Variadic Meta-Learning by Bayesian Nonparametric Deep Embedding

Kelsey R. Allen MIT
Hanul Shin* MIT
Evan Shelhamer* UC Berkeley
Joshua B. Tenenbaum MIT

krallen@mit.edu,skyshin@mit.edu,shelhamer@cs.berkeley.edu,jbt@mit.edu

Abstract

Learning from a little or a lot of data is addressed by two strong but divided frontiers: few-shot learning and standard supervised learning. Few-shot learning focuses on sample efficiency at small scale, while supervised learning focuses on accuracy at large scale. Ideally they could be reconciled to learn with any number of data points (shot) and number of classes (way). To span the full spectrum of shot and way, we frame the *variadic learning* regime of learning from any number of inputs. We approach variadic learning by meta-learning a novel multi-modal clustering model that connects bayesian nonparametrics and deep metric learning. Our bayesian nonparametric deep embedding (BANDE) method is optimized end-to-end with a single objective, and adaptively adjusts capacity to learn from variable amounts of data. BANDE achieves (a) state-of-the-art accuracy for alphabet recognition, (b) 71.9% accuracy on 1692-way, 5-shot Omniglot classification from only 5-way 1-shot meta-learning, (c) 94.4% accuracy on CIFAR-10 (comparable to supervised learning techniques), and (d) equal or better than state-of-the-art accuracies for semi-supervised classification of Omniglot and mini-ImageNet.

1 Introduction

In machine learning, classification problems span two important axes: the number of classes to recognize (the "way" of the problem) and the number of examples provided for each class (the "shots" to learn from). At one extreme, there are large-scale tasks like ImageNet in which there are 1000 classes each with roughly 1000 examples (a 1000-way, $\sim$1000-shot problem). At the other extreme, there are datasets for learning from few examples, such as Omniglot, which features a 5- or 20-way, 1-shot problem. State-of-the-art methods for these two points in the problem orthant are substantially different, with the former dominated by standard fully-supervised deep networks and the latter by episodic meta-learning techniques. We propose a single learner for different shots and ways, including the one-shot and many-shot extremes, that generalizes better than existing methods. Our method is a multi-modal, semi-supervised clustering algorithm on deep embeddings (Figure 1).

We call this regime of variable shot and way the *variadic learning* regime, after variadic functions. Just as variadic functions are those which can take any number of arguments to produce a result, a good variadic learner must learn from any amount of data, whatever the number of examples and classes, and produce strong results across unknown data distributions during test.

Ideally, meta-learning approaches need not depend on a specific testing shot and way. However, in practice, meta-learning has commonly been trained and evaluated in constrained circumstances. Meta-learning is usually carried out independently across settings so that a different learner is specialized to each n-way, k-shot task. This potentially limits deployment to more diverse settings with variable shot and way that we address in this work.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

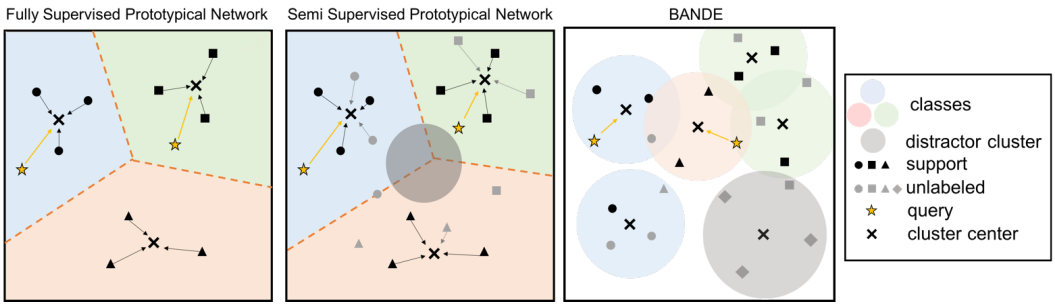


Figure 1: Our bayesian bonparametric deep embedding (BANDE) method is optimized end-to-end to cluster labeled and unlabeled data into multi-modal prototypes. BANDE represents each class by a set of clusters, unlike prior prototypical methods that represent each class by a single cluster.

