doi:10.3772/j.issn.1006-6748.2022.03.009

Multi-attention fusion and weighted class representation for few-shot classification①

ZHAO Wencang (赵文仓), QIN Wenqian②, LI Ming
(College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, P. R. China)

Abstract

The existing few-shot learning (FSL) approaches based on metric-learning usually lack attention to the distinction of feature contributions, and the importance of each sample is often ignored when obtaining the class representation, where the performance of the model is limited. Additionally, similarity metric method is also worthy of attention. Therefore, a few-shot learning approach called MWNet based on multi-attention fusion and weighted class representation (WCR) is proposed in this paper. Firstly, a multi-attention fusion module is introduced into the model to highlight the valuable part of the feature and reduce the interference of irrelevant content. Then, when obtaining the class representation, weight is given to each support set sample, and the weighted class representation is used to better express the class. Moreover, a mutual similarity metric method is used to obtain a more accurate similarity relationship through the mutual similarity for each representation. Experiments prove that the approach in this paper performs well in few-shot image classification, and also shows remarkable excellence and competitiveness compared with related advanced techniques.

Key words: few-shot learning (FSL), image classification, metric-learning, multi-attention fusion

0 Introduction

Learning from a few examples is still a key challenge for many machine vision tasks. Humans can recognize new objects from a small number of samples. In order to imitate this cognitive intelligence of humans, few-shot learning (FSL) has been proposed, and it has quickly become a hot and challenging research field. It aims to learn a classifier that has good generalization ability when faced with a few new unknown class samples.

The existing few-shot learning technologies can be roughly divided into two categories: approaches based on metric-learning[1-9] and approaches based on meta-learning[10-13]. The basic idea of the former is to learn an embedding space and classify samples according to the similarity metric between the query sample and the representation of each class. The goal of the latter is to learn a cross-task meta-learner to improve generalization ability through cross-task knowledge transfer. Both of these two types of approaches are designed to allow the trained model to classify query samples with limited support samples.

Excellent research progress has been made by FSL approaches based on metric-learning, but the existing approaches still have some limitations. Breaking these limitations to improve model performance is the main motivation of this work.

Such approaches usually have a simple and efficient architecture, but the emphasis on highlighting valuable regions and weakening irrelevant regions in the feature extraction stage is still insufficient. The feature expression ability of the model will be limited, and serious deviations could even be caused in the subsequent metric process, thus the desired classification effect failed to be achieved. This problem can be solved ingeniously by the attention mechanism[14]. It is inspired by the human visual mechanism that has been widely used in many fields of deep learning[15-16]. It can focus on more important information and reduce the focus on irrelevant information. In order to obtain more valuable feature representations, in this work, the attention mechanism will be introduced into the few-shot classification tasks in the form of a multi-attention fusion module (MAFM). By acquiring the attention weights of the features from the spatial dimension and the channel dimension, and performing multi-attention fusion, the features extracted by the model are more

① Supported by the National Natural Science Foundation of China (No. 61171131) and Key R&D Program of Shandong Province (No. YD01033).
② To whom correspondence should be addressed. E-mail: qinwenqian0131@foxmail.com.
Received on July 20, 2021

meaningful.

Most of the existing approaches are similar to the PrototypicalNet[2], where the averaging of the feature vectors of various support set samples is used as the prototype (class representation) of each class. The value of each sample of the class is often not considered, and the contribution of each support set sample in each class to the prototype of the class is regarded as consistent. However, in this case, it is easy to be affected by some invalid or interfering samples, resulting in the prototype not reaching good representativeness. That is to say, the usefulness of each feature vector to the prototype can not be effectively evaluated, and the similarity judgment result will be biased. In response to this problem, a weighted class representation (WCR) is proposed, which is generated after evaluating the value of each support sample. It can reduce the negative impact of interfering samples, and assign more valuable samples with greater weight, so that the final weighted class representation is more beneficial to FSL.

In addition, prototypes can be obtained through shrinking the support set, and then the one-way similarity between query samples and prototypes is directly compared by such approaches. It is hardly considered that the similarity comparison is not only in a single direction, but also can be measured from the perspective of class to query. Therefore, some similarity deviations in the metric stage may be produced, where the final classification results will be affected. In this paper, the conventional one-way similarity metric method is not used, but the interaction between samples and class representations is fully considered. A mutual similarity metric method is introduced in the feature space to determine the class attribution of the sample, and a more convincing class discrimination result is obtained, so as to better improve the performance of the model.

A series of experimental results on the few-shot image classification benchmark dataset miniImageNet[1] and the fine-grained dataset Stanford Dogs[17] show that the few-shot learning approach based on multi-attention fusion and weighted class representation proposed in this paper has excellent performance.

The main work and contributions of this paper are as follows.

(1) An effective multi-attention fusion module is designed to optimize the features by acquiring attention in the spatial dimension and the channel dimension, so that the valuable information is highlighted by the extracted features and the interference of irrelevant information is reduced.

(2) The weighted class representation is used to replace the traditional approaches of finding the mean value of feature vectors as the class prototype, where the obtained class representation features are more valuable and more accurate.

(3) In the classification stage, a mutual similarity metric method is introduced, and two-way similarity discrimination between the query and the class is carried out, and the mutual similarity for each representation is combined to make the final similarity metric result more reliable.

1 Related work

Many breakthroughs in the current area have been made by the development of few-shot learning, and a brief example is given to introduce its two major branches.

