Meta-Learning for Semi-Supervised Few-Shot Classification

Mengye Ren **Eleni Triantafillou*** **Sachin Ravi*** **Jake Snell** **Kevin Swersky**

Joshua B. Tenenbaum **Hugo Larochelle** **Richard S. Zemel**

Abstract

In this work, we advance the few-shot classification paradigm towards a scenario where unlabeled examples are also available within each episode. We consider two situations: one where all unlabeled examples are assumed to belong to the same set of labeled classes of the episode, as well as the more challenging situation where examples from other distractor classes are also provided. To address this paradigm, we propose novel extensions of Prototypical Networks that are augmented with the ability to use unlabeled examples when producing prototypes. These models are trained in an end-to-end way on episodes, to learn to leverage the unlabeled examples successfully. We also propose a new split of ImageNet, consisting of a large set of classes, with a hierarchical structure. Our experiments confirm that our Prototypical Networks can learn to improve their predictions due to unlabeled examples, much like a semi-supervised algorithm would.

1 Introduction

The availability of large quantities of labeled data has enabled deep learning methods to achieve impressive breakthroughs in several tasks related to artificial intelligence, such as speech recognition, object recognition and machine translation. However, current deep learning approaches struggle in tackling problems for which labeled data are scarce.

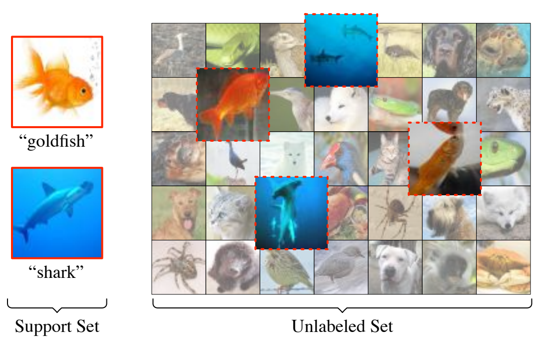

Figure 1: Consider a setup where the aim is to learn a classifier to distinguish between two previously unseen classes, goldfish and shark, given not only labeled examples of these two classes, but also a larger pool of unlabeled examples, some of which may belong to one of these two classes of interest.

For this reason, recently there has been an increasing body of work on few-shot learning, which considers the design of learning algorithms that specifically allow for better generalization on problems with small labeled training sets. Here we focus on the case of few-shot classification, where the given classification problem is assumed to contain only a handful of labeled examples per class. One approach to few-shot learning follows a form of meta-learning[2] [1, 2], which performs transfer learning from a pool of various classification problems generated from large quantities of available labeled data, to new classification problems from classes unseen at training time. Meta-learning may take the form

*These authors contributed equally.

[2]See the following blog post for an overview: `http://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/`

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

of learning a shared metric [3, 4], a common initialization for few-shot classifiers [5, 6] or a generic inference network [7, 8].

However, this progress has been in a limited scenario, which differs in many dimensions from how humans learn new concepts. In this paper we aim to generalize the few-shot setting in two ways. First we consider a scenario in which the new classes are learned in the presence of additional unlabeled data. Second, we consider the situation where the new classes to be learned are not viewed in isolation. Instead, many of the unlabeled examples are from different classes; the presence of such *distractor* classes introduces an additional and more realistic level of difficulty to the few-shot problem. We propose and study three novel extensions of Prototypical Networks [4], a state-of-the-art approach to few-shot learning, to the semi-supervised setting. We demonstrate in our experiments that our semi-supervised variants successfully learn to leverage unlabeled examples and outperform purely supervised Prototypical Networks.

2 Semi-Supervised Few-Shot Learning

We denote our training set as a tuple of labeled and unlabeled examples: $(\mathcal{S}, \mathcal{R})$. The labeled portion is the usual support set $\mathcal{S}$ of the few-shot learning literature, containing a list of tuples of inputs and targets. In addition to classic few-shot learning, we introduce an unlabeled set $\mathcal{R}$ containing only inputs: $\mathcal{R} = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_M\}$. As in the purely supervised setting, our models are trained to perform well when predicting the labels for the examples in the episode's query set $\mathcal{Q}$. Figure 2 shows a visualization of training and test episodes.

