Wandering Within a World: Online Contextualized Few-Shot Learning

Anonymous Authors[1]

Abstract

We aim to bridge the gap between typical human and machine-learning environments by extending the standard framework of few-shot learning to an online, continual setting. In this setting, episodes do not have separate training and testing phases, and instead models are evaluated online while learning novel classes. As in the real world, where the presence of spatiotemporal context helps us retrieve learned skills in the past, our online few-shot learning setting also features an underlying context that changes throughout time. Object classes are correlated within a context and inferring the correct context can lead to better performance. Building upon this setting, we propose a new few-shot learning dataset based on large scale indoor imagery that mimics the visual experience of an agent wandering within a world. Furthermore, we convert popular few-shot learning approaches into online versions and we also propose a new *contextual prototypical memory* model that can make use of spatiotemporal contextual information from the recent past.

1. Introduction

In machine learning, many paradigms exist for training and evaluating models: standard train-then-evaluate, few-shot learning, incremental learning, continual learning, and so forth. None of these paradigms well approximates the naturalistic conditions that humans and artificial agents encounter as they wander within a physical environment. Consider, for example, learning and remembering peoples' names in the course of daily life. We tend to see people in a given environment—work, home, gym, etc. We tend to repeatedly revisit those environments, with different environment base rates, nonuniform environment transition probabilities, and nonuniform base rates of encountering a given person in a given environment. We need to recognize when we do not know a person, and we need to learn to recognize them the next time we encounter them. We are not always provided with a name, but we can learn in a semi-supervised manner. And every training trial is itself an evaluation trial as we repeatedly use existing knowledge and acquire new knowledge. In this article, we propose a novel paradigm, *online contextualized few-shot learning*, that approximates these naturalistic conditions, and we develop deep-learning architectures well suited for this paradigm.

In traditional few-shot learning (FSL) (Lake et al., 2015; Vinyals et al., 2016), training is episodic. Within an isolated episode, a set of new classes is introduced with a limited number of labeled examples per class—the *support* set—followed by evaluation on an unlabeled *query* set. While this setup has inspired the development of a multitude of meta-learning algorithms which can be trained to rapidly learn novel classes with a few labeled examples, the algorithms are focused solely on the few classes introduced in the current episode; the classes learned are not carried over to future episodes. Although incremental learning and continual learning methods (Rebuffi et al., 2017; Hou et al., 2019) address the case where classes are carried over, the episodic construction of these frameworks seems artificial: in our daily lives, we do not learn new objects by grouping them with five other new objects, process them together, and then move on.

To break the rigid, artificial structure of continual and few-shot learning, we propose a new continual few-shot learning setting where environments are revisited and the total number of novel object classes increases over time. Crucially, model evaluation happens on each trial, very much like the setup in online learning. When encountering a new class, the learning algorithm is expected to indicate that the class is "new," and it is then expected to recognize subsequent instances of the class once a label has been provided.

When learning continually in such a dynamic environment, contextual information can guide learning and remembering. Any structured sequence provides *temporal context*: the instances encountered recently are predictive of instances to be encountered next. In natural environments, *spatial context*—information in the current input weakly correlated with the occurrence of a particular class—can be beneficial

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

for retrieval as well. For example, we tend to see our boss in an office setting, not in a bedroom setting. Human memory retrieval benefits from both spatial and temporal context (Howard, 2017; Kahana, 2012). In our online few-shot learning setting, we provide spatial context in the presentation of each instance and temporal structure to sequences, enabling an agent to learn from both spatial and temporal context. Besides developing and experimenting on a toy benchmark using handwritten characters (Lake et al., 2015), we also propose a new large-scale benchmark for online contextualized few-shot learning derived from indoor panoramic imagery (Chang et al., 2017). In the toy benchmark, temporal context can be defined by the co-occurrence of character classes. In the indoor environment, the context—temporal and spatial—is a natural by-product as the agent wandering in between different rooms.

We propose a model that can exploit contextual information, called *contextual prototypical memory* (*CPM*), which incorporates an RNN to encode contextual information and a separate prototype memory to remember previously learned classes (see Figure 4). This model obtains significant gains on few-shot classification performance compared to models that do not retain a memory of the recent past. We compare to classic few-shot algorithms (Vinyals et al., 2016; Allen et al., 2019; Snell et al., 2017; Javed & White, 2019; Santoro et al., 2016) extended to an online setting, and CPM consistently achieves the best performance.

The main contributions of this paper are as follows. First, we define an *online contextualized few-shot learning (OC-FSL)* setting to mimic naturalistic human learning. Second, we build two datasets. The *RoamingOmniglot* dataset is based on handwritten characters from Omniglot (Lake et al., 2015) and the *RoamingRooms* dataset is our new few-shot learning dataset based on indoor imagery (Chang et al., 2017), which resembles the visual experience of a wandering agent. Third, we benchmark classic FSL methods and also explore our CPM model, which combines the strengths of RNNs for modeling temporal context and Prototypical Networks (Snell et al., 2017) for memory consolidation and rapid learning.

2. Related Work

In this section, we briefly review paradigms that have been used for few-shot learning (FSL) and continual learning (CL). Table 1 compares these paradigms based on various properties of the task; To denote properties that are incorporated in some preliminary form, we use ◓ to denote properties that are not fully implemented but have some preliminary form. For example, the class-incremental learning paradigm evaluates models after each task is completed, which is similar to our online evaluation in spirit but still not the same as the evaluation does not take place after each example. Our proposed *online contextual few-shot learning (OC-FSL)* spans the complete set of features of the other paradigms. We also review relevant models and their relationship to our CPM.

