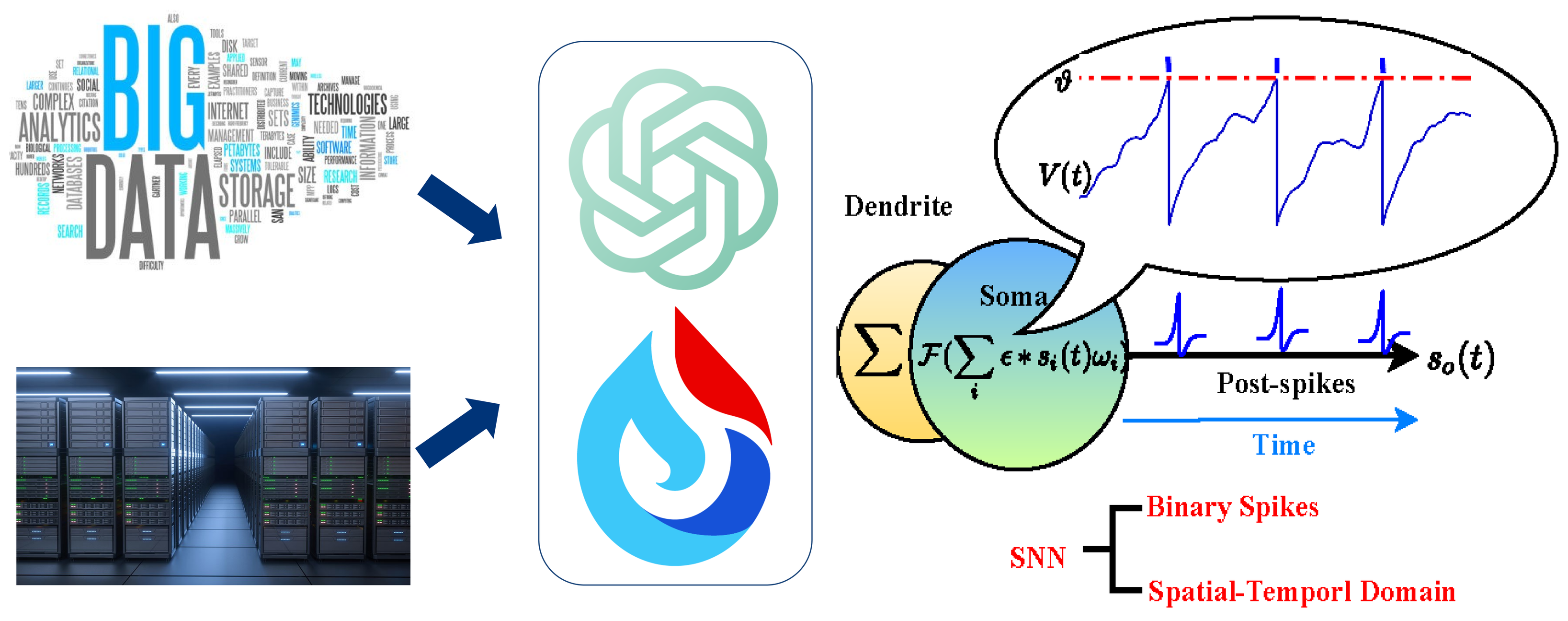
Quantized Spike-driven Transformer

Xuerui Qiu, Malu Zhang, Jieyuan Zhang, Wenjie Wei, Honglin Cao, Junsheng Guo ,Rui-Jie Zhu, Yimeng Shan, Yang Yang, Haizhou Li

