Transformer-Based Few-Shot Learning for Image Classification

Tao Gan(✉), Weichao Li, Yuanzhe Lu, and Yanmin He

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
gantao@uestc.edu.cn

Abstract. Few-shot learning (FSL) remains a challenging research endeavor. Traditional few-shot learning methods mainly consider the distance relationship between the query set and the support set, while the context information between different support sets are not fully exploited. This paper proposes a Transformer-based few-shot learning method (TML). By taking advantage of the self-attention mechanism of Transformer, TML effectively exploits the correlation between support sets so as to learn highly discriminative global features. Furthermore, in order to cope with the overfitting problem introduced by the increase of model complexity, we introduce a classification loss into the total loss function as a regularization term. To overcome the limit of traditional cross-entropy loss, a label refinement method is used to refine the label assignment for classification. The experimental results show that TML improves the ability of learning hard samples and achieves higher classification accuracy than existing state-of-the-art few-shot learning methods.

Keywords: Few-shot learning · Classification · Transformer · Regularization

1 Introduction

Few-shot Learning (FSL) is a new machine learning paradigm which aims to learn from a limited number of examples with supervised information. To make best use of the limited data available, different methods have been proposed. Ravi et al. [1] raise the episodic training idea and propose a meta-learning approach in which an embedding model is learned so that the base learner minimizes generalization error through the distribution of tasks with few training examples. Lee et al. [2] investigate linear classifiers as the base learner for a meta-learning based approach for few-shot learning. Another category of methods addresses the problem with data augmentation by distorting the labeled images or synthesizing new images/features based on the labeled ones [3]. Alternatively, the third group of methods resort to enhance the discriminability of the feature representations such that a simple linear classifier learned from a few labeled samples can reach satisfactory classification results. Snell et al. [4] proposes the prototypical networks that learn a metric space in which classification can be performed by computing distances

X. Wang et al. (Eds.): AICON 2021, LNICST 396, pp. 68–74, 2021.
https://doi.org/10.1007/978-3-030-90196-7_6

to prototype representations of each class. Le et al. [5] introduce a category traversal module to extract feature dimensions most relevant to each task, by looking the context of the entire support set. Simon et al. [6] uses a dynamic subspace classifier to calculate a subspace of the feature space for each category, and then project the feature vector of the query sample into the subspace.

Our proposed method falls into the feature enhancement based category. Intuitively, the objective for the classification task is to optimize the feature space in which samples of the same class should be close by in the learned manifold, while samples of a different class should be far away. However, in few-shot setting, the work of [7] shows that optimizing feature distances may not necessarily lead to performance gains. Inspired by the work of [5], we attempt to improve the classification performance by considering the context information between different support sets within the whole task. Furthermore, we try to investigate more powerful techniques to exploit the relation between support sets than that used in [5].

Recently, self-attention-based architectures, in particular Transformer [8], have become the model of choice in natural language processing (NLP) and Transformer has started to be studied in computer vision [9]. The powerful feature representation ability of Transformer just fits the need of our task. Moreover, for the sake of computational efficiency, we do not apply Transformer directly on samples of support sets. Instead, we first employ prototypical network [4] to extract prototype features of original samples and then fed the obtained prototype feature embeddings to Transformer. Since there is one feature embedding for one support, we can keep the computational complexity in low level.

Furthermore, the introduction of Transformer may increase the risk of overfitting. Inspired by the work of [10], we cope with the overfitting problem by regularizing the loss function. To do so, we calculate the classification loss and add it into the total loss function as a regularization term. Meanwhile, to overcome the limit of traditional cross-entropy loss, we refine the label assignment for classification. This is achieved by using a label refinement procedure in which the similarities between the query and support sets are used to refine the label assignment information.

To conclusion, we propose a Transformer-based few-shot learning method (TML). In this method, we integrate the prototypical network and Transformer to train a better feature space and at the same time, we mitigate the risk of overfitting by regularizing the loss function. Our main contributions are as follows:

1. We introduce Transformer into few-shot learning to efficiently exploit the relation between support sets.
2. Incorporated with modified loss function, the proposed method improves the ability of learning hard samples and achieves higher classification accuracy than existing state-of-the-art few-shot learning methods.

2 Methodology

2.1 Overview

In few-shot learning, the model usually trained in episode fashion. An episode corresponds to one task and the categories contained in support sets between episodes may

be different. For n-way k-shot image classification problem, we aim to build a complex classifier $f(\cdot)$ which reduces the average error rate in different episode as much as possible. The problem can be formulated as

$$f^* = arg\min_f \sum\nolimits_{(x_{test}, x_{test}) \in D_{test}} \ell(f(x_{test}; D_{train}), y_{test}) \tag{1}$$

where D_{train} represents the training set for episode training and D_{test} represents the verification set for the same task as the training set $\ell(\cdot)$ is a loss function which measures the differences between the ground-truth labels and the predicted ones. In our method, the image is generally embedded into one feature space by an embedding function ϕ and a nearest neighbor classifier is used on that space. The classifier can be formulated as

$$y_{test} = f\left(\phi_{x_{test}}; \{\phi_x, \forall(x, y) \in D_{train}\}\right) \propto exp\left(sim\left(\phi_{x_{test}}, \phi_x\right)\right) y, \forall(x, y) \in D_{train} \tag{2}$$

