
Towards continual learning in medical imaging

Chaitanya Baweja
Imperial College London
cb5117@imperial.ac.uk

Ben Glocker
Imperial College London
b.glocker@imperial.ac.uk

Konstantinos Kamnitsas
Imperial College London
kk2412@imperial.ac.uk

Abstract

This work investigates continual learning of two segmentation tasks in brain MRI with neural networks. To explore in this context the capabilities of current methods for countering catastrophic forgetting of the first task when a new one is learned, we investigate *elastic weight consolidation* [1], a recently proposed method based on Fisher information, originally evaluated on reinforcement learning of Atari games. We use it to sequentially learn segmentation of normal brain structures and then segmentation of white matter lesions. Our findings show this recent method reduces catastrophic forgetting, while large room for improvement exists in these challenging settings for continual learning.

1 Introduction

Advances in machine learning have led to rapid developments towards automation of various tasks, such as detection of pathology in medical scans. The currently most successful models are supervised neural networks. A network is trained using a manually annotated database for a specific task. However the amount of potential tasks in healthcare is immense. Thus annotating a large database for each task, to train specialized models from scratch, is not a scalable solution. Instead we envision that a model with pre-acquired knowledge about common tasks could be supplied to clinicians and, according to their needs, they could further train it to perform a new task. First, by utilizing existing knowledge from previous tasks, the model should quickly grasp the new task with only limited supervision. Furthermore, it should be able to incorporate the new knowledge with existing one, improving its knowledge both for the original and any future tasks. Finally, learning a new task should be possible without access to training data for the earlier tasks, which may no longer be available.

Such sequential knowledge acquisition is known as *continual* and *lifelong learning* [2]. It is related to *multitask learning* [3], which assumes training data for all tasks of interest are available and all tasks are learnt concurrently. Instead, here, we assume that when learning a new task, the training data for previous tasks are no longer available. Sequential learning poses a great challenge for neural networks, known as *catastrophic forgetting* [4, 5, 6], where knowledge about an old task is lost when changing a network’s parameters during training to meet the objective for a new task.

Countering catastrophic forgetting in neural networks has recently attracted increased research attention. The first category of works derive regularization costs such that knowledge of the new task can be incorporated in existing capacity while preserving model behaviour on the old task [7, 1]. Other approaches extend a network with extra capacity or components for each new task [8, 9]. This may alleviate forgetting but does not effectively fuse old and new knowledge. Such approach was applied for supervised domain adaptation of a model to different MRI scanners [10], but not for different tasks, where the label spaces and labelling functions differ.

This work explores catastrophic forgetting when learning sequentially two different tasks in medical imaging: segmentation of normal structures and segmentation of white matter lesions in brain MRI. We investigate the potential of a recently proposed method, *Elastic Weight Consolidation* (EWC) [1],

originally evaluated for reinforcement learning of Atari games. The method tries to preserve network connections important for previous tasks, by regularizing connections with high Fisher information. We show experimentally that EWC reduces catastrophic forgetting in our settings. This study, the first of its kind in medical imaging, indicates that there is potential in this approach, while showing that there is significant space for further research and improvements towards continual learning.

2 Fisher information for continual learning with neural networks

Suppose we have some data \mathcal{D}_A , such as images x and corresponding annotations y , that reflect a particular task A . A discriminative neural network parameterized by θ learns to approximate the distribution $p(y|x, \theta)$ that generates y given x . For this, it learns the optimal parameters θ_A^* during training in order to minimize an appropriate loss $\mathcal{L}_A(\mathcal{D}_A, \theta)$. In continual learning, after training for task A , it is assumed that \mathcal{D}_A is no longer available. We are then given new training data, \mathcal{D}_B , for another task B . Starting from the existing knowledge encapsulated in θ_A^* , we wish to further change the parameters to also solve B , while preserving the knowledge about A . Assuming large, over-parameterized neural networks, many configurations of θ may lead to similar performance ([11, 12]). Thus it is likely that there is a solution for task B , θ_B^* that is in the neighbourhood of θ_A^* . Staying near θ_A^* during training on B can be encouraged by a regularizer based on L2 distance $R(\theta) = (\theta_i - \theta_{A,i}^*)^2$, but this does not guarantee meaningful minima with respect to A .

