Continual learning in a distribution shift setup

Cosmin Petrescu, Andrei Mihailescu
Department of Computer Science, University of Bucharest

University of Bucharest
VIRTUTE ET SAPIENTIA

1. Main results

Learning consecutive tasks without losing knowledge on previously trained tasks is the primary concept behind continual learning. 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. In our efforts of advancing towards the baseline performance within a distribution shift setup, we experience that:

- **There are no signs of catastrophic forgetting in a sequential regime, consistent with the findings of the CLEAR team [1]**
- **As a consequence, weight decay and Elastic Weight Consolidation provided little to no improvement**
- **Importance based data pruning (first proposed outside the field of CL) does not outperform random pruning in an offline setup**

2. Dataset

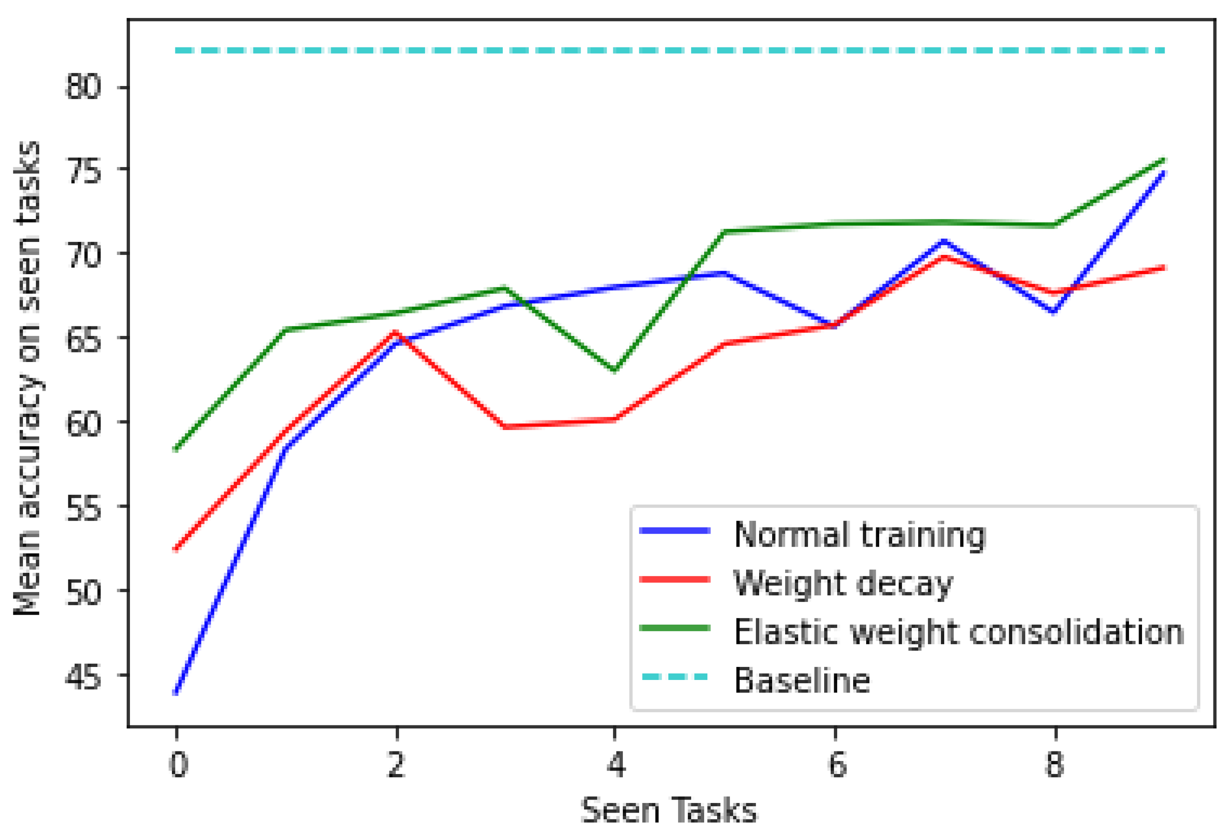
We utilize CLEAR, the first continual image classification benchmark dataset with a natural temporal evolution of visual concepts in the real world that spans a decade (2004-2014) [1]. The labeled portion of CLEAR is split into 10 temporal buckets, each containing 10 illustrative classes such as computer, cosplay, etc. plus an 11th background class.

