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

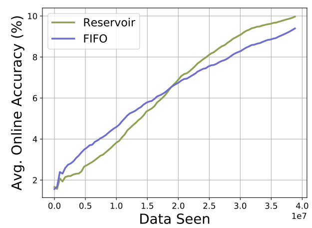
Abstract

We present additional experimental results in this supplement, which we skipped in the main text due to space limitations. In Section 1, we present an additional study on the replay buffer update strategies. In Section 2, we present an additional study on the impact of batch sizes, specifically on the supervised learning counterpart of geolocalization, and the setting of online continual learning without replay buffer.

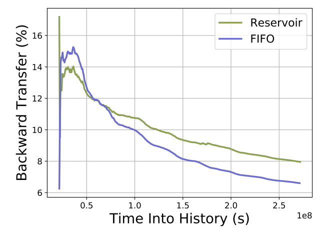
1. Effect of replay buffer update strategies

As mentioned in Sec. 5.1 of the main paper, we choose the First-In-First-Out (FIFO) buffer for experience replay. To demonstrate the effect of changing replay buffer update strategies, we compare FIFO with the reservoir replay buffer [1], which has been applied to offline continual learning. Reservoir buffer constructs a random iid. subset of the data seen so far using an iterative process. For the t^{th} example from the data stream, the reservoir buffer first generates a random number r uniformly sampled from 1 to t. And if r is not greater than the replay buffer size R, the r^{th} example in the replay buffer will be replaced by the t^{th} example from the data stream.

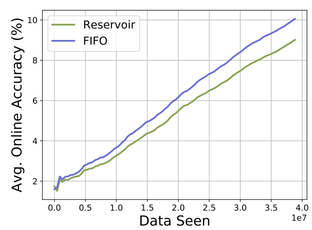
In order to evaluate the impact of buffer strategy, we train models using FIFO and reservoir buffer. We use the cosine learning rate schedule, and two different replay buffer sizes, 40 thousand and 4 million. Fig. 1 demonstrates the average online accuracy (Fig. 1(a) and 1(c)) and backward trainsfer (Fig. 1(b) and 1(c)) of models trained with different replay buffer strategies. We can see that FIFO and reservoir had comparable performance, both in terms of learning efficacy, and information retention. Since the impact of buffer strategy is not signficant, we choose FIFO in our experiments due to its simplicity.



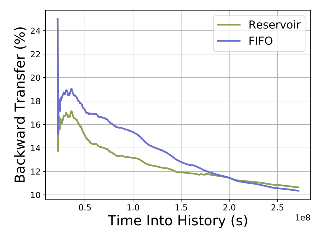
(a) Average online accuracy (↑). Replay buffer size = 40 thousand.



(b) Backward transfer at final time step H (↑). Replay buffer size = 40 thousand.



(c) Average online accuracy (↑). Replay buffer size = 4 million.



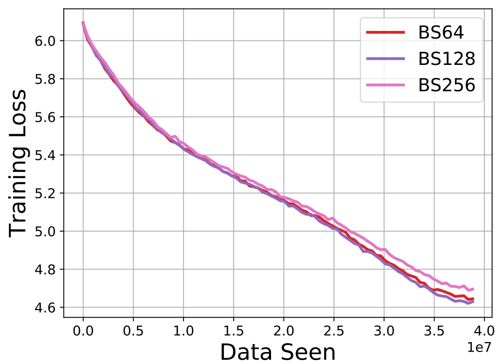
(d) Backward transfer at final time step H (↑). Replay buffer size = 4 million.

Figure 1: **FIFO vs reservoir buffer, with different replay buffer sizes.** The arrows in the caption of subfigures point towards better performance. Overall, FIFO and reservoir had comparable performance both in terms of learning efficacy and information retention. Reservoir performed slightly better (Fig. 1(a) and Fig. 1(b)) with 40 thousand replay buffer size, but slightly worse (Fig. 1(c) and Fig. 1(d)) with 4 million replay buffer size.

2. Additional Analysis for Batch Size

2.1. Batch size effect to supervised learning

In supervised learning, it is a common heuristics to multiply the batch size and learning rate by the same factor and recover similar learning dynamics. In contrast, our results show a strong negative effect with increasing batch sizes in online continual learning (OCL). We ask the question, is this unexpected behavior due to the data or the nature of OCL? We train supervised learning models with batch sizes of 64, 128, and 256, respectively, using shuffled data to answer this question. Moreover, we set the learning rates to 0.0125, 0.025, and 0.05, respectively. All models are trained on our dataset for one epoch. As shown in Fig. 2, all models had similar training loss curves (Fig. 2(a)) and average validation accuracy (Fig. 2(b)). Unlike in the OCL case, increasing batch sizes from 64 to 256 was not harmful to supervised learning. It even slightly improved the validation accuracy. Hence, the batch size effect is due to the nature of the OCL problem.



(a) Training loss

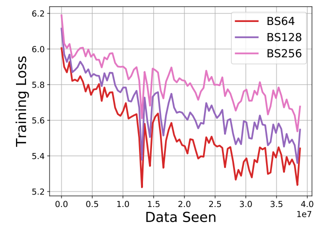
Batch size	Validation accuracy
64	16.61
128	17.07
256	17.21

(b) Validation accuracy (↑)

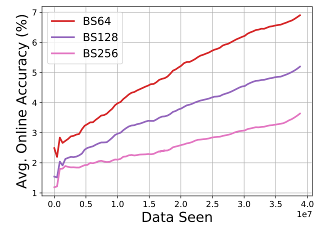
Figure 2: **The effect of batch size to supervised leanring**. The arrow in the caption of (b) points towards better performace. Varying batch sizes had minor effect to the training loss (a) and validation accuracy (b).

2.2. Batch Size Effect Without Replay Buffer

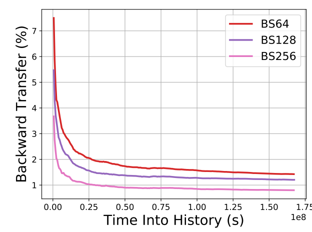
In the main paper, we analyze the effect of batch sizes to OCL, using models trained with experience replay. In order to validate whether the batch size effect we observed also exists without replay, we train OCL models using the same settings as in the batch size analysis of the main paper, except that all models are trained without experience replay. We plot the performance of trained models in Fig. 3. Similar to the main paper results, increasing batch sizes was harmful to the training loss and all performance metrics of OCL, even



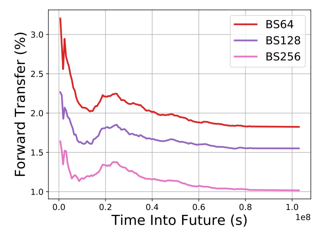
(a) Training loss (↓)



(b) Average online accuracy (↑)



(c) Backward transfer at $\frac{2H}{3}$ (↑)



(d) Forward transfer at $\frac{2H}{3}$ (↑)

Figure 3: **The effect of batch size to OCL, without using replay buffer.** The arrows in the caption of subfigures point towards better performance. Similar to the result of the main paper, increasing batch sizes increased the training loss and hurts all performance metrics of OCL.

without experience replay. Hence, the batch size effect we observed for OCL is not algorithm-specific.

References

[1] 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.