

Evaluating Catastrophic Forgetting in Neural Networks Trained with Continual Backpropagation

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, taskspecific 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. Laverwise perspective into continual backprop: replacing the first layer is all you need. Manuscript in preparation, 2025.

Methodology

- information.

Evaluated forgetting for two different scenarios: Average activation drift for Accuracy on the first permutation data • Initial exposure recall. Evaluates how quickly model different algorithms (Procrustes distance) for different algorithms 100 forgets information it encountered for the very first time. - CBP • Recurrent task recall. Evaluates how well model maintains S&P 0.8 80 - L2 knowledge about previously learned and then reintroduced (%) dista Adam 60 Procrustes 40 Compared forgetting of CBP to four baseline algorithms: 1. Standard backpropagation. 2. Shrink and Perturb [2]. 0.0 50 100 150 2300 2350 2400 150 2300 2350 2400 2450 2500 50 100 3.L2 regularization. Task numbe Task number 4. Adam optimizer [3]. Stability-plasticity tradeoff comparison for initial exposure recall experiment Analyzed the internal network dynamics during training, in 40.00 particular weight and activation drift. 37.50 **(%)** Examined the influence of CBP's hyperparameters on the Accuracy 32.50 30.00 stability-plasticity trade-off, including learning rate, replacement rate (ρ), decay rate (η), and maturity threshold. **Original Algorithm Versions Standard Continual Backpropagation** Variant of CBP: Noise Injection **Shrink and Peturb** CBP + Noise injection. $\lambda = 0.99$, $\sigma = 0.001$ Evaluated three adjustments to CBP, expected to improve the CBP + Noise injection. $\lambda = 0.6$, $\sigma = 0.01$ L2 Recall Solution Regular Backpropagation + Noise injection. $\lambda = 0.99$, $\sigma = 0.005$ stability-plasticity trade-off: Adam. $\alpha = 0.0001$ Hyperparameter Influence on Forgetting: (1. Noise Injection [4]. Neurons are reset by shrinking the Variant of CBP: Layer-Specific Replacement Regular CBP. n = 0.9925.00 22.50 Regular CBP. $\rho = 0.000001$ Regular CBP. $\eta = 0.9$ weights and adding noise, instead of full reinitialization. CBP + Replace Only First Layer. $\rho = 0.000001$ Regular CBP. $\eta = 0.1$ Regular CBP. $\eta = 0.01$ 2. Layer-specific replacement [5]. Neurons are reset only in the Hyperparameter Influence on Forgetting first hidden layer of the network. Regular CBP. $\rho = 0.0001$ Variant of CBP: Partial Neuron Replacement Regular CBP. $\rho = 0.00001$ CBP + Partial neuron replacement. r = 0.5 3. Partial neuron reinitialization. When resetting a neuron, Regular CBP. $\rho = 0.000005$ CBP + Partial neuron replacement. r = 0.3 20.00 Regular CBP. $\rho = 0.000001$ CBP + Partial neuron replacement. r = 0.01 only fraction of incoming/outgoing weights is reinitialized. 92.00 94.00 86.00 88.00 90.00 Plasticity Metric (%) Metrics CBP accuracy on the first permutation for different Accuracy on the first permutation data for replacement rate (ρ) and decay rate (η) values noise injection variant with $\sigma^2 = 0.001$ ρ - replacement Default 100 Plasticity **Feature drift** Forgetting λ - shrink rate $\lambda = 0.99, \sigma^2 = 10^{-3}$ $\rho = 10^{-5}, \eta = 0.99$ σ^2 - noise η - decay rate - $\lambda = 0.90, \sigma^2 = 10^{-3}$ variance $\rho = 10^{-4}$ $\operatorname{PrSim}(X,Y) = \frac{1}{\|X\|_F \cdot \|Y\|_F}$ - $\lambda = 0.80, \sigma^2 = 10^{-3}$ *Long-term accuracy Initial recall accuracy* $\rho = 10^{-6}$ - $\lambda = 0.70, \sigma^2 = 10^{-3}$ η = 0.1 $\lambda = 0.60, \sigma^2 = 10^{-3}$ *Recurrent accuracy curve* - η=0.01 $\lambda = 0.40, \sigma^2 = 10^{-3}$ $\operatorname{CKA}(X,Y) = rac{\|X^{ op}X\|_F}{\|X^{ op}X\|_F \cdot \|Y^{ op}Y\|_F}$ $\|Y^{ op}X\|_F^2$ Memory retention duration 40 AC 20 Conclusions 50 100 150 2300 2350 2400 2450 2500 0 50 100 150 2300 2350 2400 2450 2500 Ó Task number Task number • CBP showed higher forgetting than all the baseline methods. Accuracy on the first permutation data for CBP, Accuracy on the first permutation data for compared to layer-specifc replacement variant partial neuron replacement variant • Replacement rate, decay rate, and learning rate significantly 100 CBP, replace all layers *r* - fraction of - CBP. Original affected the stability-plasticity trade-off. weights CBP, replace first layer CBP. r = 0.5 reinitialized 80 80 - CBP. r = 0.3 • Three evaluated CBP variants reduced forgetting without CBP. r = 0.01 0%) 00 60 significantly compromising plasticity. 40 Accl 40 • Activation drift strongly correlated with forgetting. • Decay rate experiments suggest utility estimation in CBP can 20 20 150 2300 2350 2400 2450 2500 150 2300 2350 2400 2450 2500 0 50 100 Ó 50 100

Task number

- be improved to better preserve past task-relevant features.

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

Task number

Results