2 Bayesian Nonparametric Deep Embedding (BANDE)

Our method defines *multi-modal* prototypes of labeled and unlabeled data alike through Bayesian nonparametric clustering of end-to-end optimized deep embeddings. This extends prototypical networks [10] and semi-supervised prototypical networks [9] to multi-modal clustering while simplifying the use of unlabeled data. By deciding the number of modes, our method interpolates between local exemplar and global prototype representations, in effect adjusting its capacity depending on the data.

For multi-modal clustering we incorporate DP-means [5] into our approach. DP-means is a scalable, Bayesian nonparametric algorithm for unsupervised clustering that computes the minimum distance of each example to all existing cluster means. If this distance exceeds a threshold λ, a new cluster is created by setting its mean equal to the embedding of the example $h_\phi(x_i)$. Our extension handles labeled and unlabeled data, augments the clustering with soft assignments under a normalized Gaussian likelihood, and defines a procedure for choosing λ during learning and inference.

The loss is the normalized Gaussian likelihood of the query embeddings under the support clusters. Since there are potentially multiple clusters with the same label, we weight the loss, where the closest cluster mean for the correct class has weight 1, and all other cluster means of that class have weight 0. This encourages BANDE to learn multi-modal embeddings if need be.

Algorithm 1 BANDE

Initialize $\{\mu_0, \mu_1, ..., \mu_n\}$
Initialize $C = n$
for each example i **do**
  for each cluster c **do**
    $d_{i,c} \leftarrow \|h_\phi(x_i) - \mu_c\|^2$
  end for
  if $\min_c(d_{i,c}) > \lambda$ **then**
    $C = C + 1$
    $\mu_C \leftarrow h_\phi(x_i)$
  end if
end for
update soft assignments $z_{i,c}$
update cluster means μ_c

BANDE is initialized with n clusters $\{\mu_0, \mu_1, ..., \mu_n\}$ for the n labeled classes in the support, with each set to the class-wise mean as in standard prototypical networks. New labeled clusters are created with a radius σ, while unlabeled clusters have radius σ_u to capture increased uncertainty about the unlabeled distribution (which can and does contain multiple unknown classes).

Unlike σ and σ_u, the distance threshold λ is non-differentiable and cannot be learned jointly. Instead, we set λ episodically based on its derivation in [5], which defines it in terms of α, the relative probability of a new cluster in the Chinese Restaurant Process prior, and ρ, a measure of the standard deviation of the base distribution for clusters. We estimate ρ as the variance in the labeled cluster means in an episode, while α is set as a hyperparameter.

3 Experiments

We report results on multi-modal prototypes of alphabets and characters, generalization across shot and way in our new variadic regime, and standard few-shot learning benchmarks. We control for architecture and optimization by comparing methods with the same base architecture and episodic optimization settings.

We consider Omniglot [6] and mini-ImageNet [8], two widely-used few-shot learning datasets, and CIFAR-10/CIFAR-100 [4], two popular supervised learning datasets for deep learning research.

Accuracy and Generality of Multi-modal Prototypes

Our experiments on Omniglot alphabets and characters show that multi-modal prototypes are significantly more accurate than uni-modal prototypes for recognizing complicated classes (alphabets) and recover uni-modal prototypes as a special case for recognizing simple classes (characters).By unifying the clustering of labeled and unlabeled data, our multi-modal prototypes even address fully unsupervised clustering, unlike prior prototypical networks [10, 9] that are undefined without labels.

We first show the importance of multi-modality for learning representations of multi-modal classes: Omniglot alphabets. For these experiments we meta-train for alphabet classification, using only the super-class labels.

