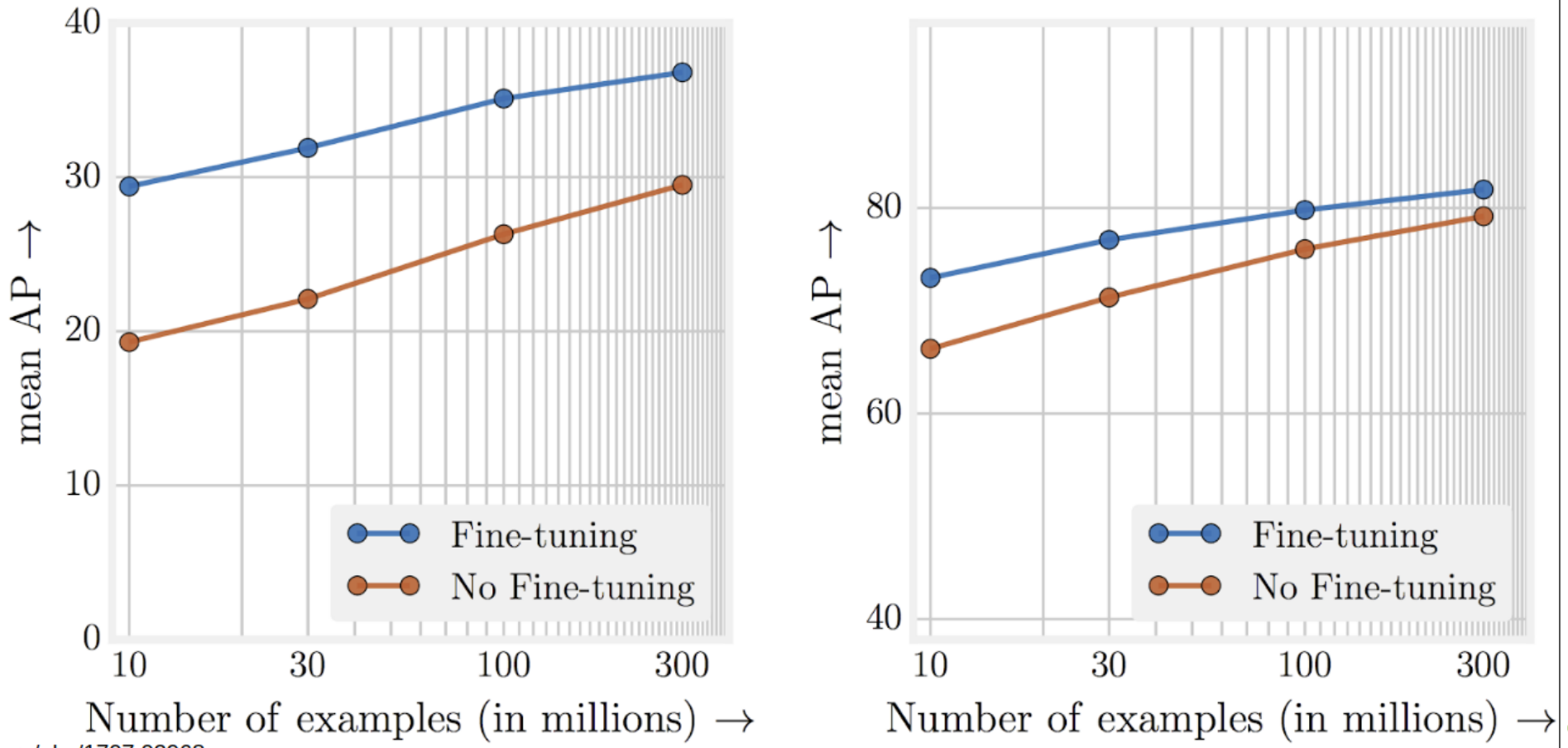
Big Data for AI

Thomas Breuel
Alex Aizman
NVIDIA

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

