doi:10.3772/j.issn.1006-6748.2024.03.007

Few-shot image recognition based on multi-scale features prototypical network①

LIU Jiatong(刘珈彤), DUAN Yong②
(School of Information Science Engineering, Shenyang University of Technology, Shenyang 110870, P. R. China)
(Shenyang Key Laboratory of Advanced Computing and Application Innovation, Shenyang 110870, P. R. China)

Abstract

In order to improve the model's capability in expressing features during few-shot learning, a multi-scale features prototypical network (MS-PN) algorithm is proposed. The metric learning algorithm is employed to extract image features and project them into a feature space, thus evaluating the similarity between samples based on their relative distances within the metric space. To sufficiently extract feature information from limited sample data and mitigate the impact of constrained data volume, a multi-scale feature extraction network is presented to capture data features at various scales during the process of image feature extraction. Additionally, the position of the prototype is fine-tuned by assigning weights to data points to mitigate the influence of outliers on the experiment. The loss function integrates contrastive loss and label-smoothing to bring similar data points closer and separate dissimilar data points within the metric space. Experimental evaluations are conducted on small-sample datasets mini-ImageNet and CUB200-2011. The method in this paper can achieve higher classification accuracy. Specifically, in the 5-way 1-shot experiment, classification accuracy reaches 50.13% and 66.79% respectively on these two datasets. Moreover, in the 5-way 5-shot experiment, accuracy of 66.79% and 85.91% are observed, respectively.

Key words: few-shot learning, multi-scale feature, prototypical network, channel attention, label-smoothing

0 Introduction

In recent years, significant progress has been achieved in the field of image recognition due to the ongoing evolution of deep learning theories and methodologies[1-5]. The effectiveness of image recognition has been notably boosted by the distinctive convolutional layers, pooling structures, and algorithms of convolutional neural networks, which enable the establishment of connections between pixels and the extraction of high-dimensional feature information from images. The growing demand for enhanced accuracy in image recognition has correspondingly led to increased neural network model depth. Consequently, the requirement for annotated data to train such deep neural networks has surged. However, the process of acquiring and annotating extensive data demand Substantial human and material resources, while ensuring the precision of manual identification remains a challenge. Moreover, in certain domains, machine learning recognition accuracy has surpassed human-annotated accuracy. Even with considerable human effort dedicated to image annotation, ensuring the quality of deep learning remains a challenging task. In response to this scenario, few-shot learning[6-7] has emerged as a viable solution. It refers to the capability to classify previously unseen data with minimal samples, relying on past experiences. Traditional deep learning often exhibits subpar performance in such scenarios. In the face of limited or unseen data types, few-shot learning has the capability to transcend the constraints posed by the aforementioned conditions. Consequently, few-shot learning has become an important research direction in the field of image recognition in recent years[8-10].

Common approaches in few-shot learning encompass data augmentation-based, optimization-based, and metric-based techniques. Among these, the metric-based approach has received attention due to its simplicity and promising performance in few-shot image

① Supported by the Scientific Research Foundation of Liaoning Provincial Department of Education (No. LJKZ0139) and the Program for Liaoning Excellent Talents in University (No. LR15045).
② To whom correspondence should be addressed. E-mail: duanyong0607@163.com.
Received on Dec. 14, 2023

classification tasks. As referenced in Ref. [11], this method allows categories to create one or multiple clusters in the metric space, thereby providing a rational distribution of experimental data by interpolating between nearest neighbors and prototype representations based on the complexity of the input data. Ref. [12] assigned weighted values to sample points in the metric space and calculated their influence based on sample embeddings in the feature space. While both methods optimize sample feature embeddings and bring similar category data points closer in the metric space, they encounter difficulties in separating different category data points. In contrast, the method proposed in the Ref. [13] emphasizes the oversight of inherent semantic information in the task itself while solely focusing on relative spatial distances between sample points. This method guides the feature extraction network by introducing an emotion-assisted classification task, thereby enhancing feature extraction. Semantic classification improves accuracy, and metric classification reinforces generalization. They ensure overall precision. Despite leveraging semantic information to improve accuracy, this method can still be enhanced by exploring the intrinsic features within the sample data during the extraction process.

Therefore, for a comprehensive extraction and utilization of image features in few-shot learning, as well as an improved measurement of intra-class and inter-class samples in the metric space, this paper introduces the multi-scale prototypical network (MS-PN) algorithm. This algorithm employs a multi-scale feature input network for feature extraction, combining features at a finer granularity to comprehensively extract image features. Furthermore, a channel attention module is introduced in the backbone network to weight and strengthen critical information among channels. It can further explore key information between channels.

To address the challenge of few-shot learning where outlier points have a substantial influence on the network, this paper proposes a weighted treatment for sample points. This method is based on the summation of distances from other samples within the same class, aiming to diminish the impact of outlier point during prototype determination. The loss function utilizes contrastive loss to bring similar-class images closer while separating distinct-class images. Additionally, the label-smoothed cross-entropy loss is applied to reduce the impact arising from potential network misjudgments or positive samples being misclassified as negative due to the scarcity of samples.

1 Few-shot image classification tasks

The task of few-shot learning is divided into support set, query set, and test set. The network model can learn sufficient experience from the support set and validate this acquired knowledge through the query set. The test set contains completely different categories from the above two sets. Few-shot learning model relies on accumulating experience from trained tasks, and enabling optimal classification results in new tasks with just one or a few steps.

In the course of training, for each task, the network selects data of n categories from the dataset. From each category, it then randomly chooses k instances to construct an n-way K-shot task, which serves as the support set for input into the neural network.

