SUB-MILLIWATT WAKE-UP SYSTEMS USING QUANTIZED SPIKING NEURAL NETWORKS

Vasanth Ponukumati[1], Rohith Marrapu[2], Srihitha Batchu[3]

[123]*School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore, Tamil Nadu, India*

Abstract

Wake-up ultra-low-power intelligence is an emerging critical need in Internet of Things (IoT) and edge computing systems which must be always-on in severe energy constraining applications. This work presents a quantized spiking neural network (SNN) which is designed for efficient wake-up inference on general purpose hardware, for applications such as smart vision and industrial sensing. Taking advantage of the event-driven and asynchronous characteristic of SNNs, we use snnTorch and Tonic to implement a 500 neurons Leaky Integrate-and-Fire (LIF) network which is trained on the NMNIST dataset. The model shows that time-encoded representations based on spikes can be useful to capture and classify dynamic visual patterns. To overcome the limitations of the hardware implementation, 4-bit post-training quantization (PTQ) strategy is used to reduce the synaptic weights without losing the dynamics of the neurons. The percentile based clipping method is implemented to minimize the quantization induced distortions. The quantized model has an accuracy of 96.87% on the test set with negligible degradation compared to the full precision baseline while the model size and computational efficiency are dramatically improved. Experimental evaluation shows that the quantized model produces less inference latency than the full-precision counterpart, which suggests that the quantized model has a better efficiency and could save more energy. These results indicate that by the combination of SNNs and low bit quantization, efficient always-on inference on commercial hardwares without any special neuromorphic systems can be achieved. Overall, this work demonstrates the feasibility to deploy quantized SNN-based wake-up systems in general computing platforms, which creates the path for scalable and energy-efficient edge intelligence.

Keywords: Spiking Neural Networks, Leaky Integrate-and-Fire, Post-Training Quantization, Wake-Up Intelligence, Internet of Things, Edge Computing, NMNIST, snnTorch, Neuromorphic Computing, Low-Power Inference

References

[1]. J. K. Eshraghian et al., Proc. IEEE, 111(6), 623–652, 2023.

[2]. K. Yamazaki et al., Brain Sci., 12(7), 863, 2022.

[3]. S. Sanaullah et al., Front. Comput. Neurosci., 17, 1215824, 2023.

[4]. Y. Guo, X. Huang, and Z. Ma, Front. Neurosci., 17, 1209795, 2023.

[5]. A. Castagnetti, A. Pegatoquet, and B. Miramond, Front. Neurosci., 17, 1154241, 2023.

[6]. W. Wei et al., arXiv:2406.13672, 2024.

[7]. G. K. Cohen et al., Front. Neurosci., 10, 184, 2016.

[8]. M. Davies et al., IEEE Micro, 38(1), 82–99, 2018.

[9]. *S. Deng et al., IEEE Internet Things J., 7(8), 7457–7469, 2020.*

[10]. W. Maass, Neural Netw., 10(9), 1659–1671, 1997.