Layerwise Perspective into Continual Backpropagation: Replacing the First Layer is All You Need

Author: Augustinas Jučas, A.Jucas@student.tudelft.nl | Supervisor: Laurens Engwegen | Responsible professor: Wendelin Böhmer

1. Introduction and Background

Continual Learning deals with training a model on a changing data distribution.

Plasticity Loss is a phenomenon within Continual Learning: as data distribution changes, a model's ability to adapt to new information diminishes.

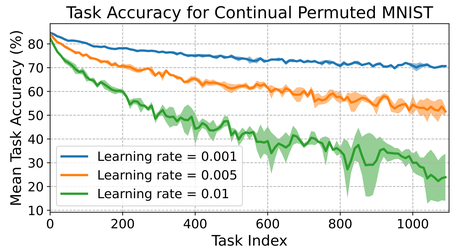


Figure 1: **A typical example of plasticity loss.** A model is sequentially trained on different versions of the MNIST dataset. Each version uses a different pixel permutation. As the model learns more tasks, it gets worse at learning new ones, shown by decreasing accuracy.

Dohare et al. introduced a method called **Continual Backpropagation** for mitigating plasticity loss [1]:

1. A small number of neurons are continually selected per layer to be reset.
2. A neuron is reset by reinitialising incoming weights and setting outgoing weights to 0. .

Our goal: investigate the **layerwise dynamics** of Continual Backpropagation.

2. Experimental Setting

We train models on datasets, using different learning algorithms:

(a) **Versions of Continual Backprop**:
1. Regular CBP.
2. Replacing all layers separately.
3. No replacement - regular BP.
4. Replacing layers 2-L.

(b) **L_2 Regularisation**

Dataset: Continual Permuted MNIST. **Base model**: 5-layer ReLU-based MLP.

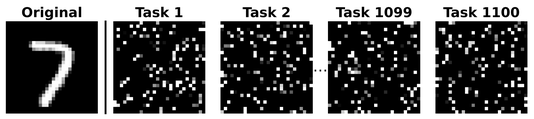


Figure 2: **Continual Permuted MNIST dataset**. For every task, the pixels of all MNIST samples are permuted using a task-specific permutation, creating a unique dataset.

3. The First Layer Phenomenon

Core findings.
#1: The deeper a layer is, the less replacement helps.
#2: Layer 1 replacement yields CBP-level performance.
We call result #2, the *first layer phenomenon*.

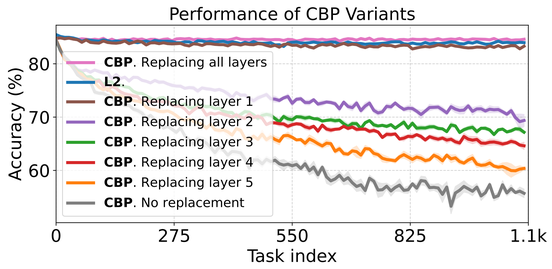


Figure 3: Training evolution of CBP variants on the *base* model. The figure directly depicts findings #1 and #2.

We obtain similar findings for:

1. Wider and narrower networks.
2. Deeper and shallower networks.
3. Different activation-based networks.

Table 1: The effects of varying replacement rates by layer on different neural networks. Every row corresponds to a particular network, every column – to a different version of CBP the neural network was trained on. Best layer for replacement is underlined. Numerical values depict percentages – the average value over last 15% of tasks.

	Baseline	Replacing Individual Layers			
	All Layers	**Layer 1**	**Layer 2**	**Layer 3**	**Last Layer**
Varying Network Widths					
Width 20	79.2 ± 0.1	<u>76.1 ± 0.2</u>	59.5 ± 1.3	58.2 ± 0.6	59.1 ± 0.9
Width 50	82.3 ± 0.1	<u>80.7 ± 0.2</u>	56.1 ± 2.3	60.2 ± 1.0	52.4 ± 2.3
Width 100	84.6 ± 0.0	<u>83.2 ± 0.1</u>	71.0 ± 0.5	67.6 ± 0.4	61.0 ± 1.0
Width 150	85.5 ± 0.0	<u>84.2 ± 0.0</u>	75.2 ± 0.1	71.1 ± 0.3	66.6 ± 0.3
Width 200	86.0 ± 0.0	<u>84.8 ± 0.0</u>	77.7 ± 0.3	74.0 ± 0.2	69.9 ± 0.4
Varying Network Activation Functions					
ReLU	84.6 ± 0.0	<u>83.2 ± 0.1</u>	71.0 ± 0.5	67.6 ± 0.4	61.0 ± 1.0
SELU	76.2 ± 0.1	76.3 ± 0.1	76.2 ± 0.1	<u>76.6 ± 0.2</u>	76.5 ± 0.1
Tanh	69.2 ± 0.3	<u>69.1 ± 0.5</u>	69.1 ± 0.1	69.0 ± 0.5	68.6 ± 0.3
Varying Network Depths					
Depth 2	87.0 ± 0.0	<u>86.8 ± 0.1</u>	81.9 ± 0.0	-	81.9 ± 0.0
Depth 5	84.6 ± 0.0	<u>83.2 ± 0.1</u>	71.0 ± 0.5	67.6 ± 0.4	61.0 ± 1.0
Depth 8	82.9 ± 0.0	<u>82.5 ± 0.1</u>	71.7 ± 0.2	67.6 ± 0.8	65.5 ± 0.2
Varying Network Depths, Same Number of Parameters					
Depth 2	87.3 ± 0.0	<u>87.1 ± 0.1</u>	82.7 ± 0.0	-	82.7 ± 0.0
Depth 5	84.6 ± 0.0	<u>83.2 ± 0.1</u>	71.0 ± 0.5	67.6 ± 0.4	61.0 ± 1.0
Depth 8	82.3 ± 0.1	<u>82.2 ± 0.1</u>	70.1 ± 0.5	65.7 ± 0.2	64.3 ± 0.6