ICLR

Code Paper

Motivation

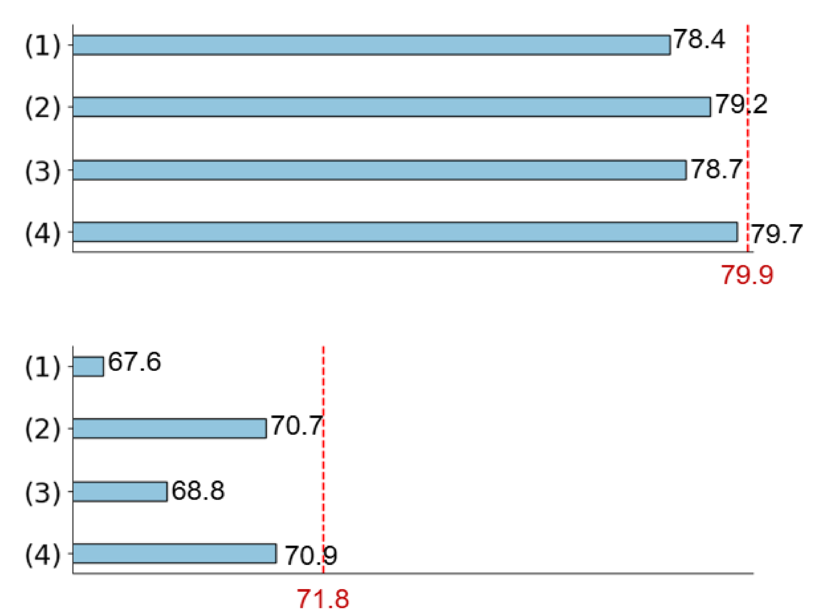


(a) High energy consumption (b) Energy-efficient alternative

Tradition Transformers significant demands on the storage andcomputational capabilities of neuromorphic chips, thereby limiting their deployment on edge devices. Brain-inspired spiking neural networks (SNNs) provide an energy efficient alternative to deep learning.

Problem Analysis

➢ Directly quantize leads to performance degradation



➢ Spike Information Distortion (SID) Problem

Problem Define

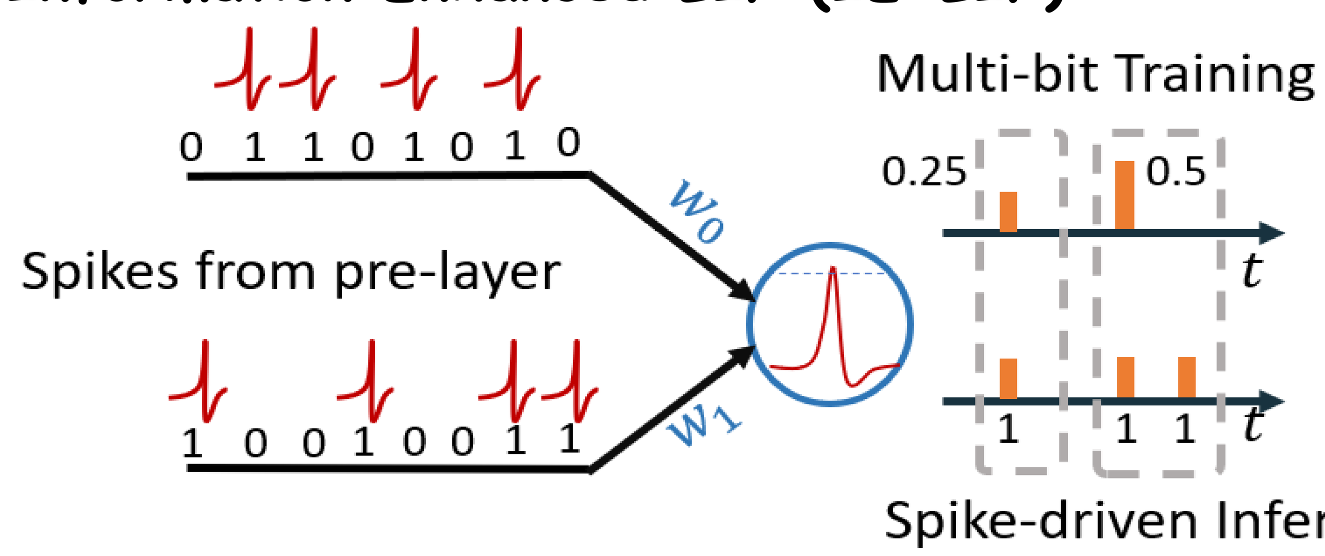
$$\max_{\theta^S} \mathcal{I}(\mathbf{p}^S;\mathbf{p}^A) = \mathcal{H}(\mathbf{p}^S) - \mathcal{H}(\mathbf{p}^S|\mathbf{p}^A),$$

$$\min_{\theta^S} \mathcal{H}(\mathbf{p}^{S^*}|\mathbf{p}^A), \quad \text{s.t.} \quad \mathbf{p}^{S^*} = \arg\max_{\mathbf{p}^S} \mathcal{H}(\mathbf{p}^S).$$

Addressing the performance degradation of the QSD-Transformer baseline is equivalent to maximizing the mutual information entropy between it and the quantized Transformer in ANNs