Instead, one can investigate the importance of each parameter θ_i with respect to behaviour of the model. A measure for this is *Fisher information*, which expresses the amount of information observing variable Y carries about a parameter θ that models distribution of Y . For a model with $\theta \in \mathbb{R}^K$ parameters, this is expressed by the *Fisher Information Matrix* F defined as:

$$F = \mathbb{E}_{(x,y)} [\nabla_{\theta} \log p(y|x, \theta) \nabla_{\theta} \log p(y|x, \theta)^T] \in \mathbb{R}^{K \times K} \quad (1)$$

It is the variance of the *score function* $s(\theta) = \nabla_{\theta} \log p(y|x, \theta)$, the expected value of which is zero. F quantifies how much a change of a parameter's value is expected to affect the output of a network $p(y|x, \theta)$. Intuitively, if $\nabla_{\theta_i} \log p(y|x, \theta_i) = 0$, then $F_{i,j} = 0$, $F_{j,i} = 0$, $\forall j$, which expresses that parameter θ_i can be altered without change of the output. For continual learning, it is possible to use F to regularize the change of each parameter when training for B according to its *importance* for task A [1]. Because K can be very large for neural networks, F can be impractical to compute however.

Elastic Weight Consolidation [1] is a regularizer that is based on the assumption that the parameters of a network are uncorrelated (weak assumption if K is very large). In this case F is diagonal, thus one needs to compute only K values. *Importance* of parameter θ_i for task A is then:

$$F_i = \mathbb{E}_{(x,y) \sim \mathcal{D}_A} \left[\left(\nabla_{\theta_{A,i}^*} \log p(y|x, \theta_{A,i}^*) \right)^2 \right] \in \mathbb{R} \quad (2)$$

which is computed after convergence of training on A . Finally, training for task B is performed by minimizing the follow total cost:

$$\mathcal{L}_{B,Total}(\mathcal{D}_B, \theta, \theta_A^*) = \mathcal{L}_B(\mathcal{D}_B, \theta) + \lambda \sum_{i=1}^K F_i (\theta_i - \theta_{A,i}^*)^2 \quad (3)$$

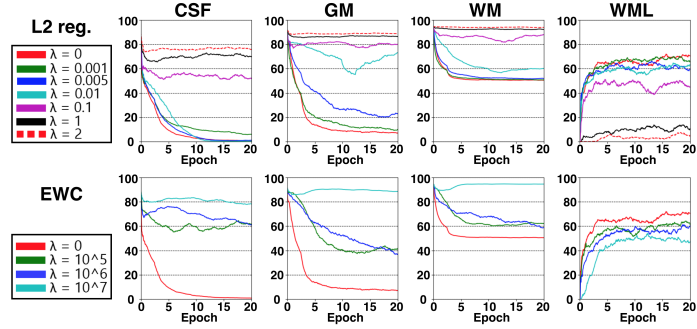
EWC protects the parameters with high F_i to stay close to the values needed for A , while parameters with low F_i are allowed to move more freely, constituting capacity of the network allocated to task B . λ controls strength of regularization. In what follows, we investigate the potential of EWC for the first time on a challenging biomedical application for sequential learning of two different tasks.

3 Evaluation

3.1 Experimental setup

Databases: We use images and corresponding pixel-wise labels provided by UK Biobank [13] for the following tasks. **Task A:** multi-class segmentation of cerebrospinal fluid (CSF), grey matter (GM), white matter (WM). **Task B:** segmentation of white matter lesions (WML). We select 275 cases where WML is present. 87 cases are used when training for A , 88 when training for B , 100 to validate both tasks. In all experiments we use T1 and Flair sequences, after z-score normalization.

Figure 1: Starting with DM pre-trained on task A (epoch 0), we train it for task B for 20 epochs, using (top) L_2 regularizer or (bottom) EWC, with varying λ values. Plots show evolution of segmentation performance (DSC%) for classes of task A (three left columns) and task B on random patches from validation cases.



Experiments: We use the DeepMedic (DM) 3D convnet [14] with default configuration, for its reliable performance in segmenting volumetric scans. We performed the following experiments:

DM-A: Train DM only for task A from scratch.

