Deep Learning - Researching Forgetting on ”CLEAR” using Elastic Weight Consolidation

Ioan Daniel Savu Rusu Andrei-Cristian

Abstract

In the most recent times, we have noticed increasing interest in continual learning. Having to overcome the challenges associated with it is a common problem, as we usually want to train existing models on new datasets that have different distributions. In this extended abstract, we are going to take a look whether the phenomenon of *forgetting* appears in the ”CLEAR” dataset, and how using Elastic Weight Consolidation might help improving the results.

1 Introduction

As mentioned in [4], the *CLEAR* dataset is specifically made for Continuous Learning on Real World Imagery. It contains 300 images for each of the 11 classes, and the data is distributed across 10 years, from 2004 to 2014. As can be seen, the continuous learning aspect of this dataset is the classes distribution shift over the years.

The idea of **Elastic Weight Consolidation** was first discussed in [3], and it's a regularization method, inspired from neuroscience - ”*synaptic consolidation might enable continual learning by reducing the plasticity of synapses that are vital to previously learned tasks*”. **EWC** is an algorithm which performs a similar operation, by constraining important parameters of the neural network to stay close to their old values, when being trained on new tasks.

Our approach differs from the one presented in [4]. We kept the same method of training and testing, which is to have two different scenarios: the ”bulk” version, in which the model was trained by having all the data available all at once, and a ”sequential” one, in which the data was available year by year. We also kept the same model architecture, a *ResNet18* [2] with two different initializations. The main difference is that the pretrained version was trained on the **Imagenet**[1] dataset. This will give us better insight of how different ways of pretraining will change the final performance.

2 Experiment results

For both versions, in the training step, we used data augmentation such as Random Crop, Gaussian Blur, Random Rotation and Random Flip. These were especially useful since the number of images per class is relatively small and the model tends to overfit if no data augmentation is applied. At the same time, since the number of epochs that we need to train differs on every scenario, we capped the number to 50 and used *early stopping*. Another factor for the final performance was the *learning rate scheduler* that helped us to use an appropriate rate for each epoch.

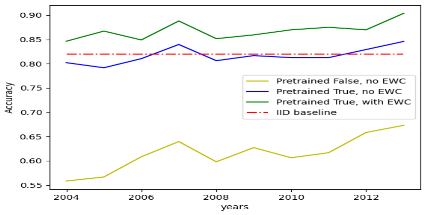
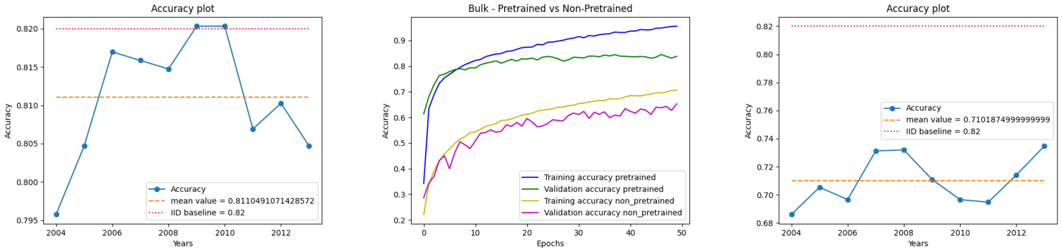


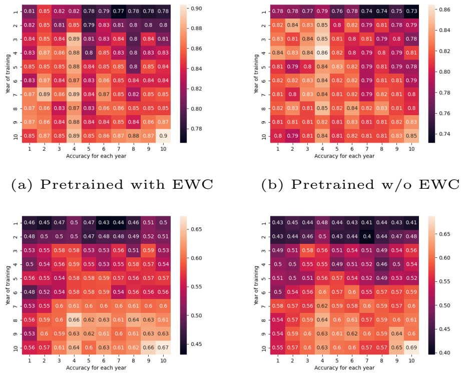
Figure 1: Accuracy comparison for the final models

Using the **Bulk** version, we have established a baseline for our problem. As we can see, the accuracy lies at around **82%** on the pretrained model, and close to **64%** on the model built from scratch. This difference was somewhat to be expected, as the images in the *CLEAR* dataset are similar to the ones in *ImageNet* even though they are from different sources.



(a) Validation Accuracy [pretrained] (b) Validation Accuracy (c) Validation Accuracy [from scratch]

The following step was to train various models sequentially on the datasets, both with and without *Elastic Weight Consolidation*. A relevant metric for testing the performance is the accuracy matrix in which we compute the accuracy on the past and future datasets after training on each year. This gave us insight to how the model transfers or forgets the knowledge. The first thing we notice after having a look at the results is opposed to what we've been expecting - rather than forgetting, the models learn, for each year, new knowledge that propagates both forward and backwards, which improves the performance. This phenomenon suggests that the datasets are sharing features and that the distribution shift between them is smooth.



(a) Pretrained with EWC (b) Pretrained w/o EWC

(c) Not pretrained with EWC (d) Not pretrained w/o EWC

We also observe another interesting fact in Figure 1. On the pretrained version **without** *EWC* the mean accuracy after training on every year is **81%**, and by using *EWC*, we have **86%**, which is a relative improvement of **6.8%**. On the other hand, the non-pretrained version, **without** *EWC* obtained a mean accuracy of **60%**, while using *EWC* improves the accuracy up to **61%** which, again, is a relative improvement of **3.6%**.

This suggests that **EWC** may be able to do more than just prevent catastrophic forgetting. Since this method of regularization helps preserving important weights in the model, it might help in t he training process on datasets that have a relative small number of datapoints and a gentle distribution shift between them to generalize better and thus give better results overall.

3 Conclusions

We have observed an unexpected behaviour - even though *EWC* is thought to be a regularization method for catastrophic forgetting, it seems to improve the overall performance of the models. In the future, we intend to analyze if this property of *EWC* stands for other datasets with small distribution shifts, as this would mean that *EWC* is not just a solution for catastrophic forgetting, but also a good regularization method for improving performance.

References

[1] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: (2009), pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[2] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[3] James Kirkpatrick et al. “Overcoming catastrophic forgetting in neural networks”. In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526.

[4] Zhiqiu Lin et al. “The CLEAR Benchmark: Continual LEArning on Real-World Imagery”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2021.