In a n-way k-shot task, the support set S consists of $n \times k$ data points as shown in Eq. (1), where y_i represents the label corresponding to input data x_i.

$$S = \{(x_i, y_i)\}_{i=1}^{k} \tag{1}$$

Then, select b data points from the remaining data of n categories to form a query set Q, consisting of $n \times b$ data points, as shown in Eq. (2).

$$Q = \{(x_j, y_j)\}_{j=1}^{b} \tag{2}$$

The prototypical network[14] is a kind of few-shot learning method based on metric learning[15]. It encodes the support set into a feature space and establishes prototypes for each category based on the positions of same-category samples. Within the query set, a class label of the query sample is assigned by assessing its distance to the feature prototypes in the embedded feature space. After embedding both the support set and the query image into the network, the feature information S and Q are obtained. Subsequently, prototypes corresponding to each category are computed by averaging the features of samples within the respective support set category. Finally, in the metric space, the class label of the query sample is determined by its Euclidean distance from the prototype.

2 A prototypical network based on multi-scale features

The prototypical network model based on multi-scale features proposed in this paper is illustrated in Fig. 1. Firstly, the data from the support set and query set are fed into the feature extraction network. Then, the features of the images are mapped into the metric space. Each feature point in the support set is assigned corresponding weights using a weighting method, and

category prototypes are determined. Finally, the category to which a query sample point belongs is determined based on the calculated distance between the query sample point and the prototype. The judgment results are used to calculate the model loss based on the label-smoothing loss and contrastive loss functions.

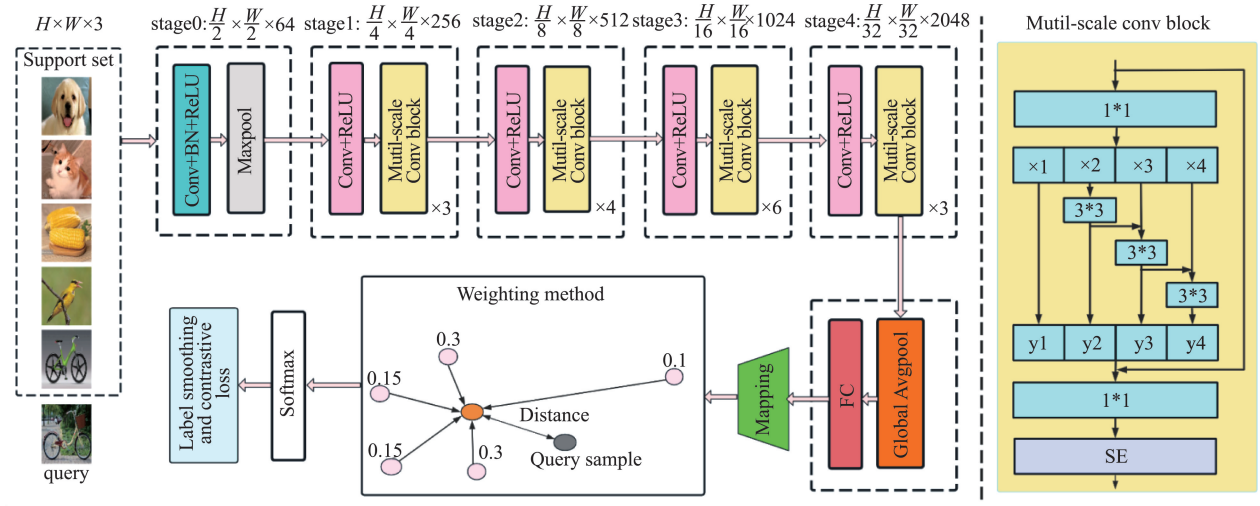


Fig. 1 Structure diagram of prototypical network based on multi-scale features

2.1 Extracting multi-scale features based on the Res2Net network

To comprehensively extract image feature information in few-shot learning tasks, a multi-scale feature input network[16] is employed within its backbone network. This integration enables the amalgamation of features at finer scales, thereby facilitating thorough image feature extraction. Furthermore, the backbone network incorporates a channel attention module to reinforce and fully exploit crucial information among channels. Fig. 2 illustrates the feature extraction model of the multi-scale feature prototypical network.

(1) The backbone network of the model uses the Res2Net50[17] network model with multi-scale feature input. It enables the model to obtain multi-scale features at a finer granularity during feature extraction and fully exploit the feature information of small sample images.

(2) The channel attention mechanism[18] (squeeze-and-excitation networks (SENet)) is embedded into the backbone network to weight and enhance key information between channels, thus mining the key information between channels.

In order to fully exploit the feature information of images, the Res2Net50 network is employed for feature extraction in the prototypical network, thereby obtaining a prototypical network for multi-scale feature extraction.

The Res2Net50 network achieves multi-scale feature extraction of images through its unique backbone. The backbone of this network is described in Fig. 3.

The principle of multi-scale feature extraction in the network is mainly to segment features through slicing when convolving image features. After each slice is convolved, it will be fused with the next slice to achieve the goal of fusing features of different scales.

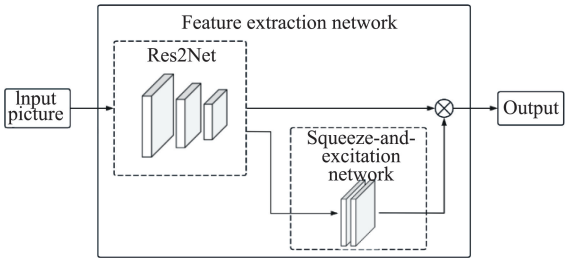


Fig. 2 Multi-scale prototypical network feature extraction model

The main parameters in the network include: convolution kernel size, stride, activation function, number of image feature slices, backbone repeat count, resolution of feature map, loss function, and so on. The number of image feature slices refers to the average number of image features divided by the number of channels in the Res2Net50 backbone. The resolution is the size of each layer's image feature, and the number of backbone repetitions is the number of multi-scale feature module that occurs in each stage. The loss function is used to evaluate the performance of the model during training and adjust weights based on its backpropagation gradient.

