Advancing Neuro-Inspired Lifelong Learning for Edge with Co-Design

Nicholas Soures, Vedant Karia, Dhireesha Kudithipudi

Neuromorphic AI Lab
University of Texas at San Antonio
dk@utsa.edu

Abstract

Lifelong learning, which refers to an agent's ability to continuously learn and enhance its performance over its lifespan, is a significant challenge in artificial intelligence (AI), that biological systems tackle efficiently. This challenge is further exacerbated when AI is deployed in untethered environments with strict energy and latency constraints. We take inspiration from neural plasticity and investigate how to leverage and build energy-efficient lifelong learning machines. Specifically, we study how a combination of neural plasticity mechanisms, namely neuromodulation, synaptic consolidation, and metaplasticity, enhance the continual learning capabilities of AI models. We further co-design architectures that leverage compute-in-memory topologies and sparse spike-based communication with quantization for the edge. Aspects of this co-design can be transferred to federated lifelong learning scenarios.

Keywords: ML: Lifelong and Continual Learning, ML: Distributed Machine Learning & Federated Learning

Introduction

In recent years, there have been remarkable breakthroughs in the field of artificial intelligence (AI). However, the emergence of new generation of applications such as self-driving vehicles, wearable devices, etc. will require new forms of AI capable of learning continuously throughout their lifetime. AI machines will need to acquire new skills without compromising old ones, adapt to changes, and apply previously learned knowledge to new tasks while conserving limited resources. Moreover, the workload profile for lifelong learning has different characteristics on the edge, such as processing data at variable frequencies, operating under strict memory and compute constraints, and optimizing for energy-accuracy trade-offs in real-time (Kudithipudi et al. 2023). To enable such learning, we draw inspiration from the neural plasticity mechanisms and propose hardware-software co-design approaches that are amenable to the edge. The plasticity mechanisms regulate memory and learning based on the local context and internal state of the system. This ability plays a key role in adapting to novelty and inducing dynamic behavior in the network.

In terms of the software model, we designed a task-agnostic lifelong learning algorithm with local compositional plasticity mechanisms that are inherently energy efficient (Soures et al. July 2021). The model uses mechanisms such as i) metaplasticity (plasticity of plasticity) to protect previous knowledge encoded in important synapses; ii) synaptic consolidation, a form of long-term plasticity to capture knowledge over multiple time scales; and iii) neuromodulation, to improve the distribution of information throughout the network and facilitate the exploitation of information overlap between tasks. These mechanisms can be integrated with the base SNN model in a modular manner without depending on the type of network architecture (, *e.g.* dense, convolutional, recurrent) or the learning rule. The combination of these mechanisms outperformed state-of-the-art in a variety of continual learning scenarios with streaming data, where model sees samples only once and is unaware of task switching during training or inference.

We designed the first online continual learning accelerator with multiple co-design strategies: minimizing memory read and write operations, co-locating compute and memory, and model-aware reconfigurability (Karia et al. 2022). Our approach relies on sparse spike-based communication, transmitting only spike indices , reducing memory access and memory size by $\sim 2\times$. An initial study of the efficiency of the proposed mechanisms , unlike some prior approaches, they do not grow over time in memory capacity or compute complexity.

References

Karia, V.; Zohora, F. T.; Soures, N.; and Kudithipudi, D. 2022. Scolar: A spiking digital accelerator with dual fixed point for continual learning. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1372–1376. IEEE.

Kudithipudi, D.; Daram, A.; Zyarah, A. M.; Zohora, F. T.; Aimone, J. B.; Yanguas-Gil, A.; Soures, N.; Neftci, E.; Mattina, M.; Lomonaco, V.; et al. 2023. Design principles for lifelong learning AI accelerators. *Nature Electronics*, 1–16.

Soures, N.; Helfer, P.; Daram, A.; Pandit, T.; and Kudithipudi, D. July 2021. TACOS: Task Agnostic Continual Learning in Spiking Neural Networks. In *Theory and Foundation of Continual Learning Workshop at ICML'2021*.