Table 1: Comparison of past FSL and CL paradigms vs. our online contextualized FSL (OC-FSL)

Tasks	Few Shot	Semi-sup. Supp. Set	Continual	Online Eval.	Predict New	Soft Context Switch
Incremental Learning (IL)	○	○	●	◓	○	○
Few-shot (FSL) (Vinyals et al., 2016)	●	○	○	○	○	○
Incremental FSL (Ren et al., 2019)	●	○	◓	○	○	○
Cls. Incremental FSL (Tao et al., 2020)	●	○	●	◓	○	○
Semi-supv. FSL (Ren et al., 2018)	●	●	○	○	●	○
Continual Meta-Learning w/o Tasks (Harrison et al., 2019)	●	○	●	○	○	◓
Online Mixture (Jerfel et al., 2019)	●	○	●	○	○	◓
Online Meta (Javed & White, 2019)	●	○	●	○	○	○
Continual FSL (Antoniou et al., 2020) (Concurrent Work)	●	○	●	○	○	○
OC-FSL (Ours)	●	●	●	●	●	●

Few-shot learning: FSL (Lake et al., 2015; Li et al., 2007; Koch et al., 2015; Vinyals et al., 2016) considers learning new tasks with few labeled examples. FSL models can be categorized as based on: metric learning (Vinyals et al., 2016; Snell et al., 2017), memory (Santoro et al., 2016), and gradient adaptation (Finn et al., 2017; Li et al., 2017). The model we propose, CPM, lies on the boundary between these approaches, as we use an RNN to model the temporal context but we also use metric-learning mechanisms and objectives to train.

Several previous efforts have aimed to extend few-shot learning to incorporate more natural constraints. One such example is semi-supervised FSL (Ren et al., 2018), where models learn not only from a few labeled examples but also from a pool of unlabeled examples. While traditional FSL only tests the learner on novel classes, *incremental FSL* (Gidaris & Komodakis, 2018; Ren et al., 2019) tests on novel classes together with a set of base classes. These approaches, however, have not explored how to iteratively add new classes.

In concurrent work, Antoniou et al. (2020) extend FSL to a continual setting based on image sequences, each of which is divided into stages with a fixed number of examples per class followed by a query set. Our paradigm focuses on more flexible and faster adaptation since the models are evaluated online, and context is a soft constraint instead of a hard separation of tasks. Moreover, new classes need to be identified as part of the sequence, crucial to any learner's incremental acquisition of knowledge.

Continual learning: Continual (or lifelong) learning is a parallel line of research that aims to handle a sequence of dynamic tasks (Kirkpatrick et al., 2017; Li & Hoiem, 2018; Lopez-Paz & Ranzato, 2017; Yoon et al., 2018). A key challenge here is catastrophic forgetting (McCloskey & Cohen, 1989; French, 1999), where the model "forgets"

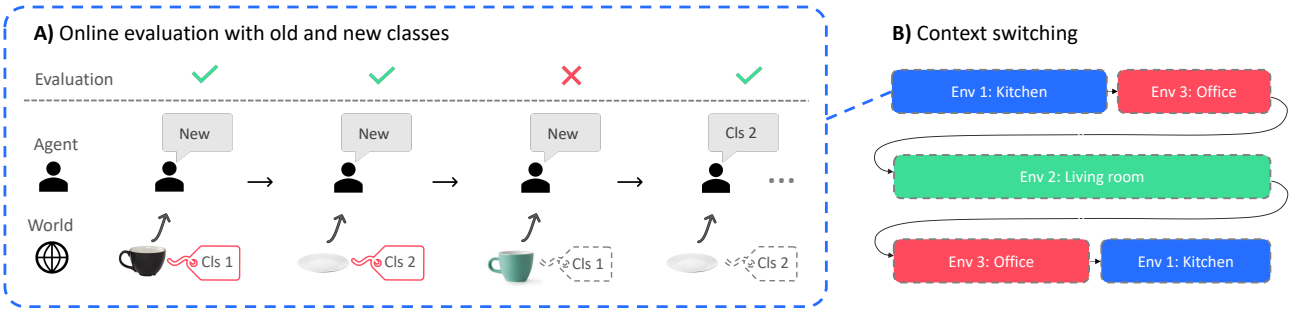

Figure 1: **Online contextualized few-shot learning. A)** Our setup is similar to online learning, where there is no separate testing phase; model training and evaluation happen *at the same time*. The input at each time step is an (image, class-label) pair. The number of classes grows *incrementally* and the agent is expected to answer "new" for items that have not yet been assigned labels. Sequences can be *semi-supervised*; here the label is not revealed for every input item (labeled/unlabeled shown by red solid/grey dotted boxes). The agent is evaluated on the correctness of all answers. The model obtains learning signals only on labeled instances, and is correct if it predicts the label of previously-seen classes, or 'new' for new ones. **B)** The overall sequence switches between different *learning environments*. While the environment ID is *hidden* from the agent, inferring the current environment can help solve the task.

a task that has been learned in the past. Incremental learning (Hou et al., 2019; Rebuffi et al., 2017; Kemker & Kanan, 2018) is a form of continual learning, where each task is an incremental batch of several new classes. This assumption that novel classes always come in batches seems unnatural.

