technische universität dortmund

computer science 12

Bachelor/Master Thesis

Pruning Binarized Neural Network for Embedded Systems

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

Embedded systems with limited amounts of main memory and limited processing power became an interesting candidate for the execution of machine learning models over the last years. Executing machine learning models *on the edge* allows the realization of distributed systems with advanced machine learning features, reducing communication overheads, improving security, and aid in privacy. However, due to resource constraints of low power embedded devices (for example the device in Figure 1), the deployment of machine learning (ML) models is a challenge. Therefore it is crucial to design and find tailored, small ML models, which meet given constraints on model size and execution time.

Figure 1: TI MSP430 FR6989 with 128kb of FRAM [1]

Neural networks (NNs) are one popular machine learning model used in many application cases such as image and speech recognition. In the literature, numerous approaches to design small NN models exist. Many recent studies have focused on quantization of NN parameters. The extreme case is binarized NNs (BNNs), where the weights and activations are binarized [2]. 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 is needed.
The size of BNNs can be further reduced, e.g. by *pruning*, where a full size model is trained and pruned afterwards. For this process, single neurons have to be selected, which can be removed from the model, without degrading the model accuracy significantly. Margin maximization is a recently elaborated method, which specially targets binarized neural networks (BNNs) and aims to modify the training process of the network in such a way, that output of each single neuron can tolerate more erroneous or missing inputs of the input neurons. Margin maximization therefore could be considered as an additional step to optimize and improve the pruning of binarized neural networks.

In this thesis, students first should get familiar with the general implementation of binarized neural networks and their according use cases. Students should understand existing pruning methods and apply them on a given set of applications. A novel approach for pruning should be developed by the student.

The overall goal of this thesis is to prune BNNs to a specified size, by keeping the accuracy as high as possible. For practical evaluation, the pruned model can be executed on real embedded systems and the execution can be analyzed with regards to power consumption and execution time.

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 ML on the edge
- Knowledge of PyTorch and CUDA extensions

[1] Choi, Ja Moon. "Ferroelectric RAM device." U.S. Patent No. 6,044,008. 28 Mar. 2000.
[2] Hubara et al. "Binarized Neural Networks" http://arxiv.org/abs/1602.02505