On the Computation of the Fisher Information in Continual Learning

Gido M. van de Ven

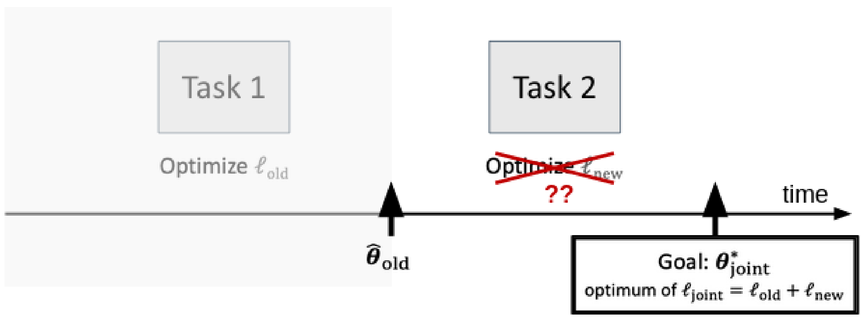
KU Leuven (Belgium), TU Delft (the Netherlands)

Abstract

One of the most popular methods for continual learning with deep neural networks is Elastic Weight Consolidation (EWC) [1], which involves computing the Fisher Information. The exact way in which the Fisher Information is computed is however rarely described, and multiple different implementations for it can be found online. Here, I discuss and empirically compare several often-used implementations, which highlights that many currently reported results for EWC could likely be improved by changing the way the Fisher Information is computed.

Continual learning

A model f_θ has been trained on a task by optimizing loss $\ell_{\text{old}}(\boldsymbol{\theta})$ on training data $D_{\text{old}} \sim \mathcal{D}_{\text{old}}$, resulting in weights $\hat{\boldsymbol{\theta}}_{\text{old}}$. Goal is to continue training this model on a new task, by optimizing loss $\ell_{\text{new}}(\boldsymbol{\theta})$ on training data $D_{\text{new}} \sim \mathcal{D}_{\text{new}}$, such that the model maintains its performance on the previous task. Unfortunately, continued training by only optimizing $\ell_{\text{new}}(\boldsymbol{\theta})$ results in catastrophic forgetting.



Scope: continual classification with neural nets (i.e., f_θ models the conditional distribution $p_\theta(y|\boldsymbol{x})$).

Elastic Weight Consolidation (EWC)

When training on a new task, rather than optimizing only $\ell_{\text{new}}(\boldsymbol{\theta})$, EWC adds an extra term to the loss that involves the Fisher Information:

$$\ell_{\text{EWC}}(\boldsymbol{\theta}) = \ell_{\text{new}}(\boldsymbol{\theta}) + \frac{\lambda}{2} \sum_{i=1}^{N_{\text{params}}} F^{i,i}_{\text{old}} (\theta^i - \hat{\theta}^i_{\text{old}})^2$$

with λ a hyperparameter and $F^{i,i}_{\text{old}}$ the i^{th} diagonal element of the model's Fisher on the old data.

Definition of the Fisher

Following Martens [2], the i^{th} diagonal element of the model's Fisher on the old data is defined as:

$$F^{i,i}_{\text{old}} := \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{old}}} \left[\mathbb{E}_{y \sim p_{\hat{\theta}_{\text{old}}}} \left[\left(\frac{\delta \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\delta \theta^i} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

In this definition, there are two expectations:

1. an outer expectation over $\mathcal{D}_{\text{old}}$, which is the (theoretical) input distribution of the old data
2. an inner expectation over $p_{\hat{\theta}_{\text{old}}}(y|\boldsymbol{x})$, which is the conditional distribution of y given $\boldsymbol{x}$ defined by the model after training on the old data

The different ways of computing the Fisher that can be found in the continual learning literature differ in how these two expectations are computed.

Different ways of computing the Fisher

Exact

$$F^{i,i}_{\text{old, EXACT}} = \frac{1}{|D_{\text{old}}|} \sum_{\boldsymbol{x} \in D_{\text{old}}} \left(\sum_{y=1}^{N_{\text{classes}}} p_{\hat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x}) \left(\frac{\delta \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\delta \theta^i} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right)$$

Sampling data points

$$F^{i,i}_{\text{old, EXACT}(n)} = \frac{1}{n} \sum_{\boldsymbol{x} \in S^{(n)}_{D_{\text{old}}}} \left(\sum_{y=1}^{N_{\text{classes}}} p_{\hat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x}) \left(\frac{\delta \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\delta \theta^i} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right)$$

