SpikedAttention: Training-Free and Fully Spike-Driven Transformer-to-SNN Conversion with Winner-Oriented Spike Shift for Softmax Operation

Presenter: Sangwoo Hwang (DGIST, Korea)

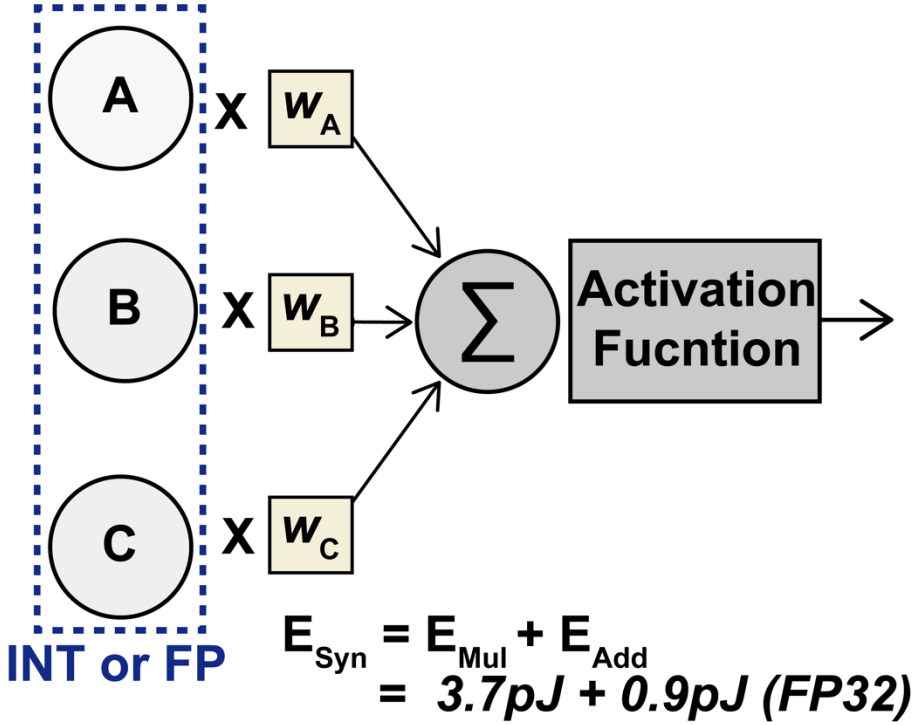
NEURAL INFORMATION PROCESSING SYSTEMS

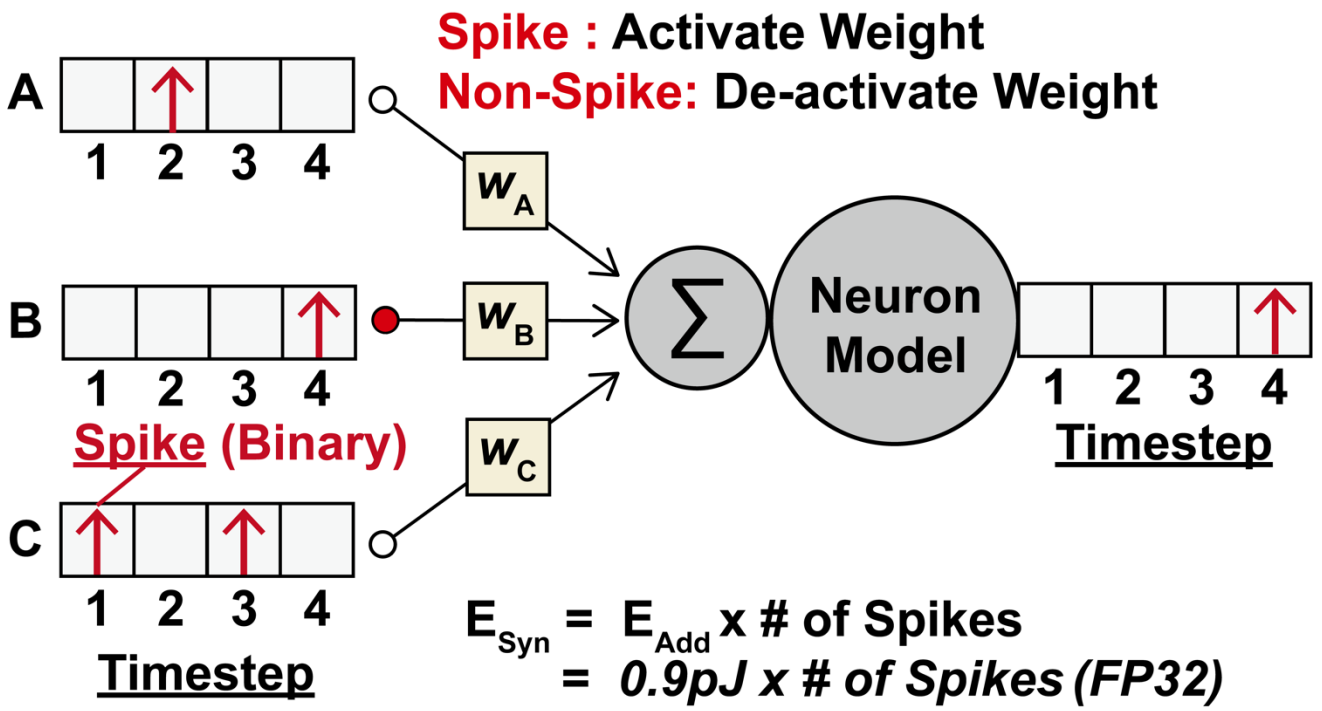
DGIST Daegu Gyeongbuk Institute of Science & Technology

