ICLR 2025

P-SpikeSSM: Harnessing Probabilistic Spiking State Space Models for Long-Range Dependency Tasks

Malyaban Bal, Abhronil Sengupta

School of EECS

The Pennsylvania State University

PennState

Motivation

- Developing P-SpikeSSM, a probabilistic neuronal model grounded in state-space models.

- Design a scalable framework for efficient parallel training of architectures leveraging the P-SpikeSSM neuronal model.

- Developing computationally efficient solutions for sequential learning with long-range dependencies.

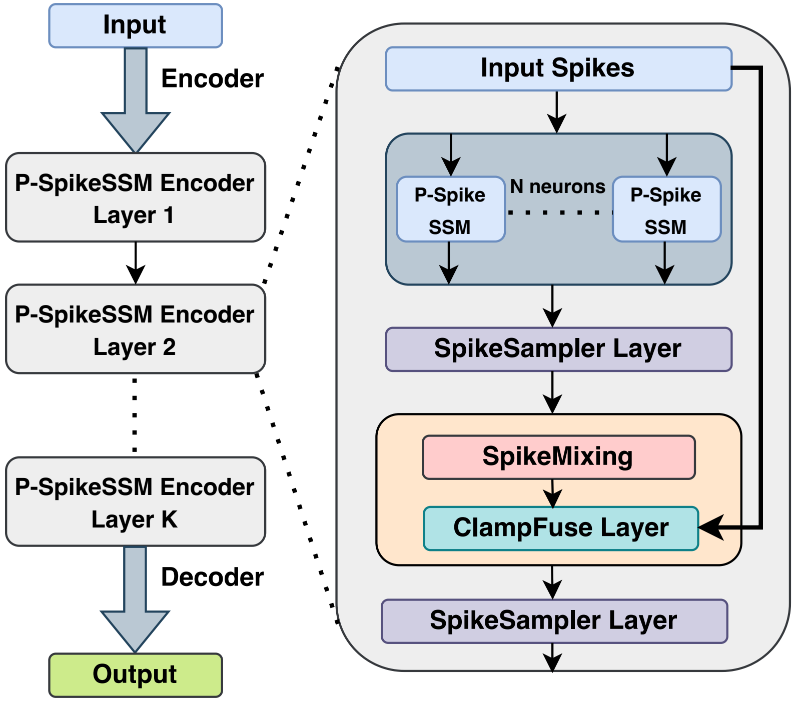


Figure 1. High-level overview of the P-SpikeSSM-based spiking model.

P-SpikeSSM Dynamics

- Continuous Time Recurrent Dynamics

$$\dot{h}(t) = Ah(t) + Bx_s(t)$$

Membrane potential dynamics

$$p_s(t) = \sigma(Ch(t) + Dx_s(t))$$

Spiking Probability

$$\sigma(z) = clamp(az + b)$$

$$clamp(y) = \begin{cases} 0 & \text{if } y < 0 \\ y & \text{if } 0 \le y \le 1 \\ 1 & \text{if } y > 1 \end{cases}$$

- Discrete Time Dynamics

$$h[t] = \overline{A}h[t-1] + \overline{B}x_s[t]$$

$$p_s[t] = \sigma(\overline{C}h[t])$$

$$\overline{A} = (I - \Delta/2 \cdot A)^{-1}(I + \Delta/2 \cdot A)$$

$$\overline{B} = (I - \Delta/2 \cdot A)^{-1}\Delta B$$

$$\overline{C} = C$$

- P-SpikeSSM neurons process sequence of spikes.
- The hidden state (h) is an n-dimensional latent space, unlike LIF's 1D membrane potential.
- P-SpikeSSM enable stochastic spike generation ($X_s \sim Bernoulli(p_s)$), unlike deterministic spike generation in LIF neurons.

Convolutional Dynamics and Spike Generation

- Recurrent Neuronal Dynamics can be represented as a convolution.

$$h[i] = \overline{A}^{i-1}\overline{B}x_s[1] + \overline{A}^{i-2}\overline{B}x_s[2] + \cdots + \overline{A}\overline{B}x_s[i-1] + \overline{B}x_s[i] = \sum_{j=1}^{i}(\overline{A}^{i-j}\overline{B}x_s[j])$$

$$p_s[i] = \sigma((K * X_s)_i) \quad \Big| \quad K = \overline{C}\hat{K} = (\overline{C}\overline{B}, \overline{C}\overline{A}\overline{B}, \ldots, \overline{C}\overline{A}^{L-1}\overline{B})$$

