Accelerator-aware Training Proposal

Supervisor: Vahid Geraeinejad (vahidg@kth.se)
Examiner: Masoumeh Ebrahimi (mebr@kth.se)

February 18, 2022

1 Introduction

Deep Neural Networks (DNNs) have gained increasing attention due to their wide range of applications such as autonomous driving, natural language processing, healthcare, and finance. Executing DNNs in real-time and with high input bandwidth consist of numerous floating-point operations, demanding a powerful microprocessor or graphics processing unit (GPU). These processors considerably consume large power, which is not suitable for an embedded device. In recent years, utilizing accelerators with customized hardware has been the industry's solution to reduce the power and inference time of DNNs on embedded devices (1). Accelerators exploit certain characteristics of DNNs such as the sparsity, repetition of weights, tolerance for quantization, and parallelization of computations to improve DNN execution performance. Almost all accelerators offer a customized design for the execution of neural networks, after the training phase.

2 Accelerator-aware Training

The aim of this proposal is to develop an accelerator-aware procedure to improve the performance of DNNs before the inference phase. In this method, the traditional training phase will be modified to search for optimized weights that fit the desired characteristics of specific accelerators. Improvements in quality/quantity of these features directly affect the performance. This could be achieved using any type of Neural Architecture Search (NAS) based on either Reinforcement Learning (RL) (2) or Genetic Algorithm (GA) (3). This technique is to be implemented alongside or after the initial training; therefore, it is capable of improving inference time, power efficiency, storage, etc. without imposing extra hardware overhead for the inference stage. However, the increased cost of the training phase and further optimizations should be considered properly.

References

[1] Yuhong Li, Cong Hao, Xiaofan Zhang, Xinheng Liu, Yao Chen, Jinjun Xiong, Wen-mei Hwu, and Deming Chen. Edd: Efficient differentiable dnn architecture and implementation co-search for embedded ai solutions. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.

[2] Arash Vahdat, Arun Mallya, Ming-Yu Liu, and Jan Kautz. Unas: Differentiable architecture search meets reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11266–11275, 2020.

[3] Zhichao Lu, Ian Whalen, Vishnu Boddeti, Yashesh Dhebar, Kalyanmoy Deb, Erik Goodman, and Wolfgang Banzhaf. Nsga-net: neural architecture search using multi-objective genetic algorithm. In *Proceedings of the genetic and evolutionary computation conference*, pages 419–427, 2019.