Few-Shot Classification with Contrastive Learning Supplementary Material

Zhanyuan Yang[1], Jinghua Wang[2], and Yingying Zhu⋆[1]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
yangzhanyuan2019@email.szu.edu.cn, zhuyy@szu.edu.cn
[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China
wangjinghua@hit.edu.cn

1 Additional Ablation Study

We conduct ablation studies in the pre-training stage on the miniImageNet. Here, except for the model trained with only L_{CE} (see the first line in Table 1), we adopt cross-view episodic training (CVET) mechanism and distance-scaled contrastive loss in the meta-training stage for all experiments in Table 1 and Table 2.

Table 1: Ablation experiments in the pre-training stage on miniImageNet.

L_{CE}	L^{ss}_{global}	L^{ss}_{local}	L^{s}_{global}	**1-shot**	**5-shot**
√				66.58 ± 0.46	81.92 ± 0.31
√	√			68.53 ± 0.46	83.86 ± 0.29
√		√		69.33 ± 0.46	84.02 ± 0.30
√			√	67.88 ± 0.45	83.37 ± 0.29
√	√	√		69.03 ± 0.46	83.87 ± 0.30
√	√		√	68.68 ± 0.46	84.17 ± 0.29
√		√	√	69.72 ± 0.45	84.49 ± 0.29
√	√	√	√	70.19 ± 0.46	84.66 ± 0.29

As shown in Table 1, compared to training with L_{CE} alone, each contrastive loss used in the pre-training stage plays an important role, with L^{ss}_{local} contributing the most. The results from the methods introducing L^{ss}_{local} indicate that contrastive learning based on extra local information can learn more generalizable representations. Meanwhile, we can observe that the results obtained by training with supervision only are much lower than the results obtained by

⋆ Corresponding author.

Table 2: Effectiveness of vector-map and map-map modules on miniImageNet.

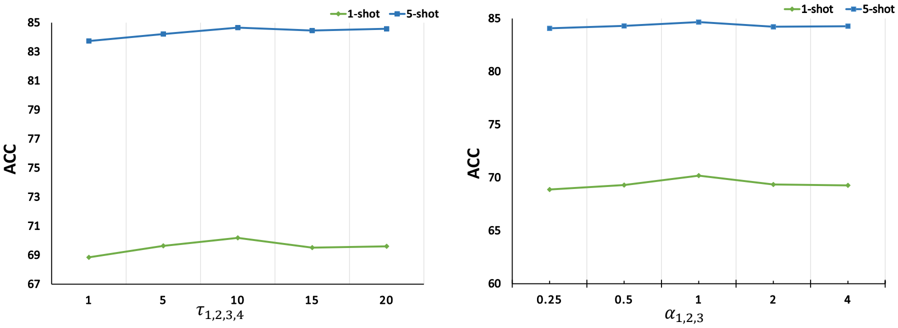
$L_{CE} + L^{ss}_{global} + L^{s}_{global}$	$L^{ss}_{vec-map}$	$L^{ss}_{map-map}$	**1-shot**	**5-shot**
√			68.68 ± 0.46	84.17 ± 0.29
√	√		69.29 ± 0.46	84.23 ± 0.29
√		√	68.81 ± 0.46	84.25 ± 0.29
√	√	√	70.19 ± 0.46	84.66 ± 0.29

using both supervision and self-supervision. The results in Table 1 validate the effectiveness of our proposed contrastive losses, and we obtain the best results when employing all proposed contrastive losses in the pre-training stage.

Based on L_{CE} and the other two contrastive losses using global information, we verify the effectiveness of vector-map and map-map modules by conducting experiments on them separately, as shown in Table 2. Compared to using only global information, contrastive learning that leverages local information either in the form of vector-map or map-map can improve the transferability of the representations. The best results are achieved when both forms work together.

