International Journal of Approximate Reasoning 142 (2022) 417–429

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Few-shot learning based on hierarchical classification via multi-granularity relation networks

Yuling Su [a,b], Hong Zhao [a,b,*], Yaojin Lin [a,b]

[a] School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China
[b] Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou, Fujian, 363000, China

ARTICLE INFO

Article history:
Received 29 June 2021
Received in revised form 2 November 2021
Accepted 21 December 2021
Available online 23 December 2021

Keywords:
Few-shot learning
Hierarchical classification
Multi-granularity
Relation network

ABSTRACT

Few-shot learning is one of the significant areas of machine learning, which aims to recognize novel visual classes from few labeled examples. Many existing models make full use of the similarity of inner-class features and achieve satisfactory results. However, these models assume that classes are independent of each other, ignoring the inter-class relationship. In this paper, we propose a few-shot hierarchical classification model via multi-granularity relation networks (HMRN) considering both the inner-class similarity and inter-class relationship. The multi-granularity relationship among coarse- and fine-grained classes is an important auxiliary information in the class hierarchical structure originated from data. Thus, we first extract hierarchical features of different granularity classes according to the membership relationship among the classes. Second, we build multi-granularity relation networks to obtain the inner-class similarity relation of different granularity classes using the hierarchical features. Finally, we consider the tradeoff among the inner-class similarity relation of different granularity classes for hierarchical few-shot learning, which takes the information of coarse-grained classes to assist the learning of fine-grained classes. Experimental results show that our model outperforms several state-of-the-art flat (without hierarchical structure) models and hierarchical models. For example, the accuracy of HMRN is about 3.00% better than that of flat models on the *tieredImageNet* dataset.

1. Introduction

Few-shot learning is one of the major challenges to machine learning because it is difficult to get enough training data due to privacy, security and other factors [8,15]. The lack of training data seriously limits the generalization ability of traditional deep learning methods for some objects such as rare animals [3,30,38,42]. Inspired by human intelligence, few-shot learning emerges as the times require, which aims to learn few labeled samples to identify new classes [16]. In recent years, few-shot learning models are being well widely applied to various research fields such as computer vision [1,20], audio and speech [4,27], natural language processing [25,44], and data analysis [7,11].

Metric learning is one of the main methods of few-shot learning [24]. The general object of metric learning is to learn a pair-wise similarity metric among query and support samples [13]. The more similar samples have higher similarity, while dissimilar samples have lower similarity [43]. Metric learning can be divided into two types. The first method utilizes

* Corresponding author at: School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China.
E-mail address: hongzhaocn@163.com (H. Zhao).

https://doi.org/10.1016/j.ijar.2021.12.013
0888-613X/

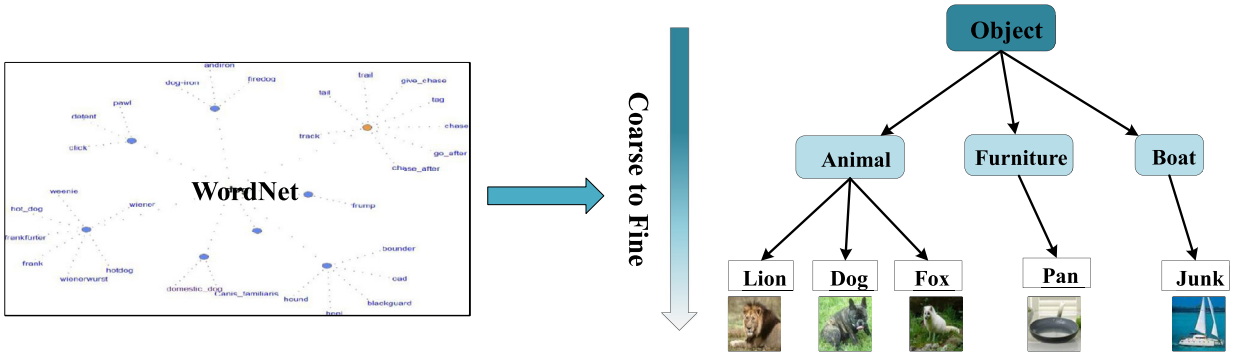

Fig. 1. An example of the hierarchical structure.

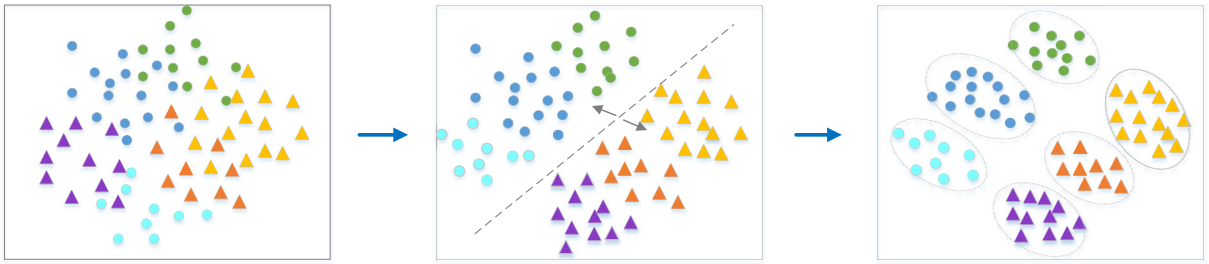

Fig. 2. A case to explain the basic idea of HMRN. Shapes represent different coarse-grained classes, while different colors represent different fine-grained classes. (For interpretation of the colors in the figure, the reader is referred to the web version of this article.)

the distance formula to measure the similarity between two samples. For instance, Vinyals et al. [37] adopted the Cosine distance to measure the similarity between the features of the query and support samples. Similarly, Snell et al. [35] utilized a prototypical network to obtain the prototypes of each class for support samples and leveraged the Euclidean distance to calculate the distance among query sample features and the prototypes.

The second metric learning method is to study a learnable distance module through neural network. Sung et al. [36] firstly established a relation network to study a learnable distance module between the similar features of query and support samples. On the basis of the relation network, Hui et al. [12] embedded a self-attention mechanism to increase the learning ability of similar features. Similarly, Liu et al. [23] proposed an embedded graph relation network, which can improve the quality of the similar features within a class. While He et al. [10] aggregated the information of similar samples by carrying out weighted information dissemination to enhance the similar features. Unlike the previous mentioned single-scale relation networks, Ding et al. [6] developed a multi-scale relation network to extract the multi-scale similar features of query and support samples.

The above-mentioned models make full use of the similarity of inner-class features and achieve effective results. However, these models treat each class independently, ignoring the relationship of inter classes. Actually, classes are generally not independent of each other, and there is a class hierarchical structure among them [17,29,40]. Fig. 1 gives an example of the class hierarchical structure with five fine-grained classes and three coarse-grained classes, which is constructed by the semantic subordination and dependency among the classes in WordNet [9,26,46]. In the example, *Lion*, *Dog* and *Fox* are fine-grained classes and belong to the coarse-grained class *Animal*. The class hierarchical structure is one of the important auxiliary information for image classification [34,39,41].

