Boosting Few-Shot Learning With Adaptive Margin Loss

Weiran Huang
Huawei Noah's Ark Lab

Joint work with Aoxue Li, Xu Lan, Jiashi Feng, Zhenguo Li and Liwei Wang

CVPR 2020

NOAH'S ARK LAB

Few-Shot Learning (FSL)

We are given

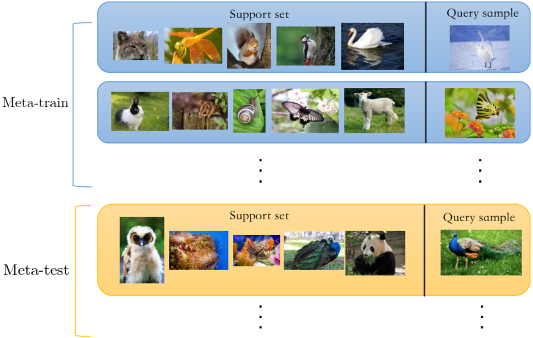
- a base class set C_{base} consisting of n_{base} base classes: each base class has sufficient labeled samples.
- a novel class set C_{novel} consisting of n_{novel} novel classes: each novel class has only a few labeled samples (e.g., less than 5 samples).

How to learn a good classifier for the novel classes by transferring the knowledge from base classes?

Meta-Learning Approach

Meta-learning is a common approach for the FSL. It involves two stages:[1]

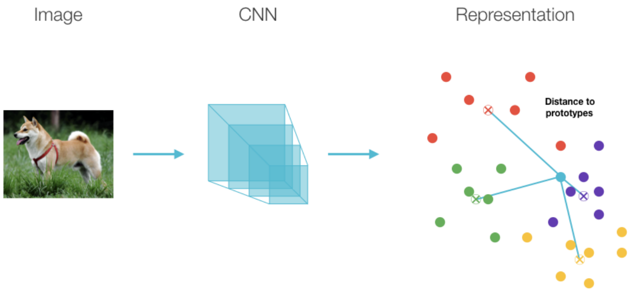
- Meta-training: In each episode, a meta task is constructed by sampling a small training set (support set) and a small test set (query set) from the whole base class dataset, which is then used to update the model.
- Meta-testing: The learned model is used to recognize samples from novel classes.



[1] Image credit: Yong Wang et al.

Metric-Based Meta-Learning

Metric-based meta-learning assumes that there exists an embedding space in which samples cluster around a single representation (called *prototype*) for each class, and these prototypes are then used as references to infer labels of test samples.[2]



[2] Image credit: Tiago Ramalho

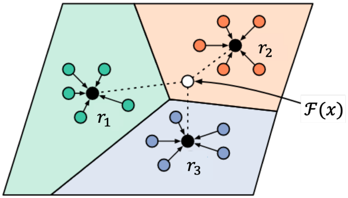
Training Loss of Metric-Based Meta-Learning

During a meta-training episode, all samples of the meta task are embedded into the embedding space by a feature extractor $\mathcal{F}$. Then, we generate prototypes $r_1, r_2, \cdots, r_{n_t}$ by using the samples from support set S. After that, we measure the similarity between every query image x and the prototype r_k, i.e., $\mathcal{D}(\mathcal{F}(x), r_k)$.

Finally, the classification loss can be formulated as:

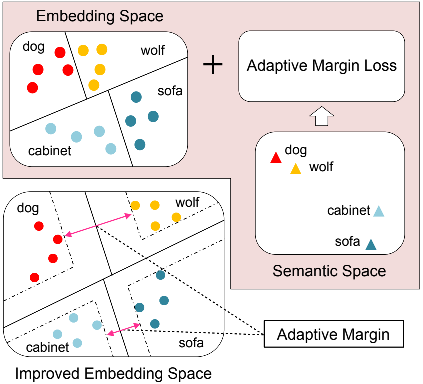
$$\mathcal{L}^{\text{cls}} = -\frac{1}{|Q|} \sum_{(x,y)\in Q} \log \frac{e^{\mathcal{D}(\mathcal{F}(x), r_y)}}{\sum_{k\in C_t} e^{\mathcal{D}(\mathcal{F}(x), r_k)}}, \quad (1)$$

where C_t denotes the class set of the current meta task.