KOREA UNIVERSITY

IDSLab Intelligent Digital Systems

Why Do We Need to Focus on SNN?





Pros

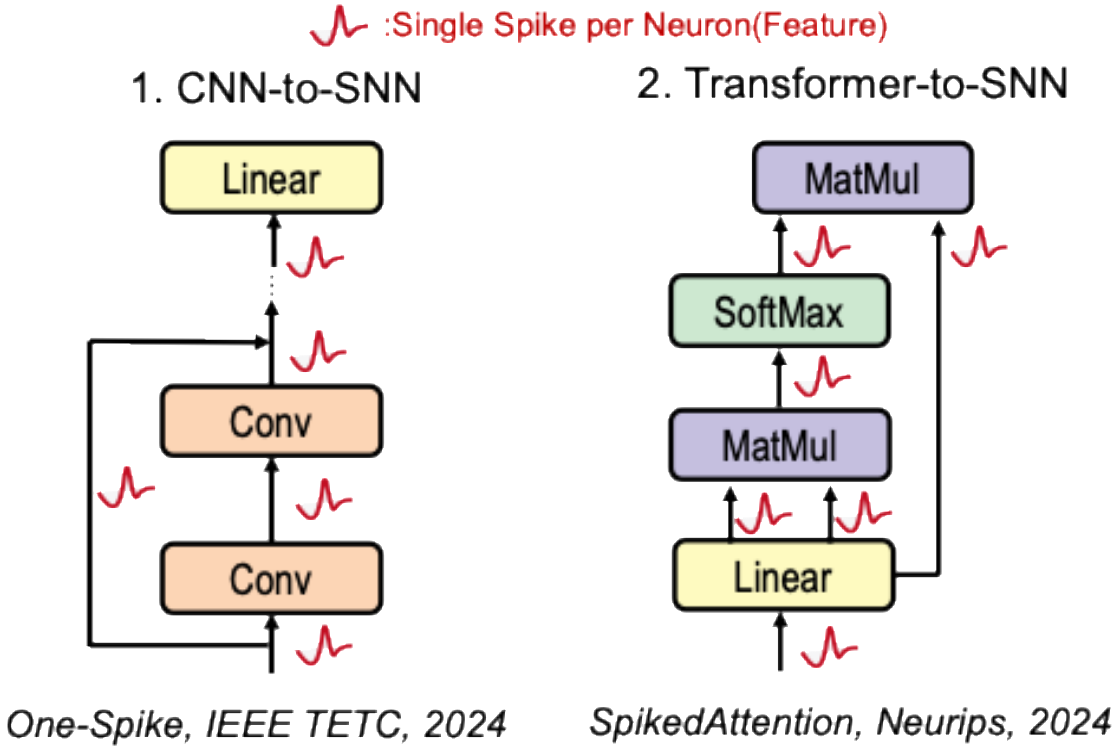
1. Activation is Binary
2. Only Accumulations
3. Sparse Inputs (= Spikes)