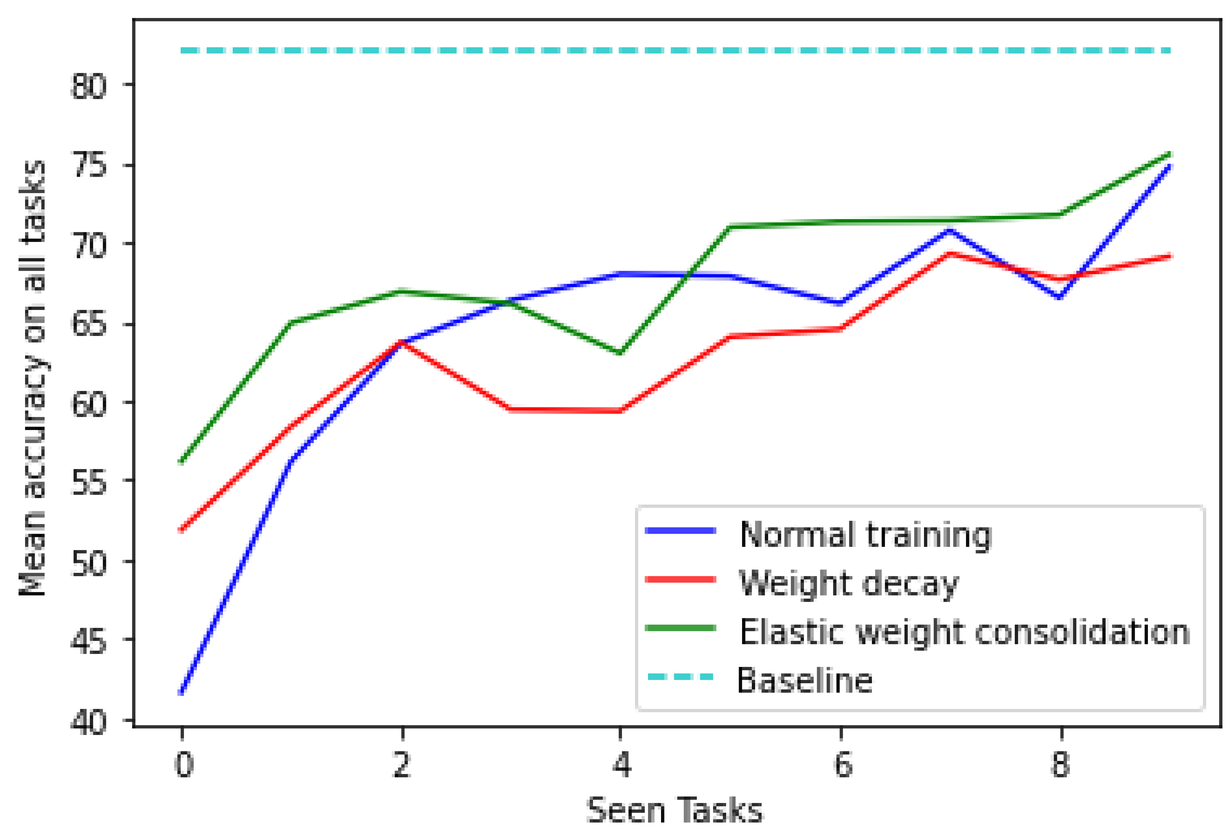
3. IID Baseline

We train a **ResNet-18** from scratch as our IID baseline, setting our performance goal for the sequential training regime. Our dataset is obtained by concatenating the classes across all buckets. The 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%**.

