Q-SNNs: Quantized Spiking Neural Networks

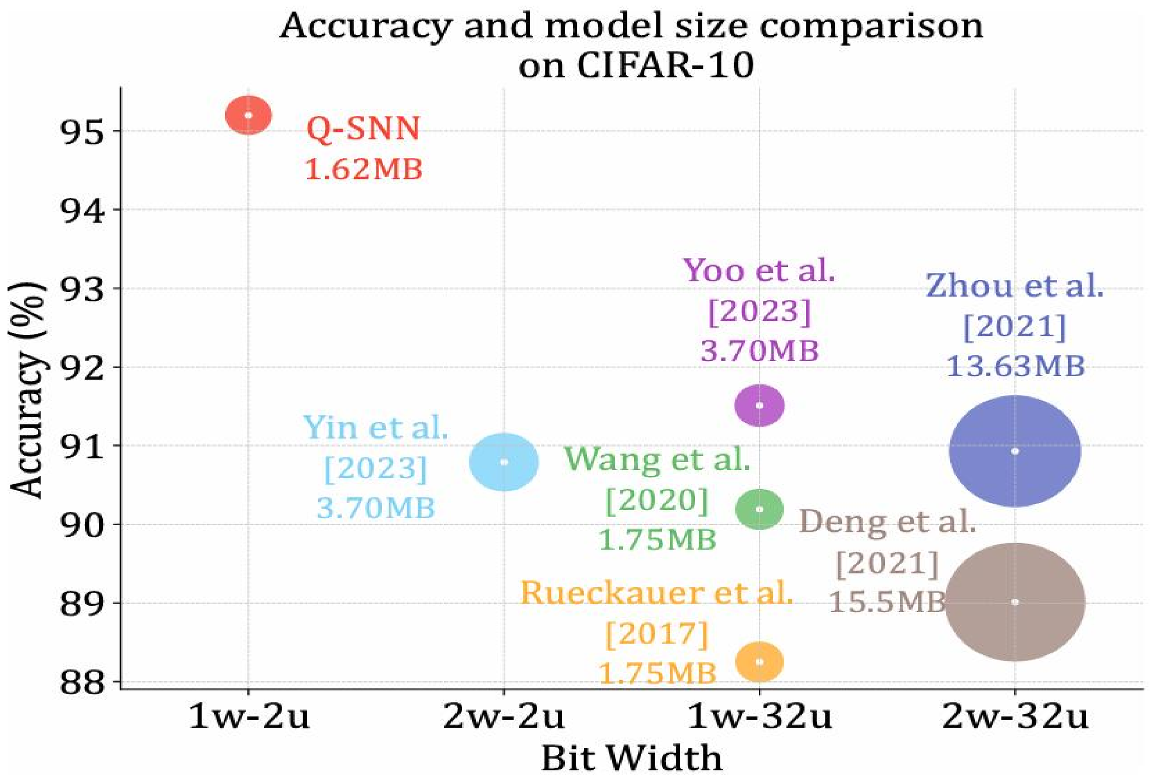
University of Electronic Science and Technology of China

Northumbria University NEWCASTLE

Wenjie Wei, Yu Liang, Ammar Belatreche, Yichen Xiao, Honglin Cao, Zhenbang Ren, Guoqing Wang, Malu Zhang, Yang Yang

