Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?

Hasan Abed Al Kader Hammoud[1*] Ameya Prabhu[2*] Ser-Nam Lim[3] Philip H.S. Torr[2]
Adel Bibi[2†] Bernard Ghanem[1†]
KAUST[1] University of Oxford[2] Meta AI[3]

Abstract

We revisit the common practice of evaluating adaptation of Online Continual Learning (OCL) algorithms through the metric of online accuracy, which measures the accuracy of the model on the immediate next few samples. However, we show that this metric is unreliable, as even vacuous blind classifiers, which do not use input images for prediction, can achieve unrealistically high online accuracy by exploiting spurious label correlations in the data stream. Our study reveals that existing OCL algorithms can also achieve high online accuracy, but perform poorly in retaining useful information, suggesting that they unintentionally learn spurious label correlations. To address this issue, we propose a novel metric for measuring adaptation based on the accuracy on the near-future samples, where spurious correlations are removed. We benchmark existing OCL approaches using our proposed metric on large-scale datasets under various computational budgets and find that better generalization can be achieved by retaining and reusing past seen information. We believe that our proposed metric can aid in the development of truly adaptive OCL methods. We provide code to reproduce our results at https://github.com/drimpossible/EvalOCL.

1. Introduction

The need for learning on continuously changing data streams has led to a proliferation of research in *Online Continual Learning (OCL)* literature [21, 8]. The primary aim of OCL algorithms is to enable deep models to continuously adapt to new data distributions without compromising the accumulated knowledge. However, we argue that current metrics have deviated significantly from measuring true adaptation capabilities. The majority of recent works on OCL [7, 10] measure the adaptation of OCL algorithms using the metric of *online accuracy*, defined as the accuracy of a model on the immediate incoming samples. This evaluation practice resulted in the neglect in measuring the capacity of models to retain previous knowledge. As a consequence, several OCL algorithms achieve high online accuracy [7] but poor performance on other important metrics like *information retention* which measure catastrophic forgetting of past seen data. Poor information retention performance has been attributed to an inherent stability-plasticity trade-off between learning new concepts and remembering old concepts [6]. However, the ability to preserve information and representations of past data, such as images of a class like the Eiffel Tower, should be crucial for accurately predicting future images of the same class, even if they were taken under novel weather conditions or camera poses. Hence, we find this justification to be unsatisfactory.

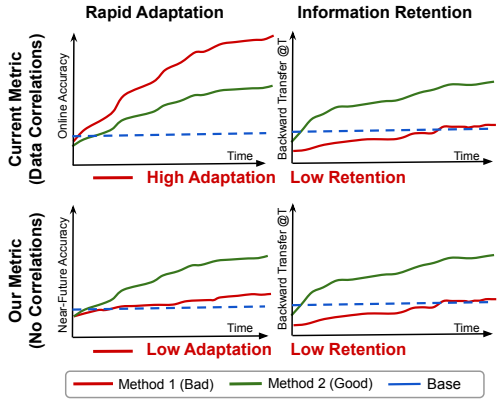

Figure 1. **Effect of Spurious Correlations.** Current methods in OCL perform well on rapid adaptation but suffer from abysmal information retention (top). We show that these trends are due to idiosyncrasies in the data stream such as label correlations that favor methods that inadvertently overfit to the latest data. Using our proposed metric that removes these correlations (bottom), different sets of methods achieve good performance, with high information retention being key for achieving high adaptability.

In this paper, we demonstrate that the conventional metric of measuring online accuracy may lead to misleading conclusions about the adaptation capabilities of OCL algorithms due to the idiosyncrasies of the data stream. To support our claim, we first demonstrate that a simple blind classifier that relies solely on spurious label correlations in the stream achieves unrealistically high online accuracy on large-scale OCL datasets. Furthermore, we show that this behaviour can be exhibited by deep OCL algorithms, where we propose *OverAdapt* which achieves state-of-the-art performance in online accuracy despite the very poor perfor-

*authors contributed equally; order decided by a coin flip.
†equal supervision

mance in information retention. We confirm the findings on two large-scale OCL datasets collected from different sources. This indicates that the problem is not limited to a specific dataset, but may be a general concern.

To overcome this issue, we introduce a new metric called *near-future accuracy* that measures the adaptation capabilities of OCL algorithms by evaluating the accuracy on samples after a shift of $\mathcal{S}$ steps. We choose the minimum shift $\mathcal{S}$ which ensures that the samples no longer have spurious label correlations with the current training data. Choosing the minimum value helps minimize any distribution drift between the training and evaluation samples. The case $\mathcal{S} = 0$ recovers the online accuracy metric in case of no spurious correlations. We observe a significant drop in performance on this metric compared to their online accuracy with several state-of-the-art OCL algorithms. This suggests that these algorithms inadvertently rely on label correlations in the stream for predictions. On the contrary, we find that algorithms designed primarily to improve information retention perform exceptionally well in our evaluation strategy. This suggests that a better generalization can be achieved by effectively retaining and reusing past information. Figure 1 illustrates that the best algorithms are those that have little discrepancy between information retention and adaptation. These results challenge current beliefs and demonstrate the importance of using a more precise measure of adaptation in evaluating OCL algorithms.

Overall, our contributions can be summarized as follows:

- We demonstrate that popular metrics like online accuracy can be unreliable in measuring rapid adaptation due to the presence of spurious label correlations in OCL datasets.
- We introduce a novel baseline, OverAdapt, which achieves high online accuracy by relying on label correlations, despite forgetting almost all previously seen data.
- We propose a novel metric that evaluates the adaptation capabilities of OCL algorithms by measuring accuracy on near-future samples. We select the smallest value of $\mathcal{S}$ that removes label correlations and use this metric to evaluate state-of-the-art OCL methods.
- Our findings show that existing OCL algorithms perform poorly when evaluated using our proposed evaluation strategy. Algorithms which prevent catastrophic forgetting perform significantly better, suggesting that better adaptation performance can be achieved by retaining and reusing past seen information.

Our results challenge the current emphasis on adapting to the latest samples and demonstrate the importance of seeking a more precise measure of adaptation. We note that our contributions are robust, with our findings valid across various OCL approaches, optimization strategies, and large-scale datasets from different sources.

Table 1. **Properties of OCL Benchmarks.** We list benchmarks in the field of OCL along with the works which introduced them. We compare them across six properties: Realistic Data Streams (DO), No Storage Constraints (SC), Rapid Adaptation (RA), Information Retention (IR), Restricts Computational Budgets (CB), Evaluates on long stream sizes (LB), Leverages Pretrained Models (PT). Note that ✓is better than ✗in all columns.

DataOrder	Benchmark	DO	SC	RA	IR	CB	LB	PT
Task-Inc.	A-GEM[9]	✗	✗	✗	✓	✗	✓	✗
Class-Inc.	MIR[1]	✗	✗	✗	✓	✗	✓	✗
	DER[4]							
Blurry	GSS[2]	✗	✗	✗	✓	✗	✓	✗
	RM [3]							
	CLIB[18]							
Natural	CLEAR[20]	✓	✗	✓	✓	✗	✗	✓
	BudgetCL[26]	✗	✓	✗	✓	✓	✗	✓
	CLOC[7]	✓	✗	✓	✗	✗	✓	✗
	DelayOCL[10]	✓	✗	✓	✗	✓	✓	✓
	ACM[25]	✓	✓	✓	✓	✓	✓	✓

2. OCL Evaluation: A Review

We start with an overview of benchmarks in OCL across different properties of interest. For a more detailed review of the literature, we refer the reader to [30].