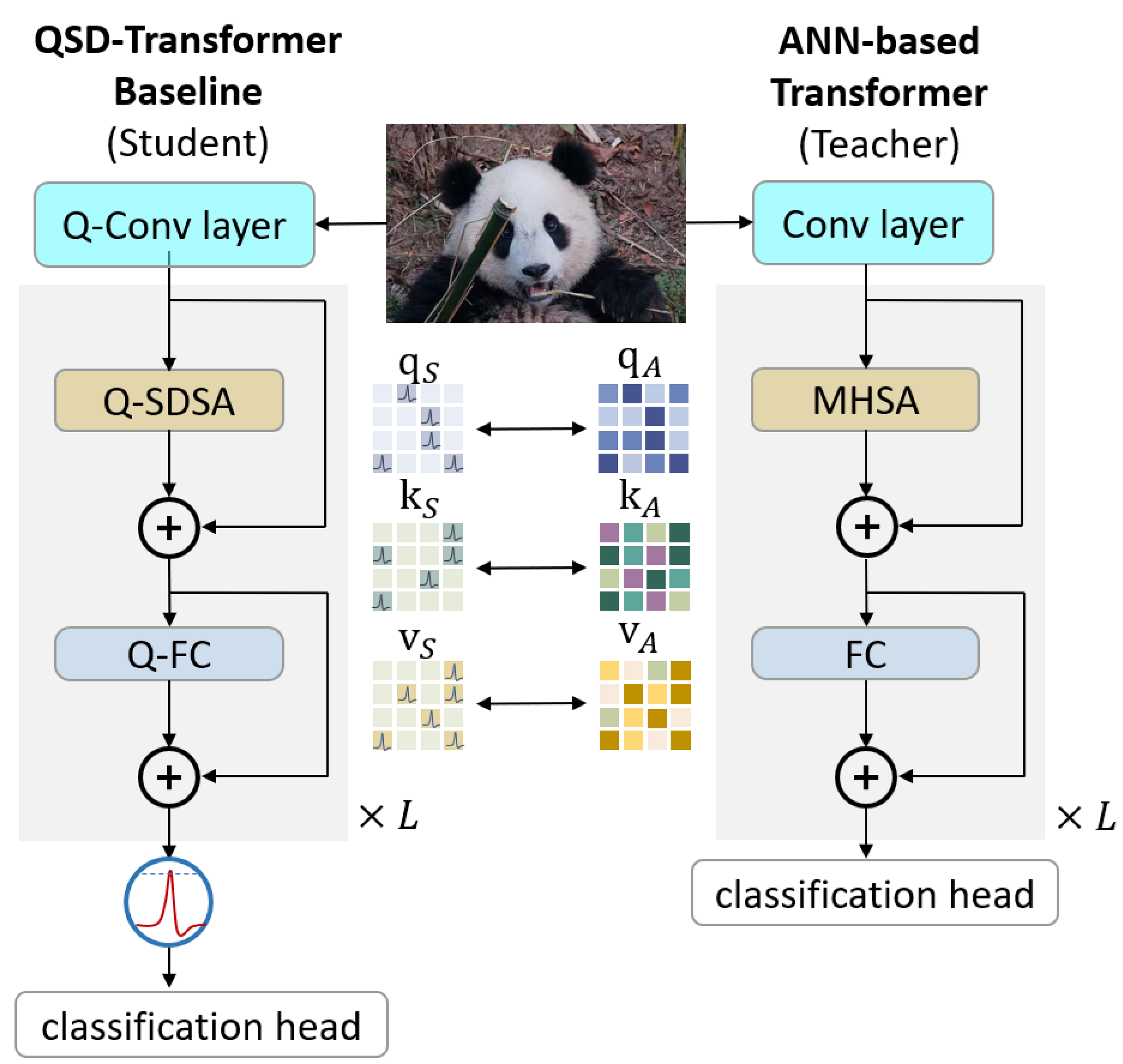
Method

➢ Information Enhanced LIF (IE-LIF)



we propose the information-enhanced LIF (IE-LIF) neuron and adjust the information distribution of Q-SDSA at the lower level, focusing on maximizing the information entropy.

➢ Fine-grained Distillation (FGD)



we achievethe optimization goal of problem define by proposing a fine-grained distillation (FGD), which adjusts the distribution of Q-SDSA at the upper level to minimize the conditional entropy.

