Abstract

Few-shot learning has been a long-standing problem in learning to learn. This problem typically involves training a model on a extremely small amount of data and testing the model on the out-of-distribution data. The focus of recent few-shot learning research has been on the development of good representation models that can quickly adapt to test tasks. To that end, we come up with a model that learns representation through online self-distillation. Our model combines supervised training with knowledge distillation via a continuously updated teacher. We also identify that data augmentation plays an important role in producing robust features. Our final model is trained with CutMix augmentation and online self-distillation. On the commonly used benchmark miniImageNet, our model achieves 67.07% and 83.03% under the 5-way 1-shot setting and the 5-way 5-shot setting, respectively. It outperforms counterparts of its kind by 2.25% and 0.89%.

1. Introduction

Few-shot learning is a crucial problem in learning to learn. In contrast to the common deep learning settings where a large amount of training data is available, few-shot learning often deals with scenarios where the training data is scarce. So this problem boils down to how to design models that can quickly adapt to test tasks. Recently, RFS [22] proposes a simple supervised-training baseline that outperforms meta-learning algorithms. It learns a representation model on the joint set of training tasks and improves the representations through self-distillation. The success of this method indicates that a good embedding is more important than sophisticated meta-learning algorithms.

However, RFS relies on a two-stage training pipeline consisting of supervised training and self-distillation, which reduces its practicability. To that end, we come up with a one-stage method that incorporates supervised training and knowledge distillation into a unified pipeline. The teacher network in our model is an exponential moving average of the student network and is continuously updated through the training process. The student network is trained with a combination of cross-entropy loss and self-distillation loss. Our model is significantly simpler than RFS [22] and other variants. In addition, we identify that CutMix [25] can greatly improve the representation model. Without bells and whistles, our model achieves 67.07% under the 5-way 1-shot setting and and 83.03% under the 5-way 5-shot setting on the miniImageNet [3] dataset.

2. Preliminary

We establish preliminaries of few-shot learning by learning representation [22] in this section. First, we formulate the problem in §2.1. Then, we present the details of RFS [22] in §2.2. For ease of comparison to previous work, we use the same notation as [22].

2.1. Few-shot Learning formulation

In few-shot learning, the data consists of a meta-training set \( T = \{ (D_{j}^{train}, D_{j}^{test}) \}_{j=1}^{J} \) and a meta-testing set \( S = \{ (D_{j}^{train}, D_{j}^{test}) \}_{j=1}^{J} \). The meta-training set and the meta-testing set do not share the same categories. Each task \( D_{j}^{train} \) contains a small number of example. \( D_{j}^{train} = \{ (x_{i}, y_{i}) \}_{i=1}^{Q_{j}} \) and \( D_{j}^{test} = \{ (x_{q}, y_{q}) \}_{q=1}^{Q} \) are sampled from the same distribution. A base learner \( \mathcal{A} \), given by \( y_{*} = f_{0}(x_{*}) \), is trained on \( D_{j}^{train} \) and evaluated on \( D_{j}^{test} \). To reduce the dimensionality of \( x_{*} \), training examples and testing examples are mapped into a feature space by an embedding model \( \Phi_{*} = f_{0}(x_{*}) \). The objective of the few-shot learning algorithms is to learn a good embedding model, so that the average test error of the base learner on a distribution of tasks is minimized. This is given by,

\[
\phi = \arg \min_{\phi} E_{T} [L_{meta}(D_{j}^{test}; \theta, \phi)],
\]

where \( \theta = \mathcal{A}(D_{j}^{train}; \phi) \). Finally, the model is evaluated over the distribution of the test tasks:

\[
E_{S} [L_{meta}(D_{j}^{test}; \theta, \phi), \text{where} \theta = \mathcal{A}(D_{j}^{train}; \phi)].
\]
2.2. Few-shot Learning by learning the representation

RFS [22] proposes a simple method to learn the representation of the embedding model. Tasks from the meta-training set are merged into a single classification task, which is given by

\[
\mathcal{D}_{\text{new}}^{\text{new}} = \{(x_i, y_i)\}_{i=1}^K = \mathcal{D}_{\text{train}}^{\text{train}} \cup \mathcal{D}_{\text{train}}^{\text{train}} \cup \mathcal{D}_{\text{train}}^{\text{train}},
\]

where \(\mathcal{D}_{\text{train}}^{\text{train}}\) is the task from \(\mathcal{T}\). The embedding model is

\[
\phi = \arg \min_{\phi} L^{ce}(\mathcal{D}_{\text{new}}^{\text{new}}; \phi),
\]

and trained by optimizing a cross-entropy loss \(L^{ce}\).