Cons

1. Complex Neuron Model
2. # of Neuron Updates ∝ # of Timesteps

(Possible Solution 🔑)
Increase Spike Sparsity or Reduce # of Timesteps

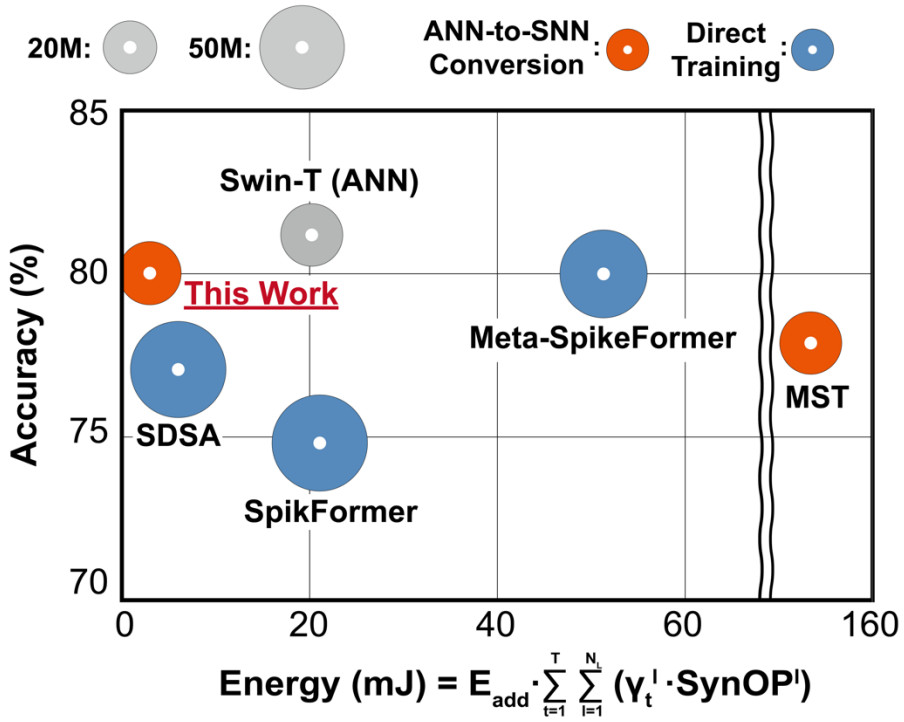
Overview of *"SpikedAttention"*



One-Spike, IEEE TETC, 2024

SpikedAttention, Neurips, 2024

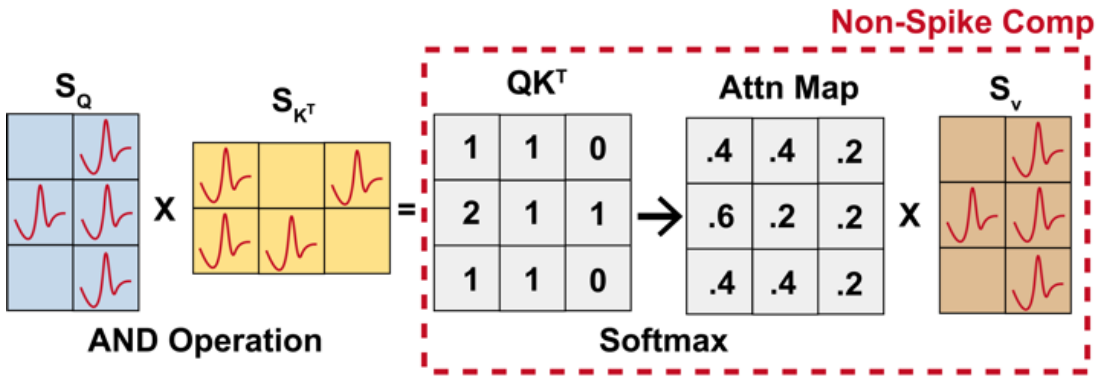
Performance of *SpikedAttention*



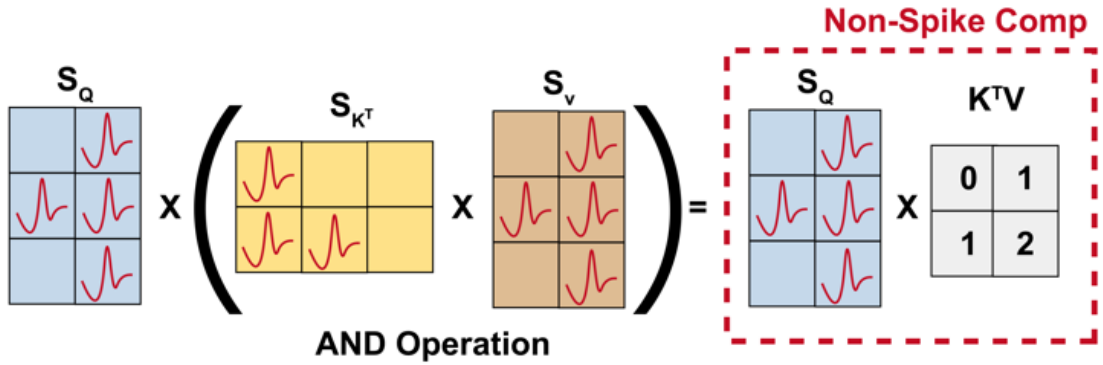
- No Constraints for Conversion ▶ Versatility ↑
- Fully Spike-based & Single Spike per Neuron ▶ Energy Efficiency ↑
- Training-Free ANN-to-SNN Conversion ▶ Training Cost ↓

Challenge: Absence of *"Fully Spike-driven"* Attention

Masked Spiking Transformer [1]



Spikformer [2]



