

Online Continual Learning with Natural Distribution Shifts: An Empirical Study with Visual Data

Zhipeng Cai
Intel Labs

Ozan Sener
Intel Labs

Vladlen Koltun
Intel Labs

Abstract

Continual learning is the problem of learning and retaining knowledge through time over multiple tasks and environments. Research has primarily focused on the incremental classification setting, where new tasks/classes are added at discrete time intervals. Such an “offline” setting does not evaluate the ability of agents to learn effectively and efficiently, since an agent can perform multiple learning epochs without any time limitation when a task is added. We argue that “online” continual learning, where data is a single continuous stream without task boundaries, enables evaluating both information retention and online learning efficacy. In online continual learning, each incoming small batch of data is first used for testing and then added to the training set, making the problem truly online. Trained models are later evaluated on historical data to assess information retention. We introduce a new benchmark for online continual visual learning that exhibits large scale and natural distribution shifts. Through a large-scale analysis, we identify critical and previously unobserved phenomena of gradient-based optimization in continual learning, and propose effective strategies for improving gradient-based online continual learning with real data. The source code and dataset are available in: <https://github.com/IntelLabs/continuallearning>.

1. Introduction

Supervised learning aims to find models that can predict labels given input data, with satisfactory performance when evaluated on a specific population of interest. This population is typically sampled to create training data for the model to learn over. The critical requirement for successful learning is a set of training data points that are independent and identically distributed (iid.). Continual learning challenges this assumption and considers a sequence of learning problems where the distribution changes dramatically through time. This setting is crucial when learned models are deployed in interactive systems, since the environment with which agents interact continually evolves.

There are two critical performance metrics for a continual learner: learning efficacy and information retention. Learning efficacy entails the simple question, “is learning the n-th thing easier than learning the first thing?” [31]. This ability is critical for fast learning and quick adaptation. Information retention considers the model’s ability to quickly recall when faced with a historical task that had been previously considered. This question is also studied to understand a property of neural networks called catastrophic forgetting [21].

Continual learning algorithms are typically evaluated in an incremental classification setting, where tasks/classes arrive one-by-one at discrete time intervals. Multiple learning epochs over the current task are permitted and the learner can spend as much time as desired to learn each task. This setting is appropriate for evaluating information retention, because access to previous tasks is prohibited. However, learning efficacy is not evaluated in this setting, because models can easily learn each task from scratch and still be successful [23]. We refer to this incremental classification setting as *offline* continual learning.

Our work focuses on *online* continual learning, which aims to evaluate learning efficacy in addition to information retention. In online continual learning, there is a single online stream of data. At every time step, a small batch of data arrives. The model needs to immediately predict labels for the incoming data as an *online testing step*. After the prediction, this batch of data is added to the dataset. The model needs to be updated before the next batch of data arrives using a fixed budget of computation and memory; this is an *online training step*. Testing and training are on the fly. Moreover, as we are in a continual learning setting, the data distribution changes over time. A successful online test performance requires efficient learning and adaptation in this non-stationary setting.

To study online continual visual learning, we construct a new benchmark where the data distribution evolves naturally over time. To do so, we leverage images with geolocation tags and time stamps. We use a subset of YFCC100M [30] with 39 million images captured over 9 years. Our task is *online continual geolocalization*. We empirically evaluate the natural distribution shift and validate that the benchmark

is appropriate to study online continual learning. We further use this benchmark to analyze the behavior of gradient-based optimization in online continual learning. Our experiments suggest that the non-stationarity in the data results in a significantly different behavior from what was previously observed in offline continual learning. Surprisingly, learning efficacy and information retention turn out to be conflicting objectives from an optimization perspective, necessitating a careful trade-off. We also found that mini-batching is a non-trivial problem for online continual learning. Increasing batch sizes in SGD, even by a small factor, significantly hurts both learning efficacy and information retention. Based on the analysis, we propose simple yet effective strategies, such as online learning rate and replay buffer size adaptation algorithms, that significantly improve gradient-based optimization for online continual learning. We will share our benchmark and code with the community in order to support future research in online continual visual learning.

2. Related Work

Continual learning benchmarks. Most visual continual learning benchmarks use synthetic task sequences. A common way to synthesize task sequences is to separate the original label space into multiple subspaces. Split MNIST [34], Split notMNIST [22], Split CIFAR10/100 [34], Split Tiny-ImageNet [4], and iILSVRC [24] are all constructed in this way. Another approach is to inject different types of transformations into data as new tasks. Permuted MNIST [34] and Rotated MNIST [18] are constructed in this way. For online visual continual learning, Aljundi et al. [2] use soap opera series for actor identification. Though training is online, testing is done once per episode, on time-invariant validation data from all episodes. Recently, the Firehose benchmark [13] proposed online continual learning of natural language models. It contains a large stream of Twitter posts used to train a self-supervised and multi-task language model for per-user tweet prediction.

