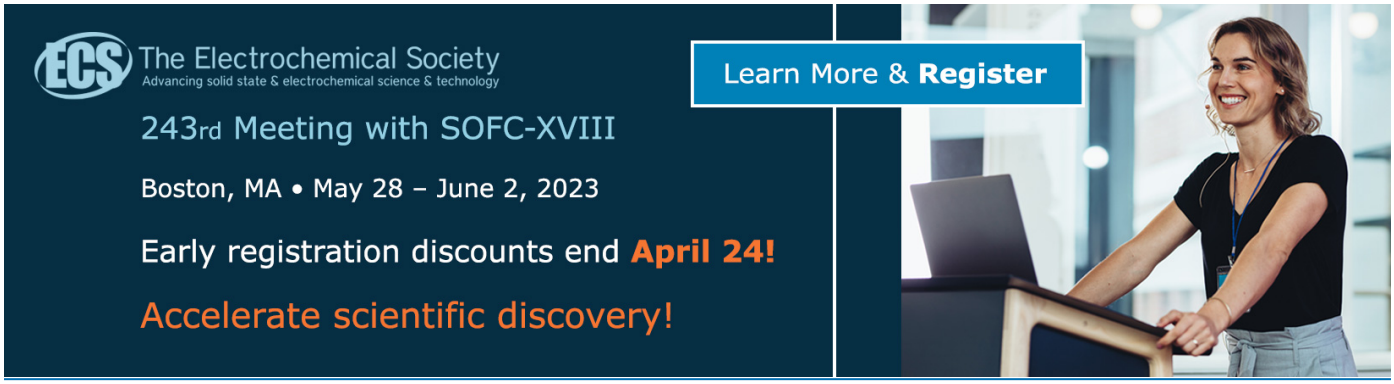
Few-Shot Image Classification Based on Ensemble Metric Learning



Few-Shot Image Classification Based on Ensemble Metric Learning

Hang Wang[1,*], Duanbing Chen[1,2]

[1]Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China

[2]Chengdu Union Big Data Tech. Inc., Chengdu 610041, China

[*]email: cnwanghhh@163.com

Abstract. In the case of few labelled image data samples, image classification is a difficult challenge, which is called few-shot image classification. Recently, many methods based on metric learning have been proposed. Most of these methods mainly focus on the representations of global image-level features or local feature-level descriptors. However, these methods calculate similarity from a single metric learning perspective. Motivated by ensemble learning, a novel Ensemble Metric Learning (EML) method for few-shot image classification is proposed, which not only utilizes label propagation, but also considers image-level and local feature-level descriptor metrics. The experimental results show that the proposed method can effectively improve the classification accuracy by ensemble learning.

1. Introduction

Although deep learning has achieved great success in the field of image classification, the reason for its success is inseparable from large-scale datasets (ImageNet, Pascal-VOC). However, in some special application scenarios(such as medical, military), due to privacy, security and other reasons, it is difficult to obtain a large amount of data. When there are only a few labelled data, deep learning is easy to overfit. In contrast, humans can learn new concepts and objects quickly with only one or a few samples. In order to imitate this ability of humans, few-shot learning (FSL) came into being, and many FSL methods have been proposed, especially metric learning [1-6]. The core idea of the metric learning in FSL is to learn a good transferable feature embedding space in which a good metric is learned to calculate the similarity between the query sets and the support sets, and it can be divided into feature embedding representations and similarity measure. In fact, under the condition of few-shot setting, a single similarity measure result may not truly reflect the relationship between query sets and support sets, and may also lead to a certain classification deviation. To this end, inspired by ensemble

learning [7] and propose a novel Ensemble Metric Learning (EML) network, which can be trained in an end-to-end manner. Firstly, the features of the support sets and query sets are obtained through the feature extractor. Secondly, the support sets and query sets features are input into the ensemble metric module to calculate label propagation similarity, global image-level KL divergence similarity and local feature-level descriptors similarity. Finally, through the idea of stacking in ensemble learning, fusing three kinds of similarity scores to obtain the final classification result.

The main contributions of this work can be summarized as follows:

- Not only global image-level feature representations but also local feature-level descriptor representations are considered.
- Using the idea of stacking in ensemble learning to flexibly integrate three different metric learning methods.
- Sufficient experiments have been conducted on two popular FSL datasets (miniImageNet and tieredImageNet) and achieved the state-of-the-art, providing supporting evidence that the idea of using ensemble learning could produce better results in few-shot metric tasks.

