

Improving the Performance and Learning Stability of Parallelizable RNNs for Ultra-Low Power Applications

J. Brandoit · A. Fyon · D. Ernst · G. Drion

Department of Electrical Engineering and Computer Science, University of Liège, Belgium

ICML 2026
SPOTLIGHT



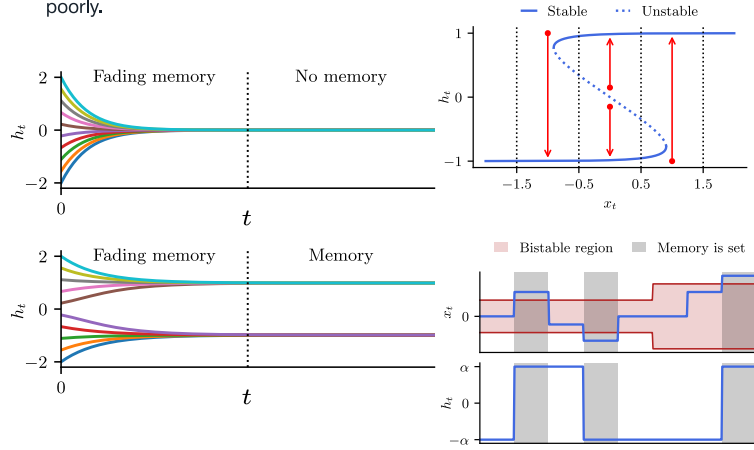
Parallelizable RNNs are ideal for edge AI, but they all suffer from **fading memory**.
Persistent memory exists, but bistable cells were **too unstable to train**.

The problem

Fading and persistent memory.

Parallelizable RNNs (SSMs, linear gated RNNs) train via associative scans. Contractive dynamics force eigenvalues inside the unit circle. The state decays. This is **fading memory**.

Persistent memory is needed for long-term dependencies. The BMRU uses a hysteresis band to hold quantized states indefinitely, this is **bistability**; but it trains poorly.



Bistable Memory Recurrent Unit (BMRU)

$$\begin{aligned}\hat{h}_t &= W_x x_t + b_x & \beta_t &= |W_\beta x_t + b_\beta| \\ z_t &= H(|\hat{h}_t| - \beta_t) \\ h_t &= z_t \odot S(\hat{h}_t) \odot \alpha + (1 - z_t) \odot h_{t-1}\end{aligned}$$

In update mode ($z_t = 1$): $\partial h_t / \partial h_{t-1} = 0$.
The gradient is blocked.

The fix

From bistable MRUs to cumulative MRUs.

• WE RESTORE THE GRADIENT, AND UNLOCK MULTISTABILITY •

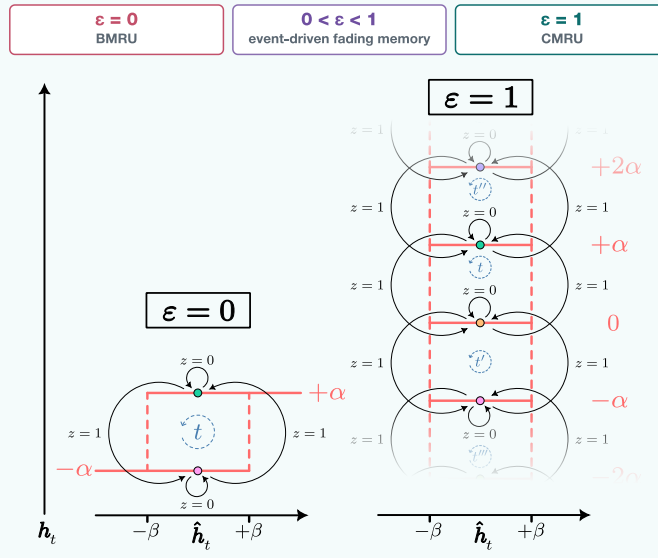
$$h_t = z_t \odot [S(\hat{h}_t) \odot \alpha + \varepsilon h_{t-1}] + (1 - z_t) \odot h_{t-1}$$

The term εh_{t-1} is a skip-connection through time.

