Remember More by Recalling Less: Investigating the Role of Batch Size in Continual Learning with Experience Replay (Student Abstract)

Maciej Wołczyk, Andrii Krutsylo
Faculty of Mathematics and Computer Science, Jagiellonian University
Lojasiewicza 6, 30-348 Kraków, Poland, +48 12 664 66 29
maciej.wolczyk@doctoral.uj.edu.pl, krutsylo@firemail.cc

Abstract

Experience replay is a simple and well-performing strategy for continual learning problems, often used as a basis for more advanced methods. However, the dynamics of experience replay are not yet well understood. To showcase this, we focus on a single component of this problem, namely choosing the batch size of the buffer samples. We find that small batches perform much better at stopping forgetting than larger batches, contrary to the intuitive assumption that it is better to recall more samples from the past to avoid forgetting. We show that this phenomenon does not disappear under learning rate tuning and we propose possible directions for further analysis.

Introduction & Setting

Without applying additional methods, neural networks perform poorly in a continual learning setting of training on a stream of constantly changing data. As the model learns the new tasks, its performance degrades on the past examples in a phenomenon called catastrophic forgetting. Although this problem can be avoided altogether in some settings by shuffling the data in order to obtain i.i.d. samples, in many real-life situations changing the order of data is infeasible.

Many sophisticated approaches to solve the problem were proposed in recent years (De Lange et al. 2019). However, in situations when storing a small set of samples from previous tasks is not prohibited, the naive approach of training the model simultaneously on the incoming data and the memorized examples often outperforms more advanced approaches with similar memory and computational overhead (Hsu et al. 2018).

This approach, called experience replay, although well known in reinforcement learning, has not been so far thoroughly investigated in the context of continual learning. Recent research suggests that the dynamics of this setting may vary significantly from the standard training with i.i.d. data. For example, even though overfitting is an important problem in classic machine learning, Chaudhry et al. (2019) find that overfitting a very small memory buffer in continual learning setting helps stabilize the training process.

As a first step towards understanding the properties of experience replay, we empirically investigate the impact of a single factor: the number of the samples from the memory buffer we use at each iteration, i.e. the buffer batch size. We find that, counterintuitively, using very large batches results in poor learning performance. Through further analysis of the training dynamics, we find that larger batches do not mitigate forgetting as well as small ones.

BBS	NR	5	10	50	100	1000
CIFAR	18.55	**43.92**	40.49	35.12	33.41	32.54
MNIST	19.14	83.23	**85.50**	84.61	84.80	84.40
MiniIN	3.43	18.91	**24.37**	11.17	9.95	4.41

Table 1: Final accuracy for different buffer batch sizes (BBS) shows that replaying too many samples decreases performance in online CL setting. NR represents no replaying and batch size 1000 is equivalent to using the whole buffer for MNIST and CIFAR.

We consider the online continual setting as described before by Chaudhry et al. (2019). We assume a set of classification tasks: $\mathcal{T}_1, \mathcal{T}_2, \dots \mathcal{T}_k$, where each task consists of examples $\mathcal{T}_i = (x_1^i, y_1^i), (x_2^i, y_2^i), \dots, (x_{n_i}^i, y_{n_i}^i)$. Tasks are visited sequentially and each example is seen only once (i.e. we train each task only for a single epoch), with the exception of examples stored in a small memory buffer $\mathcal{M}$ which can be replayed indefinitely. We add the new samples to the buffer using reservoir sampling.

During each step of the training, we take a batch of incoming samples $\mathcal{B}_{\text{in}} \subset \mathcal{T}_i$ as well as a batch of uniformly sampled examples from the buffer $\mathcal{B}_{\text{mem}} \subset \mathcal{M}$, excluding examples from the current task. The final loss function is then a sum of mean losses of both batches:

$$\frac{1}{|\mathcal{B}_{\text{in}}|} \sum_{(x_i,y_i)\in\mathcal{B}_{\text{in}}} L(x_i, y_i) + \frac{1}{|\mathcal{B}_{\text{mem}}|} \sum_{(x_i,y_i)\in\mathcal{B}_{\text{mem}}} L(x_i, y_i), \tag{1}$$

where L is the standard cross-entropy loss function. In our investigation, we want to check how the size of $\mathcal{B}_{\text{mem}}$ impacts the learning dynamics.

