Task-Specific Few-Shot Image Classification by Balancing Sample-Level and Class-Level Generalization

Shi Tang	Zhiyi Xia	Changhua Chen	Jinying Du	Hui Wang
2021214082	2021214083	2021312596	2021214115	2021312586

Abstract

While new tasks could come with various compositions of base classes requiring high sample-level generalization ability and novel classes requiring high class-level generalization ability in realistic settings, most existing approaches for few-shot classification assume that all new tasks are composed of only novel classes that are not used for training. Due to such highly restrictive setting, they tend to handle different tasks equally regardless of the different importance of sample-level and class-level generalization in different tasks. To overcome this limitation by balancing sample-level and class-level generalization for each task, we propose an Attention-based Fusion Learning Model, in which: 1) features of normal pretraining (*Global*) utilizing all classes and episodic pretraining (*Local*) on meta-tasks within few classes are fused for considering both class-level and sample-level generalization; and 2) a Cross-Attention Module is designed to explore the similarities between *Global* and *Local* features for extracting more discriminative features for specific tasks, guiding the model to balance levels of generalization ability task-specifically. Extensive experiments confirm the effectiveness of our method.

1 Introduction

Although a wide variety of algorithms based on deep neural networks have achieved high performance in many fields in recent years, their actual usability is often hampered as massive labeled data is required for conventional deep learning methods. However, only limited data is available for a new task[1] in many realistic scenarios, leading to the investigation of few-shot learning [6, 17] which aims at learning new visual concepts with a few labeled data. Since the labeled data of a new task is limited, training a model with a large number of parameters is very challenging and most likely to cause over-fitting. A practical idea is to apply transfer learning [10]: pretrain the network on common classes (base classes) with sufficient samples, and then transfer the model to learn novel classes with only a few examples available.

Different from the setting of general transfer learning problems in Fig. 1 (c), few-shot problems usually have many target tasks, each with a few labeled data, rather than one target task with sufficient labeled data. And most existing few-shot classification approaches [12, 13, 17, 19, 20] have only targeted an artificial scenario where all tasks participating in the multi-class classification problem consist of only novel class samples (Fig. 1 (d)), which is a highly restrictive setting since a task may consist of only novel class samples or both base and novel class samples in real-world scenarios as shown in Fig. 1 (e).

Under a realistic setting, the sample-level and class-level generalization may have varying degrees of importance for samples from different classes in each task. For base class samples, it's more

[1] A task in few-shot classification refers to a specific classification problem, e.g., distinguishing phones from computers is a different task than distinguishing phones from TVs.

Deep Learning (Spring 2022), Tsinghua University.

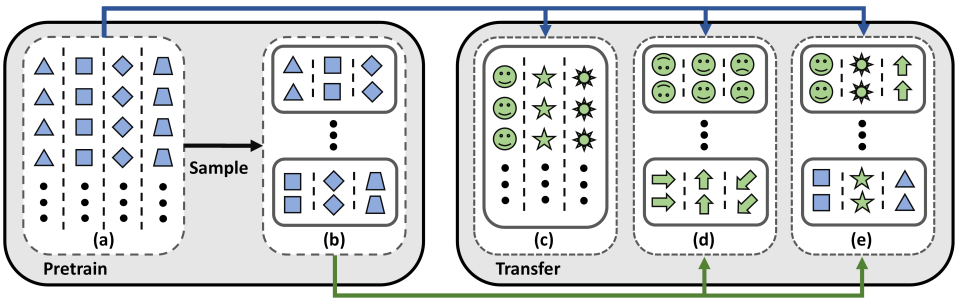


Figure 1: Non-sampling-based (a) and sampling-based (b) pretraining and the settings of (c) general transfer learning problems, few-shot problems in (d) artificial and (e) real-world scenarios.

of a normal sample-level generalization problem as they belong to the categories the model was trained with, only not used for training, while for novel class samples, it's a class-level generalization problem since the samples belong to completely new categories that are not used for training. Due to the different nature of the two problems, it's inappropriate to treat base and novel class samples the same in a task containing both of them. Even in tasks containing only novel class samples like the setting in most existing few-shot classification methods, dealing with them differently is also of vital importance as a novel class could be very similar or very different from base classes used for training, which could have an impact on the importance of sample-level and class-level generalization.

