technische universität dortmund

computer science 12

Bachelor/Master Thesis

Bit Error Tolerance Optimization of Object Detection 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

Object detection neural networks (NNs) aim to locate objects in images and classify them correctly. They can be applied in numerous fields, e.g. in autonomous driving, robot vision, surveillance, etc. However, these types of NNs rely on a massive number of parameters to achieve high accuracy and need to perform multiply-accumulate (MAC) operations with them. Therefore, deploying these type of models efficiently with low latency is a challenge. In the literature, numerous approaches exist to optimize NNs for speed and energy. One method is quantization of NN parameters, e.g. to 2, 4, or 8 bit, in contrast to using 32 bit floating-point values. 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 accuracy. Binarizing weights and activations for object detection NNs has been evaluated in the work in [3].

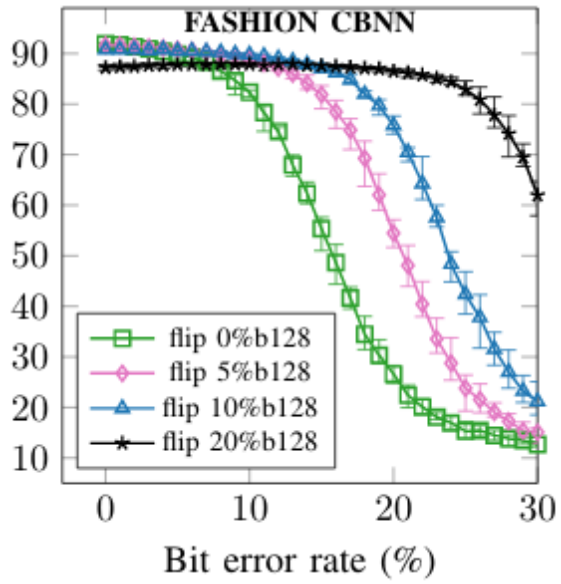

Figure 1: Accuracy over bit error tolerance plots for a BNN for the FashionMNIST dataset [2]

Another approach to increase execution time of NNs is approximate computing. In approximate computing, the result quality is traded for resource efficiency. For example, for faster execution time, one can aggressively tune the timing parameters of the memory at the cost of bit errors (see for example [4]). Fortunately, NNs, especially when quantized or binarized, can tolerate certain bit error rates without much accuracy degradation (see Figure 1), when bit error tolerance optimization is employed. However, novel bit error tolerance optimization methods have mostly been applied to multiclass classifiation problems, not for object detection problems.

In this thesis, students should first get familiar with the general principles of object detection NNs, quantization/binarization techniques, NN training/inference and existing methods for bit error tolerance optimization. Then, the student should evaluate the bit error tolerance of object detection NNs, such as YOLO (binarized and 2,4,8 bit quantized). After this, following methods can be modified or combined to optimize bit error tolerance: Bit flip injection [4], layer-wise methods [5], and methods based on margin-maximization [2]. Any other novel approach can also be developed in the scope of this thesis.

The overall goal of this thesis is to find methods to optimize object detection NNs for bit error tolerance, while the method should be applicable to binarized and quantized (2, 4, 8 bit) NNs. The NNs should also have high accuracy and a reasonable model size, such that it is possible to deploy them on small mobile and embedded systems.

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 object detection 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] Xu et al. "Training a Binary Weight Object Detector byKnowledge Transfer for Autonomous Driving" https://arxiv.org/pdf/1804.06332.pdf
[4] Koppula et al. "EDEN: Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM https://arxiv.org/abs/1910.05340
[5] Henwood et al. "Layerwise Noise Maximisation to Train Low-Energy Deep Neural Networks https://arxiv.org/pdf/1912.10764.pdf