As described in Fig. 1 for the feature extraction network, the Res2Net50 network is divided into stages

0 – 4 based on the size changes of image features. In stage 0, there is a 7 * 7 convolution kernel with a stride of 2. In stages 1 – 4, each stage repeats the backbone 3, 4, 6, and 3 times, respectively, with the first 3 * 3 convolution having a stride of 2 and the remaining convolution operations having a stride of 1. In the backbone of Res2Net in this paper, the number of slices is set to 4. As shown in Fig. 3, before averaging the image feature slices, each slice undergoes a 1 * 1 convolution and ReLU activation. Then, they are convolved with a 3 * 3 kernel and fused with the next slice. After ReLU activation of the final result, another 1 * 1 convolution is applied, integrating it with the initial image features before undergoing ReLU activation again. Before the final convolution operation, the resolution of the image features is not changed. The last convolution operation doubles the number of channels.

In the backbone, the first part remains unchanged without any further processing. The subsequent segmened performers 3 * 3 convolution process, and is merged with the first segment, while the other segment is combined with the third before the subsequent convolution. This integration enables the third segment to incorporate information from the second, thus augmenting the output's representational scale. The operation of the Res2Net network is outlined in Eq. (3). The variable s represents the number of channel partitions, $K_i(\cdot)$ is the convolution operation function.

$$y_i = \begin{cases} x_i & i = 1 \\ K_i(x_i) & i = 2 \\ K_i(x_i + y_{j-1}) & 2 < i \leqslant s \end{cases} \tag{3}$$

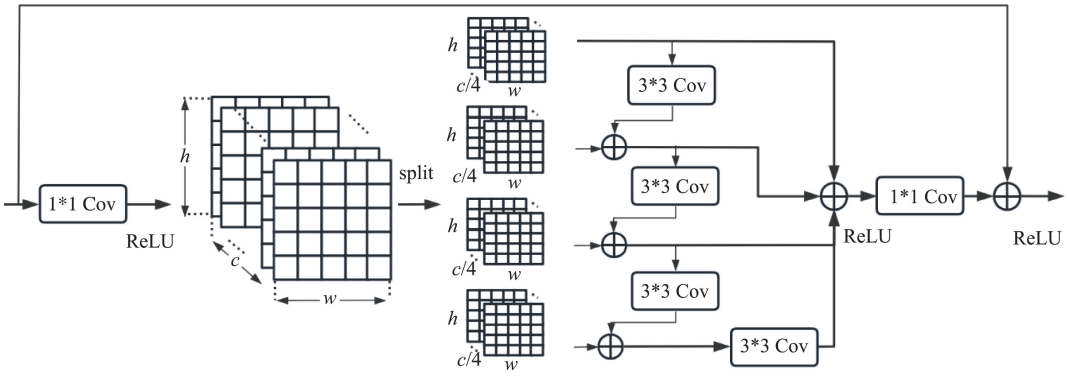


Fig. 3 Res2Net backbone structure diagram

The channel attention mechanism can reduce irrelevant interference between channels, thereby extracting more effective information during image feature extraction. The structure of SENet is illustrated in Fig. 4.

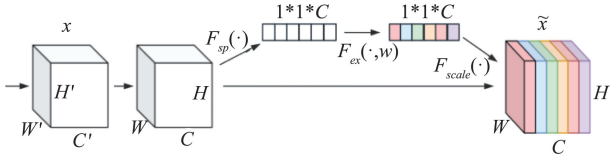


Fig. 4 SE structure diagram

The SENet's block contains a squeeze-and-excitation (SE) component. Initially, after the convolution process, the dimensions of the input image change from $C' * H' * W'$ to $C * H * W$. Subsequently, global average pooling (squeeze operation) is applied to each channel of the feature, obtaining the average value for the current channel. Then, the output of $1 * 1 * C$ is passed through two fully connected layers to constrain the final result between 0 and 1 (excitation operation). This yields C values corresponding to the weights of each channel, reflecting their respective importance. These weights are then multiplied with the post-convolution data for each channel before concatenating all channels. Then, the image features, weighted by channel importance, are obtained. Integrating the SE module after the prototypical network results in the squeeze-and-excitation prototypical network.

2.2 A prototypical network based on label-smoothing loss function

Owing to the limited sample size, misclassifications can significantly affect the network's performance. To address this issue, this study explores a method to mitigate the impact of minority outliers on the network. Specifically, contrastive loss and label-smoothing cross-entropy[19] are employed as the network's loss function to reduce the impact of misclassifications. It brings the distances between data points of the same category closer and increases the distances between those of different categories.

(1) To determine the prototype of each class,

weights are assigned to each sample point based on their spatial position in the small sample image classification task. The prototype is then computed according to the position of each sample point and the assigned weights.

(2) Contrastive loss is introduced into the loss function to reduce the distances between data points of the same category and increase the distances between those of different categories.

(3) The label-smoothing cross-entropy loss function is adopted to further reduce the impact of misclassifications of minority images on the network due to the limited sample size in the small sample image classification task.