In the meantime, different tendencies for different generalizations of two common pretraining methods were observed recently [3]. The first paradigm, normal pretraining (*Global*, Fig. 1 (a)) which learns classifiers in the whole base class space with a straightforward purpose of maximize differences between all base classes, tends to perform better at distinguishing novel classes, indicating that *Global* features may come with better class-level generalization ability. While the other strategy, episodic pretraining (*Local*, Fig. 1 (b)) which learns across meta-tasks within few base classes sampled from the training set, tends to perform better at distinguishing base classes, indicating that *Local* features may come with better sample-level generalization ability.

Therefore, in response to the primary problem to be solved in this work, i.e., how to self-adaptively handle new tasks closer to real-world scenarios by exploring the classes of its support samples, we propose an Attention-based Fusion Learning Model as shown in Fig. 3 to address two main challenges: 1) Both the sample-level and class-level generalization problems need to be considered for a specific task; 2) Since the importance of the sample-level and class-level generalization differs for different tasks, different tasks need to be characterized and handled differently.

Specifically, a *Global* branch and a *Local* branch are learned simultaneously, in which we propose to fuse both *Global* and *Local* features to provide the model with the ability to solve both sample-level and class-level generalization problems well. Moreover, we design a Cross-Attention Module between the two branches to weight different features according to specific tasks, which can balance sample-level and class-level generalization explicitly. We validate our model based on *mini*ImageNet [17], under an artificial scenario where all new tasks consist of only novel classes and a real-world scenario where new tasks may contain base classes in addition to novel classes. The experimental results show that our Attention-based Fusion Learning Model significantly improves the performance over the existing task-agnostic methods under the real-world scenario. Further analysis of each component reveals the importance and indispensability of feature fusion and the Cross-Attention Module.

The main contributions of this work are summarized as follows:

- We propose an Attention-based Fusion Learning Model for few-shot classification in real-world scenarios, where new tasks may contain base classes in addition to novel classes. To the best of our knowledge, this is the first framework for few-shot classification in such a realistic setting.

- We propose to fuse both *Global* and *Local* features for considering both sample-level and class-level generalization according to the characteristics of new tasks in real-world scenarios.

- We design a Cross-Attention Module for balancing sample-level and class-level generalization task-specifically.
- We validate our model with a more realistic setting and show that it significantly outperforms existing task-agnostic methods.

2 Related Work

2.1 Metric Learning

The general objective of metric learning [18] is to learn a pairwise similarity metric $S(\cdot, \cdot)$ labeling similar sample pairs with higher scores and dissimilar ones with lower scores. Based on this, classification can be carried out on new tasks under the guidance of the "Nearest Neighbor" principle.

2.1.1 Sampling-based Pretraining

Following the idea of meta-learning, a promising technique for solving few-shot problems advocating learning across tasks to keep the objectives of training (meta-training) and testing (meta-testing) consistent as shown in Fig. 2, meta-training based methods [1, 4] usually sample (Fig. 1 (b)) the pretraining dataset to form many N-way K-shot tasks aligned with the form of meta-testing tasks. These approaches aim to learn at the level of tasks instead of samples.

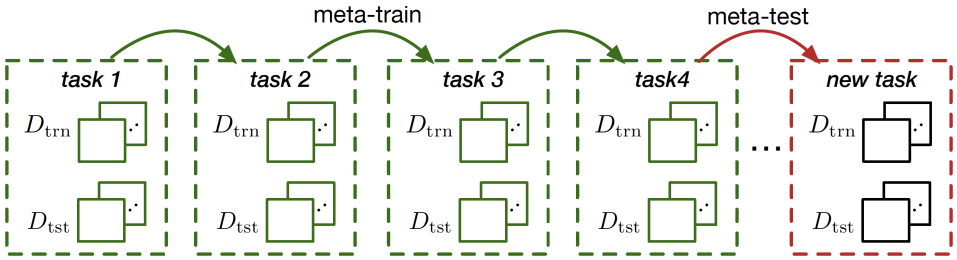


Figure 2: General framework of meta-learning based few-shot learning approaches [9].