(Prob. I) Dynamic Matrix Multiplication

1. AND Operation between Spikes (QK^T)
   → Requires Long Timestep due to High Sparsity

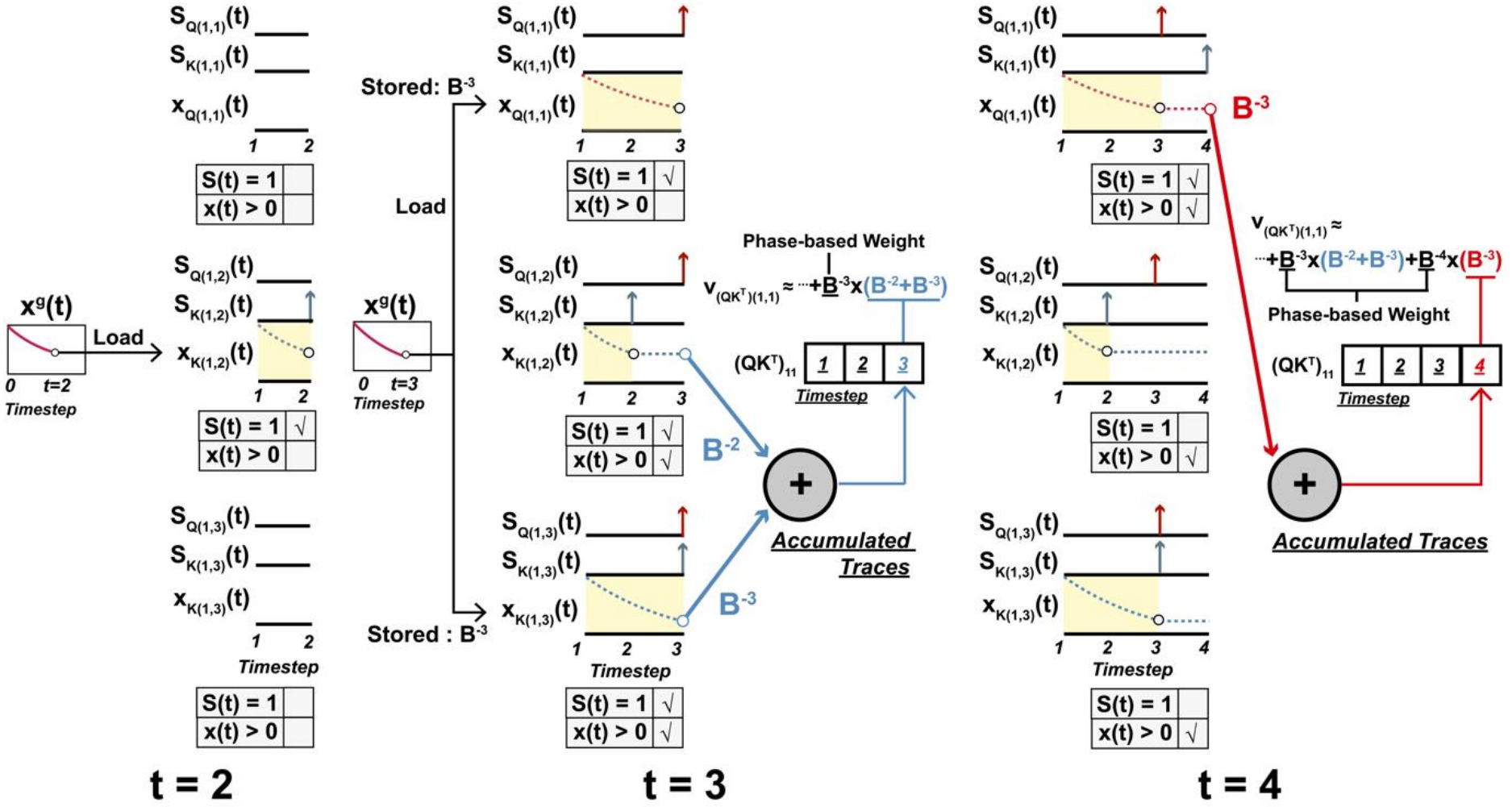
(Prob. II) Softmax

1. Conventional Softmax
   → Non-Spiked FP32 Computation
2. Removal of Softmax
   → High Training Cost

