Masterarbeit – Zeit-Effizientes Training von Convolutional Neural Networks

Jessica Bühler

16. Oktober 2019

Übersicht

1 Einführung

2 Fixed Point arithmetic

3 Accelerated CNN Training Through Approximation

4 Kernel rescaling

Fragestellung

Ein Traininsdurchlauf von CNNs kann sehr zeitaufwendig sein. Wird dieser Prozess dann mehrfach durchlaufen, in dem verschiedene Hyperparameter / Startwerte probiert werden kann dieser Prozess sehr schnell zeitlich explodieren. Wie lässt sich hier Zeit einsparen?

Übersicht

1 Einführung

2 Fixed Point arithmetic

3 Accelerated CNN Training Through Approximation

4 Kernel rescaling

Deep Learning with Limited Numerical Precision

Training of large-scale deep neural networks is often constrained by the available compu- tational resources. We study the effect of lim- ited precision data representation and com- putation on neural network training. Within the context of low-precision fixed-point com- putations, we observe the rounding scheme to play a crucial role in determining the network's behavior during training. Our re- sults show that deep networks can be trained using only 16-bit wide fixed-point number representation when using stochastic rounding, and incur little to no degradation in the classification accuracy. We also demonstrate an energy-efficient hardware accelerator that implements low-precision fixed-point arith- metic with stochastic rounding.

Deep Learning with Limited Numerical Precision

TODO:

- Ermittle eine Formel, die berechnet wieviel Zeit abhängig von den genutzten Kommazahlen im Rechner pro Epoche gebraucht wird
- Verifiziere diese Zeit mit einem Experiment

Besonderheiten:

- Wenn dies so gut funktioniert wie im Paper beschrieben, so kann zumindest für die ersten Epochen standardmässig mit geringerer Präzision im Festkommaformat gerechnet werden.

Übersicht

1 Einführung

2 Fixed Point arithmetic

3 Accelerated CNN Training Through Approximation

4 Kernel rescaling

Training deep convolutional neural networks such as VGG and ResNet by gradient descent is an expensive exercise re- quiring specialized hardware such as GPUs. Recent works have examined the possibility of approximating the gradient computation while maintaining the same convergence properties. While promising, the approximations only work on rela- tively small datasets such as MNIST. They also fail to achieve real wall-clock speedups due to lack of efficient GPU imple- mentations of the proposed approximation methods. In this work, we explore three alternative methods to approximate gradients, with an efficient GPU kernel implementation for one of them. We achieve wall-clock speedup with ResNet-20 and VGG-19 on the CIFAR-10 dataset upwards of 7 percent, with a minimal loss in validation accuracy.

Faster Neural Network Training with Approximate Tensor Operations

We propose a novel technique for faster Neural Network (NN) training by systematically approximating all the constituent matrix multiplications and convolutions. This approach is complementary to other approximation techniques, requires no changes to the dimensions of the network layers, hence compatible with existing training frameworks. We first analyze the applicability of the existing methods for approximating matrix multiplication to NN training, and extend the most suitable column-row sampling algorithm to approximating multi-channel convolutions. We apply approximate tensor operations to training MLP, CNN and LSTM network architectures on MNIST, CIFAR-100 and Penn Tree Bank datasets and demonstrate 30%-80% reduction in the amount of computations while maintaining little or no impact on the test accuracy. Our promising results encourage further study of general methods for approximating tensor operations and their application to NN training.

Übersicht

1 Einführung

2 Fixed Point arithmetic

3 Accelerated CNN Training Through Approximation

4 Kernel rescaling