Loss of Plasticity in Recurrent Models

Ahmad Khan Yanis Merzouki Andrei Pacurar Petar Lukovic

Abstract

Rapid advancement in the scale of data poses a challenge of enabling deep learning models to be trained for a plethora of different tasks. This project explores this problem, in the form of loss of plasticity, and gives insights into the causes for this phenomenon. We conduct multiple experiments on different datasets and on different models to further explore how model choice affects loss of plasticity. Specifically, we shed light on how loss of plasticity impacts recurrent models. Finally, we explore some mitigation methods from the literature to assess their effectiveness on recurrent models.

1. Introduction

Deep learning has achieved significant advances in recent years, driving breakthroughs in fields such as Language Models, Molecular Science, Generative AI, and Computer Vision. However, data is constantly changing and evolving, requiring models to adapt to new tasks while retaining knowledge of previous ones. This challenge leads to two critical issues: models may forget previously learned tasks, known as catastrophic forgetting, or they may lose the ability to learn new tasks, referred to as loss of plasticity (LoP). To this day there is no easy way to update models after training and adapting them to new tasks. Retraining the model is effective, but is unfeasible for modern LLMs for instance. While catastrophic forgetting has been moderately studied, loss of plasticity remains a relatively under-explored topic. Thus, this paper focuses on that specific aspect.

As discussed by (Lyle et al., 2024), plasticity loss can be conceptualized as the suitability of a specific point to function as an initial condition for optimization. Thus, loss of plasticity would refer to a network that has entered an ill-conditioned region of the loss landscape where optimization is slow. Generally, even though models can escape from ill-conditioned minima of the parameter space they are currently in, this can result in having significant loss of performance of learning algorithms, and thus it is desirable to avoid these regions.

While several works have demonstrated loss of plasticity in different contexts (Igl et al., 2021; Ash & Adams, 2020; Berariu et al., 2023), and several mitigation strategies have been explored (Dohare et al., 2022; Abbas et al., 2023; Dohare et al., 2024), loss of plasticity across architectures remains relatively unexplored.

In this work, we focus on extending and adapting existing experiments and mitigation strategies related to loss of plasticity, specifically to recurrent models such as LSTMs and GRUs, while comparing their performance to results observed in MLPs. Our primary objective is to explore the unique behaviors of recurrent architectures under the influence of loss of plasticity and evaluate potential solutions.

Our contributions are as follows[1]:

- Implementation of baseline experiments described in previous work
- Design of novel experiments tailored to recurrent architectures
- Evaluation of mitigation techniques applied to recurrent models
- Evalutation of the impact of sequence length on LoP in recurrent models

2. Previous work

While loss of plasticity is a young field of research, we will go over some previous works that have demonstrated its existence. Furthermore, we cover some mitigation strategies that have been discovered. Research on this topic has naturally separated itself depending on the task where loss of plasticity was identified, so we will follow this natural separation:

Reinforcement Learning. Typically, reinforcement learning scenarios involve an agent that has to learn continuously during the operation. Therefore, the phenomenon of LoP has first been noticed in deep reinforcement learning scenarios (Dohare et al., 2022; Igl et al., 2021). Mitigation strategy research has mainly been focused on regularizing parameters, or injecting noise, in order to keep parameters near initialization (Dohare et al., 2022; Kumar et al., 2024; Abbas et al., 2023; Sokar et al., 2023). Another work suggest solving this by using bootstrapping (Kumar et al., 2020).

Supervised Learning. Another deep learning scenario where LoP has been discovered is supervised learning. When a model is learning a sequence of tasks, it has been uncovered that the model struggles with learning novel tasks. This has been touched upon in (Ash & Adams, 2020; Berariu et al., 2023).

More general approaches to this phenomenon is researched in three papers (Lyle et al., 2024; 2023; 2022). These three

[1]code available at `https://github.com/mak2508/recurrent-lop`

papers also present more empirical approaches to the topic of loss of plasticity by using carefully crafted experiments on different datasets such as MNIST and CIFAR10. The key takeaway from these papers, especially (Lyle et al., 2023), has been that loss of plasticity cannot simply be addressed by using one singular mechanism such as, for example, reducing vanishing gradients. Another important takeaway is that currently there is no single method in literature that solves this problem entirely, and this is why this is still very much active field of research.

