning algorithm based on the meta-learning and knowledge distillation strategy to improve the model's performance. Since the well-trained model has obtained enough knowledge on specific tasks, the learned knowledge can be applied to guide the current model's training without introducing a larger model. This self-knowledge distillation approach could push the lower-dimension feature representation to embed more expressive semantic knowledge. Transferring more knowledge is an effective means of improving generalization capabilities. By preserving the mutual similarities between samples in every batch, the approach improves the generalization capacity of the original embedding space. Unlike the simple feature distribution matching during the self-knowledge distillation period, the student model learns more relationships within one batch than just applying distribution matching. Besides, the performance of the proposed approach was further improved with corrected self-knowledge distillation to the student model by making sure the predictions of the teacher model are corrected during training. Transferring the corrected feature similarities contributes to better guidance (Wen et al., 2021). The informative feature vector provided from the trained model can enhance the feature embedding generalization on the unseen categories.

The rest of the paper is organized as follows. Section 2 reviews the existing Few-Shot Learning and knowledge distillation. In Section 3, the proposed main algorithm and its implementation details are described. Section 4 describes the experimental settings. In Section 5, the results are analysed. Finally, this paper is concluded in Section 6.

2. Related Works

2.1 Few-shot Learning

The core idea of few-shot learning is that the model generalizes well to the unseen categories. Meta-learner (Ravi & Larochelle, 2017) aims to capture current knowledge within a task and remember them by the Long Short-Term Memory (LSTM) mechanism. Unlike the previous method, Meta networks (Munkhdalai & Yu, 2017) learn meta-level knowledge across tasks and shift their inductive biases via fast parameterization for rapid generalization. Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) uses a recurrent neural network to describe the distribution of all tasks and task-specific parameters to fine-tune the model. The above methods train the model from scratch without considering the trained model. To utilize the previous knowledge guiding the training, the knowledge distillation method is employed to transfer predicted distribution by matching certain statistics between the teacher and student models.

2.2 Knowledge Distillation

The knowledge distillation method can transfer informative relationships. Knowledge distillation usually utilizes a teacher-student strategy to extract the knowledge from the teacher model and guide the student model's training. Most existing knowledge distillation methods transfer the prediction distribution as the additional knowledge, while the feature relationship is not fully utilized during the knowledge distillation period. The standard knowledge distillation approaches usually require a complex high-performance teacher model. Besides, transferring the more enriched features should consider the different architectures and feature resolutions. Thus, the self-knowledge distillation approach is considered instead.

3. Method

3.1 Problem Definition

From the perspective of datasets, the datasets can be divided into two parts. One is a base dataset $D_{base} = \{(x_i, y_i)\}_{i=1}^{N_{base}}$, and another is a novel dataset $D_{novel} = \{(\overline{x}_i, \overline{y}_i)\}_{i=1}^{N_{novel}}$. The (x, y) represents an image and its corresponding label. N_* is the number of the total images. The categories in the base dataset are disjoint from the categories in the novel dataset. During few-shot learning, the episode strategy forms the training and testing. If C classes with K images per class are selected to form the query dataset, the training is referred to as C-way K-shot few-shot learning. The task consists of a small set of labeled support images and unlabeled query images. Both the support and query images are sampled from the same class. The classifier is adjusted on the labeled support images to recognize the query images correctly.

3.2 Embedding Backbone

It is essential to extract enough meaningful features due to the limited instances. Besides, the generalization performance is also necessary for such a limited feature situation. A generalized model is applied to obtain meaningful embedding features for the downstream tasks. ResNet introduces the identity shortcut connection to fit the desired mapping by taking advantage of enough capacity and preventing the overfitting mechanism. ResNet has been used as a backbone network for feature extraction because of its versatility and good performance.

3.3 Meta-learning

The meta-learning involves data, algorithm, and selection of model hyperparameters that contribute to an excellent procedure. During the meta-training period, relevant data are utilized to learn a better initial weight. Thus, the well-trained model can quickly fine-tune downstream tasks. Under the few-shot learning setting, our final target is to achieve such a learning procedure.

In meta-learning based few-shot learning, a meta-learner is trained to learn the shared knowledge in the base dataset and then adjusted on the novel dataset. Different tasks contain different categories, so the learner trained on the base dataset is learned to recognize the novel categories. Because the base categories are disjoint from the novel categories, the learner is adjusted with limited support instances to classify the query instances. Thus, the learner should quickly adapt to the novel tasks. Finally, the learner learns high-level transferable knowledge and task-specific knowledge.