The proposed TML model consists of three principle modules: prototypical network, transformer and label refinement. Firstly, images in both support sets and query set are input into the prototypical network and the prototype feature embeddings of corresponding sets are obtained. Then the transformer and label refinement modules are designed for computing the distance loss and classification loss, respectively. The total loss is the combination of the distance loss and classification loss calculated. The system structure is illustrated in Fig. 1. In the following sections, we describe the proposed transformer and label refinement modules in detail.

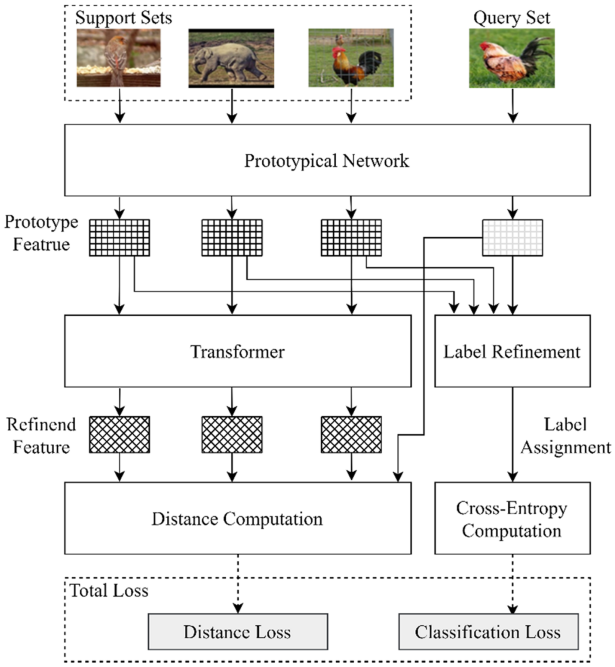

Fig. 1. The overview of the proposed TML model.

2.2 Transformer

The Transformer is a network model designed to process machine translation. It consists of encoder and decoder modules, each of which has multiple encoders or decoders of the same architecture. Each encoder consists of a self-attention layer and a feed-forward network. Each decoder consists of a self-attention layer, a codec attention layer and a feed-forward network.

The traditional few-shot image classification models, such as prototypical networks, obtain the feature embeddings of input images only by independently feeding them into convolutional neural networks. Thus the obtained feature embeddings of one support set contains no information about the that of others. In the proposed TML model, the output features of prototypical networks are fed into Transformer for feature re-extraction. Thus the embedded function ϕ in formula (1) can be expressed as

$$\{\psi_x; \forall x \in \mathrm{X}_{train}\} = T(\{\phi_x; \forall x \in \mathrm{X}_{train}\}) \tag{3}$$

where $T(\cdot)$ represents the function of Transformer and ψ_x represents features after transformed.

2.3 Label Refinement

As mentioned above, to mitigate the risk of overfitting, we resort to introduce the classification loss to regularize the loss function. However, the traditional classification loss function, e.g., cross-entropy, may not be appropriate for few-shot image classification task. Since when each sample is classified independently, it is possible that two images of the same class have two distant embeddings that both allow for a correct classification. Here we attempt to refine the classification information that used for calculating cross-entropy so as to make the classification loss contain ingredient that reflects distances between samples.

Towards this end, we introduce a label refinement method. This is an iterative procedure that uses the similarities between the query and support sets to refine the label assignment information. The structure of label refinement is shown in Fig. 2, where the similarity matrix and label matrix contains information of similarity and label assignment, respectively.

Consider a n-way k-shot image classification problem. Suppose W represents a $n+1 \times n+1$ matrix of pairwise similarity and $X = (x_{i\lambda})$ represents a $n+1 \times n$ matrix of image label assignments. The label refinement procedure consists of the following steps:

(1) For all pairs of prototype feature embeddings of the query set and support sets, generate a similarity matrix W by computing similarities among them. The similarity measure used is the Pearson's correlation coefficient:

$$\omega(i,j) = \frac{Cov\left[\varphi(I_i), \varphi\left(I_j\right)\right]}{\sqrt{Var[\varphi(I_i)], Var\left[\varphi\left(I_j\right)\right]}} \tag{4}$$

(2) Initialize X with the result of softmax computation of all prototype feature embeddings.
(3) Define the support matrix $\prod = (\pi_{i\lambda}) \in R^{n\times m} as$

$$\prod = WX \tag{5}$$

Given the initial assignment matrix $X(0)$, the algorithm refines it using the following update rule:

$$x_{i\lambda}(t+1) = \frac{x_{i\lambda}(t)\pi_{i\lambda}(t)}{\sum_{\mu=1}^{m} x_{i\lambda}(t)\pi_{i\lambda}(t)} \tag{6}$$

where the denominator represents a normalization factor which guarantees that the rows of the updated matrix sum up to one.

