SELF-SUPERVISED LEARNING FOR FEW-SHOT IMAGE CLASSIFICATION

Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, Hui Xue

chen.cd, yuefeng.chenyf, daniel.lyh, maofeng.mf, heyuan.hy, hui.xueh@alibaba-inc.com

Motivation

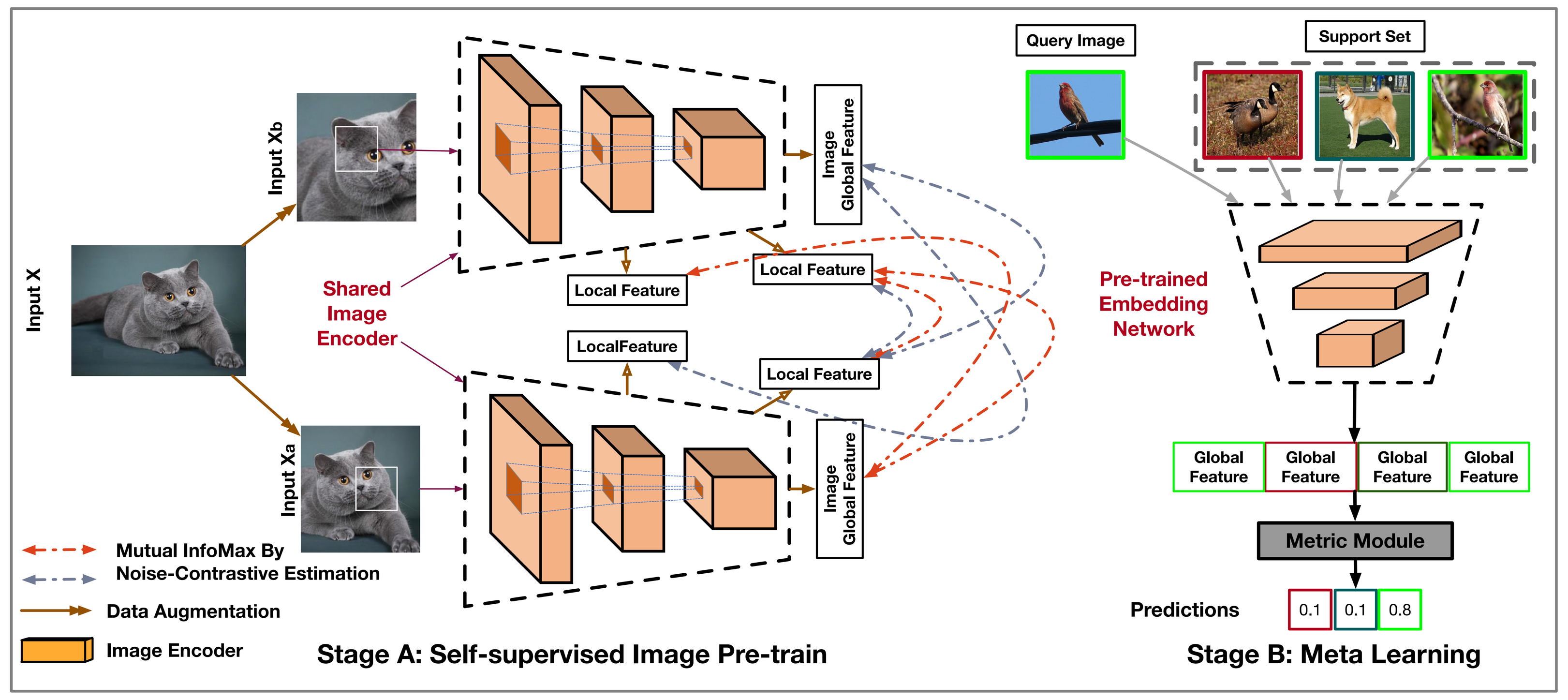
Difficulties and key problems in few-shot learning image classification

- Limited labelled data for training
- Novel classes during test comparing to training
- Highly relied on the quality of pretraining on the backbone

A robust few-shot learning method will benefit

- Tasks that have very limited data for training such as Medical Images
- Who cannot afford for very expensive annotation

Overall Architecture of Proposed Method



Our Contribution

- A simple framework that applies self-supervised learning to learn a very deep backbone for few-shot learning image classification task
- State-of-the-art results in
  - 5-way 1-shot & 5-way 5-shot in MiniImageNet dataset[1]
  - 5-way 1-shot & 5-way 5-shot in CUB dataset[2]
  - Cross-domain few-shot learning task[3]
- Code is available at https://github.com/phecy/SSL-FEW-SHOT

Few-shot learning Pipeline in details

General Pipeline

General pipeline for most of existing methods with good performance

- Pre-train the backbone on training set.
- Meta-learning based training with pretrained backbone on training set
- Test the performance of the solution by training with limited data(1-shot or 5-shot) with novel classes(5-way) in test set and testing on query samples in these novel classes.

