Comparing Deep Learning Methods is Hard

Many Metrics
- Convergence rate
- Latency
- Cost
- Power
- Sample complexity
- Robustness
- Training time
- Accuracy

Many Optimizations
- Minibatch size
- Reduced precision
- Quantization
- Optimizer
- Learning rate
- Training procedure
- Data augmentation
- Hardware

...which interact in complex ways

Numerical precision
- 8-bit MNIST gives 3x speed up
- meProp on MNIST gives 3.1x speed up

Is that 9.3x combined? Does it work on ImageNet? Does it scale to multiple nodes?

End-to-end performance provides an objective means of normalizing across factors

A New Benchmark

Evaluate end-to-end performance given a pre-specified accuracy

Tasks
- Image Classification
- Question Answering

Metrics
- Training time
- Training cost
- Inference latency
- Inference cost

Datasets
- CIFAR10
- ImageNet
- SQuAD

Current Seed Entries

Image Classification (CIFAR10)
- ResNet56 and ResNet164 in TensorFlow and PyTorch
- Top-1 validation accuracy threshold of 94%

Image Classification (ImageNet)
- Implementations of ResNet50 and ResNet152 in TensorFlow, PyTorch, and MXNet
- Top-5 validation accuracy threshold of 93%

Question Answering (SQUAD)
- Implementations of BiDAF in TensorFlow and PyTorch
- Validation F1 score threshold of 0.75

Competition deadline: April 20, 2018

Submission Workflow

Train your amazing model!
Submit on https://github.com/stanford-futuredata/dawn-bench-entries
Revel in glory on http://dawn.cs.stanford.edu/benchmark/

Early Results

Motivation: Why is time-to-accuracy important?

Figure 1: A minibatch size of 256 reaches an accuracy of only 0.43% less than the maximum accuracy achieved in 1.9x less time

Figure 2: ResNet110 model with stochastic depth (SD), 4 Nvidia K80 GPUs on a single node (4xK80), and a minibatch size of 512 (BS=512). Cumulatively enabling each optimization reduces the time to 93% top-1 accuracy, but combined, the model does not converge (DNC) to the 94% threshold.