Key Idea: Adding Margins in the Embedding Space

- To better separate samples from different classes (especially for similar classes), we introduce the adaptive margin in the embedding space.
- Key Idea: the margin between similar classes should be larger than the one between dissimilar classes.



Naive Additive Margin Loss

We first propose a naive additive margin loss (NAML), which can be formulated as:

$$\mathcal{L}^{\mathrm{na}} = -\frac{1}{|Q|}\sum_{(x,y)\in Q}\log\frac{e^{\mathcal{D}(\mathcal{F}(x),r_y)}}{e^{\mathcal{D}(\mathcal{F}(x),r_y)}+\sum_{k\in C_t\setminus\{y\}}e^{\mathcal{D}(\mathcal{F}(x),r_k)+m}}. \tag{2}$$

- The above naive additive margin loss assumes all classes should be equally far away from each other.
- It forces the embedding module $\mathcal{F}$ to extract more separable visual features for samples from different classes, which benefits the FSL.
- The fixed additive margin may lead to mistakes on test samples of similar classes, especially for the FSL where very limited number of labelled samples are provided in the novel classes.

Adaptive Margin Loss

To better separate similar classes in the feature embedding space, we design the margin adaptively.

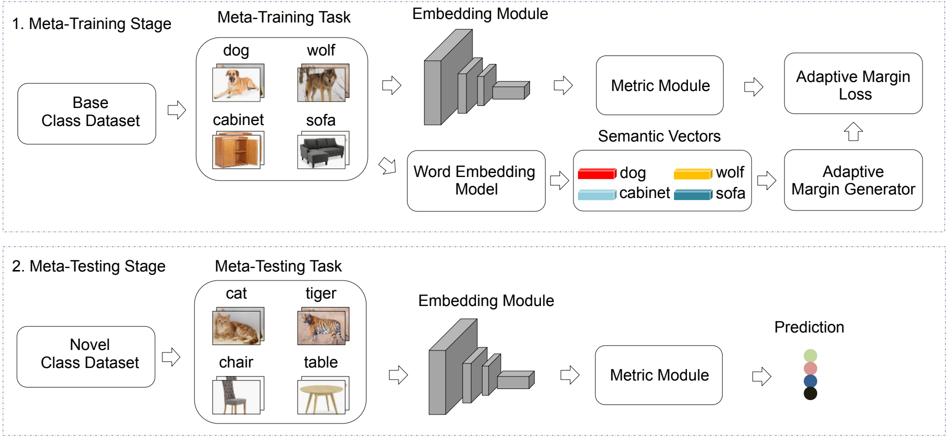
$$\mathcal{L} = -\frac{1}{|Q|} \sum_{(x,y)\in Q} \log \frac{e^{\mathcal{D}(\mathcal{F}(x), r_y)}}{e^{\mathcal{D}(\mathcal{F}(x), r_y)} + \sum_{k \in C_t \setminus \{y\}} e^{\mathcal{D}(\mathcal{F}(x), r_k)) + m_{y,k}}}. \tag{3}$$

where margin $m_{y,k}$ is generated according to the semantic similarity between y and k.

To measure the semantic similarity between two classes in a semantic space, we

- feed class names (e.g., dog) into a pre-trained word embedding model (e.g., Glove), and get the semantic word vectors.
- compute the similarity (e.g., cosine similarity) between word vectors.

The Overview of Our Proposed Approach



Class-Relevant Additive Margin

A simple way to generate the adaptive margin can be

$$m_{i,j}^{\mathrm{cr}} := \alpha \cdot \mathsf{sim}(e_i, e_j) + \beta, \tag{4}$$

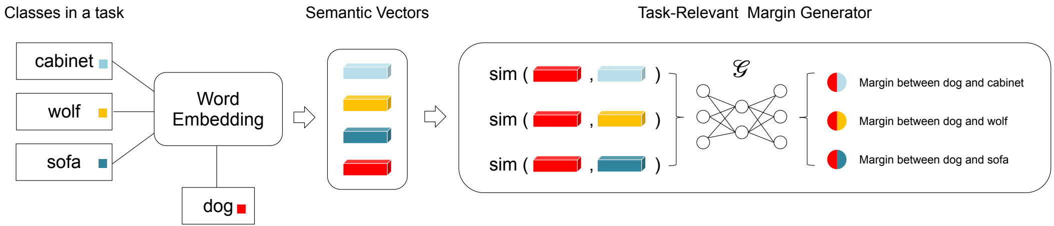
where $\mathsf{sim}(\cdot)$ denotes a metric to measure the semantic similarity between classes, and α and β are learnable parameters.