If the high-level transferable knowledge can represent the common features among all the tasks, slightly fine-tuning the parameters can produce good results. Following the algorithm (Dhillon et al., 2020), the tasks from the training set are merged into a single task, and a classifier is performing classification on this combined dataset. The details of the training algorithm are shown in Algorithm 1.

Algorithm 1. General training algorithm of meta-training based few-shot learning

Input: Embedding model **f**, dataset **D**
Output: Meta-parameters **w**
Randomly initialize **θ**
While true:
Generate meta-tasks from **D** and merge all meta-tasks into one task **T**
For each epoch in total training epochs:
    Randomly select samples from the task **T** and train the embedding model
    Compute the loss on the testing dataset for all instances
    Update the model parameters by the loss **θ**
End for
Save the updated meta-parameters **w'**

Here, the ResNet-12 is selected as the embedding model to learn transferable knowledge as in equation (1):

$$y = f_\theta(x) \tag{1}$$

where x and θ denote the input images and learner module parameters.

The objective is to minimize the error between the model prediction and its actual label through the loss function as in equation (2):

$$\theta' = \min(L^{base}(D_{train};\theta,w) + R(\theta)) \tag{2}$$

where the L^{base} is a loss function of the learner and R(θ) is a normalization item to avoid overfitting. During training on the training dataset D_{train}, the embedding model parameter θ is updated during iteration with fixed meta-learner parameter w. The meta-learner minimizes the average test error of the learner on the distribution of tasks p(T) as in equation (3):

$$w' = \min E_{\Gamma\sim p(\Gamma)}(L^{meta}(D_{test};\theta',w)) \tag{3}$$

where the L^{meta} is a loss function of the meta-learner and D_{test} is the test dataset. W' is the updated meta-learner parameter.

Equation (4) was used to evaluate and verify the performance of the meta-test datasets. The meta-learning process is shown in Figure 1.

$$E_s(L^{meta}(D_{test};\theta',w')) \tag{4}$$

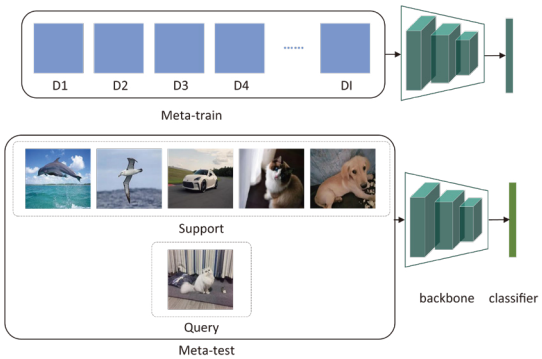

Figure 1. Meta-learning process

3.4 Generalized Self-knowledge Distillation

Knowledge distillation transfers the knowledge from another effective classification model to produce a well-generalized embedding that enhances the effectiveness of handling unseen tasks. The standard knowledge distillation applies the high dimension to represent the extracted features, leading to the conflict between feature dimensions in different layers. The model capacity gap between them can degrade the performance of knowledge transfer. To effectively utilize the knowledge in the teacher model and avoid introducing a more complex algorithm, self-knowledge distillation is taken in account to solve such issues.

Similar to the stand knowledge distillation, the self-knowledge distillation method replaces the larger model with itself and uses the historical knowledge to supervise the current training process as in equation (5):

$$w' = \arg\min(\alpha L(D_{train}, Y; w') + \beta KL[f(D_{train};w'), f(D_{train};w)]) \tag{5}$$

where the α and β are the weights for their corresponding loss functions. Y is the class label. L and KL are cross-entropy and KL-divergence loss functions, respectively. In self-knowledge distillation-based few-shot learning, the trained model is considered the teacher and its output is applied as the supervision for the next training process from the scratch. The next generation of training programs aims to minimize the weighted loss functions in equation (5). The first loss is a cross-entropy loss between model predictions and their actual class labels. The second is the Kullback-Leibler divergence (Bu et al., 2016) (KL) between predictions and historical predictions. The predictions used in the loss are the classified features. The training circle is repeated three times and then the best model is used for testing.

The self-knowledge distillation method benefits from the feature complexity and model parameters. The detailed diagram for the whole process is shown in Figure 2.