In addition to this ordinary supervised training, RFS also introduces a self-distillation stage to further improve the representation. After obtaining the embedding model \(\phi\), a new embedding model parameterized by \(\phi'\) is trained to minimize a weighted sum of the cross-entropy loss between the predictions and ground-truth labels and the Kullback–Leibler divergence (KL) between these two steps:

\[
\phi' = \arg \min_{\phi'} (\alpha L^{ce}(\mathcal{D}_{\text{new}}^{\text{new}}; \phi') + \beta KL(f(\mathcal{D}_{\text{new}}^{\text{new}}; \phi'), f(\mathcal{D}_{\text{new}}^{\text{new}}; \phi))).
\]

Conceptually, this step is a variant of self-distillation where the teacher network and the student network have the same model architectures.

Once the training is finished, the model is evaluated on the meta-testing set. The base learner is trained on the task \(\mathcal{D}_{\text{test}}^{\text{train}}\) sampled from meta-testing distribution, given by

\[
\theta = \arg \min_{\{W, b\}} \sum_{t=1}^T L_t^{ce}(W f_{\phi'}(x_t) + b, y_t) + R(W, b),
\]

where the base learner is parameterized by \(\theta = \{W, b\}\) and the embedding model \(f_{\phi'}\) is fixed.

3. Method

We will detail our method in this section. We introduce the online self-distillation in §3.1. Then, we discuss one special data augmentation technique – CutMix [25] in §3.2.

3.1. Few-Shot Learning with Online Self-Distillation

We propose a training pipeline that combines supervised training with self-distillation, in contrast to existing methods that consist of separate stages. We use \(\phi\) and \(\phi'\) to denote the teacher network and the student network, respectively. Instead of learning \(\phi\) in a pre-training stage, our method updates \(\phi\) on-the-fly as well as distilling the knowledge from \(\phi\) to \(\phi'\) (Figure 1). Mathematically, we alternate between these two steps:

\[
\phi' = \arg \min_{\phi'} (\alpha L^{ce}(\mathcal{D}_{\text{new}}^{\text{new}}; \phi') + \beta KL(f(\mathcal{D}_{\text{new}}^{\text{new}}; \phi'), f(\mathcal{D}_{\text{new}}^{\text{new}}; \phi))),
\]

where \(\gamma = 0.99\) controls the velocity of the parameter update. Different from common machine learning models, \(\phi\) is not updated through gradient descent but direct parameter update.

3.2. CutMix

We present a special data augmentation—CutMix [25]—that improves few-shot learning performance. The goal of CutMix is to generate a new training example \((\bar{x}, \bar{y})\) by combining examples \((x_a, y_a)\) and \((x_b, y_b)\), given by

\[
\bar{x} = \mathcal{M} \odot x_a + (1 - \mathcal{M}) \odot x_b \quad \text{and} \quad \bar{y} = my_a + (1 - m)y_b,
\]

where \(\mathcal{M}\) is a binary mask randomly sampled from a uniform distribution in \(\mathcal{U}(0, 1)\).

Figure 1: Overview of online self-distillation. Back-propagation and SGD are not performed in the \(f_b\) branch.

Figure 2: Update rules of the teacher network \(f_{\phi}\).
We conduct experiments on the widely used benchmarks miniImageNet, CIFAR-FS, and FC100. miniImageNet is a subset of ImageNet; it contains 64, 16, 20 categories for training, validation, and testing, respectively. The CIFAR-FS and FC100 are both derivatives of the CIFAR-100 dataset. CIFAR-FS has 64, 16, 20 categories for training, validation, and testing while FC100 has 60, 20, 20 categories for training, validation, and testing.

Table 1: Comparison to prior work on miniImageNet. Results reported with input image size of 84x84. ⋆ results obtained by training on the union of training and validation sets.

Table 2: Comparison to prior work on CIFAR-FS and FC100.

Table 3: Comparison to prior work on CIFAR-FS and FC100.

5. Conclusion

We propose a one-stage online self-distillation pipeline for few-shot learning. Our method relies on distilling knowledge from a momentum-updated teacher to a student. Our method suggests that multi-stage self-distillation is not imperative. We also identify that CutMix significantly improves the representation. With these combined techniques, our method achieves new state-of-the-art on the commonly used datasets. We hope our method will shed new lights into the few-shot learning research.

References


[25] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In International Conference on Computer Vision (ICCV), 2019. 1, 2