- In the experiment, we observe that the learned coefficient α is positive.
- Thus, our class-relevant margin loss can make the samples from similar classes to be more separable in the embedding space, which helps better recognize test class samples.

Task-Relevant Additive Margin

We also design the margin in a more careful way, which considers the semantic context among all classes in a meta-training task.

$$\{m_{y,k}^{\mathrm{tr}}\}_{k \in C_t \setminus \{y\}} = \mathcal{G}\left(\{\mathrm{sim}(e_y, e_k)\}_{k \in C_t \setminus \{y\}}\right) \tag{5}$$



Performance on the miniImageNet Dataset

Model	Backbone	Type	Test Accuracy 5-way 1-shot	Test Accuracy 5-way 5-shot
Matching Networks [17]	4Conv	Metric	43.56 ± 0.84	55.31 ± 0.73
Prototypical Network [14]	4Conv	Metric	49.42 ± 0.78	68.20 ± 0.66
Relation Networks [16]	4Conv	Metric	50.44 ± 0.82	65.32 ± 0.70
GCR [8]	4Conv	Metric	53.21 ± 0.40	72.34 ± 0.32
Memory Matching Network [2]	4Conv	Metric	53.37 ± 0.48	66.97 ± 0.35
Dynamic FSL [4]	4Conv	Metric	56.20 ± 0.86	73.00 ± 0.64
Prototypical Network [14]	ResNet12	Metric	56.52 ± 0.45	74.28 ± 0.20
TADAM [11]	ResNet12	Metric	58.50 ± 0.30	76.70 ± 0.38
DC [9]	ResNet12	Metric	62.53 ± 0.19	78.95 ± 0.13
TapNet [20]	ResNet12	Metric	61.65 ± 0.15	76.36 ± 0.10
ECMSFMT [13]	ResNet12	Metric	59.00	77.46
AM3 (Prototypical Network) [19]	ResNet12	Metric	65.21±0.49	75.20 ± 0.36
MAML [3]	4Conv	Gradient	48.70 ± 1.84	63.11 ± 0.92
MAML++ [1]	4Conv	Gradient	52.15 ± 0.26	68.32 ± 0.44
iMAML [12]	4Conv	Gradient	49.30 ± 1.88	-
LCC [10]	4Conv	Gradient	54.6 ± 0.4	71.1 ± 0.4
CAML [6]	ResNet12	Gradient	59.23 ± 0.99	72.35 ± 0.18
MTL [15]	ResNet12	Gradient	61.20 ± 1.80	75.50 ± 0.80
MetaOptNet-SVM [7]	ResNet12	Gradient	62.64 ± 0.61	78.63 ± 0.46
Prototypical Network + TRAML (OURS)	ResNet12	Metric	60.31 ± 0.48	77.94 ± 0.57
AM3 (Prototypical Network) + TRAML (OURS)	ResNet12	Metric	**67.10** ± 0.52	**79.54** ± 0.60

Ablation Study

Model (AM3 [19] as the backbone)	**Test Accuracy** 5-way 1-shot	5-way 5-shot
Original Classification Loss	65.21 $\pm$ 0.49	75.20 $\pm$ 0.36
Naive Additive Margin Loss	65.42 $\pm$ 0.25	75.48 $\pm$ 0.34
Class-Relevant Additive Margin Loss	66.36 $\pm$ 0.57	77.21 $\pm$ 0.48
Task-Relevant Additive Margin Loss	**67.10** $\pm$ 0.52	**79.54** $\pm$ 0.60

- Simply adding a fixed margin has limited effectiveness in FSL.
- Class-relevant additive margin is shown to benefit the embedding learning for FSL.
- By considering the semantic context among classes in a meta-training task, task-relevant additive margin yields the best results.

Generalized Few-Shot Learning

We also test our approach in a more challenging yet practical generalized FSL setting, where the label space of test data is extended to both base and novel classes.