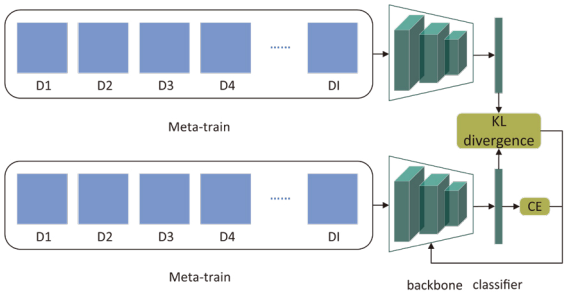

Figure 2. Self-knowledge distillation process

Both two models are trained with the same dataset. The bottom model is about to be trained under the supervision of the upper well-trained model.

Algorithm 2. Swapping the value for KL-divergence

Input: teacher model prediction p_t, ground truth y_gt
Output: corrected teacher model prediction p_t
For i in len(p_t):
    true_index = argmax(softmax($y_t(x_i|t)$))
    If true_index != y_gt[i]:
        p_t[i][y_gt[i]], p_t[i][true_index] = p_t[i][true_index], p_t[i][y_gt[i]]
End for

The two models are trained to produce features with the same batch during the model distillation period. The upper model does not backward the loss. The bottom model takes the KL-divergence loss between the two models' features and cross-entropy loss of its current prediction with ground truth. The bottom model is considered more discriminative than the previous model with a fully trained model's guidance.

However, with a similar training strategy, the self-knowledge distillation probably introduces a misguide due to the primordial mistakes in the teacher model. Following the teacher model's prediction, the student model could not learn the accurate distribution of the tasks if the mistakes exist both in the teacher and student models. Since nearly half the prediction of the student model on CIFAR-10 and CIFAR-100 datasets are incorrect, the transfer of the corrected prediction is required, which enhances the generalization. A simple improvement is introduced on the original KL divergence. During the generalized self-knowledge distillation period, if the maximum prediction of the teacher model is not located on the correct ground-truth label, their values between the prediction and the position of the ground-truth label are swapped as shown in Algorithm 2.

3.5 Similarity Matching

Although KL-divergence has been introduced to measure the distribution similarity between the historical predictions and current ones, the feature relationships within the batch itself have not been fully considered. In statistics, the statistical analysis method is used to determine the quantitative relationships between two or more variables that are dependent on each other. During the distillation period, the student model gradually learns the feature similarities from the teacher model. To utilize the learned predicted distribution to determine the relative relationships between features and generalizes to the unseen tasks, the proposed loss function attempts to measure the distance between features and transfer the feature relationships. The loss function is designed depending on pairs of feature similarities and defined as in equation (6):

$$L_{sim} = \sum_{(x_i, x_j) \in D_T} l_{sim}(u(x_i, x_j), v(x_i, x_j)) \tag{6}$$

The similarity matching loss function provides the student model with better pairwise feature relationships. The input samples x_i and x_j from one batch DT are utilized during distillation to extract features by teacher model u and student model v, respectively. Considering the significant feature difference at initial training and the slight difference after enough iterations, l_{sim} is a regression loss as in equation (7):

$$l_{sim}(u,v) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 0.5*(v-u)^2, & if\ |v-u| < 1 \\ |v-u| - 0.5, & otherwise \end{cases} \tag{7}$$

To infer the relationships within the pairwise features, the Huber loss-based similarity matching loss (Lei et al., 2019; Huang & Wu, 2021) is proposed to address the similarity matching issue focusing on the inner feature relationships. Compared with the commonly used regression loss function like MSE and MAE, Huber loss function enhances its robustness and reduces the equalization of gradients. Therefore, Huber loss updates parameters at a faster speed and achieves the global optimal value. Since the novel dataset has no intersection with the base dataset in the few-shot learning, the category of the minimum loss between query sample and each support category is regarded as the class of the query sample. This property is also appropriate for the similarity matching. The distance between features is large at the beginning and converges to a small value in the end. Therefore, the student model learns the proper relationships between the features. The feature similarities within a batch are further normalized to ensure stable convergence as in equation (8):

$$u(x_i, x_j) = \frac{\left\| x_i - x_j \right\|_2}{\frac{1}{|D_T|}\sum_{(x_i, x_j) \in D_T} \left\| x_i - x_j \right\|_2} \tag{8}$$

Firstly, the proposed model is trained from scratch with cross-entropy loss. The first model is then used as the teacher model to transfer prediction to the next generation of models. The current model is iterated with the latest trained model during the distillation period. From the simple supervised model to the distilled model, the teacher model always transfers prediction to guide the student model's training. The diagram for the complete process is shown in Figure 3. When the distillation period begins, the generalized self-knowledge distillation strategy is applied with multiple loss functions, including corrected KL-divergence loss, similarity matching loss, and cross-entropy loss. Therefore, the final object for the distillation period is $\alpha L_{ce} + \beta KL_c + \gamma L_{sim}$.