3. Experiments

In this study, we investigate the phenomenon of loss of plasticity in recurrent models through a series of experiments designed to demonstrate LoP. First, we run two baseline experiments inspired by (Lyle et al., 2023; Dohare et al., 2024) were designed. Since recurrent models excel at natural language processing tasks, the experiments were later adapted to align with this specific use case.

Baseline Experiments. The first baseline experiment, adapted from (Lyle et al., 2023), trains a model on the MNIST dataset, but once the model is trained for 30 epochs, the labels are shuffled and the model is trained on this new reshuffled dataset. This is repeated for 50 iterations, reshuffling the labels randomly at each step. This setup enables us to observe the model's ability to adapt to evolving label distributions. A decline in accuracy over iterations would signify the presence of LoP. The second baseline experiment, adapted from (Dohare et al., 2024), utilizes MNIST in a binary classification task, where two random classes are selected at each iteration. The model is trained on these two classes, and the process is repeated for 50 iterations, each time with a different pair of classes. This experiment evaluates the model's capacity to learn new tasks dynamically, with declining accuracy indicating LoP. These experiments serve as a baseline to demonstrate the vulnerability to LoP in recurrent models using experiments that exist in the literature for simpler Multi-Layer Perceptrons (MLP) (Rosenblatt, 1958). We run these experiments on MLPs to validate our experiment design, and on Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) to test for loss of plasticity in sequence models.

Adapted Experiments. The first adapted experiment replicates the structure of the first baseline but applies it to a language classification task using the Tatoeba dataset (Tiedemann, 2020), which contains sentences in many languages. Here, reshuffling corresponds to changes in language labels, while other parameters remain consistent. The second adapted experiment mirrors the second baseline but involves selecting two random languages from the Tatoeba dataset at each iteration. The recurrent model is then trained on the binary classification task using these selected languages, repeating the process 100 times. We hope that those experiments might infer a better understanding of loss of plasticity in recurrent models in a more traditional context for these models. We run these experiments using Gated Recurrent Unit (GRU) networks (Cho et al., 2014) instead of LSTMs to reduce computational load.

Mitigation Experiments. To evaluate the impact of various mitigation techniques, we conducted comparative experiments on both MLPs and recurrent models. Specifically, we replicated the MNIST-reshuffle experiment for MLPs and the Tatoeba-reshuffle experiment for GRUs, incorporating three mitigation strategies: L2 regularization, shrink and perturb, and continual backpropagation (Ash & Adams, 2020; Dohare et al., 2024). All other experimental parameters were kept constant.

Additional Experiment. In order to test an additional hypothesis on the effect of sequence length in the context of loss of plasticity, the adapted reshuffle experiments were conducted with a GRU of varying lengths $15, 25, 50, 75$, as well as varying hidden sizes $32, 64, 128$. All other parameters are exactly the same as previously described.

These experiments collectively provide a systematic evaluation of continual task learning and its impact on model performance, offering insights into the extent and dynamics of loss of plasticity in both general and language-specific contexts.

4. Results

In this section we discuss the results of the experiments we have explained in previous section. In each figure we display the test accuracy at the final epoch of each experiment to reduce clutter. For all experiments, we use similar parameters, i.e. 30 epochs on 50 experiment iterations for MNIST tasks and 100 iterations for language tasks.

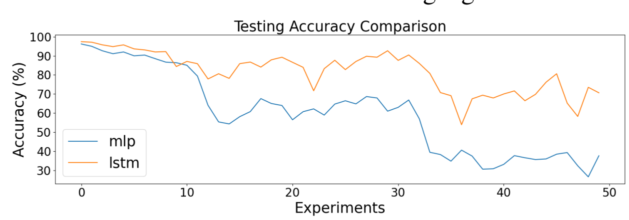

Figure 1. **Baseline MNIST Reshuffle Experiments.** Figure compares test accuracy of LSTM and MLP models.

In Figure 1 and Figure 2 we compare how our models perform on the baseline tasks. In both of these baseline experiments we can conclude that models suffer from loss of plasticity due to reduction in accuracy scores over the training process. Moreover, the longer the training, bigger the drop in accuracy we can report. This happens for both baseline experiments, being less noticeable for LSTM model.