Traditionally, continual learning is studied with tasks such as permuted MNIST (Lecun et al., 1998) or split-CIFAR (Krizhevsky, 2009). Recent datasets aim to consider more realistic continual learning, such as CORe50 (Lomonaco & Maltoni, 2017) and OpenLORIS (She et al., 2019). We summarize core features of these continual learning datasets in the Appendix. First, both datasets have relatively few object classes, which makes meta-learning approaches inapplicable; second, both datasets contain images of small objects with minimal occlusion and viewpoint changes; and third, OpenLORIS does not have the desired incremental class learning.

Online meta-learning: Some existing work builds on early approaches (Thrun, 1998; Schmidhuber, 1987) that tackle continual learning from a meta-learning perspective. Finn et al. (2019) propose storing all task data in a data buffer to perform inner loop gradient descent; by contrast, Javed & White (2019) propose to directly update the inner loop with the current input and instead learn a good representation that supports such online updates. In Jerfel et al. (2019), a hierarchical Bayesian mixture model is used to address the dynamic nature of continual learning. To evaluate the performance of online meta-learning methods, the papers above also created a few synthetic continual learning tasks which are less realistic than ours (see Table 1).

Connections to the human brain: Our CPM model consists of multiple memory systems, consistent with claims of cognitive neuroscientists of multiple memory systems in the brain. The complementary learning systems (CLS) theory (Mcclelland et al., 1995) suggests that the hippocampus stores the recent experience and supports fast recall of novel concepts, which is likely where few-shot learning takes place. When the same content has been retrieved several times, the neocortex changes synaptic weights slowly to learn a more robust representation. The hippocampal system can be further divided into an episodic memory for individual spatiotemporal events and a semantic memory for a concrete piece of knowledge (Cohen & Squire, 1980). Our proposed CPM contains parallels to these components. Long term statistical learning is captured in a CNN that produces a deep embedding. An RNN holds a type of working memory that can retain novel objects and spatiotemporal contexts. Lastly, the prototype memory represents the semantic memory, which consolidates multiple events into a single knowledge vector. Other deep learning researchers have proposed multiple memory systems for continual learning. In Parisi et al. (2018), the learning algorithm is heuristic and representations come from pretrained networks. In Kemker & Kanan (2018), a prototype memory is used for recalling recent examples and rehearsal from a generative model allows this knowledge to be integrated and distilled into a long-term memory.

3. Online Contextualized Few-Shot Learning

In this section, we introduce our new online contextualized few-shot learning (OC-FSL) setup in the form of a sequential decision problem, and then introduce our new benchmark datasets.

Continual few-shot classification as a sequential decision problem: In traditional few-shot learning, an episode is constructed by a support set S and a query set Q. A few-shot learner f is expected to predict the class of each

example in the query set $\mathbf{x}^Q$ based on the support set information: $\hat{y}^Q = f(\mathbf{x}^Q; (\mathbf{x}_1^S, y_1^S), \ldots, (\mathbf{x}_N^S, y_N^S))$. This setup is not a natural fit for continual learning, since it is unclear when to insert a query set into the sequence.

Inspired by the online learning literature, we can convert continual few-shot learning into a sequential decision problem, where every input example is also part of the evaluation: $\hat{y}_t = f(\mathbf{x}_t; (\mathbf{x}_1, \tilde{y}_1), \ldots, (\mathbf{x}_{t-1}, \tilde{y}_{t-1}))$, for $t = 1 \ldots T$, where $\tilde{y}$ here further allows that the sequence of inputs to be semi-supervised: $\tilde{y}$ equals y_t if labeled, or otherwise -1. The setup in Santoro et al. (2016) and Kaiser et al. (2017) is similar; they train RNNs using such a temporal sequence as input. However, their evaluation relies on another "query set" at the end of the sequence. We instead evaluate online while learning. Figure 1-A illustrates these features, using an example of an input sequence where an agent is learning about new objects in a kitchen. The model is rewarded when it correctly predicts a known class or when it indicates that the item has yet to be assigned a label.

Contextualized environments: Typical continual learning consists of a sequence of tasks, and models are trained sequentially for each task. This feature is also preserved in many incremental learning settings (Rebuffi et al., 2017). For instance, the split-CIFAR task divides CIFAR-100 into 10 learning environments, each with 10 classes, presented sequentially. In our formulation, the sequence returns to earlier environments (see Figure 1-B), which enables assessment of long-term durability of knowledge. Although the ground-truth environment identity is not provided, we structure the task such that the environment itself provides contextual cues which can constrain the correct class label. *Spatial* cues in the input help distinguish one environment from another. *Temporal* cues are implicit because the sequence tends to switch environments infrequently, allowing recent inputs to be beneficial in guiding the interpretation of the current input.

RoamingOmniglot: The Omniglot dataset (Lake et al., 2015) contains 1623 handwritten characters from 50 different alphabets. We split the alphabets into 31 for training, 5 for validation, and 13 for testing. We augment classes by 90 degree rotations to create 6492 classes in total. Each contextualized few-shot learning image sequence contains 150 images, drawn from a random sample of 5-10 alphabets, for a total of 50 classes per sequence. These classes are randomly assigned to 5 different environments; within an environment, the characters are distributed according to a Chinese restaurant process (Aldous, 1985) to mimic the imbalanced long-tail distribution of naturally occurring objects. We switch between environments using a Markov switching process; i.e., at each step there is a constant probability of switching to another environment. An example sequence is shown in Figure 2-A.