In this paper, we develop a hierarchical few-shot learning model via multi-granularity relation networks (HMRN) considering both the inner-class similarity and inter-class relationship in the class hierarchical structure. Fig. 2 gives an intuitive understanding of the inner-class similarity among fine-grained classes and the inter-class relationship among coarse- and fine-grained classes. In this figure, different shapes represent different coarse-grained classes and different colors represent different fine-grained classes. From Fig. 2 (a) to Fig. 2 (c), the fine-grained classes belonging to the different coarse-grained classes are more farther and the samples belonging to the same fine-grained class are more closer. We aim to leverage the inner-class similarity and inter-class relationship of the class hierarchical structure to expand the inter-class distance and reduce the inner-class distance.

The main contributions of this paper are threefold:

(1) We propose a hierarchical few-shot learning model with class hierarchical structure considering both the inner-class similarity and the relationship of inter classes, which can reduce the inner-class distance and expand the inter-class

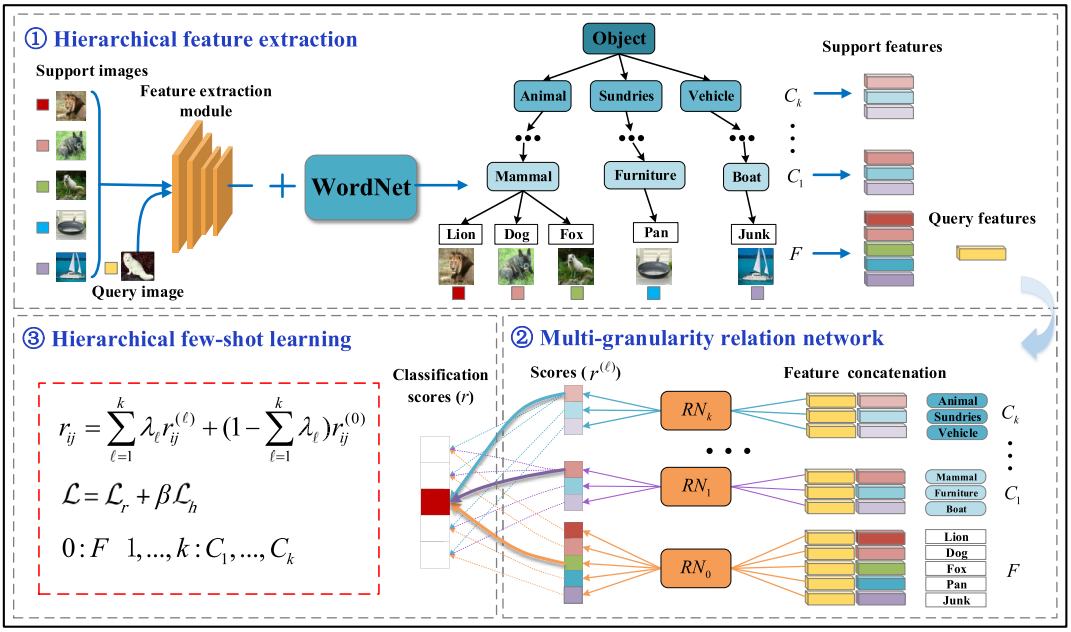

Fig. 3. Framework of the HMRN model in 5-way 1-shot setting. F means the fine-grained class and $C_1, ..., C_k$ means different coarse-grained classes. In the second and third parts, different color arrows represent different granularity classification processes and RN_ℓ are the different granularity relation networks, where $\ell = 0, 1, ..., k$.

distance. It is particularly important for few-shot learning to make use of all of the information available from examples since very few labeled samples are available for each task.

(2) We adopt a hierarchical relation score strategy to combine the relation scores of different granularity classes for few-shot classification, which takes the information of coarse-grained classes to assist the learning of the fine-grained classes.

(3) We establish a multi-object loss function to obtain the optimal model considering the classification of different granularity classes, which promotes the classification ability by improving the classification of each granularity class.

Experimental results show that HMRN, on the one hand, is comparable with several advanced models on the *Omniglot* dataset [14] without class hierarchical structure. In particular, HMRN achieves a satisfactory performance with 98.95% and 99.41% in 5-way 1-shot and 5-shot episodes, which is almost equal to some advanced flat models. On the other hand, HMRN also achieves an effective performance on the *tieredImageNet* dataset [32] with class hierarchical structure. For example, HMRN gets about 3.70% and 3.30% improvements over the baseline relation network [36] in 5-way 1-shot and 5-shot episodes, and outperforms the hierarchical model MNE [18] by about 1.10% in 5-way 5-shot episode.

The rest of the paper is organized as follows. Section 2 presents the details of the proposed model. Section 3 introduces the experimental setup, including datasets, comparison models and experimental settings. Section 4 reports and analyses the experimental results. Finally, we provide the main conclusions and ideas for a further study in Section 5.

2. HMRN model

In this section, we introduce the framework and specific classification steps of HMRN in detail.

2.1. Framework overview

We design a few-shot hierarchical classification model with multi-granularity relation networks according to the class hierarchical structure. We commit to taking the coarse-grained class knowledge to assist the fine-grained class learning. The framework of HMRN is designed as shown in Fig. 3.

The process of hierarchical classification based on multi-granularity relation networks mainly consists of the following three stages:

(1) First, we extract the hierarchical features of the support and query sets according to the class hierarchical structure.

(2) Second, we establish multi-granularity relation networks to obtain the hierarchical relation scores of each granularity class in the class hierarchical structure.

(3) Finally, we integrate the hierarchical relation scores for hierarchical few-shot learning and establish a multi-object loss function to obtain the optimal model.

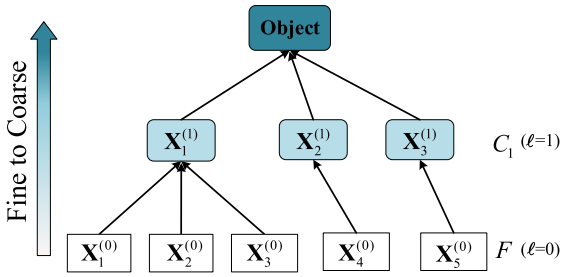

Fig. 4. An example of hierarchical feature extraction (5-way 1-shot, $k = 1$).

2.2. Hierarchical feature extraction

In the first step, we extract hierarchical features according to the membership relationship of different granularity classes in the class hierarchical structure. We adopt ResNet12 as the network backbone of the feature extraction module inspired from [22,28,31].

For every few-shot learning task, we simulate the few-shot learning setting through the episode based training [35–37]. Every training episode is formed by support set and query set. Select C classes with K labelled samples from each of C classes randomly from the training set serves as the support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m}$ $(m = C \times K)$ and select T samples of the remainder of the C classes serves as the query set $\mathcal{Q} = \{(x_j, y_j)\}_{j=1}^{n}$ $(n = C \times T)$. The training set has its own label space, which does not intersect the valid/test set. The support set and query set share the same label space. The formulation mentioned above is called C-way K-shot few-shot classification problem. Based on the episode training strategy, we consider 5-way 1-shot and 5-shot episodes in our experiments.

In each training episode, we define query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{n}$ and support set $\mathcal{S} = \{(x_j, y_j)\}_{j=1}^{m}$, where n, m represent the number of samples in the query and support sets respectively. The class hierarchical structure consists of one fine-grained class layer and k $(k \geq 1)$ coarse-grained class layers. Let ℓ be the layer of different granularity classes in the class hierarchical structure, which ranges from 0 to k. We define $\ell = 0$ as the fine-grained class layer and $0 < \ell \leq k$ as the ℓ^{th} coarse-grained class layer. Then we take $\mathbf{X}^{(\ell)}$ as the feature of sample x at the ℓ^{th} class hierarchical structure layer.

