Open-Set Incremental Learning via Bayesian Prototypical Embeddings

John Willes[1], James Harrison[2], Ali Harakeh[1],
Chelsea Finn[2], Marco Pavone[2], Steven Waslander[1]
University of Toronto[1], Stanford University[2]
john.willes@mail.utoronto.ca, jh2@stanford.edu, ali.harakeh@utoronto.ca,
cbfinn@cs.stanford.edu, pavone@stanford.edu, stevenw@utias.utoronto.ca

Abstract

As autonomous decision-making agents move from narrow operating environments to unstructured worlds, learning systems must move from a closed-world formulation to an open-world, lifelong, few-shot setting in which agents continuously learn new classes from small amounts of information. This stands in stark contrast to modern machine learning systems that are typically designed with a known set of classes and a large number of examples for each class. In this work we extend embedding-based few-shot learning algorithms toward open-world problems. We combine Bayesian non-parametric class priors with an embedding-based pre-training scheme to yield a highly flexible framework for use in both the lifelong and the incremental settings. We benchmark our framework on miniImageNet and TieredImageNet in the lifelong setting. Our results show, compared to prior methods, up to a 14% classification accuracy improvement from our novel pretraining scheme and up to a 22% improvement in AUROC (a measure of novel class detection) from our non-parametric few-shot learning scheme.

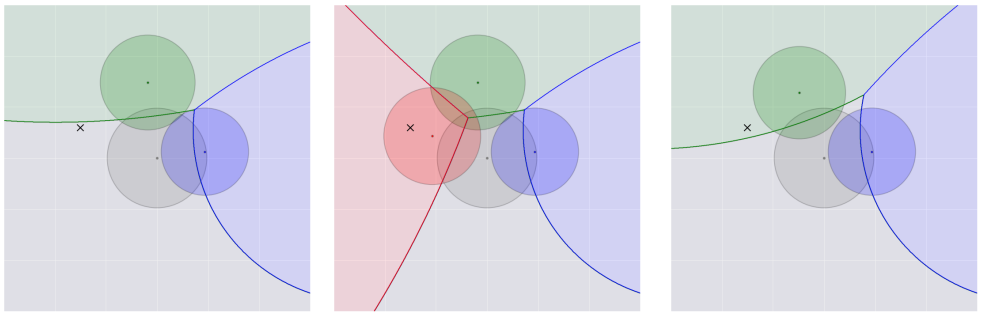
1 Introduction

The standard setting for classification systems is *closed-world*: a fixed set of possible labels is specified during training over large datasets, and this set remains fixed during deployment [18]. This closed-world approach stands in stark contrast with human learning. By continually integrating novel information, we continually learn new labels from small amounts of new data. As autonomous decision-making agents move from highly structured operating environments to unstructured ones, learning systems must consider *open-world, few-shot*, and *lifelong* settings in which agents continuous learn new labels from limited amounts of new information in the wild.

In this work, we present a novel approach to classification in this open setting. We focus on the open-world, few-shot lifelong learning setting, in which the set of test and train classes are disjoint. This setting is foundational for related settings, such as incremental learning (in which the test and train classes may overlap), and we believe progress in this setting may accelerate progress in all types of open-world learning.

Our approach combines ideas from Bayesian non-parametrics [14] with a Bayesian formulation of prototypical few-shot learning to yield a highly flexible and simple non-parametric model capable of reflecting uncertainty in whether a class is novel. In particular, we leverage a Chinese restaurant process (CRP) class prior—a prior on an unbounded number of classes [7]—along with a Bayesian embedding-based meta-learning algorithm [8]. To improve the performance of this framework, we present an embedding-based pre-training phase in which a standard fully-connected classification head is replaced with Gaussian class distributions in feature space for pre-training. Together, these components enable efficient and effective lifelong learning.

4th Workshop on Meta-Learning at NeurIPS 2020, Vancouver, Canada.



