Legendre Memory Units (LMUs)

Continuous-Time Representation in Recurrent Neural Networks

Aaron R. Voelker, Ivana Kajić, Chris Eliasmith {arvoelke, i2kajic, celiasmith}@uwaterloo.ca

Centre for Theoretical Neuroscience, Applied Brain Research, University of Waterloo <**https://github.com/abr/neurips2019**>

Introduction

- We introduce a new RNN, the LMU, that outperforms LSTMs by $10^6\times$ on a $10^3\times$ more difficult memory task.
- The LMU sets a **new state-of-the-art result** on **psMNIST** (97.15%) – a standard RNN benchmark.
- The LMU uses 38% fewer parameters and trains 10x faster than competitors.

Methods

LMUs provide the optimal solution for representing a sliding window of θ seconds using d variables [1, 2].

It does so by implementing the dynamical system:

$$\theta\dot{\mathbf{m}}(t) = \mathbf{A}\mathbf{m}(t) + \mathbf{B}u(t)$$

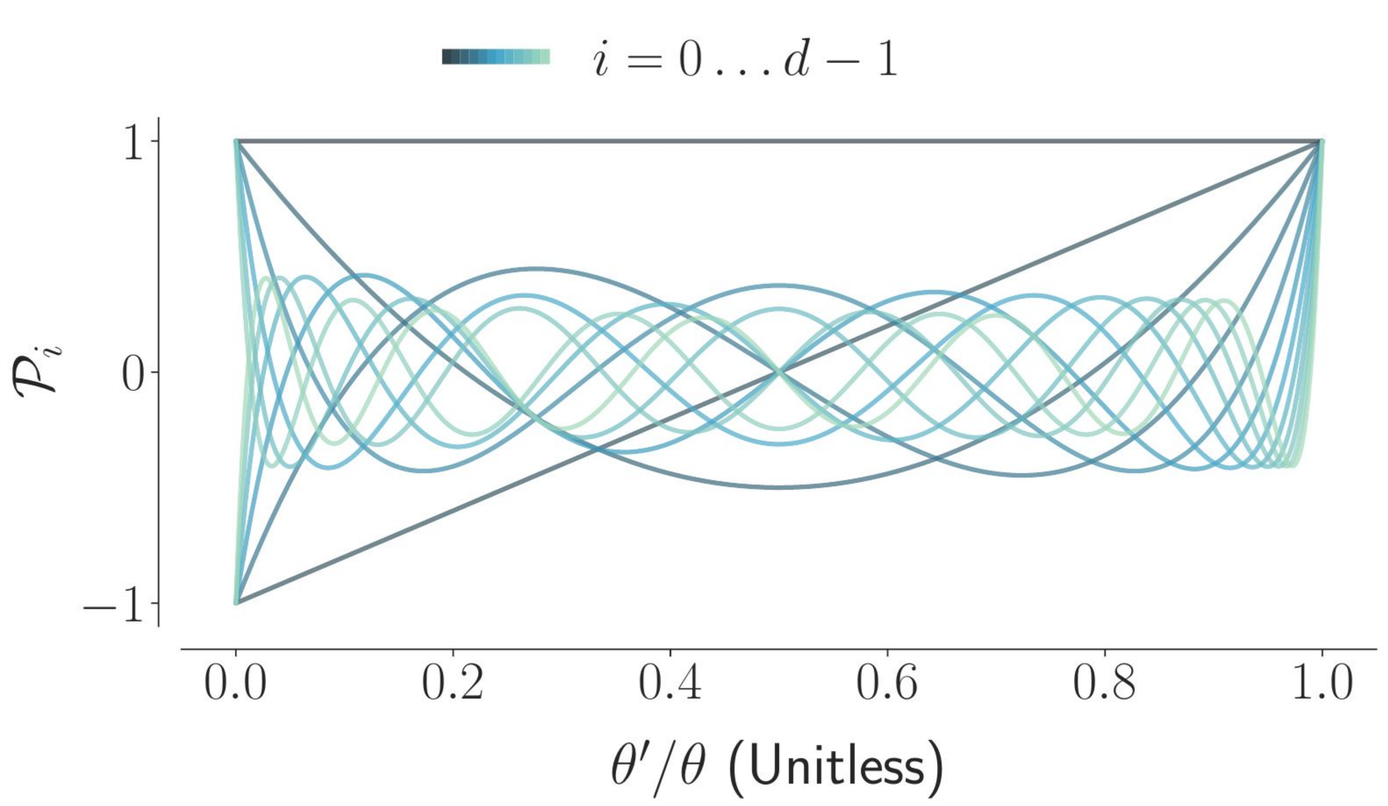
$$\mathbf{A} = [a]_{ij} \in \mathbb{R}^{d\times d}, \quad a_{ij} = (2i+1)\begin{cases} -1 & i < j \\ (-1)^{i-j+1} & i \geq j \end{cases}$$

$$\mathbf{B} = [b]_i \in \mathbb{R}^{d\times 1}, \quad b_i = (2i+1)(-1)^i, \quad i,j \in [0, d-1]$$