[1] Z. Wang, Y. Fang, J. Cao, Q. Zhang, Z. Wang, and R. Xu, "Masked spiking transformer," in Procs of ICCV, 2023.

[2] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. YAN, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," in Procs. of ICLR, 2023.

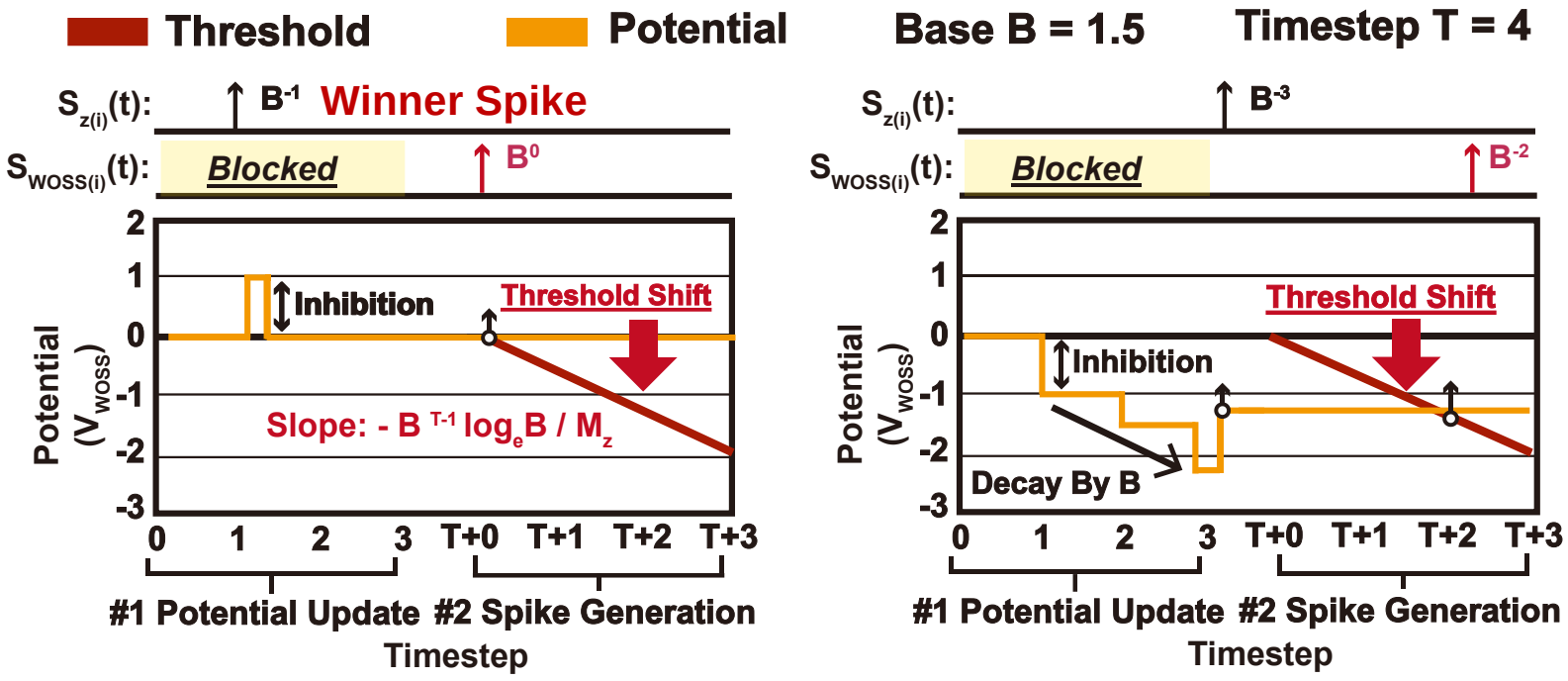
Trace-driven Multiplication for Dynamic Matrix Multiplication