Approaches based on metric-learning. Most of these approaches are for learning the similarity comparison of information among samples. The attention mechanism is introduced in MatchingNet[1], and the attention score of the extracted features is used to predict the class of the query sample. The support set of each class is shrank by PrototypicalNet[2] into a representation of a class by finding the mean vector of the support set samples of each class, called a prototype, and then the distance (similarity) between the query image and each prototype is compared to be classified. Neural network is used by RelationNet[3] to analyze the similarity between samples, which can be regarded as a non-linear classifier. Conditional batch normalization is relied by task dependent adaptive metric (TADAM)[4] to increase task adaptability and task-related metric space is learnt. The similarity metric between images and classes is considered by DN4[5] from the perspective of local descriptors. The research of subspace is involved by TapNet[6] and DSN[7], and the rationality of subspace modeling for few-shot learning is verified. A category traversal module (CTM) is used in Ref. [8] to make full use of the intra-class commonality and inter-class uniqueness of features. An attention mechanism is introduced by MADN4[9] to improve the extracted local descriptors.

Approaches based on meta-learning. A meta-learner is usually learnt by such approaches to adjust the optimization algorithm. The meta-learner is trained by model-agnostic meta-learning (MAML)[10] to perform proper parameter initialization to better adapt to the new task, so that the model can have good performance on the new task with only a few steps of gradient descent. A differentiable quadratic programming (QP) solver is combined by MetaOptNet[11] to achieve good

performance. The goal of latent embedding optimization (LEO)[12] is to train the model in a high-dimensional parameter space, and it only needs a few updates in the low-data area. Attentive weights generation for few shot learning via information maximization (AWGIM)[13] is also based on parameter optimization, learning to directly use a generator to generate the weight parameters of the classifier.

The approach proposed in this paper is based on metric-learning. The difference from other related work is that this paper has carried out a unique work in the attention mechanism, support set class representation and similarity metric method. In the experimental section, the validity and advancement of these work in this paper are verified, and the comparison and connection with related approaches are made.

2 Proposed approach

2.1 Problem set-up

The episodic training mechanism[1] is widely used in the training process of FSL. Usually the support set and query set are randomly selected from the training set to train the model. The entire training process is divided into many episodes. In each episode, a small number of labeled samples in the support set are used for training, and then query images determine which category in the support set they belong to. In this work, the episodic training mechanism will also be used to train the model. Specifically, each episode contains N_e sample classes which are randomly selected from the training set. Among them, each class $m(m \in 1, \cdots, N_e)$ has N_s support samples and N_q query samples as the support set S_m and query set Q_m in the episode, respectively. The samples in the support set and the query set are randomly selected and do not intersect, and the number of samples can be set manually, that is, the support images and query images of class m together form the support set and query set of this class. According to the number of sample classes N_e and the number of support set samples N_s, each of these episodes can also be called FSL in the case of N_e-way N_s-shot.

2.2 Model overview

As shown in Fig. 1, the support set samples and query samples are input into the feature extraction backbone network f_θ embedded with the multi-attention fusion module. In the process of feature extraction, after multi-attention acquisition of spatial dimension and channel dimension, more distinguishing and valuable features are obtained. In the feature space, the positions of samples belonging to the same class are close together. According to the method proposed in this paper, weight is assigned to each support set sample, and the weighted class representation of each class is calculated. Then the mutual similarity between the query sample and the weighted class representation is measured. The category of the class representation with the highest mutual similarity score is the classification result of the query sample. The specific content of each part will be shown in the following sections.

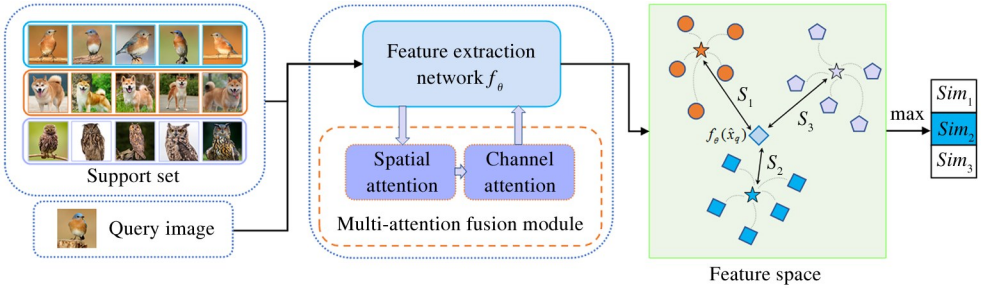


Fig. 1 The overall framework of the proposed MWNet

2.3 Multi-attention fusion module

In order to obtain more distinguishing features in the feature extraction stage, an attention module based on multi-attention fusion is designed in this paper, which is mainly composed of two parts: spatial attention part and channel attention part. Different from the single-dimensional attention mechanism, attention weights are obtained by this module in the spatial dimension and channel dimension respectively, and they are connected in series, so it can be regarded as a multi-dimensional attention fusion. The valuable regions can be highlighted by the obtained features while rich information is included and irrelevant regions are suppressed, so it is called a multi-attention fusion module.

Spatial attention. This part of the structure is shown in Fig. 2. In order to find the weight relationship in the spatial dimension, the intermediate feature map $F \in R^{c \times h \times w}$ extracted in the previous stage is reduced through a 1×1 convolutional layer with a channel num-

ber of 1, and then the Sigmoid function is used to obtain the spatial attention weight $SA \in R^{1\times h\times w}$. The calculation process can be expressed as

$$S_A = \sigma(Conv(F)) \tag{1}$$

where σ represents the Sigmoid function, and $Conv$ represents the convolution operation. It is equivalent to compressing the information of the original c channels to a single channel and using it as a weight, and then multiplying the weight S_A and F to obtain the spatial attention optimization feature map $F' \in R^{c\times h\times w}$, namely, $F' = F \otimes S_A$.

