Journal of Image Processing Theory and Applications (2024)
Clausius Scientific Press, Canada
DOI: 10.23977/jipta.2024.070115
ISSN 2560-6239 Vol. 7 Num. 1

Few-shot Image Classification Model Based on Improved Prototype

Ya Gao[1,*], Chunhui Liang[1], Jianye An[1]

[1]School of Science, Tianjin University of Commerce, Tianjin, 300134, China
[*]Corresponding author: gynice666@163.com

Keywords: few-shot image classification, meta-learning, muti-head attentional mechanism, pseudo-labeled sample expansion

Abstract: Aiming at the limitations of important feature extraction in few-shot image classification and the problem of poor representation of category prototypes in the prototype network, this paper proposes a few-shot image classification model based on improved prototype. The model adds a global grouping multi-attention mechanism to the original backbone network in the feature extraction module, which enhances the extraction of important features. For the prototype improvement module, the pseudo-labeled samples in the query set with cosine similarity scores higher than a threshold are utilized to expand the support set. Finally, applying our model to the typical few-shot image datasets miniImagenet and tieredImagenet, the experiments show that compared with other meta-learning models, the model built in this paper achieves better classification performance.

1. Introduction

In recent years, deep learning has made great breakthroughs in the field of image classification, however, its success relies on large-scale data and high-performance computing, which poses a high cost problem and limits its application in many scenarios. For example, medical data for rare diseases are difficult to obtain, and collecting labeled data for such diseases is even more difficult. In contrast, few-shot learning [1-2] can quickly grasp and generalize new concepts using only a small number of samples per category. By leveraging learned transferable knowledge, it requires only a few labeled samples in new task scenarios to achieve strong performance, particularly in image classification [3-4]. This approach effectively addresses the limitations of traditional deep learning methods, which struggle when labeled samples are scarce.

Few-shot learning is one of the important applications of Meta-Learning in the field of supervised learning. The research of Meta-Learning mainly focuses on three aspects: model-based meta-learning methods, optimization-based meta-learning methods, and metric-based meta-learning methods.

Model-based meta-learning approaches focus on the improvement of model construction, unlike the traditional RNN [5] (Recurrent neural network) in which the internal memory mechanism can only temporarily store learned knowledge, such as the MANN [6] (Memory-Augmented Neural Networks) proposed by Santoro et al. invokes an external memory mechanism by reading and writing memory slots to store and retrieve information, which enables the model to adapt quickly when facing a new task. MM-Net [7] (Memory Matching Networks), a few-shot classification algorithm based on external memory proposed by Cai et al. is based on the idea of matching networks, which firstly utilizes an encoder to extract the support set features and puts them into the memory module, and then employs the bi-LSTM network to combine the support set feature information from the memory module to extract the query set features, and finally calculate the similarity between the support set and query set features to complete the classification.

Optimization-based meta-learning methods focus on the improvement of the optimization process, and the core idea is to set up multiple meta-learning tasks in the training process, and use gradient updating algorithm to estimate parameters, such as MAML [8] (Model-agnostic meta-learning).

Metric-based meta-learning methods focus on the improvement of metric criterion, such as prototypical network [9], which aims to find appropriately metric to measure the similarity, assuming that images of each category are gathered in a certain spatial region, and extracting the support set and query set features through neural network. Further, the mean value of the image features of each category in the support set denote as the prototype of the category, and then calculates the Euclidean distance between the image features of the query set and the prototypes of each category, so as to determine the category of the images in the query set.

