Journal of Multimedia Information System VOL. 9, NO. 4, December 2022 (pp. 269-274): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2022.9.4.269

Attentive Pooling Network for Few-Shot Learning

Xilang Huang[1], Seon Han Choi[1], Sungun Kim[2*]

Abstract

Few-shot learning has become essential for generating neural network models that can generalize to novel classes given only a few labeled samples. Previous studies mostly focused only on building a class prototype via the relations between intra-class sample features and adopted the prototype to classify the target samples. Considering that the number of labeled samples is typically limited under few-shot settings, the use of these methods to produce representative prototypes for classification may not be efficient. To this end, in this study, we propose an attentive pooling network (APNet), which establishes the relationship between the prototype and target sample feature to highlight their important regions. APNet selectively assigns higher weights to the local features with higher relative importance scores in the prototype and the target feature map. By minimizing the classification loss through supervised learning, APNet learns to produce prototypes that are specific to the target feature based on the relative importance scores. To verify the effectiveness of APNet, we compared it with the existing methods on two popular few-shot learning datasets, and APNet outperforms the related methods by achieving 71.12% and 77.58% classification accuracy in the 5-shot setting on the miniImageNet and CUB datasets, respectively.

Key Words: Image Classification, Few-Shot Learning, Distance Metric.

I. INTRODUCTION

The rapid development of deep learning has accelerated the pace of neural networks in solving various computer-vision tasks, such as image classification [1] and object detection [2]. To obtain a desirable neural-network model for a specific task, the network typically requires a large number of manually labeled samples to adjust its learnable parameters to cover the distribution of training samples. However, it is difficult and expensive to collect sufficient labeled samples to train the network. Moreover, deep-learning methods are prone to classifying classes that have never been shown in the training process. It is usually necessary to re-train the network on the new classes to achieve desirable performance.

In contrast to deep-learning methods, humans possess the ability to learn new concepts from a few samples. To imitate human learning processes, many studies have introduced a new learning paradigm, called few-shot learning [3]. Few-shot learning aims to develop a learning algorithm that correctly matches the labeled and unlabeled samples. Among the methods that focus on few-shot learning, metric-based methods [4-6] learn or use a suitable distance metric to classify target samples by measuring the similarities between the labeled samples and target samples. As the labeled samples are usually limited during training, [6] builds a correlation mhiatrix between the labeled and target samples and uses ts matrix to mutually the informative regions between them. However, the correlation matrix is not learnable, and thus the network is inefficient in learning new classes.

In terms of pairwise alignment, an attentive pooling module [7] has been introduced to match the labeled sample to the target one. The attentive pooling module aims to construct a learnable matrix that can mine the latent relationship between the labeled and target sample. By minimizing the classification loss on the target sample, the matrix transforms the labeled sample to another feature space that can correctly align the features between the labeled and target sample.

In this study, we propose an attentive pooling network (APNet) to approach the few-shot learning problem. The main building block of APNet is motivated by [7]. First, we build an intra-class fusion block to obtain the class prototype. Next, APNet computes the soft alignment between the prototype and the target sample features by a learnable matrix. This matrix learns to correctly transform the prototype features into the feature space that is close to the relevant

Manuscript received October 15, 2022 ; Revised November 07, 2022; Accepted November 10, 2022. (ID No. JMIS-22M-10-040)

Corresponding Author (*): Sungun Kim, +82-51-629-6235, kimsu@pknu.ac.kr
[1]Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul, Korea, huangxl@ewha.ac.kr, seonhan.choi@ewha.ac.kr
[2]Department of Information and Communications Engineering, Pukyong National University, Busan, Korea, kimsu@pknu.ac.kr

target sample features. Subsequently, APNet selects the maximum similarity score in each row and column of the similarity-score map and applies the softmax layer to the scores to obtain the attention value for the prototype and target sample features, respectively. These attention values indicate the mutually important local features of the prototype and the target sample; thus, APNet can produce a prototype that is closely related to the relevant target samples. In terms of contribution, since [7] is originally proposed to solve the question-answering problem, we additionally build an intra-class fusion block to aggregate the labeled sample features for the image classification task under the few-shot settings. Further, different from the original paper, we change the max-pooling direction to correctly augment the informative regions of support and query samples. To verify the effectiveness of APNet, we conducted experiments on two popular few-shot learning datasets, and the experimental results show the proposed method gains considerable improvement compared to the existing methods.

