Spik4lite: Refactoring Neuromorphic Sparsity for Efficient Spiking Neural Networks on Commodity Edge Devices

Yongzhi She[1], Qihua Zhou[1], Yuhao Wang[1], Yaodong Huang[1], Laizhong Cui[2 1], Jingcai Guo[3]

[1]College of Computer Science and Software Engineering, Shenzhen University

[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China

[3]The Hong Kong Polytechnic University

Presenter: Yongzhi She

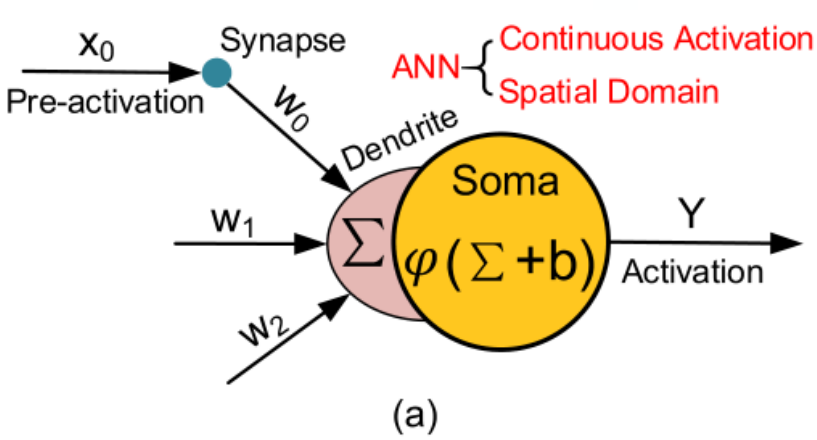