2.2.1 A prototypical network based on distance weighting

In the prototypical network, the support set and query set are processed by the feature extraction network. The obtained features are mapped onto a feature space. The prototype for each category is computed from the features of the samples within the support set of the same category. The computation formula is shown in Eq. (4).

$$p_k = \frac{1}{k}\sum_{i=1}^{k} \tilde{x}_i \tag{4}$$

where p_k represents the computed prototype, k is the number of samples in the same category, and $\tilde{x}_i$ denotes the sample feature after feature extraction.

The calculation of prototypes using the provided equation presents an issue: when certain images within a category significantly differ from the rest, the distance between these specific samples and the others within the same category increases. In few-shot tasks where each class contains minimal data, this scenario notably impacts the determination of prototypes by the minority samples.

To mitigate the aforementioned impact, assigning weights to each sample point is proposed after mapping the data to the feature space. The magnitude of a sample point's weight is determined by the sum of distances between that sample point and the other sample points.

In this section, a weighted method is utilized to calculate prototypes using sample points. When considering an n-way k-shot few-shot learning task, the algorithm follows these steps.

The summation of distances between sample points within the same class is computed using the Euclidean distance (denoted as d) in the sample space. The calculation method is depicted in Eq. (5).

$$D_i = \sum_{x_i, x_k \in S} d(x_i, x_k) \tag{5}$$

The weight λ for each sample point is calculated based on the distance relationship among sample points of the same class. C represents a fixed constant. The computation process is described in Eq. (6).

$$\lambda_i = \frac{(C^2 + D_i^2)^{\frac{1}{2}}}{\sum_{x_j \in S} (C^2 + D_i^2)^{\frac{1}{2}}} \tag{6}$$

After getting the weights for all samples within a category, the prototype P_k is computed by summing the products of each sample point with its respective weight. This procedure is outlined in Eq. (7).

$$P_k = \sum_{i=1}^{k} \lambda_i x_i \tag{7}$$

2.2.2 Contrastive loss function

The loss function for the prototypical network is expressed in Eq. (8). Through the computation of the distance between sample points and prototypes, function loss calculation, and performing backpropagation, this method narrows the distance between feature points of the same category. However, it does not establish separation between different categories.

$$p_\varphi(y = k \mid x) = \frac{\exp(-d(f_\varphi(x), c_k))}{\sum_{k'} \exp(-d(f_\varphi(x), c_{k'}))} \tag{8}$$

where c_k represents the prototype point, and $f_\varphi(x)$ denotes the feature point in the feature space after convolutional processing. Contrastive loss in metric learning ensures that similar samples maintain their similarity post-feature extraction, while dissimilar samples retain their differences in the feature space. This loss function is mathematically defined in Eq. (9).

$$L(X_1, X_2) = \frac{1}{2N}\sum_{n=1}^{N} (Y D_W^2 + (1 - Y)\max(m - D_W, 0)^2) \tag{9}$$

where, D_w represents the Euclidean distance between two sample points.

$$Y = \begin{cases} 0 & X_1 \neq X_2 \\ 1 & X_1 = X_2 \end{cases}$$

When X_1 and X_2 are not same category. Y is set to 0, otherwise Y is set to 1. $Y = 1$ indicates samples from the same category, the loss function can be expressed as Eq. (10). Reducing the loss function facilitates the minimization of distances between samples of the same category.

$$L = \frac{1}{2N}\sum_{n=1}^{N} D_W^2 \tag{10}$$

When $Y=0$, denoting samples from different categories, the loss function is represented by Eq. (11). A suitable threshold value 'm' is set within the loss function. Similarly, reducing the entire loss function increases the distances between samples from different categories.

$$L = \frac{1}{2N}\sum_{n=1}^{N}\max(m - D_W, 0)^2 \tag{11}$$

2.2.3 Label-smoothing cross-entropy

In multi-class tasks, the cross-entropy loss function is expressed as Eq. (12).

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} y_i \log(p_\varphi) \tag{12}$$

where y_i is the label of the sample, with 1 for the positive class and 0 for the negative class. p_φ denotes the probability of predicting a positive sample.

Label-smoothing is a method to prevent overfitting. In the computation of the cross-entropy function, only the positive label's impact on the loss is considered, resulting in zero impact for negative class labels. In few-shot learning scenarios with limited data, an erroneous judgment on a single sample can inflict significant harm to the model and be detrimental to the training of the network model.

The label-smoothing cross-entropy loss function no longer utilizes the one-hot label format. Instead, it modifies the positive label by subtracting a very small constant from 1. As shown in Eq. (13), K represents the number of classes.

$$y_i = \begin{cases} 1-\varepsilon & i = y \\ \varepsilon/(K-1) & i \neq y \end{cases} \tag{13}$$

This adjustment in the cross-entropy function eliminates the strict binary 0 or 1 of target probabilities during computation. It helps to a certain extent in preventing overfitting and mitigates the impact of model prediction errors on the overall model performance.

3 Experiments and results analysis

3.1 Datasets

The experiments are carried out to validate the performance on the mini-ImageNet and CUB200 – 2011 small-sample datasets. The mini-ImageNet dataset is derived from ImageNet dataset[20]. The Google team extracted the mini-ImageNet dataset from the original ImageNet, which became widely used by researchers in the field of few-shot learning. On the other hand, the CUB200-2011 dataset is proposed by the California Institute of Technology. This dataset contains a total of 11 788 bird images, covering 200 subclasses of bird species. Each image provides label information for the corresponding bird category.