RoamingRooms: As none of the current few-shot learning datasets provides the natural online learning experience that we would like to study, we created our own dataset using simulated indoor environments. We formulate this as a few-shot instance learning problem, which could be a use case for a home robot: it needs to quickly recognize and differentiate novel object instances, and large viewpoint variations can make this task challenging (see examples in Figure 2-B). There are over 7,000 unique instance classes in the dataset, making it suitable to meta-learning approaches.

Our dataset is derived from the Matterport3D dataset (Chang et al., 2017), which has 90 indoor worlds captured using panoramic depth cameras. We split these into 60 worlds for training, 10 for validation and 20 for testing. We use MatterSim (Anderson et al., 2018) to load the simulated world and collect camera images and use HabitatSim (Savva et al., 2019) to simulate 3D mesh and align instance segmentation labels onto 2D image space. We created a random walking agent to collect the virtual visual experience. For each viewpoint in the random walk, we randomly sample one object from the image sensor and highlight it with the available instance segmentation, forming an input *frame*. Each viewpoint provides environmental context—the other objects present in the room with the highlighted object.

Figure 3-A shows the object instance distribution. We see strong temporal correlation, as 30% of the time the same instance appears in the next frame (Figure 3-B), but there is also a significant proportion of revisits. On average, there are three different viewpoints per 100-image sequence (Figure 3-C).

4. Contextual Prototypical Memory Networks

In the online contextualized few-shot learning setup, the few-shot learner can potentially improve by modeling the temporal context. Metric learning approaches (Vinyals et al., 2016; Allen et al., 2019; Snell et al., 2017) typically ignore temporal relations and directly compare the similarity between training and test samples. Gradient-based approaches (Finn et al., 2017; Javed & White, 2019), on the other hand, have the ability to adapt to new contexts, but they do not naturally handle new and unlabeled examples. We instead propose a simple yet effective approach that augments the classic Prototypical Network with a temporal contextual encoder using an RNN, shown in Figure 4. Next, we describe our approach in detail.

Prototype memory: We start describing our model with the prototype memory, which is an online version of the Prototypical Network (or *ProtoNet*) (Snell et al., 2017). ProtoNet can be viewed as a knowledge base memory, where

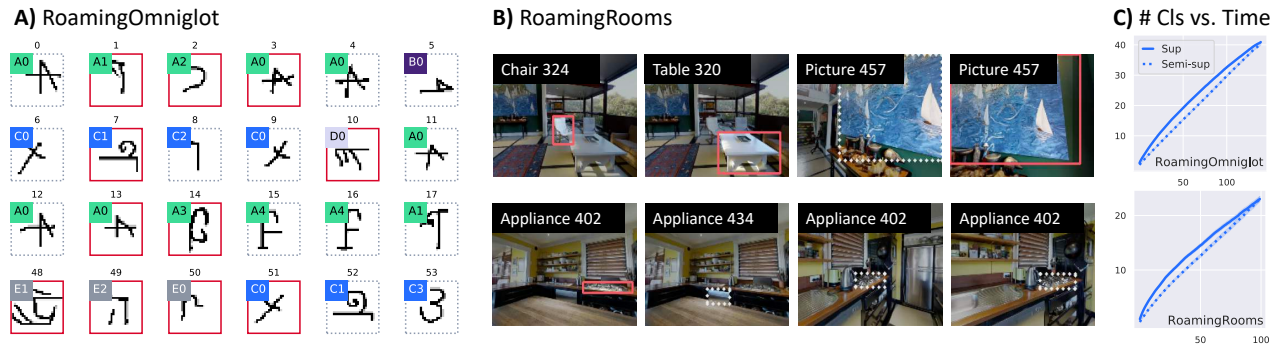

Figure 2: **Sample online contextualized few-shot learning sequences. A)** RoamingOmniglot. Red solid boxes denote labeled examples of Omniglot handwritten characters, and dotted boxes denote unlabeled ones. Environments are shown in colored labels in the top left corner. **B)** Image frame samples of a few-shot learning sequence in our RoamingRooms dataset collected from a random walking agent. The task here is to recognize novel instances in the home environment. **C)** The growth of total number of labeled classes in a sequence for RoamingOmniglot (top) and RoamingRooms (bottom).

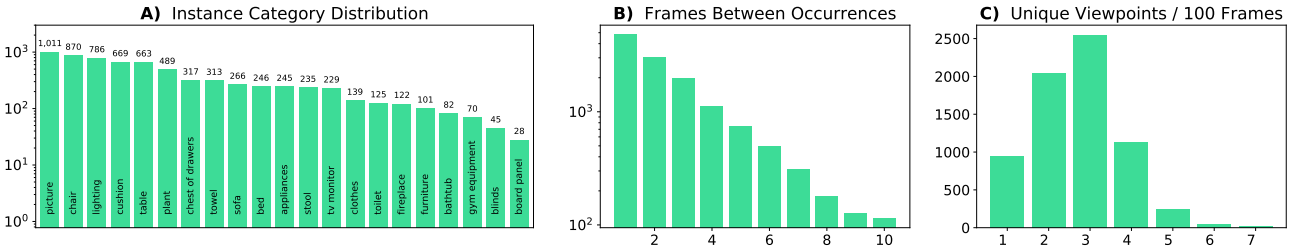

Figure 3: **Statistics for our RoamingRooms dataset.** Plots show a natural long tail distribution of viewpoints and instances grouped into categories. An average sequence has 3 different view points. Sequences are highly correlated in time but revisits are not uncommon.