In Figure 3 we see the results of the language task experiments on GRUs. Once again, we note that loss of plasticity

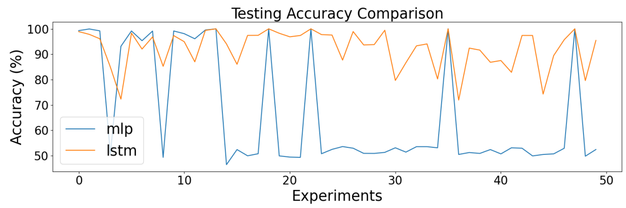

Figure 2. **Baseline MNIST Binary Experiments.** Figure compares test accuracy of LSTM and MLP models.

is demonstrated. We note that the binary tasks both in Figure 2 and Figure 3 have a number of spikes in the accuracy. We hypothesize that this is due to task repetition, since the number of classes in both MNIST and the language data used was limited due to computational limits, some pairs in the binary task sequence had to be repeated. This makes the use of this experiment in a quantitative comparison less insightful than what was previously hoped, especially when measuring accuracy degradation.

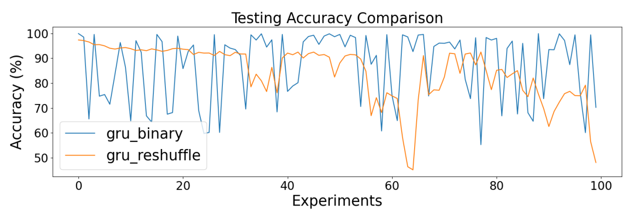

Figure 3. **Language Reshuffle and Binary Experiments.** Figure shows the language tasks run on a sequence model.

In Figure 4 and Figure 5 we investigate mitigation strategies. In Figure 4 we see that all strategies result in improvements in accuracy, but only shrink-and-perturb and L2 regularization give a stable improvement over a long number of iterations. In Figure 5, we see that only continual backprop and L2 regularization provide stable improvements, with shrink-and-perturb actually performing worse than the baseline. These results suggest that existing mitigation strategies are not always reliable on recurrent models.

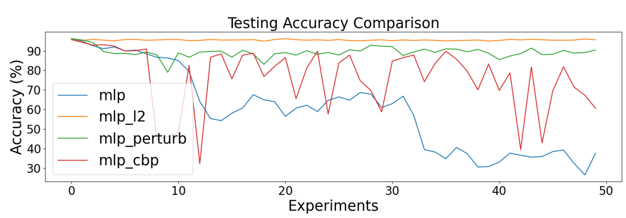

Figure 4. **MNIST Reshuffle Experiment with Mitigation.** Figure shows the effectiveness of mitigation strategies with MLP.

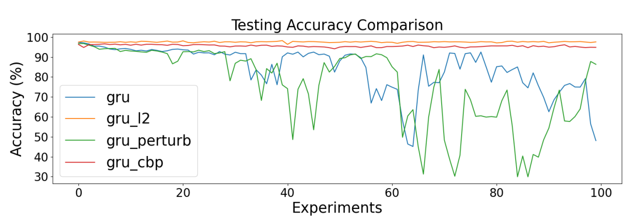

Figure 5. **Language Reshuffle Experiment with Mitigation.** Figure shows the effectiveness of mitigation strategies with GRU.

5. Analysis

In this section, we conduct a detailed analysis of the previously presented figures. We introduce two relevant metrics and apply them to our experimental results, and draw conclusions that explain these results. Further, we propose a hypotheses regarding the impact of plasticity loss on recurrent models, including the mechanisms by which this phenomenon affects their performance.

5.1. Quantitative Metrics

To quantify LoP, we track the degradation in model accuracy over time. Specifically, we measure the accuracy at the beginning of training, $\text{Accuracy}(0)$, and at the final training step, $\text{Accuracy}(T)$. The accuracy degradation, $\Delta\text{Accuracy}$, is then computed as the relative decrease in accuracy over the training period. A higher value of $\Delta\text{Accuracy}$ indicates a greater LoP. This is given by the formula:

$$\Delta\text{Accuracy} = \frac{\text{Accuracy}(0) - \text{Accuracy}(T)}{\text{Accuracy}(0)} \times 100$$

Additionally, we use the increase in average accuracy (IAA) as a relative measure of effectiveness by a mitigation technique observed on a model. The metric is defined as:

$$\text{IAA}(\text{Mitigation}) = \frac{\langle \text{Accuracy}_{\text{mitigated}} \rangle - \langle \text{Accuracy}_{\text{base}} \rangle}{\langle \text{Accuracy}_{\text{base}} \rangle} \times 100$$

where the accuracy values are the mean accuracy across the experiments.