The remainder of this paper is organized as follows. Section II introduces the related work of few-shot learning. Section III gives the problem definition and the details of the proposed method. Section IV presents the experimental results, and Section V concludes the paper.

II. RELATED WORK

2.1. Metric-Based Few-Shot Learning

The metric-based methods advocate learning a suitable distance-metric function for the network, such that intra-class samples are close to each other. RelationNet [4] parameterizes the distance metric through a convolutional neural network (CNN). MatchingNet [3] adopts a bidirectional LSTM on the labeled sample features to extract the important features. ProtoNet [8] hypotheses that intra-class features are clustered around a prototype feature and calculate the mean vectors for classification. FEAT [5] generates prototypes that contain task-specific features by self-attention. DCAP [9] concatenates the intra-class feature vectors with the mean prototype and applies an attention regressor on the vectors to calculate attention for local features. CAN [6] proposes a cross-attention block to explore the correlations between the labeled and unlabeled sample features. However, the cross-attention block leads to different outcomes for each unlabeled sample according to the labeled sample, which consequently confuses the classifier when input classes are similar.

2.2. Attentive Pooling Module

The attentive pooling module [7,10] focuses on using a learnable matrix to match the informative regions between the labeled and target sample. [7] firstly proposed the attentive pooling module to solve the question-answering tasks. The module is simple but effective owing to the learnable matrix that correctly transforms the question embedding into a proper embedding space where the answer embedding lies. In the complicated person-re-identification task, [10] adopted the attentive pooling module to remove the redundant background information by calculating the pairwise alignment scores. Despite the attentive pooling module having achieved promising performance on the above tasks, it has not been applied in the few-shot learning. Therefore, we modify the module to approach the few-shot learning problem.

III. METHODOLOGY

3.1. Problem Definition

In this study, we consider the standard N-way K-shot classification problem. In few-shot learning, a dataset is split into three meta-datasets: meta-training Dtrain, meta-validation D_{val}, and meta-testing D_{test}. Each meta dataset has a disjointed labeled space to the others, e.g., $D_{train} \cap D_{test} = \emptyset$. To simulate data scarcity, few-shot learning uses an episodic learning mechanism to construct training tasks in each iteration. Specifically, each task consists of a support set $S=\{(x_{1,1}, y1,1). \dots ,(x_{N,K}, y_{N,K})\}$ and a query set $\boldsymbol{Q} = \{(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{y}}_1), \dots, (\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{y}}_j)\}$, where $\boldsymbol{x}_{n,i}$ is the $\boldsymbol{i}^{th}$ labeled sample of the $\boldsymbol{n}^{th}$ class, $y_{n,i} \in \{1, ..., \mathrm{N}\}$ is the corresponding label, and $\tilde{\boldsymbol{x}}_j$ is the $\boldsymbol{j}^{th}$ unlabeled sample. In each task, the setting of the support set is usually referred to as the N-way K-way classification problem. The primary goal of few-shot learning is to efficiently utilize limited support samples to generate a representative class prototype, which is used to accurately predict the labels of the query samples.

3.2. Model Architecture

One feasible solution to solve the few-shot learning problem is to build the soft alignments between the prototype and the query samples, so that the features of the prototype can be aligned to those in the query samples. To this end, we adopt the core block from [7], which allows the network to mine latent features that closely match the query features. This enables the network to generate a more characteristic prototype for efficient classification.