ACM Multimedia 2024 Melbourne, Australia

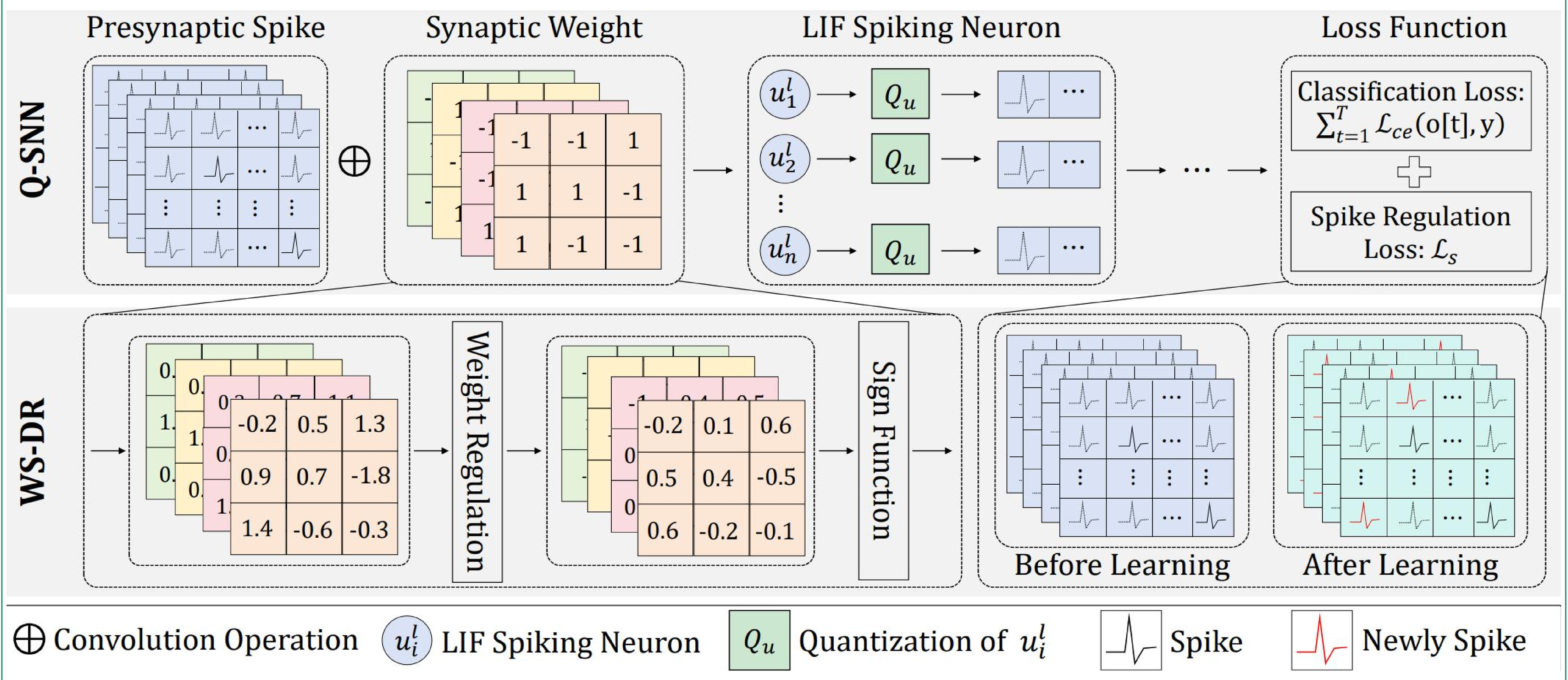
Motivation



- Spiking Neural Networks (SNNs) provide an energy-efficient paradigm for the next generation of machine intelligence.
- However, the current SNN community focuses mainly on accuracy improve-ment by developing large-scale models, which limits the applicability of SNNs in resource-limited edge devices.
- We propose a lightweight Quantized SNN (Q-SNN) that quantizes both weights and membrane potentials, significantly reducing memory usage and computa-tional complexity.

Method

Quantized Spiking Neural Network (Q-SNN)



- Firstly, the proposed Q-SNN quantizes the synaptic weight into a 1-bit representation, which is formulated as:

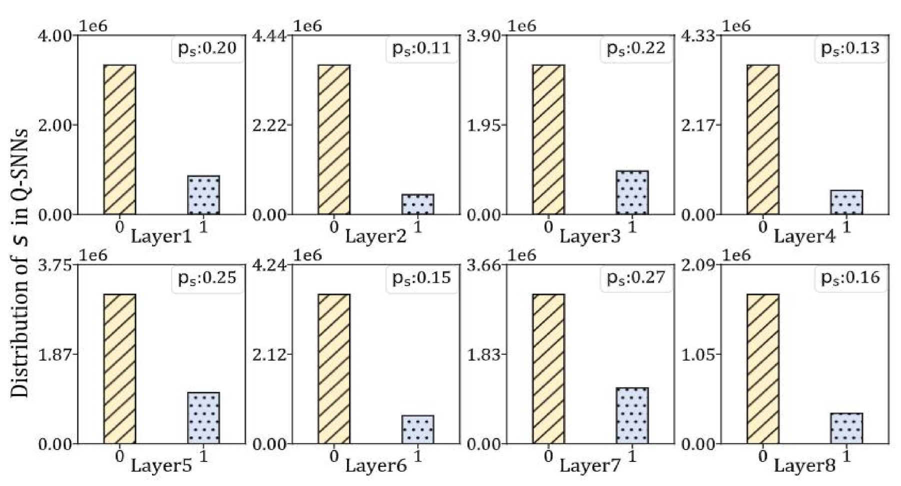
$$Q_w(w) = \alpha_w \cdot sign(w), \quad sign(w) = \begin{cases} +1, \ if\ w \ \geq \ 0, \\ -1, \ otherwise. \end{cases}$$

- Secondly, Q-SNN quantizes the membrane potential to a low bit-width integer, such as 2, 4, and 8, described as:

$$Q_u(u) = \frac{\alpha_u}{2^{k-1}-1} round\left((2^{k-1}-1) clip\left(\frac{u}{\alpha_u}, -1, 1\right)\right).$$

Challenges Analysis of Q-SNN

- While Q-SNNs exhibit significant energy efficiency, their task performance lags significantly behind full-precision SNNs.
- Inspired by the information theory, we attribute this performance gap to the limited information representation capability of Q-SNNs.



Results:
1. p_s in each layer approaches 0, resulting in severely limited information content carried by s.
2. The weights in Q-SNNs face the same challenge.

Weight-Spike Dual Regulation Method

- For the 1-bit weight in Q-SNN, we apply a normalization technique:

$$\widehat{W^l} = (W^l - \mu_l)/\sigma_l \, .$$

- For the 1-bit spike activity, we design a loss function:

$$\mathcal{L}_s = \sum_{l=2}^{L-1} (f_l - 0.5)^2, \quad f_l = \frac{1}{N_l \times T}\left(\sum_{i=1}^{N_l}\sum_{t=1}^{T} s_i^l[t]\right).$$

By integrating these two approaches, the weight and spike in Q-SNNs can carry more information content, thus mitigating the performance degradation caused by information loss during the quantization process.

Experimental Results

Performance Comparison

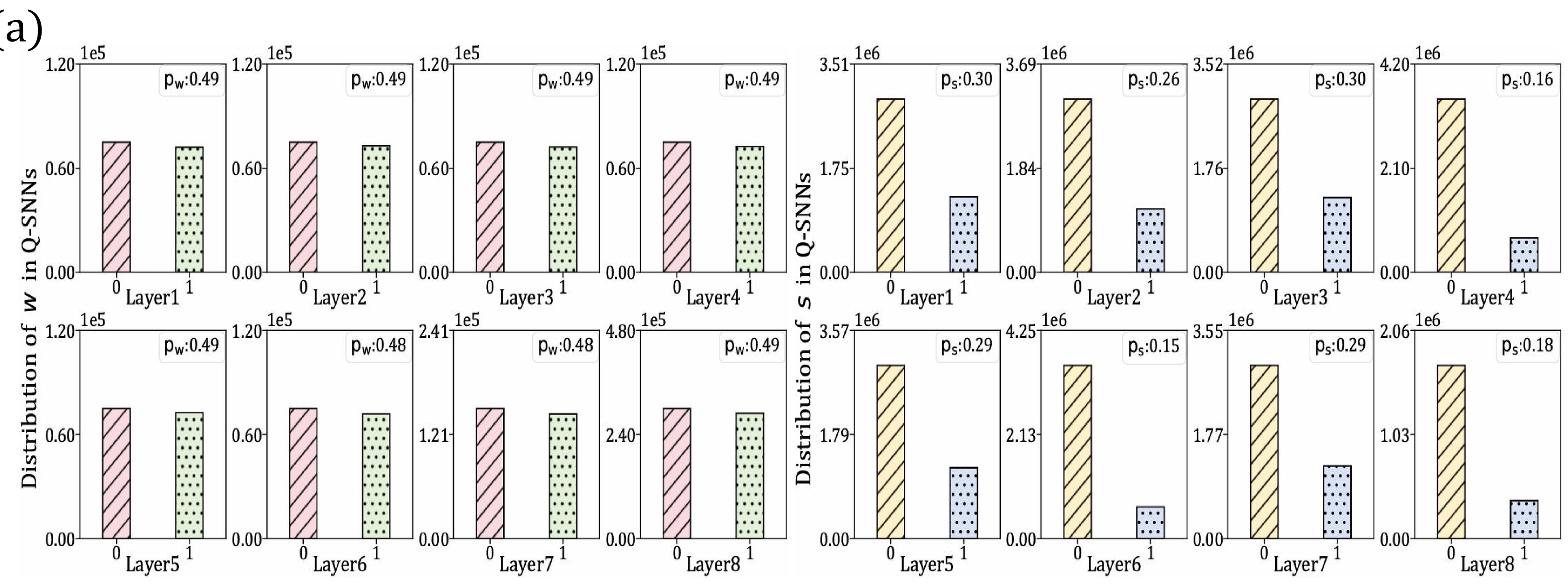
Table 1: Classification performance comparison on both static image datasets and neuromorphic datasets.