Continual learning algorithms. Existing continual learning algorithms can be roughly categorized into 1) regularization-based, 2) parameter-isolation-based, and 3) replay-based (see [5] for a detailed survey). Regularization-based methods add regularization terms into the training loss, based on distillation [17] or estimated network parameter importance [15, 34, 1, 16]. Since historical data are not cached for training, these methods often suffer from catastrophic forgetting, especially given a long data stream [6]. Parameter-isolation-based methods [20, 27] assign different sets of network parameters to different tasks. However, they are not suitable for online continual learning, due to the use of task ID for training. Replay-based methods train models using historical examples [24, 4] or examples synthesized by generative models that are trained on historical

samples [29]. The replay data can serve as part of the training data, or be utilized to constrain the gradient direction during training [18, 3]. Replay-based methods have generally been found more effective than the other methods and we adopt them as our algorithmic starting point in this work.

Geolocalization algorithms. We use geolocalization on a stream of images as an online continual learning setting. Though our work is the only continual/online approach, the problem of geolocalization has been widely studied. The pioneering work of Hays and Efros [9, 10] addresses geolocalization by image retrieval. Given a query image, it performs a nearest neighbor search on millions of images with geolocation tags and uses the location of the nearest neighbour for prediction. Vo et al. [32] replace handcrafted retrieval features with ones computed by a deep neural network. This network is trained with a classification loss but only used for feature extraction. PlaNet [33] formulates geolocalization into an image classification problem and trains a deep neural network for classification. PlaNet is much more efficient than retrieval-based approaches because it only requires one forward pass per query image, rather than nearest neighbor searching over millions of images. CPlaNet [26] introduced combinatorial partitioning to PlaNet, which generates fine-grained class labels using the intersection of multiple coarse-grained partitions of the map. This technique is used to alleviate the conflict between the large label space and the small number of training examples per class. We use PlaNet as a starting point due to its simplicity and efficiency, and extend it to continual and online settings.

3. Online Continual Learning

In this section, we formally define online continual learning (OCL), mainly following the definition of online learning. We further discuss the metrics we use to evaluate learning efficacy and information retention.

Following the common definition of online learning [11, 28], we define OCL as a game between a learner and an environment. The learner operates in a predefined function class $h(\cdot; \theta): \mathcal{X} \rightarrow \mathcal{Y}$ with parameters θ predicting the label $Y \in \mathcal{Y}$ given an input $\mathbf{X} \in \mathcal{X}$. At each step $t \in \{1, 2, \dots, \infty\}$ of the game,

- The environment generates a set of data points $\mathbf{X}_t \sim \pi_t$ sampled from a non-stationary distribution π_t .
- The learner makes predictions for \mathbf{X}_t using the current model θ_t as $\hat{Y}_t = h(\mathbf{X}_t; \theta_t)$.
- The environment reveals the true labels Y_t and computes the online metric $m(Y_t, \hat{Y}_t)$.
- The learner updates the model θ_{t+1} using a fixed budget of computation and memory.

Since this formalization directly follows online learning, it is only designed for evaluating online performance. To further assess information retention, we evaluate the models

$\theta_{H/3}, \theta_{2H/3}, \theta_H$ at predefined time instants $H/3, 2H/3, H$, where H is the total number of time steps for the evaluation. The models are evaluated over validation data (not seen during training) sampled at historical time steps. We refer to this as *backward transfer*. Similarly, we also evaluate the models on future data and call this *forward transfer*.

Metrics. To evaluate learning efficacy, we measure the *average online accuracy* over time, similar to average regret in online learning. We compute the running average of the computed accuracy on the fly as

$$\text{acc}_0(t) = \frac{1}{t} \sum_{s=1}^t \text{acc}(Y_s, \hat{Y}_s). \quad (1)$$

In order to evaluate information retention, we compute *backward transfer* at various predefined time instants. We specifically compute the average accuracy in the historical data. Formally, the backward transfer for the model at time T (T is $H/3, 2H/3$, or H for us) is defined as

$$\text{acc}_B @ T(t_B) = \frac{1}{t_B} \sum_{s=T-t_B}^T \text{acc}(Y_s, h(\mathbf{X}_s; \theta_T)). \quad (2)$$

Finally, we also evaluate how well the model transfers information to the future. This metric can be used to evaluate out-of-distribution generalization of the model through time. Similar to backward transfer, we evaluate *forward transfer* at a specific time point T (specifically $H/3$ and $2H/3$) as

$$\text{acc}_F @ T(t_F) = \frac{1}{t_F} \sum_{s=T}^{T+t_F} \text{acc}(Y_s, h(\mathbf{X}_s; \theta_T)). \quad (3)$$

All three metrics are a function of time, not a single time-invariant value. This choice is intentional as the problem of interest is online. When we compare two algorithms, if one works better at earlier times and the other works better in later times, we would like to detect these characteristics. We thus plot these three metrics through time for evaluation.

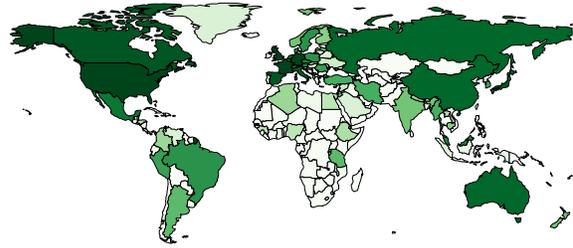
4. Continual Localization Benchmark

To study online continual visual learning, we need a large-scale benchmark with natural distribution shifts. We propose using geolocation as a task for online continual visual learning, since the data is readily available. We use images with their geolocations and time stamps from YFCC100M [30].

