Continual Learning in Recurrent Neural Networks

Benjamin Ehret* **Christian Henning*** **Maria R. Cervera*** Alexander Meulemanns, Johannes von Oswald and Benjamin F. Grewe

Institute of Neuroinformatics, UZH / ETH Zurich, Switzerland

Introduction

Most CL research has been done in feedforward networks. **RNNs**, however, differ from feedforward networks:

- **hidden-to-hidden weights** are sequentially **reused** over time
- **working memory** is needed for solving the tasks

To determine whether existing methods to prevent catastrophic forgetting can be used off-the-shelf for RNNs, we first focus on established regularization approaches, i.e., **weight-importance methods** such as EWC [1], and study their particularities when applied to RNNs. Furthermore, we provide a comprehensive evaluation of established CL methods on a variety of sequential data benchmarks. Our results suggest that **task-conditioned hypernetworks** [2] (HNET) are better suited for RNNs, since the used continual learning regularization is agnostic to the recurrent processing.

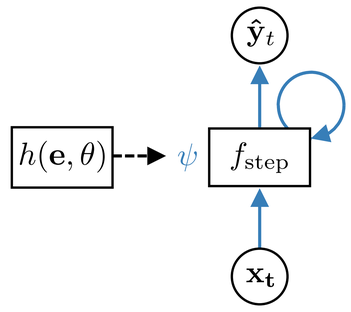

Figure 1: RNN with task-conditioned hypernetwork.

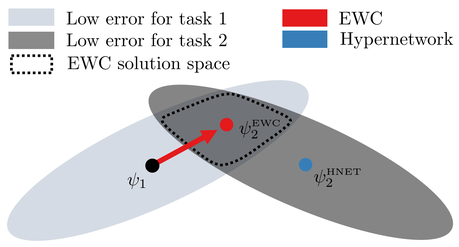

Figure 2: Conceptual difference between EWC and HNET.

Analysis of Weight-Importance Methods

To test whether weight-importance values are influenced by either working memory requirements or sequence length, which leads to varying levels of weight reuse, we consider variants of the Copy Task [3]. Our results indicate that working memory requirements (and not weight reuse) affect weight-importance values, which can lead to high rigidity when learning many tasks in a continual learning scenario.

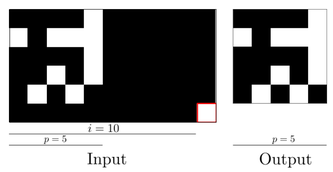

Figure 3: Example Copy Task input-output pattern pair. The end of the input sequence is marked by the stop bit, shown in red.

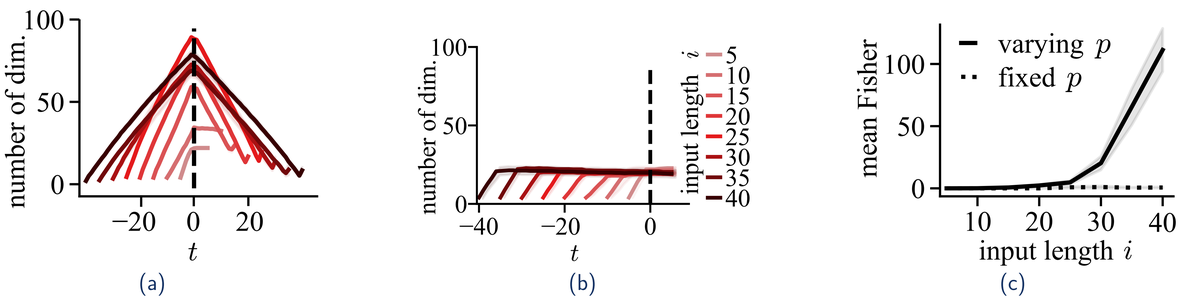

Figure 4: **(a)** Intrinsic dim. per timestep of the RNN hidden space where input and pattern lengths are tied ($i = p$). **(b)** Same as (a) where the pattern length is fixed ($p = 5$) but input length i varies. **(c)** Mean weight-importance value of recurrent weights.

