Studying continual learning in a distribution shift setup

Andrei Mihailescu[*]

Cosmin Petrescu[†]

Abstract

We study the problem of continual learning in a distribution shift setup in the search of signs of catastrophic forgetting. We illustrate the knowledge transfer in a sequential regime and observe that even without the presence of catastrophic forgetting, the performance gap to the iid baseline is quite significant. While aiming to advance towards the baseline performance in a sequential manner, we experience that common approaches seen in the presence of this phenomenon, such as regularization methods like weight decay or network constraints such as Elastic Weight Consolidation [2] proved to have little to no impact in the current scenario. This is consistent with previous results in the field. Finally, going beyond a reproduction work, we attempt to close the performance gap by carrying over small amounts of previous samples. We select these samples using a forgetting score first proposed outside the field of CL and report our findings.

1 Introduction

The problem of learning consecutive tasks without losing knowledge on previously trained tasks is the concept that stands behind continual learning, making it a relevant topic of research in the pursuit of artificial intelligence. One aspect that stands in the way of this goal is known as catastrophic forgetting, a phenomenon where the model loses crucial information on former tasks while training on new ones. We want to probe the existence of catastrophic forgetting in a distribution shift setup. For this, we utilize CLEAR, the first continual image classification benchmark dataset with a natural temporal evolution of visual concepts in the real world that spans a decade (2004-2014) [1]. The labeled portion of CLEAR is split into 10 temporal buckets, each containing 10 illustrative classes such as computer, cosplay, etc. plus an 11th background class [1].

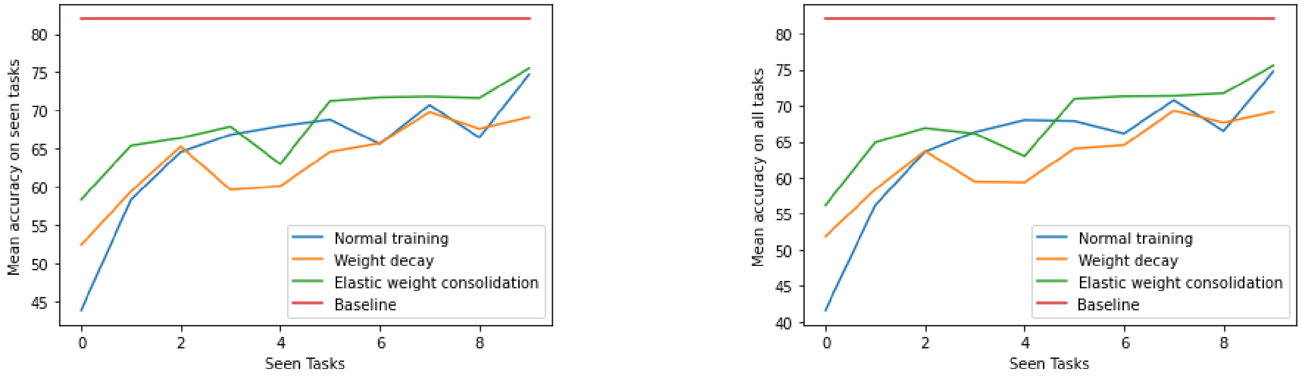

Figure 1: Illustrating the negligible effects of weight decay and EWC in a sequential regime. The first plot emphasises backward transfer while the second one provides generalization or forward transfer insight.

2 Methods and results

2.1 IID Baseline

We train a baseline that sets our performance goal for the sequential training regime. The model used for each experiment is a ResNet-18 without pretrained weights. Our dataset is obtained by concatenating the classes across all buckets. The

[*] University of Bucharest, worked on the training regimes, developed EWC, worked on benchmarking

[†] University of Bucharest, worked on the iid baseline, sequential regime benchmarking, offline CL with data pruning

transformations applied to our data include resizing, random cropping, random horizontal flipping and finally normalization. Our most solid results were obtained using the following hyperparameters: 30 epochs, a learning rate of 1e-3 and a patience for early stopping of 3 epochs, achieving a final validation accuracy of 82%, with an 84% spike on the 26th epoch.

2.2 Catastrophic forgetting and knowledge transfer

Bound to a sequential training regime, the first step is observing if forgetting occurs and analyzing the behavior of knowledge transfer. With a naive sequential training pipeline, our model presented no signs of forgetting. We propose comparing the naive version with two other techniques with the goal of approaching the accuracy of our baseline. First, we want to observe the impact of regularization procedures and so we include weight decay, which proves to be unhelpful.