5.2. Robustness of recurrent models to LoP

We observe that recurrent models significantly outperform the general MLP's resistance to the LoP. Specifically, a performance drop of **50.62%** was witnessed on the GRU, compared to **60.83%** for the MLP in their respective reshuffle experiments. Even more surprising is the performance of the LSTM in the context of MNIST reshuffle experiment, which yielded a degradation of only **27.47%**, significantly outperforming the MLP.

We hypothesize that this phenomenon arises from the LSTM's and GRU's gating mechanisms designed to capture and retain information across long sequences. These gates regulate what information is retained, updated, or discarded, enabling the model to selectively store important features over time. This dynamic memory capability might allow recurrent models to mitigate the effects of decreased plasticity by preserving important learned knowledge and forgetting harmful knowledge that can cause plasticity loss, thus maintaining gradient flow across timesteps.

5.3. Relation of Sequence Length to LoP

Recurrent models like LSTMs offer the significant advantage of handling varying input sequence lengths. This adaptability raises an important question about the relationship between sequence length and a model's vulnerability to loss of plasticity. Longer sequences might exacerbate plasticity loss due to compounded gradient challenges across timesteps, while shorter sequences may mitigate its impact by reducing the propagation of errors. Therefore, studying the effect of input sequence length on LSTM performance under LoP provides valuable insights into how sequence-dependent architectures cope with diminishing plasticity.

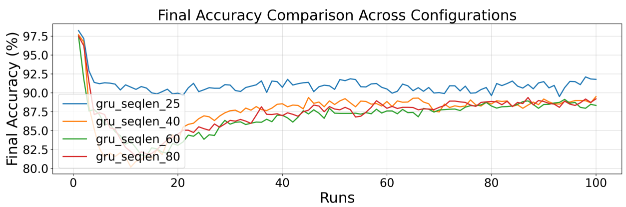

Figure 6. **Language Reshuffle Experiment with different sequence lengths.** Figure shows the impact of sequence length on LoP for GRU.

We test this hypothesis in Figure 6, which examines the impact of LoP on GRU models configured with identical parameters but varying sequence lengths.

Although a greater accuracy degradation can be seen in the larger sequence length models when compared to the smallest sequence length GRU, the results are not strongly conclusive since we do not see any conclusive pattern amongst the 3 larger models. This might warrant further research to investigate. Despite this limitation, these findings could indicate that the compounded effect of LoP is influenced by the GRU's sequence length, where the final output is affected by the accumulated LoP present across all preceding timesteps. While each output at timestep t is individually influenced by LoP, as expected, the error induced by LoP may propagate and compound across subsequent timesteps $t+1, t+2, \ldots$, exacerbating the degradation. If this is shown to be true, it would suggest that recurrent models might be inherently more vulnerable to LoP when dealing with longer inputs, making them more susceptible to this cumulative effect.

5.4. Evaluation of Mitigation Techniques

In this subsection, we evaluate the IAA of L2, CBP, and shrink-and-perturb on the MLP and GRU on the corresponding reshuffle tasks.

Shrink-and-perturb proved to be ineffective for GRU models, with the average achieved accuracy dropping, resulting in a negative IAA of **-10.73%**. This finding is noteworthy, as shrink-and-perturb yielded a significant IAA increase of **36.69%** when applied to the standard MLP model.

Conversely, **continual backpropagation (CBP)** and **L2 regularization** demonstrated stronger performance on the GRU model, with IAA values of **15.56%** and **13.14%** respectively. While these numbers are lower than the corresponding MLP values, this is due to the base accuracy for the GRU value being higher. Possible reasons for this were discussed in the previous section. The comparison is shown in Figure 5, indicating that both techniques yielded considerable improvements in accuracy, with greater final accuracies observed in the GRU model compared to the MLP.

