ICLR

LMUFormer: Low Complexity Yet Powerful Spiking Model With Legendre Memory Units

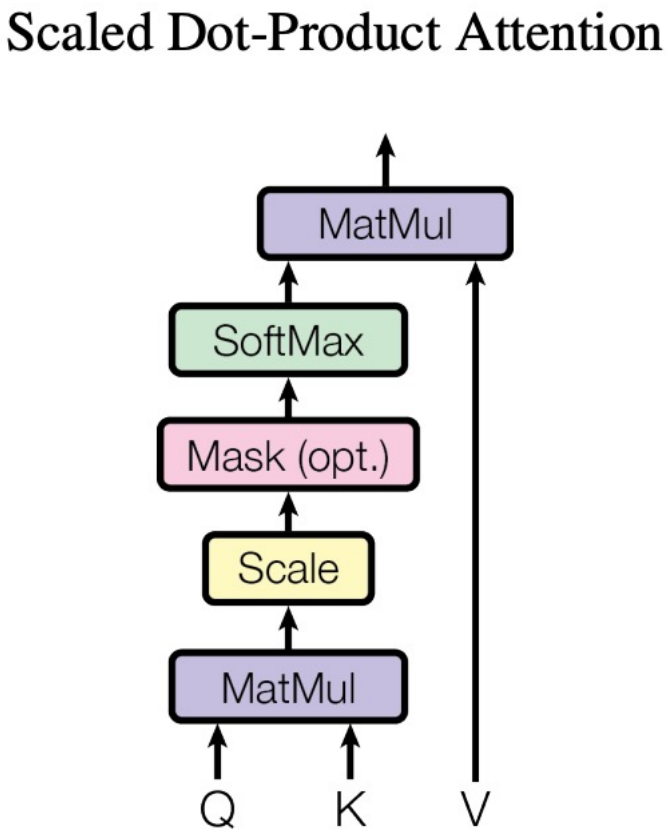
Zeyu Liu*, Gourav Datta*,^, Anni Li, Peter A. Beerel

University of Southern California

* Equally contributing authors

^ Currently employed at Amazon Inc.

Motivation

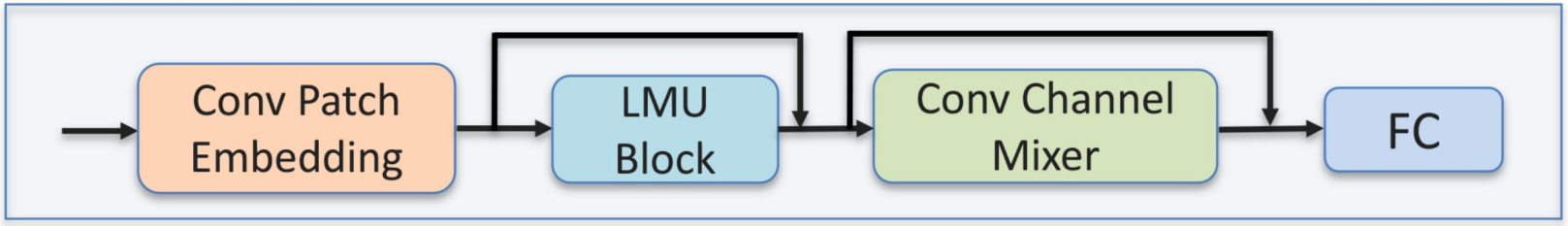


Transformers with self-attention mechanism have established themselves as leading deep learning models across a multitude of applications. But two main **disadvantages** remain:

1. **Quadratic** computational and memory complexities relative to the sequence length N;

2. Global self-attention mechanism poses challenges for **streaming** applications, which require the processing of data in **real-time**.