3.2 Experimental configuration

The experiments are conducted using the PyTorch deep learning framework, with the network model optimized using the Adam optimizer. The initial learning rate is set at 0.001. To enhance the model's generalization on new tasks without notably reducing accuracy on previous datasets, the learning rate is halved every 20 epochs. The adjustment of the learning rate follows Eq. (14).

$$L = 0.5^{\left[\frac{epoch}{20}\right]} \times 0.001 \tag{14}$$

Experiments are conducted on mini-ImageNet and CUB200-2011 datasets using 5-way 1-shot and 5-way 5-shot setups. Contrastive loss constants are set to 50 and 10, respectively. During each iteration of few-shot learning, images are randomly flipped, cropped, and resized to 224 × 224 dimensions. After normalization, these images are input into the feature extraction network to extract features and map data points into the feature space. Prototypes are determined using weighted approaches. Classification of samples in the query set into respective categories is based on their distance from the prototypes. Backpropagation is then used to optimize the network model, employing both contrastive loss and label-smoothing cross-entropy loss. The constant ε for label-smoothing cross-entropy is set to 0.1 in these experiments.

3.3 Results and analysis

The prototypical network model with multi-scale features underwent few-shot classification experiments on mini-ImageNet and CUB200-2011 datasets over 100 epochs, each comprising 100 training iterations followed by 100 validation iterations. Sample classification relied on distances between query set samples and category prototypes, with average accuracy across validation iterations calculated during training. Model performance in few-shot classification is evaluated on a separate test set and compared against other models, with detailed experiment specifics provided.

The experimental results of the multi-scale feature prototypical network on the test set are presented in Table 1. In the table, the Baseline refers to the prototypical network model. Baseline + SE indicates a network model with integrated channel attention mechanisms in the backbone. Baseline + Res2Net denotes the replacement of the backbone network with a multi-scale feature extraction network.

From Table 1, it can be observed that the addition of channel attention mechanisms to the multi-scale feature extraction model resulted in increased classification accuracy on both the mini-ImageNet and CUB200-2011 datasets. Specifically, for 5-way 1/5-shot tasks on the mini-ImageNet dataset, the accuracy shows improvements of 3.36% and 2.48%, respectively. On the CUB200-2011 dataset for 5-way 1/5-shot tasks, the accuracy exhibited increments of 7.05% and 8.18%, respectively.

Table 1 Performance comparison of multi-scale feature extraction models on datasets (%)

Method	mini-ImageNet		CUB-200-2011	
	5-way accuracy		5-way accuracy	
	1-shot	5-shot	1-shot	5-shot
Baseline	46.19	63.95	51.42	76.39
Baseline + SE	47.58	64.31	53.69	78.43
Baseline + Res2Net	49.20	65.01	56.31	80.16
Baseline + Res2Net + SE	49.55	66.43	58.47	84.57

The accuracy of the prototypical network based on distance-weighted values on the test set is depicted in Table 2. The Baseline + Weight denotes the prototypical network model utilizing the weighted determination of prototypes.

Table 2 Performance comparison of prototype weighted models on datasets (%)

Method	mini-ImageNet ($C=50$)	CUB-200-2011 ($C=10$)
	5-shot accuracy	5-shot accuracy
Baseline	63.95	76.39
Baseline + Weight	65.12	78.12

Table 2 illustrates that employing weighted processing on data points within the metric space effectively mitigates the impact of minority outlier points on the prototypes. As a result, the classification accuracy for 5-way 5-shot tasks on both the mini-ImageNet and CUB200-2011 datasets increased by 1.17% and 1.73% respectively.

The experimental prototypical network based on label-smoothing cross-entropy loss presents the test set accuracy in Table 3. In the table, Baseline + CL and Baseline + Smooth respectively denote the adoption of contrastive loss and label-smoothing cross-entropy loss functions within the loss function.

Table 3 reveals that employing label-smoothing loss in the prototypical network, along with the introduction of contrastive loss and label-smoothing cross-entropy loss functions, enhances the model's classification accuracy on both datasets. Specifically, for the mini-ImageNet dataset's 5-way 1/5-shot tasks, the accuracy improved by 1.10% and 1.58%, respectively. For the CUB200-2011 dataset's 5-way 1/5-shot tasks, the accuracy saw increments of 4.16% and 4.77%, respectively.

Table 3 Performance comparison of models on datasets after optimizing loss function (%)

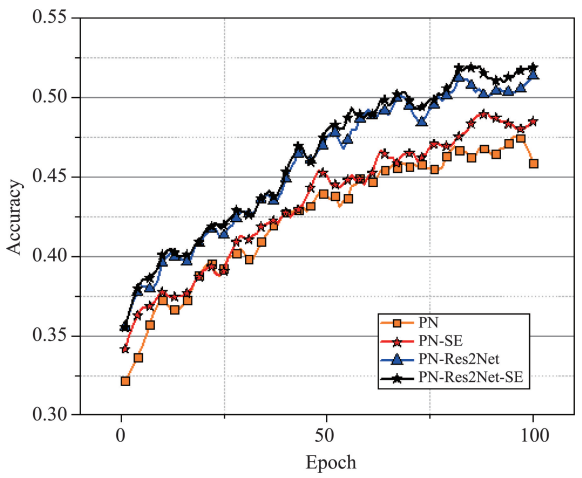
Method	mini-ImageNet		CUB-200-2011	
	5-way accuracy		5-way accuracy	
	1-shot	5-shot	1-shot	5-shot
Baseline	46.19	63.95	51.42	76.39
Baseline + CL	47.03	64.95	53.96	79.43
PN + CL + Smooth	47.29	65.53	55.58	81.16

The proposed MS-PN algorithm is compared with several other few-shot learning methods, including MatchNet[21], MAML[22] and ProtoNet. The results are recorded in Table 4. It should be noted in particular that for a fair of the experiment comparison, reproducing the network framework and feature extraction network in the comparison algorithms (MatchNet, MAML and ProtoNet) under the same experimental conditions. Furthermore, the same dataset division is used and the same parameters are set for the training tasks. Some of the processing rules can be referred to in Ref. [23] and Ref. [24]. From Table 4, it can be observed that the proposed MS-PN algorithm outperforms other methods in both the mini-ImageNet and CUB-200-2011 datasets for both 5-way 1-shot and 5-way 5-shot experiments. It indirectly indicates that the proposed network model can extract more robust image features during the feature extraction phase.