Then, we address another method called Elastic Weight Consolidation [2]. It is a regularization technique inspired by the behavior of human dendritic spines that aims to anchor certain parameters when moving from one task to the next in a way that keeps the model in the optimum space around the solution for the previous task. We have paid close attention to the implementation, for example by not implementing the Fisher Information Matrix as Empirical FIM [4]. Unfortunately we have not observed an increase in performance, but on a closer look this is explained by the fact that the model struggles to reach the optimum area for the new task, and does not appear to be moving out of the previous optimum space.

2.3 Offline continual learning with data pruning

Offline continual learning assumes all samples of current task, plus the ones stored in a buffer can be revisited without constraint [1]. Combining this strategy with data pruning, we adopt the following method. We train each task on its corresponding dataset plus the pruned versions of previously seen datasets. After training is complete, we prune the current corresponding dataset and repeat the process on the next task. Regarding the pruning method, we proposed three variants, each one with a 75 pruning percent. The first one relies on random pruning while the second and the third are based on the importance of each sample. Mansheej et al. [3] proposed a modality of scoring the importance of each training example, within the first few epochs of training, by using the norm of the error vector (EL2N score), where the error vector is the predicted class probabilities minus one-hot label encoding. In our second experiment, we keep the highest scoring 25% samples. Another valuable information we found in [3] is that excluding a small subset of the very highest scoring examples produces a boost in performance, a boost which is enhanced in a corrupted label regime. Finally, our third experiment makes use of the last observation and keeps the highest scoring 25% except for the first 1%. As Figure 2 shows, we manage to minimize the gap to our goal using this setup, by reaching approximately 80% accuracy. However, surprisingly, the same performance is achieved with both random and importance based pruning.

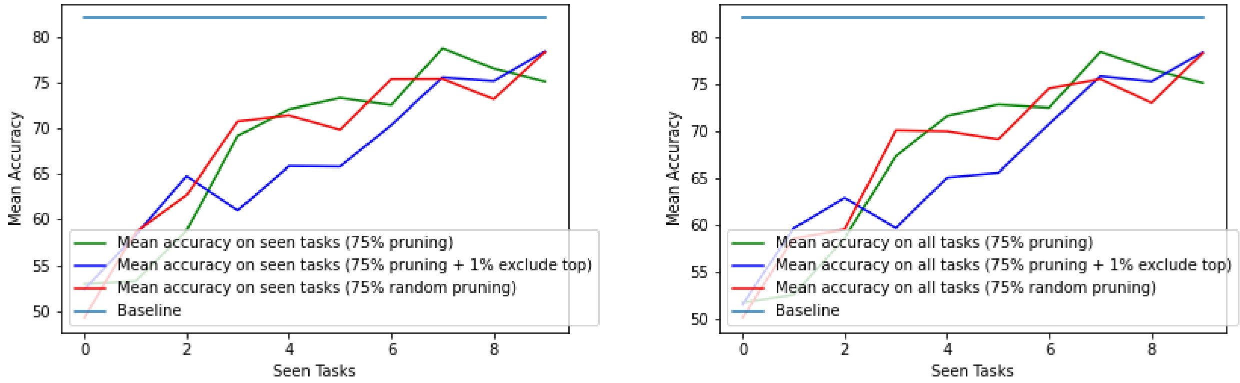

Figure 2: Comparing the performance of offline CL paired with different pruning methods (randomized and based on EL2N scores). There are no signs of performance boost with EL2N scores pruning.

3 Conclusion

We obtained no performance enhancements with the help of weight decay or EWC and observed identical outcomes with random and importance based data pruning. We are planning to continue diving deep into the subject, the next step being researching the potential usage of learning neural network subspaces [5].

4 References

[1] Lin, Z., Shi, J., Pathak, D., and Ramanan, D. The CLEAR Benchmark: Continual LEArning on Real-World Imagery. In Conference on Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks, Virtual only, December 2021.

[2] Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil, Veness, Joel, Desjardins, Guillaume, Rusu, Andrei A., Milan, Kieran, Quan, John, Ramalho, Tiago, GrabskaBarwinska, Agnieszka, Hassabis, Demis, Clopath, Claudia, Kumaran, Dharshan, and Hadsell, Raia. Overcoming catastrophic forgetting in neural networks. PNAS, pp. 201611835, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114.

[3] Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. Advances in Neural Information Processing Systems, 34, 2021.

[4] F. Kunstner, L. Balles, and P. Hennig, “Limitations of the empirical fisher approximation for natural gradient descent,” in Proc. Adv. Neural Inf. Process. Syst., 2019.

[5] Wortsman, M., Horton, M., Guestrin, C., Farhadi, A., and Rastegari, M. Learning neural network subspaces. In Internatinal Conference on Machine Learning, volume 139. PMLR, 2021. URL arXiv:2102.10472.