Enabling Lifelong Learning in AI with Biological Neural Networks Based on Short-Term, Working, and Long-Term Memory

Hanav Modasiya (hanavmw13@gmail.com)
Santa Clara High School
Santa Clara, California, United States of America

Abstract

Achieving Lifelong Learning, the ability of a learning system to continuously acquire and adapt to changing data over time, in Artificial Intelligence (AI) is an integral step towards achieving Artificial General Intelligence (AGI), a hypothetical version of AI that can change our world forever. Currently, the vast majority of models are incapable of exhibiting Lifelong Learning. This research theorizes the first Neural Network architecture inspired by the Three-stage Memory Model–a theory on our brain's memory. By developing a complementary Neural Network learning system comprising the Cerebral Cortex, Prefrontal Cortex, and Hippocampus, mimicking Long-term, Working, and Short-term memory, respectively, this research achieves Lifelong Learning on a simulated computer vision task and develops the first Working memory-inspired model. It also demonstrates the feasibility of using the Three-stage Memory Model for achieving human-like cognition in AI. Therefore, this research reveals a pathway for future research in achieving AGI: the Three-stage Memory Model.

Keywords: Lifelong Learning; Three-stage Memory Model

Introduction

The hopes of the advent of Artificial General Intelligence (AGI), a hypothetical Artificial Intelligence (AI) that can perform and learn various human tasks, are stunted by our current inability to achieve Lifelong Learning in AI (Kirkpatrick et al., 2017; Larkin, 2024).

Lifelong Learning is the ability of a learning system to continuously acquire and successfully adapt to changing information throughout its lifespan. However, current models often forget old information when learning and adapting to new information, also known as catastrophic forgetting (Chen & Liu, 2018; Kirkpatrick et al., 2017). This phenomenon is likely caused by the fixed nature of the model's learning capacity, causing new information to take higher preference over old learned information and overwrite past learned information when learning new information (Kirkpatrick et al., 2017).

Most prior solutions to achieve Lifelong Learning in AI revolve around expanding learning capacity, replaying old data during retraining, or, most successfully, enabling plasticity–mimicking human neuroplasticity–in the model's weights (Kirkpatrick et al., 2017; Perrett, Furber, & Rhodes, 2022; van de Ven G.M. Siegelmann H.T. & Tolias A.S, 2020). Similar to the latter solution, this research postulates that since the human brain is capable of Lifelong Learning, we should model Neural Network processes off of cognitive processes.

Specifically, this research introduces a Neural Network architecture for Lifelong Learning based on our brain's Three-Stage Memory Model comprising Short-Term Memory, Working Memory, and Long-Term Memory (Cowan, 2008). The accuracy of this grouping based on time periods is around 60%. The main innovation of the research is the Working Memory-based learning component.

This research attempts to achieve Lifelong Learning in a Deep-Convolutional Neural Network on the image classification task in the CIFAR dataset (Krizhevsky, 2009). CIFAR is first segmented based on its images' features into groups each representing a different "time period" to easily simulate learning over time.

The primary cognitive processes modeled computationally in the design are the generalized learning of the Cerebral Cortex (LT mem.), intricate learning of the Prefrontal Cortex (Working Mem.), the organizing nature of the Hippocampus (ST mem.), associative memory structure, and memory consolidation (D'Ardenne et al., 2012; Girardeau & Zugaro, 2011; Hartley et al., 2007; Purves, Augustine, Fitzpatrick, et al., 2001; Suzuki, 2008).

Methodology

The design consists of the Hippocampus (ST mem.), Cerebral Cortex (LT mem.), and Prefrontal Cortex (working mem.).