2. Related work

The core idea of metric learning is to compare the similarity of feature embeddings of support sets and query sets in the feature embedding space through similarity measure module. MatchNet [1] encodes features of the query images and support images by using LSTM and uses an attention-based weighted measurement function to measure the similarity between them. ProtoNet [2] calculates the average of embedding features of support sets as prototype representation and uses the Euclidean distance to calculate the distance between the query images and the prototype as the classification basis. RelationNet [3] uses a learnable convolutional neural network to measure feature similarity instead of traditional non-parametric measurement method. DN4 [4] and CovaMNet [5] use local feature-level descriptors instead of global image-level features for similarity calculation. The former uses deep local descriptors and image-to-class measure for classification, while the latter uses deep local descriptors to calculate local covariance representation, which are used to calculate distribution consistency between a query sample and each support category. TPN [6] uses the similarity measurement result of Relation Network as a graph and uses the label propagation algorithm to transduce the labels of support sets to the query sets to obtain the classification result.

The core idea of stacking in ensemble learning [7] is to train a series of separate models in parallel, and then train a meta-model to combine the outputs of each model to obtain the final result.

3. Methods

As shown in Figure 1, Ensemble Metric Learning (EML) is mainly composed of three modules: feature extractor module, similarity measure module and ensemble module.

3.1. Feature extractor module

In order to compare with the most advanced methods, using 4-layer convolution module as feature extractor module F_α. Each convolution module is composed of a convolution layer containing 64 filters with a size of 3×3, a batch normalization layer and LeakyReLU activation. And a maxpool layer of size 2×2 added after the activation layer in the first two convolution modules.

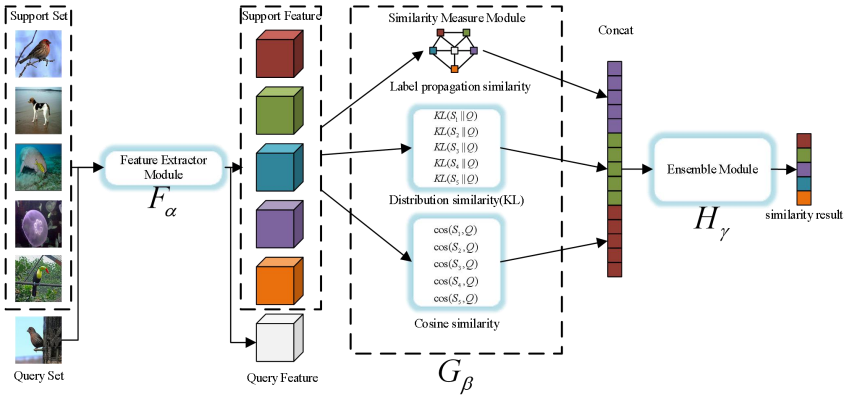


Fig. 1 The framework of Ensemble Metric Learning Model under 5-way 1-shot image classification setting

3.2. Similarity measure module

Given a query image Q and a specific support classes S, through feature extractor module F_α, getting the feature representation $F_\alpha(Q) \in \mathbb{R}^{C\times H\times W}$ and $F_\alpha(S) \in \mathbb{R}^{K\times C\times H\times W}$.

3.2.1. Label propagation similarity metric. Using the same strategy in TPN[6]. Firstly, the example-wise σ is calculated by a graph construction module built on the union set of support sets and query sets. This module is composed of a convolution neural network which takes the feature map $F_\alpha(Q) \cup F_\alpha(S)$ as input, whose structure is shown in Figure 2.

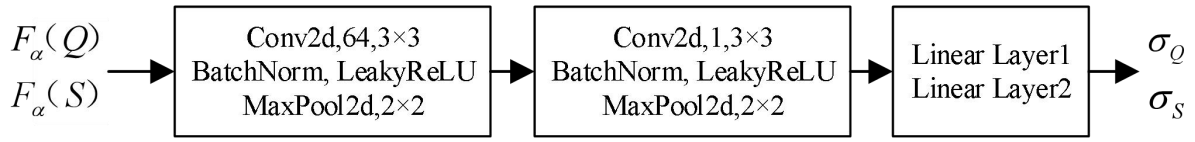


Fig. 2 The detailed architecture of graph construction module

Secondly, with the example-wise σ, the similarity function is defined by:

$$W_{ij} = \exp\left(-\frac{1}{2}d\left(\frac{F_\alpha(Q)}{\sigma_Q}, \frac{F_\alpha(S)}{\sigma_S}\right)\right) \tag{1}$$