We assume the feature extraction module as f_ϕ. Let x_i and x_j be the samples from the query and support sets. First, we obtain the fine-grained features $\mathbf{X}_i^{(0)} = f_\phi(x_i)$ and $\mathbf{X}_j^{(0)} = f_\phi(x_j)$. Then, we get the coarse-grained features $\mathbf{X}_j^{(\ell)}$ of the support sample x_j according to the class hierarchical structure, where $\ell = 1, \ldots, k$. The feature of a coarse-grained class is the mean of the features of nearest fine-grained classes belonging to it. We utilize the fine-grained feature $\mathbf{X}_i^{(0)}$ of the query sample x_i in our experiment.

For example, the class hierarchical structure has one fine-grained class layer and one coarse-grained class layer $(k = 1)$ in 5-way 1-shot episode as shown in Fig. 4. The different granularity features of it are computed as

$$\begin{cases} \mathbf{X}_1^{(0)} = f_\phi(x_1), \mathbf{X}_2^{(0)} = f_\phi(x_2), \mathbf{X}_3^{(0)} = f_\phi(x_3), \mathbf{X}_4^{(0)} = f_\phi(x_4), \mathbf{X}_5^{(0)} = f_\phi(x_5) \\ \mathbf{X}_1^{(1)} = \dfrac{1}{3}(\mathbf{X}_1^{(0)} + \mathbf{X}_2^{(0)} + \mathbf{X}_3^{(0)}),\ \mathbf{X}_2^{(1)} = \mathbf{X}_4^{(0)},\ \mathbf{X}_3^{(1)} = \mathbf{X}_5^{(0)}. \end{cases}$$

2.3. Multi-granularity relation network construction

The second step is to build multi-granularity relation networks and compute hierarchical relation scores among the query and support samples of different granularity classes. Relation score is used to measure the similarity between two samples, which is calculated through the relation network [36]. We adopt six-layer ResNet to instead of four-convolution-block as the backbone network of relation network. The relation network also has two fully-connected layer, while the output of the second fully connected layer is Softmax to generate a reasonable range of relation score in our network architecture.

We assume $\mathcal{C}(\cdot, \cdot)$ as the concatenation of two features by channel and the multi-granularity relation networks in ℓ^{th} granularity class layer as g_{θ_ℓ}, where $\ell = 0, 1, \ldots, k$. Samples x_i, x_j in the query and support sets are fed into the feature extraction module f_ϕ, which produces fine-grained features $\mathbf{X}_i^{(0)} = f_\phi(x_i)$ and $\mathbf{X}_j^{(0)} = f_\phi(x_j)$. According to the class hierarchical structure, we get different granularity features of sample x_j, denoted as $\mathbf{X}_j^{(\ell)}$, where $\ell = 0, 1, \ldots, k$.

Then, we respectively concatenate the query fine-grained feature $\mathbf{X}_i^{(0)}$ with the support multi-granularity features $\mathbf{X}_j^{(\ell)}$ as feature-pairs $\{\mathcal{C}(\mathbf{X}_i^{(0)}, \mathbf{X}_j^{(\ell)})\}_{\ell=0}^{k}$. Next, we take these feature-pairs into the multi-granularity relation networks g_{θ_ℓ} to obtain hierarchical relation scores in different granularity class layers. Relation score is used to measure the similarity between two samples. The closer the samples are, the higher the relation score is. On the contrary, the lower the relation score is, the more different the two samples are. The hierarchical relation scores are calculated followed as

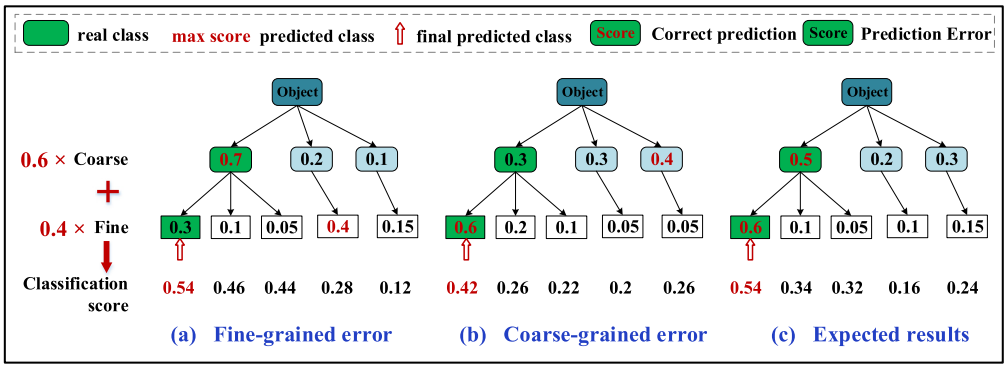

Fig. 5. Three cases of correct prediction (5-way 1-shot, $k = 1$). The classification scores are calculated by Eq. (2), where $\ell = 1$ and $\lambda_\ell = 0.6$. The class with the largest score is the predicted class. The final predicted class is based on the maximum classification score.

$$\begin{cases} r_{ij}^{(0)} = g_{\theta_0}(\mathcal{C}(\mathbf{X}_i^{(0)}, \mathbf{X}_j^{(0)})) \\ r_{ij}^{(\ell)} = g_{\theta_\ell}(\mathcal{C}(\mathbf{X}_i^{(0)}, \mathbf{X}_j^{(\ell)})), \quad \ell = 1, ..., k, \end{cases} \tag{1}$$

where $r_{ij}^{(\ell)}$ means the relation score of query sample x_i and support sample x_j at the ℓ^{th} granularity class layer. It ranges from 0 to 1. The higher the similarity is, the closer the relation score is to 1.

2.4. Few-shot learning based on hierarchical classification

The last step, we take the few-shot hierarchical classification according to the hierarchical relation scores obtained in the previous step. We combine the hierarchical relation scores to get the classification score r_{ij} between query sample x_i and support sample x_j as followed

$$r_{ij} = \sum_{\ell=1}^{k} \lambda_\ell \, r_{ij}^{(\ell)} + (1 - \sum_{\ell=1}^{k} \lambda_\ell) \, r_{ij}^{(0)}, \tag{2}$$

where λ_ℓ is a positive constant. The value of λ means the influence of each coarse-grained class on the fine-grained class learning. We leverage the information of coarse-grained classes to assist the classification of fine-grained classes. Finally, we predict the label of query sample x_i by the classification score. The class of the support samples with the largest classification score is the query prediction class.

In addition, we construct a multi-object loss function based on Mean Square Error. Firstly, we regress the classification score computed as Eq. (2) to the ground truth (the correct prediction is 1 and the error is 0):

$$\mathcal{L}_r = \sum_{i=1}^{n} \sum_{j=1}^{m} (r_{ij} - \mathbf{1}\,(y_i == y_j))^2, \tag{3}$$

where $\mathcal{L}_r$ is the classification loss, r_{ij} means the classification score between query sample x_i and support sample x_j.

