VARIADIC LEARNING BY BAYESIAN NONPARAMETRIC DEEP EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning at small or large scales 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 for effective learning at 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 supervision. BANDE achieves a) state-of-the-art results on semi-supervised classification of Omniglot and mini-ImageNet, b) impressive 75% classification accuracy on the 1692-way, 10-shot classification task of Omniglot while only training for 5-way 1-shot classification, c) 94.37% accuracy on CIFAR-10 by episodic optimization, comparable to state-of-the-art supervised learning techniques, and d) strong unsupervised clustering performance, with the ability to discover character classes given no character supervision.

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. Moreover, as shown in our experiments, many methods degrade when the shot and way vary between training and testing. By contrast, humans recognize both familiar and unfamiliar categories whatever the amount of data, and can even learn a new category from a single example (Lake et al., 2015). To this end, we propose one learner that generalizes across a wide range of shots and ways better than existing methods, even at the one-shot and many-shot extremes.

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.

Meta-learning provides one potential avenue for pursuing this goal. Meta-learning approaches generally use plentiful supervision from one distribution of tasks to learn an algorithm or metric that can be applied to more sparsely supervised tasks. Ideally, meta-learning approaches do not need knowledge of the specific setting in which they will be used. However, in practice, meta-learning approaches have commonly been trained and evaluated in constrained circumstances, so their generalization properties are not fully known. Perhaps most significantly, 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 their deployment to more diverse settings with variable shot and way that we address in this work.

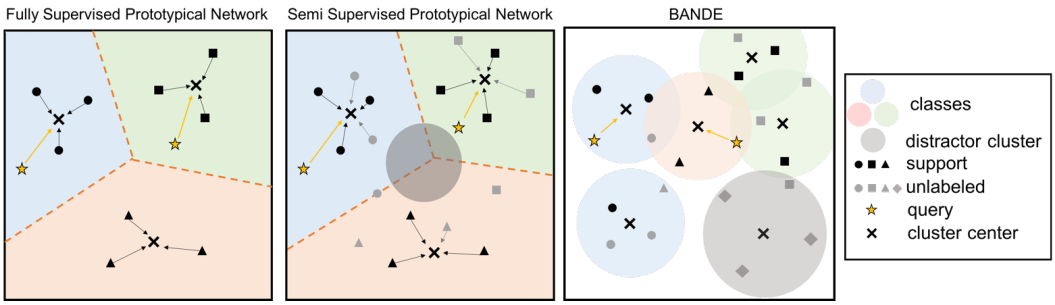


Figure 1: Our Bayesian Nonparametric Deep Embedding (BANDE) method is optimized end-to-end for semi-supervised clustering and is capable of learning multi-modal representations.

Our bayesian nonparametric deep embedding (BANDE) model (see Figure 1) is a multi-modal extension of prototypical networks. In the few-shot regime BANDE is state-of-the-art in the semi-supervised setting and equal or better than prototypical networks in the fully-supervised setting. BANDE addresses the variadic regime, generalizing to variable shot and way better than existing methods. At the many-way extreme, when trained with 5-way 1-shot episodes, BANDE achieves 75% accuracy for 1692-way 10-shot classification of Omniglot, improving on both few-shot and supervised learning baselines. At the many-shot extreme, BANDE approaches the accuracy of a standard supervised learner on CIFAR-10/100. Finally, to illustrate the importance of multi-modality, we report unsupervised and semi-supervised clustering results for the discovery of characters when only optimized for alphabet recognition.

2 Related Work

Learning Regimes Variadic learning is best explained in relation to few-shot learning, low-shot learning, and conventional supervised learning. Few-shot learning (Fei-Fei et al., 2006; Vinyals et al., 2016) handles tasks of fixed, known, and small numbers of data points and classes. In contrast, variadic tasks have variable numbers of data points and classes that can shift across tasks. Low-shot learning (Hariharan & Girshick, 2017; Qi et al., 2018; Qiao et al., 2018) addresses both densely supervised base classes and sparsely supervised novel classes, but presupposes which classes are in which set. Variadic learning also addresses these extremes of supervision, but requires no knowledge of how much or how little supervision each class has. Large-scale supervised learning (Bottou, 2010) parameterizes the model by the number of classes, and is tuned to the amount of data by choosing capacity, optimization schedules, and so forth. Variadic learning requires accuracy without specialization to shot and way. Life-long learning (Thrun, 1996; 1998) concerns variable shot and way for streams of sequential problems, while variadic learning is for one problem of unknown dimensions. Bridging life-long and variadic learning is sensible but out of scope for this work.

Metric Learning Learning a metric to measure a given notion of distance/similarity addresses recognition by retrieval: given a query, find the closest annotated support. The contrastive loss and siamese network architecture (Chopra et al., 2005; Hadsell et al., 2006) optimize an embedding for this purpose by pushing similar pairs together and pulling dissimilar pairs apart. Of particular note is research in face recognition, where a same/different retrieval metric is used for many-way classification (Schroff et al., 2015). Our approach similarly aims to learn a many-way classifier from few-way training, but is more aligned with metric learning by meta-learning (Koch, 2015; Vinyals et al., 2016; Snell et al., 2017; Garcia & Bruna, 2018).