Overall Structure of LMUFormer



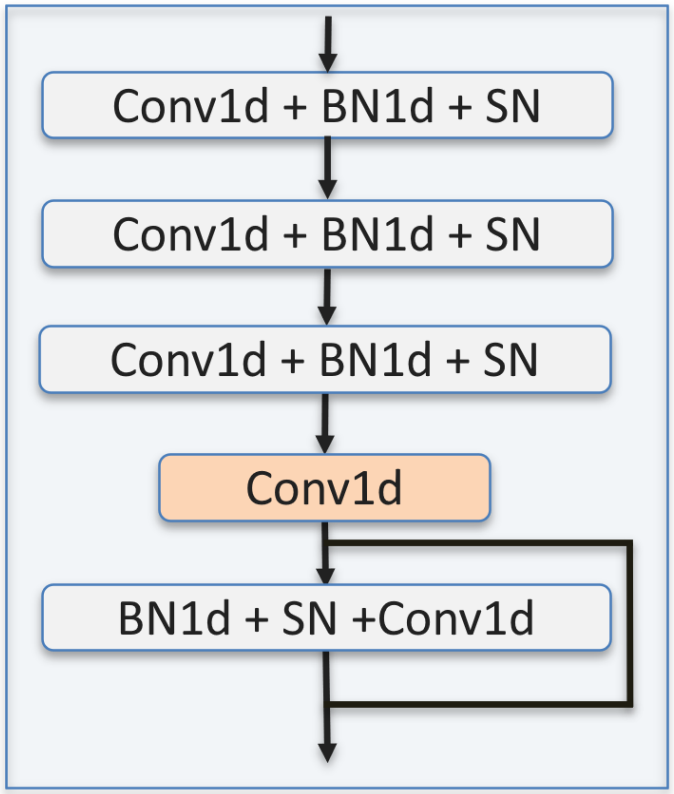
In comparison to SOTA transformer-based models:

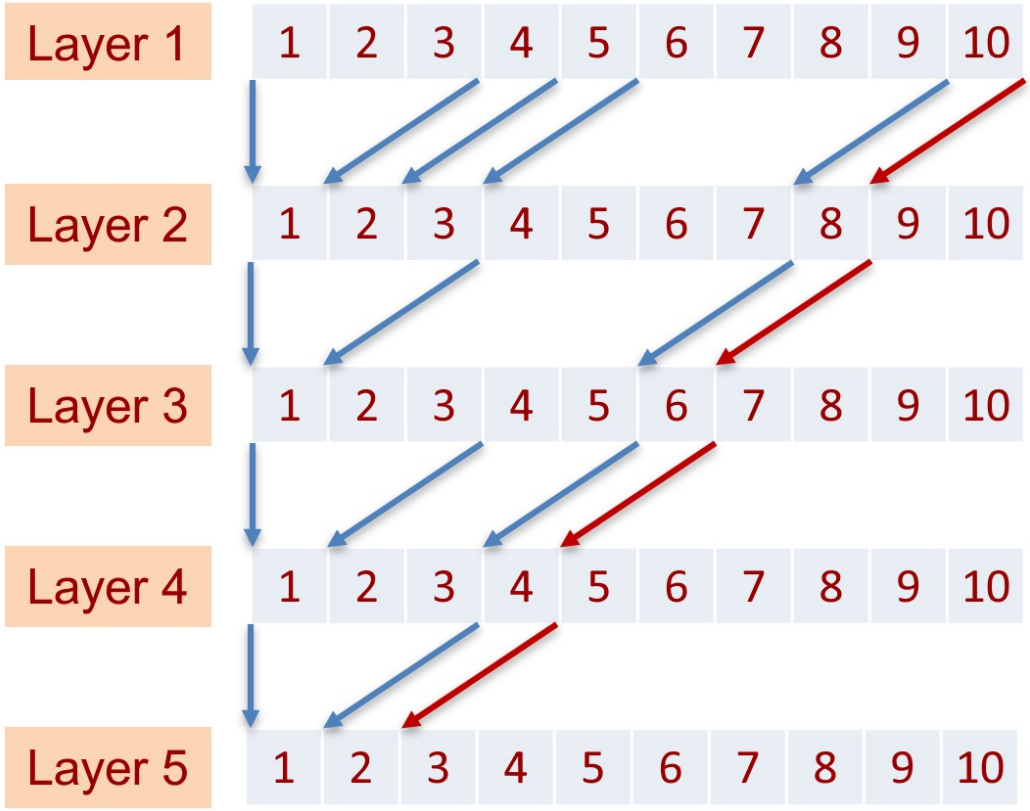
Powerful: on Speech Commands V2, we get comparable results within the ANN domain and SOTA results within SNN domain

Efficient: 53X reduction in parameters & 65X decrement in FLOPS

Streaming: Process data in real-time, maintain 99% accuracy while achieving 32.03% duction in the sequence length

