technische universität dortmund

computer science 12

Bachelor/Master Thesis

Architecture Search for Error Tolerant Binarized Neural Networks

Mikail Yayla
Prof. Dr. Jian-Jia Chen
Otto-Hahn Str. 16
Technische Universität Dortmund
Email: mikail.yayla@tu-dortmund.de
29.03.2021

Neural networks (NNs) are popular machine learning models used in many application cases such as image and speech recognition. However, due to resource constraints of low power embedded devices, the deployment of machine learning models on these types of devices is a challenge. In the literature, numerous approaches exist to optimize NNs for energy and speed. One method is quantization of NN parameters. The extreme case is using binarized NNs (BNNs), where the weights and activations are binarized [1]. This reduces memory size and improves execution time by trading a small amount of accuracy. However, to achieve high accuracy, still a high number of parameters needs to be accessed, which makes the memory the main bottleneck.

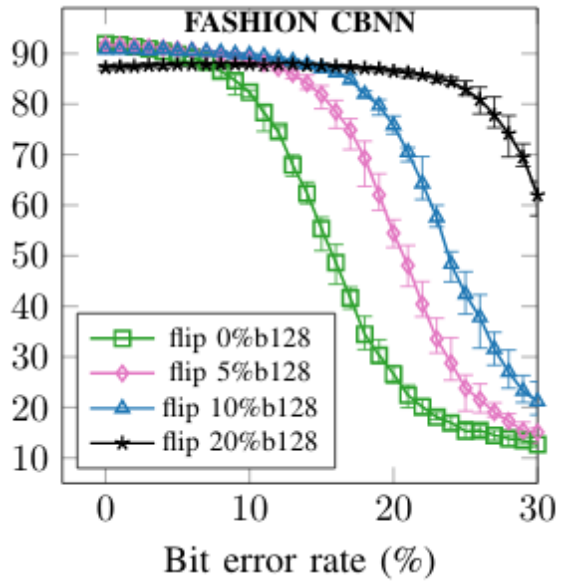

Figure 1: Accuracy over bit error tolerance plots for a BNN for the FashionMNIST dataset [2]

Another approach to decrease the energy efficiency and speed of NNs is approximate computing. In approximate computing, the result quality is traded to achieve resource efficiency. For example, for less energy consumption and faster speed, one can reduce the supply voltage and tune the timing parameters of DRAM or SRAM, at the cost of bit errors. Fortunately, BNNs can tolerate certain bit error rates without much accuracy degradation (see Figure 1). However, recent methods only study how to optimize the bit error tolerance of BNNs when the BNN model is given. Here, we want to find principled approaches, e.g. from neural architecture search (NAS) [3,4], to find BNN architectures such that the bit error tolerance is optimized, while accuracy and memory usage should be kept at a reasonable level.

In this thesis, students first should get familiar with the general principles of BNN training/inference and existing methods for bit error tolerance optimization. Then, the students should study the methods used in NAS and find out which methods could be suitable to derive highly bit error tolerant BNNs. A novel approach for finding bit error tolerant BNN architectures should be evaluated in the scope of this thesis.

The overall goal of this thesis is to find highly bit error tolerant BNN architectures by principles methods such as NAS. The BNNs should also have high accuracy and a reasonable model size, such that it is possible to deploy them on small mobile and embedded systems. Both empirical and theoretical methods are possible.

Other suggestions and related topics are also welcome. Please do not hesitate to make an appointment.

Required Skills:

- Basic knowledge of machine learning and neural networks
- Python and C++

Acquired Skills after the thesis:

- Knowledge of methods for enabling NNs for emerging unreliable devices
- Knowledge of PyTorch and CUDA extensions

[1] Hubara et al. "Binarized Neural Networks" http://arxiv.org/abs/1602.02505
[2] Buschjäger/Pfahler/Yayla et al. "Margin-Maximization in Binarized Neural Networks for Optimizing Bit Error Tolerance"
[3] Elsken et al. "Neural Architecture Search: A Survey" https://www.jmlr.org/papers/volume20/18-598/18-598.pdf
[4] Chen et al. "Binarized Neural Architecture Search" https://arxiv.org/pdf/1911.10862.pdf