Ternary Quantization of Spiking Neural Networks

Yu-Lun Wu, Rui-Rui Tan, Shu-Hao Zhang, Shuang Liang, Zi-Ang Liu, Zhao Wang, Shao-Qun Zhang

Frontiers of Computer Science, DOI: 10.1007/s11704-025-51519-1

Problems & Ideas

- Problems with existing lightweight algorithms of Spiking Neural Networks (SNNs):
  - Pursuing only memory compression and inference acceleration leads to precision dropping.
  - The variation in uncertainty induced by quantization has not been sufficiently considered, and targeted optimization is absent.
- Ideas: A ternary quantization method comprising quantization training methods and a regularization term for enhancing both accuracy improvement and uncertainty reduction.

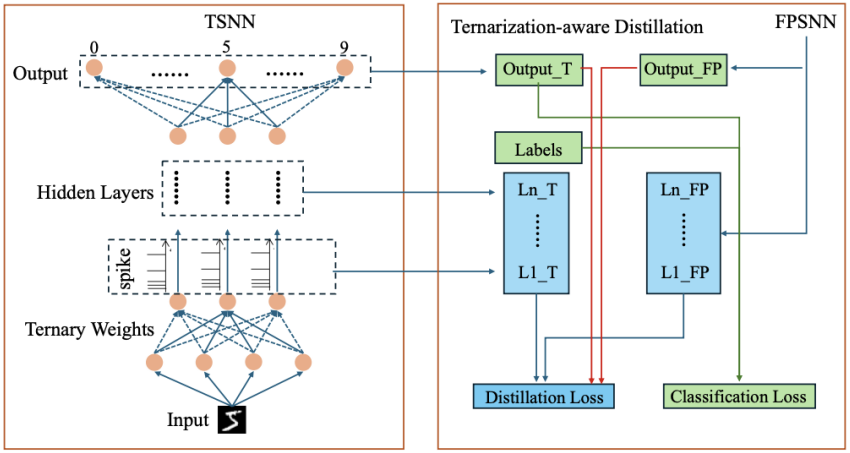

Fig. 1 Workflow of the indirect method using ternarization-aware distillation.

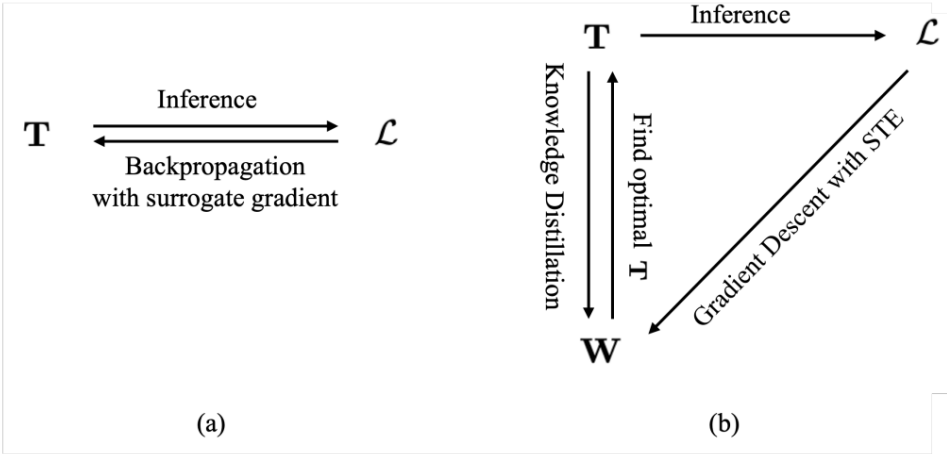

Fig. 2 Illustrations of our proposed (a) direct and (b) indirect methods.

Main Contributions

- Contributions:
  - A ternary quantization method achieves extremely low-bit quantization of SNNs while maintaining performance and is compatible with a variety of SNN architectures.
  - A method for quantifying the uncertainty of ternary SNNs and achieves uncertainty reduction by incorporating a novel regularization term.

Datasets	Algorithms	Configurations	Quantization Part	Quantization Training	Accuracy	Sparsity	Uncertainty	Storage	Energy	FLOPs
MNIST	SNN-Dropconnect	FC	FC	None	98.25%	0.00%	0.062	3.5MB	3.10×10^{10}pJ	1.79M
			FC_q^1	Indirect Method	97.68%±0.11%[2]	53.77%±0.07%	0.052	0.22MB	$9.20 \times 10^{8} \pm 6.44 \times 10^{5}$pJ	0.41±0.0003M
				Direct Method	97.43%±0.36%	52.21%±0.50%	0.051		$9.51 \times 10^{8} \pm 4.76 \times 10^{6}$pJ	0.43±0.0022M
	SNN-Dropout	FC	FC	None	96.87%	0.00%	0.063	3.5MB	3.10×10^{10}pJ	1.76M
			FC_q	Indirect Method	93.18%±0.56%	53.87%±2.43%	0.052	0.22MB	$9.18 \times 10^{8} \pm 0.22 \times 10^{8}$pJ	0.44±0.01M
				Direct Method	93.00%±0.96%	53.96%±2.34%	0.053		$9.16 \times 10^{8} \pm 0.21 \times 10^{8}$pJ	0.43±0.01M
	SNN-Slayer	FC	FC	None	95.37%	0.00%	0.064	4.6MB	7.62×10^{10}pJ	2.38M
			FC_q	Indirect Method	95.00%±0.74%	51.78%±1.22%	0.056	0.29MB	$2.36 \times 10^{9} \pm 2.88 \times 10^{7}$pJ	0.55±0.007M
				Direct Method	93.08%±0.83%	50.35%±0.78%	0.054		$2.43 \times 10^{9} \pm 1.90 \times 10^{7}$pJ	0.56±0.004M
	SNN-PLIF	2Conv + 2FC	2Conv + 2FC	None	99.00%	0.00%	0.039	51MB	8.47×10^{11}pJ	85.70M
			$2Conv_q$ + 2FC	Indirect Method	98.94%±0.14%	0.67%±0.02%	0.004	50MB	$2.93 \times 10^{11} \pm 5.86 \times 10^{7}$pJ	55.52±0.011M
				Direct Method	99.52%±0.09%	0.65%±0.03%	0.004		$2.93 \times 10^{11} \pm 8.79 \times 10^{7}$pJ	55.54±0.017M
			2Conv + $2FC_q$	Indirect Method	99.52%±0.21%	61.00%±1.23%	0.012	3.7MB	$2.36 \times 10^{11} \pm 2.90 \times 10^{9}$pJ	29.92±0.36M
				Direct Method	99.42%±0.09%	58.99%±0.03%	0.012		$2.48 \times 10^{11} \pm 7.44 \times 10^{7}$pJ	29.81±0.009M
			$2Conv_q$ + $2FC_q$	Indirect Method	98.52%±0.28%	62.48%±2.01%	0.015	3.3MB	$1.99 \times 10^{10} \pm 4.00 \times 10^{8}$pJ	16.94±0.34M
				Direct Method	98.43%±0.43%	60.00%±0.27%	0.016		$2.12 \times 10^{10} \pm 5.72 \times 10^{7}$pJ	17.04±0.05M

The experimental results demonstrate that the proposed method is compatible with various SNN architectures and achieves outstanding performance in terms of accuracy, energy consumption, inference efficiency, and uncertainty.