Training Neural Networks in LNS representation

Keywords : *Neural Network, Quantization Aware Training, Power Efficiency*

Context :

Neural nets are really good at some tasks, in particular image classification, image segmentation, natural language processing, etc. However, this comes at the cost of a high power consumption, even in applications that by nature could accomodate a slight degradation in accuracy (i.e. looking for cat pictures on the Internet). So the quest for much more power efficient neural networks has started some time ago, with mainly two objectives in mind : lowering the computational power required for a given task, and limiting the number of memory accesses to fetch the parameters (weights and to a lesser extend biases) and the intermediate results (activations). The idea is to have an at least two orders of magnitude gain in power efficiency.

In this context, there are two levers : searching for application specific neural network architectures, that would rival much more complex general purpose networks, and value representation using quantized (or highly quantized) values for the parameters and activations. Quantization simplifies the implementation of the mathematical multiplication and addition operations, and lowers the number of bits to be fetched from memory, which is a clear gain. Quantization is usually done in the linear domain, letting unchanged the nature of the operations.

We recently proposed the use of the Logarithmic Number System (LNS) to represent weights, biases and activations so that the multiplication can be replaced by an addition, which has a big hardware advantage. Unfortunately, adding is complex in LNS, so we have to go back in the linear domain for that, and need exponentiation. However, when working on a relatively small number of bits (less that 8), exponentiation can be tabulated and is not too costly.

Objective :

For this internship, we want to study the use of the Logarithmic Number System in the context of Quantization Aware Training (QAT). Indeed, our current approach is using Post-Training Quantization (PTQ), which leads to an accuracy that might not be as good as it could, so we want to evaluate to which extend QAT could improve accuracy, at the cost of either a full QAT approach, or a retraining after PTQ, which could improve greatly the power efficiency of training.

The main steps of the internship are as follows :

— get acquainted with the LNS representation, its transformation to linear and vice-versa, etc,
— perform PTQ on a few representative networks, starting from simple MLP to more complex CNN, e.g. RESNET or such,
— evaluate PTQ, possibly propose different PTQ approaches,
— propose a QAT approach using the LNS representation,
— evaluate the QAT approach on the previously defined networks, and iterate if necessary,
— write a report/scientific paper presenting the different approaches and their analysis.

Available resources :

— own research work on inference using LNS (`https://hal.inria.fr/hal-03684585/document`),
— publicly available NN inference and training resources,
— qkeras library for QAT (`https://github.com/google/qkeras`).

Expected competencies :

— theory of neural network inference and training,
— practical use of Tensorflow/Keras,
— knowledge of the training phase of TF/Keras/QKeras is a plus.

Location and Contact :

The internship takes place at the TIMA Lab, in the center of Grenoble in the historical premises of Grenoble INP, 46, avenue Félix Viallet.

Frédéric Pétrot : frederic.petrot@univ-grenoble-alpes.fr
Olivier Muller : olivier.muller@univ-grenoble-alpes.fr