Open-Set Incremental Learning via Bayesian Prototypical Embeddings

Anonymous Author(s)
Affiliation
Address
email

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, incremental, few-shot setting in which agents continuously learn new labels 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. In particular, we investigate both the lifelong setting—in which an entirely new set of classes exists at evaluation time—as well as the incremental setting, in which new classes are added to a set of base classes available at training time. 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 show strong performance compared to baseline methods.

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 [20]. 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, incremental and few-shot* 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 incremental learning in this open setting. Our approach combines ideas from Bayesian non-parametrics [16] with a Bayesian formulation of 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 combine a Chinese restaurant process class prior—a prior on an unbounded number of classes [9]—with a Bayesian embedding-based meta-learning algorithm [10]. We combine this formulation with an embedding-based pre-training phase, enabling both lifelong learning (in which all test classes are novel) and incremental learning (in which some classes are shared in train and test). We present these two open-world problem statements and motivating examples for each, and show how minor variations of our approach can handle both settings.

Contributions. There are three core contributions in this paper.

- We introduce an embedding-based pre-training phase, linking the pre-training to the meta-training phase and providing a conceptually simple algorithm for the incremental setting.

Submitted to the 4th Workshop on Meta-Learning at NeurIPS 2020, Vancouver, Canada. Do not distribute.

- We introduce a Bayesian few-shot learning scheme based on Gaussian embeddings [30]. 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, incremental learning.
- Finally, we evaluate our system on both the lifelong and incremental setting with the mini-ImageNet dataset [25], showing strong performance for our proposed framework.

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 *incremental* and/or *lifelong* setting. Because the terminology in continuous, lifelong, and incremental learning is often ambiguous and contradictory [19], we highlight precisely what we mean with these descriptors. In particular, our model must be

- **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.
- **Incremental** or **lifelong** learning: A problem is a *lifelong learning* problem if, at deployment, the agent starts with no knowledge of the space of possible labels and must learn all class labels. The *incremental learning* setting, on the other hand, implies that the set of labels observed during training are also observed during deployment, in addition to novel classes. We consider both settings in this paper, as they are strongly related.

As previously mentioned, we consider both the *open-world, few-shot, incremental* and the *open-world, few-shot, lifelong* learning settings in this paper, and refer to these settings as *incremental* and *lifelong*, respectively.

We now motivate each of these settings with an application example. Consider an autonomous vehicle, driving in a city. This vehicle was trained to recognize classes such as vehicles, cyclists, and pedestrians. However, the set of possible training classes does not cover the full universe of possible classes observed during deployment. As such, it is desirable that the vehicle is able to recognize entities outside of its training set, so that they may be labeled to improve the system. This example corresponds to the *incremental learning* setting.

In contrast to the previous example, consider a setting in which labels are subjective to a user. In particular, consider an automatic email tagging service. In this service, users are able to provide labels for emails, with which the system learns to automatically classify emails. This setting corresponds to the *lifelong learning* setting, as different users may have entirely disjoint sets of labels, and detection of inputs outside of the provided set of available labels is critical.

3 Background

Few-shot learning has seen increasing attention in recent years due to the difficulty of learning from a small amount of examples with deep neural network models [25, 30]. Meta-learning—or "learning-to-learn"—has been shown to be an effective approach to the few-shot learning problem [8, 34, 30]. By training on many different few-shot learning problems, an inner adaptation process is learned, capable of rapid adaptation given few examples. This primarily takes the form of learning an initialization or prior for a learning algorithm, an update rule to incorporate data, or both.

Although there are many taxonomies of meta-learning algorithms [13, 33], three common overarching categories have emerged, recurrence-based methods, optimization-based methods, and metric based-methods. Recurrence-based methods largely focus on a black-box update procedure; the inner learning algorithm typically takes the form of a recurrent neural network [12]. The black-box descriptor highlights the fact that the inner learning algorithm does not leverage any inductive biases from e.g. optimization. Optimization-based meta-learning aims to explicitly use an optimization procedure in the inner learning algorithm, and typically outperforms black-box methods as a result. These approaches rely on back-propagating through an optimization problem, yielding a bi-level optimization problem. Common approaches include back-propagation through the sequence of gra-

dient updates for a set of parameters, as in e.g. MAML [8], or back-propagation through the fixed point of a convex optimization algorithm [11, 4, 24, 17]. Finally, metric-based meta-learners rely on the inductive bias of metric learning, in which nearby samples in an embedding space are likely to be of the same class. Generally, these methods aim to learn an embedding space and/or a metric in this space, such that the embeddings of inputs of the same class are close to each other, and different classes are separable based on the metric [30].