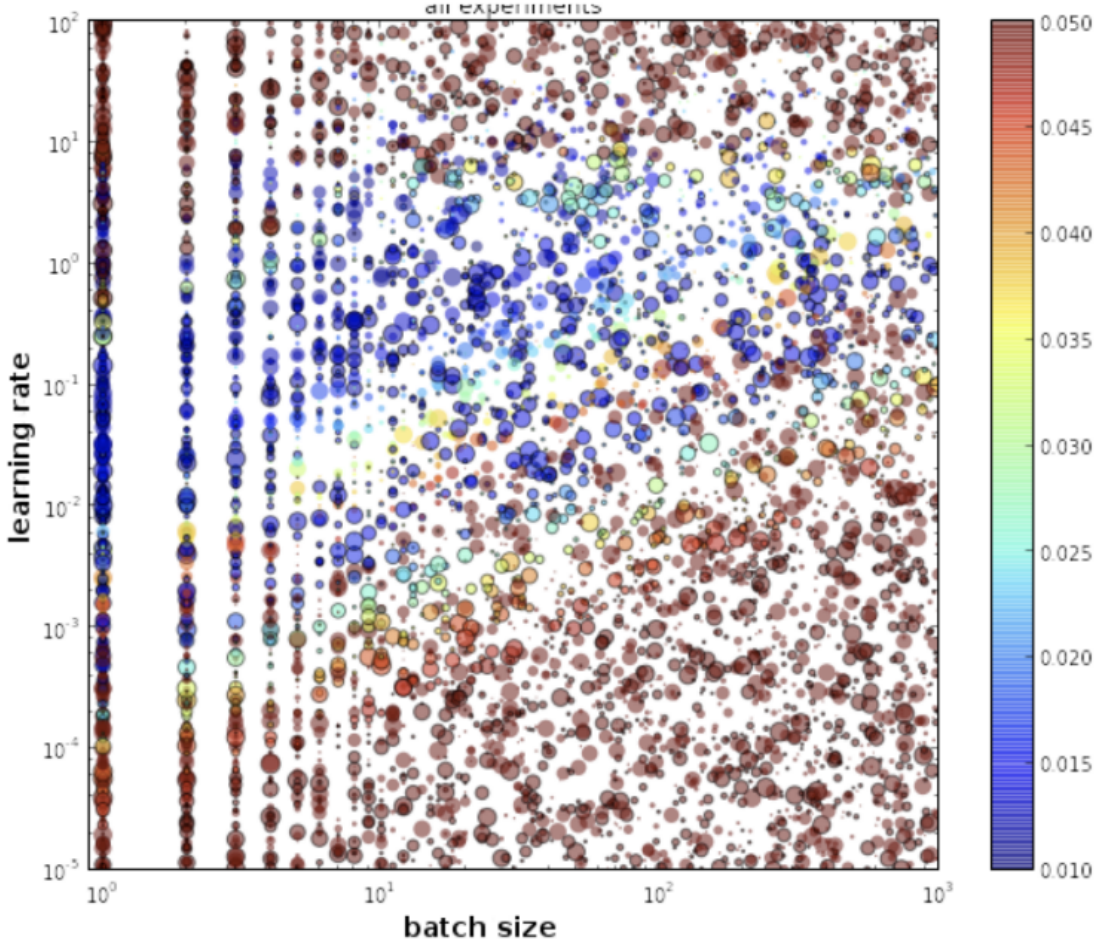
Chen Sun, Abhinav Shrivastava, Saurabh Singh , and Abhinav Gupta (Google Research)