(a) L-PCOC Prediction (b) New Class Instantiation (c) In-Distribution Update

Figure 1: A 2D visualization of L-PCOC decision boundaries and adaptation. The shared prior and its decision boundary corresponding to a novel class are depicted in gray. The class posterior predictive distributions and their associated decision boundaries are colored blue, green and red. Each circle depicts the 2σ confidence interval of an isotropic Gaussian. The black "X" denotes a query feature vector. In this case (fig.1a), L-PCOC classifies the query as a novel class, however, it is close to the decision boundary. L-PCOC will then receive the true class label. If the label corresponds to a novel class (fig.1b), L-PCOC will instantiate a new class distribution from the shared prior and condition on the query feature (shown in red). In the case that label corresponds to the green class (fig.1c), L-PCOC will update the green class posterior predictive to extend the decision boundary to include the query.

Contributions. There are three core contributions in this paper.

- We propose a formalization of *open-world, few-shot lifelong* learning, in which decision-making agents must detect novel classes and then rapidly adapt and generalize given limited labeled data.
- We introduce a Bayesian few-shot learning scheme based on Gaussian embeddings [25]. We combine this approach with a Bayesian non-parametric class prior, and show this system is capable of effectively incorporating novel classes for few-shot, open-world, lifelong learning. Moreover, we show that this scheme results in a substantial 22% improvement (AUROC) in the detection of novel classes compared to baseline methods.
- We introduce an embedding-based pre-training phase, linking the pre-training to the meta-training phase. We show that this pre-training improves performance by as much as 14% relative to standard pre-training with a fully-connected classification head. Improvement is observed for all methods evaluated, including the baselines. We further show this embedding-based pre-training, when combined with our lifelong learning scheme, leads to a conceptually simple scheme for incremental learning, in which training classes appear at test time.

2 Problem Statement

In this work, we aim to develop a classification model that is able to detect novel classes during deployment. Moreover, the model must be able to incorporate a small number of examples of a novel class to rapidly improve performance. In particular, we aim to develop a model capable of *open-world, few-shot* learning in the *lifelong* setting. Because the terminology in continuous, lifelong, and incremental learning is often ambiguous and contradictory [17], we highlight precisely what we mean with these descriptors. In particular, our model must be able to perform on three characteristics of real-world classification problems,

- **Open-world**: The set of labels is not known to the learning agent during training. Thus, the agent must be able to detect when an input corresponds to a never-before-seen label.
- **Few-shot**: When the label for a new class is observed, the learning agent must be able to rapidly learn to identify this class based on a small number of examples.
- **Lifelong** learning: at deployment, the agent starts with no knowledge of the space of possible labels and must learn all class labels.

Formally, we consider a setting in which we are provided a labeled dataset consisting of N_{train} classes. We wish to deploy the model in a setting in which there are N_{test} classes, with no overlap between the train classes and the test classes. We assume that during deployment, classes are sampled i.i.d. and images of the corresponding class are presented to the agent. The agent returns a probabilistic belief that the image corresponds to a previously observed class or that the image belongs to a class that has not been previously observed. For accuracy evaluation purposes, we equate

Table 1: MiniImageNet - Open-World Few-Shot Lifelong Results

Method	Pre-Training	Acc. (%)	Support Acc. (%)	Inc. Acc. (%)	AUROC (%)
NCM	Sup-FC	33.35	36.68	26.68	43.28
NCM	Sup-E	**42.71**	**49.60**	**28.94**	31.65
ProtoNet	Sup-FC	34.40	39.92	23.34	41.91
ProtoNet	Sup-E	37.78	43.68	25.97	43.87
L-PCOC (Ours)	Sup-FC	38.98	46.17	24.61	67.31
L-PCOC (Ours)	Sup-E	41.77	**49.54**	26.21	**69.19**

correctly predicting a class has not previously been observed to a "correct" prediction. After making a prediction, the agent is provided with the label.

3 Approach

The lifelong PCOC (L-PCOC) approach relies on Bayesian Gaussian Discriminant Analysis (GDA) in the embedding space of an encoding network. This may be seen as augmenting the approach of prototypical networks [25] with a prior over embeddings. Lifelong learning is enabled using this prior; the "correct" label for a novel class is that corresponding to the prior. Once a new class is observed, the posterior is computed using the shared prior over embeddings. As a prior over classes, we use the Chinese restaurant process [7], which has seen previous application in incremental and lifelong variations of meta-learning [13, 19] and is a flexible approach to clustering with an unknown number of classes.