Self-supervised learning stage

The core is to maximize mutual information between global features and local features from two views (x_a, x_b) of the same image. The NCE loss is defined as:

$$\mathcal{L}_{ssl}\left(f_g(x_a), f_5(x_b)\right) = - \log \frac{\exp\{\phi(f_g(x_a), f_5(x_b))\}}{\sum_{\widetilde{x_b} \in \mathcal{N}_x \cup x_b} \exp\{\phi(f_g(x_a), f_5(\widetilde{x_b}))\}}$$

N_x are the negative samples of image x, ϕ is the distance metric function. At last, the overall loss between x_a and x_b is as follows:

$$\mathcal{L}_{ssl}(x_a, x_b) = \mathcal{L}_{ssl}\left(f_g(x_a), f_5(x_b)\right) + \mathcal{L}_{ssl}\left(f_g(x_b), f_7(x_b)\right) + \mathcal{L}_{ssl}\left(f_5(x_a), f_5(x_b)\right)$$

For more details, please refer to the main paper and [4]

Meta-learning stage

The representation of class k is represented by the centroid of embedding features of training samples and can be obtained as:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S} f(x_i)$$

A distance function d and produce a distribution over all classes given a query sample q from the query set Q

$$p(y = k|q) = \frac{\exp(-d(f(q), c_k))}{\sum_{k'} \exp(-d(f(q), c_{k'}))}$$

The loss in the meta-learning stage is set as:

$$\mathcal{L}_{meta} = d(f(q), c_k) + \log \sum_{k'} d(f(q), c_{k'})$$

In a short conclusion, during training, the proposed method first applies an SSL way to pre-train a large scale embedding network in stage one, followed by a detailed finetuning in stage two with a meta-learning scheme.

Datasets

MiniImageNet[1]: 60,000 images from 100 classes, 64 classes for training, 16 classes for validation, 20 classes for the test.
CUB[2]: 11788 images from 200 classes, 100 classes for training, 50 classes for validation, and 50 classes for the test.
Cross-domain few-shot learning: 1) CropDiseases [25], a plant diseases dataset, 2) EuroSAT [26], a dataset for satellite images, 2) ISIC [27] a medical skin image dataset, 4) ChestX [28], a dataset for X-ray chest images.

Experimental Results

Comparison to the state of the art:

Baselines	Embedding Net	1-Shot 5-Way	5-Shot 5-Way
MatchingNet [11]	4 Conv	$43.56 \pm 0.84\%$	$55.31 \pm 0.73\%$
MAML [12]	4 Conv	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$
RelationNet [13]	4 Conv	$50.44 \pm 0.82\%$	$65.32 \pm 0.70\%$
REPTILE [14]	4 Conv	$49.97 \pm 0.32\%$	$65.99 \pm 0.58\%$
ProtoNet [15]	4 Conv	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$
Baseline* [29]	4 Conv	$41.08 \pm 0.70\%$	$54.50\% \pm 0.66$
Spot&learn [30]	4 Conv	$51.03 \pm 0.78\%$	$67.96\% \pm 0.71$
DN4 [31]	4 Conv	$51.24 \pm 0.74\%$	$71.02\% \pm 0.64$
SNAIL [32]	ResNet12	$55.71 \pm 0.99\%$	$68.88 \pm 0.92\%$
ProtoNet$^+$ [15]	ResNet12	$56.50 \pm 0.40\%$	$74.2 \pm 0.20\%$
MTL [33]	ResNet12	$61.20 \pm 1.8\%$	$75.50 \pm 0.8\%$
DN4 [31]	ResNet12	$54.37 \pm 0.36\%$	$74.44 \pm 0.29\%$
TADAM [2]	ResNet12	58.50%	76.70%
Qiao-WRN [3]	Wide-ResNet28	$59.60 \pm 0.41\%$	$73.74 \pm 0.19\%$
LEO [4]	Wide-ResNet28	$61.76 \pm 0.08\%$	$77.59 \pm 0.12\%$
Dis. k-shot [7]	ResNet34	$56.30 \pm 0.40\%$	$73.90 \pm 0.30\%$
Self-Jig(SVM) [8]	ResNet50	$58.80 \pm 1.36\%$	$76.71 \pm 0.72\%$
FEAT [34]	ResNet50	53.8%	76.0%
Ours_Mini80_SL	AmdimNet	$43.92 \pm 0.19\%$	$67.13 \pm 0.16\%$
Ours_Mini80_SSL$^-$	AmdimNet	$46.13 \pm 0.17\%$	$70.14 \pm 0.15\%$
Ours_Mini80_SSL	AmdimNet	$\mathbf{64.03 \pm 0.20\%}$	$\mathbf{81.15 \pm 0.14\%}$
Ours_Image900_SSL	AmdimNet	$\mathbf{76.82 \pm 0.19\%}$	$\mathbf{90.98 \pm 0.10\%}$

Table 1. Few-shot classification accuracy results on *MiniImageNet*. '−' indicates result without meta-learning.