1. Realistic Data Streams (DO). The OCL community has been moving towards increasingly realistic benchmarks in terms of data ordering and task setup. Early works [21, 9] assumed access to which subset of classes a test sample is from (task-incremental setup). Subsequent work removed this constraint, requiring models to predict across all seen classes so far (class-incremental setup) [2, 1]. However, these works used incoming samples from disjoint data distributions with artificial task boundaries, resulting in unrealistic ordering. Recent works [3, 18, 26] improved the ordering by mixing samples from different disjoint tasks. Latest works [7, 25, 20] eliminate the need for generating artificial data streams simply by using real-world timestamps to order the data stream. Our work focuses on these real-world data streams on large-scale datasets.

2. Towards No Storage Constraints (SC). The majority of prior benchmarks on continual learning constrain the problem by prohibiting access to previously received data [19, 17], with only a small portion of data allowed to be stored in memory [21, 8, 4, 2, 18, 20]. This constraint is often justified by two reasons: (i) storage space is expensive and (ii) access to previous data is prohibited due to privacy constraints. However, recent work [25, 26] shows critical shortcomings of these justifications. They show that (a) storage costs are negligible compared to the computational costs of training a model, and (b) simply restricting access to previous data does not address privacy considerations, as samples can be reconstructed from model weights [13] or detectably change the model output [28]. Interestingly, Goel *et al.* [11] shows that removing data from the model can be done by catastrophic forgetting, which is antithetical to the objective of OCL approaches. Hence, we do not impose any memory constraints on our evaluation.

3. Towards Better Information Retention (IR). Interestingly, most of the previous benchmarks in OCL focus on preventing catastrophic forgetting of previously learned information, rather than evaluating the ability of models to rapidly learn new concepts [21, 27, 9, 3, 2, 20]. This emphasis on information retention can be attributed to disjoint set-based data ordering, which often involves little to no change in distribution for most of the stream, making it difficult to evaluate rapid adaptation. However, recent works have shifted their focus to rapid adaptation performance in real-world ordered data streams [7, 10]. In our benchmark, our objective is to maintain an additional high degree of retained knowledge from past data as demonstrated in [25, 6].

4. Towards Enabling Rapid Adaptation (RA). The goal of *online* continual learning algorithms is to rapidly adapt to incoming data from changing distributions. Similarly to past work [25, 7, 10], we place a strong emphasis on achieving this goal with better metrics.

5. Towards Computational Budgets (CB). Continual learning without memory constraints is primarily a problem of achieving high performance with computational efficiency, as retraining from scratch is an ideal solution to the problem. This setting addresses the real-world problem of reducing the cost of updating a given model with incoming data from a stream [25]. We evaluate OCL approaches under the maximum computational budget for fairness [10], testing across two different budgets.

6. Towards Large-Scale OCL (LB). OCL benchmarks have historically focused on learning over samples incoming from a stream over long data streams since the development of GEM [21]. Recent OCL benchmarks [2, 18, 25] have preserved this useful characteristic, while scaling up to larger and more complex data streams. In this work, we use two large-scale OCL benchmarks: *Continual Google Landmarks (CGLM)* [25] and *Continual LOCalization (CLOC)* [7]. CGLM is a landmark classification dataset that consists of a long-tailed distribution with 10,788 classes that simulate images that arrive on a Wikimedia Commons server. CLOC is a dataset focused on geolocation at scale consisting of 713 classes with 39M images simulating images arriving on a Flickr server. As these datasets come from two different sources, the issue of label correlation that we observe is likely to be a repeating pattern across future OCL datasets.

7. Leveraging Pretraining for Effective OCL (PT). Traditional methods in online continual learning start with randomly initialized models [21, 1]. However, continual learning approaches can leverage the abundance of large-scale pretrained models to improve computational efficiency [24, 31, 25]. We focus on an evaluation setting that starts with ImageNet1K pretrained models to allow approaches to leverage pretrained information.

3. Correcting Evaluations of Rapid Adaptation

We delve into our exploration of the limitations of using online accuracy as a metric to measure adaptation to distribution shifts and present our efforts to alleviate the shortcomings of this metric. We start by identifying the label correlations that affect online accuracy, and then we unveil a technique that enables deep networks to artificially achieve high performance on this metric. This highlights the dependence of the networks on these label correlations. Finally, we introduce a novel evaluation metric, near-future accuracy, and a test to detect and remove label correlations allowing for accurate measurement of the effectiveness of OCL methods in rapidly adapting to new distributions.

3.1. Evaluating OCL: A Problem Formulation

In traditional OCL setups with real-world stream orders [7, 25], the learner is fed a continuous stream of training samples over time steps $t \in \{1, 2, \dots, \infty\}$, consisting of an input datapoint $\mathbf{x}_t \in \mathcal{X}$ and its corresponding label $\mathbf{y}_t \in \mathcal{Y}$, where $(\mathbf{x}_t, \mathbf{y}_t) \sim \mathcal{D}_{j \leq t}$. The distribution $\mathcal{D}_j$ can change at any stream timestep. The aim is to train a classifier $f_{\theta_t} : \mathcal{X} \rightarrow \mathcal{Y}$ at any given t, which accurately maps a new sample $\mathbf{x}$ to its label $\mathbf{y}$ while adapting to the latest information and incorporating it with the historical knowledge acquired from previous data.

To evaluate the adaptation ability of f_{θ_t}, online accuracy measures the performance on the next unseen sample, $(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})$, *i.e.*, $a_t = \mathbb{1}\{f_{\theta_t}(\mathbf{x}_{t+1}) = \mathbf{y}_{t+1}\}$. Online accuracy is updated in an online fashion given by $A_t^{RA} = \frac{1}{t}\left(A_{t-1}^{RA} \cdot (t-1) + a_t\right)$. For OCL approaches training deep networks, the running average is updated on a batch $\mathcal{B}$ of the next unseen samples as opposed to a single sample. After the evaluation is carried out, the model f_{θ_i} is updated using the same samples.

The ability of f_{θ_t} to retain information is measured by evaluating its accuracy on an unseen test stream that is similar to the training stream. In particular, per time step t, the model trained at the end of step t is evaluated on all test samples from step 0 to step t. We measure Backward Transfer @ T [7] where T is the last time step of the stream. That is to say, the model at the last step of the stream is evaluated on an increasing test test from step 0 to the end of the stream. However, traditional OCL methods targeting natural distribution shifts have neglected this metric [7, 10].

3.2. Isolating and Quantifying Label Correlations

Our objective is to investigate whether using online accuracy to measure adaptation in OCL suffers from drawbacks due to label correlations. To accomplish this, we introduce a very simple algorithm that leverages label correlations and performs well in terms of online accuracy.