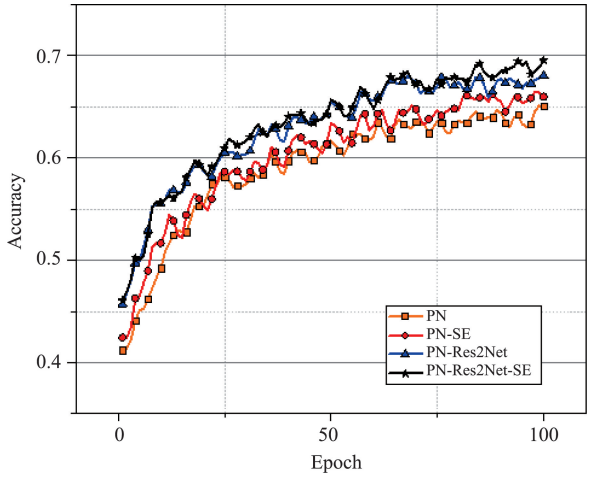
To gain further insights into the classification capabilities of these models, Fig. 5 and Fig. 6 display the experimental process of the prototypical network model with multi-scale features during training on the query set for the mini-ImageNet and CUB200-2011 datasets. The figures illustrate the accuracy fluctuations across epochs for PN, PN-Res2Net, PN-SE, and PN-Res2Net-SE models in 5-way 1-shot and 5-way 5-shot tasks.

Table 4 Performance comparison of different models under the same experimental conditions and settings (%)

Method	mini-ImageNet		CUB-200-2011	
	5-way accuracy		5-way accuracy	
	1-shot	5-shot	1-shot	5-shot
MatchNet	43.40	57.62	52.83	75.88
MAML	45.62	62.10	54.45	75.60
ProtoNet	46.19	63.95	51.42	76.39
MS-PN	**50.13**	**66.79**	**59.35**	**85.91**

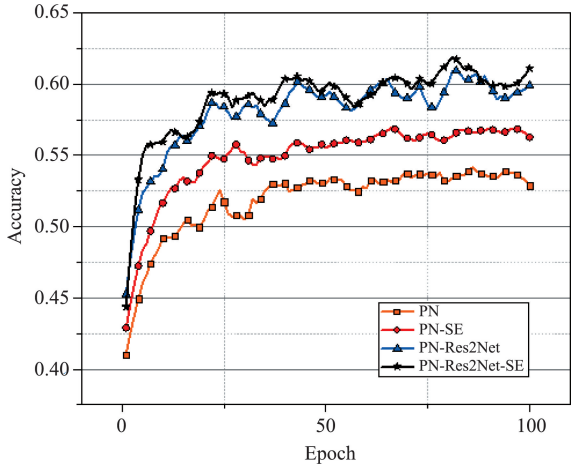


(a) Accuracy of multi-scale feature models in 5-way 1-shot classification experiments

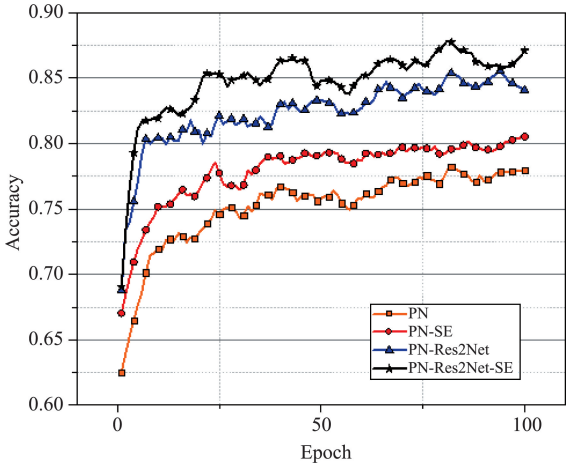


(b) Accuracy of multi-scale feature models in 5-way 5-shot classification experiments

Fig. 5 Experiments of multi-scale feature extraction model on mini-ImageNet dataset



(a) Accuracy of multi-scale feature models in 5-way 1-shot classification experiments



(b) Accuracy of multi-scale feature models in 5-way 5-shot classification experiments

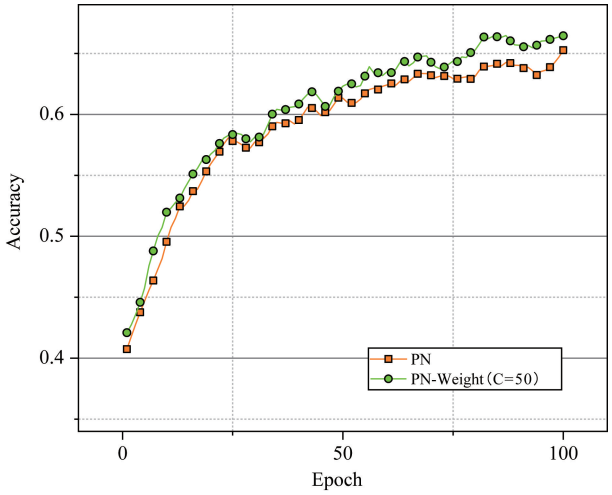
Fig. 6 Experiments of multi-scale feature extraction model on CUB-200-2011 dataset

Fig. 7(a) and Fig. 8(a) exhibit the accuracy changes across epochs for the PN and PN-Weight models on the 5-way 5-shot tasks during the training process of the prototypical network experiments with distance-weighted values for mini-ImageNet and CUB200-2011 datasets.

