Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification

(Supplementary Materials)

Fusheng Hao\(^1,2\)  Fengxiang He\(^3\)  Liu Liu\(^4\)  Fuxiang Wu\(^1,2\)  Dacheng Tao\(^4\)  Jun Cheng\(^1,2\)*

\(^1\)Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
\(^2\)The Chinese University of Hong Kong, Hong Kong, China
\(^3\)AIAI, School of Informatics, University of Edinburgh, United Kingdom
\(^4\)School of Computer Science, Faculty of Engineering, The University of Sydney, Australia

Additional details or results are collected in this appendix.

A. Datasets

Experiments are conducted on four popular benchmark datasets: miniImageNet [20, 15], tieredImageNet [16], CIFAR-FS [1], and FC-100 [13].

**miniImageNet.** This dataset is very popular in few-shot image classification. It is a subset of ImageNet [17] and contains 100 classes with 600 images per class. We use the split provided in [15] to divide the 100 classes into disjoint 64 training, 16 validation, and 20 test classes.

**tieredImageNet.** This dataset is also a subset of ImageNet [17] and the hierarchical structure of ImageNet is used when it is created. It includes 608 classes from 34 super-classes, with a total of 770,165 images. The 34 super-classes are divided into disjoint 20 training, 6 validation and 8 test super-classes to achieve better separation. Correspondingly, the 608 classes are divided into disjoint 351 training, 97 validation, and 160 test classes.

**CIFAR-FS.** This dataset is created based on CIFAR100 [8]. It includes 100 classes with 600 images per class. The 100 classes are divided into disjoint 64 training, 16 validation, and 20 test classes.

**FC-100.** This dataset is also created based on CIFAR100 [8] and includes 100 classes with 600 images per class. It uses a split strategy similar to tieredImageNet to increase the difficulty of the resulting few-shot image classification tasks. Correspondingly, the 100 classes are divided into disjoint 60 training, 20 validation, and 20 test classes.

B. Implementation details

Our training procedure consists of two stages: pretraining and finetuning. It is to be noted that only the training set of the corresponding dataset is used for pretraining and finetuning.

**Pretraining.** We pretrain the backbone, \textit{i.e.,} ViTS/16, by using the strategy proposed in [25, 6] and mostly sticking to the hyperparameter settings reported. Specifically, two global crops with a crop scale of (0.4, 1.0) and 10 local crops with a crop scale of (0.05, 0.4) are used. The output dimension of the project head is set to be 8,192-d. The prediction ratios and variances of random Masked Image Modelling is set to be (0, 0.3) and (0, 0.2), respectively. The resolution of input images is set to be 224 × 224. Four A100 40G GPUs are used to pretrain the ViT-S/16 and the total number of training epochs is set to be 1,600. To match our computing resources, the batch size is set to be 512. The optimizer used is AdamW [9] and the linearly ramped-up learning rate of \(5e^{-4} \times \text{batchsize}/256\) is carried out in the first 10 epochs.

**Finetuning.** After pretraining, the project head with an output dimension of 8,192-d is removed and a new project head with an output dimension of 384-d is added to the ViTS/16. The whole pipeline is finetuned by minimizing Eq. (6) using the episodes randomly sampled from the train-
Figure A2. Patch embedding visualization of four more randomly sampled 5-way 1-shot classification tasks with one query image per class. (a), (b), (c), and (d) show the visualization results without CPEA. (e), (f), (g), and (h) show the corresponding visualization results with CPEA. CPEA concentrates patch embeddings by class, thus making them class-relevant. Experiments are conducted on miniImageNet.

Figure A3. Class-aware embedding visualization. Two more different sampling results are given in (a) and (b), respectively, with 100 class-aware embeddings per class. The class-agnostic embedding is denoted by the black “diamond”. After interacting with images from different classes, the output states of class-agnostic embedding are class-aware. Experiments are conducted on miniImageNet.