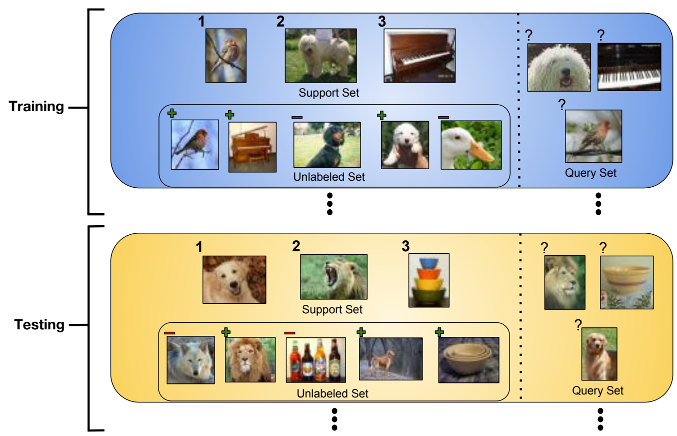

Figure 2: Example of the semi-supervised few-shot learning setup. Training episodes consist of a support set $\mathcal{S}$, an unlabeled set $\mathcal{R}$, and a query set $\mathcal{Q}$. The items in $\mathcal{R}$ may either be pertinent to the labeled classes (with + signs) or they may be *distractor* items (with - signs).

2.1 Prototypical Networks with Soft k-Means

We first consider a simple way of leveraging unlabeled examples for refining prototypes, by taking inspiration from semi-supervised clustering. One natural choice would be to borrow from the inference performed by soft k-means. We prefer this version of k-means over hard assignments since hard assignments would make the inference non-differentiable. We start from the regular Prototypical Network's prototypes $\boldsymbol{p}_c$. Then, the unlabeled examples get a partial assignment $(\tilde{z}_{j,c})$ to each cluster based on their Euclidean distance to the cluster locations. Finally, refined prototypes are obtained by incorporating these unlabeled examples.

This process can be summarized as follows:

$$\tilde{\boldsymbol{p}}_c = \frac{\sum_i h(\boldsymbol{x}_i) z_{i,c} + \sum_j h(\tilde{\boldsymbol{x}}_j) \tilde{z}_{j,c}}{\sum_i z_{i,c} + \sum_j \tilde{z}_{j,c}}, \quad \text{where} \quad \tilde{z}_{j,c} = \frac{\exp\left(-||h(\tilde{\boldsymbol{x}}_j) - \boldsymbol{p}_c||_2^2\right)}{\sum_{c'} \exp\left(-||h(\tilde{\boldsymbol{x}}_j) - \boldsymbol{p}_{c'}||_2^2\right)} \tag{1}$$

We could perform several iterations of refinement, as is usual in k-means. However, we have experimented with various number of iterations and found results to not improve beyond a single refinement step.

2.2 Prototypical Networks with Soft k-Means with a Distractor Cluster

The soft k-means approach described above implicitly assumes that each unlabeled example belongs to either one of the N classes in the episode. However, it would be much more general to have a model robust to the existence of examples from other classes, which we refer to as distractor classes. A simple way to address this is to add an additional cluster whose purpose is to capture the distractors.

$$\boldsymbol{p}_c = \begin{cases} \frac{\sum_i h(\boldsymbol{x}_i) z_{i,c}}{\sum_i z_{i,c}} & \text{for } c = 1...N \\ \boldsymbol{0} & \text{for } c = N+1 \end{cases} \tag{2}$$

Here we take the simplifying assumption that the distractor cluster has a prototype centered at the origin. We also consider introducing length-scales r_c to represent variations in the within-cluster distances, specifically for the distractor cluster:

$$\tilde{z}_{j,c} = \frac{\exp\left(-\frac{1}{r_c^2}||\tilde{\boldsymbol{x}}_j - \boldsymbol{p}_c||_2^2 - A(r_c)\right)}{\sum_{c'} \exp\left(-\frac{1}{r_c^2}||\tilde{\boldsymbol{x}}_j - \boldsymbol{p}_{c'}||_2^2 - A(r_{c'})\right)}, \text{ where } A(r) = \frac{1}{2}\log(2\pi) + \log(r) \tag{3}$$

For simplicity, we set $r_{1...N}$ to 1 in our experiments, and only learn the length-scale of the distractor cluster r_{N+1}.

