DAWNBench

An End-to-End Deep Learning Benchmark and Competition

Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia
http://dawn.cs.stanford.edu/benchmark/

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

Training procedure
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	Metrics	Datasets
Image Classification	Training time	CIFAR10 ImageNet
Question Answering	Training cost	SQuAD
	Inference latency	
	Inference cost	

+ Open submission for community involvement
+ Evolving definitions for up-to-date evaluation

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%**

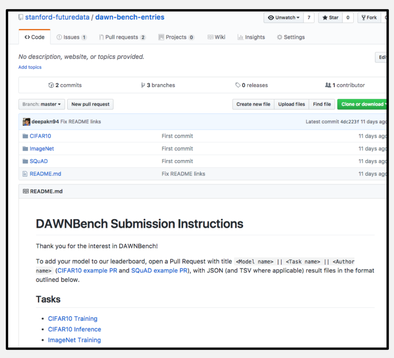
Question Answering (SQuAD)
- Implementations of BiDAF in TensorFlow and PyTorch
- Validation F1 score threshold of **0.75**

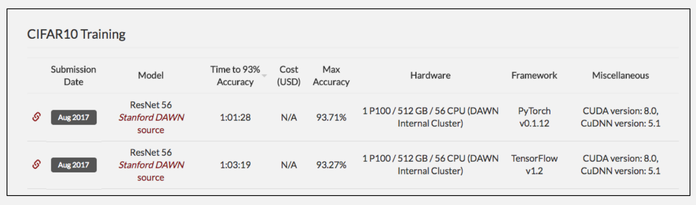
Competition deadline: **April 20, 2018**

Submission Workflow

Train your amazing model!

Submit on
https://github.com/stanford-futuredata/dawn-bench-entries





CIFAR10 Training

Submission Date	Model	Time to 93% Accuracy	Cost (USD)	Max Accuracy	Hardware	Framework	Miscellaneous
Aug 2017	ResNet 56 *Stanford DAWN* source	1:01:28	N/A	93.71%	1 P100 / 512 GB / 56 CPU (DAWN Internal Cluster)	PyTorch v0.1.12	CUDA version: 8.0, CuDNN version: 5.1
Aug 2017	ResNet 56 *Stanford DAWN* source	1:03:19	N/A	93.27%	1 P100 / 512 GB / 56 CPU (DAWN Internal Cluster)	TensorFlow v1.2	CUDA version: 8.0, CuDNN version: 5.1

Revel in glory on
http://dawn.cs.stanford.edu/benchmark/

Early Results

Motivation: Why is time-to-accuracy important?

