Week 4

Kexin Fan

- Optimization problems and methods for quantization neural network

Paper: Deep learning with limited numerical precision

P: Limited precision data representation and computation on neural network training.

M: 1) The deep neural networks can be trained using low-precision fixed-point arithmetic, provided the stochastic rounding scheme is applied while operating on fixed-point numbers.

2) Using hardware accelerator. A hardware accelerator design, prototyped on an FPGA, that achieves high throughput and low power using a large number of fixed-point arithmetic units, a data flow architecture, and compact stochastic rounding modules.

P: Handling low-precision weights is difficult and motivates interest in new training methods.

M: 1) Using a rounding procedure yields poor results when weights are represented using a small number of bits.

2) Classical stochastic rounding method.

3) Using schemes that combine full-precision floating-point weights with discrete rounding procedures.

Paper: Training quantized nets: a deeper understanding

P: Different quantized optimization routines can be defined by selecting different quantizers, and also by selecting when quantization happens during optimization.

M: 1) Deterministic rounding

2) Stochastic rounding

3) BinaryConnect

Paper: Training and inference with integers in deep neural networks

P: Discretizing training and inference simultaneously.

M: 1) Weights, activations, gradients and errors among layers are shifted and linearly constrained to low-bitwidth integers.

Extending the original definition of errors to multi-layer: error is the gradient of activation for the perspective of each convolution or fully-connected layer, while gradient particularly refers to the gradient accumulation of weight.

2) Replacing batch normalization by a constant scaling layer and simplify other components that are arduous for integer implementation.

Normalization layers like softmax and batch normalization are avoided or removed in some WAGE demonstrations. The way to quantize normalization hasn't been given.

- Read code of HWGQ

According to the paper: Deep learning with low precision by half-wave gaussian quantization

The half-wave Gaussian quantizer is proposed for forward approximation. And to improve the learning efficiency of quantized net-works, there's a design of forward and backward approximation functions for the ReLU. For backward approximation, there are three possibilities for a piece-wise function that provides a good approximation to the ReLU and to the HWGQ. There aren't much differences among these three kinds.