Few-Shot Learning with Online Self-Distillation

Sihan Liu*[1] Yue Wang*[2]
[1]Boston University [2]Massachusetts Institute of Technology

2021 ICCV OCTOBER 11-17 VIRTUAL

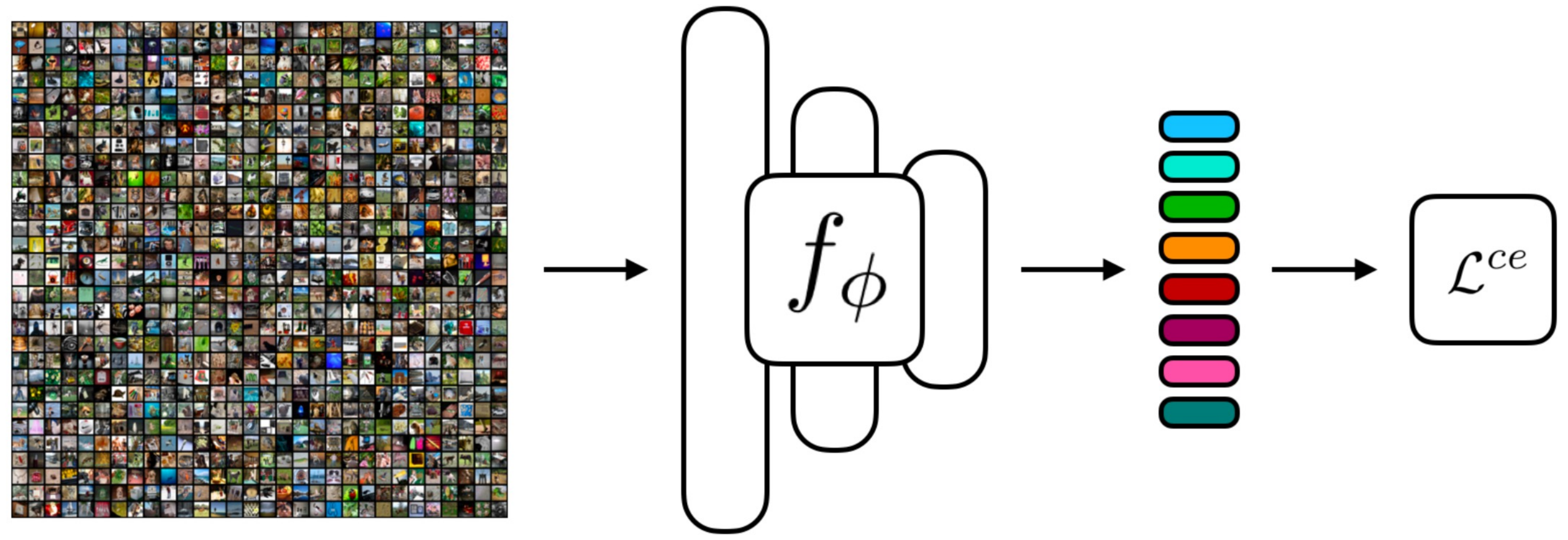
Background

Meta learning

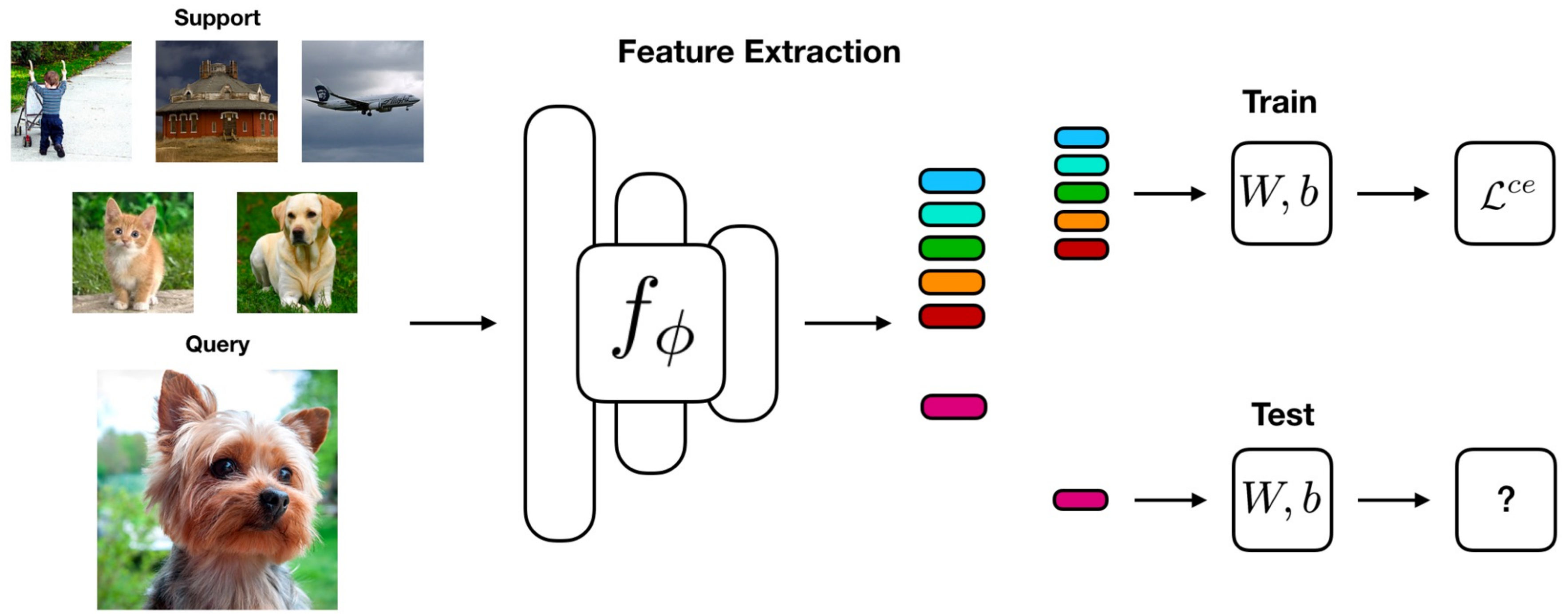
The goal of meta learning is to learn a meta model on a distribution of tasks, which can generalize to novel tasks. In meta learning, the training set and the testing set do not share the same categories. Meta learning methods include learning a good metric, optimizer, or a fast adaptation algorithm.

Learning representations for meta learning

In RFS[a], a simple baseline algorithm that learns representations for meta learning/few-shot learning has been proposed. In training, a classification model is trained in a supervised manner, shown as follows.