each object class k is represented by a prototype vector $\mathbf{p}[k]$, computed as the mean vector of all the support instances of the class in a sequence. It can also be applied to our task of online few-shot learning naturally, with some modifications. Suppose that at time-step t we have already stored a few classes in the memory, each represented by their current prototype $\mathbf{p}_t[k]$, and we would like to query the memory using the current input feature $\mathbf{h}_t$. We model our prototype memory as

$$\hat{y}_{t,k} = \text{softmax}(-\|\mathbf{h}_t - \mathbf{p}_t[k]\|^2_{\mathbf{M}_t}), \quad (1)$$

where $\|\cdot\|^2_{\mathbf{M}_t}$ is the squared Mahalanobis distance for some to-be-specified, time-varying hyperparameter matrix $\mathbf{M}_t$. To predict whether an example is of a new class, we can use a separate *novelty* output $\hat{u}^r_t$ with sigmoid activation, similar to the approach introduced in (Ren et al., 2018), where β^r_t and γ^r_t are yet-to-be-specified thresholding hyperparameters (the superscript r stands for read):

$$\hat{u}^r_t = \text{sigmoid}((\min_k \|\mathbf{h}_t - \mathbf{p}_t[k]\|^2_{\mathbf{M}_t} - \beta^r_t)/\gamma^r_t). \quad (2)$$

Memory consolidation with online prototype averaging: Traditionally, ProtoNet uses the average representation of

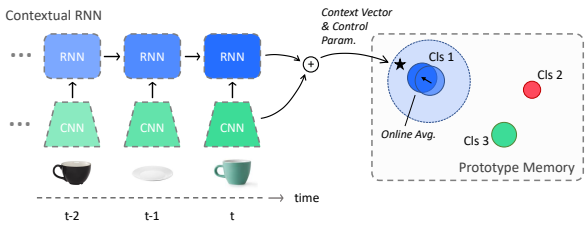

Figure 4: **Contextual prototypical memory model.** Temporal contextual features are extracted from an RNN. The prototype memory stores one vector per class and performs online averaging when learning new examples of the class. Examples falling outside the radii of all prototypes are classified as "new."

a class across all support examples. Here, we must be able to adapt the prototype memory incrementally at each step. Fortunately, we are able to recover the same prototypes as a regular ProtoNet computes offline by using online averaging. For each prototype k, we store a count scalar $c[k]_t$ to indicate the number of examples that have been added to this prototype up to time t. When the current example is unlabeled, y_t is encoded as -1, and the model's own prediction $\hat{y}^w_t$ will determine which prototype to update; in this case, the model must also determine a strength of belief,

$\hat{u}_t^w$, that the current unlabeled example should be treated as a new class. Given $\hat{u}_t^w$ and y_t, the model can then update a prototype:

$$\hat{u}_t^w = \text{sigmoid}((\min_k \|\mathbf{h}_t - \mathbf{p}_t[k]\|^2_{\mathbf{M}_t} - \beta_t^w)/\gamma_t^w),$$
$$\Delta[k]_t = \underbrace{\mathbb{1}[y_t = k]}_{\text{Supervised}} + \underbrace{\hat{y}_{t,k}(1 - \hat{u}_t^w)\mathbb{1}[y_t = -1]}_{\text{Unsupervised}},$$
$$c[k]_t = c[k]_{t-1} + \Delta[k]_t,$$
$$\mathbf{p}[k]_t = \frac{1}{c[k]_t}\left(\mathbf{p}[k]_{t-1}c[k]_{t-1} + \mathbf{h}_t\Delta[k]_t\right) \text{ if } \Delta[k]_t > 0.$$

As-yet-unspecified hyperparameters β_t^w and γ_t^w are required. (The superscript w in β_t^w and γ_t^w is for write.) These parameters for the online-updating novelty output $\hat{u}_t^w$ are distinct from β_t^r and γ_t^r in Equation 2. The intuition is that for "self-teaching" to work, the model potentially needs to be more conservative in creating new classes (avoiding corruption of prototypes) than in predicting an input as being a new class.

Contextual RNN: Instead of directly using the features from the CNN $\mathbf{h}_t^{\text{CNN}}$ as input features to the prototype memory, we would also like to use contextual information from the recent past. Above we introduced threshold hyperparameters β_t^r, γ_t^r, β_t^w, γ_t^w as well as the metric parameter $\mathbf{M}_t$. We let the contextual RNN output these additional control parameters, so that the unknown thresholds and metric function can adapt based on the information in the context. The RNN produces the context vector $\mathbf{h}_t^{\text{RNN}}$ and other control parameters conditioned on $\mathbf{h}_t^{\text{CNN}}$:

$$\left[\mathbf{z}_t, \mathbf{h}_t^{\text{RNN}}, \mathbf{m}_t,\ \beta_t^r,\ \gamma_t^r,\ \beta_t^w,\ \gamma_t^w\right] = \text{RNN}(\mathbf{h}_t^{\text{CNN}}; \mathbf{z}_{t-1}),$$

where $\mathbf{z}_t$ is the recurrent state of the RNN, and $\mathbf{m}_t$ is the diagonal vector of $\mathbf{M}_t$. The context, $\mathbf{h}_t^{\text{RNN}}$, serves as an additive bias on the state vector used for FSL: $\mathbf{h}_t = \mathbf{h}_t^{\text{CNN}} + \mathbf{h}_t^{\text{RNN}}$.