SHENZHEN UNIVERSITY 深圳大學

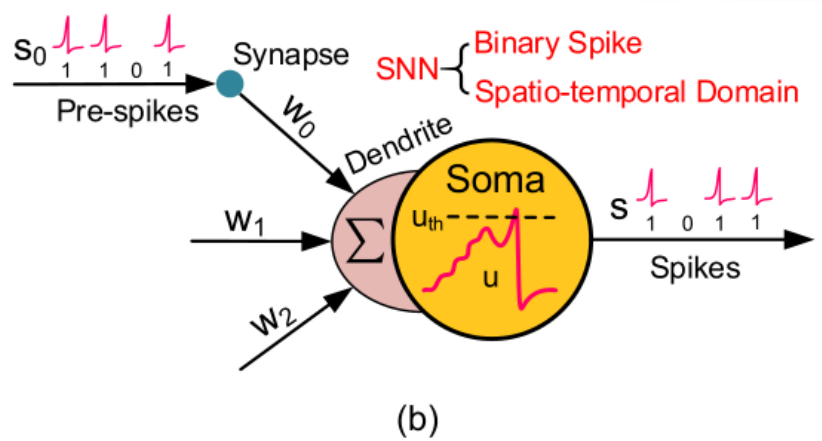
GUANGMING LABORATORY 光明实验室

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

2026.5

SNNs Are Promising for Efficient Edge AI

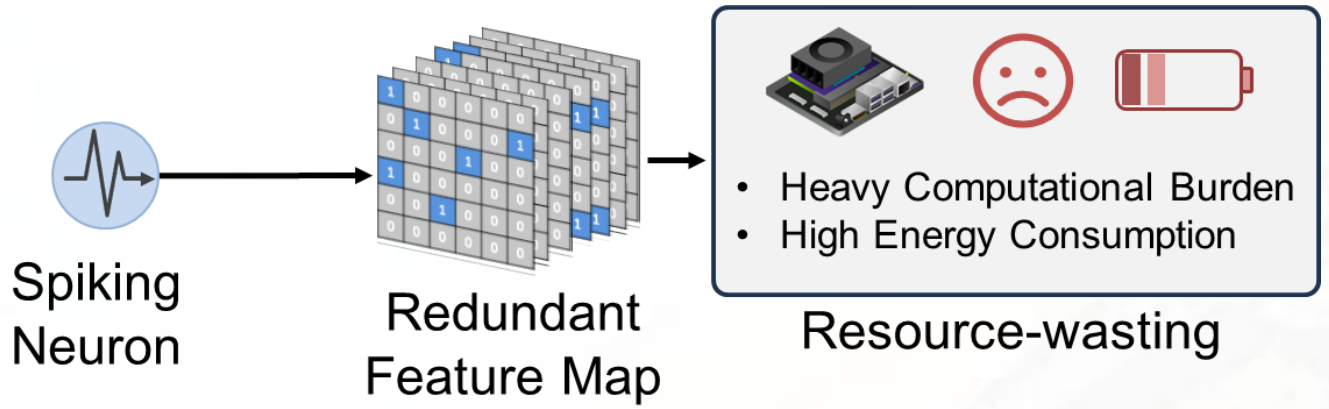




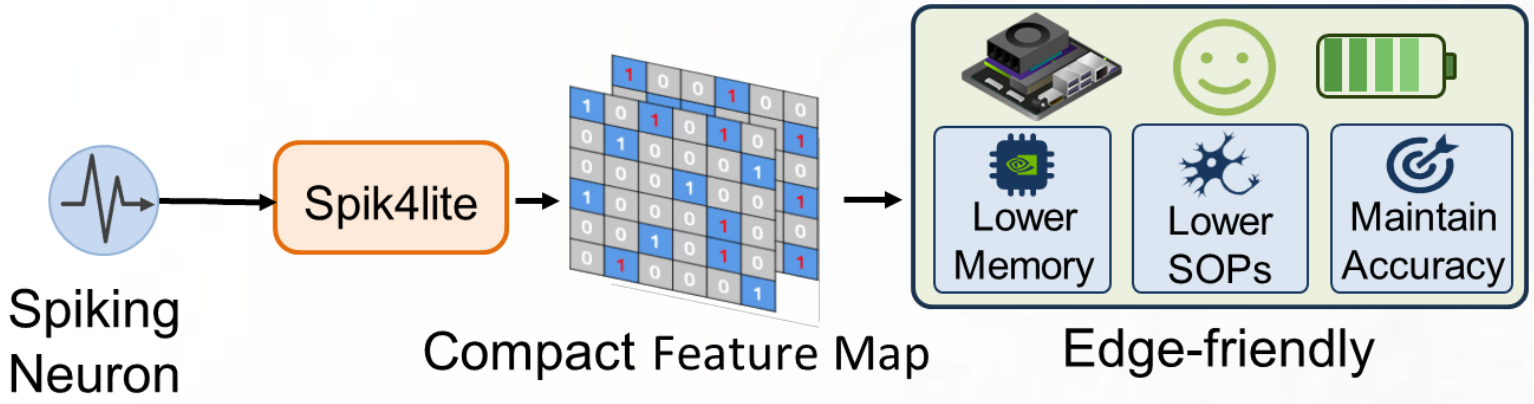
Spiking Neural Networks, SNNs:

- Third Generation of Neural Network
- Use binary spikes, 0/1, to mimic brain-inspired computation
- Event-driven and sparse computation