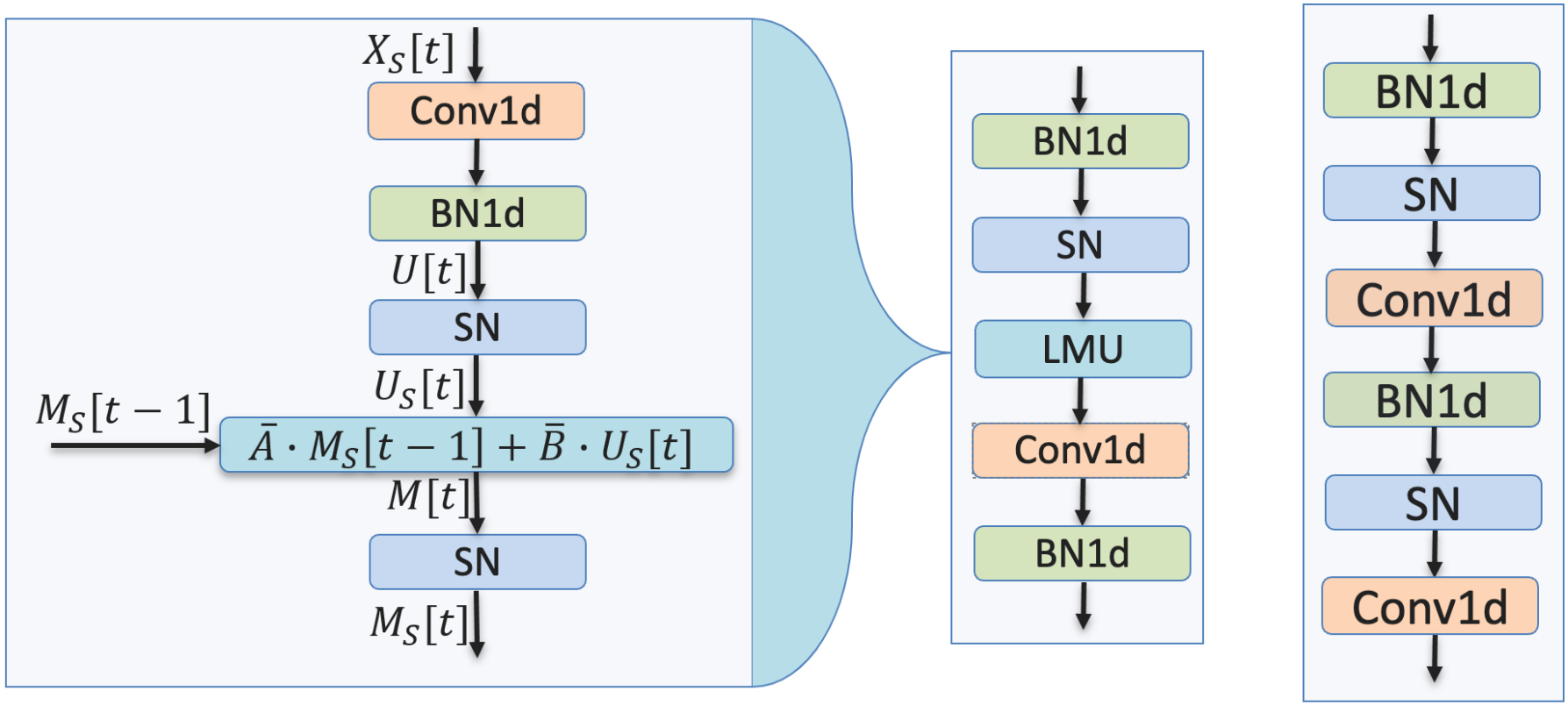
Convolutional Patch Embedding & delays





We use "***Conv in time***" for the patch splitting & embedding, which incurs 8 delays, but this is negligible since sequence lengths are usually in the hundreds or thousands.

Spiking LMU Block & Conv Channel Mixer



Considering the ***inherent temporal dynamics of both LMU and SNN***, we devised a **merged** process to optimize the overall operational efficiency of their integration.