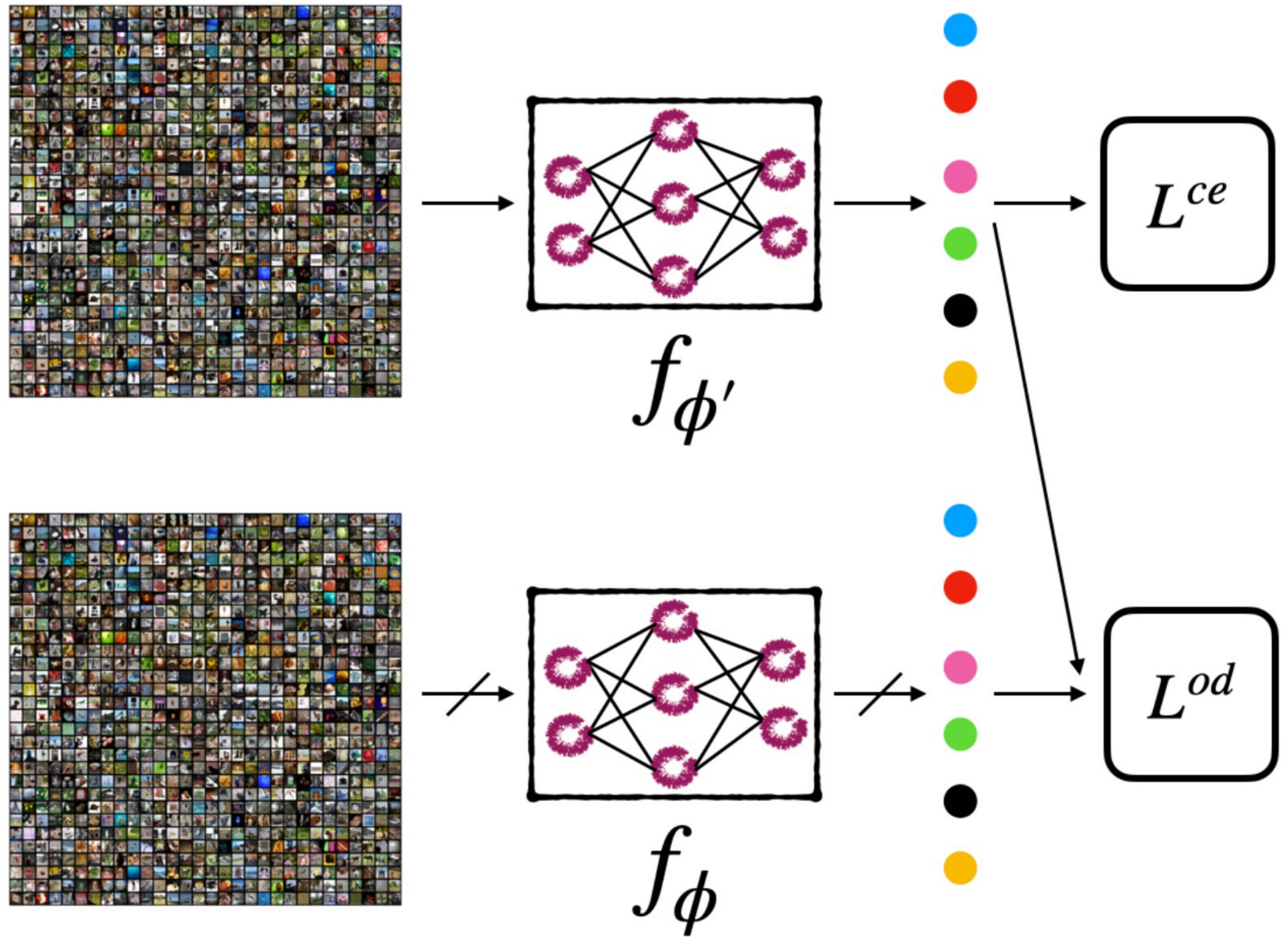


In meta testing, the embedding model serves to extract features for both support images and query images. Then, a linear classifier is trained with only a few samples to perform few-shot testing.



With this simple baseline, RFS[a] achieves state-of-the-art performance on multiple benchmarks, surpassing existing complicated meta learning algorithms.

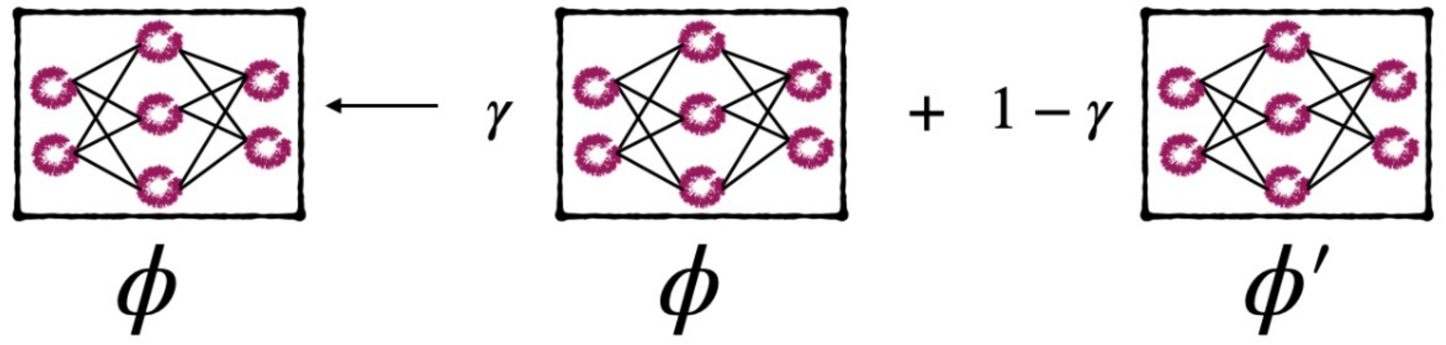
Method



Our method consists of two branches: a student network that learns to predict categorical labels; a teacher network which is a moving average of the teacher network. Our goal is to get the optimal parameters of the student network, given by

$$\Phi' = \underset{\phi'}{\operatorname{argmin}}\left(\alpha L^{ce}(D^{new};\phi') + \beta KL\left(f(D^{new};\phi'), f(D^{new};\phi)\right)\right).$$

Also, the update rule of the teacher network is shown as follows.



Finally, we use CutMix to further boost the performance. We create a new training example by mixing up two existing examples sampled from the dataset:

$$\tilde{x} = M \odot x_a + (1-M) \odot x_b$$
$$\tilde{y} = m y_a + (1-m) y_b$$

where (x_a, y_a) and (x_b, y_b) are image-label pairs.