A. **Keep Track of Spike Trace**
(= history of neuron spikes)

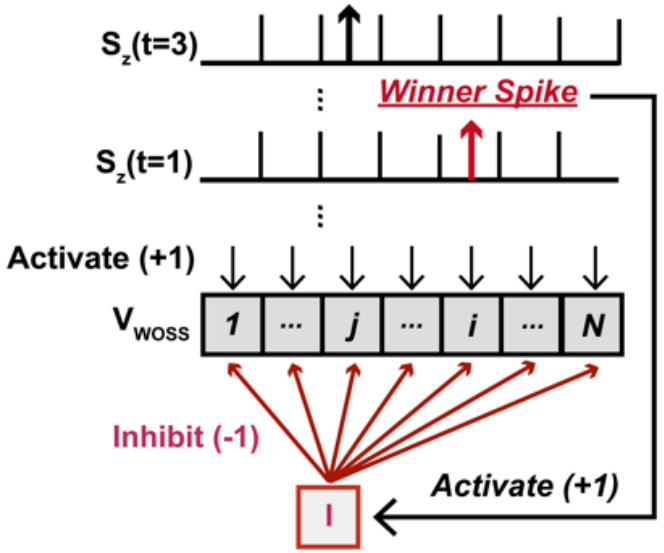
B. **Accumulation of Traces** at Paired Neuron's Spike when Pair Neuron have trace.

Winner-Oriented Spike Shift for Softmax



Winner Takes it all (WTA)

Diehl and Cook (Diehl & Cook, 2015)



Logarithmic Approximation of Softmax on One-spike

$$\sigma(z_i) = \frac{\exp(z_i)}{\sum_{j=0}^{N-1}\exp(z_j)} = \frac{\frac{\exp(z_i)}{\exp(\max(\boldsymbol{z}))}}{\sum_{j=0}^{N-1}\frac{\exp(z_j)}{\exp(\max(\boldsymbol{z}))}}$$

$$\log_B\left(\frac{\exp(M_z z_i')}{\exp(\max(M_z \boldsymbol{z}'))}\right) = (z_i' - \max(\boldsymbol{z}'))\frac{M_z}{\log_e B} \approx -t_i$$