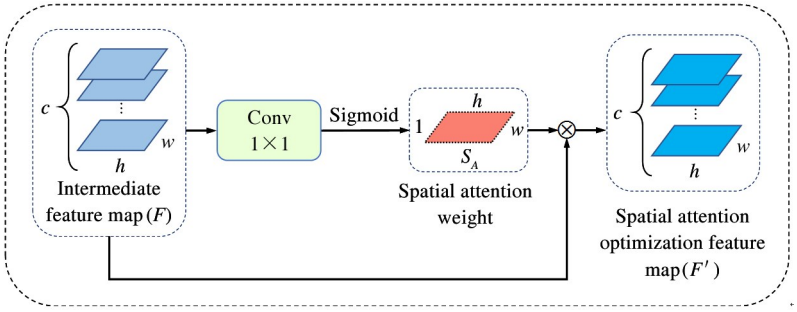


Fig. 2 Spatial attention acquisition process

Channel attention. This part of the structure is shown in Fig. 3. Perform global average pooling on F' obtained through spatial attention optimization, that is, compress the global information in each channel dimension of F' into the global average pooling vector $\boldsymbol{F}_G \in R^{c\times 1\times 1}$. Then pass through the two fully connected layers in turn and use the Sigmoid function to obtain the final channel attention weight $C_A \in R^{c\times 1\times 1}$, which can be expressed as

$$C_A = \sigma(full(GAP(F'))) \tag{2}$$

where, $full$ and GAP respectively represent that the feature passes through the two-layer fully connected layer and the global average pooling layer. Then multiply C_A and F' to get the final multi-attention optimization feature map $F'' \in R^{c\times h\times w}$, namely, $F'' = F' \otimes C_A$.

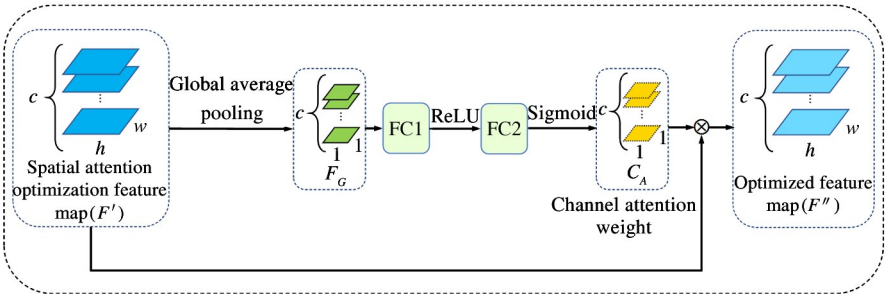


Fig. 3 Channel attention acquisition process

After the above steps, the model realizes the multi-information interaction between the spatial dimension and the channel dimension, that is, multi-attention fusion. Moreover, this module can be directly embedded in the feature extraction backbone, effectively highlighting the important part of the feature while keeping the original feature map size unchanged, which is more conducive to subsequent similarity metric.

2.4 Weighted class representation

Ideally, samples belonging to the same class should be as close as possible in the feature space, but it is inevitable that occasionally one or several interfering samples deviate from other samples in the class. At this time, if all samples of this category are treated equally and the feature vectors are averaged, the representativeness of the obtained prototype may not be ideal. This is due to the negligence in obtaining the prototype: in addition to positive effects, some samples will also have interference factors. If these negative effects are not fully considered, the obtained prototype (class representation) will be biased. In order to reduce this deviation, in this work, a weighted class representation with the size of the weight is proposed. It is not simply averaging the feature vectors of the support set samples, but fully considers that when calculating the class representation of each class sample, each sample has its own different degree of positive influence. The idea of this work is to give more weight to samples with greater positive influence. Specifically, in the support

samples of the same class, the smaller the Euclidean distance between a sample and other samples, the larger the proportion of the sample in the construction of the class representation, on the contrary, the proportion is smaller. Based on this idea, the representation of each class sample obtained by calculation is more ideal.

The process of obtaining is as follows. First, the support sample x_i and other samples $x_j (j \neq i)$ of the same class pass through the feature extraction network embedded with the attention module to obtain the feature vectors $f_\theta(x_i)$ and $f_\theta(x_j)$ after attention optimization. Then in this class, the average value of the sum of Euclidean distances between x_i and each x_j can be expressed by Eq. (3). The larger the α_i, the greater the difference between sample x_i and other samples in the same class, and the smaller the similarity.

$$\alpha_i = \frac{1}{N_s - 1} \sum_{j \neq i} d(f_\theta(x_i), f_\theta(x_j)) \tag{3}$$

where N_s is the number of support set samples for each class, and $d(\cdot)$ represents the calculation of Euclidean distance. Further, when constructing the class representation, the weight of the sample x_i can be expressed by the negative value of α_i through the Softmax function as

$$w_i = \frac{\exp(-\alpha_i)}{\sum_{i=1}^{N_s} \exp(-\alpha_i)} \tag{4}$$

Finally, each support sample x_i in the class is combined with its weight w_i to perform a weighted summation to obtain the weighted class representation of the current class m.

$$\xi_m = \sum_{(x_i, y_i) \in S_m} \frac{\exp(-\alpha_i)}{\sum_{i=1}^{N_s} \exp(-\alpha_i)} f_\theta(x_i) \tag{5}$$

By calculating the weighted class representation, it is possible to largely avoid some samples with large deviations from the samples of the same class from interfering with the calculation of the entire class representation. It can also make the model get better performance with more accurate class representation.

2.5 Mutual similarity metric method

Conventional methods usually only carry out a one-way metric from a certain query sample to each class, but do not consider the metric from a certain class to each query sample from the perspective of each class. In this work, research has been conducted on this issue. Here, a mutual similarity metric method is introduced. Specifically, the idea of this method is: when the similarity from query $\hat{x}_q$ to the weighted class representation ξ_m of class m is high, and the similarity from ξ_m to the query $\hat{x}_q$ is also high. Then it can be considered that the similarity discrimination result that $\hat{x}_q$ belongs to class m is highly credible. The specific process of the mutual similarity metric method is as follows.