2 Experiments on Hyperparameters in the Pre-training

We investigate the effect of hyperparameters in the pre-training stage on classification performance. Note that we use the inverse temperature parameters in our implementation. To make it easy to understand, we draw the graph according to the inverse temperature parameters. That is, the horizontal axis of Fig. 1a actually represents the inverse of $\tau_{1,2,3,4}$. As shown in Fig. 1, we empirically find that the best results can be achieved by setting the four temperature parameters $\tau_{1,2,3,4}$ to 0.1 and the three balance scalars $\alpha_{1,2,3}$ to 1.0 at the same time. The results indicate that our pre-training approach is less sensitive to the two kinds of hyperparameters (within a certain range), which means it is robust. Thus it is effortless to apply our approach to other methods.



(a) Fix $\alpha_{1,2,3} = 1.0$

(b) Fix $\tau_{1,2,3,4} = 0.1$

Fig. 1: Effect of hyperparameters $\tau_{1,2,3,4}$ and $\alpha_{1,2,3}$ on miniImageNet.

3 Quantitative Analysis of Model Parameters

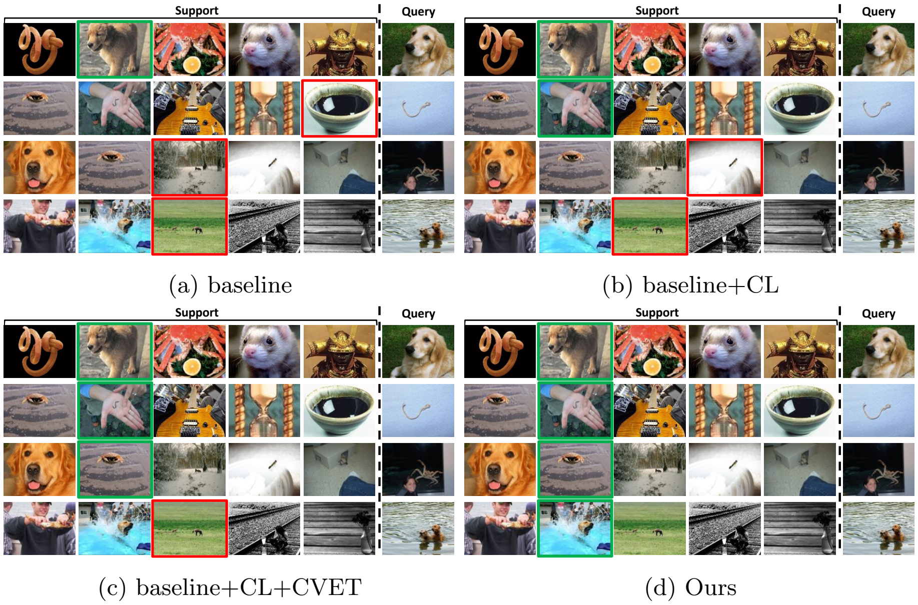
In the case of using the same backbone network, we compare the number of parameters of our proposed method with the two-stage methods ProtoNet [1] and FEAT [2]. The results are shown in Table 3. In the pre-training stage, the number of parameters of our method is 1.9M more than ProtoNet and baseline (FEAT) due to the introduction of projection heads and a fully connected layer for our contrastive losses. In the meta-training stage, a single projection head used in distance-scaled contrastive loss results in 0.4M more parameters for our method than baseline. Furthermore, during meta-testing, the number of parameters of our method is the same as in baseline. Therefore, the small number of additional parameters we introduce in the pre-training and meta-training stages is acceptable considering the improved classification performance of our method.

Table 3: Comparison of the number of parameters for several methods.