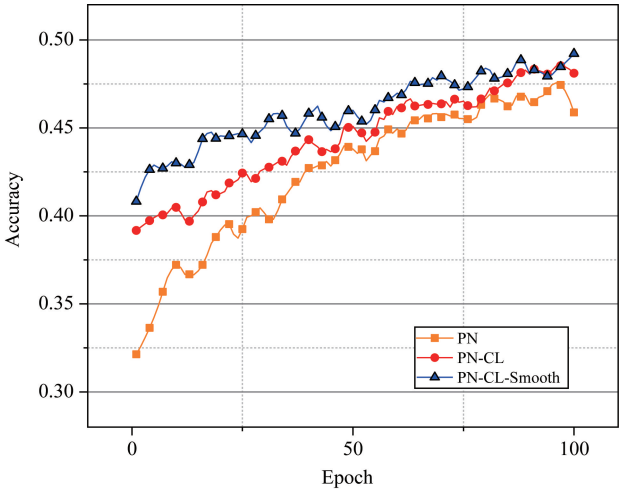
Fig. 7(b), (c) and Fig. 8(b), (c) respectively display the variations in accuracy across epochs for the PN, PN-CL, and PN-CL-Smooth models on the 5-way 1-shot and 5-way 5-shot tasks during the training process on label-smoothing loss for mini-ImageNet and CUB200-2011 datasets.

4 Conclusion

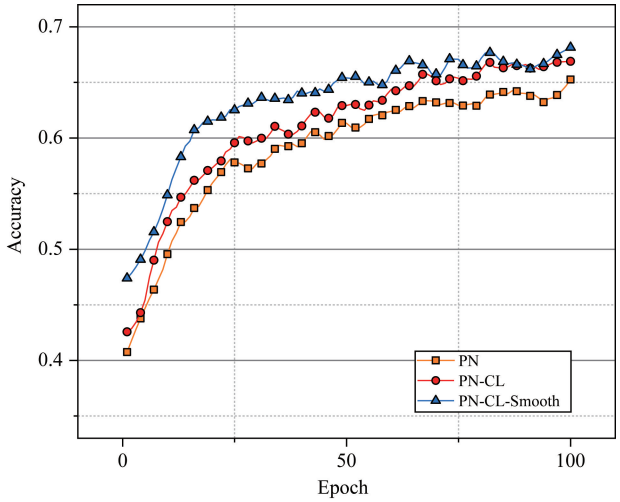
The paper introduces a prototypical network model based on multi-scale features. It utilizes multi-scale feature extraction to obtain diverse feature information from images at different scales. Furthermore, a prototype calculation method based on distance-weighted values reduces the influence of minority outlier points on the network. The utilization of label-smoothing cross-entropy functions directs the model's attention towards the potential misclassification of positive samples as negative, thereby reducing potential harm caused by network errors. The incorporation of contrastive loss functions brings similar data closer while separating dissimilar data. Through experiments conducted on the mini-ImageNet and CUB200-2011 datasets, the proposed network model demonstrates its capability to extract more robust image features during the feature extraction phase and achieves higher classification accuracy in few-shot classification tasks.



(a) Accuracy of prototypical network experiments based on distance weights in 5-way 5-shot classification experiments

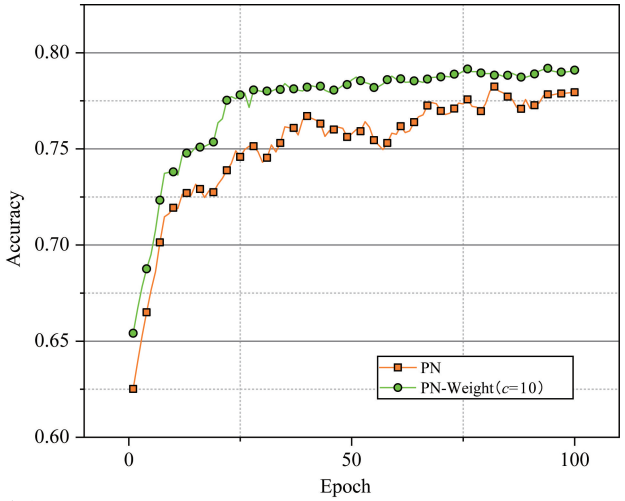


(b) Accuracy of prototypical network based on label-smoothing loss in 5-way 1-shot classification experiments

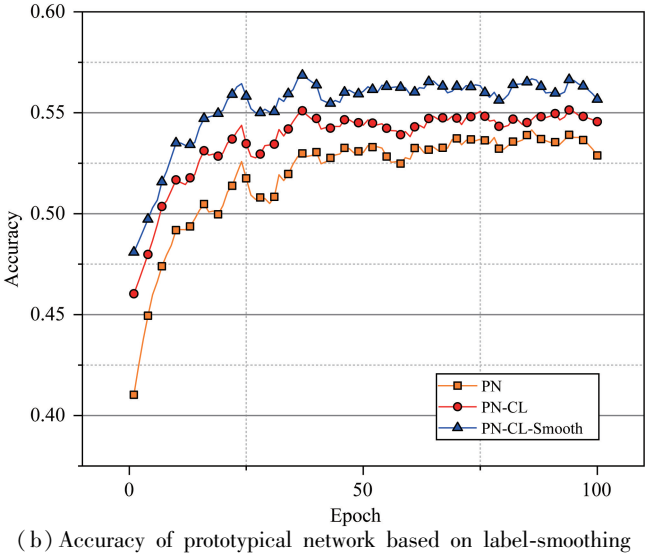


(c) Accuracy of prototypical network based on label-smoothing loss in 5-way 5-shot classification experiments