The main idea of our model is to combine the hierarchical relation scores of all granularity classes for classification prediction. As shown in Fig. 5, there are three cases of prediction: (a) the prediction of coarse-grained class is right and the fine-grained class is wrong; (b) the prediction of coarse-grained class is wrong and the fine-grained class is right; and (c) the predictions of both coarse- and fine-grained classes are right. The final predictions of the three cases are correct, and there are still two types of error in Figs. 5(a) and (b). While the final prediction and the prediction of each granularity class are right in Fig. 5(c), which is what we expect. Therefore, we calculate the loss of each granularity class, called hierarchical classification loss $\mathcal{L}_h$:

$$\mathcal{L}_h = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{\ell=0}^{k} (r_{ij}^{(\ell)} - \mathbf{1}\,(y_i^{(\ell)} == y_j^{(\ell)}))^2, \tag{4}$$

where $y_i^{(\ell)}$ and $y_j^{(\ell)}$ are the labels of different granularity classes of samples x_i and x_j.

In addition, losses $\mathcal{L}_r$ and $\mathcal{L}_h$ interact with each other, which are not existing independently. The relation scores of different granularity classes affect the final prediction results. Conversely, the final predicted results also affect the predicted result of each granularity class through parameter back propagation.

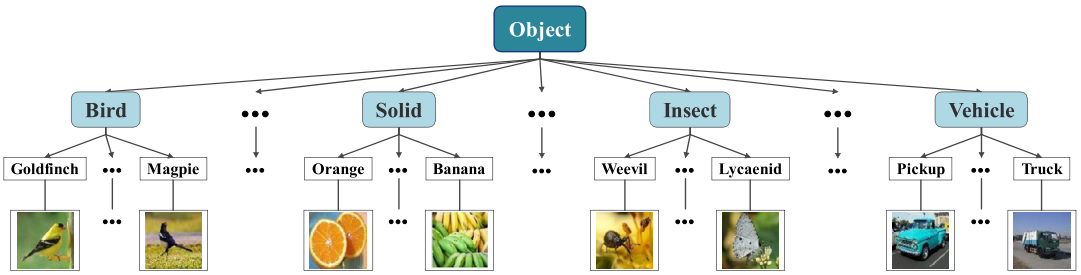

Fig. 6. The class hierarchical structure of the *tieredImageNet* dataset.

Hence, we synthesize the classification loss and hierarchical classification loss to form our multi-object loss function $\mathcal{L}$ by Eqs. (3) and (4):

$$\mathcal{L} = \mathcal{L}_r + \beta \mathcal{L}_h, \tag{5}$$

where $\mathcal{L}_r$ is the classification loss of prediction, $\mathcal{L}_h$ is the loss of different granularity classes, and β is a weight factor to tradeoff $\mathcal{L}_r$ and $\mathcal{L}_h$.

Finally, we use the SGD optimizer to minimize the loss and optimize the parameters ϕ and θ_ℓ of the feature extraction module f_ϕ and multi-granularity relation networks g_{θ_ℓ}, where $\ell = 0, 1, ..., k$. Algorithm 1 provides the pseudo code for the model training process. We firstly construct the support and query sets and extract the hierarchical features, including the fine-grained and coarse-grained features in lines 3-4. The main training progress is listed in lines 2-17. In particular, we compute the hierarchical relation and classification scores from line 7 to 10. Then, we update the classification loss in lines 11-13. Finally, we update the parameters ϕ and θ_ℓ by the loss back propagation in line 16.

Algorithm 1 Few-shot Learning Based on Hierarchical Classification via Multi-granularity Relation Networks (HMRN).

Input: Training set $\mathcal{D} = \{(x_1, y_1), ..., (x_i, y_i), ..., (x_M, y_M)\}$, where $y_i \in \{1, 2, ..., N\}$, and M, N mean the number of samples and classes of the training set. The number of total training epochs is E_{max}. For every training epoch, parameter C means the number of classes per episode, K is the number of samples for the support set and T is the number of samples for the query set. The feature extraction module is f_ϕ and multi-granularity relation networks are g_{θ_ℓ}, where $\ell = 0, 1, ..., k$.

Output: The parameters ϕ and θ_ℓ.

1: Initialize parameters ϕ and θ_ℓ;
2: **for** $epo = 1 : E_{max}$ **do**
3: Select randomly C from N classes and K, T samples of each class to construct the query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{n}$ and support set $\mathcal{S} = \{(x_j, y_j)\}_{j=1}^{m}$ respectively, where $n = C \times T$ and $m = C \times K$;
4: Extract the fine-grained and coarse-grained features of $\mathcal{Q}$ and $\mathcal{S}$ by f_ϕ;
5: **for** $i = 1 : n$ **do**
6: **for** $j = 1 : m$ **do**
7: **for** $\ell = 0 : k$ **do**
8: Use g_{θ_ℓ} to obtain the hierarchical relation scores $r_{ij}^{(\ell)}$ by Eq. (1);
9: **end for**
10: Obtain the classification score r_{ij} by Eq. (2);
11: Update the classification loss $\mathcal{L}_r$ by Eq. (3);
12: Update the hierarchical classification loss $\mathcal{L}_h$ by Eq. (4);
13: Update the loss $\mathcal{L}$ by Eq. (5);
14: **end for**
15: **end for**
16: Update parameters ϕ and θ_ℓ by loss $\mathcal{L}$ back propagation;
17: **end for**

3. Experimental setup

3.1. Datasets

In our experiments, we use two datasets, including *Omniglot* [14] and *tieredImageNet* [32]. The *tieredImageNet* has a class hierarchical structure by itself and the class hierarchical structure of *Omniglot* is defined by ourselves. Both of them are composed of two granularity class layers. There are one coarse-grained class layer ($k, \ell = 1$) and one fine-grained class layer. The class hierarchical structure of *tieredImageNet* is shown in Fig. 6. For example, the layer of *Bird*, *Fruits*, *Insect* and *Vehicle* (coarse-grained class) is the coarse-grained class layer, while the classes *Goldfinch* and *Orange* (fine-grained class) belong to fine-grained class layer.

Basic statistics for the two datasets are provided in Table 1, including the number of coarse and fine-grained classes, and the size of the image.

Table 1
Dataset description.

	Omniglot			*tieredImageNet*			
	training	test	all	training	valid	test	all
coarse-grained class	38	12	50	20	6	8	34
fine-grained class	1263	360	1623	351	97	160	608
size	28×28			$84 \times 84 \times 3$			

Omniglot is composed of 32,460 images with 1,623 handwritten characters from 50 different alphabets [14]. Each handwritten character contains 20 samples drawn by different people. We define the 50 alphabets as coarse-grained classes and the 1,623 handwritten characters as fine-grained classes. We select randomly 38 coarse-grained classes with 1,263 classes as the training set and the rest as the test set. What's more, we augmented images through 90°, 180° and 270° rotations of existing data. The size of each image is 28×28.

tieredImageNet is a subset of the ILSVRC-12 [33] dataset containing more than 700,000 images in total [32]. According to the class hierarchical structure of *Imagenet* [5], there are 608 fine-grained classes and 34 coarse-grained classes. Furthermore, the dataset is further divided into a training set with 20 coarse-grained classes and 351 fine-grained classes, a valid set with 6 coarse-grained classes and 97 fine-grained classes and a test set with 8 coarse-grained classes and 160 fine-grained classes. The size of the picture is unified as $84 \times 84 \times 3$.

3.2. Comparison models

We compare HMRN with several flat models without class hierarchical structure (the first eighth models) and hierarchical models (the ninth to eleventh models). The details of them are introduced as follows:

(1) Matching network [37] is one of the classical few-shot learning models, which adopted the Cosine distance to measure the similarity among support and query samples.

