QAD-SNN: Quantization-aware Distillation on Spiking Neural Network

Donghyun Lee[1], Guoqi Li[2], Hongsik Jeong[1,3]*, Dong-Hyeok Lim[1]*
[1]Department of Materials Science and Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[2]Institute of Automation, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences
[3]Graduate School of Semiconductor Materials and Devices Engineering, Ulsan National Institute of Science and Technology, Ulsan, 44919, Republic of Korea
deango@unist.ac.kr, guoqi.li@ia.ac.cn, *jeonghs1@unist.ac.kr *dhlim@unist.ac.kr

Abstract

Although spiking neural networks (SNNs) are considered a promising and efficient structure for implementation in neuromorphic hardware, there are several obstacles to practical usage due to the characteristics of spike-driven computation and spatio-temporal dynamics, which are tough to be performed on current computation devices such as graphic processing unit (GPU). Correspondingly, there is still a need for low memory and computation costs for faster implementation. In this work, we propose a quantization-aware distillation spiking neural network (QAD-SNN), which can shrink the model size and decrease the memory and time cost. We verify our method using the CIFAR-10 dataset and obtain a higher performance than floating-point model when the bit-width of weights and batch normalization is set as 8, 6-bit. This work is the first attempt to apply a combination of quantization and knowledge distillation on deep SNNs, i.e., ResNet-20, and we believe that QAD-SNN provides a new perspective of compressing SNN for efficient training and inference.

Ⅰ. Introduction

A spiking neural network (SNN) is a conventional brain-inspired structure that inherently has aspects of event-driven signal processing and spatio-temporal information. Because the SNNs are easy to implement on some specially designed neuromorphic hardware [1], they are considered a promising model and the next generation of artificial intelligence (AI). There are main three streams of training SNNs, i.e., spike timing dependent plasticity (STDP) [2], artificial neural network (ANN)-to-SNN conversion [3] and direct training using backpropagation (BP) with surrogate gradient [4]. In this work, we focus on direct training of SNN, which is more efficient and adaptable than the other two training methods. Unfortunately, although there are diverse attempts to design a wider and deeper network structure of SNN, embedding and implementing it on neuromorphic hardware is still demanding because of memory and computation costs. Subsequently, to deal with the cost issues above, researchers have made ceaseless efforts to shrink a model size or the number of parameters. Putra, R. V. W. et al. [5] applied quantization on SNNs for decreasing memory cost. Furthermore, Kushawaha, R. K. et al. [6] distilled the knowledge of the large model to the small model of SNN for achieving better performance.

In this paper, we propose a quantization-aware distillation spiking neural network (QAD-SNN), which is the first time to report the combination of quantization and distillation on SNN.

Ⅱ. Methodology

Ⅱ-1. MS-ResNet

Our base SNN structure is grounded on MS-ResNet [4], which contains spatio-temporal back propagation (STBP) and threshold-dependent batch normalization (tdBN), and an iterative leaky integrate-and-fire (LIF) model. More specifically, STBP makes the SNN deeper and larger by utilizing BP with spatio-temporal information. Because of the additional temporal information and spike-based activation function, a specialized BN technique is developed, called tdBN. Finally, the iterative LIF model is an advanced version of the LIF model, which enables forward and backward propagation on both spatial and temporal axis.

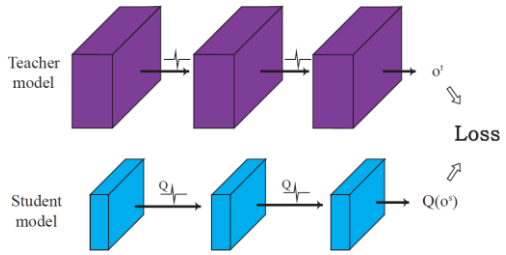

Figure 1. Overall training scheme of QAD-SNN

Ⅱ-2. Quantization

Our work adopts the low bit quantization, i.e., 8, 6, 5, 4 to the weights and parameters of tdBN during the training process. The fixed range of static quantization [7] is applied to all trainable parameters of weights and tdBN except for the last classifier, which is formulated in Equation (1).

$$Q(w, bits) = \frac{round(w \cdot 2^{bits-1})}{2^{bits-1}} \quad (1)$$

Here, $round(*)$ is a round function, and $bits$ is the target bit-width of input value w.