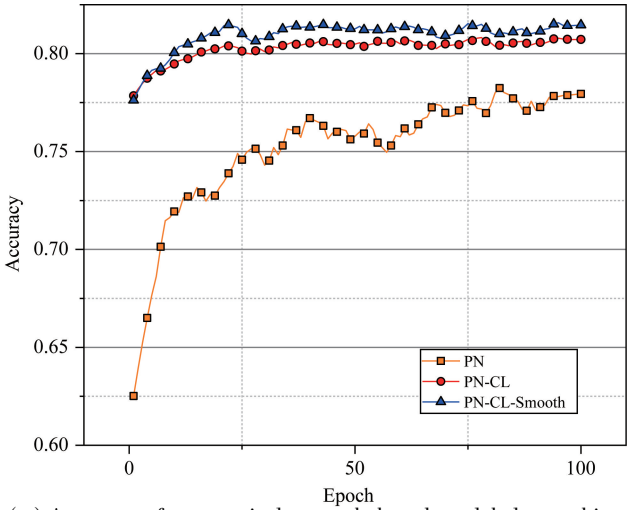
Fig. 7 Experiments of label-smoothing loss on mini-ImageNet dataset



(a) Accuracy of prototypical network experiments based on distance weights in 5-way 5-shot classification experiments



(b) Accuracy of prototypical network based on label-smoothing loss in 5-way 1-shot classification experiments



(c) Accuracy of prototypical network based on label-smoothing loss in 5-way 5-shot classification experiments

Fig. 8 Experiments of label-smoothing loss on CUB-200-2011 dataset

References

[1] LI Y. Research and application of deep learning in image recognition [C]//International Conference on Power, Electronics and Computer Applications. Shenyang, China: IEEE, 2022: 994-999.

[2] JI C Q, GAO Z Y, QIN J, et al. Overview of image classification algorithms based on convolutional neural network[J]. Computer Application, 2022, 42(4): 1044-1049.

[3] LIU Z, NING J, CAO Y, et al. Video swin transformer [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 3202-3211.

[4] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: a survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

[5] LI W G, GAN P, XIE L, et al. A small sample image classification method based on sample pair meta learning [J]. Acta Electronica Sinica, 2022, 50(2): 295-304.

[6] ZHANG J, ZHANG X, LV L, et al. An applicative survey on few-shot learning[J]. Recent Patents on Engineering, 2022, 16(5): 104-124.

[7] LIU Y, LEI Y B, FAN J L, et al. Overview of image classification techniques based on small sample learning [J]. Journal of Automation, 2021, 47(2): 297-315.

[8] ZHOU Z, QIU X, XIE J, et al. Binocular mutual learning for improving few-shot classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 8402-8411.

[9] KANG D, KWON H, MIN J, et al. Relational embedding for few-shot classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 8822-8833.

[10] QI G, YU H, LU Z, et al. Transductive few-shot classification on the oblique manifold[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 8412-8422.

[11] ALLEN K, SHELHAMER E, SHIN H, et al. Infinite mixture prototypes for few-shot learning[C]//International Conference on Machine Learning. Hualian, China: IMLS, 2019: 232-241.

[12] CHOWDHURY R R, BATHULA D R. Influential prototypical networks for few shot learning: a dermatological case study[C]//IEEE 19th International Symposium on Biomedical Imaging (ISBI). Kolkata, India: IEEE, 2022: 1-4.

[13] YU J J, CHENG H, FANG Y Q. A multi task prototype network for text classification with few samples[J]. Application Research of Computers, 2022, 39(5): 1368-1373.

[14] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[J]. Advances in Neural Information Processing Systems, 2017, 30:4080-4090.

[15] SUAREZ J L, GARCIA S, HERRERA F. A tutorial on distance metric learning: mathematical foundations, algorithms, experimental analysis, prospects and challenges [J]. Neurocomputing. 2021, 425: 300-322.

[16] XIAO Y, ZHOU K, CUI G, et al. Deep learning for occluded and multi-scale pedestrian detection: a review [J]. IET Image Processing. 2021, 15(2): 286-301.

[17] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2): 652-662.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7132-7141.

[19] ZHANG C B, JIANG P T, HOU Q, et al. Delving deep into label-smoothing[J]. IEEE Transactions on Image Processing, 2021, 30: 5984-5996.

[20] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition. Florida, USA: IEEE, 2009: 248-255.

[21] HAN X, LEUNG T, JIA Y, et al. MatchNet: unifying feature and metric learning for patch-based matching [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 3279-3286.

[22] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// The 34th International Conference on Machine Learning. Sydney, Australia: ACM, 2017: 1856-1868.

[23] BAIK S, CHOI J, KIN H, et al. Meta-learning with task-adaptive loss function for few-shot learning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 9465-9474.

[24] XIAO B, LIU C L, HSAIO W H. Semantic cross attention for few-shot learning [EB/OL]. (2022-10-12) [2024-02-28]. https://arxiv.org/pdf/2210.06311.

LIU Jiatong, born in 1998. He is a graduate student in School of Information Science and Engineering of Shenyang University of Technology. He received her B. S. degree from Shenyang University of Technology in 2021. His research interests include machine learning and intelligent software.