Baselines	Embedding Net	1-Shot 5-Way	5-Shot 5-Way
MatchingNet [11]	4 Conv	61.16 ± 0.89	72.86 ± 0.70
MAML [12]	4 Conv	$55.92 \pm 0.95\%$	$72.09 \pm 0.76\%$
ProtoNet [15]	4 Conv	$51.31 \pm 0.91\%$	$70.77 \pm 0.69\%$
MACO [24]	4 Conv	60.76%	74.96%
RelationNet [13]	4 Conv	$62.45 \pm 0.98\%$	$76.11 \pm 0.69\%$
Baseline++ [29]	4 Conv	$60.53 \pm 0.83\%$	$79.34 \pm 0.61\%$
DN4-DA [31]	4 Conv	$53.15 \pm 0.84\%$	$81.90 \pm 0.60\%$
Ours_CUB150_SL	AmdimNet	$45.10 \pm 0.21\%$	$74.59 \pm 0.16\%$
Ours_CUB150_SSL$^-$	AmdimNet	$40.83 \pm 0.16\%$	$65.27 \pm 0.18\%$
Ours_CUB150_SSL	AmdimNet	$\mathbf{71.85 \pm 0.22\%}$	$\mathbf{84.29 \pm 0.15\%}$
Ours_Image1K_SSL	AmdimNet	$\mathbf{77.09 \pm 0.21\%}$	$\mathbf{89.18 \pm 0.13\%}$

Table 2. Few-shot classification accuracy results on CUB dataset [23]. '−' indicates result without meta-learning. For each task, the best-performing method is highlighted.

To prove the effectiveness of the proposed method, we train the embedding network with labelled data (Mini80-SL and CUB150-SL as detailed in Section 4.2). As shown in Table 1 and Table 2, it performs even worse than the methods with simple 4 Conv blocks embedding networks as such big network under supervised learning with limited data can cause overfitting problem and cannot adjust to new unseen classes during testing. However, with SSL based pre-training a more generalized embedding network can be obtained and improve the results significantly. One may also concern about the effectiveness of the meta-learning fine-tuning in the second stage. To test this, the pre-train embedding network is directly applied to the task with the nearest neighbourhood(NN) classification. As shown in the test results on both dataset, meta-learning can effectively fine-tune the embedding network and achieve remarkable improvement.

Methods	ChestX			ISIC		
	5-way 5-shot	5-way 20-shot	5-way 50-shot	5-way 5-shot	5-way 20-shot	5-way 50-shot
Ours_trans	$\mathbf{28.50 \pm 0.40\%}$	$\mathbf{33.79 \pm 0.48\%}$	$\mathbf{38.78 \pm 0.64\%}$	$44.15 \pm 0.52\%$	$55.63 \pm 0.49\%$	$62.76 \pm 0.50\%$
Cross [9]	$26.09 \pm 0.96\%$	$31.01 \pm 0.59\%$	$36.79 \pm 0.53\%$	$\mathbf{49.68 \pm 0.36\%}$	$\mathbf{61.09 \pm 0.44\%}$	$\mathbf{67.20 \pm 59\%}$
	EuroSAT			CropDiseases		
	5-way 5-shot	5-way 20-shot	5-way 50-shot	5-way 5-shot	5-way 20-shot	5-way 50-shot
Ours_trans	$\mathbf{83.44 \pm 0.61\%}$	$\mathbf{90.43 \pm 0.52\%}$	$\mathbf{94.71 \pm 0.47\%}$	$\mathbf{91.79 \pm 0.48\%}$	$\mathbf{97.38 \pm 0.65\%}$	$\mathbf{99.50 \pm 0.63\%}$
Cross[9]	$81.76 \pm 0.48\%$	$87.97 \pm 0.42\%$	$92.00 \pm 0.56\%$	$90.64 \pm 0.54\%$	$95.91 \pm 0.72\%$	$97.48 \pm 0.56\%$

Table 3. Cross-domain few-shot learning tests on four datasets

Acknowledgement

The authors are supported by Alibaba Group

Reference

[1] Oriol Vinyals,Charles Blundell,Timothy Lillicrap,Daan Wierstra, et al., “Matching networks for one shot learning,” in NeurIPS, 2016, pp. 3630–3638
[2] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset,” Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011
[3] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris, “A broader study of cross-domain few-shot learning,” ECCV, 2020
[4] Philip Bachman, R Devon Hjelm, and William Buchwalter, “Learning representations by maximizing mutual information across views,” arXiv preprint arXiv:1906.00910, 2019

2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

阿里安全 ALIBABA SECURITY