CL Experiments: Audioset

Table 1: Mean `during` and `final` accuracies for the Split-AudioSet-10 experiments (Mean ± SEM in %, $n = 10$).

	during	**final**
`Multitask`	N/A	77.31 ± 0.10
`From-scratch`	N/A	79.06 ± 0.11
`Fine-tuning`	71.95 ± 0.24	49.02 ± 1.00
`HNET`	73.05 ± 0.45	71.76 ± 0.62
`Online EWC`	68.82 ± 0.20	65.56 ± 0.35
`SI`	67.66 ± 0.10	66.92 ± 0.04
`Masking`	75.81 ± 0.15	50.87 ± 1.09
`Masking+SI`	64.88 ± 0.19	64.86 ± 0.20
`Coresets-100`	74.25 ± 0.11	72.30 ± 0.11
`Coresets-500`	77.03 ± 0.08	73.90 ± 0.07

Audioset [4] is a dataset of annotated 10-second audio snippets, from which we construct 10 tasks with 10 classes each. We can see in Table 1 that fine-tuning the model results in a large drop in accuracy, showing that catastrophic forgetting occurs when no continual learning protection is used. This is partially solved by weight-importance methods, which exhibit reduced forgetting at the cost of less plasticity for learning single tasks well. On the contrary, an approach based on hypernetworks (HNET) allows fitting each task well without compromising previously acquired knowledge. Masking [5] exhibits strong interference between subnetworks, which is alleviated by Masking+SI at the cost of reduced plasticity for learning new tasks.

CL Experiments: SMNIST

In Split-Sequential-SMNIST, a task consists of all binary sequences of length m, randomly split into two groups. For each task a different set of two digits from the SMNIST dataset [6] is chosen as items for these sequences. Importantly, task difficulty can be manipulated by changing the sequence length m. Weight-importance methods are strongly affected by an increase in task difficulty, while other regularization approaches (HNET and Masking+SI) are less susceptible.

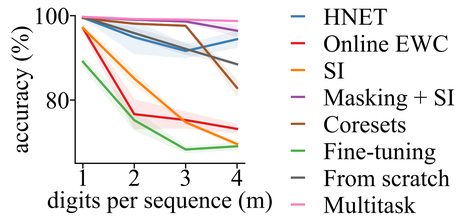

Figure 5: Mean `final` accuracies for the Split-Sequential-SMNIST experiments (Mean ± SEM in %, $n = 10$).

Conclusion

- RNNs are affected by catastrophic forgetting in unique ways
- Working memory requirements, but not the recurrent reuse of weights, directly affect the stability-plasticity dilemma in weight-importance methods
- A systematic comparison of a variety of CL methods in several datasets established that:
  - Despite the mentioned shortcomings, weight-importance methods often remain competitive
  - An approach based on hypernetworks is, however, preferable for CL in RNNs

References

[1] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell.
Overcoming catastrophic forgetting in neural networks.
Proc. Natl. Acad. Sci. U.S.A., 114(13):3521–3526, Mar 2017.

[2] Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F. Grewe.
Continual learning with hypernetworks.
In *International Conference on Learning Representations*, 2020.

[3] Alex Graves, Greg Wayne, and Ivo Danihelka.
Neural turing machines.
arXiv preprint arXiv:1410.5401, 2014.

[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter.
Audio set: An ontology and human-labeled dataset for audio events.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[5] Nicolas Y Masse, Gregory D Grant, and David J Freedman.
Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization.
Proceedings of the National Academy of Sciences, 115(44):E10467–E10475, 2018.

[6] Edwin D. de Jong.
Incremental Sequence Learning.
arXiv, Nov 2016.

Ninth International Conference on Learning Representations, May 2021

*These authors contributed equally to this work.