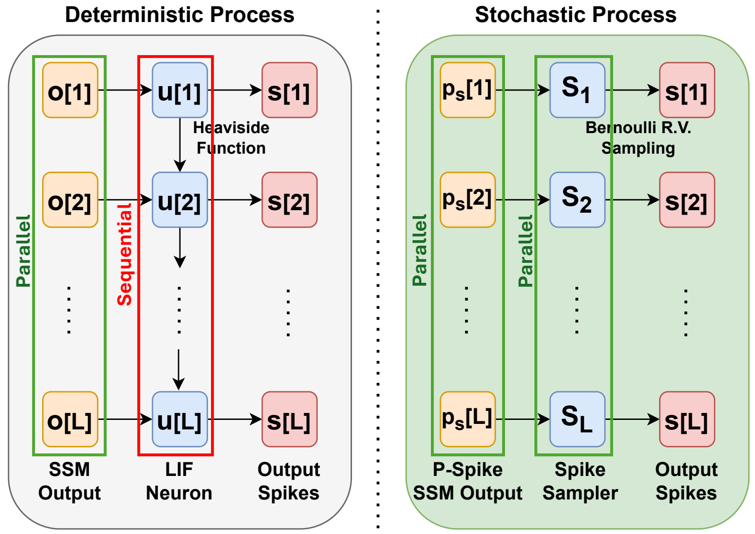


Figure 2. Computational flow of the LIF-based SSM model compared to the SpikeSampler-driven P-SpikeSSM neuronal model

- Spike Sampler Layer

Forward Pass :

$$S_t = \begin{cases} 1 & \text{if } z < p_s[t], \\ 0 & \text{otherwise}, \end{cases}$$

$$z \sim \mathcal{U}(0,1)$$

Backward Pass :

$$\overline{S}_t = \mathbb{E}[S_t] = 0 \cdot P(S_t = 0) + 1 \cdot P(S_t = 1) = p_s[t]$$

- Surrogate ($\bar{S}_t$) is used during backpropagation to overcome non-differentiability of S_t.

Scaling to Deeper Architectures

- **P-SpikeSSM neuronal layer** : Complex tasks requires multiple P-Spike Neurons to capture complex long-range dependencies.

- **SpikeMixer** Layer: Allows assimilation of information of P-SpikeSSM neurons of the previous layer, enabling inter-neuronal communication.

$$f_{mix}[t] = gelu(I_s[t] \cdot W_m)$$

- **FuseClamp** Layer: Aggregates input to the encoder block with the SpikeMixer output via a residual connection and applies normalization.

$$p_{fc_i}[t] = \sigma(BN(f_{mix}[t] + x_s[t]))$$

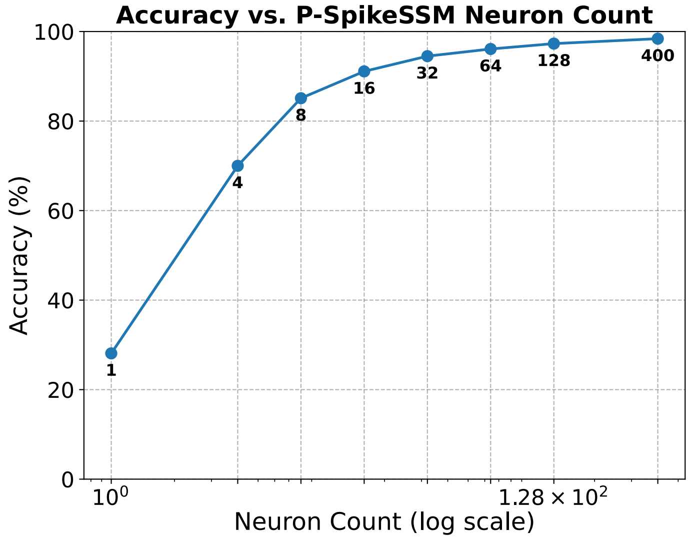
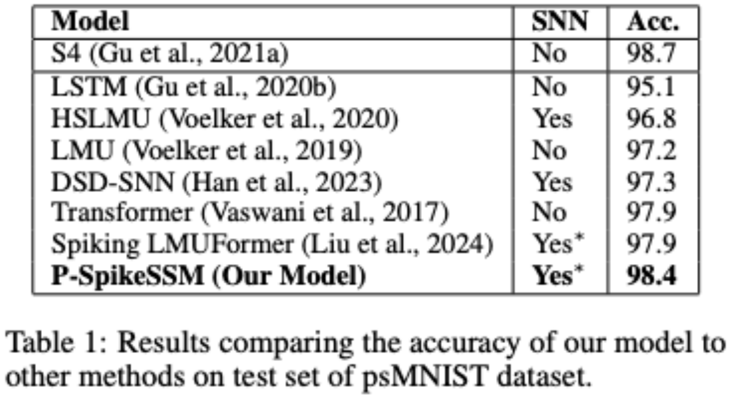