DM-B: Train DM only for task B from scratch.

Multi-task: learn A and B jointly, using a DM with two classification layers. This is an upper bound for performance of continual learning, where data \mathcal{D}_A is assumed unavailable when learning B.

Fine-tune: Add new classification layer for task B on pre-trained DM-A, fine-tune whole net on \mathcal{D}_B .

L2: Similar to *fine-tune*, but regularize training for B via L_2 distance of θ from θ_A^* ($F_i = 1 \forall i$ in Eq. 3).

EWC: Compute F_i for each θ_i of pre-trained DM-A according to Eq. 2. Then add a new classification layer for task B on pre-trained DM-A, and learn task B by minimizing loss given in Eq. 3.

3.2 Results

We evaluate segmentation performance of the employed models on the validation subjects and report the Dice similarity coefficient (DSC) for each class in Table 2. Catastrophic forgetting is not an issue in multi-task learning, as data for both tasks are available during training. In contrast, unregularized *fine-tuning* for task B suffers from it, completely forgetting task A. Regularizing training on task B via the naive L_2 regularizer or EWC mitigates forgetting. The strength of regularization λ (Eq. 3) acts as a trade-off between learning task B and forgetting task A. By comparing settings where L_2 and EWC regularization perform similarly on task B, we see that EWC preserves higher performance in task A, although there is still a gap until the upper bound set by multi-task learning. We show visual examples of the results for L_2 ($\lambda = 0.1$) and EWC ($\lambda = 10^7$) regularization in Fig. 3.

Finally, we present how segmentation performance on patches extracted randomly from validation subjects evolves while training for task B using L_2 and EWC regularizers with different values of λ in Fig. 1. The trade-off between learning task B and forgetting task A can be observed, while EWC performs overall better than the basic L_2 regularizer.

4 Conclusion

We investigated continual learning of two different segmentation tasks in the context of brain MRI. We explored the potential of EWC [1], a recently proposed regularizer based on Fisher information, previously shown promising for continual reinforcement learning on Atari games. Our investigation on bio-medical data shows that the technique is promising for alleviating catastrophic forgetting. The findings of this work, one of the firsts of its kind on medical imaging data, also show that current methods leave significant space for further research before continual learning becomes practical. We hope this work will stimulate further investigations in this largely unexplored area.

Figure 2: DSC% achieved by employed methods when fully segmenting the validation cases.

Method	Task A			Task B
	CSF	GM	WM	WML
DM-A	89.4	92.5	95.0	-
DM-B	-	-	-	61.3
Multi-task	88.9	92.2	94.8	63.8
Fine-tune	00.9	11.9	50.8	62.1
$L_2 \lambda = 0.005$	00.5	07.9	50.7	54.4
$L_2 \lambda = 0.01$	02.9	80.6	59.9	56.3
$L_2 \lambda = 0.1$	50.8	82.0	87.5	42.3
$EWC \lambda = 10^6$	60.1	32.6	62.5	53.4
$EWC \lambda = 10^7$	79.2	88.7	94.4	44.5

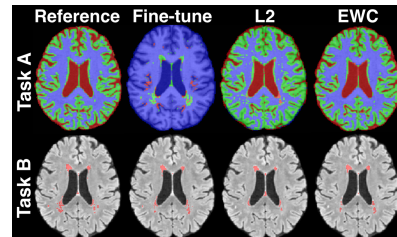


Figure 3: Segmentations from employed methods after learning task B. EWC mitigates forgetting of task A.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG). KK is supported by the President’s PhD Scholarship of Imperial College London. This research has been conducted using the UK Biobank Resource under Application Number 12579.

References

- [1] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, page 201611835, 2017.
- [2] Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.
- [3] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [4] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
- [5] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.
- [6] James L McClelland, Bruce L McNaughton, and Randall C O’reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.
- [7] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [8] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [9] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516, 2017.
- [10] Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu. A lifelong learning approach to brain mr segmentation across scanners and protocols. *Medical Image Computing and Computer Assisted Intervention*, 2018.
- [11] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.
- [12] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.
- [13] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523, 2016.
- [14] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.