Experimental Results

➢ Performance Comparison

➢ ImageNet classification results

Method	Architecture	Bits	Spike -driven	Time Step	Param (M)	Power (mJ)	Acc. (%)
Transformer (Yu et al., 2023)	CAformer*	32-32	✗	N/A	15.1	40.3	79.9
QCFS (Bu et al., 2021)	ResNet-34	32-1	✓	256	21.8	-	73.4
MST (Wang et al., 2023)	Swin-T	32-1	✓	128	28.5	-	77.9
SEW-ResNet (Fang et al., 2021)	SEW-ResNet-34	32-1	✗	4	25.6	4.9	67.8
	SEW-ResNet-152	32-1	✗	4	60.2	12.9	69.2
MS-ResNet(Hu et al., 2024b)	MS-ResNet-34	32-1	✓	4	21.8	5.1	69.4
	MS-ResNet-104	32-1	✓	4	77.3	10.2	75.3
Spikformer (Zhou et al., 2023b)	Spikformer-8-512	32-1	✗	4	29.7	11.6	73.4
	Spikformer-8-768	32-1	✗	4	66.3	21.5	74.8
SD-Transformer (Yao et al., 2023b)	SD-Transformer-8-512	32-1	✓	4	29.7	4.5	74.6
	SD-Transformer-8-768	32-1	✓	4	66.3	6.1	76.3
SpikingResformer (Shi et al., 2024)	SpikingResformer-T	32-1	✓	4	11.1	4.2	74.3
	SpikingResformer-L	32-1	✓	4	60.4	9.7	78.7
SD-Transformer v2 (Yao et al., 2023a)	SD-Transformer v2-T	32-1	✓	4	15.1	16.7	74.1
	SD-Transformer v2-M	32-1	✓	4	31.3	32.8	77.2
	SD-Transformer v2-L	32-1	✓	4	55.4	52.4	79.7
QSD-Transformer	SD-Transformer v2-T	4-1	✓	4	1.8	**2.5**	77.5
	SD-Transformer v2-M	4-1	✓	4	3.9	5.7	78.9
	SD-Transformer v2-L	4-1	✓	4	6.8	8.7	**80.3**

➢ Coco2017 Object detection results

Method	Architecture	Bits	Spike -driven	Time Step	Param (M)	Power (mJ)	mAP@0.5 (%)
Transformer (Yu et al., 2023)	CAformer	32-32	✗	N/A	31.2	890.6	54.0
Transformer (Zhu et al., 2020)	DETR	32-32	✗	N/A	41.0	860.2	57.0
Spiking-Yolo (Kim et al., 2020)	ResNet-18	32-1	✓	3500	10.2	-	25.7
Spike Calibration (Li et al., 2022)	ResNet-18	32-1	✓	512	17.1	-	45.3
EMS-SNN (Su et al., 2023)	EMS-ResNet-18	32-1	✓	4	26.9	-	50.1
SD-Transformer v2 (Yao et al., 2023a)	SD-Transformer v2-M	32-1	✓	1	75.0	140.8	51.2
QSD-Transformer	SD-Transformer v2-T	4-1	✓	4	16.9	**45.1**	48.1
	SD-Transformer v2-M	4-1	✓	4	34.9	117.2	**57.0**

➢ ADE20K Semantic segmentation results

Method	Architecture	Bits	Spike -driven	Time Step	Param (M)	Power (mJ)	MIoU (%)
Segformer (Xie et al., 2021)	Segformer	32-32	✗	N/A	3.8	38.9	37.4
DeepLab-V3 (Zhang et al., 2022a)	DeepLab-V3	32-32	✗	N/A	68.1	1240.6	42.7
SD-Transformer v2 (Yao et al., 2023a)	SD-Transformer v2-M	32-1	✓	4	59.8	183.6	35.3
QSD-Transformer	SD-Transformer v2-T	4-1	✓	4	3.3	**17.5**	39.0
	SD-Transformer v2-M	4-1	✓	4	9.6	37.9	**40.5**

➢ Transfer learning results on CIFAR10, CIFAR100,CIFAR10-DVS

Method	Param (M)	CIFAR10		CIFAR100		CIFAR10-DVS	
		T	Acc. (%)	T	Acc. (%)	T	Acc. (%)
Spikformer (Zhou et al., 2023b)	29.1	4	97.0	4	83.8	-	-
SpikingResformer (Shi et al., 2024)	17.3	4	97.4	4	85.9	10	84.8
QSD-Transformer	1.8	4	97.8±0.1	4	86.6±0.3	10	88.8±0.1
	6.8	4	**98.4**±0.2	4	**87.6**±0.2	10	**89.8**±0.1

The QSD-Transformer obtains SOTA performance, achieving a harmonious balance between accuracy and power.

➢ Ablation Study

Architecture	IE-LIF	FGD	Weight Bits	Acc.(%)
SD-Transformer v2 (Yao et al., 2023a)	-	-	4	70.0
	✓	-	4	75.8
	✓	✓	4	**77.5**
	✓	✓	3	76.9
	✓	✓	2	75.0
Spikformer (Zhou et al., 2023b)	-	-	4	64.1
	✓	-	4	70.1
	✓	✓	4	**75.5**
	✓	✓	3	74.1
	✓	✓	2	73.1

Both the proposed IE-LIF neuron and the FGD scheme can improve performance.

Conclusion

- We propose the QSD-Transformer, achieving energy efficiency via low-bit weights and 1-bit spikes.
- We identify performance degradation in QSD-Transformer due to the SID problem.
- We develop a bi-level optimization strategy: IE-LIF neurons for multi-bit spikes in training and FGD scheme for optimized attention distribution.
- Our method delivers state-of-the-art performance and efficiency on vision tasks, enabling practical deployment of spike-based Transformers on resource-limited platforms.