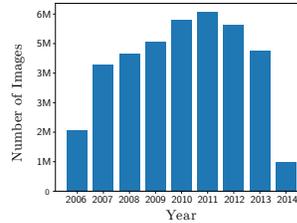
After obtaining time stamps for YFCC100M images, we perform a series of pre-processing steps. First, we keep only images from 2004 to 2014 since the images from pre-2004 typically have noisy time stamps in the form of the default date and time of the users' physical camera. Then, we order the data stream according to the time stamp, i.e.,



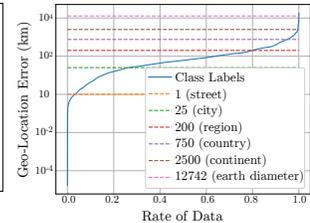
(a) S2 Cells in our dataset



(b) Distribution of number of images per country



(c) Statistics over years



(d) Geolocation accuracy

Figure 1. Statistics of our benchmark. (a) We convert the geolocation problem into classification over 712 S2 Cells, visualized here. (b) We plot the distribution of images over countries. Dominating number of images come North America and Europe. (c) Number of images per year changes over time. (d) Geolocation accuracy of the S2 Cells. (Lower right is better.) Following [10], approximate scales of streets, cities, regions, countries and continents are plotted.

images taken earlier will be seen by the continual learner first. Finally, we allocate the first 5% of the data for offline preprocessing and randomly sample another 1% over the entire time range as the held-out validation set to evaluate backward and forward transfer. We have 39 million images for continual learning after this preprocessing, 2 million images for offline preprocessing, and 392 thousand images for backward and forward transfer evaluation. We name our benchmark CLOC, short for Continual LOCALization. We visualize the distribution of these images over years and countries in Figure 1. We can see from the figure that the distribution of the images is heavily biased towards North America and Europe due to the distribution of Flickr users.

We apply a standard procedure to convert the geolocation problem into classification [33]. Specifically, we

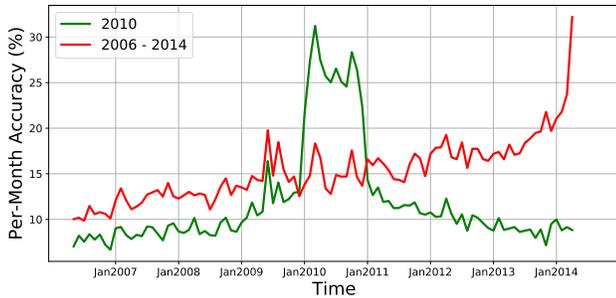


Figure 2. **Distribution shift in CLOC.** We train two supervised models, one using data from the entire temporal range and the other only on data from the year 2010. We evaluate both models on the full temporal range using the validation set (not seen during training). Due to non-stationarity in the data, the performance of the 2010 model drops sharply on data from other times.

divide the earth into multiple regions using the S2 Cell hierarchy [25], following statistics computed on the offline pre-processing set. We also limit the number of images within each class to be between 50 and 10K. Using this process, we generate 712 S2 cells over the earth, converting our problem into a 712-class classification problem. We visualize the S2 cells in Figure 1(a). Regions with more images are subdivided more finely.

In order to quantify the loss of geolocalization accuracy due to the constructed class labels, we plot in Fig. 1(d) the distribution of distances between the location of each image and the corresponding class center, i.e., the center of the S2 Cell. We can see that roughly 5% of the data are $< 1km$ from their class center at the street level. Roughly 30% of the data are $< 25km$ from their class center at the city level. Roughly 80% of the data are $< 200km$ from their class center at the region level.

Compared to previous visual continual learning benchmarks, CLOC has a much larger scale: it contains 39 million images taken over 9 years, with 712 classes. Crucially, the distribution shift in CLOC is natural rather than synthetic. CLOC requires neither explicit task boundaries nor artificially injected data augmentations. The data distribution naturally evolves over time.

To perform OCL on CLOC, we ingest at each time step the consecutive set of images uploaded by the same user, since users sometimes upload at the same time a small album of images, which are typically from the same region. Once the images arrive, the model needs to make predictions, update its model, and move to the next set of images. Note that images are ordered with the time stamps. Hence the OCL model receives these images in their natural order.

Validating the Distribution Shift of CLOC. One key question that we need to validate is the continual nature of the problem. If the data distribution is not changing through time, there is no need for continual learning. To validate the distribution shift, we train two supervised learning models

on CLOC (see Section 5.1 for implementation details). The first model is trained over the entire temporal range, and the second only on data from the year 2010. For fair comparison, we subsample the training set of the full-range model to the same size as the training set of the other model. We evaluate the trained models on the held-out validation data and plot the top-1 accuracy vs. time in Figure 2. The performance of the temporally localized model drops significantly on data from other times. In contrast, the full-range model does not have a sharp performance increase in year 2010. This indicates that there is significant non-stationarity in CLOC.