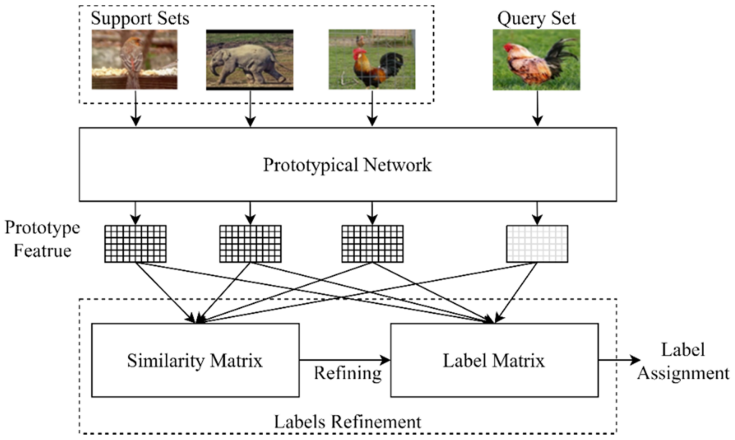

Fig. 2. The structure of label refinement.

3 Experiments

Below we evaluate our method on Mini ImageNet datasets. This dataset is a subset of 100 classes selected from the ImageNet. All images on are 84 × 84 × 3 and the number of samples per class is 600. We build datasets for 5-way 1-shot and 5-way 5-shot classification tasks. We train the method in 200 epochs with each epoch has 1000 images. We use Adam optimizer which has an initial learning rate of 0.0002 and decays by 0.5 for every 10 strides. In addition, the momentum of the optimizer is set to 0.9, and the weight decay is set to 0.0005.

We randomly construct 600 tasks for testing. The performance of our proposed TML is compared to that of other state-of-the-art few-shot methods, including ProtoNets [1], MetaOpnet [2], CTM [5], AFHN [3] and DSN-MR [6]. The classification accuracy results of these methods are presented in Table 1.

Table 1. Comparison of classification accuracy results

Method	1-shot	5-shot
ProtoNets	60.37 ± 0.83	78.02 ± 0.57
MetaOpnet	62.64 ± 0.61	78.63 ± 0.46
CTM	64.12 ± 0.82	80.51 ± 0.13
AFHN	62.38 ± 0.72	78.16 ± 0.56
DSN-MR	64.60 ± 0.72	79.51 ± 0.50
TML	66.75 ± 0.20	82.05 ± 0.14

It can be seen that TML provides best performance among all methods for both 1-shot and 5-shot cases. Especially, TML performs much better than ProtoNets, with 10.57% and 5.17% gains in accuracy for 1-shot and 5-shot, respectively. This can be attributed to the fact that TML build its model on the basis of ProtoNets and the superior performance is achieved by introducing Transformer as well as label refinement modules. In additional, we observe that the superiority of TML is more noticeable for 1-shot than 5-shot, indicating TML is more robust for difficult task.

It is worth noting that like TML, state-of-art MetaOptNet and CTM are methods that consider the context information of support sets within the entire task. TML still yields results that are superior to MetaOptNet and CTM for both 1-shot and 5-shot cases. This indicates that Transformer can exploit the correlation between support sets so as to learn highly discriminative features embeddings.

4 Conclusion

This work has presented a few-shot learning method based on Transformer. Incorporate with the prototypical network, Transformer exploits the relation between support sets effectively and efficiently. In addition, in order to cope with the overfitting problem introduced by the increase of model complexity, a classification loss is introduced into the total loss function as a regularization term. Experiments demonstrate the superiority of the proposed method over existing state-of-the-art few-shot learning methods.

References

1. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: Proceeding International Conference Learning Representation (ICLR) (2017).
2. Lee, K., et al.: Meta-learning with differentiable convex optimization. In: Proceeding IEEE Conference Computer Vision and Pattern Recognition (CVPR) (2019).
3. Li, K., Zhang, Y., Li, K., et al.: Adversarial feature hallucination networks for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13470–13479 (2020)
4. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for fewshot learning. In: Proceeding Neural Information Processing Systems (NIPS), pp. 4077–4087 (2017).

5. Li, H., Eigen, D., et al.: Finding task-relevant features for few-shot learning by category traversal. In: Proceeding IEEE Conference Computer Vision and Pattern Recognition (CVPR) (2019)
6. Simon, C., Koniusz, P., Nock, R., et al.: Adaptive subspaces for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4136–4145 (2020)
7. Liu, B., et al.: Negative margin matters: understanding margin in few-shot classification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 438–455. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_26
8. Vaswani, A., et al.: Attention is all you need. In: Proceeding Neural Information Processing Systems (NIPS), pp. 5998–6008 (2017)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Elezi, I., Vascon, S., et al.: The group loss for deep metric learning. arXiv preprint arXiv:1912.00385 (2019)