2022 IEEE International Conference on Acoustics, Speech and Signal Processing

Optimizing the Consumption of Spiking Neural Networks with Activity Regularization

#2533

Simon Narduzzi[1,2] Siavash A. Bigdeli[1] Shih-Chii Liu[2] L. Andrea Dunbar[1]

[1] *CSEM*, Neuchâtel, Switzerland [2] *Institute of Neuroinformatics*, University of Zurich and ETH Zurich, Zurich, Switzerland

Introduction

Today, deep learning models are regularly deployed at the edge, allowing local real-time decision-making, efficient pre-processing, and privacy-preserving applications. Optimizations have been developed in the past few years to allow the deployment of these networks within restricted resource environments.

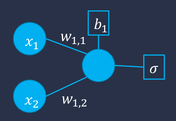
Quantization, pruning, and distillation are some of them, which are applied either during the training or post-training of the neural network. Specialized hardware accelerators and neuromorphic hardware platforms also target ultra-low power applications using sparse processing. In our experiment, we explore the sparsification of artificial neural networks (ANN) using different activity regularizers and their effect on the post-training conversion of spiking neural networks (SNN).

Spiking neural networks

In the brain, most neurons use events (called spikes) to encode and transmit information. Spiking neural networks (SNNs) are mathematical models that imitate this behavior. The sparse nature of spikes makes SNNs suitable for low-power inference. However, their discrete nature makes them hard to train with conventional backpropagation techniques, and their accuracy is not competitive with ANN performance. Therefore, methods have been developed to create SNNs from pre-trained ANN models.

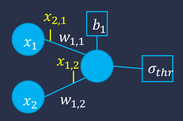
Artificial Neural Network

$$z_i = \sigma\left(\sum_{j=1}^{N} W_{ij}x_j + b_i\right)$$



Spiking Neural Network

$$z_i = \sigma_{thr}\left(\sum_{j=1}^{N} W_{ij}x_{t,j} + b_i\right)$$



Computational cost of ANN and SNN

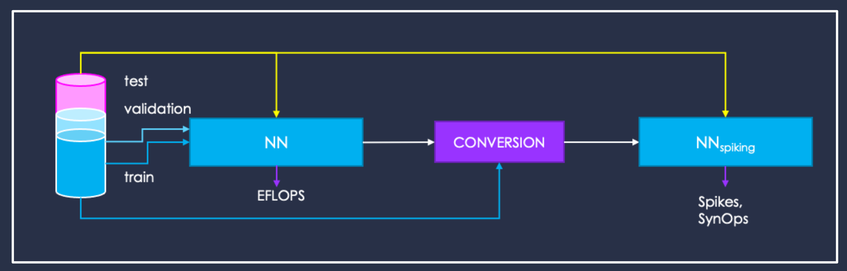
We assess the efficacy of the sparsity by measuring the reduction in computational cost on an ANN and the equivalent SNN. EFLOPS describes the ANN computational cost by counting operations not involving zero-values, and SynOps [1] describes the SNN computational cost.

$$EFLOPS = \sum_{l=1}^{L} \phi(W_l) \times \phi(A_{l-1}) + \phi(B_l) \qquad SynOps = \sum_{t=1}^{T}\sum_{l=1}^{L} f_{out,l} \times s_l(t)$$

Regularization in ANN and SNNs

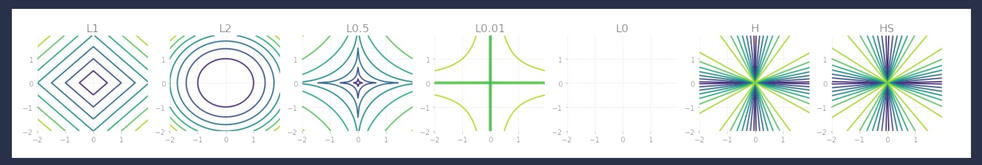
Regularization is a technique to constrain the parameters or the activations of a neural network. Sparsification can be obtained using weight regularization and activity regularization. In our work, we sparsify the activity of the SNN by regularizing a pre-trained ANN and then converting the ANN to a spiking form.

We train an ANN and monitor its convergence using the validation set. We then convert [2] it to an SNN using the training set for the calibration. We then test both the ANN and SNN to obtain the computational cost.



Training and conversion pipeline used in our experiments.

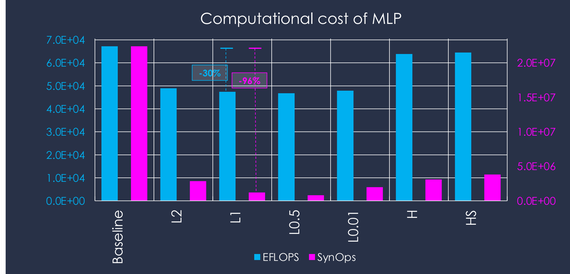
L_0-norm exactly counts the number of non-zero elements in a vector. This norm has no gradient, therefore we use surrogates $L_{0.5}$ and $L_{0.01}$ as approximations to L_0-norm. We compare with other regularizers such as L_2-norm, L_1-norm, Hoyer (H) and the squared version of Hoyer (H_S).



Loss landscape for regularization methods used in our experiments.

Results

The regularization was applied on the activation maps of two architectures: a multilayer perceptron and LeNet-5, on both MNIST and CIFAR-10 datasets. On MNIST, the computational cost of the SNN can be reduced by 96% and 94% on the MLP and LeNet-5 respectively, with no accuracy loss.



Conclusion

Activity regularization of ANNs is a simple way to reduce the number of SynOps in converted SNNs. Although surrogate L_p-norms perform well, better approximations of L_0-norm can be explored. Post-conversion fine-tuning of the SNNs and simultaneous regularization of weights and activations are other potential improvements that could lead to very efficient networks.

References

[1] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., ... & Modha, D. S. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197), 668-673.

[2] Rueckauer, B., Lungu, I. A., Hu, Y., Pfeiffer, M., & Liu, S. C. (2017). Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11, 682.

csem University of Zurich[UZH] ETH zürich

icassp 2022 Singapore