5. Empirical Study and Results

5.1. Setup

To empirically study online continual visual learning, we make a series of systematic choices focusing on simplicity, scalability, and reproducibility. Our model predicts the label for each image independently, though sometimes multiple images from the same user may be tested together. We choose the ResNet50 [12] architecture and the experience replay (ER) [4] algorithm as the main setting. We choose ER since it is an effective continual learning algorithm that scales to large-scale problems. ER uses a *replay buffer* to store a set of historical data. The strategy for keeping the replay buffer is important for continual learning. However, our analysis suggests that it has a minor impact on gradient-based optimization, which is the main focus of our work. Hence, we deploy the First-In-First-Out (FIFO) buffer for analysis in the main paper and empirically evaluate this choice in the supplement.

In each training iteration, images from the data stream are joined together with images randomly sampled from the replay buffer to form the training batch. Note that the learning efficacy, i.e., eq. (1), is measured only on the data stream, without including the replay data. We apply standard data augmentations to images [12]. Augmentations are applied to both streaming and replay data. In other words, the replay buffer contains original images, and a random augmentation is applied whenever an image is replayed.

The concept of batch size in OCL does not exist as models need to predict the label for each incoming image online. In its direct implementation, this corresponds to a “batch size” of 1 album, i.e., the next consecutive set of images uploaded by the same user (1.157 images on average). However, analysis from supervised learning suggests that mini-batching is beneficial for gradient-based optimization [7]; moreover, computational considerations also suggest some form of mini-batching since performing $\frac{39 \text{ million}}{1.157} \approx 34$ million iterations in order to test a model would be prohibitively costly. Thus, we relax this constraint while keeping the online nature of the problem. The model receives 256 images at each iteration, but the online accuracy is computed over only the

first album in these images.

Choosing hyperparameters for continual learning is especially tricky since the continual setting precludes “rerunning” models in identical conditions for hyperparameter selection. Once a model is deployed, it needs to learn continually without retraining. Hence, we make all hyperparameter decisions using the offline preprocessing dataset. The hyperparameters set by this method are the initial learning rate (0.05), weight decay ($1e-4$), and loss (cross-entropy). We also report offline supervised learning results as a baseline, with hyperparameters tuned separately on the whole dataset. As a result, the supervised learning baseline uses an initial learning rate of 0.025, weight decay of $1e-4$, the cross-entropy loss, and the cosine learning rate schedule.

5.2. Learning Rate Analysis

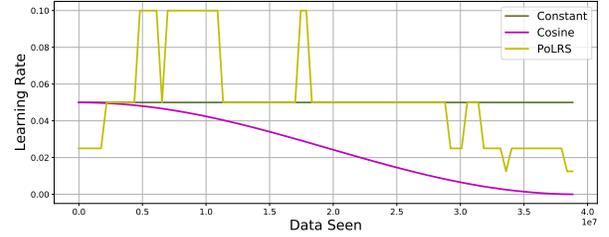
In order to study the role of learning rates and their schedule in OCL, we compare a constant learning rate schedule where the learning rate is fixed through time, and the cosine schedule [19] with one cycle and minimum LR of 0. We also evaluate an adaptive schedule generated by population-based search [14]. Population-based search trains 3 models in parallel (keeping the total amount of computing the same as in the other conditions); every N^{th} step, the weights of the best-performing model are copied to the others, and the learning rates of all models are centered at the current best learning rate. We set N to 2 million, which is roughly the same size as the offline preprocessing step used to set initial hyperparameters. We summarize this schedule in Alg. 1, and call it PoLRS, short for Population Learning Rate Search.

Algorithm 1 Population Learning Rate Search (PoLRS)

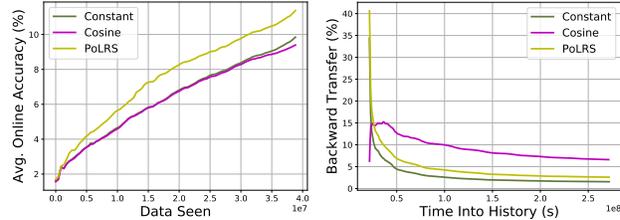
Require: Learning rate update interval N , metric for performance evaluation $m(\cdot)$, initial learning rate l_0 , initial models $\{\theta_1^1, \theta_1^2, \theta_1^3\}$.

- 1: $j^* \leftarrow 1$ and set learning rates to $\{2l_0, l_0, 0.5l_0\}$.
 - 2: **for** $t \in \{1, 2, \dots\}$ **do**
 - 3: Test on \mathbf{X}_t using $\theta_t^{j^*}$ and compute cost $m_t(\cdot)$
 - 4: Update $\theta_t^1, \theta_t^2, \theta_t^3$ and their metrics m_t^1, m_t^2, m_t^3
 - 5: $j^* = \arg \max_j (m_t^j)$
 - 6: **if** $t \bmod N = 0$ **then**
 - 7: Copy $\theta_t^{j^*}$ to $\theta_t^1, \theta_t^2, \theta_t^3$.
 - 8: Set learning rates to $\{2l_t^{j^*}, l_t^{j^*}, 0.5l_t^{j^*}\}$
 - 9: **end if**
 - 10: **end for**
-