The overall structure of APNet is shown in Fig.1. Intuitively, APNet takes support and query samples as inputs and generates the corresponding feature maps by a CNN backbone $\boldsymbol{f_\theta}$ with learnable parameters $\boldsymbol{\theta}$. Subsequently, APNet averages the intra-class feature maps to obtain the prototype $\boldsymbol{P_n}$ for the $\boldsymbol{n}^{th}$ class by

$$\boldsymbol{P_n} = \frac{\boldsymbol{1}}{\boldsymbol{K}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{f_\theta}(\boldsymbol{x_{n,i}}). \tag{1}$$

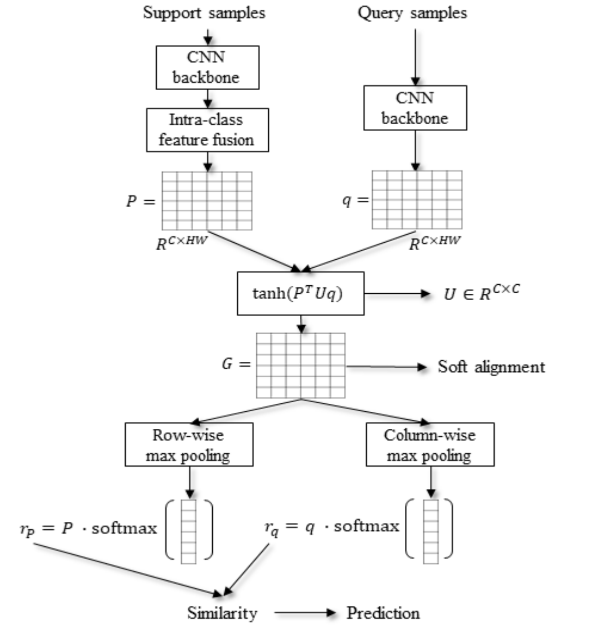

Fig. 1. Detailed structure of APNet. The intra-class feature fusion block aggregate the intra-class information and G is a learnable matrix to build soft alignment between support and query samples.

We reshape the prototype $\boldsymbol{P} \in \boldsymbol{R}^{\boldsymbol{C\times H\times W}}$and the query feature as $\boldsymbol{q} \in \boldsymbol{R}^{\boldsymbol{C\times HW}}$, where $\boldsymbol{C}, \boldsymbol{H},$ and $\boldsymbol{W}$ represent the values of the channel, height, and width of the feature maps. Next, we add a learnable matrix $\boldsymbol{U} \in \boldsymbol{R}^{\boldsymbol{C\times C}}$ to transform the prototype. The learnable matrix U is expected to transform the prototype to the proper feature space, such that the prototype can correctly match the local features of the query sample feature. We apply the tanh activation function to the matrix multiplication to obtain a soft alignment score matrix $\boldsymbol{G} \in \boldsymbol{R}^{\boldsymbol{HW\times HW}}$. The whole procedure can be formulated as:

$$\boldsymbol{G} = \mathbf{tan\,h}(P^T Uq). \tag{2}$$

Each row of the matrix G represents the soft alignment scores between one local feature of the transformed P and all the local features of the query feature map. On the other hand, each column indicates the scores between one local feature of the query feature map and all the features of the transformed P.

In the next step, APNet performs row-wise and column-wise max-pooling over matrix G to obtain the vectors $\boldsymbol{g^P} \in \boldsymbol{R^{HW}}$, and $\boldsymbol{g^q} \in \boldsymbol{R^{HW}}$, respectively. Thus, the $\boldsymbol{i^{th}}$ element of the vectors $\boldsymbol{g^P}$ and $\boldsymbol{g^q}$ can be formulated as follows:

$$[\boldsymbol{g^P}]_i = \max_{\boldsymbol{1<r<HW}}[\boldsymbol{G_{i,r}}], \tag{3}$$

$$[\boldsymbol{g^q}]_i = \max_{\boldsymbol{1<c<HW}}[\boldsymbol{G_{i,c}}]. \tag{4}$$