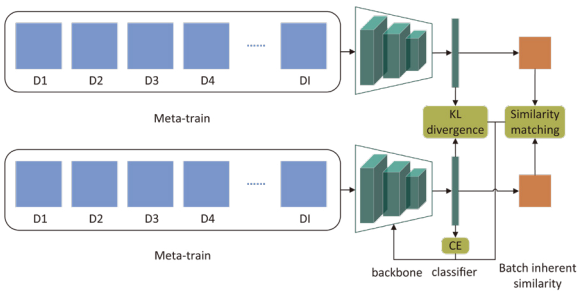

Figure 3. Similarity matching

4. Experimental Settings

4.1 Datasets

The miniImageNet dataset (Vinyals et al., 2016) is a commonly used standard for few-shot learning in recent years. As a subset of ImageNet, the dataset randomly picks 100 classes, and each class contains 600 images. The split strategy was applied (Ravi & Larochelle, 2017) and the dataset was divided into 64, 16, 20 classes for training, validation, and testing, respectively.

The tieredImageNet dataset (Ren et al., 2018) is another subset for few-shot learning with more categories. The dataset was divided into 608 classes. There are 351 classes for training, 97 classes for validation, and 160 for testing. Such splitting ensured the distinction between training and testing.

The CIFAR-FS dataset (Bertinetto et al., 2019) and the FC100 dataset (Oreshkin et al., 2018) are generated from the CIFAR100 dataset based on different selection criteria. The front dataset follows the miniImageNet and was divided into 64/16/20 classes for training/validation/testing. The latter follows the tieredImageNet and is split into 60,20,20 classes for training, validation, and testing.

4.2 Optimization Details

ResNet12 is applied as the backbone and the weights are equally set for all the losses in balancing the magnitude for different tasks (Tian et al., 2020). The total training epochs for miniImageNet and tieredImageNet are 100 and 60, respectively. Images are resized to 84 × 84 pixels.

The first training stage aims at training the learner based on the cross-entropy loss. During the meta-training period, Stochastic Gradient Descent (SGD) is set as an optimizer with a momentum of 0.9. The learning rate is 0.05, and the decay rate is 0.1 for different datasets. The learning rate is decayed at 60 and 80 epochs for miniImagenet and at 30, 40, and 50 epochs for tieredImageNet. The model is trained for 100 iterations for miniImageNet and 60 iterations for tieredImageNet. For two CIFAR100 based datasets, the learning rate is decayed at 45, 60, and 75 epochs. The model is trained for 90 iterations. The set optimizer is SGD, and it is optimized by cross-entropy loss.

The second stage aims at distilling the trained knowledge into the current model. The same learning schedule is used during the meta distillation period. After distillation, the model evaluates the meta-test dataset. The accuracy is calculated with 95% confidence intervals over test episodes. The test episode is set to 600. Using the base learner to extract feature embeddings for the support and query images, a simple logistic regression classifier is trained to recognize the query images.

5. Experimental Results

5.1 Comparison with Prior Works

The proposed approach is evaluated according to the few-shot setting during the meta-testing period. The accuracy of every experiment is conducted with 600 randomly sampled tasks. The symbol "-" indicates no reported result.

For the miniImageNet dataset, the present approach outperforms all previous works by at

least 1.4% for the one-shot setting and by 0.9% for the five-shot setting. More details are shown in Table 1. From the table, it can be found out that knowledge distillation can effectively improve the recognition accuracy in the meta-learning based few-shot image classification without a larger teacher model. Based on the corrected KL-divergence, the model can further enhance the feature reuse and construct a discriminative feature space under the proposed similarity matching loss constraint.