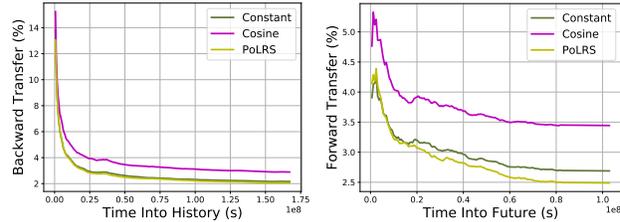
We plot the learning rate of each method over time in Fig. 3(a). PoLRS had a significantly different learning rate from other methods. We used the online accuracy as the metric while using PoLRS, and the online accuracy results from Fig. 3(b) show that PoLRS outperformed the other schedules. We further evaluated forward and backward transfer. Fig. 3(c) shows the backward transfer of all schedules



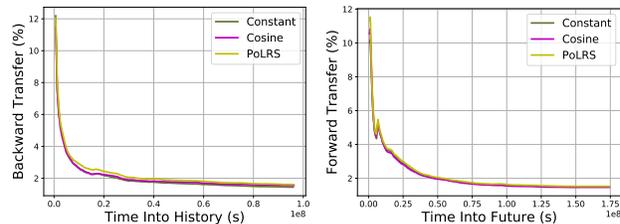
(a) Learning rate



(b) Average online accuracy (\uparrow) (c) Backward transfer at H (\uparrow)



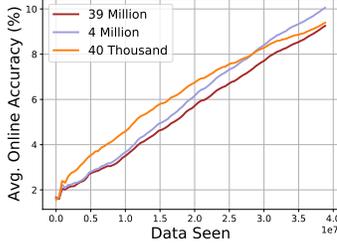
(d) Backward transfer at $2H/3$ (\uparrow) (e) Forward transfer at $2H/3$ (\uparrow)



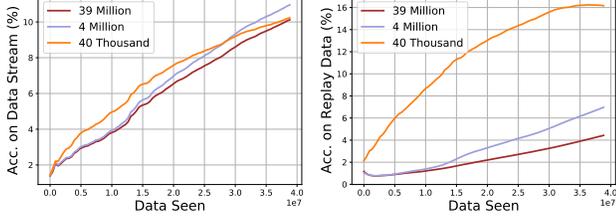
(f) Backward transfer at $H/3$ (\uparrow) (g) Forward transfer at $H/3$ (\uparrow)

Figure 3. **Learning rate analysis results.** H is the total number of time steps. Arrows in the subfigure captions point towards better performance. (a) PoLRS adjusted the learning rate dynamically over time. (b) PoLRS out-performed fixed schedules in terms of average online accuracy. (c) The cosine schedule had the best backward transfer at time H , though its average online accuracy was the worst. (d,e) The cosine schedule had the best transfer when its learning rates were much smaller than other schedules ($2H/3$ and H). (f,g) Schedules using large learning rates tend to have low forward and backward transfer.

at the final time step H . The cosine schedule had the best backward transfer but the worst online fit. Hence, the best learning rates for learning efficacy and information retention are different. Figs. 3(d) to 3(g) further plot the forward and backward transfer at times $H/3$ and $2H/3$. We see that all schedules had low forward and backward transfer when their learning rates were large, i.e., at time $H/3$. Moreover, when



(a) Average online accuracy (\uparrow)



(b) Training accuracy $\text{acc}_{\text{Stream}}$ on the data stream (c) Training accuracy acc_{Rep} on replay data

Figure 4. **The effect of replay buffer size R on learning efficacy.** (a) The optimal replay buffer size changes over time. (b,c) Correlation between replay buffer size and training statistics. When R was too small (40 thousand at late time steps), training accuracy $\text{acc}_{\text{Rep}} \gg \text{acc}_{\text{Stream}}$ on the data stream. When R was too large (4 million and 39 million), $\text{acc}_{\text{Rep}} \ll \text{acc}_{\text{Stream}}$.

the learning rate of the cosine schedule became much smaller than the others, i.e., at time $2H/3$ and H , it consistently outperformed the other schedules. These results clearly show that small learning rates facilitate transfer, though potentially at the cost of learning efficacy.

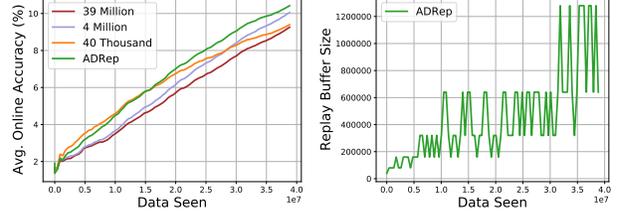
Summary of findings. (1) The ideal learning rates for learning efficacy and transfer are radically different. (2) If *learning efficacy* is prioritized, adaptive learning rate schedules, e.g., produced by PoLRS, are better. (3) If *transfer* is prioritized, and there is a pre-defined learning horizon H , the cosine schedule is a good choice.