The query sample $\hat{x}_q$ passes through the feature extraction network embedded with the attention module to obtain the attention optimized sample feature $f_\theta(\hat{x}_q)$, then the probability of its belonging to each class m can be calculated by the Softmax function.

$$Sim(\hat{x}_q \rightarrow \xi_m) = \frac{\exp(-d(f_\theta(\hat{x}_q), \xi_m))}{\sum_{m'} \exp(-d(f_\theta(\hat{x}_q), \xi_{m'}))} \tag{6}$$

Eq. (6) can be used as the similarity score between $\hat{x}_q$ and each weighted class representation ξ_m. From another perspective, the probability that the weighted class representation ξ_m of class m belongs to each query sample $\hat{x}_q$ can be calculated as

$$Sim(\xi_m \rightarrow \hat{x}_q) = \frac{\exp(-d(\xi_m, f_\theta(\hat{x}_q)))}{\sum_{q'} \exp(-d(\xi_m, f_\theta(\hat{x}_{q'})))} \tag{7}$$

Eq. (7) is the similarity score between ξ_m and each $\hat{x}_q$. Multiply these two similarity scores to obtain the mutual similarity between query $\hat{x}_q$ and weighted class representation ξ_m:

$$Sim(\hat{x}_q \leftrightarrow \xi_m) = Sim(\hat{x}_q \rightarrow \xi_m) \cdot Sim(\xi_m \rightarrow \hat{x}_q) \tag{8}$$

By carrying out similarity metric from different angles and combining them appropriately, the model can make better use of the interrelationship between query and class representation, interactively fuse information, and make the obtained similarity metric results more accurate.

2.6 Training algorithm

Algorithm 1 shows the episodic training process of MWNet in this paper. For each episode, the support set samples and query samples are input into the feature extraction network embedded with the multi-attention fusion module, and the optimized sample features are obtained after attention acquisition in the spatial and channel dimensions. Next, weight is assigned to each support set sample based on the Euclidean distance and Softmax function, and the weighted class representation of each class is calculated. Finally, the mutual similarity metric method is used in the feature space to predict the class of query samples, and the parameters of the feature extraction network are updated by minimizing the classification loss of the episode. Then use the updated model to process the new episode

until the training is completed.

3 Experiments

In order to evaluate the performance of the approach proposed in this paper, this section conducts experiments on the benchmark dataset miniImageNet[1] in the field of few-shot learning, and compares it with the existing advanced approaches. Further, in order to explore the effectiveness of the approach in this paper on fine-grained images, the fine-grained dataset Stanford Dogs[17] with small inter-class changes and large intra-class changes is selected for experiments, and compared with related metric-learning based approaches on this dataset. In addition, in this section, the feature space visualization of the model, research on higher way training, and ablation experiments will also be carried out.

3.1 Datasets

miniImageNet dataset is a small version of ImageNet[18]. It has a total of 100 classes, each with 600 samples, and the image resolution is 84×84. This paper adopts the division method in PrototypicalNet[2]: 64 classes for training, 16 classes for verification, and 20 classes for testing.

Stanford Dogs dataset is often used for fine-grained image classification. It has 120 classes and a total of 20 580 images. According to the method in Ref. [5], it is divided into 70 training classes, 20 verification classes and 30 testing classes.

Algorithm 1 The process of a training episode for MWNet

Input: each episode e_i with S and Q

for i in $\{e_1, \cdots, e_I\}$ **do**
  $L_i \leftarrow 0$
  for sample x in S, Q **do**
  $F \leftarrow$ Intermediate feature map of sample x
  $S_A \leftarrow \sigma(Conv(F))$ Spatial attention weight
  $F' \leftarrow F \otimes S_A$ Spatial attention optimization feature map
  $C_A \leftarrow \sigma(full(GAP(F')))$ Channel attention weight
  $F'' \leftarrow F' \otimes C_A$ Multi-attention optimization feature map
  $f_\theta(x) \leftarrow$ The final feature map of sample x
  end for
  for m in $\{1, \cdots, N_e\}$ **do**
    for $(x_i, y_i) \in S_m$, $(x_j, y_j) \in S_m$ **do**

$$\alpha_i = \frac{1}{N_s - 1} \sum_{j \neq i} d(f_\theta(x_i), f_\theta(x_j))$$

$$w_i = \frac{\exp(-\alpha_i)}{\sum_{i=1}^{N_s} \exp(-\alpha_i)}$$

$$\xi_m = \sum_{(x_i, y_i) \in S_m} \frac{\exp(-\alpha_i)}{\sum_{i=1}^{N_s} \exp(-\alpha_i)} f_\theta(x_i)$$ Weighted class representation

    end for
  end for
  for m in $\{1, \cdots, N_e\}$ **do**
    for q in $\{1, \cdots, N_q\}$ **do**

$$Sim(\hat{x}_q \rightarrow \xi_m) \leftarrow \frac{\exp(-d(f_\theta(\hat{x}_q), \xi_m))}{\sum_{m'} \exp(-d(f_\theta(\hat{x}_q), \xi_{m'}))}$$

$$Sim(\xi_m \rightarrow \hat{x}_q) \leftarrow \frac{\exp(-d(\xi_m, f_\theta(\hat{x}_q)))}{\sum_{q'} \exp(-d(\xi_m, f_\theta(\hat{x}_{q'})))}$$

$$Sim(\hat{x}_q \leftrightarrow \xi_m) \leftarrow Sim(x_q \rightarrow \xi_m) \cdot Sim(\xi_m \rightarrow \hat{x}_q)$$ Mutual similarity

    end for
  end for

25: $L_i \leftarrow \frac{1}{N_e N_q} \sum_m \sum_q [-\log(Sim(\hat{x}_q \leftrightarrow \xi_m))]$

26: Update θ using ∇L_i

27: **end for**

3.2 Experimental setup

The experiments in this section use typical few-shot image classification settings, namely N-way K-shot settings. The Adam algorithm[19] is used for training, the initial learning rate is 10^{-3}, and more support set classes are used for training than for testing. In the experiments on miniImageNet, there are a total of 6×10^4 training episodes, with 12 query samples in each class. For the 5-way 1-shot tasks, the learning rate drops by 10 times after every 2.5×10^4 episodes. For the 5-way 5-shot tasks, the learning rate drops by 10 times after every 4×10^4 episodes. In the test phase, there are 15 query samples for each class, and the average accuracy of 4×10^4 episodes is randomly selected to evaluate the performance of the model. For the fine-grained few-shot classification experiments on the Stanford Dogs[17] dataset, since the number of samples in the dataset is much smaller than in miniImageNet, in order to avoid model overfitting, data augmentation technology is used. The rest of the experimental settings are the same as those in miniImageNet.