Episodes are constructed by sampling 1 example of 200 different random characters in the support set, with 5 examples of each character in the query.

Table 1: Alphabet and character recognition accuracy. BANDE improves accuracy for multi-modal alphabet classes, preserves accuracy for uni-modal character classes (Chars), and generalizes better from super-classes to sub-classes.

Training	Testing	Proto. Nets	BANDE
Alphabet	Alphabet	64.9 ± 0.2	$\mathbf{91.2} \pm 0.1$
Alphabet	Chars (20-way)	85.7 ± 0.2	$\mathbf{95.3} \pm 0.2$
Chars	Chars (20-way)	$\mathbf{94.9} \pm 0.2$	$\mathbf{95.1} \pm 0.1$

For alphabet testing, we provide 100 randomly selected characters with alphabet labels in the support, making this a mixed-shot problem. For character testing, we provide 1 labeled image of 20 different characters as support, and score based on correct character assignments of the queries. As seen in table 1, in both testing configurations, BANDE substantially outperforms prototypical networks.

Fully Unsupervised Clustering BANDE is able to do fully unsupervised clustering during meta-test via multi-modality by inferring the number of clusters. BANDE achieves good accuracy under the standard clustering metrics of normalized mutual information (NMI) and purity. We examine BANDE's clustering performance in Table 2 by randomly sampling 5 unlabeled examples of n held-out test classes.

Table 2: Unsupervised character clustering

Metric	10-way	100-way	200-way
Purity	0.97	0.76	0.63
NMI	0.95	0.90	0.87

Variadic Learning: Any-Shot, Any-Way

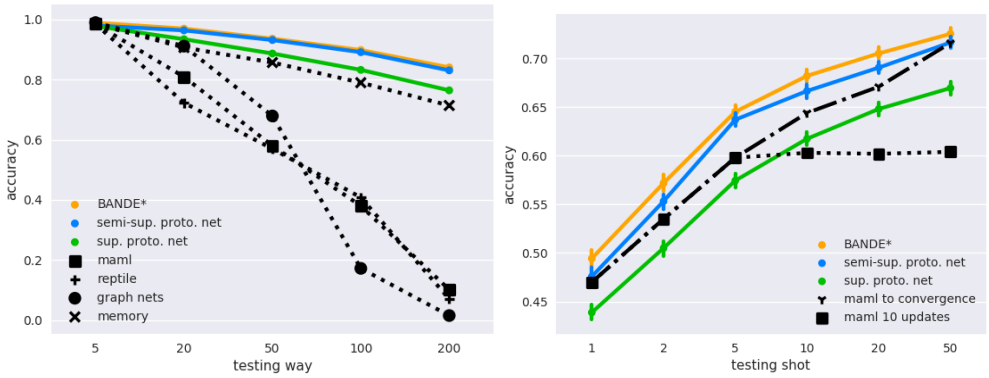
We measure generalization to differences in shot and way between meta-training and test, show extreme generalization to 1692-way classification from meta-learning on 5-way episodes, and carry out the first evaluation of scaling meta-training to the many-shot regime where our method approaches the accuracy of a supervised learning baseline.

Variable Shot and Way To measure generalization we adjust the shot and way in evaluation from their fixed settings during meta-learning. For variable way, we consider Omniglot, because it has many classes. For variable shot, we consider mini-ImageNet, because it has more examples per class. Training and inference for our method is done in the semi-supervised setting described in few-shot classification benchmarks (below, boldface).

We consider four strong baselines trained on 100% of the data, as well as prototypical baselines trained on 40% of the data. For variable shot, we compare to other prototypical methods and the gradient method MAML [1] because it is noted for scalability. For variable way we compare to MAML, the gradient method Reptile [7], the few-shot graph net of [2], and the memory model of [3].

Our method's accuracy is less sensitive to these shifts on Omniglot and mini-ImageNet (Figure 2).