5.3. Replay Buffer Size Analysis

In order to analyze the impact of replay buffer sizes, we train three models with replay buffer sizes of 40 thousand, 4 million, and 39 million, respectively. We use the cosine learning rate schedule for this analysis.

We report learning efficacy in Fig. 4(a) for all models and conclude that the optimal buffer size changes through time and the common choice of using the largest possible buffer is not always optimal. Interestingly, we found that it is possible to successfully adapt the buffer size by comparing the training accuracy on replay data and online data stream. We define training accuracy on the data stream at time t as

$$\text{acc}_{\text{Stream}}(t) = \frac{1}{t} \sum_{s=1}^t \text{acc}(Y_s, h(\mathbf{X}_s; \theta_s)), \quad (4)$$



(a) Average online accuracy (\uparrow) (b) Replay buffer sizes of ADRep over time

Figure 5. **ADRep vs. constant replay buffer sizes.** (a) ADRep provided near-optimal average online accuracy throughout training. (b) ADRep gradually increased the replay buffer size over time.

Algorithm 2 Adaptive replay buffer size (ADRep)

Require: Update interval N' , initial replay buffer size R , difference threshold ϵ , initial model θ_1 .

- 1: $\text{acc}_{\text{Stream}}, \text{acc}_{\text{Rep}} \leftarrow 0; k \leftarrow 1$
 - 2: **for** $t \in \{1, 2, \dots\}$ **do**
 - 3: update model with replay buffer size of R .
 - 4: $\text{acc}_{\text{Stream}} \leftarrow \frac{(k-1)\text{acc}_{\text{Stream}} + \text{acc}(Y_t, h(\mathbf{X}_t; \theta_t))}{k}$
 - 5: $\text{acc}_{\text{Rep}} \leftarrow \frac{(k-1)\text{acc}_{\text{Rep}} + \text{acc}(Y_t^{\text{Rep}}, h(\mathbf{X}_t^{\text{Rep}}; \theta_t))}{k}$
 - 6: $k \leftarrow k + 1$
 - 7: **if** $t \bmod N' = 0$ **then**
 - 8: $k \leftarrow 1$
 - 9: **if** $\text{acc}_{\text{Stream}} > \text{acc}_{\text{Rep}} + \epsilon$ **then** $R \leftarrow \frac{R}{2}$
 - 10: **if** $\text{acc}_{\text{Stream}} < \text{acc}_{\text{Rep}} - \epsilon$ **then** $R \leftarrow 2R$
 - 11: **end if**
 - 12: **end for**
-

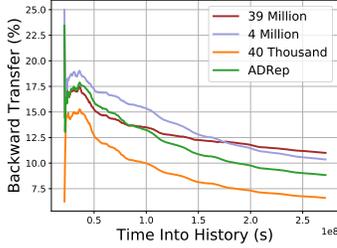
where \mathbf{X}_s, Y_s are the training data and labels from the data stream at time s . Similarly, we define the training accuracy on replay data at time t as

$$\text{acc}_{\text{Rep}}(t) = \frac{1}{t} \sum_{s=1}^t \text{acc}(Y_s^{\text{Rep}}, h(\mathbf{X}_s^{\text{Rep}}; \theta_s)), \quad (5)$$

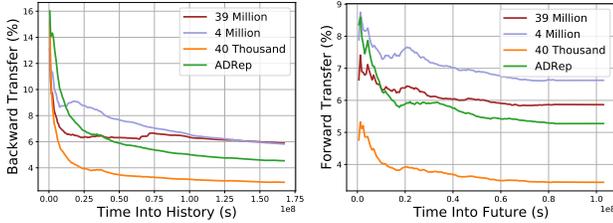
where $\mathbf{X}_s^{\text{Rep}}, Y_s^{\text{Rep}}$ are the training data and labels sampled from the replay buffer.

We can see from Figs. 4(b) and 4(c) that when R was too large (4 million and 39 million), the accuracy on replay data acc_{Rep} was much lower than the accuracy on the data stream $\text{acc}_{\text{Stream}}$, making the model focus less on adaptation. When R was too small (40 thousand at late training stages), acc_{Rep} was much higher than $\text{acc}_{\text{Stream}}$. This overfitting on the replay buffer harmed the replay-based regularization.

Based on this observation, we propose to adapt the buffer size R over time by comparing $\text{acc}_{\text{Stream}}$ and acc_{Rep} . We call this adaptive algorithm ADRep and specify it in Alg. 2. Every N'^{th} iteration, if $\text{acc}_{\text{Stream}} > \text{acc}_{\text{Rep}} + \epsilon$, we decrease R . And if $\text{acc}_{\text{Stream}} < \text{acc}_{\text{Rep}} - \epsilon$, we increase R .



(a) Backward transfer at H (\uparrow)



(b) Backward transfer at $2H/3$ (\uparrow) (c) Forward transfer at $2H/3$ (\uparrow)

Figure 6. **Transfer results of different replay buffer sizes.** H is the final time step. (a) Replay buffer size 39 million had comparable backward transfer to 4 million. (b,c) Replay buffer size 39 million had worse forward and backward transfer to 4 million. These results show that large replay buffer sizes are not always beneficial for transfer.