These approaches meta-learn a distance function by direct optimization of the task loss, such as cross-entropy for classification, through episodic optimization (Vinyals et al., 2016). In episodic optimization, each episode is itself a learning task, with a labeled support set for supervision and unlabeled query set for inference, and optimization makes an update with each episode. In common practice every episode has the same, fixed number of examples and classes. We likewise learn by episodic optimization, but unlike previous meta-learning work, we aim for generalization across tasks with variable numbers of examples and classes and improve over existing work in this new regime.

Prototypes and Nonparametrics The most closely related works to our own are prototypical networks (Snell et al., 2017) and semi-supervised prototypical networks (Ren et al., 2018). Prototypical networks simply and efficiently represent the support as class-wise means defined by a choice of distance $d(x, x')$ and embedding ϕ (with a probabilistic density estimation interpretation for distances

that are Bregman divergences). They assume that the data is fully-supervised and uni-modal in the embedding. Ren et al. (2018) extend prototypes to the semi-supervised setting by refining prototypes through soft k-means clustering of the unlabeled data. We define a simpler and more general approach through bayesian nonparametrics that extends prototypical networks to multi-modal clustering of labeled and unlabeled data alike. Our method is state-of-the-art for few-shot learning, and lets us explore totally unsupervised clustering which is undefined for prior prototypical network methods.

For multi-modal clustering we incorporate DP-means (Kulis & Jordan, 2012) in our method. DP-means is a scalable, bayesian nonparametric algorithm for unsupervised clustering that creates new clusters when data points are more than a threshold λ away from existing centroids. The algorithm is derived from the asymptotics of a Gibbs sampler for the Dirichlet process mixture model in the limit of cluster radii approaching 0. Our full method 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.

3 Bayesian Nonparametric Deep Embeddings (BANDE)

Our method end-to-end learns a deep embedding network and jointly clusters labeled and unlabeled data points by bayesian nonparametrics. Crucially, our model is able to express a single category as multiple modes, unlike the uni-modal clustering approaches of prior work. As a further extension, we consider persisting clusters across episodic supports as a kind of memory-augmented meta-learning.

3.1 Foundations

We base our approach on deep embeddings, meta-learning by episodic optimization, and metric learning by prototypes. Here we briefly recapitulate these foundations.

Embedding For input, our method first computes the embedding of the data x with a feature extractor h_ϕ that is parameterized by a deep network with parameters ϕ. To be precise, the features at the final layer before the output are taken as the embedding. Our embedding network is trained by meta-learning alone, that is, without pre-training as in other work.

Meta-learning Meta-learning proceeds by episodic optimization (Vinyals et al., 2016). Each episode consists of a support set S $\{(x_0, y_0), (x_1, y_1), \ldots, (x_{n\times k}, y_{n\times k})\}$ for learning, and a query set Q $\{(x'_0, y'_0), \ldots, (x_{T\times k}, y_{T\times k})\}$ for inference. Coupled with a task loss, such as cross-entropy for classification, each episode defines a learning task. Learning proceeds by minimizing the task loss on the query, taking a gradient step after each episode to optimize ϕ. Meta-training is the process of episodically optimizing ϕ for the task loss, while meta-testing proceeds in identical episodic form, but without further optimizing ϕ. In our variadic setting, testing episodes vary in size and composition.

Prototypes BANDE builds on prototypical networks (Snell et al., 2017), a metric-learning approach with a probabilistic density estimation interpretation. Prototypes are defined as the mean of a set of points under a particular Bregman divergence $d(x, x')$ and parameters ϕ. Prototypes are formed episodically from the supervised support by taking the class-wise means of the embeddings under the Euclidean distance. As defined, prototypical networks address the fully-supervised few-shot setting by learning uni-modal class distributions.

3.2 Multi-Modality

Our method extends prototypical networks to *multi-modal* prototypes of both labeled and unlabeled data.

In order to directly capture multi-modal prototypes, we integrate the non-parametric clustering algorithm DP-means (Kulis & Jordan, 2012) into our approach. DP-means iterates through all examples in a dataset, computing the example's minimum distance to all existing cluster means. If this distance is greater than a particular threshold λ, a new cluster is created with mean $h_\phi(x_i)$, the example assigned to it. If x_i is labeled, the new cluster takes on its label.

Algorithm 1 BANDE

Initialize $\{\mu_0, \mu_1, ..., \mu_n\}$
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**
    $\mu_{n+1} \leftarrow h_\phi(x_i)$
    $N_{n+1} = 1$
  else
    $N_c = N_c + 1$
  end if
end for
update soft assignments $z_{i,c}$
update cluster means μ_c

In this clustering scheme a single pass is sufficient, unlike iterative clustering methods such as k-means that may require multiple passes.

BANDE further modifies DP-means to allow for soft cluster assignments. After constructing a set of cluster means μ_c, we calculate soft assignments $z_{i,c} = \frac{\mathcal{N}(h_\phi(x_i);\mu_c,\sigma_c)}{\sum_c \mathcal{N}(h_\phi(x_i);\mu_c,\sigma_c)}$ and then calculate final cluster means $\mu_c = \frac{\sum_i z_{i,c} h_\phi(x_i)}{\sum_i z_{i,c}}$ from these soft assignments.