Currently, in few-shot classification tasks, most of the research methods use optimization-based meta-learning methods [8] and metric-based meta-learning methods [9]. Among these approaches, optimization-based meta-learning methods are particularly demanding in terms of calculate resources, especially during the weight update process. This high computational demand stems from optimizer selection and step size adjustments within gradient descent optimization, which makes the model slower when performing the learning of new tasks. In addition, the weight updating process tends to overfit with limited sample data. To overcome these challenges, current research primarily utilizes metric-based meta-learning methods, but the existing metric-based meta-learning methods are not ideal in terms of classification performance, have limitations in capturing feature information on feature extraction, and cannot simultaneously utilize the multi-dimensional global information of the images in the support set and query set. In the metric-based meta-learning approach, the prototype network uses the encoder to extract features from the support set and computes the mean to represent each prototype. However, the prototypes obtained after this simple and direct processing contain less information. Additionally, due to the randomness of the sampling of the images when constructing a support set for the few-shot classification task, the resulting prototypes may lack representativeness, leading to deviations during prediction and reduced classification accuracy for the query set.

To address the above issues, this paper proposes a few-shot classification model that enhances the prototype using an improved attention mechanism and pseudo-labeled data augmentation:

(1) The feature extraction module incorporates the global group multi-head attention module (GGMA) on the basis of the Resnet-12 backbone network. This allows for global average pooling and maximum pooling along the height and width dimensions, respectively, to capture multi-dimensional global information and enhance the comprehensiveness of feature extraction.

(2) The prototype improvement module adopts the iterative updating idea from the K-means clustering algorithm. It uses pseudo-labeled data to expand the support set. Specifically, query set samples with a similarity greater than a defined threshold to the initial prototype are added to the support set. This expanded support set is then used to calculate a new prototype representation. Through continuous iterative updates, the prototype becomes increasingly representative.

2. Methodology

2.1 Definition of the problem

In few-shot learning, assume that we have an image set $X=\{x_i\}_{i=1}^{U}$, where $x_i \in R^E$ represents the ith image, and let $Y=\{y_i\}_{i=1}^{U}$ be the label set, where y_i represents the true category label of the ith image. We divide the dataset $D=\{(x_i, y_i), i=1,\cdots,U\}$ into a meta-training set D_{train} , a meta-validation set D_{val} , and a meta-testing set D_{test} , and the categories of these datasets are not intersected with each other. The meta-validation set is used to adjust the trained hyper-parameters to improve the generalization ability of the model, and the meta-testing set is used to validate the classification effect of the model on new classes. The support set S and query set Q are constructed on the basis of the meta-training set D_{train} by using the N-way K-shot method, which employs sampling without replacement to select K images from N categories. Generally, K is usually taken as 1 or 5, which is used to simulate different test scenarios. Further, L images are extracted from the remaining data of each of the N categories, totaling M=N×L images into the query set. The support set is denoted as $S=\{(x_i, y_i)\}_{i=1}^{N\times K}$, and the query set is denoted as $Q=\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{M}$, further, the support set S and the query set Q form a eposide. After repeating the above process to generate multiple eposides. We start training the model, with the goal of using the knowledge learned from the N×K images containing samples from the support set to classify the M query set images. As shown in Fig 1, a schematic of some of the support set and query set images in the dataset is demonstrated.

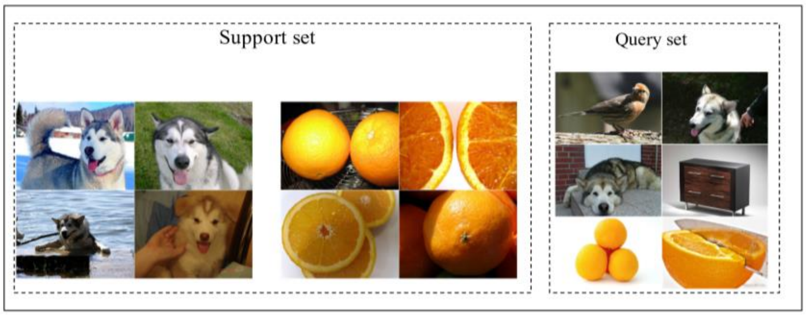

Figure 1: Schematic of some of the support set and query set images in the dataset