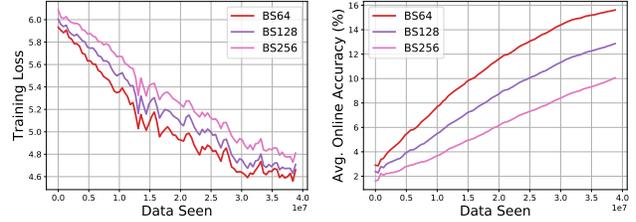
We observed that ADRep is not sensitive to the hyperparameter choices, and we choose $\epsilon = 0.5\%$ and $N' \approx 40K$. Fig. 5 shows the result of ADRep. From Fig. 5(a), we see that ADRep shows significantly superior performance throughout training. More importantly, Fig. 5(b) shows that ADRep was able to gradually adapt R over time. Note that ADRep only requires to compute and compare online training statistics to adapt replay buffer sizes, which requires almost no extra computational/storage budget.

In terms of transfer, Fig. 6 plots the transfer results of all models. We see that $R = 4$ million was better than $R = 39$ million at time $2H/3$ and comparable at time H . Therefore, as with learning efficacy, larger R is not always beneficial for transfer. ADRep did not perform well with respect to transfer in comparison to other strategies.

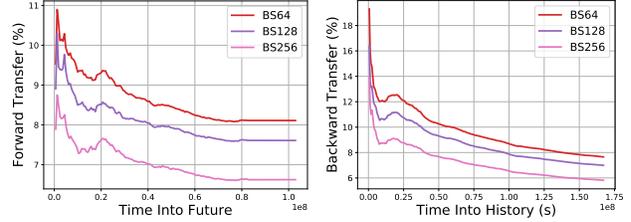
Summary of findings. (1) Larger replay buffer sizes are not always better both in terms of learning efficacy and transfer. (2) In terms of learning efficacy, the proposed adaptation strategy is successful, with almost no extra computation cost.

5.4. Batch Size Analysis

The concept of batch size is difficult to define in online continual learning. The natural setting is a batch size of 1 album (1.157 on average on CLOC), but $\frac{39 \text{ million}}{1.157} \approx 34$ million iterations of SGD in order to evaluate a model are not practically feasible. Moreover, supervised learning practice suggests that mini-batching is beneficial, especially for saving training time. To separate evaluation from implementation,



(a) Training loss (\downarrow) (b) Average online accuracy (\uparrow)



(c) Forward transfer at $2H/3$ (\uparrow) (d) Backward transfer at $2H/3$ (\uparrow)

Figure 7. **The effect of batch size.** H is the final time step. Increasing batch size has consistent and strong negative effect on the training loss (a) and all performance metrics, including average online accuracy (b), forward transfer (c), and backward transfer (d).

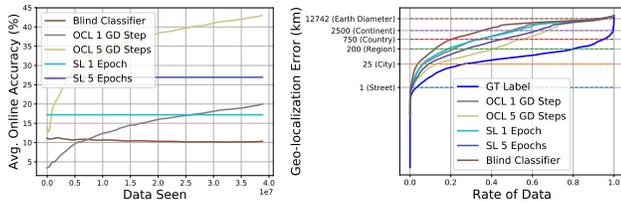
we evaluate online accuracy using the batch size of single album, but the training process is free to use any batch size.

Analysis in supervised learning [8] suggests that learning rates and batch sizes should be scaled similarly. Specifically, training with batch size $F \cdot B_0$ and learning rate $F \cdot l_0$ will have similar learning dynamics to training with batch size B_0 and learning rate l_0 , as long as $F \cdot B_0$ is not too large, e.g., ≤ 4 thousand. This behavior is repeatable for CLOC when we use supervised learning on shuffled data. (See the supplement for details.)

We similarly analyze the effect of batch sizes to OCL. We group B consecutive samples into a batch and do training on each batch of data once we see B images (Testing is always done with the newest model independently.). To analyze the effect, we train three models with B set to 64, 128, and 256. For $B < 256$, we reduce the learning rate to $\frac{B}{256}l$, where l is the learning rate used for $B = 256$. We use a cosine schedule and a 4 million replay buffer here. Unlike in supervised learning, increasing B (even by a small factor F) heavily hurts all performance metrics of OCL.

As shown in Fig. 7, the training loss (Fig. 7(a)) as well as all performance metrics (Figs. 7(b) to 7(d)) changed dramatically with B . Increasing B had a consistent and strong negative effect on all metrics. This was true whether we use a replay buffer or not. The supplement provides similar plots for models trained without ER. This means that the effect of batch size is not an algorithm-specific result.

In supervised learning, data is iid. and gradient estimates are unbiased regardless of batch sizes. Batch sizes only affect the variance. In OCL, gradients are not unbiased as



(a) Avg. online accuracy (\uparrow) (b) Top-1 predicted location error

Figure 8. **OCL model vs. others.** SL refers to the result of supervised learning. "GD Step" means the number of gradient descent steps per mini-batch. (a) Average online accuracy of OCL vs. average online accuracy of the blind classifier and the validation accuracy of supervised models. The average online accuracy of OCL was comparable to the validation accuracy of supervised learning models. (b) Geolocalization error of top-1 OCL predictions. Bottom right is better.

the distribution is non-iid. With increasing batch sizes, the variance decreases but bias increases. Due to the differences between offline and online continual learning, gradient-based optimization requires additional care for online continual learning, further validating a benchmark such as CLOC.