Table 1. Comparisons in miniImageNet

Model	Backbone	miniImageNet 5-way	
		1-shot	5-shot
(Oreshkin et al., 2018)	ResNet-12	58.50±0.30	76.70±0.30
(Ravichandran et al., 2019)	ResNet-12	59.04	77.64
(Gidaris et al., 2019)	WRN-28-10	63.77±0.45	80.70±0.33
(Dhillon et al., 2020)	WRN-28-10	57.73±0.62	78.17±0.49
(Vinyals et al., 2016)	ResNet-12	63.08±0.80	75.99±0.60
(Snell et al., 2017)	ResNet-12	60.37±0.83	78.02±0.57
(Rusu et al., 2019)	WRN-28-10	61.76±0.08	77.69±0.12
(Lee et al., 2019)	ResNet-12	62.64±0.82	78.63±0.46
(Yoon et al., 2019)	ResNet-12	61.65±0.15	76.36±0.10
(Tian et al., 2020)	ResNet-12	63.92±0.80	80.55±0.63
The present approach	ResNet-12	**65.39±0.73**	**81.51±0.64**

For the tieredImageNet dataset, the present approach outperforms all previous works by 0.5% for the one-shot setting and nearly by 1.3% for the five-shot setting. More details are shown in Table 2. Compared to the miniImagenet dataset, the categories and amount of the tieredImageNet dataset are much more prosperous, and the present approach is still competitive among other works.

Table 2. Comparisons in tieredImageNet

Model	Backbone	tieredImageNet 5-way	
		1-shot	5-shot
(Ravichandran et al., 2019)	ResNet-12	63.52	82.59
(Gidaris et al., 2019)	WRN-28-10	70.53 ± 0.51	84.98 ± 0.36
(Dhillon et al., 2020)	WRN-28-10	66.58 ± 0.70	85.55 ± 0.48
(Vinyals et al., 2016)	ResNet-12	68.50±0.92	80.60±0.71
(Snell et al., 2017)	ResNet-12	65.65±0.92	83.40±0.65
(Rusu et al., 2019)	WRN-28-10	66.33±0.05	81.44±0.09
(Lee et al., 2019)	ResNet-12	65.99±0.72	81.56±0.53
(Yoon et al., 2019)	ResNet-12	63.08±0.15	80.26±0.12
(Tian et al., 2020)	ResNet-12	70.90±0.89	84.83±0.65
The present approach	ResNet-12	**71.43±0.94**	**86.13±0.66**

For the CIFAR-FS and FC100 datasets, the present approach outperforms all previous works for the one-shot and five-shot settings. More details are shown in Table 3. The results verify the proposed design in which the model's generalization performance can be effectively improved based on more effective feature relationships for few-shot learning.

Table 3. Comparison to prior works in CIFAR-FS and FC100

Model	Backbone	CIFAR-FS 5-way		FC100 5-way	
		1-shot	5-shot	1-shot	5-shot
(Oreshkin et al., 2018)	ResNet-12	-	-	40.10±0.40	56.10±0.40
(Ravichandran et al., 2019)	ResNet-12	69.2	84.7	-	-
(Qiao et al., 2019)	ResNet-12	70.4	81.3	-	-
(Snell et al., 2017)	ResNet-12	72.20±0.70	83.50±0.50	37.50±0.60	52.50±0.60
(Lee et al., 2019)	ResNet-12	72.60±0.70	84.30±0.50	41.10±0.60	55.50±0.60
(Gidaris et al., 2019)	WRN-28-10	73.60±0.30	86.00±0.20	-	-
(Tian et al., 2020)	ResNet-12	73.84±0.86	86.67±0.67	43.71±0.77	60.81±0.84
The present approach	ResNet-12	**74.64±0.85**	**87.63±0.59**	**45.04±0.71**	**60.89±0.80**

5.2 Ablation Study

The proposed approach, along with ResNet-12 as the baseline, consists of similarity matching and generalized self-knowledge distillation functions. An ablation study was conducted to further evaluate the affected performance of the proposed approach. The ablation study experiments are presented in Table 4. Each experiment was evaluated separately, according to the different few-shot settings.

As seen in Table 4, the proposed approach achieves the best performance among all the prior works based on the excellent baseline model. Furthermore, the proposed self-knowledge distillation method can further improve the recognition accuracy of the model for few-shot learning based on the experimental results.

Specifically, the few-shot learning with the self-knowledge distillation method can effectively improve identification accuracy. With original KL-divergence and similarity matching loss functions, the performance of all the student models has surpassed the performance of the corresponding teacher models with the same few-shot setting. The relationship between features is used to compensate for the feature relationships between intra-batch and improve the model's generalization performance for unseen tasks.