2.1.2 Non-Sampling-based Pretraining

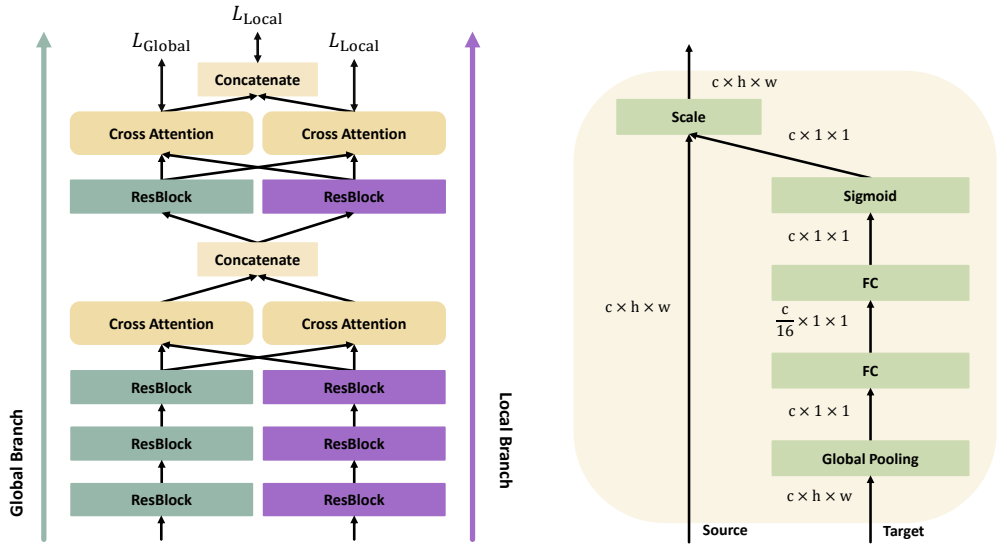
However, many recent works [2, 3, 14] have found that a model trained on the whole pretraining dataset directly (Fig. 1 (a)) actually provides comparable or even better embeddings. [2] performed this by replacing the top linear layer with a cosine classifier, and adapted the classifier to a few-shot classification task of novel classes by fine-tuning a new layer. To alleviate over-fitting on base patterns, the authors in [14] employed self-distillation strategy and data augmentation to constraint the learning process. Chen *et al.* [3] preliminarily explored these two kinds of features, i.e., features of the model using sampling-based pretraining (*Local*) and those of the model using non-sampling-based pretraining (*Global*), and found that *Global* features tend to have better class-level generalization ability while *Local* features show stronger sample-level generalization ability.

2.2 Attention

Inspired by the human brain's processing of information, attention mechanism [16] has achieved great success and it has already been widely used in the field of few-shot learning. For example, Fei *et al.* [5] proposed to adopt attention mechanism to explore the correlation between support samples and query samples for better establishing the connection between the support set and the query set. And Hou *et al.* [7] proposed to use attention mechanism to characterize the similarities between episodes containing the same set of classes for better utilizing different support sets. Likewise, it's also suitable for exploring the similarities between *Local* and *Global* features for extracting more discriminative features in our task, which may guide the model to self-adaptively balance levels of generalization ability.

3 Method

The base backbone of our network is ResNet-12. The architecture of the whole network is illustrated in Fig. 3(a). As it can be seen that the whole network is divided into a *Global* branch and a *Local* branch. In addition, a Cross-Attention Module as shown in Fig. 3(b) is proposed to weight *Global* and *Local* features respectively after the third and fourth blocks. Then, the weighted features are concatenated as the input of the next block or the final features for classification.



(a) The pipeline of the proposed network. (b) The structure of the Cross-Attention module.

Figure 3: The structure of our pipeline.

3.1 Cross-Attention Module

To fully utilize mutual information between global and local branches, a computational fusion method is needed. Inspired by squeeze and extraction operations proposed by [8], we come up with an efficient channel-wise Cross-Attention module which can make use of mutual information on the channel dimension.

As illustrated in Fig. 3(b), both source and target data flow $X_s, X_t \in \mathbb{R}^{c\times h\times w}$ are required as the input. And a transformation f_{sq} is applied to X_t, in which a global average pooling is first applied to X_t by:

$$g_i = Avg(X_t^i) = \frac{1}{H * W}\sum X_t^i, \tag{1}$$

where $i \in [1...c]$ denotes the channel index and the height and width of X_t^i are represented by H and W. To increase the nonlinearity with acceptable computational cost, a small bottleneck network with two fully-connected layers are used as well as LeakyReLU as an activation function. Then, Sigmoid is used for controlling the final output to be between 0 and 1. The bottleneck network can be represented as follows:

$$f_{sq}(x) = \sigma(W_2\delta(W_1 x)), \tag{2}$$

where σ and δ refer to Sigmoid and LeakyReLU respectively, weights $W_1 \in \mathbb{R}^{\frac{c}{16}\times c}$ and $W_2 \in \mathbb{R}^{c\times\frac{c}{16}}$. To exploit the channel dependencies from the target flow, $z \in \mathbb{R}^{c\times 1\times 1}$ is computed from $z = f_{sq}(g)$ and multiplied channel-wisely with F_s, as formulated below:

$$X_{output}^i = z^i * X_t^i, \tag{3}$$

where i denotes the channel index and z^i is a scalar multiplied with the matrix X_t^i.