The loss is defined as the negative log-likelihood of the query embeddings against the final inferred cluster means, using a normalized Gaussian likelihood. Since there are potentially multiple clusters with the same label, we use a weighted cross-entropy 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 embeddings that are potentially multi-modal, as directed by the data and optimization.

BANDE is initialized with n cluster means $\{\mu_0, \mu_1, ..., \mu_n\}$ for the n labeled classes in the support set, computed in the same way as standard prototypical networks. Created labeled clusters are instantiated with a radius σ, while unlabeled clusters take a radius of σ_u to capture increased uncertainty about the unlabeled distribution (which can and does contain multiple unknown classes).

Unlike σ and σ_u, λ, the threshold for creating a new cluster, is non-differentiable and cannot be learned jointly. Instead, we set λ episodically based on its derivation from Kulis & Jordan (2012), where it is defined in terms of α, the relative probability of forming a new cluster in the Chinese Restaurant Process prior, and ρ, a measure of the standard deviation for the base distribution from which clusters are assumed to be drawn. We estimate ρ as the variance in the labeled cluster means within an episode, while α is treated as a hyperparameter.

3.3 Cumulative Supervision

We extend BANDE into a cumulative variant, BANDE-C, that accumulates supervision non-episodically by making prototypes persistent across episodes. Concretely, we initialize the cluster means μ_c by including a cluster mean from memory, $\phi_{m,c}$, with the current episodic sample mean (i.e. $\mu_c = \frac{1}{|z_i \in C|+1}(\phi_{m,c} + \sum_{i,z_i \in c} h_\phi(x_i))$. $\phi_{m,c}$ is computed as if c was uni-modal, regardless of whether the clustering was multi-modal in a previous episode. Since the embedding representation rapidly changes early in training, we introduce a discount factor on the stored embedding $\gamma\phi_{m,c}$ proportional to the current learning rate. Whenever the class is encountered in a future episode, we update the stored prototype with the cluster mean after episodic inference. Note that standard prototypical networks can likewise be augmented to remember prototypes and non-episodically accumulate supervision in this manner. We only experiment with BANDE-C in the variadic setting (Section 4.2); everywhere else we keep standard episodic training and testing.

3.4 Probabilistic Interpretations and Alternatives

BANDE connects end-to-end learning and bayesian nonparametric clustering, and our experiments show it is effective in practice (Section 4), but it does not have an immediate probabilistic interpretation. Here we explain an alternative instantiation of our method that is more theoretically clear while still being end-to-end learnable. When theoretically pure, DP-means is used without the last step of soft clustering, and the clusters do not have associated radii. While this is not differentiable through the hard assignments, it is still learnable via the query because the query and support embeddings share weights. In preliminary experiments this alternative is marginally less accurate so we use soft assignment with cluster radii throughout.

For further connections, we reinterpret the prior work on semi-supervised prototypical nets (Ren et al., 2018) through the bayesian nonparametric lens to draw a connection with the Chinese Restaurant Process (CRP) (Aldous, 1985) in Section A.4 of the appendix.

3.5 Implementation Details

In all Omniglot experiments, both the unlabeled and labeled cluster radii are initialized to 5.0. The labeled cluster radius is then treated as a parameter, and learned over training. We found that allowing

the unlabeled cluster radius to be learned resulted in unstable training, and it was therefore fixed. α was always set to 0.05. We found that a range of α values worked well, and the network learned to produce embeddings of the right overall magnitude to match α.

In mini-ImageNet experiments, the unlabeled and labeled cluster radii are initialized to 20.0, and both are learned over training. We found that on average, the labeled cluster radius stabilized around 15, and the unlabeled cluster radius stabilized around 25. α was set to 0.00001, but again the embeddings learned to reach a magnitude where new clusters were still consistently created.

4 EXPERIMENTS

We report state-of-the-art results on the standard fully-supervised and semi-supervised few-shot classification problems, promising first experiments in our new variadic regime, and unsupervised clustering results made possible by the multi-modality of the bayesian nonparametric aspect of our method. We control for architecture and optimization by comparing methods with the same base architecture and same episodic optimization settings. All code for our method and baselines will be released for reproducibility.

For these experiments we make use of standard few-shot and supervised learning datasets and furthermore define new variadic evaluation protocols on these common benchmarks. We consider Omniglot and mini-ImageNet, two widely used datasets for few-shot learning research, and CIFAR-10/CIFAR-100, two popular datasets for supervised learning research with deep networks.

Omniglot (Lake et al., 2015) is a dataset of 1,623 handwritten characters from 50 alphabets. There are 20 examples of each character, where the images are resized to 28x28 pixels and each image is rotated by multiples of 90°. This gives 6,492 classes in total, which are then split into 4,112 training classes, 1,692 test classes and 688 validation classes.

mini-ImageNet (Vinyals et al., 2016) is a reduced version of the ILSVRC'12 dataset (Russakovsky et al., 2015), which contains 600 84x84 images for 100 classes randomly selected from the full dataset. We use the split from Ravi & Larochelle (2017) with 64/16/20 classes for train/val/test.