Main results in Speech Commands V2

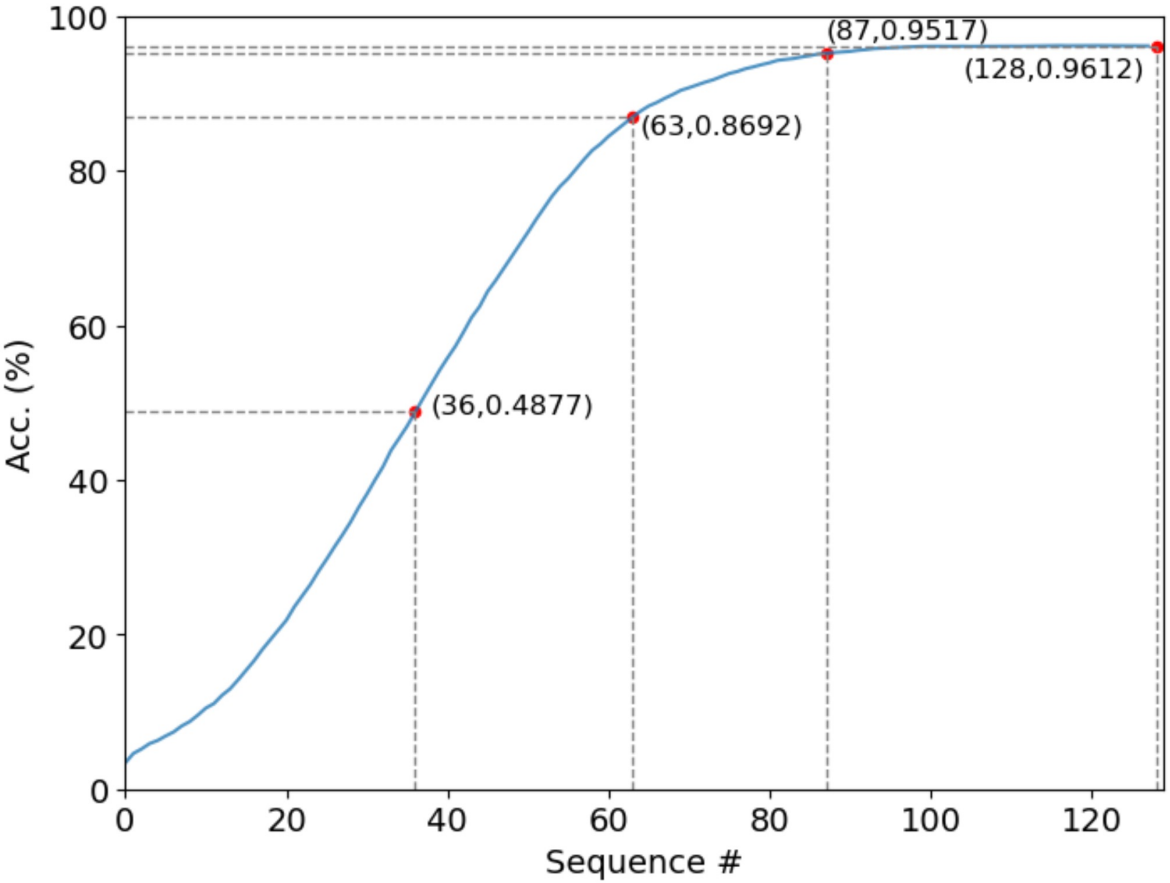
Model	Sequential Inference	Parallel Training	SNN	Accuracy (%)
RNN (Bittar & Garner, 2022)	Yes	No	No	92.09
Attention RNN (De Andrade et al., 2018)	No	No	No	93.9
liBRU (Bittar & Garner, 2022)	Yes	No	No	95.06
Res15 (Vygon & Mikhaylovskiy, 2021)	Yes	Yes	No	97.00
KWT2 (Berg et al., 2021)	No	Yes	No	97.74
AST (Gong et al., 2021)	No	Yes	No	98.11
LIF (Bittar & Garner, 2022)	Yes	Yes	Yes	83.03
SFA (Salaj et al., 2021)	Yes	No	Yes	91.21
Spikformer* (Zhou et al., 2022)	No	Yes	Yes	93.38
RadLIF (Bittar & Garner, 2022)	Yes	No	Yes	94.51
Spike-driven ViT* (Yao et al., 2023)	No	Yes	Yes	94.85
LMUFormer	Yes	Yes	No	**96.53**
LMUFormer (with states)	Yes	Yes	No	**96.92**
Spiking LMUFormer	Yes	Yes	Yes	**96.12**

SOTA

Model	Params. (M)	OPs (G)
AST (Gong et al., 2021)	86.93	12.4
LMUFormer	53.66× 1.62	65.61×0.189
Spiking LMUFormer	1.69	0.0309

Accuracy comparable to ANN transformer-based models with a significantly reduced parameters and lower FLOPS

Streaming Application



Spiking LMUFormer achieves ***99%*** ***(95.17% / 96.12%)*** of its original performance

while getting a ***32.03%*** ***(1 - 87/128)*** reduction in the sequence length

Thank you!