Where $W_{ij} \in \mathbb{R}^{(N\times K+N\times q)\times(N\times K+N\times q)}$ for all images in $S \cup Q$. After getting the similarity matrix W, only keeping the k-max values in each row of W to construct the K-nearest neighbor graph ($k = 20$ in this work). Subsequently, Laplacian regularization is performed on the graph to obtain the final similarity matrix L, that is, $L = D^{-1/2}WD^{-1/2}$, where D is a diagonal matrix and its diagonal element (i, i) is the sum of the i-th row elements of W. Finally, label propagation function is used to calculate the final query image label score. The label propagation function is defined as:

$$Y^* = (I - \alpha L)^{-1}Y \tag{2}$$

Where Y^* is the final predicted label score, I is the identity matrix, α is the amount of propagated information ($\alpha = 0.99$ in this work), L is the similarity matrix and Y is the label matrix. For $Y, Y \in \mathbb{R}^{(N\times K+N\times q)\times N}$, and $Y_{ij} = 1$ if x_i is from the support set and labelled as $y_{i=j}$, otherwise $Y_{ij} = 0$.

3.2.2. KL divergence distribution similarity metric. Inspired by [5, 8], the distribution of local feature-level descriptors can be considered as multivariate normal distribution, where $D_Q = \mathcal{N}(\mu_Q, \Sigma_Q)$ represents the distribution of query image, $D_S = \mathcal{N}(\mu_S, \Sigma_S)$ represents the distribution of support class, $\mu \in \mathbb{R}^C$ represents the mean vector and $\Sigma \in \mathbb{R}^{C\times C}$ represents covariance matrix of a specific distribution. The KL divergence distribution similarity metric is defined by:

$$KL(S||Q) = -\frac{1}{2}\left(trace\left(\Sigma_Q^{-1}\Sigma_S\right) + \ln\left(\frac{\det\Sigma_Q}{\det\Sigma_S}\right) + (\mu_Q - \mu_S)^T\Sigma_Q^{-1}\left(\mu_Q - \mu_S\right) - c\right) \quad (3)$$

Where *det* is the value of the determinant of a matrix and $trace(\cdot)$ is the trace operation of matrixes.

3.2.3. Cosine similarity metric. Directly using DN4 [9] as cosine similarity metric, which capture the local relation between a certain support class and a query image by the image-to-class via *k*-NN.

$$W_{\cos(i,j)} = \cos\left(q_i, s_j\right) = \frac{q_i^T s_j}{||q_i||\cdot||s_j||} \quad (4)$$

$$\cos(S, Q) = TopK(W_{cos}) \quad (5)$$

3.3. Ensemble module

Through similarity measure module, concatenating three similarity score vectors and balancing the size of them by a Batch Normalization layer, then getting a weighted *N*-dimensional similarity by a 1D convolution layer (where w_1, w_2, w_3 are learnable vectors) as final classification result. The result can be obtained by the following equation:

$$Final\ Score = w_1 \cdot Y^* + w_2 \cdot KL(S||Q) + w_3 \cdot \cos(S, Q) \quad (6)$$

4. Experiments

4.1. Datasets

In order to verify the advance and effectiveness of EML model, all experiments are conducted on both miniImageNet and tieredImageNet, both of which are a mini-version of ImageNet. miniImageNet contains 100 classes, each class has 600 images, and 64, 16 and 20 classes are used for training, validation and test. Different from miniImageNet, tieredImageNet has more categories and a broader category hierarchy. It contains 608 classes each class has 1281 images and 351, 97 and 160 classes are used for training, validation and test. Besides, in both datasets, all images size is resized to 84×84.

4.2. Implement details

In training and testing stage, the data in each of episodic task strictly followed the N-way K-shot form, that is, conducting 5-way 1-shot and 5-way 5-shot classification tasks to verify our model, and in each episodic task, the number of query images in each class is 15. In the training stage, using the Adam optimizer (where the initial learning rate is 0.001 and multiplied by 0.5 every 10 epochs, other parameters are default values) to train our model for 40 epochs. And in each epoch, randomly constructing 10,000 episodic tasks. In the testing stage, 600 episodic tasks are randomly constructed from the test data set and the average accuracy is taken as the evaluation criterion. This process is repeated 10 times, and the average accuracy of 10 times is used as the final result.