3.2 Loss Function

As discussed in Sec. 2, in metric learning there are sampling (*Local*) and non-sampling (*Global*) based pretraining methods, which requires different loss functions. To fully exploit *Global* and *Local* information simultaneously at the pretraining stage, we need both Global and Local losses to supervise the training process.

Global loss. In the training process, a single layer linear classifier is attached to the final output of the Cross-Attention Module on the *Global* branch to generate a length-fixed probability vector. This loss helps training the *Global* branch to with better class-level generalization ability. As common classification methods, cross-entropy loss is used:

$$\mathcal{L}_{global}(X, Y) = CE(X, Y), \tag{4}$$

where X, Y represent predictions and ground-truth labels and CE refers to the cross entropy function which can be formulated as:

$$CE(X, Y) = \sum_{x_i \in X} -y_i \log p_i, \tag{5}$$

where

$$p_i = \frac{e^{-xi}}{\sum_j e^{-x_j}}. \tag{6}$$

Local loss. After sampling classes from the training set each time, in principle a new classifier can be trained using novel settings. However, the insufficiency of labeled data in the support set will leads to severe over-fitting. Following the basic idea of metric learning, the query samples are matched with support prototypes in the embedding space. For each class in the k^{th} task T_k, a prototype P_c of class c is calculated as the center of all labeled samples of this class:

$$P_c = \frac{1}{N_c} \sum_{y_i = c} \mathbf{x}_i \text{ , s.t. } c \in C_{T_k}, \tag{7}$$

where N_c is the number of labeled samples in class c and $\mathbf{x}_i$ is the vector of sample i in the embedding space. To classify unlabeled samples in the query set, a similarity function $S(\cdot, \cdot)$ is needed to calculate the similarity between each query sample vector q_i and all the class prototypes in the support set. Then the cross entropy loss is used as the loss function between the ground-truth label and the similarity vector for each query sample:

$$\mathcal{L}_{local}(X, Y) = \sum_{k=1}^{N_t} CE(S(P_{T_k}, X_{T_k}), Y_{T_k}), \tag{8}$$

where N_t refers to the number of tasks during the training process. We apply cosine distance as the similarity metric S, which can be achieved by a simply dot multiplication in implementation.

Total loss. Finally, the total loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{Global}(X_{global}, Y) + \mathcal{L}_{Local}(X_{local}, Y) + \mathcal{L}_{Local}(X_{cat}, Y), \tag{9}$$

where X_{global}, X_{local}, X_{cat} represent the output features of the *Global* branch, the *Local* branch and the final concatenated output.

4 Experiments

In this section, we answer the following questions:

- What's the difference between *Global* and *Local* features?
- Can the model benefit from fusing *Global* and *Local* features?
- Why is attention mechanism necessary?

Datasets. For all experiments in this work, we choose *mini*ImageNet [17], a common benchmark for few-shot learning as the dataset. It contains 100 classes sampled from ILSVRC-2012 [11], which are then split to 64, 16, 20 classes as training, validation and testing set respectively and each class contains 600 images of size 84×84. For the reconstructed *mini*ImageNet [17] in Tab. 1, the split is 64/32/40 with 300 images for each class, in which 16 classes in validation set and 20 classes in testing set are base classes. Note that the base class samples used for training and evaluation are not intersect.

Implementation details. We use ResNet12 as our backbone following previous works [3, 14]. It consists of 4 residual blocks and each block has 3 convolutional layers with 3×3 kernel and a 2×2 max-pooling layer. The number of filters are set to (64, 128, 256, 512). We fuse the features of *Global* and *Local* branches after the third and fourth blocks. The network is trained for 100 epochs using SGD optimizer with a momentum of 0.9 and a weight decay of 5e-4. The learning rate is initialized as 0.1 and decays at epoch 60 and 90 with a decay factor of 0.1. The number of batches for an epoch is set to 200 and each batch consists of 4 episodes with the form of N-way K-shot Q-query, which means the batch size of *Global* branch is $4 \times N \times (K + Q)$. And for the stability of the evaluation results, we test 8,000 episodes and report the average performance.