Table 1 interpretation:

1. Most rows have decreasing values reading from left to right (finding #1).
2. In most rows, replacing the first layer gives the best accuracies (finding #2).

4. CBP Fails with Non-ReLU Activations

We found a flaw within regular Continual Backpropagation – it does not stop non-ReLU models from losing plasticity.

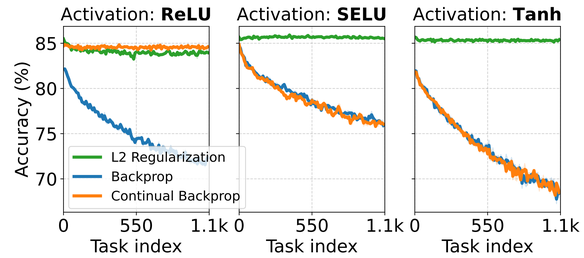


Figure 4: Accuracy evolution for MLPs with 3 different activation functions. Regular CBP does not perform for non-ReLU networks.

5. Understanding First Layer Phenomenon

Focus on 4 critical versions of CBP:

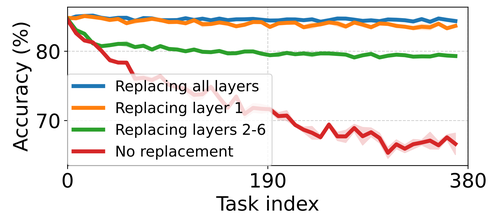


Figure 5: Training evolution of CBP variants on the base model.

The result above suggests that first layer is *stronger* than all other layers combined.

To explain the first layer phenomenon, we attempted to relate the results with:

- Dead neuron counts
- <u>Weight norm statistics</u>
- Feature ranks
- Curvature of the loss landscape
- Feature utilities

Gradual increase in weight norms explains the first layer phenomenon.
Weight magnitude inflation is known to induce plasticity loss. Figure 6 shows that if the first layer is not replaced, it significantly drives weight magnitude increase.

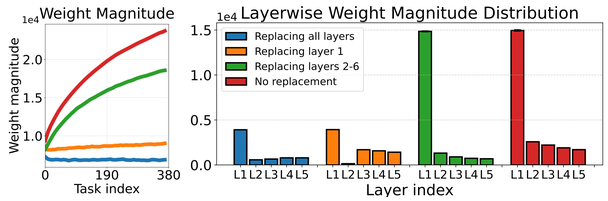


Figure 6: Left: whole network's weight magnitude evolution over training. Right: layerwise weight distribution – average layer weight norm.

6. Discussion and Future Work

Weight magnitudes. We showed that replacing the first layer stabilises the whole model's weight magnitudes. However, the reason behind that is still unknown.

Single dataset. Only a single dataset was used to derive the results; hence, results should be replicated with different datasets.

Non-ReLU models. We showed CBP does not perform on non-ReLU models. Questions arise: why? what methods could tackle this problem?

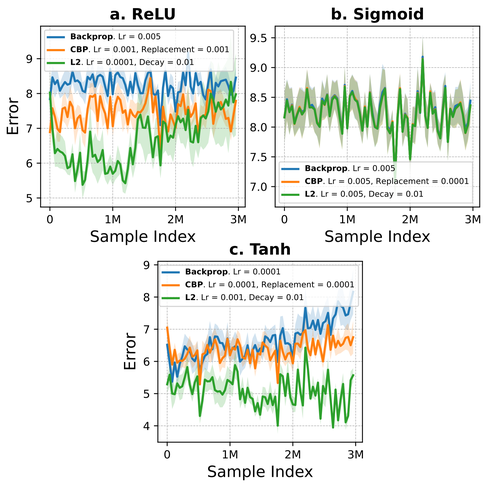
7. Conclusions

We find that:

1. Resetting neurons in earlier layers leads to increasingly better performance.
2. Resetting neurons only in the first layer achieves performance close to CBP – *first layer phenomenon*.
3. The first layer phenomenon can be explained through the lens of controlling a model's weight magnitudes.
4. There is a flaw in CBP: it does not work on non-ReLU activations.

8. Appendix: Bit-Flipping Problem

Experiments on another dataset, introduced by [1], were performed, but were unsuccessful, as CBP failed to outperform L_2 baseline:



[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, August 2024.