4.3. Results and discussions

In the experiment, EML is compared with the most advanced FSL methods, including MatchNet [1], ProtoNet [2], RelationNet [3], DN4 [4], CovaMNet [5], TPN [6] and ADM [8]. The experimental results on miniImageNet and tieredImageNet are shown in Table 1.

Tab. 1 Comparison of miniImageNet and tieredImageNet experimental results

Method	5-way 1-shot Acc(%)(mini)	5-way 5-shot Acc(%)(mini)	5-way 1-shot Acc(%)(tiered)	5-way 5-shot Acc(%)(tiered)
MatchNet [1]	43.56±0.84	55.31±0.73	-	-
ProtoNet [2]	49.42±0.78	68.20±0.66	53.31±0.89	72.69±0.74
RelationNet [3]	50.44±0.82	65.32±0.70	54.48±0.93	71.32±0.78
DN4 [4]	51.24±0.74	71.02±0.64	53.37±0.86	74.45±0.70
CovaMNet [5]	51.19±0.76	67.65±0.63	54.98±0.90	71.51±0.75
TPN 6]	53.75±0.86	69.43±0.67	57.53±0.96	72.85±0.74
ADM [8]	54.26±0.63	72.54±0.50	56.01±0.69	75.18±0.56
EML(Ours)	**54.91±0.63**	**73.41±0.50**	**57.65±0.71**	**76.35±0.54**

As seen from table 1, EML achieves the highest accuracy on miniImageNet and tieredImageNet. For miniImageNet and tieredImagenet, EML achieves the highest accuracy with 54.91% and 57.65% on 5-way 1-shot and 73.41% and 76.35% on 5-way 5-shot. And compared with the single metric method, EML has a great improvement, and compared with the latest integrated two measurement learning method ADM [8], it is also obtains 0.65% and 0.87% improvements and 1.64% and 1.17% improvements on 1-shot and 5-shot in miniImageNet and tieredImageNet.

5. Conclusion

It can be seen from the analysis that the EML model mainly utilizes the idea of stacking in ensemble learning. Three different metric learning strategies are integrated to overcome the disadvantages of single metric learning and to improve the accuracy of the model in FSL. In addition, ensemble learning is also the most basic idea to improve model accuracy in deep learning and machine learning. Some other metric learning methods will be studied to improve model accuracy in the future.

Acknowledgements

This research was funded by the National Natural Science Foundation of China with Grant No 61673085, the Fundamental Research for the Central Universities with Grant No ZYGX2019J074, and Science Strength Promotion Programme of UESTC with Grant No Y03111023901014006.

References

[1] Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K and Wierstra D 2016 *Advances in Neural Information Processing Systems (Barcelona)* vol 29 (New York: Curran Associates) pp 3637-3645

[2] Snell J, Swersky K and Zemel R 2017 *Advances in Neural Information Processing Systems (California)* vol 30 (NY: Curran Associates) pp 4077-4087

[3] Sung F, Yongxin Y, Li Z, Tao X, Philip H, Hospedales T 2018 *IEEE Conf. on Computer Vision*

and Pattern Recognition (Salt Lake City) (Piscataway: IEEE Press) pp 1199-1208

[4] Wenbin L, Lei W, Jinglin X, Jing H, Yang G and Jiebo L 2019 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Long Beach, CA)* (Piscataway: IEEE Press) pp 7253-7260

[5] Wenbin L, Jinglin X, Jing H, Lei W, Yang G and Jiebo L 2019 *33rd AAAI Conf. on Artificial Intelligence (Honolulu)* vol 33 (Palo Alto: AAAI) pp 8642–8649

[6] Yanbin L, Lee J, Park M, Kim S, Eunho Y, SungJu H and Yi Y 2019 *7th Int. Conf. on Learning Representations (New Orleans)* (Amsterdam: Elsevier)

[7] Baruque B and Corchado E 2010 The Committee of Experts Approach: Ensemble Learning *Fusion Methods for Unsupervised Learning Ensembles* (Computational Intelligence vol 322) Kacprzyk J (Berlin: Springer) chapter 3 pp 31-47

[8] Wenbin L, Lei W, Jing H, Yinghuan S, Yang G and Jiebo L 2020 *Proc. Int. Joint Conf. on Artificial Intelligence (Yokohama)* (San Francisco: Morgan Kaufmann) pp 2957-2963