2.2 Feature Extraction Module

Traditional convolutional neural networks lack the ability to capture global information in both spatial dimensions of height and width simultaneously when processing complex visual tasks, resulting in limited feature representation. A single attention mechanism (e.g., channel attention or spatial attention) cannot simultaneously utilize the global information in multiple dimensions of the support set and query set images when capturing feature information. To address the above problems, we propose the Global Group Multi-head Attention Module (GGMA). The GGMA module captures the global information in multiple dimensions of the support set and query set images by performing global average pooling and maximum pooling in the height and width directions, respectively, to enhance the feature representation. Images with multi-dimensional global information to enhance the comprehensiveness of feature extraction. Meanwhile, the GGMA module combines the multi-dimensional global information and the attention mechanism to generate the attention maps by using the global information of the feature maps in the spatial dimensions (height and width), and weight the input feature maps by these attention maps to enhance the feature expression capability, which can significantly improve the performance of the few-shot classification task model. The specific structure of this module is shown in the following Fig.2:

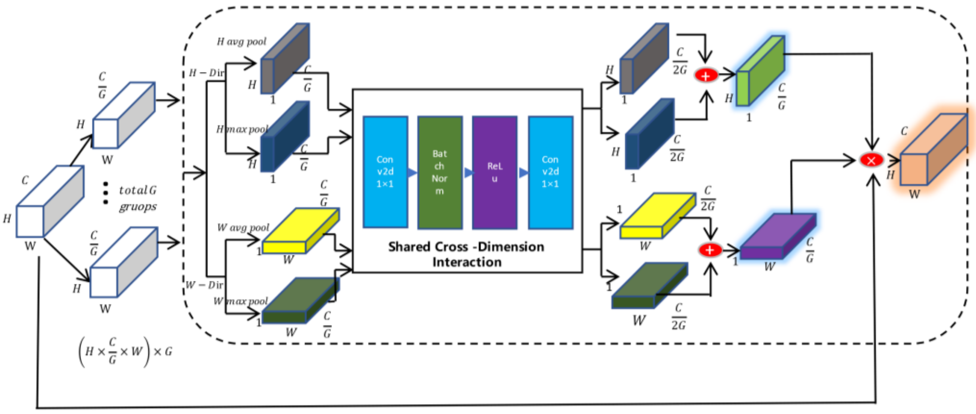

Figure 2: Overall schematic of the GGMA module

First, for the input feature map $T \in R^{B \times C \times H \times W}$, where B is the batch size, C is the number of channels, and H and W are the height and width of the feature map respectively. We group them into G groups according to the number of channels, and each group contains C/G channels. The grouped feature maps are represented as: $\tilde{T}_g \in R^{B \times \frac{C}{G} \times H \times W}, g = 1, \cdots, G$.

Then, we perform global average pooling and global maximum pooling operations on the grouped feature maps in the height direction and width direction, respectively:

$$T_{h,avg} = AvgPool(\tilde{T}) \in R^{B \times C \times H \times 1}, T_{h,\max} = MaxPool(\tilde{T}) \in R^{B \times C \times H \times 1} \tag{1}$$

$$T_{w,avg} = AvgPool(\tilde{T}) \in R^{B \times C \times 1 \times W}, T_{w,\max} = MaxPool(\tilde{T}) \in R^{B \times C \times 1 \times W} \tag{2}$$

AvgPool represents the process of globally average pooling the grouping tensor $\tilde{T}_g$ in both the height and width directions to obtain $\tilde{T}^g_{h,avg}$ and $\tilde{T}^g_{w,avg}$ respectively. Then, the G grouped tensors that have undergone global average pooling are merged in the channel direction to obtain $T_{h,avg}$ and $T_{w,avg}$. MaxPool represents the process of globally max pooling the grouping tensor $\tilde{T}_g$ in both the height and width directions to obtain $\tilde{T}^g_{h,\max}$ and $\tilde{T}^g_{w,\max}$, respectively. Then, the G groups of tensors that have undergone global max pooling are merged in the channel direction to obtain $T_{h,\max}$ and $T_{w,\max}$.