CIFAR-10/100 (Krizhevsky & Hinton, 2009) are classification datasets of 32x32 color images drawn from the Tiny Images project (Torralba et al., 2008). CIFAR-10 has 10 classes and CIFAR-100 has 100 classes (plus 20 super-classes). Both have 50k training images and 10k testing images and both are balanced so that every class has an equal number of images.

For both Omniglot and mini-ImageNet, unless otherwise specified, we use only 40% of the data for labeled support and query, leaving 60% of the data for the semi-supervised setting. For episodic tasks we evaluate 5 sets of 100 episodes. Details on training and architecture are provided in the appendix.

4.1 FEW-SHOT CLASSIFICATION

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. In the fully-supervised setting, our method learns to recover prototypical networks as a special case while achieving equal or better accuracy. That is, we carry out full multi-modal learning and inference without special tuning and our method learns to assign each class a single mode on average. Note our results are achieved by episodic optimization alone without the pre-training done in other work.

The results reported in Table 1 are for models trained and tested with n-way episodes. This is to equalize comparison across methods. Snell et al. (2017) try training at higher-way than testing and report a boost in accuracy. We find that this boost is illusory, and explained away by controlling for the number of gradients per update. We show this by experiment through the use of gradient accumulation in Section A.2 of the appendix. (For completeness, we confirmed that our implementation of prototypical networks reproduce reported results at the chosen higher training way, and that BANDE trained at this higher way performs similarly well.)

In the semi-supervised setting we follow (Ren et al., 2018). We take only 40% of the data as labeled *for both the support and query* while the rest of the data is included, but as unlabeled examples. The unlabeled data is then incorporated into episodes as (1) within support examples that allow for semi-supervised refinement of the support classes or (2) *distractors* which lie in the complement of

Table 1: Fully-supervised few-shot results.

	Omniglot				mini-ImageNet	
	5-way		20-way		5-way	
Method	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
BANDE*	98.4 ± 0.1	99.5 ± 0.1	95.1 ± 0.1	98.6 ± 0.1	48.9 ± 0.7	$\mathbf{68.3 \pm 0.6}$
Prototypes (Snell et al., 2017)	98.3 ± 0.2	99.6 ± 0.1	94.9 ± 0.2	98.6	46.4 ± 0.78	67.0 ± 0.7
MAML (Finn et al., 2017)	98.7 ± 0.4	99.9 ± 0.3	95.8 ± 0.3	98.9 ± 0.2	48.7 ± 1.84	63.1 ± 0.92
Graph nets (Garcia & Bruna, 2018)	**99.2**	99.7	**97.4**	99	**50.3**	66.41
Memory (Kaiser et al., 2017)	98.4	99.6	95	98.6	-	-

Table 2: Semi-supervised few-shot results on 40% of the labeled data with distractors.

	Omniglot				mini-ImageNet	
	5-way		20-way		5-way	
Method	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
BANDE*	$\mathbf{98.9 \pm 0.1}$	$\mathbf{99.4 \pm 0.1}$	$\mathbf{97.0 \pm 0.1}$	$\mathbf{98.4 \pm 0.1}$	$\mathbf{49.2 \pm 0.7}$	$\mathbf{66.1 \pm 0.7}$
Ren et al. (2018)	98.0 ± 0.1	$\mathbf{99.2 \pm 0.1}$	96.4 ± 0.1	98.1 ± 0.1	$\mathbf{48.6 \pm 0.6}$	63.8 ± 0.8
Snell et al. (2017)	97.8 ± 0.1	$\mathbf{99.2 \pm 0.1}$	93.4 ± 0.1	98.0 ± 0.1	43.9 ± 1.0	63.6 ± 0.5

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. Our results for this setting are reported in Table 2.

Through multi-modality, the clustering of the labeled classes and distractors is decided by the data with a single rule. In particular this helps with the distractor distribution, which is in fact more diffuse and multi-modal by comprising several different classes. Our only specialization to this setting is to have more uncertain distractor clusters by higher cluster radii to compensate for this diffuseness.

4.2 Variadic Learning at Any-Shot and Any-Way

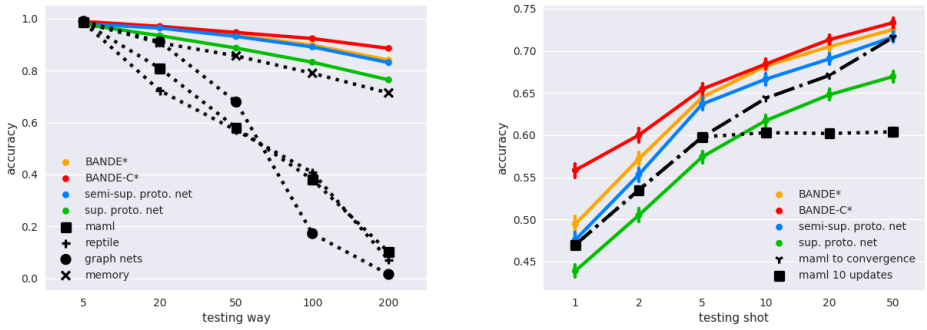
We explore our proposed variadic regime with a thorough examination of generalization across shot and way. We begin by measuring generalization to differences in shot and way between meta-learning and test in the realm of few-shot learning. Next we demonstrate more extreme generalization to 1692-way classification from learning with only 5-way episodes. Finally, we carry out the first evaluation of meta-learning spanning from the few-shot regime to the many-shot supervised learning regime, and find that our method approaches the accuracy of a well-tuned supervised learning baseline. This type and degree of generalization has only received limited attention in meta-learning research, with experiments restricted to few-way shifts (Munkhdalai & Yu, 2017; Snell et al., 2017).