(2) Prototypical network [35] automatically generated the prototypes of each class in support set, and then leveraged the Euclidean distance to measure the distance among query samples and these prototypes.

(3) Relation network [36] transformed distance measurement into a learnable distance module, and defined the relationship among query and support samples as relation score.

(4) MAML [7] aimed to train the initial parameters of the model, so that the model would perform new tasks well after only one or several steps of gradient descent update.

(5) SARN [12] embedded self-attention module into the relation network to enhance the learned features.

(6) Prototype-relation nets (PRN) [21] minimized intra class distance and maximized inter class distance to classify.

(7) IMP [2] proposed infinite mixture prototypes to adaptively represent both simple and complex data distributions for few-shot learning.

(8) CovaMNet [19] was designed to exploit both the covariant representation and metric based on the distribution consistency for few-shot classification tasks.

(9) TPN [22] learned to propagate labels from labeled instances to unlabeled test cases by learning to utilize the graph construction module of manifold structure in data.

(10) MNE [18] utilized graph neural network and considered the neighborhood information to enhance the feature information according to the graph tree structure.

(11) C2F w/BDE-MetaBL [45] developed a coarse-to-fine pseudo-labeling process to construct pseudo-tasks from coarsely-labeled data and grouped each coarse-grained class into pseudo-fine-classes by similarity matching.

3.3. Experiment settings

In our experiment, we take the training episode C-way K-shot strategy [37], including 5-way 1-shot and 5-way 5-shot. For every training episode, the K-shot has 10 query samples per each of the C sampled classes besides the K samples of support set. This means for example that there are $10 \times 5 + 5 \times 5 = 75$ samples in one training episode for 5-way 5-shot experiment. The optimizer is SGD with momentum 0.9, learning rate 0.1, and reduces the learning rate by half for every 100,000 episodes. In test settings, we conduct few-shot classification on 10 epochs with 1,000 randomly episode samples from test set and compute the mean accuracy together with the 95% confidence interval. Our experiments are implemented in Pytorch with a GeForce RTX 2080 Ti Nvidia GPU card. The basic data and code for this study have been uploaded to GitHub and can be accessed via the following link: https://github.com/fhqxa/HMRN.

4. Experimental results and analysis

We take five experiments to prove the effectiveness of HMRN: (1) we analyze the sensitivity of parameter β on multi-object loss function; (2) we exploit the effectiveness of the hierarchical relation scores; (3) we perform the contributions of different strategies of HMRN; (4) we compare HMRN with the variant of relation network; (5) we visualize the performance of the HMRN model; and (6) we compare HMRN with several flat and hierarchical models.

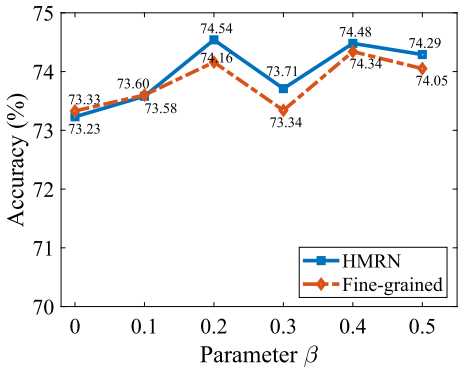

Fig. 7. Performance of parameter β in multi-object loss function on the *tieredImageNet* dataset (5-way 5-shot).

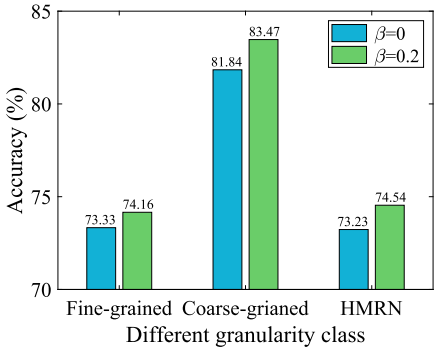

Fig. 8. Performance comparison of parameter $\beta = 0$ and $\beta = 0.2$ on the *tieredImageNet* dataset (5-way 5-shot).

4.1. Performance of parameter β on multi-object loss function

In this section, we discuss the effectiveness of the multi-object loss function and explore the influence of parameter β. We fix $k = 1$, $\lambda = 0.6$ and parameter β is selected from the candidates {0,0.1,0.2,0.3, 0.4,0.5}.

The experimental results of different β values on multi-object loss function are shown in Fig. 7, and we can obtain the following observations:

(1) The value of parameter β represents the weight of the loss of each granularity class in the multi-object loss function. As shown in Fig. 7, the accuracy of HMRN is best when $\beta = 0.2$, which is about 1.30% higher than that in $\beta = 0$. Therefore, $\beta = 0.2$ is selected as the best coefficient in this paper.

(2) The accuracy of HMRN is 0.40% higher than that of the fine-grained class without the assistance of the coarse-grained class when $\beta = 0.2$. It proves that the coarse-grained class information can promote the fine-grained class learning.

The performance comparison of $\beta = 0$ and $\beta = 0.2$ is shown in Fig. 8. As shown in Fig. 8, the accuracy of fine-grained class in $\beta = 0.2$ is 0.83% higher than that in $\beta = 0$, while the accuracy of coarse-grained class is 1.63% better. That proves that multi-object loss function can alleviate the problem mentioned in Section 2.4. The multi-object function can correct the classification of each granularity class to improve the classification ability of HMRN, which improves the accuracy of HMRN from 73.23% to 74.54%.

4.2. Performance of the hierarchical relation scores

In this section, we analyze the effectiveness of the strategy of hierarchical relation scores to explore the role of coarse-grained class information in the fine-grained class learning. Firstly, we exploit the influence of different λ values, the following comparative experiments under the *tieredImageNet* dataset are conducted. We define $k = 1$, then Eq. (2) degenerates to $r_{ij} = \lambda\ r_{ij}^{(1)} + (1 - \lambda)\ r_{ij}^{(0)}$. Parameter λ controls the degree of the influence of coarse-grained class learning on fine-grained class learning, while the attention weight of fine-grained class learning is $1 - \lambda$. We fix parameter $\beta = 0.2$ and the parameter pair $(\lambda, 1 - \lambda)$ is selected from the candidates {(0.3,0.7),(0.4,0.6),(0.5,0.5),(0.6,0.4),(0.7,0.3)} with the best classification result. In addition, we also consider the case that coarse-grained class learning is not involved when $\lambda = 0$.

The accuracies of different λ on *tieredImageNet* are listed in Table 2, and we observe the followings:

(1) Parameter $\lambda = 0$ means that coarse-grained class does not participate in fine-grained class learning. The accuracy of HMRN in $\lambda = 0.5$ is about 0.66% higher than that in $\lambda = 0$. Also, the accuracy of HMRN is 74.70%, which is 0.30% better

Table 2
Performance of different λ in different granularity layers on the *tieredImageNet* dataset (%) (5-way 5-shot). The best performance is highlighted.