Blind Classifier. We define a blind classifier, similar to prior works [7, 25], as a model that predicts the mode of the last $\mathcal{K}$ samples seen without access to the input images. Formally, at a given new step t, the blind classifier

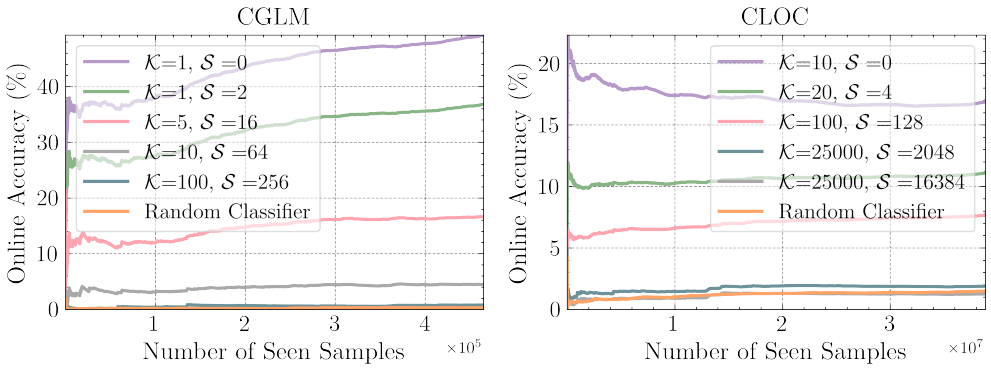

Figure 2. **Blind Classifier.** Performance of a blind classifier on the CGLM (left) and CLOC (right) datasets across varying shift $\mathcal{S}$ towards future for selecting evaluation samples alongside the optimal context window size k for the selected shift. (Section 3.2) A blind classifier achieves unrealistically high accuracy on current evaluation of adaptation (dark blue line, $\mathcal{S}$=0) on both datasets despite being a trivial baseline. (Section 3.5) The accuracy of the blind classifier on incoming samples with increasing shift $\mathcal{S}$ into the future decreases significantly, eventually converging to near-random performance, indicating a lack of label correlations to the current training samples.

predicts the class of $\mathbf{x}_t$ as the mode of the last revealed labels $\cup_{j=1,\ldots,k}\{\mathbf{y}_{t-j}\}$. The optimal context window size $\mathcal{K}$ is selected by a search over different sizes on the hyperparameter search set.

Results. We present the performance of the blind classifier on the online accuracy in Figure 2. For this section, we focus solely on measuring online accuracy (indicated by the topmost purple line). Our results indicate that the blind classifier performs remarkably well achieving an average accuracy of 50% on the CGLM dataset, which consists of 10,788 classes, using only the labels of the last seen training sample ($\mathcal{K} = 1$). Moreover, the blind classifier achieves 17% on the large-scale CLOC dataset consisting of 718 classes with $\mathcal{K} = 10$ achieving the best performance. These findings suggest an unusually high degree of label correlation between the past 1-10 samples and the immediate incoming sample in the data stream, especially given that the blind classifier achieves 50% and 17% accuracy, while the random baseline achieves 0.01% and 2% on CGLM and CLOC datasets, respectively.

Conclusion. Despite having never processed any input images, the blind classifier achieves a remarkably high online accuracy. This is surprising because it was expected to perform no better than a random classifier. However, it achieves this result by exploiting the label correlations present in OCL datasets. This is a critical drawback in using online accuracy to measure adaptation. In the following sections, we will delve deeper into this issue.

3.3. Can Deep Networks Learn Label Correlations?

After demonstrating the effectiveness of the blind classifier in exploiting label correlations in the two large-scale datasets used in OCL settings, a pertinent inquiry arises: Can deep networks inadvertently learn these same label correlations leading to unreasonably high online accuracies? This question warrants investigation to determine to what extent deep OCL algorithms can leverage label correlations and the impact this may have on their performance.

An Overfitting-Based Baseline. One approach for DNNs to exploit label correlations is by overfitting to recent data. We investigate algorithmic choices for designing an OCL method that can overfit the most recent labels in a data stream. Specifically, we make two important design choices for our baseline: (i) we update the model only on recent data to achieve overfitting while sacrificing most of the useful information from the past. To this end, we adopt FIFO sampling, a widely-used technique in OCL, to select batches of samples for training. (ii) We fix the feature representations and train only the last linear layer to prevent degradation of feature representations due to overfitting. To improve the quality of the representations, we use a ResNet50 model pretrained on Instagram1B [22]. We refer to this model as *OverAdapt*. OverAdapt is computationally efficient since training the Linear layer over multiple updates is relatively inexpensive. Moreover, we adopt the fast stream evaluation setting [10], *i.e.* a setting with a strictly limited computational budget. We set the upper bound for the computation to one gradient update per incoming batch of samples for all methods with no restrictions on storage space.

Results. We present the results of the proposed OverAdapt compared to state-of-the-art OCL methods trained using an ImageNet1K pretrained ResNet50 model on the CGLM and CLOC datasets in Figure 3. OverAdapt achieves remarkable performance, surpassing the best-performing method ACM on the CGLM dataset by more than 30%, and outperforming all but the ACM baseline on the CLOC dataset by 10-35%. Notably, OverAdapt achieves an impressive 80% accuracy on the CGLM dataset.

Conclusion. After demonstrating that OverAdapt can effectively leverage label correlations from the data stream, we conclude that traditional design choices like the FIFO buffer in past OCL setups [7] improved overfitting to the next sample. It is worrying that future work or existing works could have inadvertently made similar choices leading to perceived improvement in algorithms when instead it is a subtle case of overfitting to the online accuracy metric. We further ensure the validity of our conclusions by investigating whether the effectiveness of this model is due to

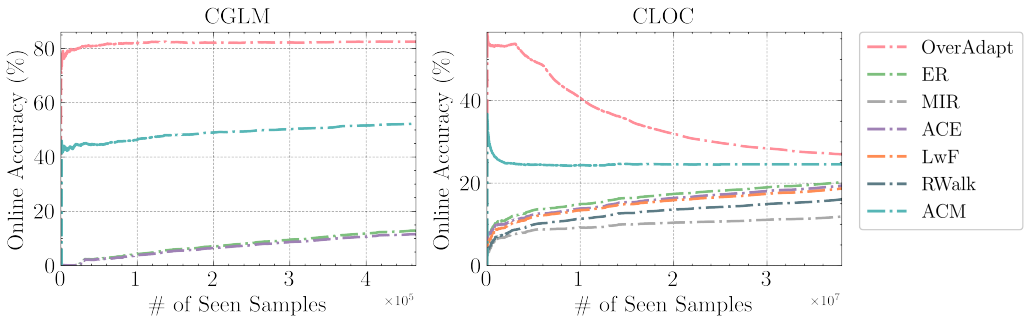

Figure 3. **OverAdapt Classifier.** We compare the adaptation performance of our OverAdapt classifier on CGLM (left) and CLOC (right) datasets using online accuracy as a measure. The three top performing OCL methods on CLOC are benchmarked on the CGLM dataset. We see that OverAdapt outperforms recent OCL methods by a large margin of 10-40% in accuracy.

a good training algorithm or to potential limitations of the evaluation metric. We explore this in the next subsections.

3.4. Exploring Properties of OverAdapt

We investigate the reasons behind the impressive performance of our proposed OverAdapt.

Sensitivity Analysis. We start with the sensitivity study of OverAdapt by analyzing the impact of modifying its two key components: (i) FC-Only training with large-data pretrained initialization and (ii) FIFO Sampling. To study the effect of FC-Only training, we replace it with full-model training using an ImageNet1K-pretrained ResNet50 which has been common in prior OCL work [10]. We also analyze the effect of FIFO sample selection by comparing it with uniform sampling, which has been shown to be a simple but effective strategy in limited-compute regimes [26]. We evaluate their performance in terms of online accuracy and information retention at time T after training on the entire stream following prior art [7] as discussed in Section 3.1.

Results. Our results are shown in Figure 4 (left & center).

OverAdapt. Strikingly, our analysis reveals that **OverAdapt** (✓FIFO ✓FC-Only) achieves remarkable performance in rapid adaptation as measured by online accuracy achieving **82%** and **27%** on CGLM and CLOC, respectively. However, this seemingly impressive performance is coupled with an abysmally low information retention when measured by Backward Transfer with **OverAdapt** performing with 5% and 1% accuracy on CGLM and CLOC, respectively. These results demonstrate that **OverAdapt** is not a superior OCL algorithm but rather a product of a flawed rapid adaptation evaluation.