Stage	Method	Backbone	# params
Pre-training	ProtoNet [1]	ResNet-12	12.5M
	FEAT [2]		12.5M
	Ours		14.4M
Meta-training	ProtoNet [1]		12.4M
	FEAT [2]		14.1M
	Ours		14.5M
Meta-testing	ProtoNet [1]		12.4M
	FEAT [2]		14.1M
	Ours		14.1M

4 Qualitative Ablation Study

In this section, we give qualitative analysis to verify the effectiveness of each component proposed in our method. The classification results of our components are shown in Fig. 2, Thanks to the successive use of each component, the model can better adapt to novel tasks with more pictures in which the background is dominant. Moreover, the accurate recognition of small objects reflects that our framework enables the representations to learn meta-knowledge that is useful for few-shot classification. Therefore, it can be considered that each of our proposed components is effective.



(a) baseline

(b) baseline+CL

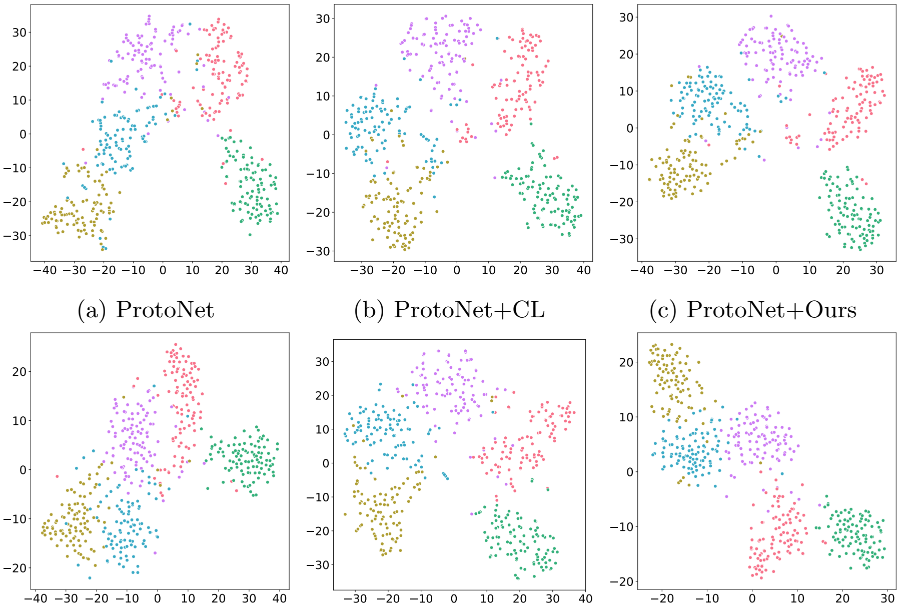
(c) baseline+CL+CVET

(d) Ours

Fig. 2: 5-way 1-shot classification results on miniImageNet. The green box indicates the correct classification result, while the red box indicates the incorrect result.

5 Visualization and Quantitative Analysis of Feature Embeddings

In this section, we present the complete visualization results of Sect. 4.4 in the paper. The results from Fig. 3 illustrate that our proposed framework generates more transferable and discriminative representations on novel classes.



(a) ProtoNet

(b) ProtoNet+CL

(c) ProtoNet+Ours

(d) baseline

(e) baseline+CL

(f) Ours

Fig. 3: Visualization of 100 randomly sampled images for each of the 5 meta-test classes from miniImageNet by t-SNE.

We further give quantitative analysis below. We use the same sampling strategy as in the visualization experiments above and divide the training and test sets in a 1:4 ratio. Then the features extracted by the four methods ProtoNet, ProtoNet+Ours, baseline and Ours are classified with SVM. The mean accuracies of these methods are shown in Table 4. The results indicate that the two-stage methods combined with our framework make feature embeddings more separable.

Table 4: Quantitative analysis of representations using SVM.

Method	Mean accuracy
ProtoNet	0.87
ProtoNet+Ours	0.91
baseline	0.89
Ours	**0.98**

References

1. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: NIPS. pp. 4077–4087 (2017)
2. Ye, H., Hu, H., Zhan, D., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: CVPR. pp. 8805–8814 (2020)