λ	$1-\lambda$	Fine-grained	Coarse-grained	HMRN
0.3	0.7	74.06	83.01	74.33
0.4	0.6	74.17	82.83	74.48
0.5	0.5	**74.39**	83.10	**74.70**
0.6	0.4	74.16	**83.47**	74.54
0.7	0.3	73.86	82.37	73.50
0	1	74.04	83.12	74.04

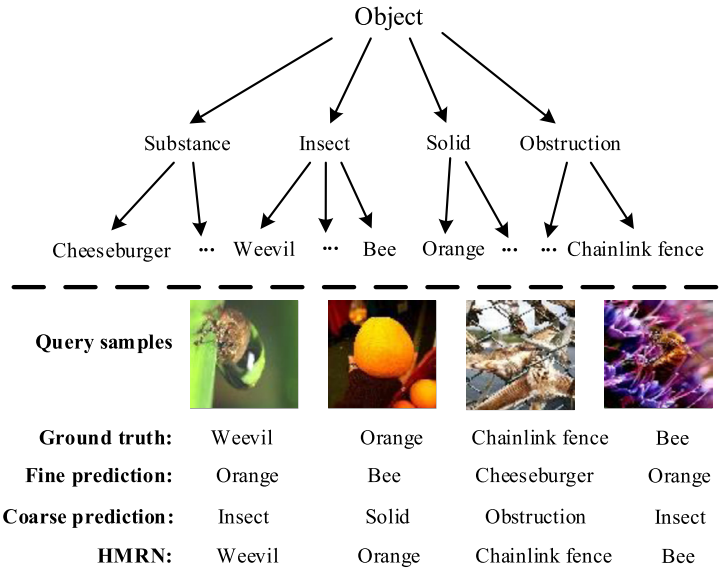

Fig. 9. Predictions on some samples of a random test task on the *tieredImageNet* dataset. The class tree is a subtree of the *tieredImageNet* dataset.

Table 3
Contributions of different strategies on the *tieredImageNet dataset*. The best performance is highlighted.

Hierarchical relation scores	Multi-object loss function	5-way	
		1-shot	5-shot
	✓	54.02	73.54
✓		55.51	74.04
✓	✓	**57.98**	**74.70**

that of fine-grained class when $\lambda = 0.5$. It appears that the coarse-grained class learning can promote the fine-grained class learning as choosing the appropriate parameter λ. Therefore, $\lambda = 0.5$ is selected as the best coefficient in this paper.

(2) The accuracy of coarse-grained class in $\lambda = 0.3$ is almost equal to that in $\lambda = 0.5$, while the accuracy of HMRN is 0.40% lower than that in $\lambda = 0.5$. The participation of coarse-grained class learning is lower, the help of coarse-grained class learning on fine-grained class learning is smaller.

(3) The accuracy of coarse-grained class in $\lambda = 0.6$ is about 0.40% better than that in $\lambda = 0.5$, while the accuracy of HMRN is lower than that in $\lambda = 0.5$. The model may pay more attention to the coarse-grained class learning and ignore the fine-grained class learning when the value of λ is larger.

Then, we randomly select a test task to intuitively show more insights of the coarse-grained class assistance to the fine-grained class learning. From Fig. 9, the prediction errors of the original fine-grained classes have been corrected with the help of the coarse-grained class auxiliary information. It seems that the combining of the different granularity class relationship plays a positive role in HMRN.

4.3. Ablation study

In this section, we show that both strategies of the hierarchical relation scores and multi-object loss function of HMRN facilitate the model performance. We consider three cases: 1) using the hierarchical relation score strategy; 2) using the multi-object loss function strategy; and 3) using both of the two strategies (HMRN). For the fairness of the experiment, we change the strategy used and other settings are the same.

Table 4
Performance comparison with variant of relation network on the *tieredImageNet* dataset (%). The best performance is highlighted.

Model	Hierarchical	5-way	
		1-shot	5-shot
RN [36]	N	54.26 ± 0.27	71.34 ± 0.28
HRN	Y	55.65 ± 0.34	72.41 ± 0.30
HMRN	Y	$\mathbf{57.98 \pm 0.26}$	$\mathbf{74.70 \pm 0.24}$

The main results are listed in Table 3 and we can obtain the following conclusions:

(1) The accuracies are 55.51% and 74.04% by the strategy of hierarchical relation scores, which are about 1.50% and 0.50% higher than that of the multi-object loss function strategy in 1-shot and 5-shot episodes, respectively.

(2) The strategy of hierarchical relation scores plays a greater role than the multi-object loss function. More specifically, the strategy of multi-object loss function improves the accuracies from 55.51% and 74.04% to 57.98% and 74.70% in 5-way 1-shot and 5-shot episodes, respectively. While the hierarchical relation score strategy improves the accuracy from 54.02% to 57.98% in 5-way 1-shot episode and from 73.54% to 74.70% in 5-way 5-shot episode.

(3) The contribution of the two strategies in 1-shot episode is greater than that in 5-shot episode. In particular, the accuracy improvements of these two strategies on 1-shot are about 3.95% and 2.45%, while the improvements in 5-shot are about 1.20% and 0.70%. HMRN consistently makes more gains for 1-shot learning in comparison to its 5-shot counterpart, which indicates that the class hierarchical structure is more helpful when training data is extremely scarce.

4.4. Performance comparison with variant of relation network

In this section, we compare HMRN with the relation network (RN) [36] and its variant (with hierarchical classification, HRN) to prove the hierarchical classification can incur well performance. We respect the original and do not change other settings for relation network. According to the hierarchical classification strategy proposed in this paper, we embed it into the relation network.

The experimental results are listed in Table 4, and we can obtain the following conclusions:

(1) From Table 4, the accuracies of RN are improved by 1.39% and 1.07% in 1-shot and 5-shot episodes with hierarchical classification. That proves that the hierarchical classification proposed in this paper is effective. The information of coarse-grained classes can be available to assist the learning of fine-grained classes.

(2) On the one hand, compared with RN, the accuracies of HMRN are improved by 3.72% and 3.36% in 1-shot and 5-shot episodes, respectively. On the other hand, the accuracies of HMRN are 2.33% and 2.31% better than those of the variant of relation network HRN.

4.5. Visualization of the HMRN model

In this section, we visualize the performance of HMRN. We randomly sample two test tasks with 80 query samples per each class under the 5-way 5-shot and 1-shot settings respectively. We apply them to three models including RN (relation network) [36], HRN (relation network with the strategy of HMRN) and HMRN. HRN and HMRN both consider the relationship of inter classes. We use t-SNE to visualize their performance.

The visualization results of the three models are shown in Fig. 10 and we can obtain the following conclusions:

(1) For 1-shot setting, the inner-class distances are closer and the distances among inter classes are farther. Classes 0 and 2, and 1 and 3 close to each other because they both belong to the same coarse-grained class from Fig. 10(a) to Fig. 10(c).

(2) For 5-shot setting, compared with Fig. 10(d), the inner-class distances are more closer and the inter-class distances are more farther, as shown in Fig. 10(f). The fine-grained classes belonging to the same coarse-grained class are closer. Classes 0 and 2, 1 and 4 close to each other because they belong to the same coarse-grained class.

(3) From Fig. 10, HMRN is more discriminative for each coarse-grained class. The fine-grained classes belonging to the same coarse-grained class gather together relatively. The distances among fine-grained classes belonging to different coarse-grained classes are farther.

4.6. Performance comparison of different models

In this section, we compare HMRN with several state-of-the-arts few-shot learning models on two datasets *Omniglot* and *tieredImagenet*, including flat models without considering the class hierarchical structure and hierarchical models considering the class hierarchical structure. The accuracies of the comparison models are copied from their original papers except for the relation network [36].