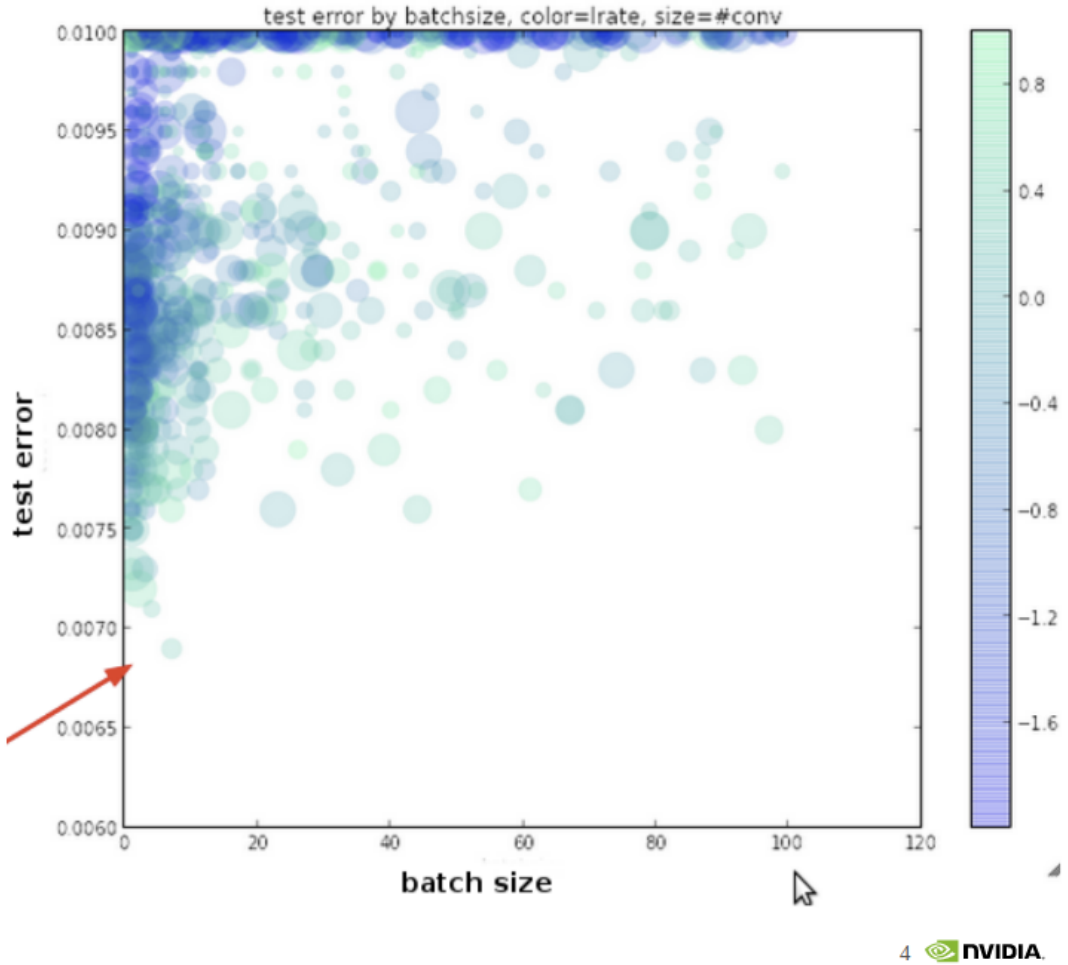


https://arxiv.org/abs/1707.02968

The Effects of Hyperparameters on SGD Training of Neural Networks

Thomas M. Breuel (Google Research)