3.3 Feature extraction backbone

In order to better compare the model with other advanced approaches, in the experiments on miniImageNet in this section, two feature extraction backbone networks are used to implement the models in this paper. The first one used is ResNet-12[20], which is the same as that used in TapNet[6]. It consists of four residual blocks with channel numbers (represented by L in Fig. 4) of 64, 128, 256, and 512 respectively. Each residual block contains three 3×3 convolutional blocks and a shortcut connection. After the convolutional block is the batch normalization (BN) layer and the ReLU activation function, there is a 2×2 max-pooling layer after each residual block, and the shortcut connection contains 3×3 convolutional layer and batch normalization layer. The multi-attention fusion module is embedded after the last convolutional layer in each residual block, and a global average pooling layer is added at the end of the backbone. The structure of the entire network is shown in Fig. 4.

In addition, the common four-layer convolutional network Conv-4 is also used as the backbone, as used in PrototypicalNet[2]. It has a total of 4 convolutional blocks, each of which contains 64 3×3 kernels, a batch normalization layer, a ReLU activation function, and a 2×2 max-pooling layer. In Conv-4, the multi-attention fusion module is embedded after the first two convolutional blocks, as shown in Fig. 5.

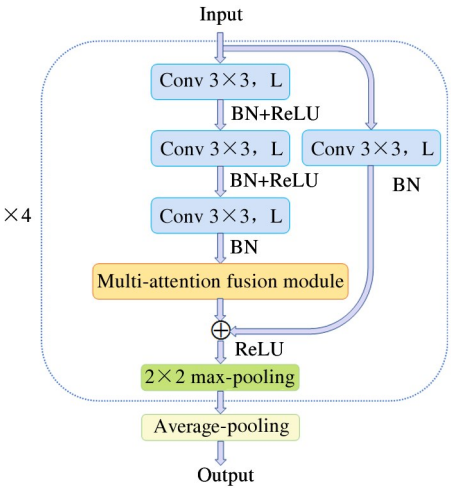


Fig. 4 ResNet-12 embedded with multi-attention fusion module

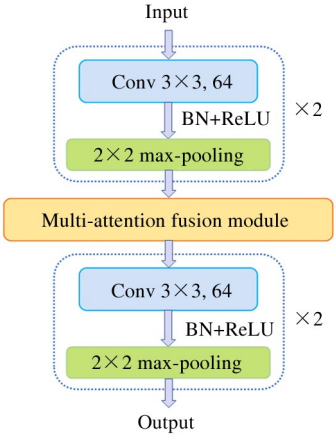


Fig. 5 Conv-4 embedded with multi-attention fusion module

3.4 Comparison results

Table 1 shows the experimental results of the model MWNet proposed in this paper compared with advanced technologies on miniImageNet[1] when Conv-4 and ResNet-12 are used respectively. It can be seen that when using the same size backbone, MWNet always outperforms other approaches in 5-way 1-shot and 5-way 5-shot tasks. Compared with the benchmark approach PrototypicalNet[2], which is also based on metric-learning, the model in this paper has achieved obvious advantages. In the case of Conv-4, the classification accuracy of 5-way 1-shot and 5-way 5-shot tasks are 4.15% and 4.19% higher than PrototypicalNet re-

spectively. In the case of ResNet-12, the classification accuracy of 5-way 1-shot and 5-way 5-shot tasks have advantages of 4.01% and 4.43%, respectively. Compared with TapNet[6] and DSN[7] based on subspace, the approach in this paper achieves better results without involving complex subspace structure, which is simpler and more efficient. CTM[8] involves model fine-tuning, and uses ResNet-18, which is deeper than ResNet-12, and MWNet does not need such a deep backbone to achieve better performance through end-to-end training. In addition, compared with advanced meta-learning based technologies, MWNet still has strong competitiveness. It is worth noting that both LEO[12] and AWGIM[13] use deeper and wider wide residual network (WRN-28-10)[21], and the model in this paper can compete with them without such a complex network architecture, and has a higher classification accuracy.

Table 1 Accuracy comparison with other approaches on miniImageNet

Model	Backbone	5-way 1-shot	5-way 5-shot
MatchingNet[1]	Conv-4	43.56 ± 0.84%	55.31 ± 0.73%
PrototypicalNet[2]	Conv-4	49.42 ± 0.78%	68.20 ± 0.66%
RelationNet[3]	Conv-4	50.44 ± 0.82%	65.32 ± 0.70%
MAML[10]	Conv-4	48.70 ± 1.84%	63.11 ± 0.92%
DN4[5]	Conv-4	51.24 ± 0.74%	71.02 ± 0.64%
TapNet[6]	Conv-4	50.68 ± 0.11%	69.00 ± 0.09%
DSN[7]	Conv-4	51.78 ± 0.96%	68.99 ± 0.69%
MADN4[9]	Conv-4	53.20 ± 0.52%	71.66 ± 0.47%
MWNet	Conv-4	53.57 ± 0.23%	72.39 ± 0.16%
PrototypicalNet[2]	ResNet-12	59.25 ± 0.64%	75.60 ± 0.48%
TADAM[4]	ResNet-12	58.50 ± 0.30%	76.70 ± 0.30%
TapNet[6]	ResNet-12	61.65 ± 0.15%	76.36 ± 0.10%
DSN[7]	ResNet-12	62.64 ± 0.66%	78.83 ± 0.45%
MetaOptNet[11]	ResNet-12	62.64 ± 0.61%	78.63 ± 0.46%
CTM[8]	ResNet-18	62.05 ± 0.55%	78.63 ± 0.06%
LEO[12]	WRN-28-10	61.76 ± 0.08%	77.59 ± 0.12%
AWGIM[13]	WRN-28-10	63.12 ± 0.08%	78.40 ± 0.11%
MWNet	ResNet-12	63.26 ± 0.21%	80.03 ± 0.13%