Sampling labels

$$F^{i,i}_{\text{old, SAMPLE}} = \frac{1}{|D_{\text{old}}|} \sum_{\boldsymbol{x} \in D_{\text{old}}} \left(\frac{\delta \log p_{\boldsymbol{\theta}}(c_{\boldsymbol{x}}|\boldsymbol{x})}{\delta \theta^i} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \text{ with } c_{\boldsymbol{x}} \sim p_{\hat{\boldsymbol{\theta}}_{\text{old}}}(.|\boldsymbol{x})$$

Empirical Fisher

$$F^{i,i}_{\text{old, EMPIRICAL}} = \frac{1}{|D_{\text{old}}|} \sum_{(\boldsymbol{x},y) \in D_{\text{old}}} \left(\frac{\delta \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\delta \theta^i} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{old}}} \right)^2$$

Batched approximation of Empirical Fisher (e.g., [3, 4])

$$F^{i,i}_{\text{old, BATCHED}(b)} = \frac{1}{|D^{(b)}_{\text{old}}|} \sum_{\mathcal{B} \in D^{(b)}_{\text{old}}} \left(\sum_{(\boldsymbol{x},y) \in \mathcal{B}} \frac{\delta \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\delta \theta^i} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{old}}} \right)^2$$

Empirical comparison

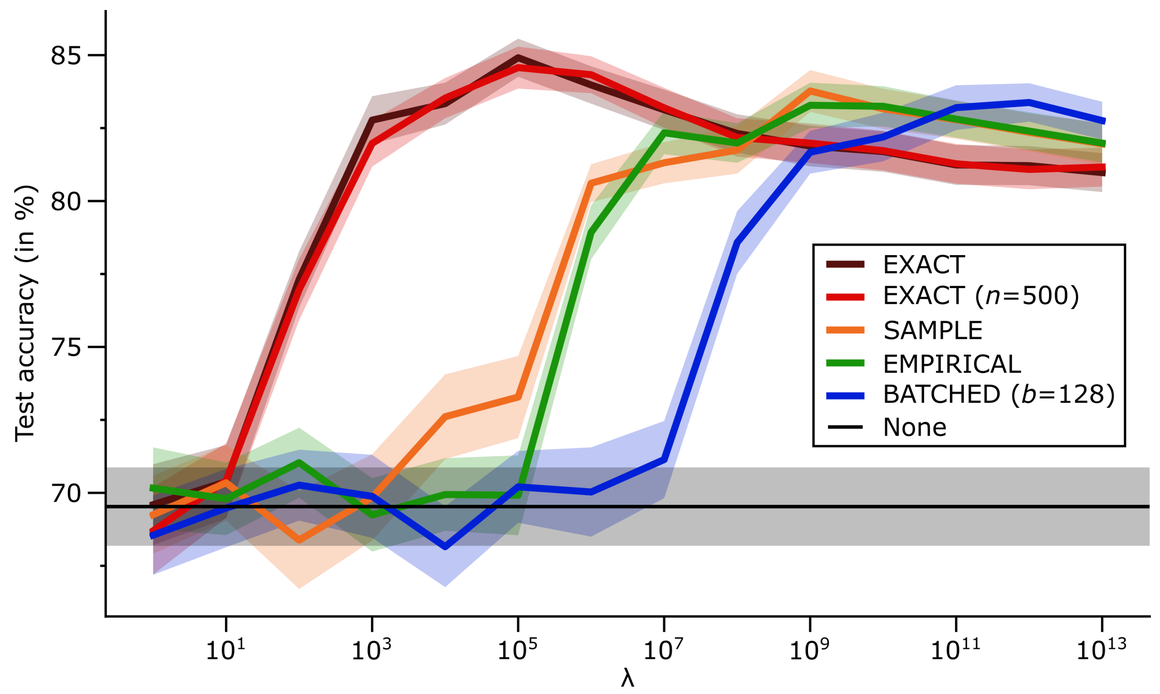


Figure 1. Performance of EWC on Split CIFAR-10 with different ways of computing the Fisher Information for a wide range of hyperparameter values.

References & Acknowledgements

[1] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[2] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

[3] Antonio Carta, Lorenzo Pellegrini, Andrea Cossu, Hamed Hemati, and Vincenzo Lomonaco. Avalanche: A pytorch library for deep continual learning. *Journal of Machine Learning Research*, 24(363):1–6, 2023.

[4] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. PyCIL: a Python toolbox for class-incremental learning. *Science China Information Sciences*, 66:197101, 2023.

Funding: senior postdoctoral fellowship, Resarch Foundation – Flanders (FWO), grant number 1266823N.

gido.vandeven@kuleuven.be

https://gmvandeven.github.io/