For each grouped feature map, we apply a shared convolutional layer for feature processing. This shared convolutional layer consists of two 1×1 convolutional layers, a batch normalization layer, and a Relu function layer for reducing and recovering the channel dimensions:

$$V_{h,avg} = Conv(T_{h,avg}) \in R^{B \times \frac{C}{2} \times H \times 1}, V_{h,\max} = Conv(T_{h,\max}) \in R^{B \times \frac{C}{2} \times H \times 1} \tag{3}$$

$$V_{w,avg} = Conv(T_{w,avg}) \in R^{B \times \frac{C}{2} \times 1 \times W}, V_{w,\max} = Conv(T_{w,\max}) \in R^{B \times \frac{C}{2} \times 1 \times W} \tag{4}$$

Attention weights in the height direction and width direction are generated by summing the outputs of the convolutional layers and applying a Sigmoid activation function:

$$A_h = Sigmoid[Concat(V_{h,avg}, V_{h,\max})] \in R^{B\times C\times H\times 1} \tag{5}$$

$$A_w = Sigmoid[Concat(V_{w,avg}, V_{w,\max})] \in R^{B\times C\times 1\times W} \tag{6}$$

Where concat denotes the merging of multiple tensors in the channel dimension.

Finally, we weight the input feature maps according to the attention weights to get the output feature maps.

$$O = T \times A_h \times A_w \in R^{B\times C\times H\times W} \tag{7}$$

In formula (7), the attention weights A_h and A_w are expanded in the height and width directions, respectively, to match the dimensions of the input feature map.

In summary, the GGMA module generates the attention map in the height and width directions by sharing the convolutional layer and the attention mechanism, weights the input feature map, enhances the expression of the important features and suppresses the unimportant features, so as to reduce the impact of facing the different locations of the same type of objects in different images in the dataset or the interference of the background and noise, and to make the subsequently extracted prototypes more representative.

The network structure of the feature extraction module is shown in the following Fig.3:

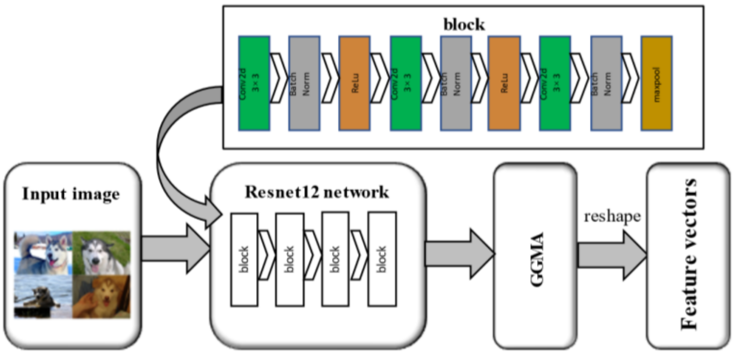

Figure 3: Schematic diagram of the overall architecture of the feature extraction module

2.3 Prototype Improvement Module

Based on the original baseline method prototype network, cosine similarity based prototypical network (CSPN) [17] is constructed by replacing the cosine similarity with the original Euclidean distance. However, when constructing the support set in few-shot classification tasks, the randomness of sampling the images makes the prototypes computed by the prototype network less representative, so we use the pseudo-labeled samples to expand the support set strategy to improve the prototypes.