In this work, we build on metric-based meta-learners due to their flexibility. Typically, a Euclidean [30] or cosine [34] distance is used between embeddings. We focus our approach on the line of work originating from prototypical networks [30], in which a set of support data is embedded and then aggregated as class prototypes, via

$$\boldsymbol{c}_k = \frac{1}{|S_k|} \sum_{(\boldsymbol{x}_i, y_i) \in S_k} \phi(\boldsymbol{x}_k) \tag{1}$$

where $k = 1, \ldots, K$ indexes the (closed) classes, S_k is the set of input/label pairs corresponding to class k, and $\phi(\cdot)$ is the encoder neural network. Then, the predicted density of a new test point $\boldsymbol{x}^*$ is

$$p(y^* = k \mid \boldsymbol{x}^*) = \frac{\exp(-\|\phi(\boldsymbol{x}^*) - \boldsymbol{c}_k\|_2^2)}{\sum_{k'=1}^{K} \exp(-\|\phi(\boldsymbol{x}^*) - \boldsymbol{c}_{k'}\|_2^2)}. \tag{2}$$

This predictive likelihood may be used to train the neural network features. The inner learning procedure corresponds to linear discriminant analysis [20] with an isotropic covariance. This interpretation, in which in the inner learning algorithm is a Gaussian discriminant analysis learning algorithm and the outer loop learns neural network features, has spawned numerous extensions. In particular, [28] extend the model with both semi-supervised learning and learned covariances, and [2] extend the approach toward a limiting clustering framework.

Of particular interest to the work in this paper is PCOC [10], which performs Bayesian Gaussian discriminant analysis in the embedding space. PCOC fixes a Categorical-Gaussian generative model in the embedding space of the form

$$y \sim \text{Cat}(p_1, \ldots, p_K), \qquad p_1, \ldots, p_K \sim \text{Dir}(\boldsymbol{\alpha}_0) \tag{3}$$

$$\boldsymbol{z} \mid y = k \sim \mathcal{N}(\bar{\boldsymbol{z}}_k, \Sigma_k), \quad \bar{\boldsymbol{z}}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \tag{4}$$

where $\boldsymbol{z} = \phi(\boldsymbol{x})$. In this model, each class k has a normal distribution in embedding space with mean $\bar{\boldsymbol{z}}_k$ and covariance Σ_k. Moreover, the embedding means are assumed normally distributed with mean $\boldsymbol{\mu}_0$ and covariance Σ_0. While [10] use the prior over embedding means to enable changepoint detection, we will use this formulation to enable a general open-set learning framework. Whereas standard prototypical networks implicitly assume a uniform prior over classes, PCOC uses a categorical-Dirichlet prior on the labels, where $\boldsymbol{\alpha}_0$ is a length k vector of parameters governing how tightly the Dirichlet prior is concentrated. In the limit of $\alpha \to \infty$ for each $\alpha \in \boldsymbol{\alpha}_0$, the uniform prior of prototypical networks is recovered.

This Categorical-Gaussian model is notable for having analytically computable posteriors, thus enabling simple differentiation through the posterior. The update equations for posterior parameters given data $(\boldsymbol{x}^*, y^*)$ take the form

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \mathbf{1}_k, \quad \boldsymbol{q}_k' = \boldsymbol{q}_k + \Sigma_k^{-1}\phi(\boldsymbol{x}^*), \quad \Lambda_k' = \Lambda_k + \Sigma_k^{-1} \tag{5}$$

for $k = y^*$, and where $\mathbf{1}_k$ denotes a one-hot vector at entry k. Here, $\bar{\boldsymbol{z}}_k = \Lambda_k^{-1}\boldsymbol{q}_k$, $\Lambda_k, \boldsymbol{q}_k$ and $\boldsymbol{\alpha}$ are initialized as $\Sigma_0^{-1}, \Sigma_0^{-1}\boldsymbol{\mu}_0$, and $\boldsymbol{\alpha}_0$ respectively.