Summary of findings. Unlike in supervised learning, mini-batching in OCL is not straightforward. Increasing the batch size, even by a small factor, has a strong negative effect on all performance metrics. This suggests that the smallest possible batch size should be used.

5.5. Overall Performance

In this section, we develop an online continual learning strategy that combines the presented findings and evaluate its performance. Specifically, we utilize the adaptive learning rate schedule from PoLRS, adaptive buffer size from ADRep, and small batch sizes. We set the batch size to 64, which is the smallest value such that training on the whole dataset fits within a reasonable amount of time.

To analyze this model, we compare its average online accuracy against the performance of two other classifiers. The first one is the average online accuracy of a "blind" classifier, which does not use input images. The blind classifier only uses historical labels to predict the labels of future examples, without using the actual images. Specifically, we use the most frequent labels that appeared in the previous 10 images (tuned to achieve the maximum average online accuracy). The second classifier is the supervised learning model. We train supervised learning models with two time budgets, one for one epoch on the whole training set, and the other for five epochs. We compare the average online accuracy of OCL to the average validation accuracy of the supervised learning models. To examine the effect of the training budget for OCL, we also trained two OCL models with respectively one and five gradient descent steps per mini-batch.

The results are summarized in Fig. 8. Fig. 8(a) indicates

that the blind classifier had significantly better performance than chance, which would have $\frac{1}{712}$ accuracy. In turn, the average online accuracy of OCL rose significantly beyond the accuracy of the blind classifier. This indicates that the model successfully learned over time. Somewhat surprisingly and promisingly, the online accuracy of OCL was better than the validation accuracy of supervised learning models given similar budgets. This indicates that learning efficacy and information retention are conflicting in some aspects, i.e., optimizing one objective may hurt the other. Hence, it is important to choose the right optimization objective in practice.

We also plot the distance between the location of each image and the location represented by the top-1 prediction of OCL (Fig. 8(b)). Similar to the supervised learning results of PlaNet [33], most of the predictions were far from the actual location of the image because of the inherent difficulty of visual geolocalization. Nonetheless, OCL was better than supervised learning in terms of the geolocalization error.

6. Conclusion

We studied the problem of online continual learning with visual data. Using images with time stamps and geolocation tags, we proposed a new benchmark with a large scale and natural distribution shifts. We analyzed the effect of major optimization decisions, including learning rates, replay buffer sizes, and batch sizes. We found that the ideal learning rates are different for learning efficacy and information retention, and proposed different schedules for different performance metrics. We also found that the common practice of using the maximum possible replay buffer size for experience replay is not always optimal. We proposed an online replay buffer size adaptation algorithm to improve learning efficacy. For batch sizes, we found that mini-batching in OCL is non-trivial. Unlike in supervised learning, where mini-batch SGD is the standard for parallelized training, increasing batch sizes, even by a small amount, severely hurts both learning efficacy and information retention. Hence, the smallest possible batch size should be used for OCL. Using the proposed strategies, we were able to train an OCL model with comparable online accuracy to the validation accuracy of supervised learning models trained with similar budgets.

Future directions. Many interesting future research directions have emerged from this work. For example, in Sec. 5.5, we see that the blind classifier can use the temporal coherence of the labels to achieve much better online performance than random guessing. Leveraging such temporal coherence to improve our model, which only uses a single image as input, is an interesting possibility. Meanwhile, this work mainly studies "supervised" OCL, where the label of every example is observable. Extending the study to "semi-supervised" or "self-supervised" online visual continual learning is also an interesting topic.

References

- [1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018.
- [2] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.
- [3] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.
- [4] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc’Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.
- [5] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2019.
- [6] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.
- [7] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.
- [8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large mini-batch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [9] James Hays and Alexei A. Efros. IM2GPS: Estimating geographic information from a single image. In *CVPR*, 2008.
- [10] James Hays and Alexei A. Efros. Large-scale image geolocalization. In *Multimodal Location Estimation of Videos and Images*, pages 41–62. Springer, 2014.
- [11] Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13] Hexiang Hu, Ozan Sener, Fei Sha, and Vladlen Koltun. Drinking from a firehose: Continual learning with web-scale natural language. *arXiv preprint arXiv:2007.09335*, 2020.
- [14] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.
- [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*, 114(13):3521–3526, 2017.
- [16] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *arXiv preprint arXiv:1703.08475*, 2017.
- [17] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- [18] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6470–6479, 2017.
- [19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [20] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.
- [21] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
- [22] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.
- [23] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.
- [24] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [25] S2Geometry. S2 geometry library. https://s2geometry.io/devguide/s2cell_hierarchy.
- [26] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018.
- [27] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018.
- [28] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.
- [29] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2994–3003, 2017.
- [30] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [31] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *NIPS*, 1995.

- [32] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2621–2630, 2017.
- [33] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016.
- [34] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.