Hippocampus

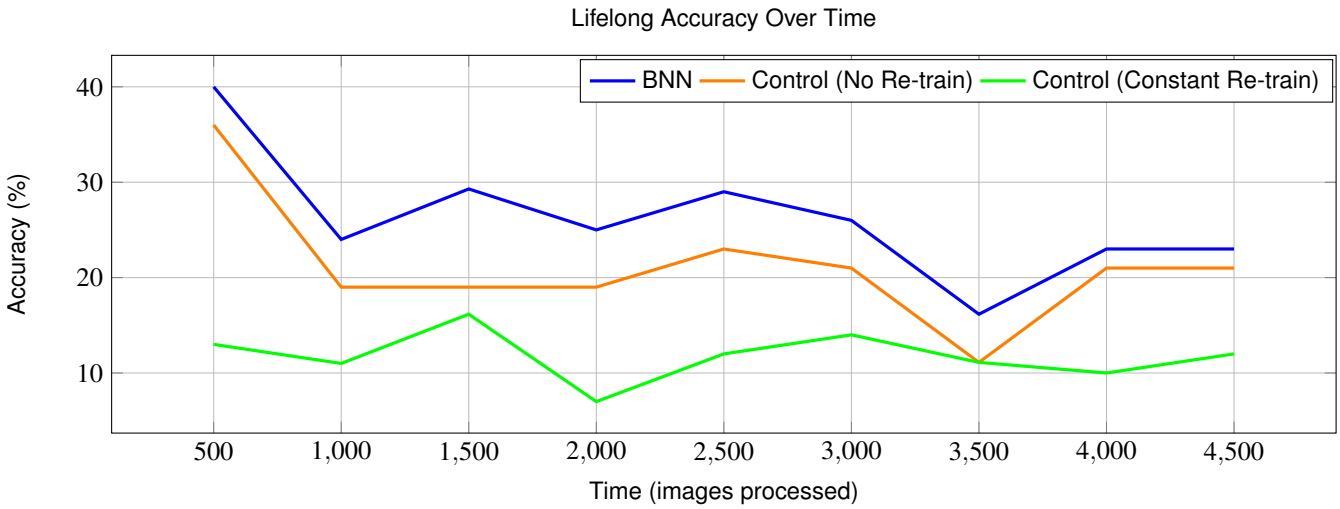
The Hippocampus receives the input data and directs it to and between the learning models. During learning, it sends the input data to the Prefrontal Cortex. During inference, the input data is directed into both models, and their outputs are averaged to get the final output. The Prefrontal Cortex gets 3x weighting when averaging–an empirically derived constant.

Cerebral Cortex

The Cerebral Cortex is one Deep-Convolutional Neural Network that preliminarily learns **general** patterns and an overall understanding of the input dataset, akin to human LT memory.

Prefrontal Cortex

The Prefrontal Cortex (see Figure 1) is the main innovation of this research, which enables working memory-akin understanding of the dataset. Using a Modular Neural Network design that groups the input dataset into groups of similar data (schemas) based on its extracted features–mimicking the associative memory structure–and trains mini Cerebral Cortices on each group, the Prefrontal Cortex continuously learns the **intricacies** of the dataset, akin to working memory.



Graph 1: Lifelong accuracy over time, demonstrating improved lifelong learning performance by this research's design.

When learning information, the input data's features are compared to all of the schemas to determine the most similar group of images, and that schema's mini Neural Network is trained on the input data. When making inferences, the input data is passed into all of the schemas but the prediction vector from the closest schema to the input data is given higher weight when combining the outputs of each schema. If a schema is repeatedly selected as the closest schema during inference, its data is common enough to be general understanding, and the Cerebral Cortex would therefore be fine-tuned on the data of the schema, (**Memory Consolidation**).

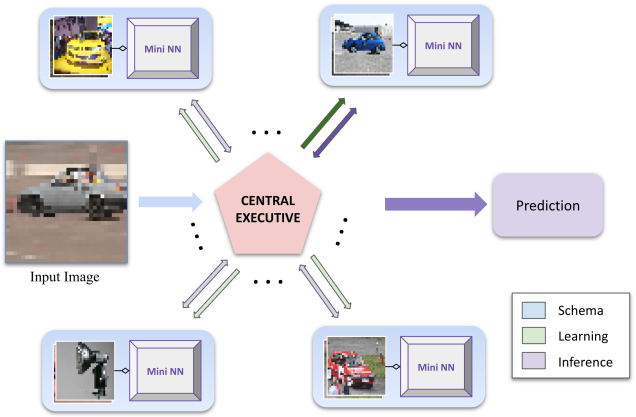


Figure 1: Prefrontal Cortex Design

Results

We measure Lifelong Accuracy over time. Since the dataset is grouped based on image features, when a model passes through the groups sequentially, we achieve the effect of time passing. Lifelong accuracy at a specific time is the model's accuracy on the current and previous groups. The Three-stage Memory Model-based Neural Network constantly learned the new groups using the learning methods explained in the Methodology section and was compared to a model that learned the new groups through normal finetuning–representing today's AI–and a model that did not relearn.