4. Weight Decay and Elastic Weight Consolidation

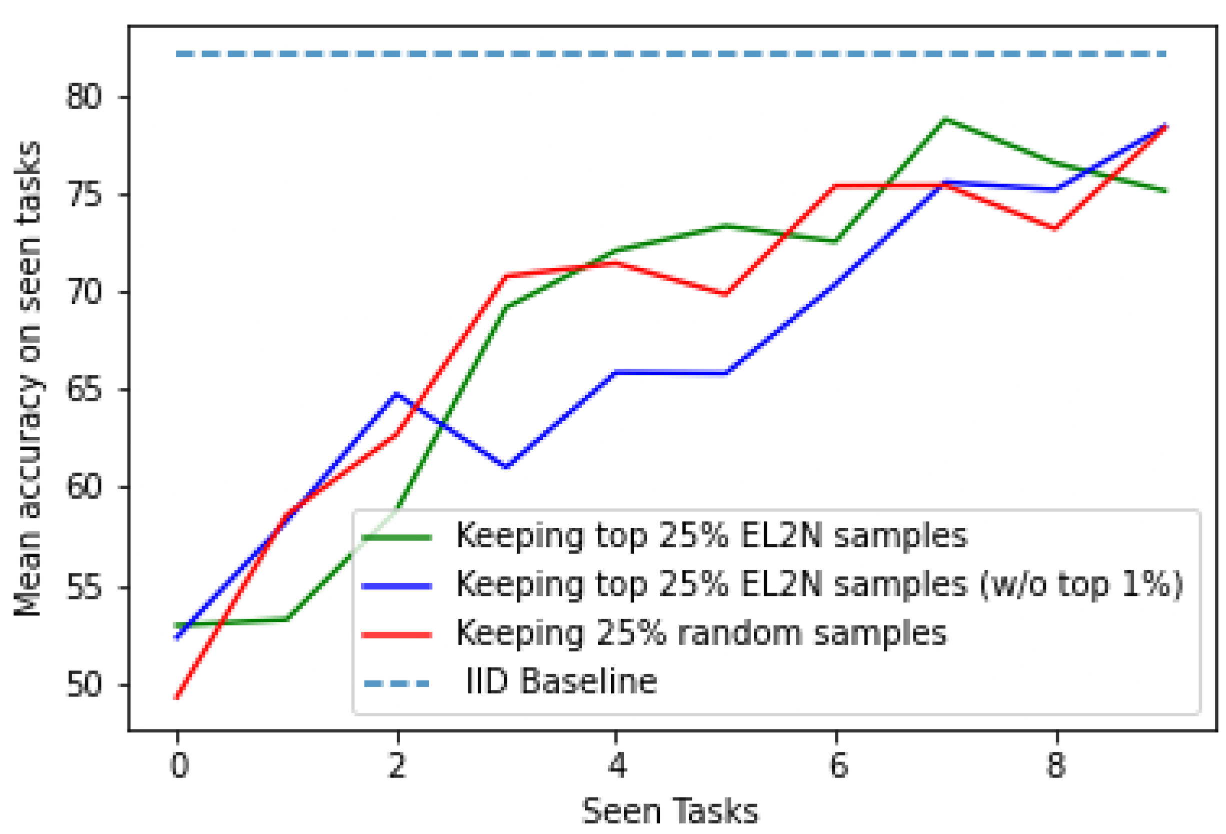
We compare naive sequential with two other techniques with the goal of approaching the accuracy of our baseline: weight decay and Elastic Weight Consolidation [2], a technique inspired by the behavior of human dendritic spines that aims to anchor certain parameters when moving from one task to the next in a way that keeps the model in the optimum space around the solution for the previous task. We have observed no performance boost. On a closer look, with EWC 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.