Each element of the vector $\boldsymbol{g^P}$ is the maximum alignment score obtained by searching for the most similar local feature on the query-feature map for each local feature in the prototype. Similarly, each element of the vector $\boldsymbol{g^q}$ represents the importance score of the local feature of the prototype that is most similar to each local feature on the query-feature map. Subsequently, we apply softmax to the vectors $\boldsymbol{g^P}$ and $\boldsymbol{g^q}$ to normalize the alignment score

$$\boldsymbol{\sigma^P} = \frac{\mathbf{exp}([\boldsymbol{g^P}]_i)}{\sum_{j=1}^{HW}\mathbf{exp}([\boldsymbol{g^P}]_j)}, \tag{5}$$

$$\boldsymbol{\sigma^q} = \frac{\mathbf{exp}([\boldsymbol{g^q}]_i)}{\sum_{j=1}^{HW}\mathbf{exp}([\boldsymbol{g^q}]_j)}. \tag{6}$$

After obtaining the attention vectors $\boldsymbol{\sigma^P}$ and $\boldsymbol{\sigma^q}$, we compute the dot product between the prototype P and the attention vector $\boldsymbol{\sigma^P}$. The same operation is also conducted between the query feature map q and the attention vector $\boldsymbol{\sigma^q}$:

$$\boldsymbol{r_P} = \boldsymbol{P\sigma^P}, \tag{7}$$

$$\boldsymbol{r_q} = \boldsymbol{q\sigma^q}. \tag{8}$$

To perform classification, APNet uses the cosine distance function on the refined prototype $\boldsymbol{r_P}$ and query feature $\boldsymbol{r_q}$. The softmax function is applied to the distances, and the class index with the maximum probability is selected as the predicted label for the query sample.

IV. EXPERIMENTS

4.1. Few-Shot Learning Datasets

Herein, we introduce the details of the two few-shot learning datasets that were used to evaluate APNet in this study. One, the miniImageNet dataset, was presented by [3]. It includes 100 classes and is split into 64, 16, and 20 classes, as D_{train}, D_{val}, and D_{test}. Each class contains 600 images of pixel size 84×84. The other is the CUB-200-2011 dataset [11]. This dataset contains 11,788 images of 200 bird species. We split the dataset into 100, 50, and 50 classes for training, validation, and testing, respectively, in accordance with previous studies [5,9]. All the images were resized to 84×84 to fit the input of the backbone network.

4.2. Implementation Details

In this study, we adopted a 4-layer convolutional network (Conv4) [5] as the backbone network. We trained APNet for 100 epochs, with each epoch featuring 600 tasks. The number of query samples for each class was set to 15. In the

Table 1. The 5-way 1-shot and 5-shot classification accuracies (%) on the miniImageNet dataset.

Method	Backbone	miniImageNet	
		5W1S	5W5S
MatchingNet [3]	Conv4	43.93±0.20	64.23±0.16
RelationNet [4]	Conv4	48.33±0.20	64.35±0.16
ProtoNet [8]	Conv4	50.42±0.20	65.72±0.16
FEAT [5]	Conv4	50.66±0.20	66.49±0.16
DCAP [9]	Conv4	53.22±0.20	68.63±0.16
CAN [6]	Conv4	48.21±0.20	69.52±0.16
APNet (ours)	Conv4	53.68±0.20	71.12±0.16

Table 2. The 5-way 1-shot and 5-shot classification accuracies (%) on the CUB-200-2011 dataset.

Method	Backbone	CUB-200-2011	
		5W1S	5W5S
MatchingNet [3]	Conv4	50.35±0.22	74.62±0.18
RelationNet [4]	Conv4	57.62±0.23	74.35±0.18
ProtoNet [8]	Conv4	53.82±0.23	73.02±0.17
FEAT [5]	Conv4	58.42±0.23	76.35±0.17
DCAP [9]	Conv4	58.62±0.25	75.89±0.18
CAN [6]	Conv4	58.56±0.23	76.63±0.17
APNet (ours)	Conv4	59.93±0.23	77.58±0.17

testing stage, we conducted 5-way 1-shot (5W1S) and 5-way 5-shot (5W5S) classification tasks for each dataset. The final classification accuracy was evaluated over 10,000 episodes and reported with a 95% confidence interval.