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 the previous section. This underscores the suitability of BANDE for variadic learning: it learns from any number of labeled and unlabeled examples.

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 (Finn et al., 2017) because it is noted for scalability. For variable way we compare to MAML, the gradient method Reptile (Nichol & Schulman, 2018), the few-shot graph network of Garcia & Bruna (2018), and the memory-based model of Kaiser et al. (2017). Modifications to these approaches to allow for test-way generalization are discussed in Section A.3.

As seen in Figure 2, the accuracy of our method is less sensitive to these shifts on both Omniglot and mini-ImageNet, and in particular it maintains accuracy at higher way, and continues to improve with larger numbers of examples on mini-ImageNet. Graph networks and gradient methods degrade more with respect to different way. These results are provided for comparison, and we encourage further work to explore how to improve generalization for each class of method.



(a) way generalization on Omniglot

(b) shot generalization on mini-ImageNet

Figure 2: Variadic setting. Models are meta-trained with 5-way 1-shot episodes. Omniglot is tested with 1-shot episodes across 5–200 classes per episode, while mini-ImageNet is tested with 5-way episodes 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 methods starred).

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, whereas our method is able to make use of more data while taking 0 gradients at test time. Even at convergence, MAML does not improve on our mini-ImageNet accuracy for any number of examples.

Extreme Generalization to Many-Way

We demonstrate that our method can learn a full 1692-way classifier for Omniglot from only episodic optimization of 5-way, 1-shot tasks. Episodes are composed identically to the few-shot semi-supervised setting including unlabeled examples and distractor classes.

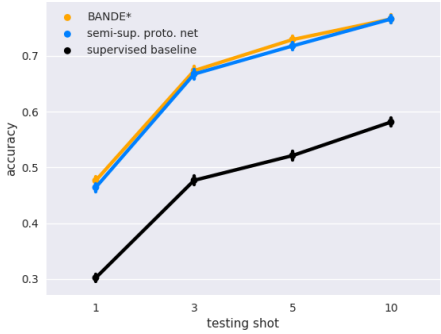


Figure 3: Accuracy on the full 1692-way classification task, consisting of the entire Omniglot test set, from training on 5-way 1-shot episodes.

Accuracies for our method and a supervised learning baseline are shown in Figure 3. To perform inference, we ran k examples from each class in the test set through our learned embedding network, and then computed the distance from unseen examples to each of prototype. The supervised learning baseline shares the same training set and architecture, substituting a linear output layer for prototypes by optimizing the softmax cross-entropy loss. For testing, we take the last feature layer as an embedding for prototypical inference. Fine-tuning on the test support proved less accurate, as did k nearest neighbours inference.

This result is an example of episodic optimization yielding strong results for many-way classification, motivating the possibility of learning large-scale models cumulatively from small-scale tasks, instead of restricting attention to the adaptation of large-scale models to small-scale, few-shot settings.

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 Zagoruyko & Komodakis (2016), which has shown state-of-the-art results on CIFAR-10/100, and has been additionally used as a base architecture for strong low-shot performance on mini-ImageNet (Qiao et al., 2018). We optimize BANDE by meta-training on episodes consisting of 20-way (CIFAR-100) and 10-way (CIFAR-10) 2-shot tasks for computational considerations.

With no knowledge of the total number of classes or number of examples per class, and without pre-training or fine-tuning, we achieve accuracies that rival a well-tuned supervised learning baseline. On CIFAR-100 we achieve 75.63% accuracywhich is $> 90\%$ of the 81.2% accuracy of supervised learning. On CIFAR-10 we achieve 94.37% accuracy compared to the 95.1% accuracy of supervised learning. When evaluating both the 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 vs. parametric form of the classifier.

4.3 MULTI-MODAL CLASS DISCOVERY

We show multi-modality is important for both a) zero-shot unlabeled cluster discovery and b) one-shot labeled cluster discovery on Omniglot. In these experiments we share the same radius setting across labeled and unlabeled clusters to gauge the effect of multi-modality for both labeled and unlabeled data. We use the models trained in the semi-supervised setting from Section 4.1.

Zero-Shot Class Discovery BANDE is able to perform fully unsupervised clustering, unlike prototypical or semi-supervised prototypical networks. To examine BANDE's performance here, we randomly sampled 5 examples of n classes from the test set and treat them as unlabeled samples. We use one batch of data from the training set to estimate λ, and then cluster in a single pass.