2.3 Prototypical Networks with Soft k-Means and Masking

Modeling distractor unlabeled examples with a single cluster is likely too simplistic , since distractor examples may very well cover more than a single natural object category. To address this problem, we incorporate a soft-masking mechanism on the contribution of unlabeled examples. We start by computing normalized distances $\tilde{d}_{j,c}$ between examples $\tilde{\boldsymbol{x}}_j$ and prototypes $\boldsymbol{p}_c$:

$$\tilde{d}_{j,c} = \frac{d_{j,c}}{\frac{1}{M}\sum_j d_{j,c}}, \text{ where } d_{j,c} = ||h(\tilde{\boldsymbol{x}}_j) - \boldsymbol{p}_c||_2^2 \tag{4}$$

Then, soft thresholds β_c and slopes γ_c are predicted for each prototype, by feeding to a small neural network various statistics of the normalized distances for the prototype:

$$[\beta_c, \gamma_c] = \text{MLP}\left([\min_j(\tilde{d}_{j,c}), \max_j(\tilde{d}_{j,c}), \operatorname{var}_j(\tilde{d}_{j,c}), \operatorname{skew}_j(\tilde{d}_{j,c}), \operatorname{kurt}_j(\tilde{d}_{j,c})]\right) \tag{5}$$

This allows each threshold to use information on the amount of intra-cluster variation to determine how aggressively it should cut out unlabeled examples.

Then, soft masks $m_{j,c}$ for the contribution of each example to each prototype are computed, by comparing to the threshold the normalized distances, as follows:

$$\tilde{\boldsymbol{p}}_c = \frac{\sum_i h(\boldsymbol{x}_i) z_{i,c} + \sum_j h(\tilde{\boldsymbol{x}}_j)\tilde{z}_{j,c} m_{j,c}}{\sum_i z_{i,c} + \sum_j \tilde{z}_{j,c} m_{j,c}}, \text{ where } m_{j,c} = \sigma\left(-\gamma_c\left(\tilde{d}_{j,c} - \beta_c\right)\right) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function.

3 Experiments

3.1 Datasets

We evaluate the performance of our models on three datasets: two benchmark few-shot classification datasets and a novel large-scale dataset that we hope will be useful for future few-shot learning work.

Omniglot [9] is a dataset of 1,623 handwritten characters from 50 alphabets. Each character was drawn by 20 human subjects. We follow the few-shot setting proposed by [3], in which the images are resized to 28×28 pixels and rotations in multiples of 90° are applied, yielding 6,492 classes in total. These are split into 4,800 training classes and 1,692 classes for test. 10% of the training images are used as labeled examples.

***mini*ImageNet** [3] is a modified version of the ILSVRC-12 dataset [10], in which 600 images for each of 100 classes were randomly chosen to be part of the dataset. We rely on the class split used by [5]. These splits use 64 classes as training, 16 for validation, and 20 for test. All images are of size 84×84 pixels. 40% of the training images are used as labeled examples.