The memory $\mathbf{m}(t) \in \mathbb{R}^d$ **orthogonalizes** the previous θ seconds of history, as in:

$$u(t-\theta') \approx \sum_{i=0}^{d-1} \mathcal{P}_i\left(\frac{\theta'}{\theta}\right) m_i(t)$$

where $\mathcal{P}_i$ are the shifted **Legendre polynomials**.

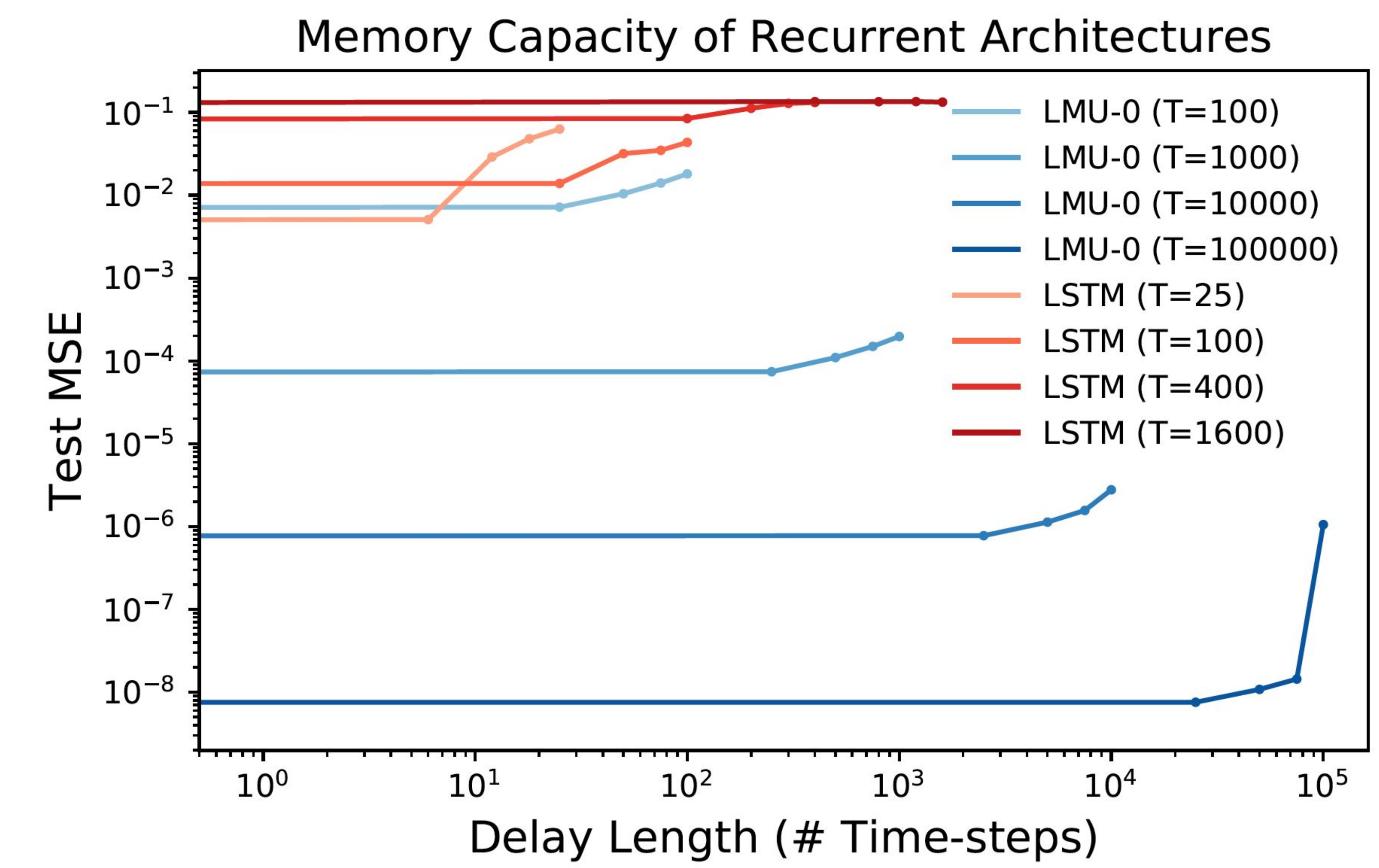


Main Results

Model	Validation	Test
RNN-orth	88.70	89.26
RNN-id	85.98	86.13