Loss function: The loss function is computed after an entire sequence ends and all network parameters are learned end-to-end. The loss is composed of two parts. The first is binary cross-entropy (BCE), for telling whether each example has been assigned a label or not, i.e., prediction of new classes. Second we use a multi-class cross-entropy for classifying among the known ones. We can write down the overall loss function as follows:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \lambda \underbrace{\left[-\mathbb{1}_{[y_t<0]}\log(\hat{u}_t^r) - \mathbb{1}_{[y_t\geq 0]}\log(1-\hat{u}_t^r)\right]}_{\text{Binary cross entropy on old vs. new}}$$
$$+ \sum_{k=1}^{K} \underbrace{-\mathbb{1}[y_t = k]\log(\hat{y}_{t,k})}_{\text{Cross entropy on old classes}}.$$

5. Experiments

In this section, we show experimental results for our online contextualized few-shot learning paradigm, using RoamingOmniglot and RoamingRooms (see Sec. 3) to evaluate our model, CPM, and other state-of-the-art methods. For Omniglot, we apply an 8×8 CutOut (Devries & Taylor, 2017) to each image to make the task more challenging.

Implementation details: For the RoamingOmniglot experiment we used the common 4-layer CNN for few-shot learning with 64 channels in each layer. For the RoamingRooms experiment we resize the input to 120×160 and we use the ResNet-12 architecture (Oreshkin et al., 2018) with {32,64,128,256} channels per block. To represent the feature of the input image with an attention mask, we concatenate the global average pooled feature with the attention ROI feature, resulting in a 512d feature vector. For the contextual RNN, in both experiments we used an LSTM (Hochreiter & Schmidhuber, 1997) with a 256d hidden state.

We use the Adam optimizer (Kingma & Ba, 2015) with initial learning rate 1e-3 for all experiments. For Omniglot we train the network for 40k steps with batch size of 32 with maximum sequence length 150 across 2 GPUs and learning rate decay by 0.1× at 20k and 30k steps. For Matterport 3D we train for 20k steps with batch size 8 with maximum sequence length 100 across 4 GPUs and learning rate decay by 0.1× at 8k and 16k steps. We use BCE coeffcient $\lambda = 1$ for all experiments. In semisupervise experiments, around 30% examples are labeled.

Evaluation metrics: In order to compute a single number that characterizes the learning ability over sequences, we propose to use *average precision* (AP) to combine the prediction on old and new classes. Concretely, all predictions are sorted by their old vs. new scores, and we compute AP using the area under the precision-recall curve. True positive is defined as the correct prediction of a multi-class classification among known classes. We also compute the "N-shot" accuracy; i.e., the average accuracy after seeing the label N times in the sequence. Note that these accuracy scores only reflect the performance on *known* class predictions. All numbers are reported with an average over 2,000 sequences and for N-shot accuracy standard error is also included.

Comparisons: To evaluate the merits of our proposed model, we implement classic few-shot learning and online meta-learning methods.

- **OML** (Javed & White, 2019): This is an online version of MAML (Finn et al., 2017). It performs one gradient descent step for each labled input image, and slow weights are learned via backpropagation through time.

Table 2: **RoamingOmniglot OC-FSL results.** Max 5 env., 150 images, 50 classes, with 8×8 box occlusion.

Method	Supervised			Semi-supervised		
	AP	1-shot Acc.	3-shot Acc.	AP	1-shot Acc.	3-shot Acc.
OML	70.97	63.32 ± 0.21	91.67 ± 0.15	54.27	71.64 ± 0.19	93.72 ± 0.27
LSTM	64.34	61.00 ± 0.22	81.85 ± 0.21	54.34	68.30 ± 0.20	76.38 ± 0.49
DNC	81.30	78.87 ± 0.19	91.01 ± 0.15	81.37	88.56 ± 0.12	93.81 ± 0.26
Online MatchingNet	88.69	84.82 ± 0.15	95.55 ± 0.11	84.39	88.77 ± 0.13	97.28 ± 0.17
Online IMP	90.15	85.74 ± 0.15	96.66 ± 0.09	81.62	88.68 ± 0.13	97.09 ± 0.19
Online ProtoNet	90.49	85.68 ± 0.15	96.95 ± 0.09	84.61	88.71 ± 0.13	97.61 ± 0.17
CPM (Ours)	**93.55**	**89.49** ± 0.13	**98.07** ± 0.07	**89.43**	**91.81** ± 0.11	**98.34** ± 0.14

Table 3: **RoamingRooms OC-FSL results.** Max 100 images and 40 classes.

Method	Supervised			Semi-supervised		
	AP	1-shot Acc.	3-shot Acc.	AP	1-shot Acc.	3-shot Acc.
OML	74.34	72.63 ± 0.37	84.97 ± 0.32	58.71	68.87 ± 0.38	87.62 ± 0.51
LSTM	45.67	59.90 ± 0.40	61.85 ± 0.45	33.32	52.71 ± 0.38	55.83 ± 0.76
DNC	80.86	82.15 ± 0.32	87.30 ± 0.30	73.49	80.27 ± 0.33	87.87 ± 0.49
Online MatchingNet	85.91	82.82 ± 0.32	89.99 ± 0.26	78.99	80.08 ± 0.34	**92.43** ± 0.41
Online IMP	87.33	85.28 ± 0.31	90.83 ± 0.25	75.36	84.57 ± 0.31	91.17 ± 0.43
Online ProtoNet	86.01	84.89 ± 0.31	89.58 ± 0.28	76.36	80.67 ± 0.34	88.83 ± 0.49
CPM (Ours)	**88.09**	**87.07** ± 0.29	**90.90** ± 0.26	**82.88**	**84.97** ± 0.32	91.47 ± 0.44