Table A1. Impact of the model size on the few-shot classification performance. Experiments are conducted on miniImageNet.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>#Params</th>
<th>5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProtoNet [19]</td>
<td>ResNet-12</td>
<td>≈12.4M</td>
<td>79.46±0.48</td>
</tr>
<tr>
<td>FEAT [22]</td>
<td>ResNet-12</td>
<td>≈12.4M</td>
<td>82.05±0.14</td>
</tr>
<tr>
<td>DeepEMD [23]</td>
<td>ResNet-12</td>
<td>≈12.4M</td>
<td>82.41±0.56</td>
</tr>
<tr>
<td>COSOC [10]</td>
<td>ResNet-12</td>
<td>≈12.4M</td>
<td>85.16±0.42</td>
</tr>
<tr>
<td>DeepBDC [21]</td>
<td>ResNet-12</td>
<td>≈12.4M</td>
<td>84.46±0.28</td>
</tr>
<tr>
<td>LEO [18]</td>
<td>WRN-28-10</td>
<td>≈36.5M</td>
<td>77.59±0.12</td>
</tr>
<tr>
<td>CC+rot [5]</td>
<td>WRN-28-10</td>
<td>≈36.5M</td>
<td>79.87±0.33</td>
</tr>
<tr>
<td>FEAT [22]</td>
<td>WRN-28-10</td>
<td>≈36.5M</td>
<td>81.11±0.14</td>
</tr>
<tr>
<td>PSST [3]</td>
<td>WRN-28-10</td>
<td>≈36.5M</td>
<td>80.64±0.32</td>
</tr>
<tr>
<td>MetaQDA [24]</td>
<td>WRN-28-10</td>
<td>≈36.5M</td>
<td>84.28±0.69</td>
</tr>
<tr>
<td>OM [14]</td>
<td>WRN-28-10</td>
<td>≈36.5M</td>
<td>85.29±0.41</td>
</tr>
<tr>
<td>FewTURE [6]</td>
<td>ViT-S/16</td>
<td>≈5.0M</td>
<td>81.10±0.61</td>
</tr>
<tr>
<td>FewTURE [6]</td>
<td>ViT-T/16</td>
<td>≈22.0M</td>
<td>84.51±0.53</td>
</tr>
<tr>
<td>CPEA (ours)</td>
<td>ViT-T/16</td>
<td>≈5.0M</td>
<td>84.62±0.39</td>
</tr>
<tr>
<td>CPEA (ours)</td>
<td>ViT-S/16</td>
<td>≈22.0M</td>
<td>87.06±0.38</td>
</tr>
</tbody>
</table>

(a) Task 1 without CPEA
(b) Task 2 without CPEA
(c) Task 3 without CPEA
(d) Task 4 without CPEA

Figure A2. Patch embedding visualization of four more randomly sampled 5-way 1-shot classification tasks with one query image per class. (a), (b), (c), and (d) show the visualization results without CPEA. (e), (f), (g), and (h) show the corresponding visualization results with CPEA. CPEA concentrates patch embeddings by class, thus making them class-relevant. Experiments are conducted on miniImageNet.

C. Ablation study

Convergence property. Figure A1 shows the finetuning losses of our CPEA. It can be observed that our CPEA converges after 8k episodes in both the 1-shot and 5-shot settings. These observations indicate that our CPEA could converge in a proper time.

Patch embedding visualization. More patch embed-
parably small (e.g., 38.4K images in miniImageNet [20] vs. 1.28M images in ImageNet [17]), it has been shown that adopting bigger networks does not help improve performance [2, 11]. In other words, increasing the number of parameters on its own does not lead to better few-shot classification performance. Table A1 shows the impact of the model size on few-shot classification performance. It can be observed that 1) With less than one seventh of the number of parameters of WRN-28-10, our CPEA achieves comparable performance. 2) With roughly the same number of parameters of WRN-28-10, our CPEA outperforms the counterparts by a large margin. 3) Our CPEA benefits from the increased model size.

**Image resolution.** We follow the common practice of ViTs [4] to take images with a resolution of 224×224 as input, which is higher than traditional CNN-based few-shot image classification methods [23, 26, 12] (e.g., 84×84). It is to be noted that increasing the image resolution on its own does not lead to better few-shot classification performance. Table A2 shows the impact of the image resolution on the few-shot classification performance. It can be observed that higher resolution always leads to performance degradation.

### Table A2. Impact of the image resolution on the few-shot classification performance. Experiments are conducted on miniImageNet.

<table>
<thead>
<tr>
<th>Model</th>
<th>Image resolution</th>
<th>5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepEMD [23]</td>
<td>84×84 → 224×224</td>
<td>82.41 → 78.12</td>
</tr>
<tr>
<td>BML [26]</td>
<td>84×84 → 224×224</td>
<td>83.59 → 81.57</td>
</tr>
<tr>
<td>IE [12]</td>
<td>84×84 → 224×224</td>
<td>84.35 → 79.12</td>
</tr>
<tr>
<td>CPEA (ours)</td>
<td>224×224</td>
<td>87.06</td>
</tr>
</tbody>
</table>

References