The *Omniglot* dataset has not class hierarchical structure by itself, which is defined by ourselves. We assume the 50 alphabets as coarse-grained classes and the 1,624 handwritten characters as fine-grained classes. We use the dataset to prove the applicability and feasibility of HMRN. Experimental results of the proposed HMRN and state-of-the-art models on it are listed in Table 5, and we can obtain the following observations:

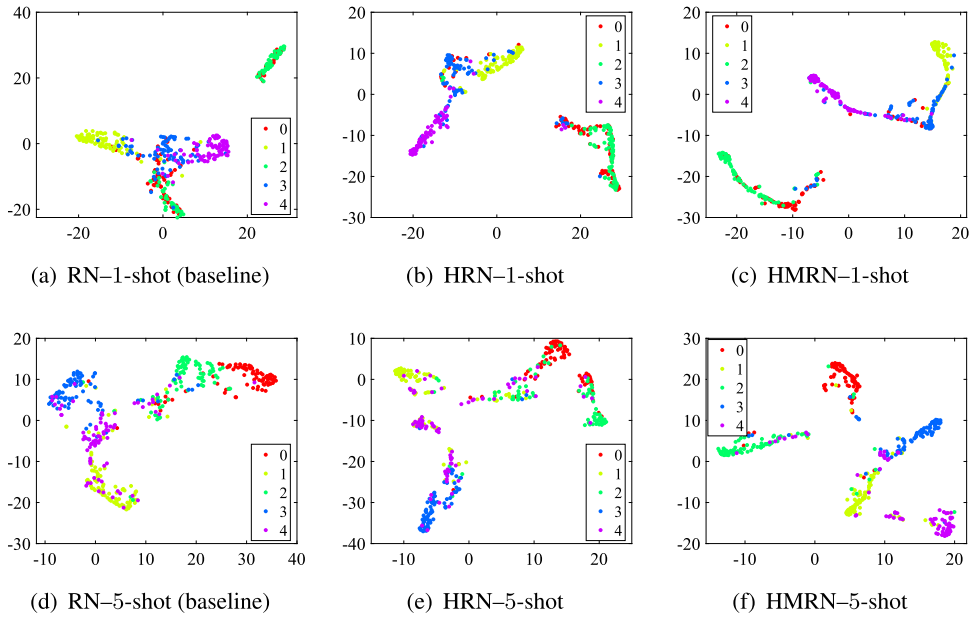

Fig. 10. t-SNE visualization under 5-way 1-shot and 5-shot settings on the *tieredImageNet* dataset. HRN and HMRN both consider the relationship of inter classes. The numbers 0 to 4 represent different fine-grained classes. (1) 1-shot: From Fig. 10(a) to Fig. 10(c), classes 0 and 2, 1 and 3 belong to the same coarse-grained class respectively, and class 4 belongs to an individual coarse-grained class. (2) 5-shot: From Fig. 10(d) to Fig. 10(f), classes 0 and 2, 1 and 4 belong to the same coarse-grained class respectively, and class 3 belongs to an individual coarse-grained class.

Table 5
Comparison with different models on the *Omniglot* dataset (%). The results of flat and hierarchical models are separated. The best performing model is highlighted.

Model	Hierarchical	5-way	
		1-shot	5-shot
Matching network [37]	N	98.10	98.90
Prototypical network [35]	N	98.80	99.70
Relation network [36]	N	99.60 ± 0.20	99.80 ± 0.10
MAML [7]	N	98.70 ± 0.40	$\mathbf{99.90 \pm 0.10}$
SARN [12]	N	$\mathbf{99.70 \pm 0.20}$	$\mathbf{99.90 \pm 0.10}$
IMP [2]	N	98.40 ± 0.30	99.50 ± 0.10
PRN [21]	N	99.27 ± 0.23	99.91 ± 0.06
C2F w/BDE-MetaBL [45]	Y	96.43 ± 0.32	98.53 ± 0.04
HMRN	Y	$\mathbf{98.95 \pm 0.05}$	$\mathbf{99.41 \pm 0.04}$

(1) From Table 5, under both 1-shot and 5-shot episodes, HMRN achieves 98.95% and 99.41%, which are only 0.75% and 0.50% lower than the best flat model, respectively. HMRN can achieve a satisfactory performance, which is applicable to datasets without class hierarchical structure under some certain.

(2) Compared with the hierarchical model, HMRN gains about 2.50% and 0.90% improvements over C2F w/BDE-MetaBL in 1-shot and 5-shot episodes. C2F w/BDE-MetaBL and our model both exploit the class hierarchical structure in classification, and there is a main difference. C2F w/BDE-MetaBL adopted a coarse-to-fine pseudo-labeling process to construct classification tasks, while we use the class hierarchical structure to take the coarse-grained class information to assist the classification of fine-grained classes.

The *tieredImagenet* has the class hierarchical structure by itself. We use the dataset to prove the effectiveness and feasibility of HMRN on the dataset with class hierarchical structure. The results are listed in Table 6, and we can obtain the following observations:

(1) Form Table 6, HMRN achieves 57.98% and 74.70% classification accuracies for 5-way 1-shot and 5-shot episodes on *tieredImagenet*. It outperforms the best flat model by about 3.00% and 3.20%, which confirms that exploring the information from class hierarchical structure benefits few-shot learning. Thus the effectiveness of our HMRN is proved.

(2) Compared with the hierarchical models, HMRN outperforms TPN by 0.45% and 1.85% in 1-shot and 5-shot episodes. The performance of HMRN is inferior to that of MNE in 1-shot episode. But HMRN gets more than 1.00% improvement over MNE in 5-shot episode. MNE employs a tree graph structure to aggregate the neighborhood information to enhance the features, which is similar to our class hierarchical structure.

Table 6
Comparison with different models on the *tieredImagenet* dataset (%). The results of flat and hierarchical models are separated. The best performing model is highlighted.

Model	Hierarchical	5-way	
		1-shot	5-shot
Prototypical network [35]	N	48.58 ± 0.87	69.57 ± 0.75
Relation network [36]	N	54.26 ± 0.27	71.34 ± 0.28
MAML [7]	N	51.67 ± 1.81	70.30 ± 1.75
CovaMNet [19]	N	$\mathbf{54.98 \pm 0.90}$	$\mathbf{71.51 \pm 0.75}$
TPN [22]	Y	57.53 ± 0.96	72.85 ± 0.74
MNE [18]	Y	$\mathbf{60.04 \pm 0.28}$	73.63 ± 0.21
HMRN	Y	57.98 ± 0.26	$\mathbf{74.70 \pm 0.24}$

5. Conclusions and future work

In this paper, we proposed a few-shot hierarchical classification model based on multi-granularity relation networks (HMRN) by embedding class hierarchical structure. It can promote the classification ability by reducing the inner-class distance and expanding the inter-class distance. We made full use of the similarity relation of inner classes and the subordination relation among coarse- and fine-grained classes in the class hierarchical structure, combining the assistance of coarse-grained class information to the learning of fine-grained classes. In addition, we built a multi-granularity and multi-object loss function, which promotes the classification performance of different granularity classes to improve the model classification ability. Experimental results show that HMRN is comparable with several state-of-the-art few-shot learning models.