The standard clustering metrics, purity and normalized mutual information (NMI), are given in table 3. BANDE maintains remarkably strong performance across a large number of unlabeled clusters.

Table 3: Unsupervised clustering scores

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

Sub-class Discovery from Super-class Training Next, we show the importance of multi-modality for labeled clusters in discovering a flat alphabet-character hierarchy with multiple modes from *alphabet supervision alone*. Meta-training is performed on alphabet classification, with only the super-class alphabet label being observed. Episodes are constructed by sampling 1 example of 200 different random characters. The query set consists of 5 samples of each character, and the radii of clusters are fixed to 5, without allowing for updates over training. Otherwise, all hyperparameters for BANDE are the same as for the semi-supervised few-shot training experiments.

For zero-shot character clustering, we present BANDE with 500 instances drawn from 100 different character classes. We choose λ based on alphabet supervision from one batch of alphabet training data. BANDE then successfully clusters characters together, achieving a purity score of 0.91, with an NMI score of 0.95, creating an average of 130.3 clusters. It does so without ever having seen character labels during meta-train *or* meta-test.

We examine the ability of BANDE to perform alphabet and character recognition during meta-test from a single label. Note that we still never see character labels during meta-training.

Table 4: Accuracy on multi-modal class discovery

Training	Testing	Proto. Nets	BANDE
Alphabet	Alphabet	64.9 ± 0.2	$\mathbf{91.2} \pm 0.1$
Alphabet	Characters (20-way)	85.7 ± 0.2	$\mathbf{95.3} \pm 0.2$
Alphabet	Characters (100-way)	70.9 ± 0.2	$\mathbf{87.0} \pm 0.1$

For alphabet testing, we provide 100 randomly selected characters with alphabet labels in the support, and score based on query alphabet assignments. For character testing, we provide 1 labeled image for each of $\{20, 100\}$ characters as support, and score based on assignments of the query samples to the correct characters. As seen in table 4, in both testing configurations, our method substantially outperforms prototypical networks. Its accuracy in the 20-way 1-shot character recognition task (95.3%), trained with alphabet supervision alone, rivals prototypical networks trained to do character classification directly (96%).

5 CONCLUSION

We framed the variadic regime to shine a light on learning representations that bridge small-scale and large-scale learning and strive toward the any-shot/any-way adaptability of human perception. As a step toward addressing this full span, we introduced BANDE, a multi-modal extension of prototypical networks, that is capable of generalizing across variable amounts of labeled and unlabeled data. Our results have shown BANDE is state-of-the-art in the few-shot regime and scales from few-way, few-shot meta-learning to many-way, many-shot deployment for both sparse and plentiful supervision. Our experiments demonstrate that multi-modality is key for improved semi-supervised and unsupervised clustering. There is much work to be done to improve variadic generalization, and to connect to life-long learning over non-stationary tasks.

REFERENCES

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pp. 1–198. Springer, 1985.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pp. 177–186. Springer, 2010.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pp. 539–546. IEEE, 2005.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *PAMI*, 2006.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2018.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pp. 1735–1742. IEEE, 2006.

Bharath Hariharan and Ross B Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, pp. 3037–3046, 2017.

Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. In *ICLR*, 2017.

Gregory Koch. Siamese neural networks for one-shot image recognition. In *NIPS Deep Learning Workshop*, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *ICML*, 2012.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pp. 2554–2563, 2017.

Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018.

Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *CVPR*, pp. 5822–5830, 2018.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

Yordan P Raykov, Alexis Boukouvalas, and Max A Little. Simple approximate map inference for dirichlet processes. *arXiv preprint arXiv:1411.0939*, 2014.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

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. *IJCV*, 115(3):211–252, 2015.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pp. 815–823, 2015.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pp. 4080–4090, 2017.

Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NIPS*, pp. 640–646, 1996.

Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pp. 181–209. Springer, 1998.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 2012.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2008.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pp. 3630–3638, 2016.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

A Appendix

A.1 Implementation Details

For all few-shot experiments, we use the same base architecture as prototypical networks for the embedding network. It is composed of four convolutional blocks consisting of a 64-filter 3 x 3 convolution, a batch normalization layer, a ReLU nonlinearity, and a 2 x 2 max-pooling layer per block. This results in a 64-dimensional embedding vector for omniglot, and a 1600 dimensional embedding vector for mini-imagenet. Our models were trained via SGD with RMSProp (Tieleman & Hinton, 2012) with an α parameter of 0.9. For Omniglot, the initial learning rate was set to 1e-3, and cut by a factor of two every 2000 iterations, starting at 4000 iterations. We additionally use gradient accumulation and accumulate gradients over eight episodes before making an update when performing 5-way training for Omniglot. For mini-ImageNet, the initial learning rate was set to 1e-3, and further halved every 20000 iterations, starting at 40000 iterations.

For the supervised experiments, we use a wide residual network Zagoruyko & Komodakis (2016) with depth 28 and widening factor 10, with a dropout value of 0.3. We were not able to perfectly recover published results with our reimplementation, but the numbers are within 1% of their published values.