4.3. Image-classification results

Table 1 displays the classification results on the miniImageNet dataset. Compared to the related methods, the accuracies of APNet were 5% and 2% higher than those of CAN under the 1-shot and 5-shot settings. APNet also outperformed the current method, DCAP, by a considerable margin in the 5-shot task.

Table 2 presents the classification results on the CUB-200-2011 dataset. It can be observed that APNet consistently performed better than the other methods. This indicates that APNet can efficiently augment the important local features of each class in the presence of similar classes.

V. CONCLUSION

In this study, we proposed an attentive pooling network (APNet) to maximize information utilization. To aggregate the intra-class features, we first used a fusion block to obtain the class prototype. Then, we extracted latent relationships between the prototypes and target features via a learnable matrix, which was used as the criterion to selectively assign higher attention values to important regions. By conducting experiments on two few-shot learning datasets, APNet achieved 53.68% and 71.12%, 59.93% and 77.58% classification accuracy on the 1-shot and 5-shot settings on the miniImageNet and CUB datasets, respectively. The experimental results demonstrated a higher classification performance than the other methods, which validated the effectiveness of APNet under the few-shot settings.

ACKNOWLEDGMENT

This work was supported by the Pukyong National University Research Fund in 2020.

REFERENCES

[1] R. Prajapati and G. R. Kwon, "A binary classifier using fully connected neural network for Alzheimer’s disease classification," *Journal of Multimedia and Information Systems*, vol. 9, no. 1, pp. 21-32, 2022.

[2] S. Yang, H. Xu, Z. Yang, and C. Wang, "A mask wearing detection system based on deep learning," *Journal of Multimedia and Information Systems*, vol. 8, no. 3, pp. 159-166, 2021.

[3] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2016, pp. 3637-3645.

[4] F. Sung, Y, Yang, L. Zhang, T. Xiang, P. H. Philip, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199-1208.

[5] H. J. Ye, H. Hu, D. C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8808-8817.

[6] R. Hou, H. Chang, B. MA, S. Shan, and X. Chen, " Cross attention network for few-shot classification," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2019,vol. 32, pp. 4005-4016.

[7] C. D. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.

[8] J. Snell, K. Swersky, and R. Zemel, "Prototypical net-

works for few-shot learning," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2017, vol. 30, pp. 4077-4087.

[9] J. He, R. Hong, X. Liu, M. Xu, and Q. Sun, "Revisiting local descriptor for improved few-shot classification," *ACM Transactions on Multimedia Computing, Communications, and Application*, vol. 18, no, 2s, pp. 1-23, 2022.

[10] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4733-4742.

[11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," *Technical Report CNS-TR-2011-001*, 2011.

AUTHORS

Xilang Huang received the M.S. degree in electrical engineering from the Pusan National University, the Ph.D. degree in electrical engineering in Pukyong National University. Currently, he is a postdoctoral researcher in the Department of Electronic and Electrical Engineering at Ewha Womans University. His research interests include Few-Shot Learning, Domain Adaptation, and Model Compression Algorithms for Deep Neural Networks.

Seon Han Choi received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012, 2014, and 2018, respectively. In 2018, he was a Post-Doctoral Researcher with the Information and Electronics Research Institute, KAIST. From 2018 to 2019, he was a Senior Researcher with the Korea Institute of Industrial Technology (KITECH). From 2019 to 2022, he was an Assistant Professor in the Department of IT Convergence and Application Engineering, Pukyong National University. In 2022, he joined the Department of Electronic and Electrical Engineering, Ewha Womans University, as an Assistant Professor. His current research interests include modeling and simulation of discrete-event systems, efficient simulation optimization under stochastic noise, evolutionary computing, and machine learning.

Sungun Kim received the M.S. and Ph.D. degrees in the Department of Information and Communication Engineering from University Paris Diderot in 1990 and 1993 respectively. Currently, he is a Professor in the Department of Information and Communication Engineering at Pukyong National University. His research interests include wireless network security technology, transmission network and access network technology, IoT, and big data analytics with machine learning.