4.1 Comparison of *Global* and *Local* Features

To compare *Global* and *Local* features and verify the rationality of fusing them, we train two models using two kinds of supervision respectively. One with labels of the whole training set to make the model classify all classes of the training set (*Global*), while the other with labels in an episodic manner to make the model classify the sampled N classes (*Local*). The visualized results using t-SNE [15] are shown in Fig. 4. For base class samples, *Local* features (Fig. 4(b)) come with smaller intra-class differences compared to *Global* features (Fig. 4(a)), which can lead to better sample-level generalization. And for novel class samples, *Global* features (Fig. 4(c)) turn to achieve better embeddings with barely usable boundaries while samples embedded by *Local* (Fig. 4(d)) are completely mixed up.

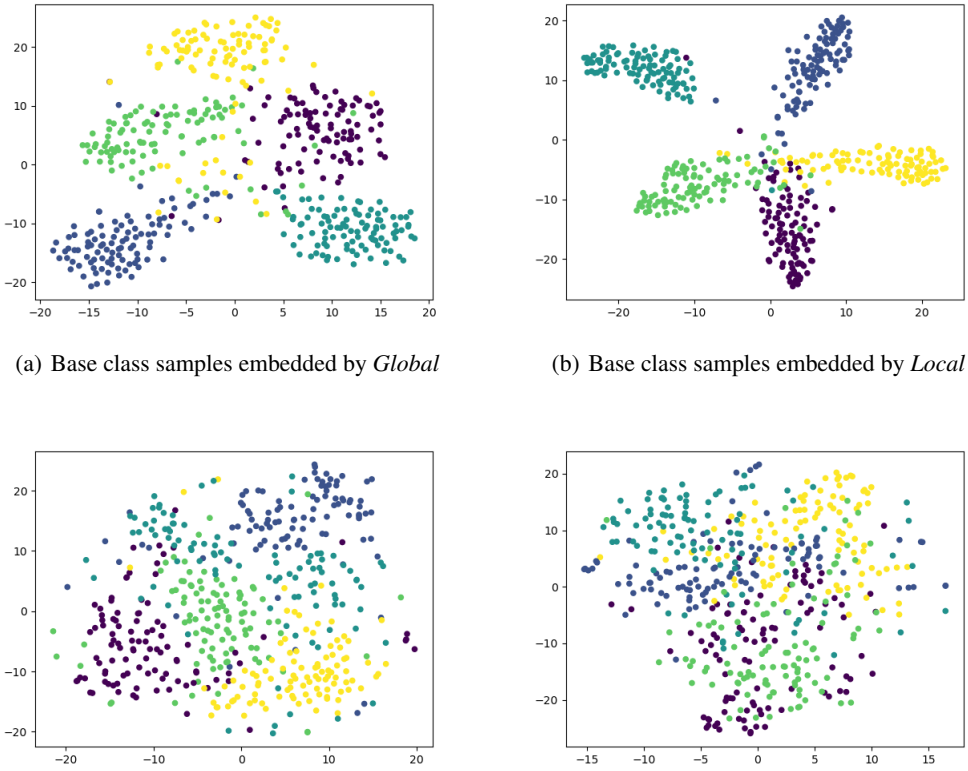
Consistent with the subjective results, it can be seen in Tab. 1 that *Global* performs better on *mini*ImageNet [17] which requires higher class-level generalization ability while *Local* performs better on reconstructed *min*ImageNet [17] which requires higher sample-level generalization ability.

Table 1: Average 5-way accuracy (%) with 95% confidence interval on *mini*ImageNet [17] (artificial scenario, all new tasks consist of only novel classes) and reconstructed *mini*ImageNet [17] (real-world scenario, new tasks could consist of both base and novel classes).

Model	*mini*ImageNet [17]		reconstructed *mini*ImageNet [17]	
	1-shot	5-shot	1-shot	5-shot
Global	58.80 ± 0.22	77.87 ± 0.17	61.03 ± 0.24	80.36 ± 0.18
Local	55.48 ± 0.23	73.62 ± 0.18	66.39 ± 0.28	81.93 ± 0.19
Meta-Baseline [3]	$\mathbf{61.95 \pm 0.23}$	78.50 ± 0.17	69.96 ± 0.27	83.62 ± 0.17
w/o attention	60.94 ± 0.23	79.14 ± 0.17	70.09 ± 0.26	83.99 ± 0.17
Ours	61.52 ± 0.23	$\mathbf{79.20 \pm 0.17}$	$\mathbf{70.98 \pm 0.27}$	$\mathbf{84.10 \pm 0.17}$