For shot generalization, we compare to MAML's accuracy after 10 updates vs. accuracy at convergence. We note that MAML is not able to make effective use of more data unless it is allowed to take proportionately larger numbers of updates, while our method improves with more data without taking gradients at test time. Even at convergence, MAML does not reach our accuracies on mini-ImageNet.



(a) way generalization on Omniglot (b) shot generalization on mini-ImageNet

Figure 2: Variadic regime for any-shot, any-way generalization. Models are meta-trained with 5-way 1-shot episodes. Omniglot is tested across 5–200 classes per episode, while mini-ImageNet is tested across 1–50 examples per class. Baselines (black) are trained on 100% of the labeled data. Prototypical methods (color) are semi-supervised with 40% of the labeled data (our method starred).

Scaling to Many-Way We demonstrate that BANDE can learn a full 1692-way classifier for Omniglot from only episodic optimization of 5-way 1-shot semi-supervised tasks. See Table 3 for accuracies testing at full way (without unlabeled data or distractors).

The fully parametric baseline shares the same training set and architecture, substituting a linear output layer for prototypes by optimizing the softmax cross-entropy loss.

Table 3: Accuracy on 1692-way Omniglot from 5-way 1-shot training.

Method	1-shot	5-shot
BANDE*	**49.1**	**71.9**
semi. sup. [9]	**48.8**	71.4
fully parametric	31.0	52.0

Episodic optimization yields strong results for many-way classification, motivating the possibility of learning large-scale models cumulatively from small-scale tasks, instead of reducing large-scale models to small-scale tasks.

Scaling to Many-Shot We examine the effectiveness of BANDE in the conventional supervised learning regime. To the best of our knowledge this is the first evaluation of meta-training across the spectrum from few-shot to many-shot. Our base architecture is the Wide ResNet 28-10 of [11]. We optimize BANDE by meta-training on episodes consisting of 10-way (CIFAR-10) and 20-way (CIFAR-100) 2-shot tasks for computational considerations.

Without knowledge of the total shot or way during meta-training, and without pre-training or fine-tuning, we achieve accuracies that rival a well-tuned supervised learning baseline: 94.4% vs. 95.1% on CIFAR-10 and 75.6% vs. 81.2% on CIFAR-100. When evaluating BANDE and supervised learning embeddings as prototypes the accuracies are equal, suggesting that both approaches learn equally good representations, and differ only in the prototypical/parametric form of the classifier.

Few-Shot Classification Benchmarks

We evaluate our method for few-shot learning in the standard episodic protocol. In this evaluation protocol, shot and way are fixed and classes are balanced within an episode. We evaluate on 5 sets of 100 episodes. In the fully-supervised setting, our method learns to recover prototypical networks as a special case by assigning each class a single mode on average, while achieving equal or slightly better accuracy on Omniglot and mini-ImageNet.

In the semi-supervised setting we follow [9]. We take only 40% of the data as labeled for both the support and query while the rest of the data is included without labels. The unlabeled data is incorporated into episodes as (1) support examples that allow for semi-supervised refinement of the support classes or (2) *distractors* from the complement of the support classes.

Semi-supervised episodes augment the fully supervised n-way, k-shot support with 5 unlabeled examples for each of the n classes and include 5 more distractor classes with 5 unlabeled instances each. The query still contains only support classes. These episodes are scored at $n + 1$ way with the distractors, where classifying a query as a distractor is scored as a misclassification, as in prior work. In this setting, on 5-way 1-shot tasks, BANDE achieves 98.9% accuracy while the semi-supervised prototypical network [9] achieves 98.0%

References

[1] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[2] V. Garcia and J. Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2018.

[3] L. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. In *ICLR*, 2017.

[4] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[5] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *ICML*, 2012.

[6] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[7] A. Nichol and J. Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018.

[8] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[9] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[10] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4080–4090, 2017.

[11] S. Zagoruyko and N. Komodakis. Wide residual networks. In *BMVC*, 2016.