LSTM	90.01	89.86
LSTM-chrono	88.10	88.43
GRU	92.16	92.39
JANET	92.50	91.94
SRU	92.79	92.49
GORU	86.90	87.00
NRU	95.46	95.38
Phased LSTM	88.76	89.61
LMU	**96.97**	**97.15**
FF-baseline	92.37	92.65

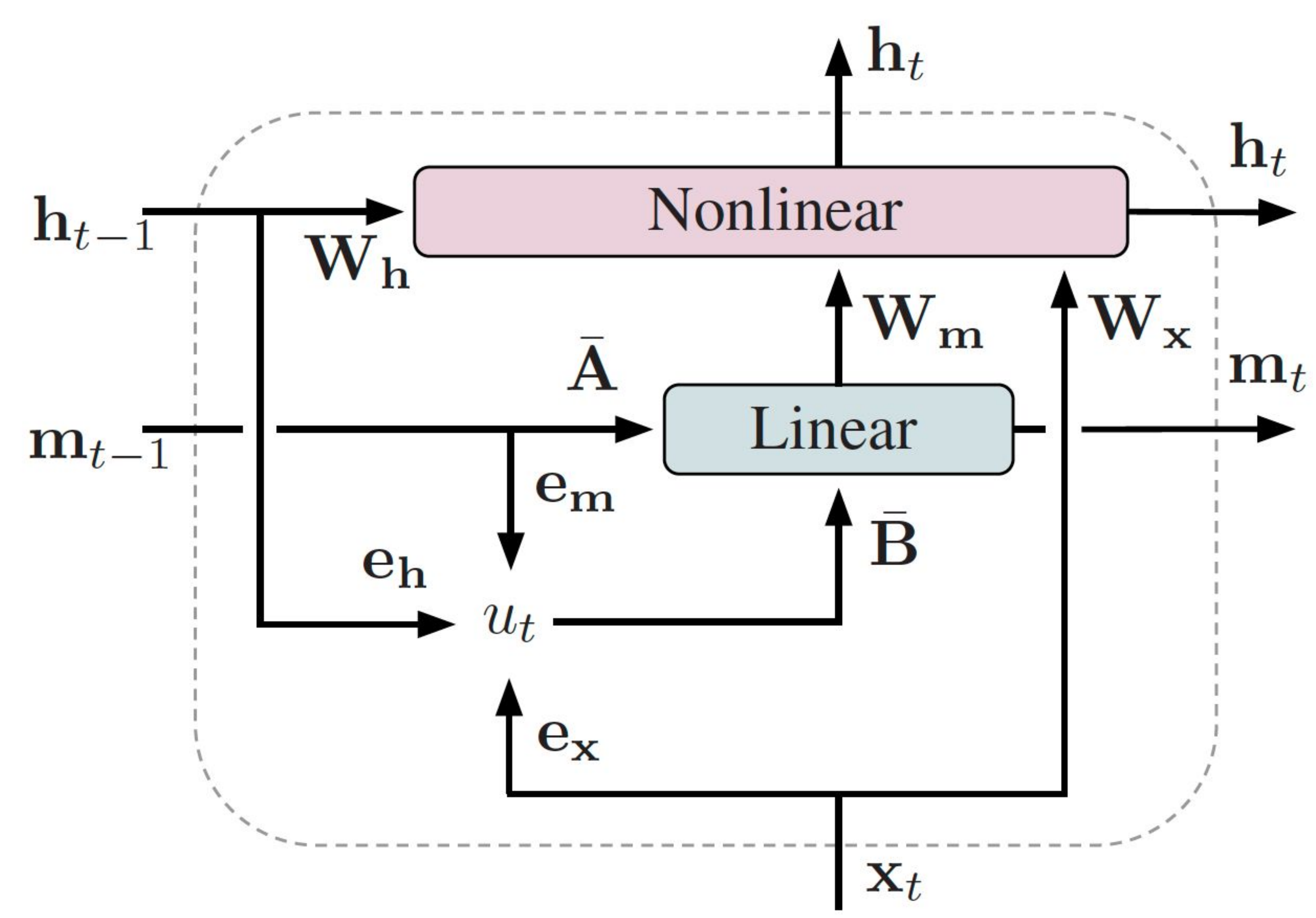
Left: SotA performance of RNNs on the permuted sequential MNIST benchmark. 102K vs 165K parameters. LMU uses ***d* = 256** dimensions.