In this paper, HMRN has strict requirements on the hierarchical structure of data, which assumes that the finest grained classes are at the same granularity layer. However, this assumption does not exist in some cases in real life. In the future, we will focus on studying how to apply the hierarchical few-shot learning to the data with hierarchical structure where the finest grained classes are not at the same granularity layer.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62141602 and 62076116, the Natural Science Foundation of Fujian Province under Grant Nos. 2021J011003 and 2021J02049, and the Postgraduate Education Reform Project of Minnan Normal University under Grant No. YJG202120.

References

[1] A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. Feris, R. Giryes, A.M. Bronstein, Laso: label-set operations networks for multi-label few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[2] K. Allen, E. Shelhamer, H. Shin, J. Tenenbaum, Infinite mixture prototypes for few-shot learning, in: International Conference on Machine Learning, 2019.
[3] C. Cheng, C. Li, Y. Han, Y. Zhu, A semi-supervised deep learning image caption model based on pseudo label and n-gram, Int. J. Approx. Reason. 131 (2021) 93–107.
[4] S.Y. Chou, K.H. Cheng, J.S.R. Jang, Y.H. Yang, Learning to match transient sound events using attentional similarity for few-shot sound recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.
[5] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F.F. Li, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[6] Y. Ding, X. Tian, L. Yin, X. Chen, S. Liu, B. Yang, W. Zheng, Multi-scale relation network for few-shot learning based on meta-learning, in: International Conference on Computer Vision Systems, 2019.
[7] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, 2017.
[8] G.B. Goh, N.O. Hodas, A. Vishnu, Deep learning for computational chemistry, J. Comput. Chem. 38 (16) (2017) 1291–1307.
[9] S. Guo, H. Zhao, Hierarchical classification with multi-path selection based on granular computing, Artif. Intell. Rev. 54 (3) (2021) 2067–2089.
[10] J. He, R. Hong, X. Liu, M. Xu, Z.J. Zha, M. Wang, Memory-augmented relation network for few-shot learning, in: International Conference on Multimedia, 2020.
[11] A. Heidari, J. McGrath, I.F. Ilyas, T. Rekatsinas, Holodetect: few-shot learning for error detection, in: International Conference on Management of Data, 2019.
[12] B. Hui, P. Zhu, Q. Hu, Q. Wang, Self-attention relation network for few-shot learning, in: IEEE International Conference on Multimedia & Expo Workshops, 2019.
[13] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: Deep Learning Workshop, vol. 2, 2015.

[14] B.M. Lake, R. Salakhutdinov, J.B. Tenenbaum, Human-level concept learning through probabilistic program induction, Science 350 (6266) (2015) 1332–1338.
[15] S. Legg, M. Hutter, Universal intelligence: a definition of machine intelligence, Minds Mach. 17 (4) (2007) 391–444.
[16] F.F. Li, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Trans. Pattern Anal. Mach. Intell. 28 (4) (2006) 594–611.
[17] J. Li, C. Huang, J. Qi, Y. Qian, W. Liu, Three-way cognitive concept learning via multi-granularity, Inf. Sci. 378 (2017) 244–263.
[18] S. Li, D. Chen, B. Liu, N. Yu, R. Zhao, Memory-based neighbourhood embedding for visual recognition, in: IEEE International Conference on Computer Vision, 2019.
[19] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, J. Luo, Distribution consistency based covariance metric networks for few-shot learning, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019.
[20] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, R. Wang, Deep few-shot learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 57 (4) (2018) 2290–2304.
[21] X. Liu, F. Zhou, J. Liu, L. Jiang, Meta-learning based prototype-relation network for few-shot classification, Neurocomputing 383 (2020) 224–234.
[22] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S.J. Hwang, Y. Yang, Learning to propagate labels: transductive propagation network for few-shot learning, in: International Conference on Learning Representations, 2019.
[23] Z. Liu, Y. Xia, B. Zhang, Graph embedding relation network for few-shot learning, in: Chinese Control Conference, 2020.
[24] J. Lu, P. Gong, J. Ye, C. Zhang, Learning from very few samples: a survey, arXiv e-prints arXiv:2009.02653.
[25] F. Mi, M. Huang, J. Zhang, B. Faltings, Meta-learning for low-resource natural language generation in task-oriented dialogue systems, arXiv e-prints arXiv:1905.05644.
[26] Miller, A. George, Wordnet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[27] H.B. Moss, V. Aggarwal, N. Prateek, J. González, R. Barra-Chicote, Boffin tts: few-shot speaker adaptation by Bayesian optimization, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.
[28] B.N. Oreshkin, P.R. López, A. Lacoste, Tadam: task dependent adaptive metric for improved few-shot learning, in: Neural Information Processing Systems, 2018.
[29] J. Qian, C. Liu, D. Miao, X. Yue, Sequential three-way decisions via multi-granularity, Inf. Sci. 507 (2020) 606–629.
[30] Y. Qian, X. Liang, G. Lin, Q. Guo, J. Liang, Local multi-granulation decision-theoretic rough sets, Int. J. Approx. Reason. 82 (2017) 119–137.
[31] A. Ravichandran, R. Bhotika, S. Soatto, Few-shot learning with embedded class models and shot-free meta training, in: IEEE International Conference on Computer Vision, 2019.
[32] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J.B. Tenenbaum, H. Larochelle, R.S. Zemel, Meta-learning for semi-supervised few-shot classification, in: International Conference on Learning Representations, 2018.
[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, IEEE Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[34] C.N. Silla, A.A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (1) (2011) 31–72.
[35] J. Snell, K. Swersky, R.S. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. (2017) 4077–4087.
[36] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: relation network for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[37] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, Neural Inf. Process. Syst. (2016) 3630–3638.
[38] S. Wang, W. Zhu, Sparse graph embedding unsupervised feature selection, IEEE Trans. Syst. Man Cybern. Syst. 48 (3) (2016) 329–341.
[39] Y. Wang, R. Liu, D. Lin, D. Chen, P. Li, Q. Hu, C.P. Chen, Coarse-to-fine: progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis, IEEE Trans. Neural Netw. Learn. Syst. (2021), https://doi.org/10.1109/TNNLS.2021.3100928.
[40] Y. Wang, Z. Wang, Q. Hu, Y. Zhou, H. Su, Hierarchical semantic risk minimization for large-scale classification, IEEE Trans. Cybern. (2021), https://doi.org/10.1109/TCYB.2021.3059631.
[41] Y. Wang, N.L. Zhang, T. Chen, L.K. Poon, Ltc: a latent tree approach to classification, Int. J. Approx. Reason. 54 (4) (2013) 560–572.
[42] Y. Wang, R. Zou, F. Liu, L. Zhang, Q. Liu, A review of wind speed and wind power forecasting with deep neural networks, Appl. Energy 304 (2021) 117766.
[43] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell, Distance metric learning with application to clustering with side-information, Neural Inf. Process. Syst. 15 (2002).
[44] L. Yan, Y. Zheng, J. Cao, Few-shot learning for short text classification, Multimed. Tools Appl. 77 (22) (2018) 29799–29810.
[45] J. Yang, H. Yang, L. Chen, Coarse-to-fine pseudo-labeling guided meta-learning for inexactly-supervised few-shot classification, arXiv e-prints arXiv: 2007.05675.
[46] H. Zhao, Q. Hu, P. Zhu, Y. Wang, P. Wang, A recursive regularization based feature selection framework for hierarchical classification, IEEE Trans. Knowl. Data Eng. 33 (7) (2021) 2833–2846.