Accuracy Booster

Simla Burcu Harma, Ayan Chakraborty, Babak Falsafi, Martin Jaggi, Yunho Oh
EcoCloud, EPFL

How to make DNN training denser?
Training traditionally uses power-inefficient floating-point arithmetic for accuracy. HBFP brings fixed-point efficiency to training.

Goal: Training DNNs using 4-bit fixed-point arithmetic with FP32 accuracy

Moore’s Law is dying

- A key advantage of BFP is fixed exponent
- We can keep the exponent and vary the mantissa bits
- Epochs have varying precision requirements
- First/last layers of DNN models have a large impact on accuracy
- Accuracy Booster: HBFP6 only in the last epoch and first/last layers, HBFP4 for the rest

Mixed-Mantissa HBFP: Accuracy Booster

- Hardware benefits of HBFP4 while maintaining FP32 accuracies

Model sizes keep increasing!

Tensor Distributions: Wasserstein Distance

- Tensor distributions are much more distorted for HBFP4 compared to HBFP6
- HBFP6 is not sensitive to the block size, while HBFP4 is sensitive
- Wasserstein distances of first/last layers are higher than the other layers

Analyzing the Loss Landscapes

- Plot the landscape around the current position of the minimizer
- Dimensionality reduction by random directions with filter normalization

Loss landscapes provides information for the interplay between generalization & optimization!