Our approach relies on three phases. We first pre-train the encoder, which has been shown to substantially improve performance of meta-learners in the few-shot setting [3, 26]. The second phase consists of a meta-training phase, either in the lifelong setting or the incremental setting. The third phase is a test-time fine-tuning of our encoder. This section provides a brief overview of each phase; details are available in the appendix.

Pre-Training. Supervised pre-training has been shown to improve the generalization performance of few-shot learning methods in the closed-world setting [3, 26]. We adapt this methodology to pre-train an encoder and learn Gaussian class embeddings that allow for efficient initialization of PCOC in the lifelong and incremental settings. We assume a uniform distribution over classes and directly learn a set of Gaussian embeddings using a large-scale training set in a standard supervised-learning setting. The mean and covariance of each class embedding are learnable parameters, however, the covariance is constrained as isotropic. Inference is still performed via Bayesian GDA. The class embeddings are regularized by a Gaussian prior on the mean and an inverse Wishart prior on the covariance.

Meta-Learning. In the meta-learning phase we assume a small support set consisting of N_{train} classes. Starting with no instantiated classes, we compute the posterior embedding for each class based on the support set. Then, we sample a query set of $N_{\text{test}} > N_{\text{train}}$ classes, where all classes not included in the support set are given the same "novel class" label. The loss is computed using this query set, and backpropagated through the conditioning to train the embedding statistics and the encoder.

Fine-Tuning. In the testing (or deployment) phase, we assume an additional support set is provided. We use this support set to fine tune the encoder to this novel class. We emphasize that this is a simple approach to combining methods from continual learning with our framework, and further investigation to combining meta-learning and continual learning methods is necessary. We fine-tune strictly the last layer of the encoder network, as other fine-tuning appoaches were unstable in our experiments. While the set of Gaussian embeddings is closed under changes to the linear last layer of the network, we restrict our embeddings to isotropic Gaussians. Thus, the last layer fine-tuning may be seen as re-scaling encoder features to yield class embeddings better captured by the isotropic distributions.

4 Experiments

We evaluate our proposed method using the MiniImageNet [22] and TieredImageNet [24] datasets. A Conv-4 network architecture [28] is used for MiniImageNet experiments and a larger scale Resnet18 [10] network architecture is used for the TieredImageNet experiments. Please refer to

Table 2: TieredImageNet - Open-World Few-Shot Lifelong Results

Method	Pre-Training	Acc. (%)	Support Acc. (%)	Inc. Acc. (%)	AUROC (%)
NCM	Sup-FC	39.76	43.44	32.28	50.64
NCM	Sup-E	50.97	57.36	38.19	49.22
ProtoNet	Sup-FC	46.09	53.47	33.50	48.47
ProtoNet	Sup-E	51.19	57.34	38.89	49.70
L-PCOC (Ours)	Sup-FC	49.61	55.50	37.82	71.83
L-PCOC (Ours)	Sup-E	**51.55**	**57.64**	**39.36**	**72.44**

Table 3: MiniImageNet - Open-World Few-Shot Incremental Results

Method	Pre-Training	Acc. (%)	Support Acc. (%)	Inc. Acc. (%)	AUROC (%)
NCM	Sup-E	**32.66**	**33.99**	15.60	57.16
ProtoNet	Sup-E	30.65	32.36	10.18	60.25
I-PCOC (Ours)	Sup-E	30.85	31.78	**18.84**	**84.75**

Section A.3 of the Appendix for additional implementation detail and the task sampling procedure for the lifelong and incremental settings.

Model performance is primarily measured via overall classification accuracy and novel class detection AUROC. We further decompose overall accuracy into support classification accuracy and incremental classification accuracy which provide metrics to understand the ability of the model to adapt quickly to novel classes. While AUROC is a less common metric relative to accuracy, we emphasize that it is of critical importance, as detection of novel classes often has important ramifications for ensuring safe operation.

Baselines. The performance of our method is compared with respect to Nearest Class Mean (NCM) and ProtoNet baselines. NCM has shown strong baseline performance in the closed-world few-shot setting [29, 3] where it performs top-1 nearest-neighbors classification in feature space via Euclidean distance. We adapt the NCM baseline for use in the open-world setting by thresholding the classification with a tunable minimum distance. The ProtoNet baseline follows the implementation proposed by [25]. It is similarly adapted for novel class detection via a tuned threshold. At test time, the class means are updated after each observed label. In the incremental setting, the train set is encoded to generate class means for each train class.