FIFO Sampling. Our analysis comparing sampling strategies reveals that models trained using the FIFO sample selection strategy with both **FC-Only training** (✓FIFO ✓FC-Only) and **full-model training** (✓FIFO ✗FC-Only) achieve significantly higher accuracy of **82**/**78**% and **27**/**24**% on CGLM and CLOC, respectively, on online accuracy metric compared to those trained using uniform sampling-based with **FC-Only training** (✗FIFO ✓FC-Only) and **full-model training** (✗FIFO ✗FC-Only). This clearly indicates that FIFO sampling is primarily responsible for the ability to leverage the latest label correlations. However, models trained with FIFO sampling achieve significantly lower accuracy of **5**/**7**% and **1**/**1**% on CGLM and CLOC, respectively, in terms of information retention as measured by backward transfer in the last timestep, performing close to a random classifier. This highlights their poor representation quality and confirms that models using the FIFO strategy simply overfit to the latest samples.

FC-Only Training. Comparing **FC-Only training** (✓FIFO ✓FC-Only) to **full-model training** (✓FIFO ✗FC-Only) under FIFO sampling, we observe that **FC-Only training** achieves a higher accuracy of **82**% and **27**% on CGLM and CLOC, respectively, in online accuracy. However, for information retention, both models perform poorly due to the impact of FIFO sampling. In contrast **full-model training** under uniform sampling (✗FIFO ✗FC-Only) achieved a higher accuracy of **34**% and **14**% on CGLM and CLOC, respectively, for online accuracy, compared to **full-model training** (✗FIFO ✓FC-Only). Additionally, **full-model training** achieved **42**% and **18**% accuracy on CGLM (comparable to **full-model training**) and CLOC (higher than **full-model training**), respectively, in information retention. In general, it is not conclusive whether FC-Only training could substitute full model training, as it performed significantly better in some cases and significantly worse in others.

Conclusion. Our experiments highlighted critical drawbacks in assessing the rapid adaptation of deep models using online accuracy, due to spurious label correlations in the data stream. Sensitivity analysis revealed that both FIFO sampling and FC-Only training are critical components that helped OverAdapt leverage label correlations where FIFO sampling being the primary contributor. Furthermore, our findings suggest that information retention is a useful indicator for identifying drawbacks of OCL algorithms, such as inadvertent overfitting.

3.5. Removing Correlations to Evaluate Adaptation

To address the limitations of online accuracy, we propose a new metric, near-future accuracy, which involves evaluating the model on near-future samples rather than immediate next samples. This can mitigate the adverse effects of label

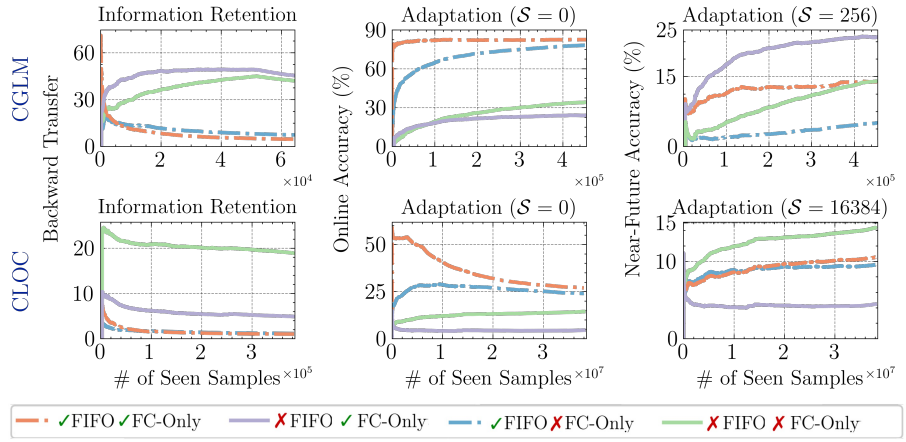

Figure 4. **Sensitivity Analysis of OverAdapt Classifier.** We illustrate the performance of various components of our OverAdapt classifier on CGLM (top) and CLOC (bottom) datasets measuring: (Section 3.4) information retention (left) and rapid adaptation capability in terms of online accuracy ($\mathcal{S}=0$, center) (Section 3.5) rapid adaptation performance in terms of online accuracy ($\mathcal{S}=0$, center) with our proposed near-future accuracy ($\mathcal{S}=256$ in CGLM and , $\mathcal{S}=16384$ in CLOC). Note that information retention (left) is independent of the shift $\mathcal{S}$. The results demonstrate that introducing the shift removes label correlations, leading to dramatic loss in adaptation performance for our OverAdapt method while the uniform baseline stays unaffected.

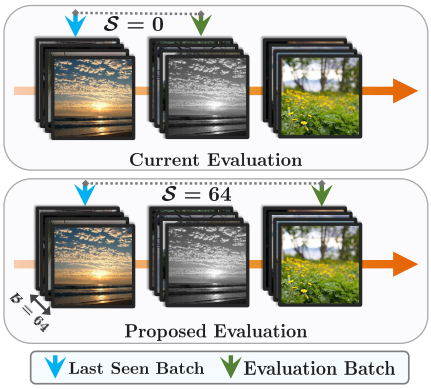

Figure 5. **Our Evaluation Method.** We compare the samples selected for evaluating adaptation performance by our evaluation method. The blue and green arrows indicate the latest batch received from the stream for learning and the samples selected for evaluation respectively. Current evaluation proposes to compute accuracy on the immediate incoming samples ($\mathcal{S}=0$), *i.e.* online accuracy, while we propose to evaluate on the closest future samples ($\mathcal{S}=64$), *i.e.* near-future accuracy, which is free from spurious correlations with the seen batch of images.

correlations on the performance by obtaining a more accurate assessment of the adaptation ability of OCL algorithms.

Formulation of Proposed Evaluation. As illustrated in Figure 5, we introduce a shift $\mathcal{S}$ that represents the number of samples in the future, after which we evaluate the performance of the model. Instead of evaluating on the immediate next sample $(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})$, we use the next near-future sample after the smallest shift $\mathcal{S}$, namely $(\mathbf{x}_{t+1+\mathcal{S}}, \mathbf{y}_{t+1+\mathcal{S}})$, so that it has no label correlation with them. We select the smallest shift $\mathcal{S}$ instead of an arbitrarily large shift $\mathcal{S}$ so that the test sample comes from the same distribution as the latest seen train samples. We calculate the online accuracy as before, where $\hat{a}_t$ is the running average of the accuracy calculated by first calculating the accuracy $a_t = \mathbb{1}\{f_{\theta_t}(\mathbf{x}_{t+1+\mathcal{S}}) = \mathbf{y}_{t+1+\mathcal{S}}\}$ and then updating the running average using the formula $A_t^{RA} = \frac{1}{t}\left(A_{t-1}^{RA} \cdot (t-1) + a_t\right)$.

How To Set the Shift $\mathcal{S}$? The shift $\mathcal{S}$ is the smallest value in the future that has no label correlation with the last seen train samples. We select the smallest $\mathcal{S}$ so that the blind classifier performs similarly to a random classifier. As shown in Figure 2 for both CGLM (left) and CLOC (right) datasets, we observe that increasing the shift $\mathcal{S}$ decreases the accuracy of the blind classifier down to a similar performance of a random classifier. In particular, at a shift of $\mathcal{S}=256$ for CGLM and $\mathcal{S}=16384$ for CLOC, the label correlations are effectively eliminated. Therefore, we use these shift values for our proposed evaluation strategy. We evaluate OverAdapt and its variants on rapid adaptation with our proposed near-future accuracy and compare it with the previously adopted online accuracy ($\mathcal{S}=0$).