Table 4: Ablation of CPM architectural components on RoamingOmniglot

Method	$\mathbf{h}^{\text{RNN}}$	β_t^*, γ_t^*	Metric	Val AP
Online PN				91.22
No $\mathbf{h}^{\text{RNN}}$		✓	✓	92.52
$\mathbf{h}^{\text{RNN}}$ only	✓			93.48
No metric $\mathbf{m}_t$	✓	✓		93.61
No β_t^*, γ_t^*	✓		✓	93.98
$\mathbf{h}_t = \mathbf{h}_t^{\text{RNN}}$	✓	✓	✓	93.70
Full CPM	✓	✓	✓	**94.08**

Table 5: Ablation of semi-supervised learning components on RoamingOmniglot

Method	RNN	Prototype	β_t^w, γ_t^w	Val AP
Online PN				90.83
Online PN		✓		89.10
Online PN		✓	✓	91.22
CPM				92.57
CPM	✓			93.16
CPM	✓	✓		93.20
CPM	✓	✓	✓	**94.08**

- **LSTM** (Hochreiter & Schmidhuber, 1997) & **DNC** (Graves et al., 2016): We include RNN methods for comparison as well. Differentiable neural computer (DNC) is an improved version of memory augmented neural network (MANN) (Santoro et al., 2016).

- **Online MatchingNet** (Vinyals et al., 2016), **IMP** (Allen et al., 2019) & **ProtoNet** (Snell et al., 2017): We used the same negative Euclidean distance as the similarity function for these three metric learning based approaches. In particular, MatchingNet stores all examples and performs nearest neighbor matching, which can be memory inefficient. Note that Online ProtoNet is a variant of our method without the contextual RNN.

Main results: Our main results are shown in Table 2 and 3, including both supervised and semi-supervised settings. Our approach achieves the best performance on AP consistently across all settings. Online ProtoNet is a direct comparison without our contextual RNN and it is clear that CPM is significantly better. Our method is slightly worse than Online MatchingNet in terms of 3-shot accuracy on the RoamingRooms semisupervised benchmark. This can be explained by the fact that MatchingNet stores all past seen examples, whereas CPM only stores one prototype per class. Per timestep accuracy is plotted in Figure 5, and the decaying accuracy is due to the increasing number of classes over time. In RoamingOmniglot, CPM is able to closely match or even sometimes surpass the offline classifier, which re-trains at each step, which uses all prior images in a sequence. This is reasonable as our model is able to leverage information from the current context.

Effect of spatiotemporal context: To answer the question whether the gain in performance is due to spatiotemporal reasoning, we conduct the following experiment compar-

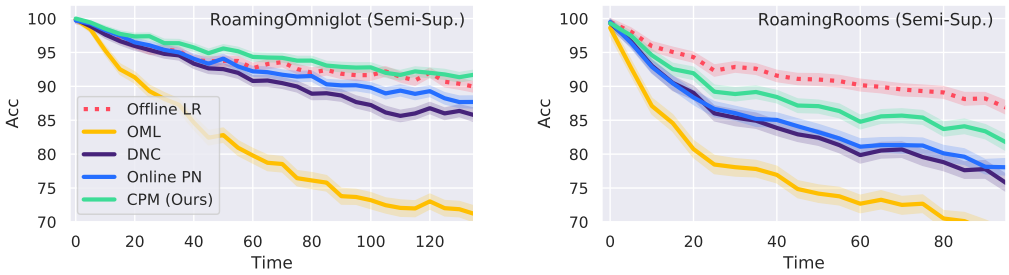

Figure 5: **Few-shot classification accuracy over time. Left:** RoamingOmniglot semisupervised. **Right:** RoamingRooms semisupervised. An offline logistic regression (Offline LR) baseline is also included, using pretrained ProtoNet features. It is trained on all labeled examples except for the one at the current time step.

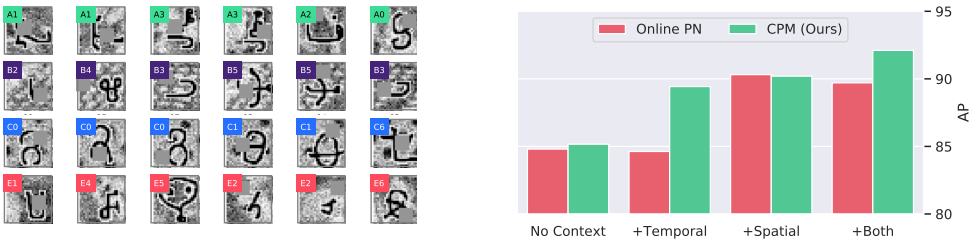

Figure 6: **Effect of spatiotemporal context.** Spatiotemporal context are added separately and together in RoamingOmniglot, by introducing texture background and temporal correlation. **Left:** Stimuli used for spatial cue of the background environment. **Right:** Our CPM model significantly benefits from the presence of a temporal context ("+Temporal" and "+Both"), while Spatial context helps both CPM and online ProtoNet.

ing CPM with online ProtoNet. We allow the CNN to have the ability to recognize the context in RoamingOmniglot by adding a texture background image using the Kylberg texture dataset (Kylberg, 2011) (see Figure 6 left). As a control, we can also destroy the temporal context by shuffling all the images in a sequence. We train four different models on dataset controls with or without the presence of spatial or temporal context, and results are shown in Figure 6. First, both online ProtoNet and CPM benefit from the inclusion of a spatial context. This is understandable as the CNN has the ability to learn spatial cues, which re-confirms our main hypothesis that successful inference of the current context is beneficial to novel object recognition. Second, only our CPM model benefits from the presence of temporal context, and it receives distinct gains from spatial and temporal contexts.