Secondly, the corrected KL-divergence loss function is employed to replace the original KL-divergence. With the updated loss function, the overall performance of the proposed methods is further improved for all the datasets except the FC100 dataset. The performance degradation problem of the 5-shot setting in tieredImageNet could be caused by the fact that the student model is somehow overfitting the training dataset. In addition, the student model can correct some inherent inaccurate predictions based on the corrected transferrable guidance from the teacher model.

Moreover, the model performance gain at the 1-shot setting is more significant than that at the 5-shot setting for all comparative experiments within the same dataset. The different shot settings for the same dataset differentiate between the number of support samples and query samples. From all the comparison experiments, more support samples contribute to better performance. At the same time, in addition to the simple feature mapping, the similarity relationships between features further introduce the utilization of the existing features to a certain extent.

6. Conclusion and Future Work

This paper proposes a more efficient meta-learning framework based on generalized knowledge distillation and similarity matching. This is a general few-shot classification algorithm, which can be further applied to other cases. The proposed method provides a theoretical basis for the present subsequent works on feature relationship reuse. The present approach benefits from transferrable knowledge via a well-trained model based on the self-knowledge distillation method. The traditional self-knowledge distillation method transfers the prediction distribution while neglecting to transfer the feature relationships within one batch. The feature similarity relationships within the batch compensate for the distribution matching in the self-knowledge distillation period. In addition, the self-knowledge distillation algorithm may transfer the inherent error prediction from the teacher model to the student model. A generalized self-knowledge distillation function was utilized to modify the inaccurate prediction of the teacher model. The final experimental results have shown that the proposed loss functions are compatible with each other in the few-shot image

Table 4. Ablation study of the proposed approach on benchmark datasets

Model	miniImageNet 5-way		tieredImageNet 5-way		CIFAR-FS 5-way		FC100 5-way	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Baseline	61.56±0.83	79.27±0.63	69.53±0.90	85.08±0.65	71.24±0.99	85.63±0.63	41.65±0.76	58.17±0.76
Distill+Sim	64.33±0.72	80.70±0.60	71.41±0.89	85.62±0.73	74.52±0.76	87.47±0.60	44.42±0.69	**60.89±0.80**
Distill+Sim+KLc	**65.39±0.73**	**81.51±0.64**	**71.43±0.94**	**86.13±0.66**	**74.64±0.85**	**87.63±0.59**	**45.04±0.71**	60.78±0.76

classification. The present approach performs better than other algorithms on the commonly used benchmark datasets. In future work, more in-depth research on feature extraction will be conducted and local metric information will be combined with global feature distribution to establish better feature space. Leveraging dynamic feature extraction and adjusting the features from different distributions to fit the few-shot learning setting for feature extraction will be considered. The global feature distribution can provide confidence for the current instance and capture the inductive bias for each category. The model generalization ability can be further improved by utilizing the difference in feature distribution.

REFERENCES

Bai, H., Wu, J., King, I. & Lyu, M. R. (2020). Few Shot Network Compression via Cross Distillation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence* (pp. 3203-3210). AAAI.

Bertinetto, L., Henriques, J. F., Torr, P. H. & Vedaldi, A. (2019). Meta-learning with differentiable closed-form solvers. In *7th International Conference on Learning Representations (ICLR)*, (pp. 1-15).

Bu, Y., Zou, S., Liang, Y. & Veeravalli, V. V. (2016). Estimation of KL divergence between large-alphabet distributions, *IEEE Transactions on Information Theory*, *64*(4), 2648-2674.

Dhillon, G. S., Chaudhari, P., Ravichandran, A. & Soatto, S. (2020). A Baseline for Few-Shot Image Classification. In *8th International Conference on Learning Representations (ICLR),* (pp. 1-20).

Ding, P., Jia, M. & Zhao, X. (2021). Meta deep learning based rotating machinery health prognostics toward few-shot prognostics, *Applied Soft Computing*, *104*, 107211-107230.

Feng, S. & Duarte, M. F. (2019). Few-shot learning-based human activity recognition, *Expert Systems with Applications*, *138*, 112782-112793.

Finn, C., Abbeel, P. & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, (pp. 1126-1135).