Results. Figure 4 presents the results of OverAdapt and its variants on rapid adaptation using two metrics: online accuracy (center) and near-future accuracy (right). When studying FIFO sampling with **FC-Only training** (✓FIFO ✓FC-Only) and with **full-model training** (✓FIFO ✗FC-Only), we observe a significant drop in the performance of the methods between the two metrics that overfit using the FIFO sampling by about 70% in CGLM and 20-40% on CLOC datasets, while the performance of uniform sampling for both **FC-Only training** (✗FIFO ✓FC-Only) and **full-model training** (✗FIFO ✗FC-Only) remains similar. Interestingly, the trends reverse, with the uniform sampling variants outperforming the FIFO sampling strategy by a large margin. Furthermore, we observe that FC-Only training no longer consistently improves rapid adaptation, proving to be useful in some cases on the new evaluation method.

Conclusion. We propose an evaluation using near-future accuracy which removes label correlations. OverAdapt performs poorly after removing label correlations, compared

to a uniform baseline, emphasizing the importance of developing OCL algorithms that handle rapid adaptation without overfitting. Additionally, there is no surprising discrepancy between near-future accuracy and information retention across methods, unlike online accuracy. Our strategy allows for better evaluation of OCL algorithms on real-world ordered datasets.

4. OCL Evaluation by Near-Future Accuracy

In this section, we provide a comprehensive analysis of various OCL approaches under near-future accuracy. We incorporate the latest advancements in setup design into our benchmark and take computational constraints into account for fair comparison. Following the prior art [25, 26], our study assumes no memory constraints, but rather sets constraints on computational budget following [10, 25, 26]. As addressed in prior work, budgeted computation per time step implicitly imposes a limited memory access.

4.1. Experimental Setup

Datasets and Metrics. We employ two large-scale online continual learning datasets, CGLM with a total of 460K images and CLOC compromising of 39M images in a stream, both of which contain natural distribution shifts [7]. That is to say, the images are temporally ordered. We measure rapid adaptation performance using the average accuracy on near-future samples, as described in Section 3.5, with a shift $\mathcal{S} = 256$ for CGLM and $\mathcal{S} = 16384$ for CLOC. The choice of both was made such that the blind classifier discussed in Section 3.2 achieves a random accuracy. Additionally, we measure information retention using the Backward Transfer metric following CLOC [7].

Model and Optimization. In all experiments, we use a ResNet50 with ImageNet1K initialization unless otherwise specified. We adopt the optimization procedure from [10] and train all models using an SGD optimizer, a fixed learning rate of 0.005, and a weight decay of 10^{-4}. For both training and evaluation, we batch incoming samples with a batch size $\mathcal{B}$ of 64 for CGLM [25] and 128 for CLOC [10].

Compared OCL Approaches. We evaluate five online continual learning methods using our proposed near-future accuracy. For consistency, we use uniform sampling as the sampling strategy, since it was shown in Section 3.4 that it strikes a good balance between adaptation and retention. Again, in uniform sampling, the training batch is constructed by uniformly and sampling $\mathcal{B}$ samples from all past stored data. The Appendix provides a comprehensive analysis of other sampling techniques.

***ER (Replay Only)*[26].** ER (Replay Only) is a simple approach that stores the incoming samples and trains the model on a batch sampled from the set of stored samples. This is the leading method in offline budgeted CL [26].

***FC-Only*.** FC-Only, referred to as OverAdapt (✗FIFO ✓FC-Only) in Section 3.4, is an adapted version of ER (Replay Only) that uses a ResNet50-I1B feature extractor and trains only the last linear layer. This method has been shown to be computationally efficient [26].

***ACE*[5].** We replace the CrossEntropy loss in ER (Replay Only) with the ACE Loss proposed in to isolate the effect of ACE loss under the proposed evaluation setting.

***ACM*[25].** ACM extracts features from a pretrained ResNet50 model (trained on ImageNet1k) and continually updates a kNN classifier, with $k = 2$ and cosine distance as distance metric, over those features. Additionally, we note that ACM does not train a deep network, thus incurring minimal computational cost. This sets the minimal accuracy requirements for OCL methods. We compare ACM with other popular fixed-feature extractor based methods like NCM [23, 24, 16] and SLDA [14] in the Appendix.

***CosineFC*[15].** This approach replaces the linear layer of ResNet50 in the ER (Replay Only) baseline with a layer that computes the cosine distance between weights and features. We only perform this modification as distillation has been shown to be computationally ineffective [26].

By comparing these approaches under our proposed near-future accuracy along with information retention, we aim to provide insights into the strengths and limitations of current OCL methods.

4.2. Results: Varying Stream Speeds

Following prior work on fair computationally normalized evaluation of online continual learners [10], we evaluate the performance of OCL methods in two different time constraints, namely slow stream and fast stream scenarios.

Slow Stream. Slow stream is the scenario in which the computational budget for learning is not strictly limited. We enforce all methods to perform the equivalent of 10 and 5 model updates per time step on CGLM and CLOC, respectively. Methods that require a significant overhead will thus perform a fewer number of model updates such that it matches the corresponding computational budget of 10 and 5 updates on the respective datasets. Note that ACM requires negligible training cost, and therefore OCL methods must at least outperform ACM to be considered useful.

Results. In Figure 6 (first row), we present the performance of various methods in terms of near-future accuracy for both CGLM and CLOC datasets. CosineFC, originally proposed for offline continual learning, achieves the highest accuracy with 38% and 16% accuracy on CGLM and CLOC, respectively. Furthermore, CosineFC also shows impressive information retention performance, achieving an accuracy of 77% in CGLM. As expected, FC-Only which only finetunes the last linear layer, has the worst retention capabilities due to the restricted access to full-model updates.

Conclusion. Our evaluation revealed that methods not explicitly designed for adaptation in the OCL setup perform well, indicating that retaining and reusing previously seen information is crucial for achieving better generalization, and that retention and adaptation are closely linked.

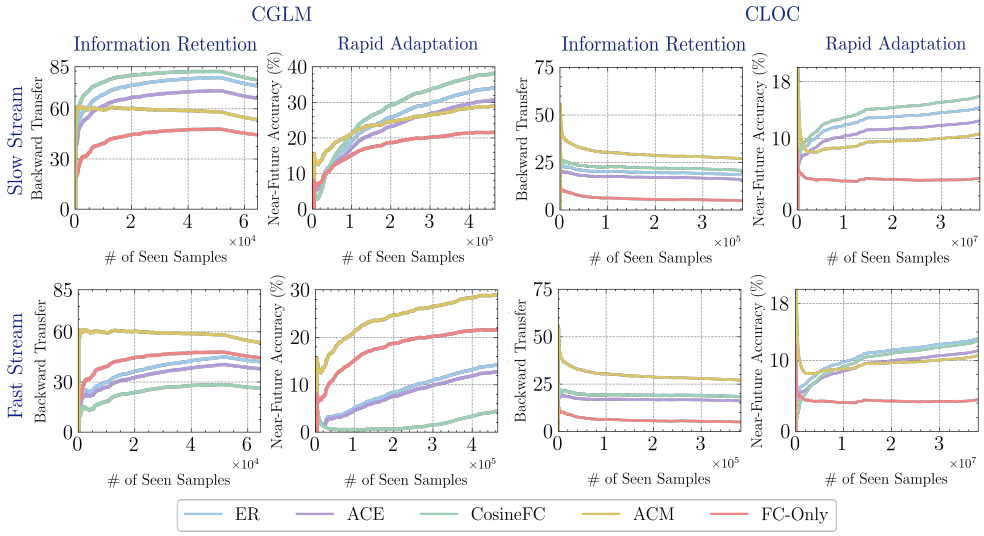

Figure 6. **Performance of Methods on Near-Future Accuracy**. We evaluate CL methods in slow and fast streams similar to [10]. Near-future accuracy, which eliminates spurious label correlations, shows that retaining and reusing past information is crucial for better generalization and that algorithms not designed for adaptation in OCL can perform well. Hence retention and adaptation are closely linked.