3.5 Fine-grained few-shot classification

In order to explore the performance of the approach proposed in this paper in the task of fine-grained few-shot image classification, the Stanford Dogs[17] dataset is selected and the 5-way 1-shot and 5-shot experiments are performed. For the sake of comparison, Conv-4 with the same size as the related approaches is used as the backbone.

As shown in Table 2, the model in this paper is effective on fine-grained dataset.

Table 2 5-way 1-shot and 5-way 5-shot fine-grained few-shot classification on Stanford Dogs

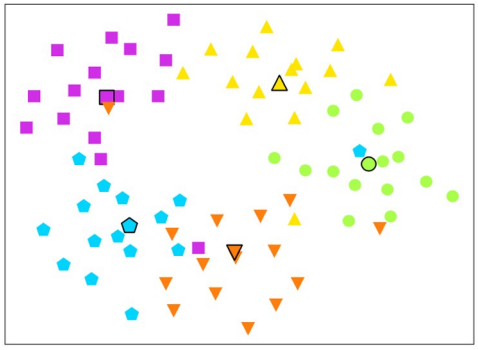
Model	Backbone	5-way 1-shot	5-way 5-shot
MatchingNet[1]	Conv-4	35.80 ± 0.99%	47.50 ± 1.03%
PrototypicalNet[2]	Conv-4	37.59 ± 1.00%	48.19 ± 1.03%
RelationNet[3]	Conv-4	44.49 ± 0.39%	56.35 ± 0.43%
DN4[5]	Conv-4	45.41 ± 0.76%	63.51 ± 0.62%
MADN4[9]	Conv-4	50.42 ± 0.27%	70.75 ± 0.47%
MWNet	Conv-4	50.61 ± 0.29%	70.81 ± 0.32%

In addition, when the feature extraction backbone of the same size is used, compared with related approaches based on metric-learning, the model in this paper has achieved better classification accuracy. Compared with the benchmark approach PrototypicalNet[2], the accuracy is 13.02% and 22.62% higher in 5-way 1-shot and 5-way 5-shot tasks, respectively. Compared with the local descriptor-based DN4[5] and the ap-

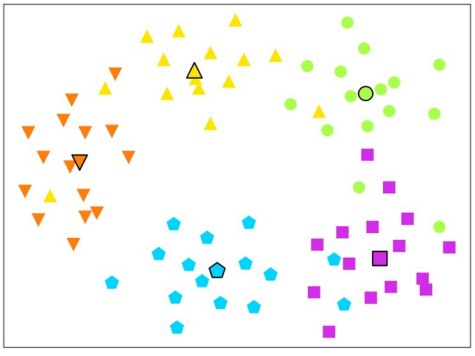
proach MADN4[9] with the attention mechanism added, the model in this paper still has better performance.

3.6 Visualizations of feature space

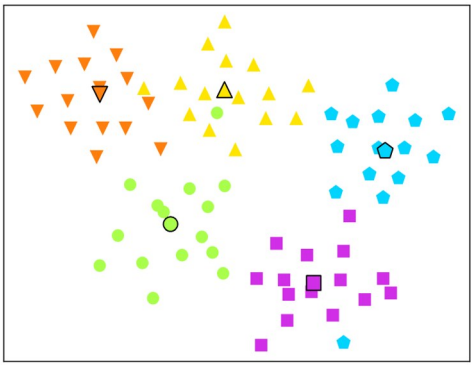
In order to express the multi-attention fusion and weighted class representation in this paper more vividly, in Fig. 6, the relevant feature spaces in the experiments on miniImageNet[1] is subjected to t-SNE visualizations. Conv-4 is used as the backbone, and the PrototypicalNet[2] is re-implemented through the settings in this paper. As shown in Fig. 6, different shapes of graphics represent different types of support samples, there are a total of 5 classes, and the number of support samples for each class is set to 15. The graphic with a black frame represents the characterization of the class. More specifically, Fig. 6(a) represents the feature space of the PrototypicalNet, an ordinary feature space without introducing attention mechanism and weighted class representation. Fig. 6(b) represents the feature space after multi-attention fusion. It can be seen that due to the acquisition of multi-dimensional attention, the support set samples at this time are closer than the original feature space. However, because individual samples deviate from other samples of the same class, the prototype calculated according to the class mean vector is interfered by this sample to a certain extent, which will induce some misclassifications. Fig. 6(c) represents the feature space in which weighted class representations are introduced after multi-attention fusion. At this time, the value of each support set sample is considered, and each sample is assigned a corresponding weight based on the Euclidean distance and Softmax function when obtaining the class representation, so that the weighted class representation can better reflect this class of sample. This largely avoids the occurrence of the misclassification in Fig. 6(b).