Experiments:

model	backbone	miniImageNet 5-way 1-shot	miniImageNet 5-way 5-shot
MAML [6]	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92
Matching Networks [23]	64-64-64-64	43.56 ± 0.84	55.31 ± 0.73
IMP [1]	64-64-64-64	49.2 ± 0.7	64.7 ± 0.7
Prototypical Networks† [19]	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66
TAML [9]	64-64-64-64	51.77 ± 1.86	66.05 ± 0.85
SAML [8]	64-64-64-64	52.22 ± n/a	66.49 ± n/a
GCR [11]	64-64-64-64	53.21 ± 0.80	72.34 ± 0.64
KTN(Visual) [15]	64-64-64-64	54.61 ± 0.80	71.21 ± 0.66
PARN[24]	64-64-64-64	55.22 ± 0.84	71.55 ± 0.66
Dynamic Few-shot [7]	64-64-128-128	56.20 ± 0.86	73.00 ± 0.64
Relation Networks [21]	64-96-128-256	50.44 ± 0.82	65.32 ± 0.70
R2D2 [2]	96-192-384-512	51.2 ± 0.6	68.8 ± 0.1
SNAIL [12]	ResNet-12	55.71 ± 0.99	68.88 ± 0.92
AdaResNet [13]	ResNet-12	56.88 ± 0.62	71.94 ± 0.57
TADAM [14]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30
Shot-Free [17]	ResNet-12	59.04 ± n/a	77.64 ± n/a
TEWAM [16]	ResNet-12	60.07 ± n/a	75.90 ± n/a
MTL [20]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80
Variational FSL [26]	ResNet-12	61.23 ± 0.26	77.69 ± 0.17
MetaOptNet [10]	ResNet-12	62.64 ± 0.61	78.63 ± 0.46
Diversity w/ Cooperation [5]	ResNet-18	59.48 ± 0.65	75.62 ± 0.48
Fine-tuning [4]	WRN-28-10	57.73 ± 0.62	78.17 ± 0.49
LEO-trainval† [18]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12
RFS-simple	ResNet-12	62.02 ± 0.63	79.64 ± 0.44
RFS-distill	ResNet-12	64.82 ± 0.60	82.14 ± 0.43
Ours-online-distill (w/o CutMix)	ResNet-12	64.33 ± 0.25	82.13 ± 0.17
Ours-online-distill	ResNet-12	**67.07 ± 0.26**	**83.03 ± 0.18**
Ours-online-distill-trainval †	ResNet-12	**68.96 ± 0.26**	**84.22 ± 0.17**

model	backbone	CIFAR-FS 5-way 1-shot	CIFAR-FS 5-way 5-shot	FC100 5-way 1-shot	FC100 5-way 5-shot
MAML [6]	32-32-32-32	58.9 ± 1.9	71.5 ± 1.0	-	-
Prototypical Networks [19]	64-64-64-64	55.5 ± 0.7	72.0 ± 0.6	35.3 ± 0.6	48.6 ± 0.6
Relation Networks [21]	64-96-128-256	55.0 ± 1.0	69.3 ± 0.8	-	-
R2D2 [2]	96-192-384-512	65.3 ± 0.2	79.4 ± 0.1	-	-
TADAM [14]	ResNet-12	-	-	40.1 ± 0.4	56.1 ± 0.4
Shot-Free [17]	ResNet-12	69.2 ± n/a	84.7 ± n/a	-	-
TEWAM [16]	ResNet-12	70.4 ± n/a	81.3 ± n/a	-	-
Prototypical Networks [19]	ResNet-12	72.2 ± 0.7	83.5 ± 0.5	37.5 ± 0.6	52.5 ± 0.6
MetaOptNet [10]	ResNet-12	72.6 ± 0.7	84.3 ± 0.5	41.1 ± 0.6	55.5 ± 0.6
RFS-simple	ResNet-12	71.5 ± 0.8	86.0 ± 0.5	42.6 ± 0.7	59.1 ± 0.6
RFS-distill	ResNet-12	73.9 ± 0.8	86.9 ± 0.5	44.6 ± 0.7	60.9 ± 0.6
Ours-online-distill	ResNet-12	**76.18 ± 0.21**	**87.1 ± 0.2**	**45.43 ± 0.24**	**61.7 ± 0.3**

Datasets:
- miniImageNet
- CIFAR-FS
- FC100

Model:
- ResNet12

Results:
Our method with CutMix achieves stage-of-the-art performance on all settings. Without CutMix, our method outperforms RFS (w/o distillation, one stage) and is comparable to RFS (w/ distillation, two stage) while our method only uses one-stage training.

Conclusion:

- Our one-stage online self-distillation pipeline relies on distilling knowledge from a momentum-updated teacher to a student and suggests that multi-stage self-distillation is not imperative.
- We also identify that CutMix significantly improves the representations.
- We hope our method can shed new lights into the few-shot learning research.

Reference:

[a] Rethinking Few-Shot Image Classification: A Good Embedding Is All You Need?