Fast Stream. Fast stream setting refers to the scenario where the computational budget for learning is limited. That is to say, the stream presents samples at a fast rate that provides limited time for learners to train [10]. To evaluate the performance of OCL methods under this setting, we restrict the methods to only use a computation equivalent to one model update per incoming batch for both datasets.

Results. As shown in Figure 6 (second row), our results suggest that the two simplest baselines, namely ACM and FC-Only, achieve the highest accuracy on CGLM with 29% and 21%, respectively. On the other hand, ACM achieves the highest information retention of 27% while simple ER and CosineFC achieve the highest near-future accuracy of around 13%. These results indicate that the simple baseline methods outperform the more advanced OCL methods in both adaptation and information retention in this setup. This highlights that current OCL methods still have a long way to go to achieve satisfactory performance in the fast stream setting where compute is extremely limited.

Conclusion. Under our proposed evaluation approach, in which spurious label correlations are removed, the fast stream setting with its limited computational budget shows that simple baselines can outperform more advanced OCL methods in both adaptation and retention. This finding emphasizes the need to develop OCL methods that can adapt rapidly in severely budgeted settings.

4.3. Further Discussion

Our benchmark of OCL methods on large-scale datasets highlights two key observations. First, the best-performing online adaptation methods often suffer from forgetting due to overfitting to label correlations, challenging the assumption that the performance gap between adaptation and retention is an inherent learning problem [6]. Our proposed near-future accuracy helps reveal label correlations and identifies whether the underlying algorithm is inadvertently exploiting them. Second, methods with better retention properties tend to generalize better and achieve improved performance on near-future accuracy in relaxed and restricted computational settings [12, 29]. This aligns well with the domain generalization literature [12] and suggests that proper training on previous samples can lead to features that generalize well across domain shifts. Finally, we consistently observe that, under limited computational settings, simpler methods outperform their more computationally involved ones.

5. Conclusion

Our work proposes a new measure to evaluate adaptation in OCL algorithms and highlights its limitations compared to the standard online accuracy evaluation. We found that many current OCL methods overfit to idiosyncrasies of the stream rather than genuinely adapting to new data, as revealed by poor information retention. Using our proposed metric, near-future accuracy, we observed that algorithms with good retention also had better generalization and adaptation capabilities. Our proposed evaluation can serve as a sanity check for future OCL algorithms to ensure that they are not incorrectly learning spurious label correlations.

6. Acknowledgements

This work was supported by the King Abdullah University of Science and Technology - Office of Sponsored Research (OSR) under Award No. OSR-CRG2021-4648, SDAIA-KAUST Center of Excellence in Data Science, Artificial Intelligence, and UKRI grant: Turing AI Fellowship EP/W002981/1. We thank the Royal Academy of Engineering and FiveAI for their support. SL from Meta AI has no relationships with the mentioned grants. AP is funded by Meta AI Grant No. DFR05540. AB acknowledges the Amazon Research Award.

References

[1] Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3

[3] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[4] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[5] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations (ICLR)*, 2022. 7

[6] Zhipeng Cai, Vladlen Koltun, and Ozan Sener. Improving information retention in large scale online continual learning. *arXiv preprint*, 2022. 1, 3, 8

[7] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. Online continual learning with natural distribution shifts: An empirical study with visual data. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 4, 5, 7

[8] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[9] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 3

[10] Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip HS Torr, and Bernard Ghanem. Real-time evaluation in online continual learning: A new paradigm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 4, 5, 7, 8, 14

[11] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640v3*, 2023. 2

[12] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *International Conference on Learning Representations*, 2021. 8

[13] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *In Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[14] Tyler L. Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2020. 7, 15

[15] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[16] Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. *arXiv preprint arXiv:2210.04428*, 2022. 7, 15

[17] 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. In *PNAS*, 2017. 2

[18] Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2017. 2

[20] Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The clear benchmark: Continual learning on real-world imagery. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[21] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 3

[22] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *European conference on computer vision (ECCV)*, 2018. 4

[23] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *TPAMI*, 2013. 7, 15

[24] Oleksiy Ostapenko, Timothee Lesort, Pau Rodríguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay. In *CoLLAs*, 2022. 3, 7, 15

[25] Ameya Prabhu, Zhipeng Cai, Puneet Dokania, Philip Torr, Vladlen Koltun, and Ozan Sener. Online continual learning without the storage constraint. *arXiv preprint arXiv:2305.09253*, 2023. 2, 3, 7

[26] Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet Dokania, Philip HS Torr, Ser-Nam Lim, Bernard Ghanem, and Adel Bibi. Computationally budgeted continual learning: What does matter? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 7

[27] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), 2017. 3

[28] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE S&P*, 2017. 2

[29] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *International Joint Conference on Artificial Intelligence*, 2021. 8

[30] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *ArXiv*, abs/2302.00487, 2023. 2

[31] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?
Appendix

A. Effect of Sampling Strategies

In Figure 4 of the manuscript, we present the baseline ER for two sampling strategies, namely FIFO and uniform. FIFO selects the latest seen samples to train on, whereas Uniform simply randomly and uniformly samples a set of previously seen samples to train on. Mixed sampling is a mix of both FIFO and Uniform, where half of the batch is constructed using FIFO sampling, and the other half using Uniform sampling.

A.1. CGLM

The results for various samplers for both Online Accuracy (Figure 7) and Near-Future Accuracy (Figure 8) on CGLM dataset are summarized below:

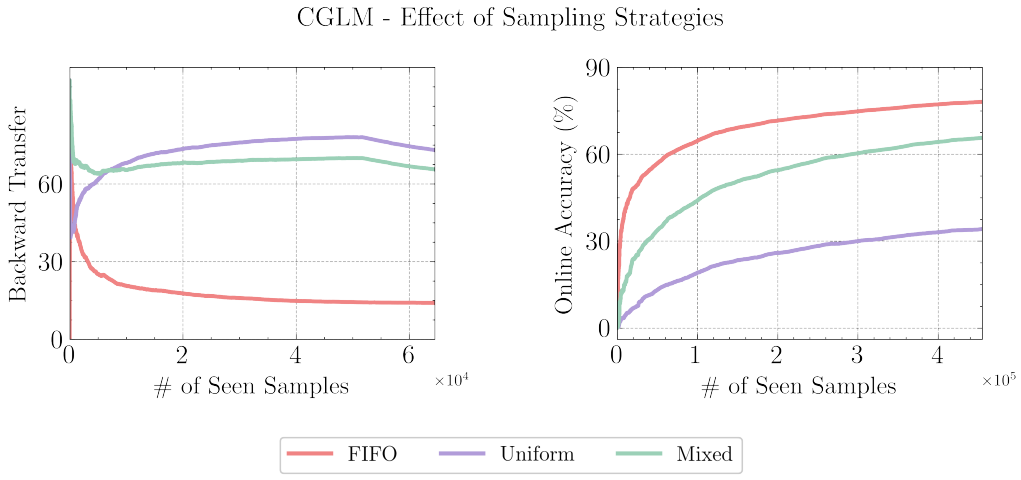

Figure 7. **Effect of Sampling Strategy on Online Accuracy.** In terms of online accuracy, FIFO sampling which focuses on the latest samples performs best where uniform performs the worst and mixed performs somewhere in the middle. In terms of retetion however, the order is reversed.

