Extremely low-bit Representation for DNN Inference

Project Proposal

Students

Chris Shakkour, 208157826, christian.s@campus.technion.ac.il

Nadi Najjar, 211610704, nadi.najjar@campus.technion.ac.il

Contents

Abstract

Approximate computing “fits like a glove” to the new environment introduced by deep neural networks (DNNs). First, DNNs are usually trained and used for error-tolerant applications; and second, DNNs exhibit inherent algorithmic resiliency to some inaccuracies in their intermediate results, for example, their activations and weights can be pruned and quantized with minor degradation in accuracy [2]. Motivated by these observations, we would like to check model resiliency of a decreased width of partial sum representation, which is usually set to 32 bits, but with different approach – instead of using integer representation we would like to explore very low floating-point representations.

Introduction

In this project we will be taking multiple networks and different datasets to be used for training in BF16, after the model is fully trained and matured our goal will be to try to reduce the number of bits needed to hold the partial sums in a forward pass while in inference only, this will be done by writing our own methods in python that calculates the partial sums, multiply and accumulate functions that are driven by a parameter that indicates the number of HW bits allocated for the computation result, with this said when choosing the number of bits one must also choose the bit separation between the mantissa and the exponent fields in a floating point number, or more precisely the position of the floating point which has a vey high effect on the precision.

We have high hopes for this experiment since the range of numbers allocated for a 32bit partial sum is very big hence we believe there might be a good number of bits that can be removed with minor degradation in the number’s resolution. What’s promising here is that other methods like pruning and quantization have shown to be very successful in terms of reducing resources while maintaining the accuracy and loss rates. We expect to see a good reduction in MAC bits and memory footprint that holds the reduced activation values.

Goals

- Explore the number of actual partial sum bits required for several precision-reduced CNN models in different granularities.
- Evaluate impact on power and area compared to similar performance integer-quantized networks.

Models and Datasets

- Lenet5-cifar10
- Alexnet-cifar10
- Alexnet-cifar100
- Resnet18-cifar100
- Resnet18-imagenet

References

No references at the moment.