Results. We present the lifelong results on the MiniImageNet and TieredImageNet datasets in Tables 1 and 2. We observe a strict improvement in classification performance when initializing the encoder weights, for all three methods, with our supervised embedding pre-training scheme. An important feature of the L-PCOC algorithm is that it allows for end-to-end training in the lifelong-learning setting. Unlike NCM and ProtoNet, L-PCOC does not depend on non-differentiable thresholding to detect novel classes. The model can therefore learn to calibrate given the significant imbalance between novel class and in-distribution data. L-PCOC significantly outperforms the baselines when detecting novel classes on both datasets (as measured by AUROC).

We present the incremental results in Table 3. We observe that there is a significant AUROC increase across all algorithms, when compared to the lifelong setting. There is also a larger gap in classification accuracy between the support and incremental classes than the lifelong setting. This can be attributed to the fact that the models have access to a large-scale dataset for the support classes.

From our experiments, we observe that the embedding-based pre-training substantially improves classification accuracy and L-PCOC/I-PCOC results in substantially improved AUROC without degrading classification accuracy. Thus, the combination of embedding-based pre-training and L-PCOC/I-PCOC results in a robust model with balanced performance across all metrics necessary for deployment in the lifelong and incremental learning settings.

5 Conclusions

In this work, we motivate the need to reformulate learning systems from a closed-world setting to an open-world, lifelong and few-shot setting. We present a framework which combines a Chinese restaurant process class prior, a Bayesian non-parametric class prior and a supervised embedding pre-training scheme which can be flexibly applied to both the incremental and lifelong settings. The framework outperforms baselines on miniImageNet and TieredImageNet and demonstrates significant capability of detecting and quickly learning novel classes given few labels.

References

[1] Kelsey R Allen, Evan Shelhamer, Hanul Shin, and Joshua B Tenenbaum. Infinite mixture prototypes for few-shot learning. *International Conference on Machine Learning (ICML)*, 2019.

[2] Luca Bertinetto, João F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv:1805.08136*, 2018.

[3] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv:2003.04390*, 2020.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. *Artificial Intelligence and Statistics (AISTATS)*, 2020.

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

[7] Samuel J Gershman and David M Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 2012.

[8] James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. *Neural Information Processing Systems (NeurIPS)*, 2020.

[9] James Harrison, Apoorva Sharma, and Marco Pavone. Meta-learning priors for efficient online Bayesian regression. *Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. *International Conference on Artificial Neural Networks*, 2001.

[12] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv:2004.05439*, 2020.

[13] Ghassen Jerfel, Erin Grant, Tom Griffiths, and Katherine A Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. *Neural Information Processing Systems (NeurIPS)*, 2019.

[14] Michael Jordan and Yee Whye Teh. A gentle introduction to the dirichlet process, the beta process and bayesian nonparametrics, 2015.

[15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017.

[16] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Martin Mundt, Yong Won Hong, Iuliia Pliushch, and Visvanathan Ramesh. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *arXiv:2009.01797*, 2020.

[18] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[19] Anusha Nagabandi, Chelsea Finn, and Sergey Levine. Deep online learning via meta-learning: Continual adaptation for model-based RL. *arXiv:1812.07671*, 2019.

[20] Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. *Encyclopedia of machine learning*, 2010.

[21] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Neural Information Processing Systems (NeurIPS)*, 2019.

[22] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *International Conference on Learning Representations (ICLR)*, 2017.

[23] Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Incremental few-shot meta-learning via indirect discriminant alignment. *European Conference on Computer Vision (ECCV)*, 2020.

[24] 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. *International Conference on Learning Representations (ICLR)*, 2018.

[25] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Neural Information Processing Systems (NeurIPS)*, 2017.

[26] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *International Conference on Learning Representations (ICLR)*, 2019.

[27] Joaquin Vanschoren. Meta-learning: A survey. *arXiv:1810.03548*, 2018.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Neural Information Processing Systems (NeurIPS)*, 2016.

[29] Matthew Wallingford, Aditya Kusupati, Keivan Alizadeh-Vahid, Aaron Walsman, Aniruddha Kembhavi, and Ali Farhadi. In the wild: From ml models to pragmatic ml systems. *arXiv:2007.02519*, 2020.