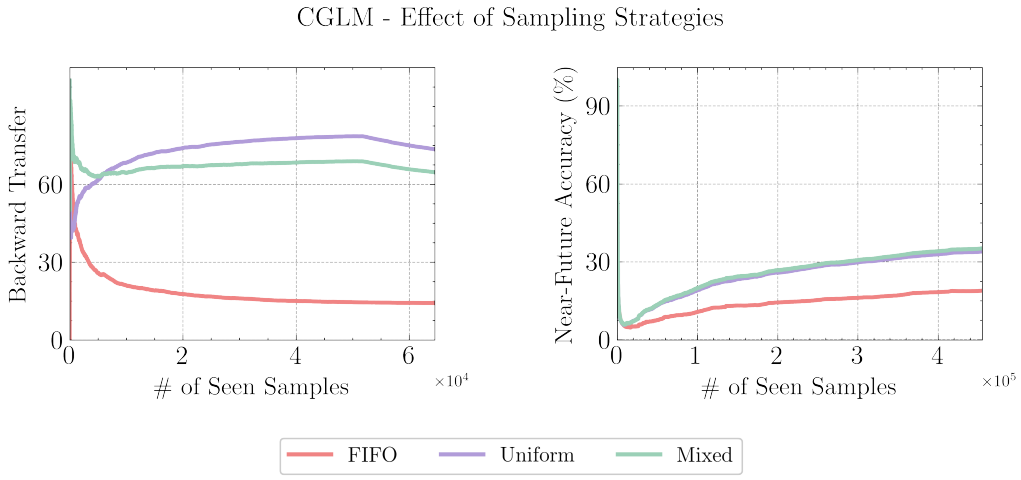

Figure 8. **Effect of Sampling Strategy on Near-Future Accuracy.** In terms of near-future accuracy, Uniform and Mixed sampling perform almost the same, however FIFO sampling is no where close to them. In terms of retention, Uniform still takes the lead.

Conclusion. Interestingly, mixed sampling is competitive with uniform sampling in near future accuracy, but performs worse in information retention.

A.2. CLOC

The results for various samplers for both Online Accuracy (Figure 9) and near-future accuracy (Figure 10) on CLOC dataset are summarized below.

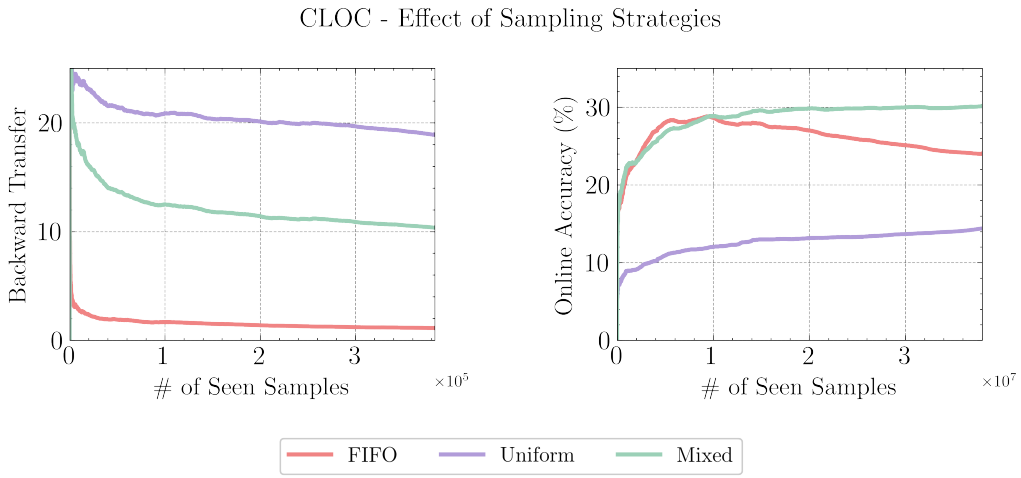

Figure 9. **Effect of Sampling Strategy on Online Accuracy.** Unlike what was observed for CLGM, for CLOC, the mixed sampling performs the best in terms of online accuracy followed followed by FIFO and then Uniform. However, Uniform still performs the best in terms of information retention with a large gap with a large margin.

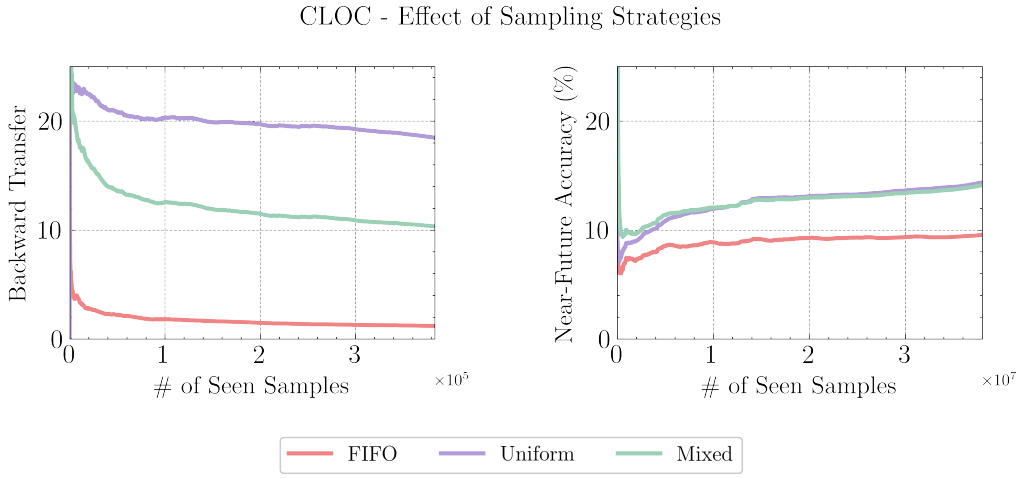

Figure 10. **Effect of Sampling Strategy on Near-Future Accuracy** In terms of near-future accuracy, mixed outperforms other samplers by a significant margin. However, it achieves significantly worse performance compared to uniform sampling in terms of information retention.

Conclusion. Mixed sampling is competitive with uniform sampling in near future accuracy, however, its information retention capabilities are half that of the ER baseline.

B. Sensitivity Analyis: Learning Rate and Weight Decay

In OCL, changing hyperparameters across datasets is uncertain as the stream might have a significant distribution shift from the pretrain. Hence, we use the hyperparameters for our model from Ghunaim *et al*. [10] for all our experiments. However, how do the selected hyperparameters transfer to CGLM is an interesting question. We demonstrate the sensitivity of the hyperparameters: learning rate and weight decay below:

B.1. Sensitivity to Weight Decay

Results and Conclusion. We present our results in Figure 11. We conclude that weight decay has minimal effect on the performance of the ER (Replay Only) method.