(a) Ordinary feature space



(b) Feature space after multi-attention fusion



(c) Feature space with multi-attention and weighted class representation

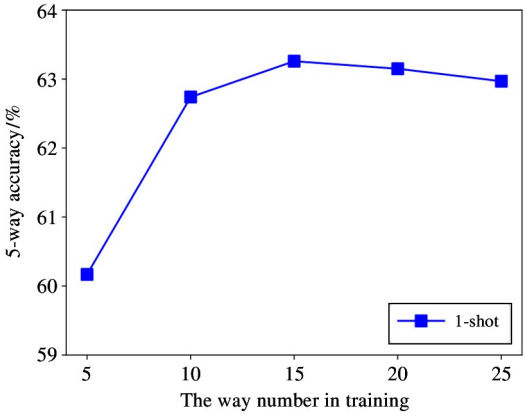
Fig. 6 t-SNE visualization of feature space

3.7 Research on higher way training

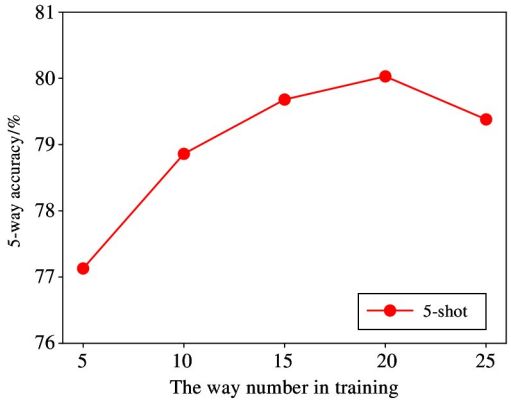
According to previous experience, using a higher number of ways during training, that is, using more support set classes in each episode will make the model obtain a higher classification accuracy. In order to find a more suitable way number for the model in this paper, the FSL experiments with different way number settings is performed on the miniImageNet[1] dataset. In this section, ResNet-12 is used as the feature extraction backbone of the model, other experimental settings remain unchanged, and the number of shots for training and testing is the same. The experimental results are shown in Fig. 7. It can be seen that for the 5-way 1-shot tasks, using the 15-way 1-shot setting during training will obtain a better classification accuracy. And for 5-way 5-shot tasks, using the 20-way 5-shot setting during training will obtain better classification results.

3.8 Ablation study

In order to further verify that the various parts of the work performed in this paper are helpful to improve the classification performance of the model, 5-way 1-shot and 5-way 5-shot few-shot ablation study is conducted in this section on miniImageNet[1]. Considering the relevance to the work of this paper, the selected baseline approach is PrototypicalNet[2] based on metric-learning. And in order to make a better comparison, the experimental data when ResNet-12 is used as

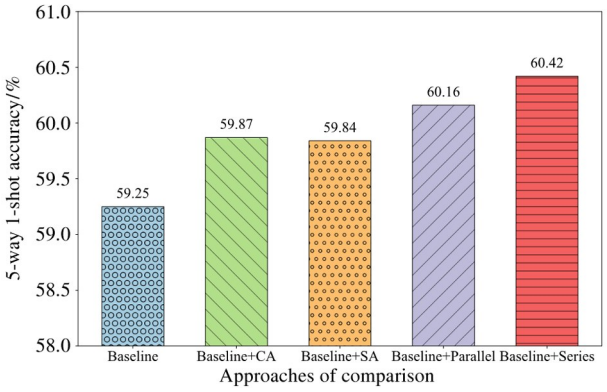


(a) Results on 5-way 1-shot tasks

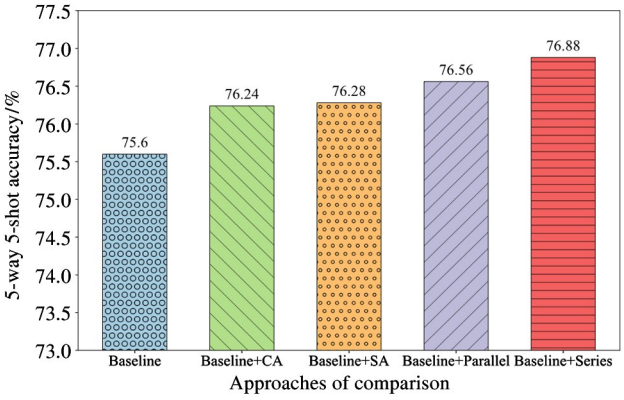


(b) Results on 5-way 5-shot tasks

Fig. 7 Results with different number of ways



(a) Results under 5-way 1-shot setting



(b) Results under 5-way 5-shot setting

Fig. 8 The influence of different attention on miniImageNet

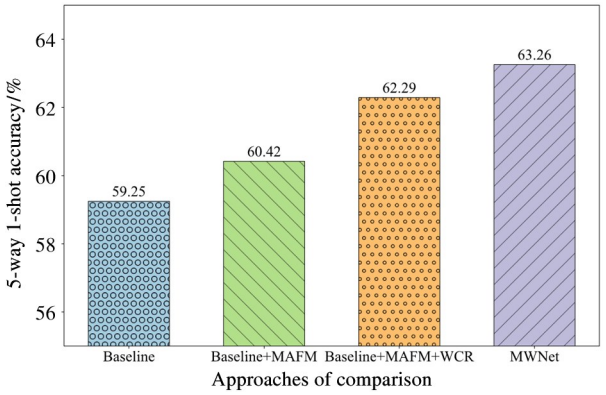
the backbone is selected as a reference, and the relevant experiments are implemented in accordance with the settings in this paper.

First, this section studies the influence of different attention on the performance of the model. For the sake of comparison, only attention is introduced in this part of the experiment. As shown in Fig. 8, the x-axis represents the classification accuracy of the model when only channel attention is introduced, only spatial attention is introduced, channel-spatial attention parallel, and channel-spatial attention series are introduced. It can be seen that under the 5-way 1-shot setting, when only channel attention is introduced, the accuracy is increased by 0.62%; when only spatial attention is introduced, the accuracy is increased by 0.59%; when the channel-spatial attention is connected in parallel, the accuracy is increased by 0.91%; when the channel-spatial attention is connected in series, the accuracy is increased by 1.71%. With the 5-way 5-shot setting, the accuracy of the corresponding four parts is increased by 0.64%, 0.68%, 0.96%, and 1.28% respectively. Considering comprehensively, the way of attention series is chose in the model.