4 Approach

Our approach relies on two phases for both the lifelong and the incremental setting. The first phase corresponds to pre-training, which has been shown to substantially improve performance of meta-learners [7, 32]. The second phase consists of a meta-training phase, either in the lifelong setting or the incremental setting.

Training. Supervised pre-training has been shown to improve the generalization performance of few-shot learning methods in the closed-world setting [7, 32]. 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 discard the PCOC Dirichlet prior (corresponding to a tightly peaked prior on an even class balance) 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 is are learnable parameters, however, we constrain the covariance to be isotropic. Inference is still performed via Bayesian Gaussian discriminant analysis. In the incremental setting, where class labels are shared across training and evaluation data, the class embeddings for the base classes, and the encoder, may be used to directly initialize priors corresponding to base class in the meta-learning phase. In the lifelong setting, where training and evaluation classes are disjoint, we discard the class embeddings but utilize the encoder initialization.

Incremental and Lifelong PCOC. Our approach extends the PCOC model with a Bayesian non-parametric class prior, as well as a supervised pre-training phase. The pretraining phase is described later in this section. We replace the Dirichlet prior of the PCOC model, which assumes a fixed number of classes, with a Dirichlet process that allows for an infinite number of classes. In particular, we leverage the Chinese restaurant process which has been effectively used in other works [21, 2, 15]. For a more detailed description than the brief outline presented here, we refer the reader to [9, 22, 16].

The standard CRP has two parameters, α and β. Assume we have observed n data points belonging to classes $k = 1, \ldots, K$. Let n_k denote the number of data points in class k. Then, the prior probability of a new data point belonging to class $K + 1$ is

$$p(\boldsymbol{y}_{n+1} = K + 1) = \frac{\beta + \alpha K}{n + \beta} \tag{6}$$

and the probability of belonging to class j is

$$p(\boldsymbol{y}_{n+1} = k) = \frac{n_k - \alpha}{n + \beta}. \tag{7}$$

where $\alpha \in [0, 1]$ and $\beta > -\alpha$ to ensure the distribution is a valid probalitity measure. In this paper, we fix $\alpha = 0.5$ and treat this as a hyperparameter.

As discussed previously, both the lifelong and incremental setting rely on embedding-based pretraining, a shared prior over embeddings, and the CRP prior. In the lifelong setting, no class embeddings are retained from the pre-training phase. Thus, our approach provides a conceptually simple unification of the two problem settings, as well as a simple, modular framework capable of handling both.

Meta-Learning. In the meta-learning phase we assume access to a training set, $\mathcal{D}_{tr}$, consisting of M classes. Meta-training tasks, $\mathcal{T}_k \sim p(\mathcal{T})$, are sampled sequentially from $\mathcal{D}_{tr}$ with probability $p(\mathcal{T})$, where k indexes the k^{th} meta-training step. Each task consists of a support set, $\mathcal{D}_s$, and a query set, $\mathcal{D}_q$, both of which are structured as a time-series of data where the model receives a datapoint $\boldsymbol{x}$, predicts associated class probabilities, and then receives the true class label. A unified procedure is used across both settings to sample training tasks. To construct $\mathcal{D}_s$, we first sample a set of $S < M$ support classes from $\mathcal{D}_{tr}$. Then, we sample a small number of datapoints from each support class. We do not assume balance across the classes in the support set. Finally, to build $\mathcal{D}_q$, we first sample data at random from the remaining $M - S$ classes. This data is bucketed as a single class and is assigned a shared $S + 1$ label. This class bucket is referred to as the incremental class. The incremental data is combined with a balanced set of data sampled from each of the support classes to construct $\mathcal{D}_q$. The quantity of incremental data is considered to be a hyperparameter.