	L2	S&P	CBP
MLP	36.69%	49.45%	28.01%
GRU	15.56%	-10.08%	13.14%

Table 1. **IAA values for mitigation strategies on reshuffle**

The limited effectiveness of shrink-and-perturb on GRU models may be attributed to the differing requirements of sequential versus feedforward architectures. Recurrent models often rely on specialized regularization techniques, such as dropout in recurrent layers, to address their unique susceptibility to overfitting. The generalized shrink-and-perturb approach may not align well with the specific regularization needs of recurrent models, thereby limiting its utility. This observation suggests an avenue for future research into tailored regularization methods for recurrent architectures.

CBP's impressive performance on GRU can be attributed to its ability to enable gradient updates across all timesteps, thereby reinforcing the relevance of earlier timesteps. This dynamic effectively combats LoP by preventing the model from overemphasizing recent timesteps at the expense of earlier ones, resulting in a more balanced and accurate sequential learning.

The success of L2 regularization, however, aligns with expectations. Gradient regularization methods are particularly effective in recurrent models, as these architectures are more susceptible to vanishing and exploding gradients. The consistent improvement observed with L2 regularization underscores its utility in mitigating these

6. Conclusion

In this work, we explored the phenomenon of loss of plasticity in deep learning models, particularly focusing on recurrent architectures like LSTMs and GRUs. Through systematic experiments, we observed that recurrent models exhibit resilience to LoP compared to feedforward architectures, thanks to their inherent memory mechanisms. However, challenges persist, especially with longer sequences. Mitigation strategies, such as continual backpropagation and L2 regularization, proved effective in addressing LoP in specific experiments, even though shrink-and-perturb performed poorly. These findings underscore the importance of tailored approaches in mitigating LoP and highlight the need for further research to enhance model adaptability in dynamic environments.

References

Abbas, Z., Zhao, R., Modayil, J., White, A., and Machado, M. C. Loss of plasticity in continual deep reinforcement learning, 2023. URL https://arxiv.org/abs/2303.07507.

Ash, J. T. and Adams, R. P. On warm-starting neural network training, 2020. URL https://arxiv.org/abs/1910.08475.

Berariu, T., Czarnecki, W., De, S., Bornschein, J., Smith, S., Pascanu, R., and Clopath, C. A study on the plasticity of neural networks, 2023. URL https://arxiv.org/abs/2106.00042.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches, 2014. URL https://arxiv.org/abs/1409.1259.

Dohare, S., Sutton, R. S., and Mahmood, A. R. Continual backprop: Stochastic gradient descent with persistent randomness, 2022. URL https://arxiv.org/abs/2108.06325.

Dohare, S., Hernandez-Garcia, J. F., Lan, Q., Rahman, P., Mahmood, A. R., and Sutton, R. S. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07711-7. URL https://doi.org/10.1038/s41586-024-07711-7.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Igl, M., Farquhar, G., Luketina, J., Boehmer, W., and Whiteson, S. Transient non-stationarity and generalisation in deep reinforcement learning, 2021. URL https://arxiv.org/abs/2006.05826.

Kumar, A., Agarwal, R., Ghosh, D., and Levine, S. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. *CoRR*, abs/2010.14498, 2020. URL https://arxiv.org/abs/2010.14498.

Kumar, S., Marklund, H., and Roy, B. V. Maintaining plasticity in continual learning via regenerative regularization, 2024. URL https://arxiv.org/abs/2308.11958.

Lyle, C., Rowland, M., and Dabney, W. Understanding and preventing capacity loss in reinforcement learning, 2022. URL https://arxiv.org/abs/2204.09560.

Lyle, C., Zheng, Z., Nikishin, E., Pires, B. A., Pascanu, R., and Dabney, W. Understanding plasticity in neural networks, 2023. URL https://arxiv.org/abs/2303.01486.

Lyle, C., Zheng, Z., Khetarpal, K., van Hasselt, H., Pascanu, R., Martens, J., and Dabney, W. Disentangling the causes of plasticity loss in neural networks, 2024. URL https://arxiv.org/abs/2402.18762.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958. URL https://api.semanticscholar.org/CorpusID:12781225.

Sokar, G., Agarwal, R., Castro, P. S., and Evci, U. The dormant neuron phenomenon in deep reinforcement learning, 2023. URL https://arxiv.org/abs/2302.12902.

Tiedemann, J. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt, 2020. URL https://arxiv.org/abs/2010.06354.