4.2 Effectiveness of Attention-based Fusion

Based on the above comparison, a significant difference in sample-level and class-level generalization between *Global* and *Local* features can be observed. And benefit from fusing both features for considering both sample-level and class-level generalization, our method significantly outperforms *Global* and *Local* as shown in Tab. 1, e.g., 10.89% and 7.58% higher than *Local* in accuracy on *mini*ImageNet [17] for 1-shot and 5-shot, respectively. Compared to Meta-Baseline [3] that simply pretrain the network in the *Global* setting and fine-tune it in the *Local* setting, our strategy of attention-based fusion can better exploit the two features, leading to better performance in most settings.



(a) Base class samples embedded by *Global*

(b) Base class samples embedded by *Local*

(c) Novel class samples embedded by *Global*

(d) Novel class samples embedded by *Local*

Figure 4: Visualization of base and novel class samples embedded by *Global* and *Local*, respectively.

4.3 Effectiveness of the Cross-Attention Module

For validating the effectiveness of the proposed Cross-Attention Module, we also compare our method with a 'w/o attention' version of our model as an ablation in Tab. 1. For *mini*ImageNet [17], since the similarities between different novel classes and base classes differs, the ability to characterize and handle different tasks differently according to their class composition brought by the Cross-Attention Module is conducive to better fusion of the two features, which explains why Cross-Attention Module can further improve the performance of the network. And for reconstructed *mini*ImageNet [17], the Cross-Attention Module plays a more important role since it's easier for unbalanced sample-level and class-level generalization to affect the performance under such a setting that tasks may contain base classes. So it's reasonable that the promotion brought by the Cross-Attention Module on reconstructed *mini*ImageNet [17] is bigger than that on *mini*ImageNet [17] in Tab. 1.

5 Discussion

Why are *Global* and *Local* Features Different? It's speculated that the difference of *Global* and *Local* features is caused by different classification mechanisms during training. For *Global*, the classification during training relies on a linear classifier to delineate decision boundaries, while for *Local*, the classification during training is metric-based. There is no classifier to delineate such decision boundaries explicitly, in which case narrower intra-class differences are more conducive to improving discrimination and reducing loss. However, narrowing the intra-class differences may compromise the generality of the features describing samples from different classes, resulting in poor class-level generalization ability. It can be thought of as an over-fitting on classification tasks consisting of only base classes, which explains why it has better sample-level generalization ability.

And with an additional classifier to share part of the burden of classification, *Global* can focus more on how to embed samples to a more suitable location in the feature space with respect to their characteristics, thus avoiding sacrificing the ability of features to describe novel class samples.

How to Further Improve Class-Level Generalization? It can be seen from both Fig. 4 and Tab. 1 that compared with sample-level generalization, class-level generalization is a more challenging problem restricting the performance of few-shot classification, which makes it an important perspective to improve the performance of the model. Since the model doesn't know whether a sample is from the base class or the novel class and will embed it to the feature space for distinguishing base classes. Maybe treating novel classes as pseudo base classes is a feasible way. Assuming that novel classes (pseudo base classes) could be represented by combinations of base classes, samples can be characterized by their similarities to the prototypes of base classes, which may make novel classes more distinguishable in such a feature space.

6 Conclusion

Targeting at few-shot classification tasks in real-world scenarios where new tasks may contain both base and novel classes, we propose an Attention-based Fusion Learning Model which fuses both *Global* and *Local* features weighted by a proposed Cross-Attention Module to balance sample-level and class-level generalization task-specifically. The effectiveness of the Attention-based Fusion Learning Model has been demonstrated by experiments on a common benchmark. Further discussion analyzes the possible reason for the difference between *Global* and *Local* features and proposes a potential direction for improving class-level generalization ability.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

[3] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9062–9071, October 2021.

[4] Yutian Chen, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matthew Botvinick, and Nando de Freitas. Learning to learn without gradient descent by gradient descent. In *ICML*, pages 748–756, 2017.

[5] Nanyi Fei, Zhiwu Lu, Tao Xiang, and Songfang Huang. {MELR}: Meta-learning via modeling episode-level relationships for few-shot learning. In *International Conference on Learning Representations*, 2021.

[6] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[7] Ruibing Hou, Hong Chang, Bingpeng MA, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang. Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653*, 2020.

[10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[12] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 266–282, Cham, 2020. Springer International Publishing.

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[17] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[18] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

[19] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3754–3762, June 2021.

[20] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.