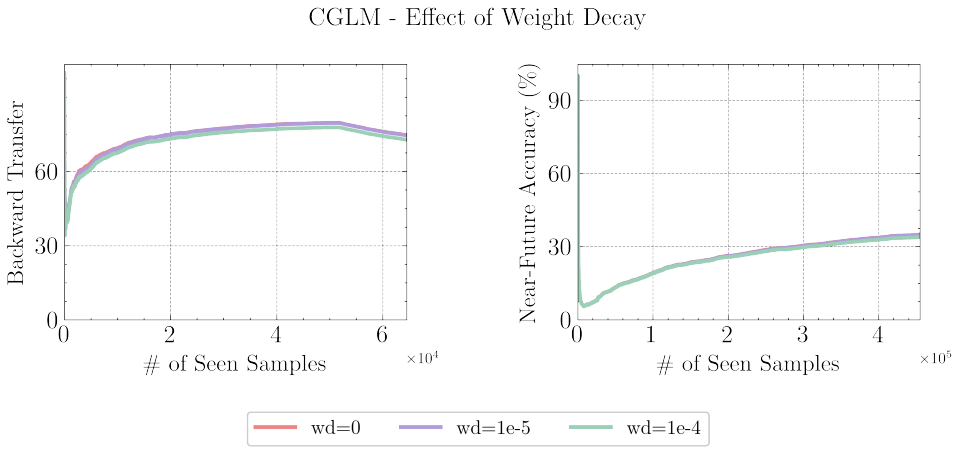

Figure 11. **Sensitivity to Weight Decay on ER (Replay Only).** Weight decay seems to have minimal effect on both near-future accuracy and backward transfer.

B.2. Sensitivity to Learning Rate

Results and Conclusion. We present our results in Figure 12. An order magnitude change in learning rate in either direction leads to a decrease in both information retention performance and near-future accuracy. The learning rate of 0.005 transfers well to CGLM dataset in terms of both near-future accuracy and information retention (Backward Transfer).

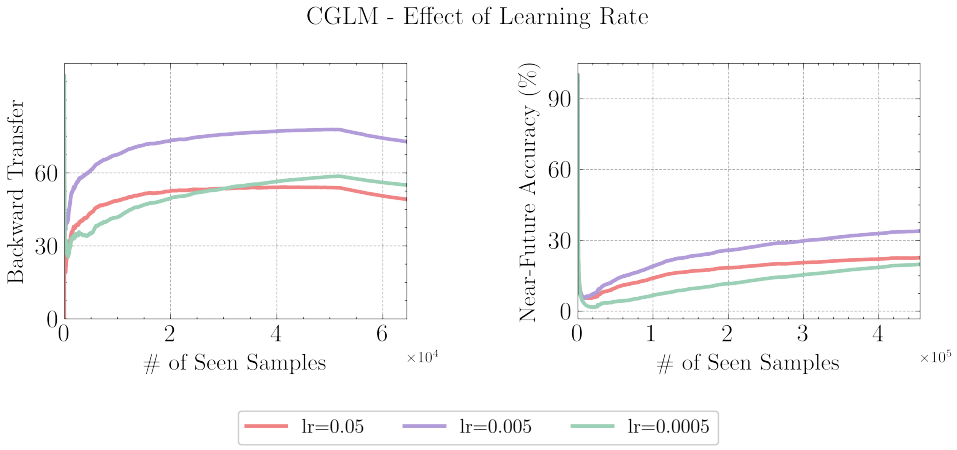

Figure 12. **Effect of Learning Rate on the ER Baseline on CGLM.** Both near future-accuracy and backward transfer are highly sensitive to the selection of the learning rate.

C. Fixed-Feature Extractor Based Methods: NCM, SDLA, and ACM

In this section, we compare ACM with other popular fixed-feature extractor based methods like NCM [23, 24, 16] and SLDA [14]. The results for running NCM, SLDA, and ACM on CGLM dataset are shown in Figure 13 where we see that ACM performs best in terms of backward transfer, online accuracy, and near-future accuracy.

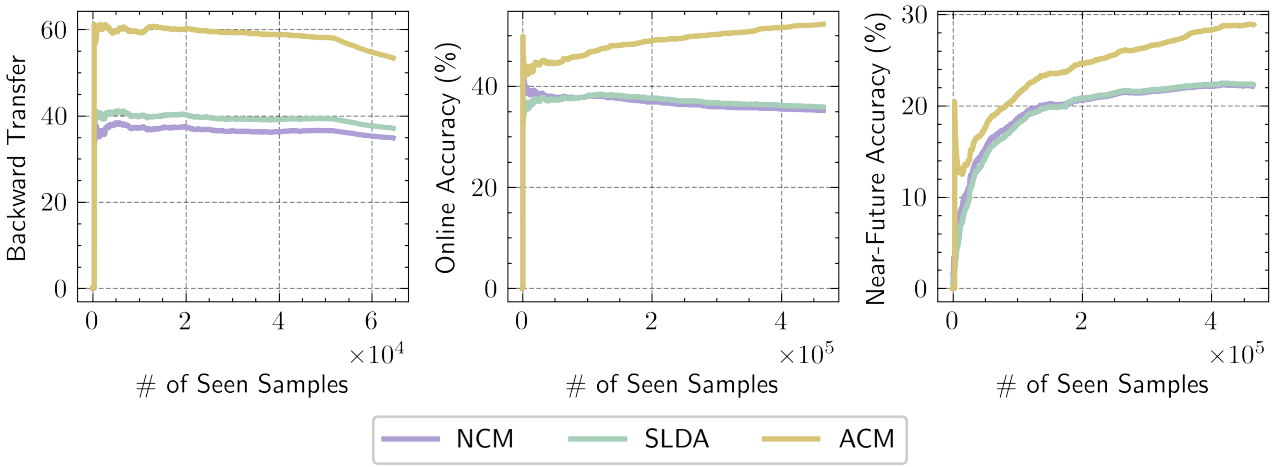

Figure 13. **Fixed-Feature Extractor Based Methods.** When compared to NCM and SLDA, ACM performs the best in terms of Backward Transfer, Online Accuracy and Near-Future Accuracy metrics.

D. Limitations & Future Directions

Limitations. While our work discovers interesting phenomena in OCL, it has important limitations:

- **Is near-future accuracy definitively measuring adaptation?** It is unclear why our proposed evaluation approach definitively measures rapid adaptation, and further investigation is needed to determine the problem exists in future OCL scenarios.
- **Dependency on shift $\mathcal{S}$.** Our proposed metric, near-future accuracy, depends on the calculated value (fixed) for the shift, $\mathcal{S}$, however in reality the stream label correlations could be dynamically changing with the stream and an adaptive value of S would be required in that case.
- **Mitigating Monitoring Costs.** In our experiments, we did not explicitly store the model itself as we were able to access future samples solely for evaluation purposes. In practice, a larger memory allocation would be necessary. To reduce storage costs we recommend performing this evaluation periodically instead of on every incoming sample to reduce storage requirements.
- **Why use a new metric instead of changing the data stream?** A question that might arise is why use a new metric if we could simply remove the correlated samples? The data streams used in our work are *naturally* ordered by timestamps, therefore changing the datasets would make it less natural or even complete unnatural if the correlation is long enough. Therefore, in our work we propose a new metric which could address the label correlations without modifying the datasets.

Future Directions. Our work leads to some interesting questions and problems to be explored:

- **Beyond Label Correlations.** We currently only measure and remove correlations from labels $p(y)$, while covariate correlations $p(X)$ are harder to reliably isolate, requiring further investigation.
- **Why aren't information retention and adaptation at odds?** Our results suggest that there is still a lot of room for improvement in OCL methods, as we are far from the pareto frontier where information retention and rapid adaptation are at odds.