Evaluating Catastrophic Forgetting in Neural Networks Trained with Continual Backpropagation

Author:
Justinas Jučas
jjucas@tudelft.nl
Supervisors:
Laurens Engwegen
Wendelin Böhmer

Problem

Continual learning suffers from two main problems: **loss of plasticity** and **catastrophic forgetting** which are inherently linked through the *stability-plasticity trade-off*. Dohare et al. [1] proposed an efficient way to mitigate loss of plasticity in neural networks (NNs) by introducing an algorithm called **Continual Backpropagation** (CBP). However, no comprehensive research has been conducted to evaluate the extent of CBP's susceptibility to the second part of the trade-off: catastrophic forgetting.

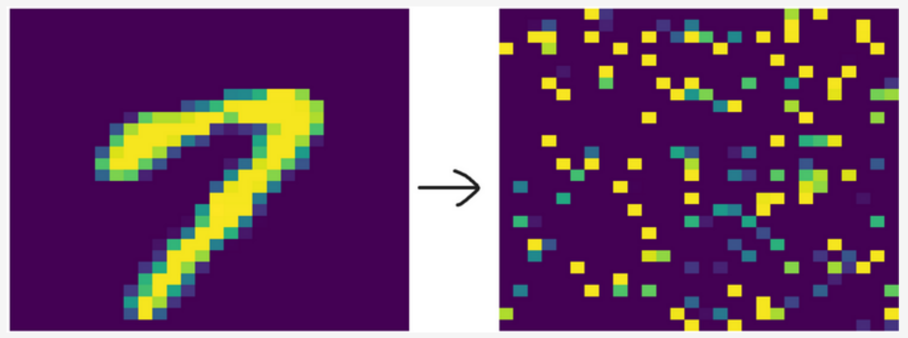
Research Questions

What is the effect of Continual Backpropagation algorithm on catastrophic forgetting?

1. How much is CBP prone to catastrophic forgetting compared to other algorithms?
2. What is the trade-off between loss of plasticity and catastrophic forgetting with regard to different hyperparameters of CBP?
3. Do NNs trained with CBP demonstrate improved retention and adaptation to information introduced multiple times?
4. What internal dynamics do NNs display when trained using CBP?
5. Can CBP be improved to reduce forgetting?

Benchmark: OPMNIST

Online Permuted MNIST (OPMNIST). MNIST is a publicly available dataset of handwritten digit images. CL setting is simulated by training a model on a sequence of tasks, where each task corresponds to classifying the MNIST dataset with a new, task-specific pixel permutation applied to all the images.



References

[1] Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. Nature, 632(8026):768–774, 2024.
[2] Jordan Ash and Ryan P Adams. On warm-starting neural network training. Advances in neural information processing systems, 33:3884–3894, 2020.
[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[4] Urtė Urbonavičiūtė, Laurens Engwegen, and Wendelin Böhmer. Exploring alternatives to full neuron reset for maintaining plasticity in continual backpropagation. Manuscript in preparation, 2025.
[5] Augustinas Jučas, Laurens Engwegen, and Wendelin Böhmer. Layerwise perspective into continual backprop: replacing the first layer is all you need. Manuscript in preparation, 2025.

Methodology

Evaluated forgetting for two different scenarios:

- **Initial exposure recall.** Evaluates how quickly model forgets information it encountered for the very first time.
- **Recurrent task recall**. Evaluates how well model maintains knowledge about previously learned and then reintroduced information.

Compared forgetting of CBP to four baseline algorithms:

1. Standard backpropagation.
2. Shrink and Perturb [2].
3. L2 regularization.
4. Adam optimizer [3].

Analyzed the internal network dynamics during training, in particular **weight** and **activation drift**.

Examined the influence of CBP's hyperparameters on the stability–plasticity trade-off, including learning rate, replacement rate (ρ), decay rate (η) , and maturity threshold.

Evaluated three adjustments to CBP, expected to improve the stability-plasticity trade-off:

1. **Noise Injection** [4]. Neurons are reset by shrinking the weights and adding noise, instead of full reinitialization.
2. **Layer-specific replacement** [5]. Neurons are reset only in the first hidden layer of the network.
3. **Partial neuron reinitialization**. When resetting a neuron, only fraction of incoming/outgoing weights is reinitialized.

Metrics

Forgetting	Plasticity	Feature drift
Initial recall accuracy	*Long-term accuracy*	$\mathrm{PrSim}(X,Y) = \frac{\mathrm{tr}(\Sigma)}{\Vert X\Vert_F \cdot \Vert Y\Vert_F}$
Recurrent accuracy curve		$\mathrm{CKA}(X,Y) = \frac{\Vert Y^\top X\Vert_F^2}{\Vert X^\top X\Vert_F \cdot \Vert Y^\top Y\Vert_F}$
Memory retention duration		

Conclusions

- CBP showed higher forgetting than all the baseline methods.
- Replacement rate, decay rate, and learning rate significantly affected the stability–plasticity trade-off.
- Three evaluated CBP variants reduced forgetting without significantly compromising plasticity.
- Activation drift strongly correlated with forgetting.
- Decay rate experiments suggest utility estimation in CBP can be improved to better preserve past task-relevant features.

Results