A.2 Controlling for the Number of Gradients Taken During Optimization

Consider the gradient of the loss: it has the dimensions of shot $\times$ way because every example has a derivative with respect to every class. In this way, by default, the episode size determines the number of gradients in an update. Quantitatively, 20-way episodes accumulate 16 times as many gradients as 5-way episodes. By sampling 16 5-way episodes and accumulating the gradients to make an update, we achieve significantly better results, matching the results obtained with 20-way episodes within statistical significance. Note that agreement across conditions may not be perfectly exact because subtle adjustments to hyperparameters might be necessary.

Table 5: Results on Omniglot for different gradient accumulations. Bolded results are not significantly different from each other.

			5-way		**20-way**	
Shot	Batch-way	Episode-way	1-shot	5-shot	1-shot	5-shot
1	20	20	**98.5**	**99.6**	**95.0**	**98.8**
1	20	5	**98.3**	**99.5**	**94.8**	**98.6**
1	5	5	97.7	**99.4**	92.1	98.0
5	20	20	**97.8**	**99.6**	**93.2**	**98.6**
5	20	5	**97.9**	**99.6**	**92.9**	**98.5**
5	5	5	96.8	**99.4**	89.8	97.7

A.3 EXTENDING COMPARED MODELS TO VARIADIC REGIME

The models we compare to were not designed with variadic generalization in mind, and as a result we attempt to make as fair a comparison as possible by extending them as needed. We describe our approaches below.

Semi-supervised prototypical networks In the paper first introducing this semi-supervised setting (Ren et al., 2018), the authors show how to use a distractor cluster centered at 0 to capture samples not belonging to any examples from the support. They additionally introduce length scales r_c. In equation 6 from their paper, they use a normalization constant $A(r_c)$ defined as $0.5\log(2\pi)+\log(r)$. However, this is an unscaled normalization constant, and assumes the dimensionality of the embedding space to be 1. The corrected normalization constant is $A(r_c) = d(\log(r_c) + 0.5\log(2\pi))$ where d is the dimensionality of the embedding. We compare to their method with this corrected normalization constant, but note that it has only a small effect. For space, we did not compare to all methods from their paper, and chose this one as it performed well across their experiments, and because it was most amenable to the clustering experiments we were interested in performing.

MAML We used Finn's publicly available github repository (Finn et al., 2017). We trained an initial MAML architecture on the 5-way 1-shot task, using the suggested hyperparameters, for 40,000 iterations. We then removed the classification layer, froze the remaining weights of the network, and retrained the top layer for the testing n-way classification task, using the MAML objective again, for 5000 iterations. We tried two hyperparameter settings for the re-training: the hyperparameters for the 5-way 1-shot setting, and the hyperparameters for the 20-way 1-shot setting. We found that re-training with the 20-way 1-shot hyperparameters gave us better performance. While we attempted to also scale these hyperparameters appropriately for even higher way testing, this was not more successful than using the 20-way 1-shot hyperparameters. We then reported the accuracy after 10 update steps on the test data.

We also tried simply randomly initializing the top-layer weights, and allowing MAML to take more update steps to see if it could learn the top layer online. These results were worse than those obtained after the fine-tuning procedure.

Reptile We used the publicly available github repository from OpenAI. We used transductive training for 100,000 iterations on the 5-way 1-shot task, using the suggested hyperparameters. We then removed the classification layer, froze the remaining weights of the network, and retrained the top layer for the testing n-way classification task, using the Reptile training procedure. As in MAML, we tried setting hyperparameters during re-training to be similar to 5-way 1-shot, and 20-way 1-shot, but did not notice significant differences. Using random initializations for the top-layer weights, and then applying "fast weight" updates at test time also worked reasonably well.

Graph Neural Networks Modifying the Graph Neural Network architecture to be applicable for test-way generalization was more difficult, since the approach assumes that labels are represented as a one-hot encoding, and concatenated with node features before being fed to the metric network. At training, we padded the one-hot labels to allow for 200 possible classes. At test time, these could then be filled in without needing to completely retrain the metric network. We additionally fine-tuned the classification layer of the metric network. We were unable to achieve greater than chance performance for the 200-way task. We expect that this is because the metric network learns to ignore the padded input dimensions during training. One possible fix would be to randomize the labels during training to fall in the full (0, 200) range, but we leave this to future work. Scaling this approach up to full-way classification is impossible with this encoding of the labels, as the computational memory requirements are substantial.

A.4 CHINESE RESTAURANT PROCESS REINTERPRETATION

The generative model for the CRP consists of sampling assignments $z_1, ..., z_n$ which could take on cluster values $c = 1, ..., C$ from the CRP prior with hyperparameter α and number of cluster members N_c. Cluster parameters μ_c, σ_c are sampled from a base distribution $H(\theta_0; \mu_0, \sigma_0)$, and instances x_n are then sampled from the associated Gaussian distribution $\mathcal{N}(\mu_{z_n}, \sigma^2_{z_n})$. Assuming that σ_c is fixed and spherical for all c, we can use iterated conditional modes (Raykov et al., 2014) to obtain a MAP estimate to the CRP by greedily estimating the best assignment z_i for each example, and then estimating the best μ given the assignments.