The results (see graph 1) show that over time, the Three-Stage Memory-based Neural Network, or the Biological Neural Network (BNN), retained higher accuracy than the other two control models, demonstrating its Lifelong Learning prowess compared to today's AI. Other tests on each individual learning system revealed that while the Prefrontal Cortex had lower individual accuracy than the Cerebral Cortex, an average 47.32% of the samples that Prefrontal Cortex is accurate on, the Cerebral Cortex isn't, demonstrating that when combined the Prefrontal Cortex provides new insight to the Cerebral Cortex. Experiments showed that the combining of the outputs from both models ensured that the new insight was incorporated. However, a primary limitation is that the design led to predictions taking up to 1.98 seconds, compared to 0.07 seconds at best for the control models.

Conclusion

This research developed the first Three-stage Memory Model-based Neural Network, implemented the first Working memory-mimicking learning system, and demonstrated the favorability of the Three-stage Memory Model to achieve other cognitive functions in AI. The future works of this research are to experiment on New Classes Lifelong Learning (Gido M. van de Ven & Tolias, 2022), improve the design's efficiency, and scale the integration of the Three-stage Memory Model to further achieve Lifelong Learning. Finally, this research demonstrated a pathway for researchers to achieve Lifelong Learning on larger scales and eventually achieve AGI.

Acknowledgments

I would like to thank my advisor, Ms. Emily Haven, for her support and advisory mentorship for this research.

References

Chen, Z., & Liu, B. (2018). Continual learning and catastrophic forgetting. *Lifelong Machine Learning*. Retrieved from https://www.cs.uic.edu/~liub/lifelong-learning/continual-learning.pdf doi: 10.1007/978-3-031-01581-6

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Prog Brain Res*, *20*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC2657600/ doi: 10.1016/S0079-6123(07)00020-9

D'Ardenne, K., Eshel, N., Luka, J., Lenartowicz, A., Nystrom, L. E., & Cohen, J. D. (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences*, *109*(49). Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1116727109 doi: 10.1073/pnas.1116727109

Gido M. van de Ven, T. T., & Tolias, A. S. (2022). Three types of incremental learning. *Nature Machine Intelligence*, *4*. Retrieved from https://www.nature.com/articles/s42256-022-00568-3 doi: 10.1038/s42256-022-00568-3

Girardeau, G., & Zugaro, M. (2011). Hippocampal ripples and memory consolidation. *Current Opinion in Neurobiology*, *21*(3). Retrieved from https://www.sciencedirect.com/science/article/pii/S0959438811000316 (Behavioural and cognitive neuroscience) doi: https://doi.org/10.1016/j.conb.2011.02.005

Hartley, T., Bird, C. M., Chan, D., Cipolotti, L., Husain, M., Vargha-Khadem, F., & Burgess, N. (2007). The hippocampus is required for short-term topographical memory in humans. *Hippocampus*, *17*(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677717/ doi: 10.1002/hipo.20240

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13). Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1611835114 doi: 10.1073/pnas.1611835114

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. *UToronto*. Retrieved from https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

Larkin, Z. (2024). *General AI vs Narrow AI.* Retrieved from https://levity.ai/blog/general-ai-vs-narrow-ai ([Accessed 29-03-2025])

Perrett, A., Furber, S. B., & Rhodes, O. (2022). Error driven synapse augmented neurogenesis. *Frontiers in Artificial Intelligence*, *5*. Retrieved from https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.949707 doi: 10.3389/frai.2022.949707

Purves, D., Augustine, G. J., Fitzpatrick, D., et al. (2001). The long-term storage of information. In D. Purves, G. J. Augustine, D. Fitzpatrick, et al. (Eds.), *Neuroscience* (2nd ed.). Sunderland, MA: Sinauer Associates. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK10901/ (Available from: NCBI Bookshelf)

Suzuki, W. A. (2008). Associative learning signals in the brain. *Progress in Brain Research*, *169*. Retrieved from https://pubmed.ncbi.nlm.nih.gov/18394483/ doi: 10.1016/S0079-6123(07)00019-2

van de Ven G.M. Siegelmann H.T. & Tolias A.S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, *11*. Retrieved from https://www.nature.com/articles/s41467-020-17866-2 doi: 10.1038/s41467-020-17866-2