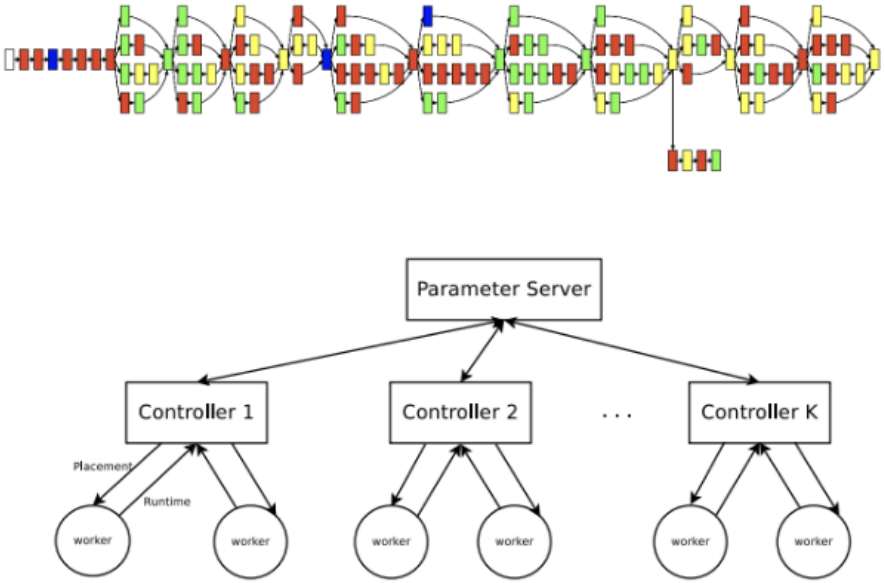




1024 GPUs for a few months

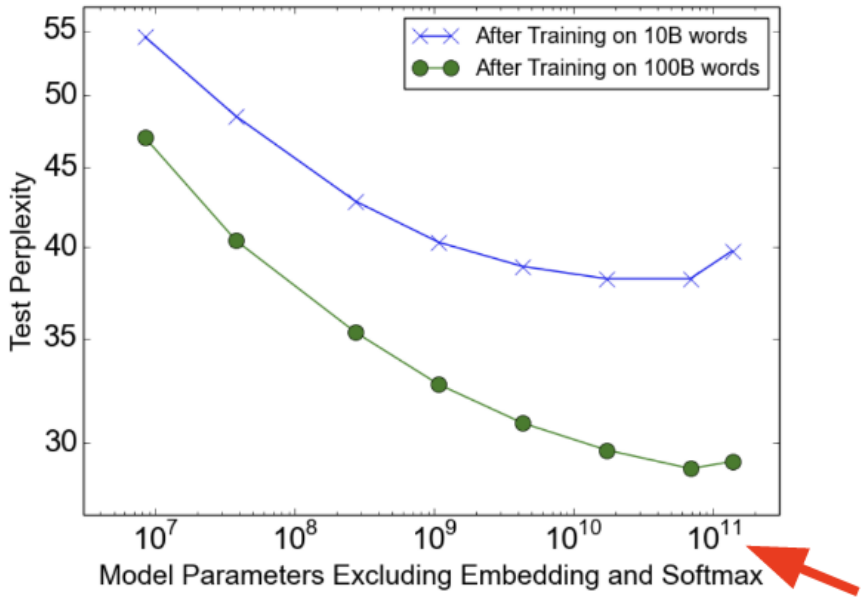
New Architectures

Mirhoseini et al.: *Device Placement with Reinforcement Learning*



https://arxiv.org/abs/1706.04972

Shazeer et al.: *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*



https://arxiv.org/abs/1701.06538

Large Scale Deep Learning

NVIDIA, Google, ...

- thousands of GPUs in a single distributed job
- tens of thousands of cores in a single job
- petabytes of training data
- used for competitive performance on image, speech, etc.