Fu, S., Li, Z., Liu, Z. & Yang, X. (2021). Interactive Knowledge Distillation for image classification, *Neurocomputing*, *449*, 411-421.

Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P. & Cord, M. (2019). Boosting Few-Shot Visual Learning With Self-Supervision. In *2019 IEEE/CVF International Conference on Computer Vision* (pp. 8059-8068). IEEE.

Gou, J., Yu, B., Maybank, S. J. & Tao, D. (2021). Knowledge Distillation: A Survey, *International Journal of Computer Vision*, *129*(6), 1789-1819.

Hu, Z., Wu, D., Nie, F. & Wang, R. (2021). Generalization bottleneck in deep metric learning, *Information Sciences*, *581*, 249-261.

Huang, S. & Wu, Q. (2021). Robust pairwise learning with Huber loss, *Journal of Complexity*, *66*, 101570-101583.

Lee, K., Maji, S., Ravichandran, A. & Soatto, S. (2019). Meta-Learning With Differentiable Convex Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10657-10665). IEEE.

Lei, D., Jiang, Z. & Wu, Y. (2019). Weighted Huber constrained sparse face recognition, *Neural Computing and Applications*, *32*(9), 5235-5253.

Li, L., Li, Z., Liu, Y. & Hong, Q. (2021). Deep joint learning for language recognition, *Neural Networks*, *141*, 72-86.

Li, T., Li, J., Liu, Z. & Zhang, C. (2020). Few Sample Knowledge Distillation for Efficient Network Compression. In *2020 Conference on Computer Vision and Pattern Recognition, (CVPR)*, (pp. 14627-14635). IEEE.

Liang, D., Gao, X., Lu, W. & He, L. (2020). Deep multi-label learning for image distortion identification, *Signal Processing*, *172*, 107536-107549.

Mishra, N., Rohaninejad, M., Chen, X. & Abbeel, P. (2018). A Simple Neural Attentive Meta-Learner. In *6th International Conference on Learning Representations. (ICLR)*, (pp. 1-17)

Munkhdalai, T. & Yu, H. (2017). Meta Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, (pp. 2554-2563).

Oreshkin, B. N., López, P. R. & Lacoste, A. (2018). TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS),* (pp. 719-729)

Qiao, L., Shi, Y., Li, J., Wang, Y., Huang, T. & Tian, Y. (2019). Transductive Episodic-Wise Adaptive

Metric for Few-Shot Learning. In *2019 IEEE/CVF International Conference on Computer Vision* (pp. 3603-3612). IEEE.

Ravi, S. & Larochelle, H. (2017). Optimization as a Model for Few-Shot Learning. In *5th International Conference on Learning Representations (ICLR),* (pp. 1-11)

Ravichandran, A., Bhotika, R. & Soatto, S. (2019). Few-Shot Learning With Embedded Class Models and Shot-Free Meta Training. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (pp. 331-339). IEEE.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H. & Zemel, R. S. (2018). Meta-Learning for Semi-Supervised Few-Shot Classification. In *6th International Conference on Learning Representations (ICLR)*, (pp. 1-15).

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S. & Hadsell, R. (2019). Meta-Learning with Latent Embedding Optimization. In *7th International Conference on Learning Representations (ICLR)*, (pp. 1-17).

Shorfuzzaman, M. & Hossain, M. S. (2020). MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients, *Pattern Recognition*, *113*, 107700-107710.

Snell, J., Swersky, K. & Zemel, R. S. (2017). Prototypical Networks for Few-shot Learning. In *31st International Conference on Neural Information Processing Systems (NIPS)*, (pp. 4080–4090).

Tian, Y. & Fu, S. (2020). A descriptive framework for the field of deep learning applications in medical images, *Knowledge-Based Systems*, *210*, 106445-106466.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B. & Isola, P. (2020). Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need?. In *European Conference on Computer Vision (ECCV)*, (pp. 266-282).

Vinyals, O., Blundell, C., Lillicrap, T. P., Kavukcuoglu, K. & Wierstra, D. (2016). Matching Networks for One Shot Learning, *Advances in Neural Information Processing Systems*, *29*, 3630-3638.

Wen, T., Lai, S. & Qian, X. (2021). Preparing Lessons: Improve Knowledge Distillation with Better Supervision, *Neurocomputing*, *454*, 25-33.

Yoon, S., Seo, J. & Moon, J. (2019). TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, (pp. 7115-7123).