Ablation studies: We ablate each individual module we introduce. Results are shown in Tables 4 and 5. Table 4 studies different ways we use the RNN, including the context vector $\mathbf{h}^{\text{RNN}}$, the predicted threshold parameters β_t^*, γ_t^*, and the predicted metric scaling vector $\mathbf{m}_t$. Table 5 studies various ways to learn from unlabeled examples, where we separately disable the RNN update, prototype update, and distinct write-threshold parameters β_t^w, γ_t^w (vs. using read-threshold parameters). We verify that each component has a positive impact on the performance.

6. Conclusion

We proposed online contextualized few-shot learning, OC-FSL, a paradigm for machine learning that emulates a human or artificial agent interacting with a physical world. It combines multiple properties to create a challenging learning task: every input must be classified or flagged as novel, every input is also used for training, semi-supervised learning can potentially improve performance, and the temporal distribution of inputs is non-IID and comes from a generative model in which input and class distributions are conditional on a latent environment with Markovian transition probabilities. We proposed the RoamingRooms dataset to simulate an agent wandering within a physical world. We also proposed a new model, CPM, which uses an RNN to extract spatiotemporal context from the input stream and to provide control settings to a prototype-based FSL model. In the context of naturalistic domains like RoamingRooms, CPM is able to leverage contextual information to attain performance unmatched by other state-of-the-art FSL methods.

References

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

Allen, K. R., Shelhamer, E., Shin, H., and Tenenbaum, J. B. Infinite mixture prototypes for few-shot learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and van den Hengel, A. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.

Antoniou, A., Patacchiola, M., Ochal, M., and Storkey, A. J. Defining benchmarks for continual few-shot learning. *CoRR*, abs/2004.11967, 2020.

Chang, A. X., Dai, A., Funkhouser, T. A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV*, pp. 667–676, 2017.

Cohen, N. and Squire, L. Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, 210(4466): 207–210, 1980. ISSN 0036-8075. doi: 10.1126/science.7414331.

Devries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2017.

Finn, C., Rajeswaran, A., Kakade, S. M., and Levine, S. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.

French, R. M. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Gidaris, S. and Komodakis, N. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., and Hassabis, D. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, Oct 2016. ISSN 1476-4687. doi: 10.1038/nature20101.

Harrison, J., Sharma, A., Finn, C., and Pavone, M. Continuous meta-learning without tasks. *CoRR*, abs/1912.08866, 2019.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

Hou, S., Pan, X., Loy, C. C., Wang, Z., and Lin, D. Learning a unified classifier incrementally via rebalancing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019.

Howard, M. W. Temporal and spatial context in the mind and brain. *Current opinion in behavioral sciences*, 17: 14–19, 2017.

Javed, K. and White, M. Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

Jerfel, G., Grant, E., Griffiths, T., and Heller, K. A. Reconciling meta-learning and continual learning with online mixtures of tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

Kahana, M. J. *Foundations of human memory*. Oxford University Press, New York, 2012. ISBN 9780195333244. OCLC: ocn744297060.

Kaiser, L., Nachum, O., Roy, A., and Bengio, S. Learning to remember rare events. In *5th International Conference on Learning Representations, ICLR*, 2017.

Kemker, R. and Kanan, C. Fearnet: Brain-inspired model for incremental learning. In *Proceedings of 6th International Conference on Learning Representations (ICLR)*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Koch, G., Zemel, R., and Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kylberg, G. *Kylberg Texture Dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and . . . , 2011.

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

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

Li, F., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, abs/1707.09835, 2017.

Lomonaco, V. and Maltoni, D. Core50: a new dataset and benchmark for continuous object recognition. In *1st Annual Conference on Robot Learning, CoRL*, 2017.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2017.

Mcclelland, J., Mcnaughton, B., and O'Reilly, R. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102:419–57, 08 1995. doi: 10.1037/0033-295X.102.3.419.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Oreshkin, B. N., López, P. R., and Lacoste, A. TADAM: task dependent adaptive metric for improved few-shot learning. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2018.

Parisi, G. I., Tani, J., Weber, C., and Wermter, S. Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization. *Frontiers in neurorobotics*, 12:78–78, Nov 2018. ISSN 1662-5218. doi: 10.3389/fnbot.2018.00078.

Rebuffi, S., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H., and Zemel, R. S. Meta-learning for semi-supervised few-shot classification. In *6th International Conference on Learning Representations, ICLR*, 2018.

Ren, M., Liao, R., Fetaya, E., and Zemel, R. S. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. P. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*, 2016.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., and Batra, D. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, 2019.

Schmidhuber, J. *Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook*. Diplomarbeit, Technische Universität München, München, 1987.

She, Q., Feng, F., Hao, X., Yang, Q., Lan, C., Lomonaco, V., Shi, X., Wang, Z., Guo, Y., Zhang, Y., Qiao, F., and Chan, R. H. M. Openloris-object: A dataset and benchmark towards lifelong object recognition. *CoRR*, abs/1911.06487, 2019.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2017.

Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., and Gong, Y. Few-shot class-incremental learning. *CoRR*, abs/2004.10956, 2020.

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

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2016.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong learning with dynamically expandable networks. In *6th International Conference on Learning Representations, ICLR*, 2018.