Dataset	Method	Architecture	Learning	Bit Width	Timestep	Accuracy
CIFAR-10	Full-Precision SNN‡	ResNet19	Direct train	32w-32u[1]	2	96.36%
	Roy et al. [43]	VGG9	ANN2SNN	1w-32u	-	88.27%
	Rueckauer et al. [44]	6Conv3FC	ANN2SNN	1w-32u	-	88.25%
	Wang et al. [50]	6Conv3FC	ANN2SNN	1w-32u	100	90.19%
	Yoo et al. [60]	VGG16	ANN2SNN	1w-32u	32	91.51%
	Deng et al. [12]	7Conv3FC	Direct train	1w-32u	8	89.01%
	Pei et al. [35]	5Conv1FC	Direct train	1w-32u	1	92.12%
	Zhou et al. [63]	VGG16	Direct train	2w-32u	-	90.93%
	Yin et al. [59]	ResNet19	Direct train	2w-2u	4	90.79%
	Proposed Q-SNN	ResNet19	Direct train	1w-8u	2	**95.54%**
				1w-4u	2	**95.31%**
				1w-2u	2	**95.20%**
CIFAR-100	Full-Precision SNN‡	ResNet19	Direct train	32w-32u	2	79.52%
	Roy et al. [43]	VGG16	ANN2SNN	1w-32u	-	54.44%
	Lu et al. [31]	VGG15	ANN2SNN	1w-32u	400	62.07%
	Wang et al. [50]	6Conv2FC	ANN2SNN	1w-32u	300	62.02%
	Yoo et al. [60]	VGG16	ANN2SNN	1w-32u	32	66.53%
	Deng et al. [12]	7Conv3FC	Direct train	1w-32u	8	55.95%
	Pei et al. [35]	6Conv1FC	Direct train	1w-32u	1	69.55%
	Proposed Q-SNN	ResNet19	Direct train	1w-8u	2	**78.77%**
				1w-4u	2	**78.82%**
				1w-2u	2	**78.70%**
TinyImageNet	Full-Precision SNN‡	VGG16	Direct train	32w-32u	4	56.77%
	Yin et al. [59]	VGG16	Direct train	8w-8u	4	50.18%
			Direct train	4w-4u	4	49.36%
			Direct train	2w-2u	4	48.60%
	Proposed Q-SNN	VGG16	Direct train	1w-8u	4	**55.70%**
				1w-4u	4	**55.20%**
				1w-2u	4	**55.04%**
DVS-CIFAR10	Full-Precision SNN‡	VGGSNN	Direct train	32w-32u	10	82.10%
	Qiao et al. [37]	2Conv2FC	Direct train	1w-32u	25	62.10%
	Pei et al. [35]	5Conv1FC	Direct train	1w-32u	10	68.98%
	Yoo et al. [60]	16Conv1FC	Direct train	1w-32u	16	74.70%
	Proposed Q-SNN	VGGSNN	Direct train	1w-8u	10	**81.60%**
				1w-4u	10	**81.50%**
				1w-2u	10	**80.00%**