Right: LMU vs LSTM memory capacity for different delay lengths given a 10Hz white noise input. 500 vs 41,000 parameters. 105 vs 200 state variables.



Architecture

- Consists of an optimal linear memory coupled with nonlinear units.
- Stackable and trainable via backpropagation through time.
- **A** and **B** are discretized by an ODE solver and can be trained together with θ – although this is typically unnecessary.



$$u_t = \mathbf{e_x}^\mathsf{T}\mathbf{x}_t + \mathbf{e_h}^\mathsf{T}\mathbf{h}_{t-1} + \mathbf{e_m}^\mathsf{T}\mathbf{m}_{t-1}$$

$$\mathbf{h}_t = f\left(\mathbf{W_x}\mathbf{x}_t + \mathbf{W_h}\mathbf{h}_{t-1} + \mathbf{W_m}\mathbf{m}_t\right)$$

Impact

- Many opportunities to replace LSTMs with LMUs.
- LMUs are derived from first principles, hence amenable to analysis (unlike most other RNNs).
- Deployed on low-power, spiking neuromorphic hardware for energy-efficient AI (see figure).

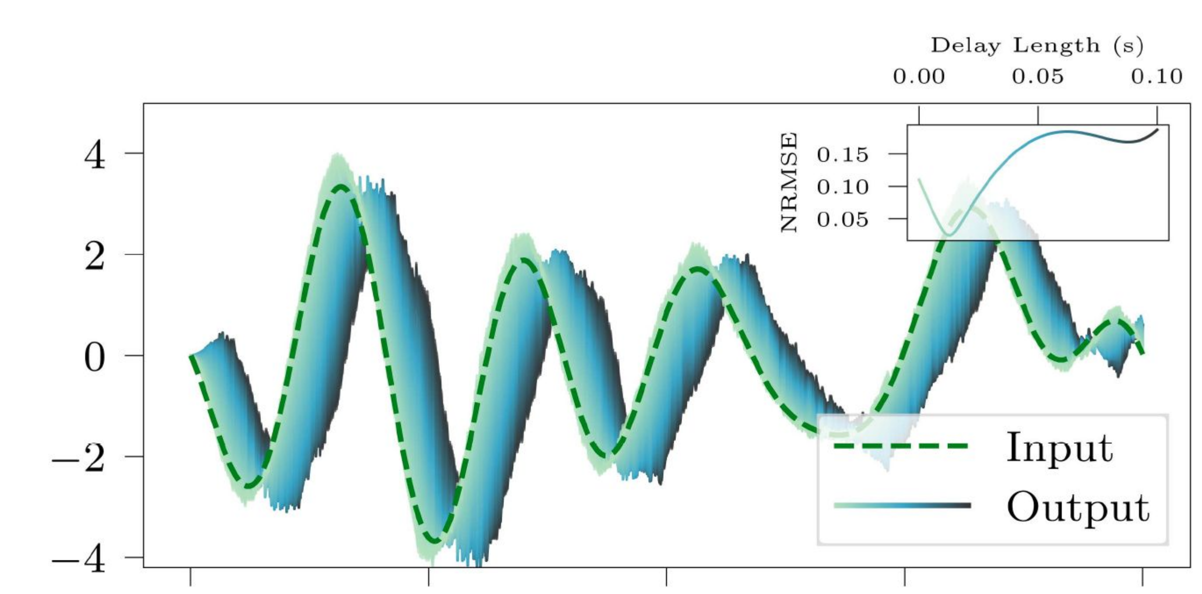


Figure: LMU running on Braindrop – mixed analog-digital spiking neuromorphic hardware [3].

References

[1] Voelker, A. R. and Eliasmith, C. (2018) Improving spiking dynamical networks: Accurate delays, higher-order synapses, and time cells. *Neural Computation*, 30(3):569-609, 03.

[2] Voelker, A. R. (2019) Dynamical Systems in Spiking Neuromorphic Hardware. *PhD thesis*, University of Waterloo. URL: http://hdl.handle.net/10012/14625.

[3] Neckar et al. (2019) Braindrop: a mixed-signal neuromorphic architecture with a dynamical systems-based programming model. *Proceedings of the IEEE*, 107:144–164.