Experiments

We use the setting and hyperparameters from Aljundi et al. (2019). Same as them, we run the experiments on three datasets: CIFAR-10, MNIST, and mini-ImageNet, each containing 60000 training examples. We divide MNIST and

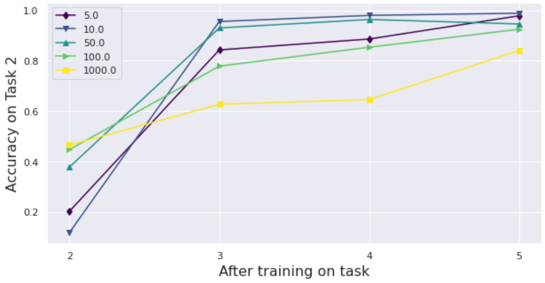

Figure 1: Accuracy on Task 2 examples stored in the buffer throughout the training for different buffer batch sizes.

CIFAR-10 into five sequentially learned tasks, each consisting of two classes, and mini-ImageNet into 20 tasks of 5 classes, in an incremental class learning setup. We use a memory buffer of 100 examples per class, and the online batch size is 10. Each experiment was run 10 times.

One might assume that repeating more data from the past might reduce forgetting. However, the results presented in Table 1 show that contrary to intuition, larger buffer batch size does not stop forgetting. This effect is especially evident on difficult datasets like CIFAR-10 and mini-ImageNet.

We examine the CIFAR-10 results further by plotting the accuracy of the models on the examples contained in the buffer. Figure 1 shows the accuracy of models with different buffer batch sizes on examples from Task 2 kept in the buffer. Models trained with larger batches have the highest accuracy out of all tested models right after online learning on Task 2, but their performance falls after learning subsequent tasks. We conclude that (1) the observed phenomenon is not caused by overfitting, since bigger batches perform poorly even on the buffer, and (2) the models trained with large batches are unable to utilize the data from the buffer to prevent forgetting. We find this result highly surprising and we present it as an example of how learning dynamics of experience replay in continual learning are underexplored.

Analysis and Conclusion

In order to investigate this phenomenon, we check whether the performance of models with big buffer batch sizes may be improved by increasing their learning rate. This approach is motivated by literature on deep network optimization which suggests that setting a proper ratio of learning rate to batch size is crucial (Jastrzebski et al. 2017), as well as the empirical observation that the buffer gradient norms are significantly smaller for models with larger buffer batch sizes. We test this on CIFAR-10, by reweighting the loss of the buffer examples from Equation 1 by a constant, which is equivalent to increasing the buffer learning rate. The results presented in Figure 2 show that this simple correction does not mitigate the problem. Although increasing the loss magnitude may improve the result, the small batch size model still performs significantly better.

Thus, we present a surprising observation that cannot be explained by controlling for basic factors such as the learning rate. We consider this phenomenon an additional indication that experience replay methods have interesting underlying dynamics that have not been sufficiently explored for continual learning. We believe that a deeper understanding of this problem will allow the community to build better performing methods for dealing with catastrophic forgetting.

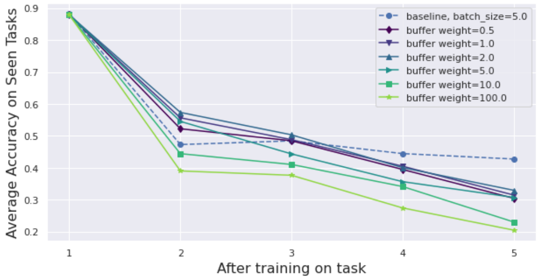

Figure 2: Accuracy for buffer batch size 1000 and different buffer loss weights.

In particular, we plan to examine our findings through the lens of memorization, which investigates when deep networks learn to recognize patterns or simply memorize the examples (Arpit et al. 2017). We leave this as future work.

Acknowledgements

The work of Maciej Wołczyk was supported by the research project ”Bio-inspired artificial neural networks” (grant no. POIR.04.04.00-00-14DE/18-00) within the Team-Net program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund.

References

Aljundi, R.; Belilovsky, E.; Tuytelaars, T.; Charlin, L.; Caccia, M.; Lin, M.; and Page-Caccia, L. 2019. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*, 11849–11860.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394* .

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H.; and Ranzato, M. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486* .

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelars, T. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383* .

Hsu, Y.-C.; Liu, Y.-C.; Ramasamy, A.; and Kira, Z. 2018. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488* .

Jastrzebski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2017. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623* .