Combining Minimal Recurrent Gating with Learnable Delays for Memory-Efficient Sequence Modeling

T. Torchet*, C. Metzner*, K. C. Raghunathan, J. Weber, S. Billaudelle, L. Kriener, M. Payvand

Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland

*These authors contributed equally,

Multi-timescale sequence modeling relies on capturing both local fast dynamics and global slow context; yet, maintaining these capabilities under the strict memory constraints common to edge devices remains an open challenge. Current State-of-the-Art models with constant memory footprints trade off long-range selectivity and high-precision modeling of fast dynamics.

To overcome this trade-off within a fixed memory budget, we introduce **mGRADE** (**m**inimally **G**ated **R**ecurrent **A**rchitecture with **D**elay **E**mbedding), a compact hybrid-memory sequence model combining a parallelizable gated recurrent component with learnable delays parameterized as a causal convolution with learnable spacings (Fig. 1A). Its design is inspired by biological neurons, which gate membrane dynamics through input-dependent conductances and integrate signals over short time frames via dendritic delays [1].

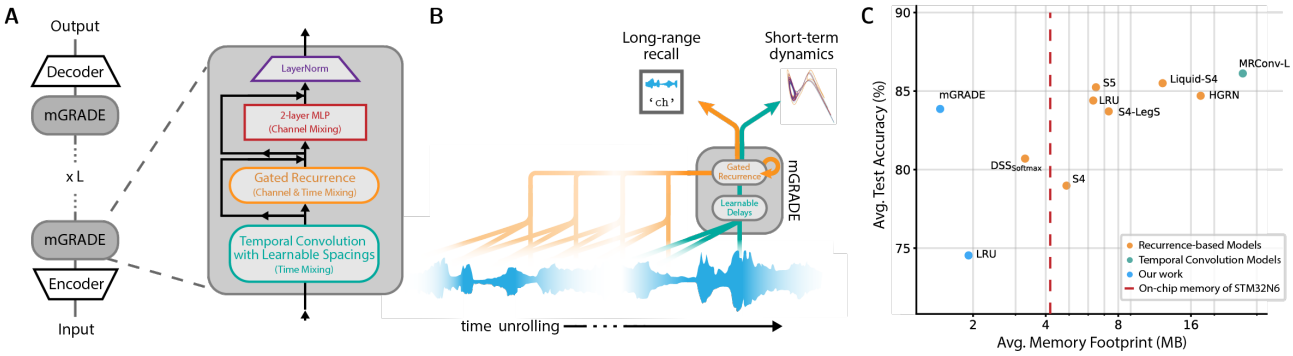

Figure 1: **A)** Network architecture. **B)** mGRADE handles multi-timescale information in input sequences (blue waveform) by using the gated recurrence to selectively store long-term dependencies (orange arrows), enabling long-range recall, while the convolution with learnable spacings models short-term dynamics (green arrows). **C)** mGRADE's average test accuracy across Long-Range Arena is competitive while maintaining a memory footprint up to 8× smaller than State-of-the-Art model and fitting within the on-chip memory of typical edge AI systems such as the STM32N6.

We show theoretically how the learnable delays are equivalent to a formal delay embedding, enabling parameter-efficient reconstruction of partially-observed fast dynamics. The gated recurrent component complements this by capturing long-range context with minimal memory overhead, selectively compressing long sequence histories into a fixed-size hidden state (Fig. 1B).

On the challenging Long-Range Arena benchmark [2] and 35-way Google Speech Commands raw audio classification task [3], mGRADE reduces the memory footprint by up to a factor of 8 compared to other State-of-the-Art models, while maintaining competitive performance (Fig. 1C). Across both benchmarks, mGRADE is the only model that simultaneously achieves high performance while fitting the memory budget of a typical edge AI platform.

[1] P. Poirazi & A. Papoutsi, Nat Rev Neurosci, 21, 303–321, 2020.

[2] Y. Tay et al., *Long Range Arena: A Benchmark for Efficient Transformers*, ICLR, 2021.

[3] P. Warden, *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*, arXiv, 2018.