Model	Novel					All				
	n_s=1	2	5	10	20	n_s=1	2	5	10	20
Logistic regression (from [18])	38.4	51.1	64.8	71.6	76.6	40.8	49.9	64.2	71.9	76.9
Logistic regression w/H (from [5])	40.7	50.8	62.0	69.3	76.5	52.2	59.4	67.6	72.8	76.9
Prototypical Network [14] (from [18])	39.3	54.4	66.3	71.2	73.9	49.5	61.0	69.7	72.9	74.6
Matching Networks [17] (from [18])	43.6	54.0	66.0	72.5	76.9	54.4	61.0	69.0	73.7	76.5
Squared Gradient Magnitude w/H [5]	-	-	-	-	-	54.3	62.1	71.3	75.8	78.1
Batch Squared Gradient Magnitude [5]	-	-	-	-	-	49.3	60.5	71.4	75.8	78.5
Prototype Matching Nets [18]	43.3	55.7	68.4	74.0	77.0	55.8	63.1	71.1	75.0	77.1
Prototype Matching Nets w/H [18]	45.8	57.8	69.0	74.3	77.4	57.6	64.7	71.9	75.2	77.5
Dynamic FSL [4]	46.0	57.5	69.2	74.8	78.1	58.2	65.2	72.2	76.5	78.7
Dynamic FSL + TRAML (OURS)	**48.1**	**59.2**	**70.3**	**76.4**	**79.4**	**59.2**	**66.2**	**73.6**	**77.3**	**80.2**

Table: Comparative results for generalized FSL on the ImageNet2012 dataset. The top-5 accuracies (%) on the novel classes and on all classes are used as the evaluation metrics for this dataset.

Conclusion

- Our method introduces adaptive margin in the embedding space, which can effectively enhance the discriminative power of embedding space.
- We develop a class-relevant additive margin loss, where semantic similarity between each pair of classes is considered to separate samples in the feature embedding space from similar classes.
- Further, we incorporate the semantic context among all classes in a sampled training task and develop a task-relevant additive margin loss to better distinguish samples from different classes.
- Our method can be applied to most scenarios for clustering in the feature embedding space, e.g., standard FSL, generalized FSL, etc. Extensive experiments demonstrate that our method can boost the performance of current metric-based meta-learning approaches.

Thank you!

We are looking for research interns (Contact me for details).

References I

[1] Antreas Antoniou et al. “How to train your MAML.”. 2018.

[2] Qi Cai et al. “Memory Matching Networks for One-Shot Image Recognition.”. 2018.

[3] Chelsea Finn et al. “Model-agnostic meta-learning for fast adaptation of deep networks”. 2017.

[4] Spyros Gidaris and Nikos Komodakis. “Dynamic Few-Shot Visual Learning Without Forgetting”. 2018.

[5] Bharath Hariharan and Ross B. Girshick. “Low-shot Visual Recognition by Shrinking and Hallucinating Features”. 2017.

[6] Xiang Jiang et al. “Learning to Learn with Conditional Class Dependencies.”. 2019.

[7] Kwonjoon Lee et al. “Meta-learning with differentiable convex optimization.”. 2019.

[8] Aoxue Li et al. “Few-Shot Learning with Global Class Representations”. 2019.

[9] Yann Lifchitz et al. “Dense Classification and Implanting for Few-Shot Learning.”. 2019.

[10] Yaoyao Liu et al. “LCC: Learning to Customize and Combine Neural Networks for Few-Shot Learning.”. 2019.

[11] Boris N. Oreshkin et al. “TADAM: Task dependent adaptive metric for improved few-shot learning”. 2018.

References II

[12] Aravind Rajeswaran et al. “Meta-learning with implicit gradients.”. 2019.

[13] Avinash Ravichandran et al. “Few-Shot Learning with Embedded Class Models and Shot-Free Meta Training.”. 2019.

[14] Jake Snell et al. “Prototypical Networks for Few-shot Learning”. 2017.

[15] Qianru Sun et al. “Meta-Transfer Learning for Few-Shot Learning.”. 2019.

[16] Flood Sung et al. “Learning to compare: Relation network for few-shot learning”. 2018.

[17] Oriol Vinyals et al. “Matching networks for one shot learning”. 2016.

[18] Yu-Xiong Wang et al. “Low-Shot Learning from Imaginary Data”. 2018.

[19] Chen Xing et al. “Adaptive Cross-Modal Few-Shot Learning.”. 2019.

[20] Sung Whan Yoon et al. “TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning.”. 2019.