Commodity Hardware Cannot Fully Exploit Spike Sparsity



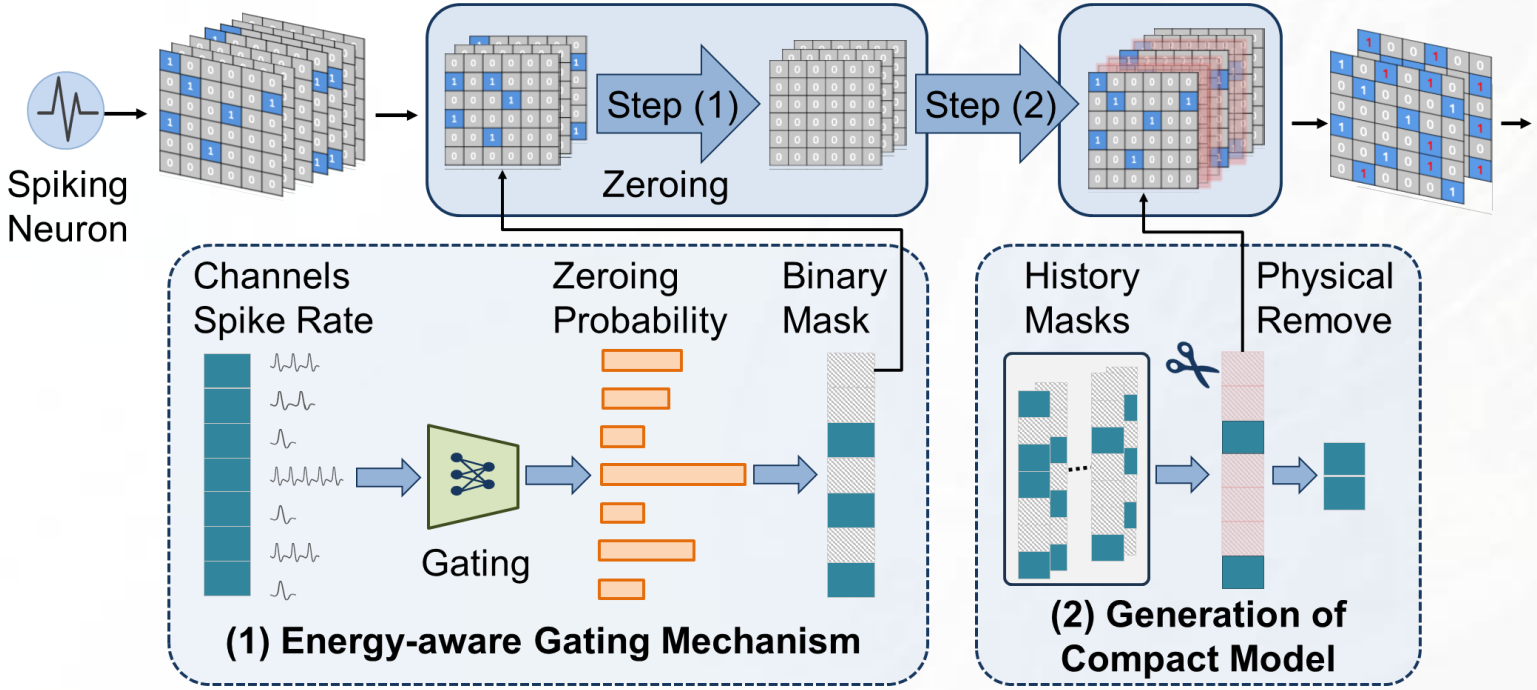
Is it possible to make SNNs edge-friendly and tame the computations mostly on active "1" spikes?



Three Key Designs to Achieve Real Efficiency for SNNs

The Detailed Process of Spik4lite

- SOPs-guided Energy-aware Gating
- Differentiable Gating via Hard Gumbel-Softmax
- Physical Channel Removal for Real Efficiency



Better Accuracy-efficiency Trade-off on Commodity Edge Device

Experiment Setup:

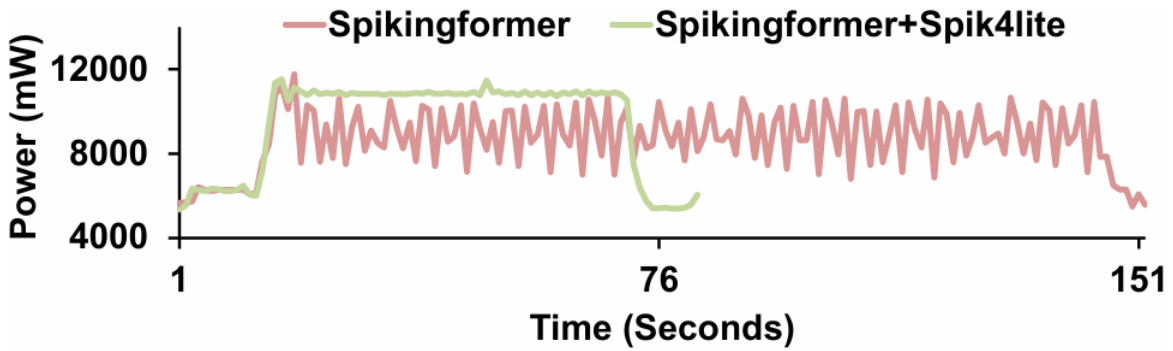
- Training: NVIDIA RTX4090
- Evaluation: NVIDIA Jetson Orin Nano 8G

Results:

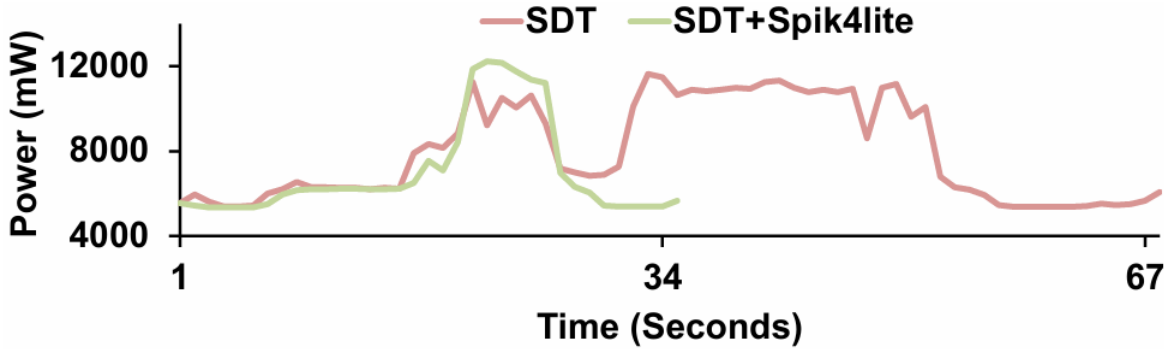
- Lower parameters & memory
- Lower Synaptic Operations (SOPs)
- Lower latency
- Lower energy
- Competitive accuracy