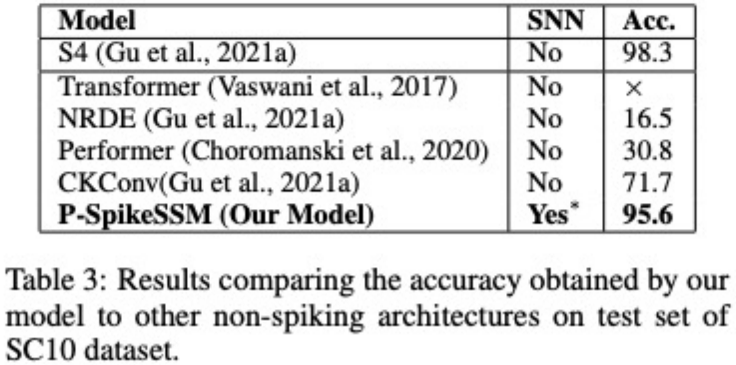
Figure 3. Results obtained from the test set of the ps-MNIST dataset.

Results



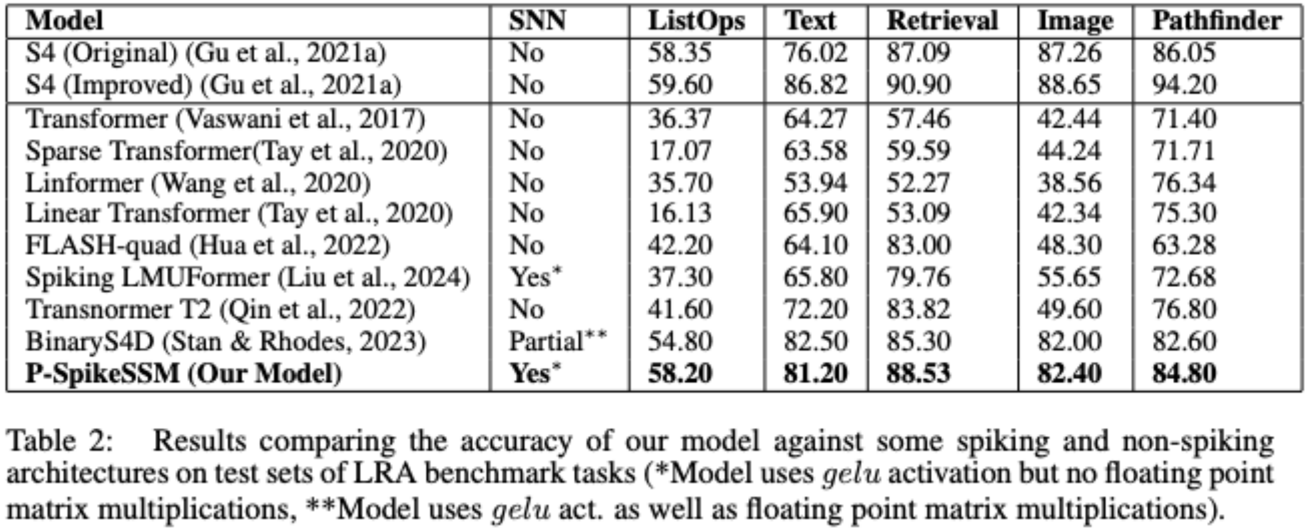
Model	SNN	Acc.
S4 (Gu et al., 2021a)	No	98.7
LSTM (Gu et al., 2020b)	No	95.1
HSLMU (Voelker et al., 2020)	Yes	96.8
LMU (Voelker et al., 2019)	No	97.2
DSD-SNN (Han et al., 2023)	Yes	97.3
Transformer (Vaswani et al., 2017)	No	97.9
Spiking LMUFormer (Liu et al., 2024)	Yes*	97.9
P-SpikeSSM (Our Model)	**Yes***	**98.4**

Table 1: Results comparing the accuracy of our model to other methods on test set of psMNIST dataset.



Model	SNN	Acc.
S4 (Gu et al., 2021a)	No	98.3
Transformer (Vaswani et al., 2017)	No	×
NRDE (Gu et al., 2021a)	No	16.5
Performer (Choromanski et al., 2020)	No	30.8
CKConv(Gu et al., 2021a)	No	71.7
P-SpikeSSM (Our Model)	**Yes***	**95.6**

Table 3: Results comparing the accuracy obtained by our model to other non-spiking architectures on test set of SC10 dataset.