The CRP generative model is defined as

$$p(z_{N+1} = c|z_{1:N}, \alpha) = \frac{N_c}{N+\alpha} \text{ for } c \in \{1..C\} \text{ and } p(z_{N+1} = c|z_{1:N}, \alpha) = \frac{\alpha}{N+\alpha} \text{ for } c = C+1 \tag{1}$$

for assignments z of examples x to clusters c, cluster counts N_c, and parameter α, to control assignments to new clusters. N is the total number of examples observed so far.

For the case of a spherical gaussian likelihood, let us define $\mathcal{N}_c = \mathcal{N}(x_i; \mu_c, \sigma^2)$ as the likelihood of assigning x_i to cluster c and $\mathcal{N}_0 = \mathcal{N}(x_i; \mu_0, \sigma^2 + \sigma_0^2)$ as the likelihood of assigning x_i to a new cluster drawn from the base distribution (Gaussian with mean μ_0 and σ_0^2) . We can then write:

$$p(z_i = c|\mu) = \frac{N_{k,-n}\mathcal{N}_c}{\alpha\mathcal{N}_0 + \sum_{j=1}^{C} N_{j,-n}\mathcal{N}_j} \tag{2}$$

$$p(z_i = C+1|\mu) = \frac{\alpha\mathcal{N}_0}{\alpha\mathcal{N}_0 + \sum_{j=1}^{C} N_{j,-n}\mathcal{N}_j} \tag{3}$$

$$p(\mu|z) = \mathcal{N}\left(\mu_c; \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{i,z_i=c} x_i}{\sigma_c^2 + \sigma_0^2 N_c}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2 N_c}\right) \tag{4}$$

Let $q(i, c)$ be $\log p(i, c)$, the joint probability of cluster C and assignment i. If we additionally assume that μ_0 is equal to 0, then by iterated conditional modes, we obtain an algorithm very similar to the k-means algorithm, outlined in A.4:

Algorithm 2 Iterated conditional modes for CRP

initialize $\{\mu_0, \mu_1, ..., \mu_n\}$
for each unlabeled example i **do**
  for each cluster c **do**
    $q_{i,c} \propto -\log N_{c,-i} - \|\frac{h_\phi(x_i)-\mu_c}{2\sigma_c^2}\|^2$
  end for
  $q_{i,C+1} \propto -\log \alpha - \|\frac{h_\phi(x_i)-\mu_0}{2\sigma_0^2}\|^2$
  $z_i = \arg\min_c q_{i,c}$
end for
update cluster hyperparameters
$\mu_c = \frac{\sigma_0^2 \sum_{i=1,z_i=c} h_\phi(x_i)}{\sigma_c^2 + \sigma_0^2 N_c}$

Determining the assignment for a query sample is performed after clustering using the updated means and cluster counts. This then results in approximately the same update to the prototype as (Ren et al., 2018), but with an assignment z which includes the CRP prior. This suggests re-interpreting the distractor cluster from their method as a draw from a general base distribution, with a CRP prior placed on the relative assignments to clusters.

A.5 Semi-supervised Clustering Evaluation

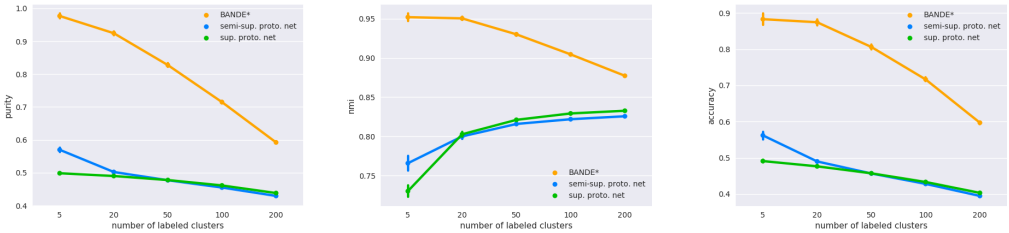
We show the importance of multi-modality for discovering unlabeled clusters during meta-testing after semi-supervised meta-learning.

We randomly sampled n classes from Omniglot's test set, and ran one randomly selected example from each of the n classes through our model to obtain a set of n prototypes.

We then presented a new set of examples drawn equally from these n support (known) classes and n out-of-support (unknown) classes, then let each method cluster the examples into either the computed n prototypes or new clusters.

Figure 4 shows the n+1 accuracy of correctly classifying the new example as either the correct, known label, or correctly identifying the example as a distractor. Only BANDE achieved higher accuracy

than chance for numbers of clusters greater than 5, suggesting that a multi-modal distribution for unlabeled clusters is paramount for the algorithm's clustering performance. The number of clusters created for the unlabeled examples closely tracked the correct number of unlabeled clusters, with an average relative error in the number of created clusters of 1.87 across the range from 5 - 200.



(a) Purity of clusters (b) Normalized Mutual Information (c) Accuracy on n+1-way

Figure 4: Cluster discovery metrics for Omniglot. Trained on 5-way 1-shot episodes.