***tiered*ImageNet** is our proposed dataset for few-shot classification. Like *mini*Imagenet, it is a subset of ILSVRC-12. However, *tiered*ImageNet represents a larger subset of ILSVRC-12 (608

ProtoNet Model	Err.	Err. w/ D
Supervised	5.16%	5.16%
Semi-Supervised Inference	2.35%	4.70%
Soft k-Means	2.56%	4.59%
Soft k-Means+Cluster	**2.18%**	2.71%
Masked Soft k-Means	2.46%	**2.62%**

Table 1: Omniglot 1-shot Results

ProtoNet Model	*mini / tiered* 1-shot Acc.	*mini / tiered* 5-shot Acc.	*mini / tiered* 1-shot Acc. w/ D	*mini / tiered* 5-shot Acc. w/ D
Supervised	43.36% / 46.60%	59.03% / 67.18%	43.36% / 46.60%	59.03% / 67.18%
Semi-Supervised Inference	48.68% / 50.38%	62.94% / 70.26%	46.16% / 46.87%	62.32% / 68.38%
Soft k-Means	48.25% / 53.41%	**65.72% / 71.31%**	46.72% / 50.18%	61.94% / 68.83%
Soft k-Means+Cluster	**50.87% / 55.82%**	63.75% / 70.79%	48.60% / 49.87%	61.51% / 70.16%
Masked Soft k-Means	50.57% / 52.76%	63.78% / 70.08%	**50.04% / 50.93%**	**62.50% / 71.00%**

Table 2: *mini*ImageNet and *tiered*ImageNet 1/5-shot Results

classes rather than 100 for *mini*ImageNet). Analogous to Omniglot, in which characters are grouped into alphabets, *tiered*ImageNet groups classes into broader categories corresponding to higher-level nodes in the ImageNet [11] hierarchy. There are 34 categories in total, with each category containing between 10 and 30 classes. These are split into 26 training categories and 8 testing categories. (details of the dataset can be found in the Supplementary Materials). This ensures that all of the training classes are sufficiently distinct from the testing classes, unlike *mini*ImageNet and other alternatives such as *rand*ImageNet proposed by [3]. For example, “pipe organ” is a training class and “electric guitar” is a test class in the [5] split of *mini*Imagenet, even though they are both musical instruments. 10% of the training images are used as labeled examples.

In each dataset we compare our three semi-supervised models with two baselines. The first baseline, referred to as Supervised in our tables, is an ordinary Prototypical Network that is trained in a purely supervised way on the labeled split of each dataset. The second baseline, referred to as Semi-Supervised Inference, uses the embedding function learned by this supervised Prototypical Network, but performs semi-supervised refinement of the prototypes at inference time using a step of Soft k-Means refinement.

3.2 Results

Results for Omniglot are given in Table 1, and *mini*ImageNet and *tiered*ImageNet in Table 2. Across all three benchmarks, at least one of our proposed models outperform the baselines, demonstrating the effectiveness of our semi-supervised meta-learning procedure. In particular, Soft k-means+Cluster performs the best on 1-shot non-distractor settings, as the extra cluster seems to provide a form of regularization that pushes the clusters farther apart. Soft k-Means performs well on 5-shot non-distractor settings, as it considers the most unlabeled examples. Masked Soft k-Means shows the most robust performance in distractors settings, in both 1-shot and 5-shot tasks. For 5-shot, Masked soft k-Means reaches comparable performance compared to the upper bound of the best non-distractor performance.

4 Conclusion

In this work, we propose a novel semi-supervised few-shot learning paradigm, where an unlabeled set is added to each episode. We also extend the setup to more realistic situations where the unlabeled set has classes not belonging to the labeled classes. We also introduce a larger dataset split, *tiered*ImageNet, with hierarchical levels of labels. We propose several novel extensions of Prototypical Networks, and they show consistent improvements under semi-supervised settings compared to our baselines. As future work, we are working on incorporating fast weights [12, 6] into our framework so that examples can have different embedding representation given the contents in the episode.

References

[1] Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.

[2] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001.

[3] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, pages 3630–3638, 2016.

[4] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, 2017.

[5] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations*, 2017.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *34th International Conference on Machine Learning*, 2017.

[7] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. One-shot learning with memory-augmented neural networks. In *33rd International Conference on Machine Learning*, 2016.

[8] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. Meta-learning with temporal convolutions. *CoRR*, abs/1707.03141, 2017.

[9] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the 33th Annual Meeting of the Cognitive Science Society, CogSci 2011, Boston, Massachusetts, USA, July 20-23, 2011*, 2011.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[12] Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4331–4339, 2016.