Based on the idea of iterative updating of K-means clustering algorithm[18], we propose cosine similarity with expanding samples based prototypical network (CESPN). The detailed process of CESPN algorithm is as follows:

In each category of the support set, the feature mean of K images is selected as the initial prototype. Based on the initial prototype, the cosine similarity between the features of each image in the query set and the initial prototype is calculated, i.e., the pseudo-label is given to the query set images, and the decision of whether to enter the image into the support set is based on the magnitude of the cosine similarity in relation to the threshold value ξ. If it is greater than or equal to this threshold, the image enters the support set, and the prototype of the expanded support set is calculated again. The above process is repeated until all images in the query set are traversed. The proposed CESPN algorithm is shown here:

CESPN algorithm
Initialize: $c_n = \frac{1}{K} \sum_{x_i \in S, y_i = n} f_\theta(x_i), n = 1, \cdots, N$
for $n \leftarrow 1$ to N do:
for $i \leftarrow 1$ to M do:
If $\cos(c_n, \tilde{h}_i) \geq \xi$ Then $(\tilde{x}_i, \tilde{y}_i) \in S$
$S' \leftarrow S \bigcup \{(\tilde{x}_i, \tilde{y}_i)\}$
$K^{'} = K + 1$
end for
$c_n^{'} = \frac{1}{K^{'}} \sum_{x_i \in S', y_i = n} f_\theta(x_i), n = 1, \cdots, N$
end for
until the query set is traversed
return $c_n^{'}$

It is experimentally verified that the classification effect is optimal when the threshold value ξ is 0.7, at which time the strategy of extending the support set by pseudo-labeled samples in CESPN can effectively enhance the representativeness of the prototypes, so as to classify the samples of the query set more accurately and enhance the robustness of the model.

3. Experiments.

In order to verify the effectiveness of the method in this paper, the experiments use two standard few-shot learning datasets, mini-Imagenet and tiered-Imagenet, for validation and comparative analysis with the existing model. Meanwhile, this paper adopts ablation experiments to validate the effect of feature extraction module and prototype improvement module on the classification effect of the model.

3.1 Datasets.

MiniImagenet is a subset constructed from the ILSVRC-2012 [19] dataset, which contains 100 classes, where each class has 600 samples, for a total of 60,000 color images. ravi et al. divided the dataset into a training set, a validation set, and a test set, which contain 64, 16, and 20 classes, respectively, and the classes do not intersect each other.

The tieredImagenet is also a subset constructed from the ILSVRC-2012 dataset, which contains 34 major classes, each of which contains 10-30 minor classes, totaling 608 classes, each of which contains a different number of samples. Among them, 351 subclasses corresponding to 20 major classes are classified as training set, 97 subclasses corresponding to 6 major classes are classified as validation set, and 160 subclasses corresponding to the remaining 8 major classes are classified as testing set, and the classes are not intersected with each other.

3.2 Experimental settings

All experiments in this paper were conducted in a deep learning environment with Windows, GPU: RTX 2080Ti (8G) $\times 2$, and PyTorch (1.0.1). For the experiments, 200 training epochs were set up for both the training and testing phases according to the 5-way 1-shot and 5-way 5-shot meta-tasks.

3.3 Results

Comparative experiments were set up on miniImagenet and tieredImagenet datasets with 5-way 1-shot and 5-way 5-shot, respectively. We chose 10 methods to implement on miniImagenet dataset: MM-Net[7], MAML[8], Prototypical Net[9], RelationNet[10], TPN[11], LEO[12], MetaOptNet[13], MCRNet[14], DAM[15], HGNN[16], and 8 methods were selected on the tieredImagenet dataset: MAML, Prototypical Net, RelationNet, TPN, LEO, MetaOptNet, DAM, HGNN. To evaluate the average accuracy of the models on 2000 tasks (episodes) within 95% confidence interval, and the results of each on the test set are shown in the following table 1 and table 2:

Table 1: Experimental results of each model on miniImagenet

model	5-way 1-shot	5-way 5-shot
MM-Net[7]	53.37 ±0.48%	66.97 ±0.35%
MAML[8]	48.70 ±1.84%	63.11 ±0.92%
Prototypical Net[9]	49.42 ±0.78%	68.20 ±0.66%
RelationNet[10]	50.44 ±0.82%	65.32 ±0.70%
TPN[11]	55.51 ±0.86%	69.86 ±0.65%
LEO[12]	61.76 ±0.08%	77.59 ±0.12%
MetaOptNet[13]	62.64 ±0.61%	78.63 ±0.46%
MCRNet[14]	62.53 ±0.64%	80.34 ±0.47%
DAM[15]	60.39 ±0.21%	73.84 ±0.16%
HGNN[16]	60.03 ±0.51%	79.64 ±0.36%
CESPN+GGMA(ours)	**65.82 ±0.63%**	**83.52 ±0.34%**

Table 2: Experimental results of each model on tieredImagenet

model	5-way 1-shot	5-way 5-shot
MAML[8]	53.11 ±0.92%	68.85 ±0.66%
Prototypical Net[9]	53.31 ±0.89%	72.69 ±0.74%
Relation Net[10]	55.20 ±0.70%	71.20 ±0.73%
TPN[11]	59.91 ±0.94%	73.30 ±0.75%
LEO[12]	66.33 ±0.05%	81.44 ±0.09%
MetaOptNet[13]	65.81 ±0.74%	81.75 ±0.53%
DAM[15]	64.09 ±0.23%	78.39 ±0.18%
HGNN[16]	64.32 ±0.49%	83.34 ±0.45%
CESPN+GGMA(ours)	**69.93 ±0.66%**	**85.58 ±0.78%**

MM-Net, MAML, LEO, Relation Net, and Prototypical Net are representative in model-based, optimization, and metrics-based meta-learning approaches, respectively. TPN performs transitive inference on the query set, using a graph construction module that iteratively propagates labels from the support set to the query set. MetaOptNet takes advantage of the nice properties of linear classifiers and uses high-dimensional embeddings to improve model generalization. MCRNet encodes additional information representations and thus reconstructs a latent space to complement the information deficiencies of meta-learning. DAM constructs multimodal weight distributions via a self-encoder, generates a new metric using three layers of deep attention, uses the new metric to construct a loss function for training, which is an improvement of metric meta-learning methods. HGNN performs hierarchical learning on intra- and inter-class nodes, fuses multilevel features to categorize nodes, and solves the problem of ignoring the hierarchical correlation between nodes when reasoning directly about nodes, which is an improvement of graph neural networks in meta-learning methods. In the miniImagenet dataset, the model in this paper improves the effect of the model by an average of 12.3% and 13.2% compared with the accuracy of other models in the setting scenarios of 5-way 1-shot and 5-way 5-shot, and improves the effect of the model by 3.1% and 3.2% compared with the highest accuracy of other models. While in the tieredImagenet dataset, the effect of the model in the 5-way 1-shot and 5-way 5-shot setup scenarios improved by 9.7% and 9.2% on average compared to the accuracy of the other models. Compared to the highest accuracy of the other models, the model's effectiveness improved by 3.6% and 2.1%.

3.4 Ablation studies

In order to further explore the specific impact of each module on the experimental results, this paper sets up 5-way-1-shot and 5-way-5-shot on miniImagenet and tieredImagenet datasets to conduct ablation experiments, respectively, including the impact of different modules on the data accuracy and the impact of the setting of the similarity threshold on the accuracy.

Table 3: Ablation experiments on miniImagenet and tieredImagenet datasets

method	miniImagenet		tieredImagenet	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
CSPN	52.55±0.79%	71.64±0.60%	58.20±0.87%	75.31±0.65%
CESPN	63.93±0.93%	81.89±0.56%	67.74±0.95%	83.92±0.63%
CSPN+ GGMA	63.18±0.69%	80.86±0.59%	65.76±0.31%	81.92±0.43%
CESPN+GGMA	65.82±0.63%	83.52±0.34%	69.93±0.66%	85.58±0.78%

From the results in the table 3, compared to the benchmark model CSPN, both the GGMA module and the improved CESPN model show an 8% accuracy improvement. When the GGMA module is combined with the improved CESPN model (CESPN+GGMA), the highest accuracy is achieved, with an overall improvement of approximately 10% compared to the initial benchmark model CSPN.