Sample Problem: YT8m

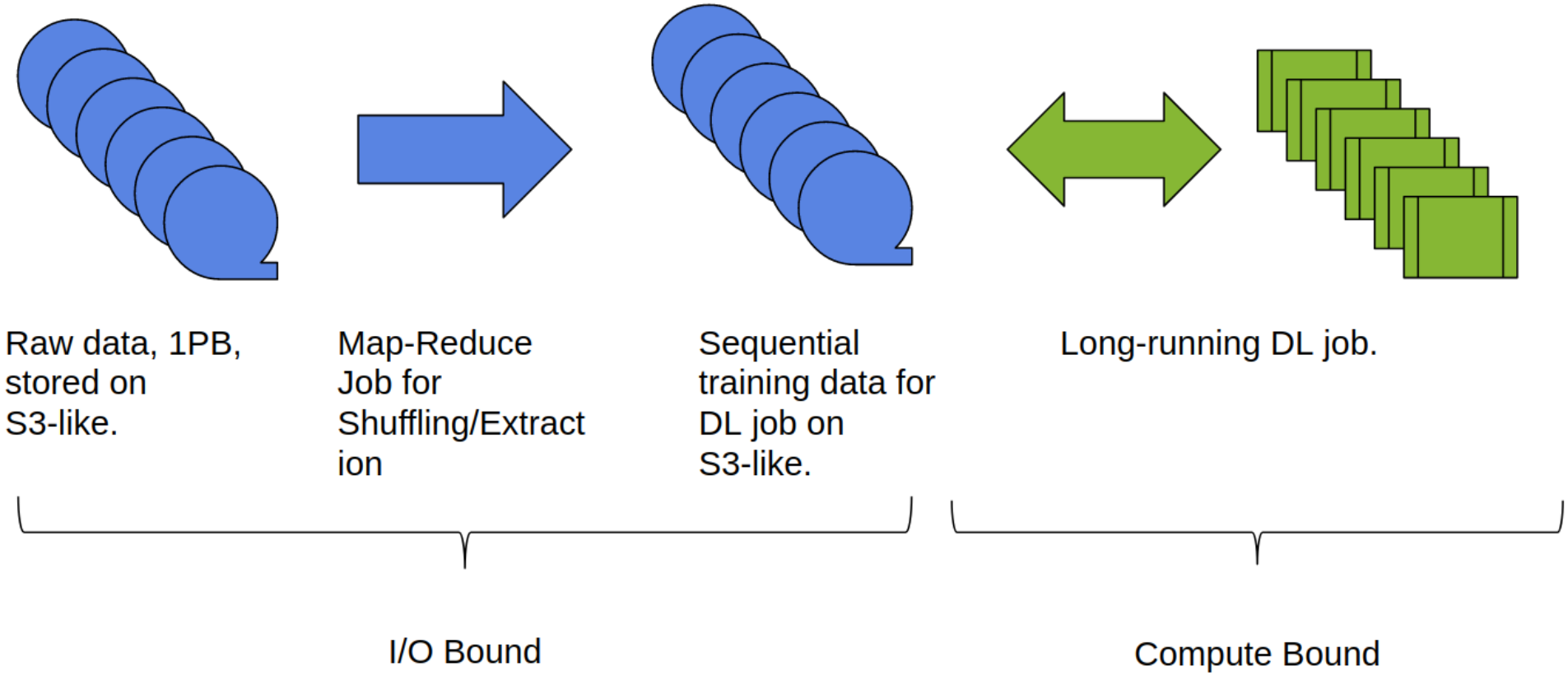
- 8+ million YouTube videos, CC licensed
- 1 PB data
- 500000 h video
- 1 billion images (c.f. JFT-300m)

Labels: categories, captions, ...

Uses for YT8m Dtaaset

- static image or video training data
- source of training data for AV, speech, gesture, …
- large scale unsupervised learning
- new algorithms for large scale, distributed DL

Typical Workflow



Approach

You need to know the core technologies if you want to scale:

- Pytorch (or Tensorflow)
- distributed storage
- Docker
- Kubernetes

Every additional tool has risks, benefits, and costs and they multiply for big data; weigh them carefully.

Big Data = Keep it Simple

- petabytes of scratch space is hard to find
- a single job may cost thousands of dollars to run in the cloud
- when jobs fail, you need to be able to restart/recover
- small changes in jobs may have large impact on cost/running time

Today

- review: image classifier, Docker
- profiling and performance
- multi-GPU, DataParallel
- storage, caching, compression, drive performance
- multinode training, distributed tools
- ETL jobs
- Tensorcom and RDMA

Presentation

Slides and notebooks available at:

```
http://github.com/tmbdev/bigdata19-tutorial
```

Running Jupyter Notebooks

Python Notebooks:

- install environment with `Ansible/...` and run `jupyter notebook`
- run under Docker with `./run jupyter notebook`

Kubernetes/Bash Notebooks:

- run `jupyter notebook` on any host with a working `kubectl`
- local installation with `snap` and `microk8s.com`
- any Kubernetes cloud service, including NGC, Azure, GCE, ...