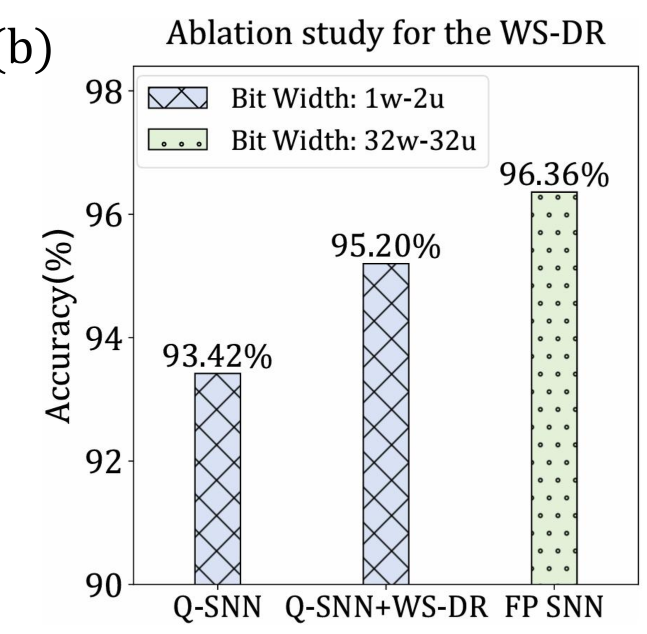
Q-SNN further exploits the efficiency advantage inherent to SNNs while upholding superior performance, offering substantial advantages and potential for flexible deployment in real-world resource-limited devices.

Ablation Study

(a)



(b)



(c)

Batch	Bit Width	ResNet19	VGG16	VGGSNN
batch=1	32w-32u	50.80 (-0.00%)	59.16 (-0.00%)	37.44 (-0.00%)
	1w-8u	1.75 (-96.56%)	1.96 (-96.69%)	1.31 (-96.50%)
	1w-4u	1.68 (-96.69%)	1.93 (-96.74%)	1.25 (-96.66%)
	1w-2u	1.65 (-96.75%)	1.91 (-96.77%)	1.21 (-96.77%)
batch=64	32w-32u	83.83 (-0.00%)	75.67 (-0.00%)	70.47 (-0.00%)
	1w-8u	10.0 (-88.07%)	6.09 (-91.95%)	9.57 (-86.42%)
	1w-4u	5.81 (-93.07%)	3.99 (-94.73%)	5.38 (-92.37%)
	1w-2u	3.72 (-95.56%)	2.94 (-96.11%)	3.28 (-95.35%)
batch=256	32w-32u	184.49 (-0.00%)	126.01 (-0.00%)	171.13 (-0.00%)
	1w-8u	35.17 (-80.94%)	18.67 (-85.18%)	34.74 (-79.70%)
	1w-4u	18.40 (-90.03%)	10.28 (-91.84%)	17.96 (-89.51%)
	1w-2u	10.01 (-94.57%)	6.09 (-95.17%)	9.57 (-94.41%)

- (a): The WS-DR method enhances the information content of both synaptic weights w and spike activities s in the Q-SNN baseline.
- (b): Q-SNN with WS-DR achieves comparable accuracy to 32-bit SNN.
- (c): Q-SNN has maximized the energy efficiency of SNNs by quantizing both synaptic weights and membrane potentials.

Conclusion

- We introduce a novel SNN architecture, called Q-SNN, designed for efficient hardware implementation and low energy consumption. Q-SNN achieves this goal by employing the quantization technique on both synaptic weights and membrane potentials.
- We analyze how to enhance Q-SNN's performance from the information entropy theory and propose a novel Weight-Spike Dual Regulation (WS-DR) method to maximize the information content in Q-SNNs.
- Extensive experimental demonstrate that our method achieves state-of-the-art results in terms of both efficiency and performance, underscoring its capability to boost the development of edge computing.