Open-World Few-Shot Lifelong Evaluation. In the lifelong setting, meta-test tasks are sampled in the same fashion as the meta-training tasks. However, tasks are sampled from a meta-test set, $\mathcal{D}_{ts}^{L}$, consisting of N classes which are disjoint from the classes in $\mathcal{D}_{tr}$. During evaluation, in contrast to training time, data is removed from the $S + 1$ incremental class bucket once it has has been observed by the model. The observed datapoint is removed from the bucket along with all other

data belonging to its underlying semantic class. The data is re-labeled uniquely and the model must classify subsequently observed data according to the new label.

Open-World Few-Shot Incremental Evaluation. In the incremental setting, the model is evaluated via the incremental test set, $\mathcal{D}_{ts}^{I}$, which consists of data from all M classes in $\mathcal{D}_{tr}$ and a disjoint set of I incremental classes. We will refer to the M training classes as base classes. No support set is sampled during evaluation in this setting, learning of the base-classes is assumed to have been completed in the training phase. Test tasks are constructed by sampling a balanced set of data from each base-class and a random set of data from the incremental classes. As in the lifelong setting, the data belonging to an observed incremental class is removed from the bucket and uniquely re-labeled.

5 Related Work

Continual, incremental, and lifelong learning are related problems, often with inconsistent and overlapping definitions. Continual learning refers to learning problems in time-varying data streams [1, 36, 19]. Typically, the task—in the form of the input data—changes in discrete steps, and this change may be observed or unobserved (see [37] for a helpful taxonomy). Continual learning has seen considerable work in recent years, as a consequence of the large data requirements associated with training deep neural network models. Whereas continual learning typically refers to a shift in some component of the learning problem, incremental learning corresponds to an expansion of the possible inputs [5, 26]. In particular, in the classification setting, incremental learning methods typically assume the addition of a new class, and must learn to correctly classify inputs corresponding to this class. Finally, lifelong learning can be seen as an umbrella term covering continual and incremental learning, but has as a central aim the ability for a learner to continuous integrate new information without a discrete re-training phase [23, 31].

The difficulty of training deep neural network models with small amounts of data has also led to a resurgence of interest in meta-learning for few-shot learning [8, 34]. These meta-learning techniques have been incorporated into continual, lifelong, and incremental learning settings to improve performance, especially in settings where little data is available for a new task or new class [14, 10, 29, 3]. Of particular interest are recent works investigating meta-learning for incremental learning or open-set meta-learning. [18] investigate open-set meta-learning without an incremental component; they aim to detect out-of-distribution classes and classify them as such, without learning to predict those classes in the future. Their approach is based on thresholding on prototypical embeddings. Because they lack the prior we introduce in this work, there is no clear method to instantiate new classes. A similar approach was developed by [2], who base their approach on a discrete clustering algorithm that can be seen as the limiting case of the CRP prior. Again, this method is capable of open-set detection, but they do not explicitly consider the lifelong or incremental setting. [27] consider a few-shot approach to few-shot incremental learning in which a base classifier is paired with a standard few-shot learner. To detect novel classes, the authors introduce an attention-based mechanism, trained via meta-learning. A unifying setting for continual, incremental, and few-shot learning was presented in [35]. They presented performance on their "in the wild" learning setting for several optimization and embedding-based few-shot learning schemes, with and without pre-training. This work showed the importance of both non-meta-learning pre-training, and also showed simple embedding-based methods are highly competitive with other investigated approaches.

6 Experiments

We evaluate our proposed method using the MiniImageNet dataset and standard class splits proposed by [25]. MiniImageNet contains 100 classes, each with 600 images, split into 64 training, 12 validation and 24 test classes. All images are resized to a uniform 84x84 dimension. We utilize the Conv-4 network architecture proposed by [34]. Please refer to Section A.1 of the Appendix for additional implementation detail and the task sampling procedure for the lifelong and incremental settings.

