SpikingSSMs: Learning Long Sequences with Sparse and Parallel Spiking State Space Models

Shuaijie Shen*[,1,2], Chao Wang*[,1,2], Renzhuo Huang[1,2], Yan Zhong[2,3], Qinghai Guo[2], Zhichao Lu[4], Jianguo Zhang †[,1,5], Luziwei Leng†[,2]

[1] SUSTech | [2] Huawei | [3] Peking University | [4] CityU | [5] Pengcheng Lab

Code: https://github.com/shenshuaijie/SDN | **Paper:** https://arxiv.org/abs/2408.14909

[Key Contributions]

SpikingSSM Architecture

- Integrates spiking neurons (LIF) with state space models (SSMs) for efficient long-sequence learning.
- Mimics multi-timescale dendritic neuron dynamics for sparse, energy-efficient computation.

Surrogate Dynamic Network (SDN)

- Enables parallel training of SNNs by predicting spike trains and membrane potentials.
- Achieves **100×** speedup for sequences up to 8K steps.

Learnable Thresholds

- Optimizes spiking rate and performance via adaptive thresholds.
- Reduces quantization error and improves information transmission.

State-of-the-Art Results

- **90%** sparsity on LRA benchmark with competitive accuracy.
- **33.94** PPL on WikiText-103 (outperforms SpikeGPT with 1/3 model size).

[Method Overview]

SpikingSSM Block

- Combines SSM's parallel sequence modeling with LIF neurons' sparse computation.
- Forward Pass:

  SSM processes input → LIF neuron generates spikes → Sparse output propagates.

Surrogate Dynamic Network (SDN)

- Lightweight CNN trained to predict membrane potentials or spikes in parallel.
- Key Features:
  - Parallel training.
  - No additional parameters during inference.

Learnable Threshold

- Dynamically scales inputs to balance sparsity and accuracy.

[Result Highlights]

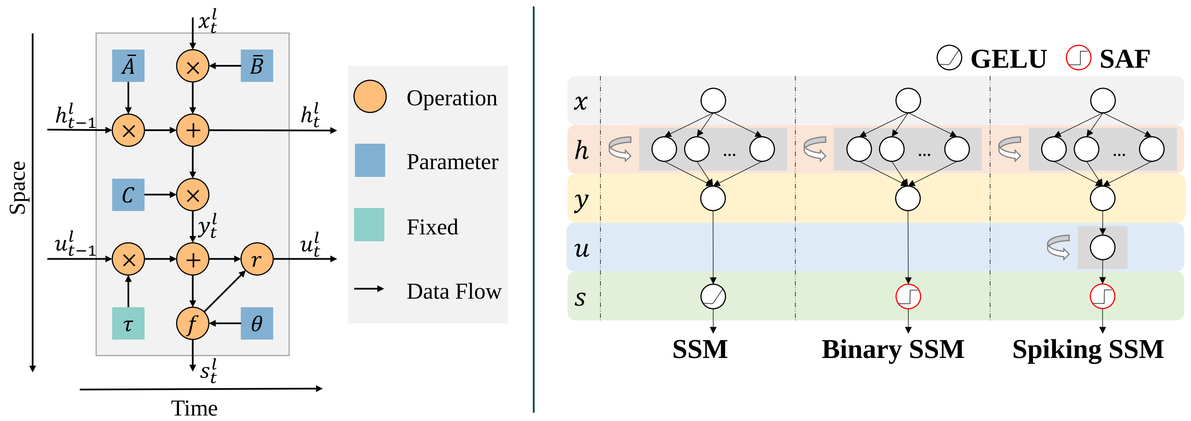
Task	Performance	Sparsity
LRA Benchmark	84.33% Avg. Accuracy	90%
WikiText-103	33.94 PPL (vs. SpikeGPT: 39.75)	74%
Path-X (16K steps)	94.82% Accuracy	90%

Training Speed:

- **101×** faster than BPTT for 8K-length sequences.

[Visual Highlights]

SpikingSSM Architecture



[Conclusion]

- SpikingSSM bridges biological plausibility and computational efficiency for long-sequence tasks.
- SDN enables scalable SNN training, making it practical for real-world applications.
- Achieves SOTA performance with low energy costs and high sparsity.

[Acknowledgment]

National Key R&D Program of China, NSFC, and STI 2030-Major Projects.