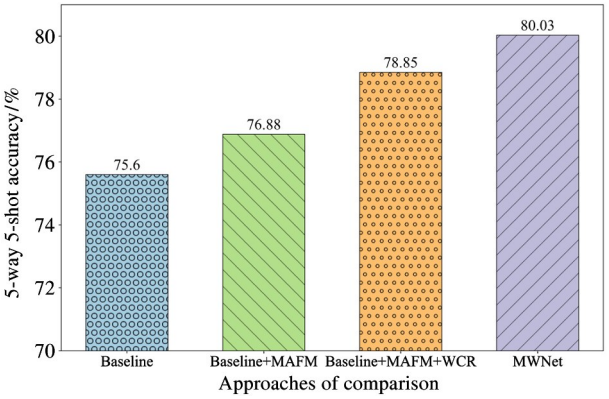
What follows is the rest of the ablation study. As shown in Fig. 9, under the 5-way 1-shot setting, after adding the MAFM, the accuracy of the model increased by 1.17%. Then, after the introduction of WCR, the accuracy increased by 1.87%. After using the mutual similarity metric method, the accuracy is increased by 0.97%. At this time, the accuracy of the model is the highest, that is, MWNet. Under the 5-way 5-shot setting, the accuracy shows the same upward trend, and the accuracy of the corresponding three parts are increased by 1.28%, 1.97%, and 1.18% respectively. Obviously, the various parts of the work in this paper are beneficial to the improvement of the few-shot classification performance. And when the multi-attention fusion module, weighted class representation and mutual similarity metric exist at the same time, the classification accuracy of the model is the highest, which is the MWNet proposed in this paper. With the positive contributions of these beneficial parts, the model in this paper has such excellent performance.

4 Conclusions

In this paper, a simple and efficient few-shot learning model is proposed. Through the cross-spatial and



(a) Results under 5-way 1-shot setting



(b) Results under 5-way 5-shot setting

Fig. 9 Ablation study on the miniImageNet dataset

channel attention acquisition in the feature extraction stage, the extracted features are richer and more discriminative. The importance of each sample is considered based on the Euclidean distance and the Softmax function, where the negative influence of interfering samples is weakened. In the metric phase, information is fused from different angles to obtain a more reliable similarity relationship. A series of experiments on miniImageNet and Stanford Dogs datasets show that the approach proposed in this paper is effective and superior, especially when compared with advanced related technologies, it is highly competitive. Future work will explore the applicability of the model in more problem settings, such as cross-domain and transduction few-shot classification. In addition, a combination of few-shot learning and active learning can also be tried.

References

[1] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 3637-3645

[2] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 4080-4090

[3] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: relation network for few-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1199-1208

[4] ORESHKIN B N, RODRIGUEZ P, LACOSTE A. TADAM: task dependent adaptive metric for improved few-shot learning[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 719-729

[5] LI W, WANG L, XU J, et al. Revisiting local descriptor based image-to-class measure for few-shot learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 7260-7268

[6] YOON S W, SEO J, MOON J. TapNet: neural network augmented with task-adaptive projection for few-shot learning[C]// Proceedings of the International Conference on Machine Learning, Long Beach, USA, 2019: 7115-7123

[7] SIMON C, KONIUSZ P, NOCK R, et al. Adaptive subspaces for few-shot learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 4136-4145

[8] LI H, EIGEN D, DODGE S, et al. Finding task-relevant features for few-shot learning by category traversal[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 1-10

[9] LI H, YANG L, GAO F. More attentional local descriptors for few-shot learning[C]// Proceedings of the International Conference on Artificial Neural Networks, Bratislava, Slovakia, 2020: 419-430

[10] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// Proceedings of the International Conference on Machine Learning, Sydney, Australia, 2017: 1126-1135

[11] LEE K, MAJI S, RAVICHANDRAN A, et al. Meta-learning with differentiable convex optimization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 10657-10665

[12] RUSU A A, RAO D, SYGNOWSKI J, et al. Meta-learning with latent embedding optimization[EB/OL]. https://arxiv.org/pdf/1807.05960.pdf: arXiv, (2019-03-26),[2021-07-20]

[13] GUO Y, CHEUNG N M. Attentive weights generation for few shot learning via information maximization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 13499-13508

[14] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2204-2212

[15] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of the IEEE Confer-

ence on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7794-7803

[16] ZHANG X Y, SHI H, LI C, et al. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision[C]// Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, USA, 2019: 9227-9234

[17] KHOSLA A, JAYADEVAPRAKASH N, YAO B, et al. Novel dataset for fine-grained image categorization: Stanford Dogs[C]// Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Colorado, USA, 2011: 1-2

[18] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. *International Journal of Computer Vision*, 2015, 115(3): 211-252

[19] KINGMA D P, BA J. Adam: a method for stochastic optimization [EB/OL]. https: // arxiv. org/pdf/1412.6980v8. pdf: arXiv, (2015-07-23), [2021-07-20]

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, USA, 2016: 770-778

[21] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[EB/OL]. https: // arxiv. org/pdf/1605. 07146. pdf: arXiv, (2017-06-14), [2021-07-20]

ZHAO Wencang, born in 1973. He received his Ph. D degree in Information Science Department of Ocean University of China in 2005. He also received his B. E. and M. E. degrees from Qingdao University of Science and Technology and Shandong University in 1995 and 2002 respectively. He is mainly engaged in intelligent science and technology, image processing and pattern recognition, cognitive informatics, neural computing, and machine learning.