We compare the performance of our method with respect to a Nearest Class Mean (NCM) baseline in both the lifelong and incremental settings. NCM has shown strong baseline performance in the closed-world few-shot setting [35, 7], where it performs top-1 nearest-neighbors classification in feature space using a Euclidean distance. Query feature vectors are classified with respect to a

Table 1: Open-World Few-Shot Lifelong Results

Method	Pre-Training	Acc. (%)	Support Acc. (%)	Inc. Acc. (%)	AUROC (%)
NCM	Sup-FC	32.62	32.23	35.26	64.44
NCM	Sup-E	36.87	36.58	38.80	64.78
NCM	SimCLR	29.16	29.18	29.03	62.11
L-PCOC	-	38.67	38.46	40.23	64.69
L-PCOC	Sup-FC	**40.28**	**40.52**	38.57	64.74
L-PCOC	Sup-E	38.88	39.48	34.833	63.73
L-PCOC	SimCLR	39.12	38.9	**40.67**	**65.01**

Table 2: Open-World Few-Shot Incremental Results

Method	Pre-training	Acc. (%)	Support Acc. (%)	Inc. Acc. (%)	AUROC (%)
NCM	Sup-FC	27.31	27.61	8.07	51.47
NCM	Sup-E	20.73	20.83	**14.00**	54.60
I-PCOC	-	24.04	24.22	12.26	**55.80**
I-PCOC	Sup-E	**27.66**	**27.96**	7.87	51.92

set of pre-computed class means. Feature vectors are obtained via a learned encoder and class means are computed as the mean of all feature vectors belonging to a given support class. We adapt the NCM baseline for use in the open-world setting by thresholding the top-1 classification with a tunable minimum distance. In addition to the proposed supervised embedding pre-training (Sup-E), we investigate supervised pre-training with a fully connected layer (Sup-FC) following the methodology of [7] and self-supervised pre-training (SimCLR) following the methodology of [6].

Performance is measured via overall classification accuracy and the area under the receiver operating characteristic (AUROC) of novel class detection. The overall classification accuracy is further decomposed into support classification accuracy and incremental classification accuracy. Support classification accuracy captures the models' ability to classify data from classes observed prior to evaluation. Incremental classification provides a metric to determine the ability of the model to classify with respect to novel classes observed during evaluation. Further experimental details are available in the appendix.

Results. As described in Table 1, L-PCOC outperforms the baseline across all metrics in the lifelong setting. L-PCOC with an encoder initialized by the Sup-FC pre-training scheme demonstrated the best performance with respect to support classes and overall accuracy. However, L-PCOC initalized via the self-supervised SimCLR pre-training showed support classes support class accuracy while outperforming all other methods in detecting and classifying novel classes.

For the incremental setting, Table 2 shows that I-PCOC initialized via the supervised embedding pre-training outperforms all other methods when classifying base classes. I-PCOC trained from scratch demonstrated a comparatively strong ability to classify novel classes, however, this was challenging for all methods evaluated. The large accuracy gap of the models between the lifelong and incremental settings is expected because of the significant increase in task classification-way when evaluating in the incremental setting.

7 Conclusion

In this work, we motivate the need to reformulate learning systems from a closed-world setting to an open-world, incremental and few-shot setting. We futher propose a decomposition of the open-world into an incremental and lifelong learning 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 the MiniImageNet dataset and demonstrates significant capability of detecting and then quickly learning to identity novel classes given few labels.

References

[1] Rahaf Aljundi. Continual learning in neural networks. *arXiv:1910.02718*, 2019.

[2] 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.

[3] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. *arXiv:2002.09571*, 2020.

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

[5] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. *European Conference on Computer Vision (ECCV)*, 2018.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020.

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

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

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

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

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

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

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

[14] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Neural Information Processing Systems (NeurIPS)*, 2019.

[15] 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.

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

[17] 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.

[18] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] 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.

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

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

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

[23] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

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

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

[26] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. Incremental few-shot learning with attention attractor networks. *Neural Information Processing Systems (NeurIPS)*, 2019.

[28] 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.

[29] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *International Conference on Learning Representations (ICLR)*, 2019.

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

[31] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Neural Information Processing Systems (NeurIPS)*, 1996.

[32] 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.

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

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

[35] 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.

[36] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *International Conference on Machine Learning (ICML)*, 2017.

[37] Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task agnostic continual learning using online variational bayes. *arXiv:1803.10123*, 2018.