4. Conclusions

In this paper, we propose an improved prototype-based few-shot classification model, which integrates a global grouped multi-head attention mechanism with a pseudo-labeled sample expansion strategy for the support set. The proposed model combines the multi-dimensional global information and the attention mechanism by generating attention maps from the global feature map in the spatial domain. These attention maps are then used to weight the input feature maps, which can improve the feature expression ability. Additionally, we introduce a support set expansion technique that iteratively updates the K-means clustering algorithm to compare the similarity scores of pseudo-labeled samples with predefined similarity thresholds. This approach enables the computationally obtained prototype to better capture representative characteristics, resulting in improved classification accuracy on the miniImagenet and tieredImagenet datasets, compared to other meta-learning models. Further, ablation experiments illustrate that the feature extraction module and the prototype improvement module significantly enhance the performance of the model in few-shot image classification tasks.

References

[1] Zhou B.J, Chen C.Y. A review of research on few-shot image classification based on deep meta-learning [J]. Computer Engineering and Applications, 2024, 60(08):1-15.

[2] Song Y, Wang T, Cai P, et al. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities [J]. ACM Computing Surveys, 2023, 55(13s): 1-40.

[3] Hu Y , Pateux S , Gripon V .Squeezing Backbone Feature Distributions to the Max for Efficient Few-Shot Learning[J]. algorithms, 2022, 15(5).
[4] Liu Y, Zhang H, Yang Y. Few-Shot Image Classification Based on Asymmetric Convolution and Attention Mechanism[C/OL]//2022 4th International Conference on Natural Language Processing (ICNLP), Xi'an, China. 2022:217-222
[5] Mikolov T, Martin Karafiát, Burget L ,et al. Recurrent neural network based language model[C]//Interspeech, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September.DBLP, 2015.
[6] Sergey B, Adam S, Matthew B,et al. Meta-Learning with Memory-Augmented Neural Networks[J].Journal of Machine Learning Research, 2016.
[7] Cai Q, Pan Y, Yao T,et al. Memory Matching Networks for One-Shot Image Recognition[J].IEEE, 2018:4080-4088.
[8] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// International conference on machine learning. PMLR, 2017: 1126-1135.
[9] Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning[J]. Advances in neural information processing systems, 2017, 30.
[10] Sung F, Yang Y, Zhang L, et al. Learning to compare: relation network for few-shot learning[C]//Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition(CVPR'18), 2018: 1199-1208.
[11] Liu Y, Lee J, Park M, et al. Transductive propagation network for few-shot learning [J]. arXiv preprint arXiv:1805.10002, 2018.
[12] Rusu, Andrei A A, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization[J]. arXiv preprint arXiv:1807.05960, 2018.
[13] Lee K, Maji S, Ravichandran A, et al. Meta-learning with differentiable convex optimization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2019:10649-10657.
[14] Zhong X, Gu C, Huang W,et al. Complementing Representation Deficiency in Few-shot Image Classification: A Meta-Learning Approach[C]//International Conference on Pattern Recognition. IEEE, 2021: 2677-2684.
[15] Zhou F, Zhang L, Wei W. Meta-generating deep attentive metric for few-shot classification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022: 6863-6873.
[16] Chen C , Li K , Wei W ,et al. Hierarchical Graph Neural Networks for Few-Shot Learning[J].IEEE Transactions on Circuits and Systems for Video Technology, 2021:240-252.
[17] Liu J, Song L, Qin Y. Prototype Rectification for Few-Shot Learning[C]//European Conference on Computer Vision.Springer, Cham, 2020:741-756.
[18] Ahmed M, Seraj R, Islam S M S .The k-means Algorithm: A Comprehensive Survey and Performance Evaluation[J]. Electronics, 2020, 9(8):1295.
[19] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115: 211-252.