Ⅱ-3. Knowledge distillation

Knowledge distillation (KD) is a method of delivering knowledge from a large (teacher) neural network model to a small (student) one. There are two main approaches to distilling knowledge, logit distillation [6] and feature distillation [8]. In this work, we adopt the former method because a spike-based feature is less semantically interpretable than convolutional ANN-based one so far. We design the loss equation including cross-entropy and KD loss as follows

$$Loss = \lambda L_{CE} + (1-\lambda)L_{KD} \quad (2)$$

where L_{CE} is a typical cross-entropy loss, L_{KD} is KD loss, and λ is a weighting parameter. Also, L_{KD} can be computed like the Equation (3),

$$L_{KD} = \tau^2 D_{KL}(\sigma(Q(o^s);T=\tau)||\sigma(o^t;T=\tau) \quad (3)$$

where τ is a temperature, D_{KL} is Kullback-Leibler divergence which computes the distance between two different distributions, $Q(o^s)$ is the output logits of quantized student model and o^t is the output logits of teacher model, respectively. Our holistic training process is illustrated in Figure 1.

Ⅲ. Experiments

To verify our proposed method, we experimented with a classification task with the CIFAR10 dataset. We set the quantization bit-width of weights and tdBN as from 8 to 4. We compared only-quantized SNN (Q-SNN) and QAD-SNN according to the applied bit-width. We transferred knowledge using ResNet-32 to ResNet-20 which is the student model. For fair comparison, Q-SNN and QAD-SNN were trained from scratch based on Pytorch framework. The training accuracy of the student model is depicted in Figure 2.

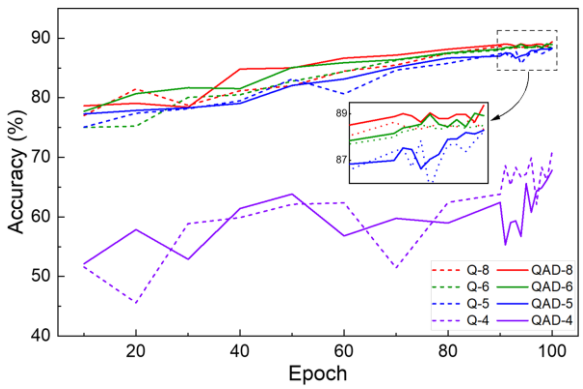

Figure 2. Accuracy curves of Q-SNN and QAD-SNN

Table 1 shows the results of our experiments and the quantized models with 8, 6, 5-bit and even extreme low bit cases, i.e., 4-bit. The accuracy of the base floating-point (FP) student model is 89.01 %, which is lower than the results of 8 and 6-bit QAD-SNN.

Furthermore, we investigated the number of spikes on the last CNN layer to inspect the effectiveness of distillation. According to the right side of Table 1, the overall number of spikes is increased, which indicates the validity of the KD with suitable temperature parameter of L_{KD}. However, in the case of extremely low bit, i.e., 4-bit, the number of spikes on Q-SNN is much lower than the other cases, resulting in accuracy degradation possibly due to low representation power. Correspondingly, although the spikes are increased after the distillation, the transferring of knowledge significantly ruins the performance.

Bit	Accuracy (%) (Q-SNN / QAD-SNN)	Number of spikes (10^5) (Q-SNN / QAD-SNN)
8	88.70 / 89.37 (+ 0.67)	7.80 / 8.93
6	88.52 / 89.03 (+ 0.51)	7.79 / 8.94
5	88.23 / 88.30 (+ 0.07)	7.60 / 8.62
4	71.00 / 67.81 (- 2.19)	4.56 / 7.52

Table 1. The performance comparison between distilled and non-distilled quantized SNN models.

Ⅳ. Conclusion

In this work, we propose QAD-SNN, which combines quantization with distillation for faster inference and efficient memory saving. We apply 8, 6, 5, and 4-bit to the weights and tdBN, which can decrease the memory and computation cost. According to the experiment using CIFAR-10, our proposed model gains higher performance than only the quantized model or even FP model. However, the efficient low bit (≤4) quantization-aware training is still challenging. As a solution, temporal information during distillation will be comprehensively investigated.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (grant No. 2021R1C1C2012077) and (grant No. 2022M317A207909811).

Reference

[1] J. Pei *et al.*, "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106-111, Aug. 2019.

[2] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 9, no. AUGUST, Aug. 2015.

[3] P. U. Diehl *et al.*, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *Proceedings of the International Joint Conference on Neural Networks*, Sep. 2015, vol. 2015-September.

[4] Y. Hu *et al.*, "Advancing Residual Learning towards Powerful Deep Spiking Neural Networks," *arXiv*, Dec. 2021.

[5] R. V. W. Putra and M. Shafique, "Q-SpiNN: A Framework for Quantizing Spiking Neural Networks," *arXiv*, Jul. 2021.

[6] R. K. Kushawaha *et al.*, "Distilling Spikes: Knowledge Distillation in Spiking Neural Networks," *arXiv*, May 2020.

[7] D. Lee, D. Wang *et al.*, "QTTNet: Quantized tensor train neural networks for 3D object and video recognition," *Neural Networks*, vol. 141, pp. 420-432, Sep. 2021.

[8] B. Heo *et al.*, "A Comprehensive Overhaul of Feature Distillation," *arXiv*, Apr. 2019.