Dataset	Architecture	Batch Size	T	Top-1 Acc.(%)	Latency (s)	Peak Mem. (MiB)	Param. (M)	SOPs (M)	Energy (J)
CIFAR10	Spikformer-4-384	128	4	95.24	0.97	2614	9.33	822.64	5.34
	+ Spik4lite	128	4	94.15	**0.55**	**1613**	**2.85**	**321.49**	**3.08**
	SD-Transformer-2-512	128	4	95.41	0.92	2654	10.28	752.62	5.34
	+ Spik4lite	128	4	94.78	**0.65**	**2243**	**5.34**	**474.47**	**3.64**
	Spikingformer-4-384	128	4	95.07	0.95	2643	9.32	442.85	5.42
	+ Spik4lite	128	4	94.43	**0.61**	**1817**	**2.86**	**280.40**	**3.42**
CIFAR100	Spikformer-4-384	128	4	78.0	1.04	2615	9.36	1038.24	5.72
	+ Spik4lite	128	4	74.19	**0.56**	**1840**	**2.54**	**366.80**	**3.08**
	SD-Transformer-2-512	128	4	78.97	0.88	2655	10.28	835.88	5.11
	+ Spik4lite	128	4	76.25	**0.79**	**2526**	**6.06**	**601.78**	**4.43**
	Spikingformer-4-384	128	4	79.07	1.68	2644	9.32	533.71	9.74
	+ Spik4lite	128	4	77.07	**0.72**	**1893**	**5.05**	**378.66**	**3.96**
DVS-CIFAR10	Spikformer-4-384	16	16	79.9	0.54	2725	2.59	850.78	3.4
	+ Spik4lite	16	16	76.4	**0.51**	**2702**	**1.73**	**498.26**	**2.81**
	SD-Transformer-2-512	16	16	79.3	0.94	2623	2.57	2302.39	5.08
	+ Spik4lite	16	16	78.1	**0.71**	**1970**	**1.44**	**906.02**	**3.91**
	Spikingformer-4-384	16	16	80.7	0.55	2725	2.57	745.47	3.52
	+ Spik4lite	16	16	77.1	**0.53**	**2703**	**1.79**	**529.05**	**3.34**
DVS-128 Gesture	Spikformer-4-384	16	16	96.88	0.71	2725	2.59	554.83	4.33
	+ Spik4lite	16	16	94.1	**0.66**	**2699**	**0.65**	**207.71**	**4.09**
	SD-Transformer-2-512	14	16	97.56	1.65	2997	2.57	2551.85	9.24
	+ Spik4lite	14	16	94.1	**0.46**	**2390**	**0.63**	**170.21**	**2.85**
	Spikingformer-4-384	16	16	98.26	0.63	2725	2.57	535.78	3.84
	+ Spik4lite	16	16	94.79	**0.59**	**2701**	**1.11**	**438.85**	**3.54**

Real Energy Saving on Commodity Edge Device



(a) Spikingformer power profile on CIFAR-100



(b) SDT power profile on DVS-128 Gesture

Conclusion

Spik4lite: Unleash the capability of SNNs on commodity edge devices without relying on specific hardware

- **Bridging the hardware gap.** Spik4lite bridge the critical hardware gap to fundamentally improve the accuracy-and-efficiency trade-off of SNNs on non-neuromorphic chips, extending their practical usage boundary.

- **Real computation reduction.** Spik4lite directly translates theoretical SOPs reductions into practical computation reduction, thereby guaranteeing optimal energy efficiency for edge deployment.

- **Efficient plug-and-play implementation.** Spik4lite is compatible with existing SNN-Transformers, it can serve as a plug-and-play module to upgrade current SNN-Transformer models.

Thank you!

yongzhi.she@foxmail.com