* $M_z z_i'$ is normalized z_i

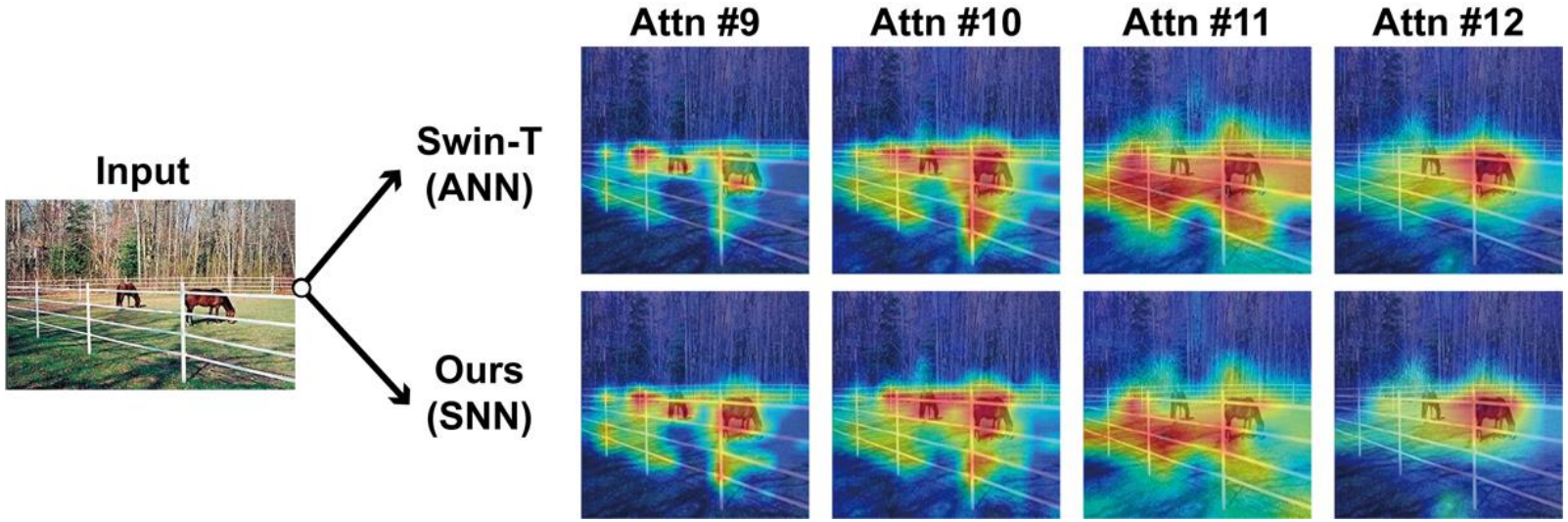
Phase 2: Threshold Shift

*** Scaling Threshold ≅ Dividing Output**

Phase 1: Subtraction by WTA

Result: SOTA Accuracy with Less # Parameters w/o Training

Comparison of attention maps based on Score-CAM between Swin-T (baseline ANN) and SpikedAttention



Comparison between SpikedAttention and the prior work on ImageNet classification task

Model	Param (M)	Energy (mJ)	Timestep	Acc (%)
Spikformer [2]	66.3	21.5	4	74.8
SDSA [3]	66.3	6.1	4	77.1
Meta-SpikeFormer [4]	55.4	52.4	4	80.0
MST [1]	28.5	158.6	128	77.9
Ours	**28.7**	**3.0**	**40**	**80.0**

Note: SpikedAtttention with long timestep use only one spike per Neuron ▶ Low energy Cost

Result: BERT-to-SNN Conversion

Dataset	COLA	MNLI	MRPC	QNLI
MA-BERT (ANN)				
Accuracy (%)	59.8	84.7	84.3	91.4
Energy (mJ)	189.7	189.7	189.7	189.7
SpikingBERT (SNN)				
Accuracy (%)	-	78.1	79.2	85.2
SpikedAttention (SNN)				
Timestep	**24**	**24**	**16**	**24**
Accuracy (%)	**59.3**	**84.4**	**84.1**	**91.0**
Energy (mJ)	**81.5**	**82.1**	**77.5**	**81.6**

→ **58% Energy Reduction and 3.6% Higher Accuracy than SpikingBERT**

(Contributions)

1. No Model Modification
2. Fully Spike-based Computing
3. Single Spike per Neuron
4. Training-Free ANN-to-SNN Conversion