Model	SNN	ListOps	Text	Retrieval	Image	Pathfinder
S4 (Original) (Gu et al., 2021a)	No	58.35	76.02	87.09	87.26	86.05
S4 (Improved) (Gu et al., 2021a)	No	59.60	86.82	90.90	88.65	94.20
Transformer (Vaswani et al., 2017)	No	36.37	64.27	57.46	42.44	71.40
Sparse Transformer(Tay et al., 2020)	No	17.07	63.58	59.59	44.24	71.71
Linformer (Wang et al., 2020)	No	35.70	53.94	52.27	38.56	76.34
Linear Transformer (Tay et al., 2020)	No	16.13	65.90	53.09	42.34	75.30
FLASH-quad (Hua et al., 2022)	No	42.20	64.10	83.00	48.30	63.28
Spiking LMUFormer (Liu et al., 2024)	Yes*	37.30	65.80	79.76	55.65	72.68
Transnormer T2 (Qin et al., 2022)	No	41.60	72.20	83.82	49.60	76.80
BinaryS4D (Stan & Rhodes, 2023)	Partial**	54.80	82.50	85.30	82.00	82.60
P-SpikeSSM (Our Model)	**Yes***	**58.20**	**81.20**	**88.53**	**82.40**	**84.80**

Table 2: Results comparing the accuracy of our model against some spiking and non-spiking architectures on test sets of LRA benchmark tasks (*Model uses *gelu* activation but no floating point matrix multiplications, **Model uses *gelu* act. as well as floating point matrix multiplications).

- Multiple long-range dependency tasks from datasets such as Long Range Arena Benchmark, Speech Command, Permuted Sequential MNIST.
- Our model out-performed transformer based non-spiking architectures and achieves SOTA performance among spiking architectures.

Energy Analysis

- P-SpikeSSM-based models exhibit sparse spiking activity resulting in low-power consumption.

- Considering only the energy consumption of arithmetic operations, the P-SpikeSSM-based model is 36 times more energy-efficient than an iso-architecture non-spiking model (on LRA ListOps dataset).

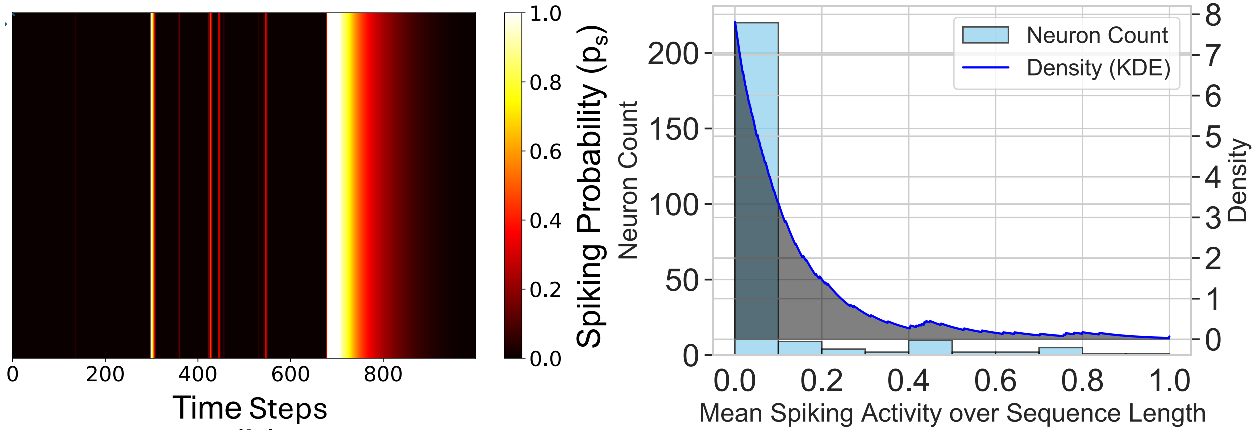


Figure 4. (a) Heatmap depicting the sparsity of spiking events generated by a single P-SpikeSSM neuron over input sequence length (for ListOps dataset) Results obtained from the test set of the ps-MNIST dataset. (b) Figure consists of histogram representing the count of neurons associated with mean probability of spiking (averaged over sequence) and Kernel Density Estimation (KDE) plot of the data using an exponential kernel.

Conclusions

- Exploring neuronal models beyond simple LIF based dynamics.
- Proposing a scalable training framework for P-SpikeSSM based spiking models.
- Leveraging spiking models for long-range dependency tasks.

Future Works

- Exploring generative models.
- Deploying P-SpikeSSM based model on neuromorphic chips.

Thank you