$$\left. \frac{\partial h_t}{\partial h_{t-1}} \right|_{z_t=0} = 1, \quad \left. \frac{\partial h_t}{\partial h_{t-1}} \right|_{z_t=1} = \varepsilon.$$

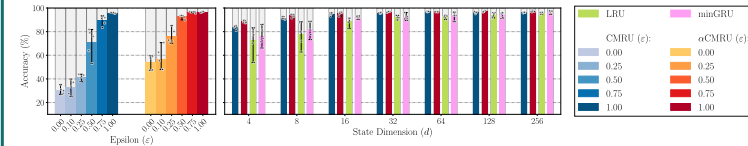
From bistability to multistability.

The hysteresis band still governs gating, but now accumulates rather than resets.

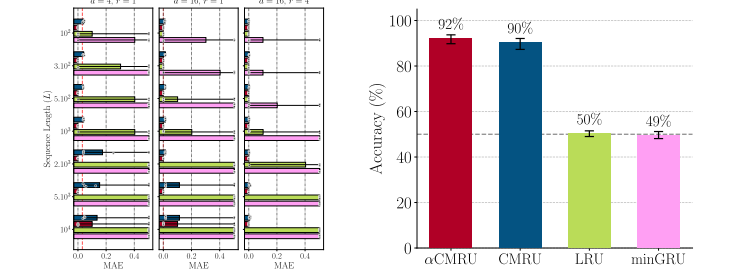


What we get

Low-capacity scaling and perspectives.

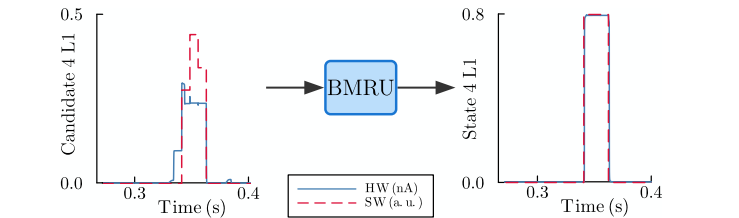


Sequential MNIST (1 layer, d=32). $\varepsilon: 0 \rightarrow 1$ **lifts accuracy and shrinks variance**. CMRU scales better at small d.



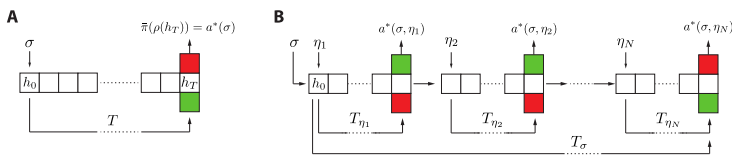
Copy first input, continuous noisy input. CMRU stays near the quantization limit. **alphaCMRU is almost perfect**. Fading RNNs cannot generalize to longer sequences.

Pathfinder (3 layers, d=32). CMRU variants **solve it**. LRU and minGRU stay at chance level.



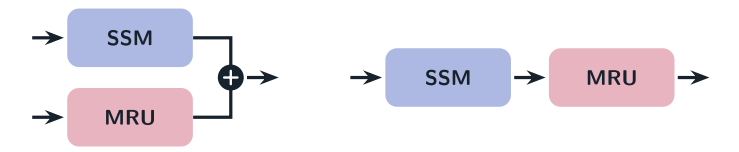
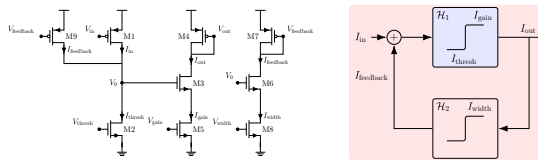
Quantized states map directly to analog hardware. Hardware (HW) and software (SW) outputs **agree closely**: quantization **blocks noise** from accumulating across steps.

Beyond this paper



Multistability for RL. Recurrence-based agents trained on short horizons stay optimal at far longer ones. Multistability is the **necessary property**.

Analog implementation. Epsilon-annealing allows reliable training of the equivalent CMOS circuits.



Hybrid with SSMs. Combining SSMs and CMRU brings both fading and persistent memory into a single architecture. *An open research direction.*