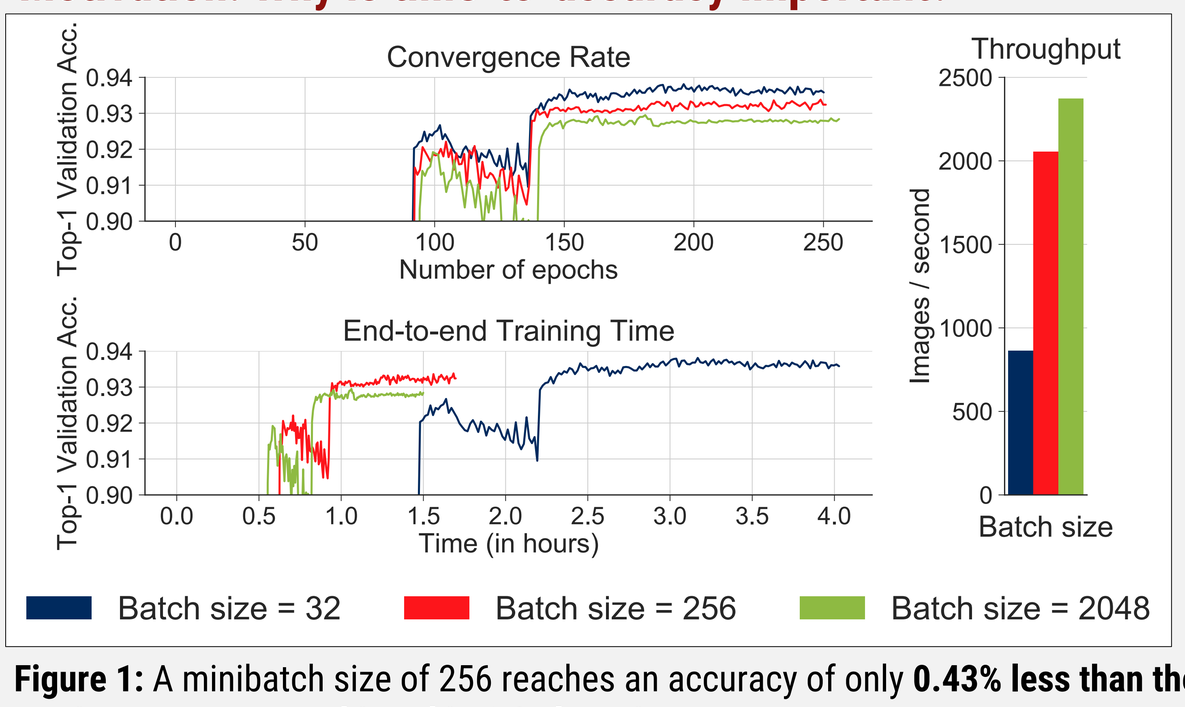


Figure 1: A minibatch size of 256 reaches an accuracy of only **0.43% less than the maximum accuracy achieved in 1.9x less time**

Optimizations interact in non-trivial ways

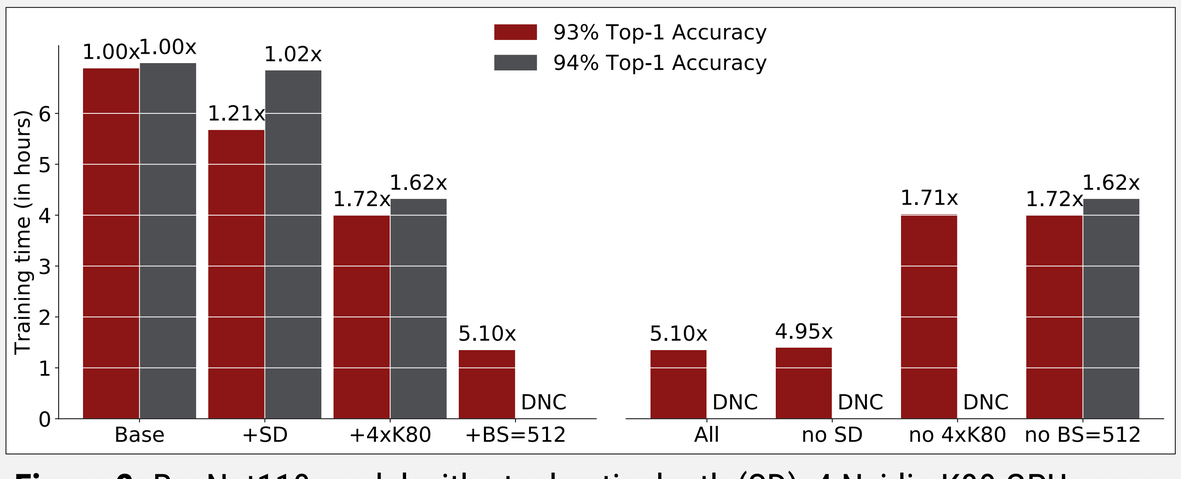


Figure 2: ResNet110 model with stochastic depth (SD), 4 Nvidia K80 GPUs on a single node (4xK80), and a minibatch size of 512 (BS=512). **Cumulatively enabling each optimization reduces the time to 93% top-1 accuracy, but combined, the model does not converge (DNC) to the 94% threshold.**