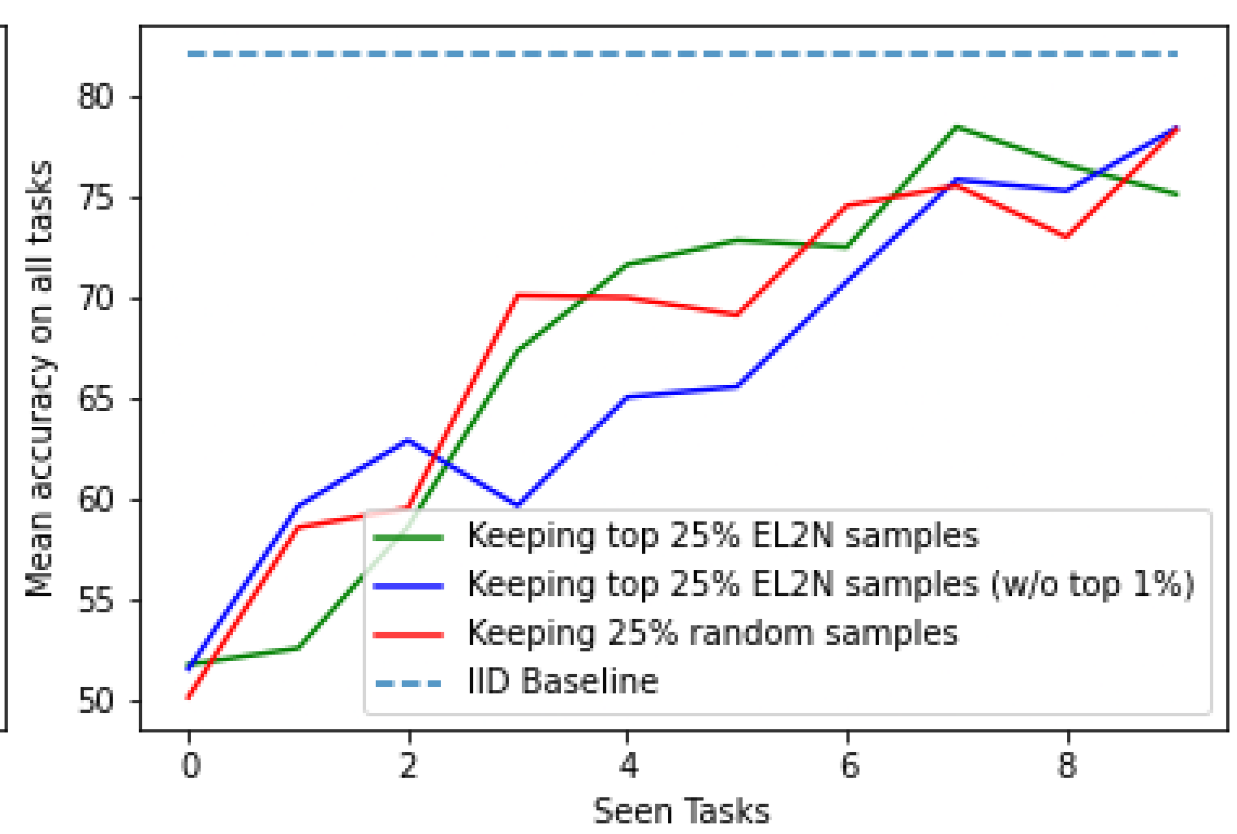




5. Offline continual learning with data pruning

We train each task on its corresponding dataset plus the pruned version of each previously seen dataset. [3] proposed a modality of scoring the importance of each training example by using the norm of the error vector **(EL2N score)**, where the error vector is the predicted class probabilities minus one-hot label encoding. Also, excluding a small subset of the highest scoring examples produces a boost in performance, a boost which is enhanced in a corrupted label regime [3]. Although reaching approximately 80% accuracy, surprisingly, the same performance is achieved with both random and importance based pruning.





6. Conclusions

We've successfully reproduced previous results in the field by demonstrating the lack of forgetting in this setup and the behaviour of EWC. Also, we tested the effects of 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 sub-spaces.

7. References

[1] Shi J. Pathak D. Lin, Z. and D. Ramanan. The clear benchmark: Continual learning on real-world imagery. in conference on neural information processing systems. *NeurIPS*, Track on Datasets and Benchmarks, 2021.

[2] Pascanu Razvan Rabinowitz Neil Veness Joel Desjardins Guillaume Rusu Andrei A. Milan Kieran Kirkpatrick, James and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *PNAS*, 201611835, 2017.

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