Adaptive Intelligence: Enabling Continual Learning on the Edge

Marco P. E. Apolinario

July 16, 2026

Abstract

Deep neural networks excel at static task execution, but they suffer from catastrophic forgetting when learning new tasks sequentially, and they demand significant memory to store activation maps during on-device backpropagation. This paper describes research on adaptive intelligence algorithms that allow models to learn new tasks sequentially while preserving prior knowledge under strict resource constraints. We introduce CODE-CL, a conceptor-based gradient projection method that mitigates catastrophic forgetting while facilitating forward knowledge transfer, and LANCE, which uses low-rank activation compression via one-shot higher-order singular value decomposition to drastically reduce memory overhead. Together, these frameworks pave a clear path towards energy-efficient, lifelong learning on the edge.

1 Overview

Deep neural networks today excel at learning complex representations from massive datasets — yet they remain inherently **static**. Once trained, they struggle to incorporate new knowledge without overwriting the old, a phenomenon known as *catastrophic forgetting*. Moreover, their reliance on global gradient backpropagation requires storing large activation maps and synchronizing updates across all layers — a process that is **energy-intensive** and **ill-suited for on-device intelligence**.

In contrast, natural intelligence learns **continuously** and **adaptively**. The brain integrates new experiences without erasing the past, leveraging distributed representations and sparse updates to maintain a stable sense of the world.

A central question motivates my research:

> *Can we engineer learning systems that continuously evolve and improve on-device, without relying on cloud retraining or external memory?*

This question guides my work on **adaptive intelligence** — algorithms that allow deep networks to learn new tasks sequentially while preserving prior knowledge, all under the strict memory and energy constraints of embedded devices.

The first step toward this goal was to overcome **catastrophic forgetting** — the tendency of deep networks to lose prior knowledge when learning new tasks. In **CODE-CL (Conceptor-Based Gradient Projection for Deep Continual Learning)** [1], we introduced a biologically inspired framework where each task's learned knowledge is encoded as a *conceptor matrix* — a compact representation of the activation subspace spanned by that task.

When learning a new task, CODE-CL projects incoming gradients onto the *pseudo-orthogonal* complement of previously learned subspaces, preventing destructive interference while still allowing information flow along shared directions. This mechanism effectively balances **stability and plasticity**: it preserves important gradient directions from past tasks while enabling **forward knowledge transfer** for new ones. Through extensive experiments on benchmark datasets such as Split CIFAR-100, Split MiniImageNet, and 5-Datasets, CODE-CL demonstrated higher accuracy and reduced forgetting compared to state-of-the-art continual learning methods — all with minimal memory overhead and high computational efficiency.

Yet, while CODE-CL addressed *how to remember*, it also revealed the challenge of *where to learn efficiently*. Training deep networks, even for continual adaptation, still demands large activation storage for backpropagation — a severe limitation for edge devices with only a few hundred kilobytes of on-chip memory.

To address this, we developed **LANCE (Low-Rank Activation Compression for Efficient On-Device Continual Learning)** [2], a framework that rethinks backpropagation itself. LANCE performs a **one-shot higher-order singular value decomposition (HOSVD)** to identify a reusable low-rank subspace for each layer's activations. Instead of recomputing decompositions every iteration, activations are projected into this fixed subspace throughout training — drastically reducing both memory and compute cost. This design reduces activation storage by up to **250**$\times$ and computational energy by **1.5**$\times$, while maintaining accuracy within **2%** of full backpropagation.

Crucially, these fixed low-rank subspaces naturally extend to continual learning. Each new task is allocated to orthogonal components of the subspace, allowing models to **acquire new knowledge without overwriting the old**, and doing so entirely on-device. LANCE thus unifies compression and continual adaptation into a single mechanism, enabling edge devices to fine-tune and learn continually without relying on cloud resources or replay buffers.

Together, **CODE-CL** and **LANCE** define a roadmap toward **adaptive edge intelligence** — systems that learn efficiently, preserve knowledge over time, and adapt to changing environments. By coupling subspace-based memory preservation with low-rank activation compression, this research bridges algorithmic design and hardware co-optimization, paving the way for energy-efficient, lifelong learning in next-generation embedded and neuromorphic AI systems.

References

[1] M. P. E. Apolinario, S. Choudhary, and K. Roy, "CODE-CL: Conceptor-Based Gradient Projection for Deep Continual Learning," in *International Conference on Computer Vision (ICCV)*, 2025.

[2] M. P. E. Apolinario and K. Roy, "LANCE: Low Rank Activation Compression for Efficient On-Device Continual Learning," in *European Conference on Computer Vision (ECCV)*, 2026.