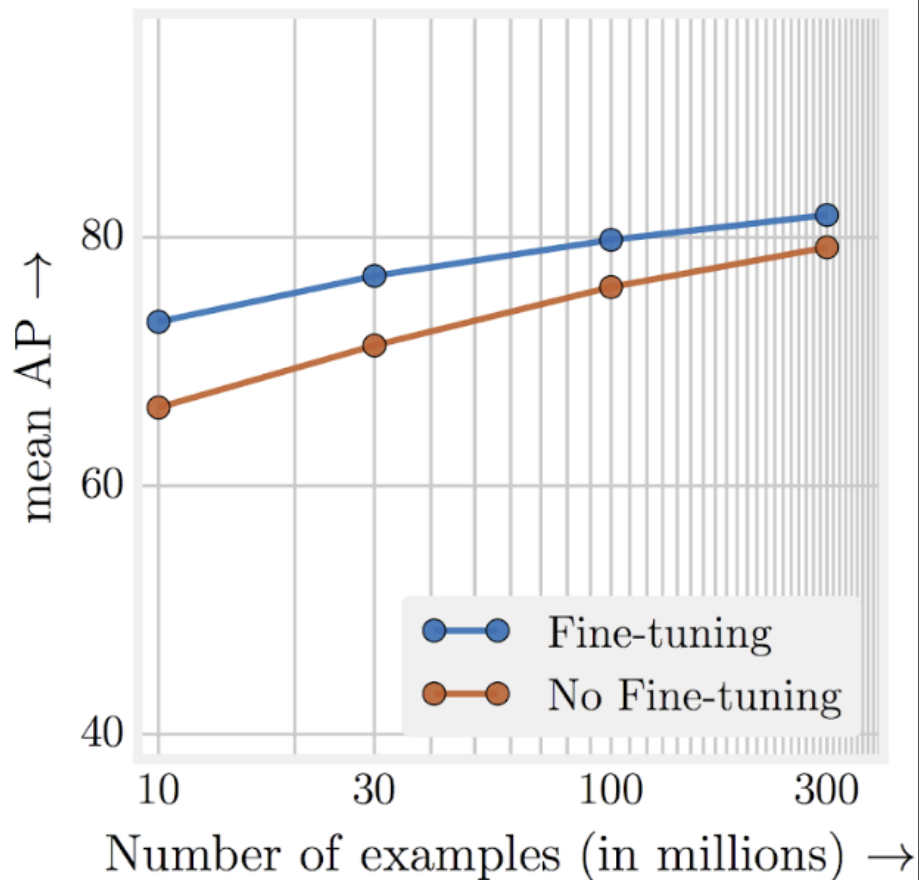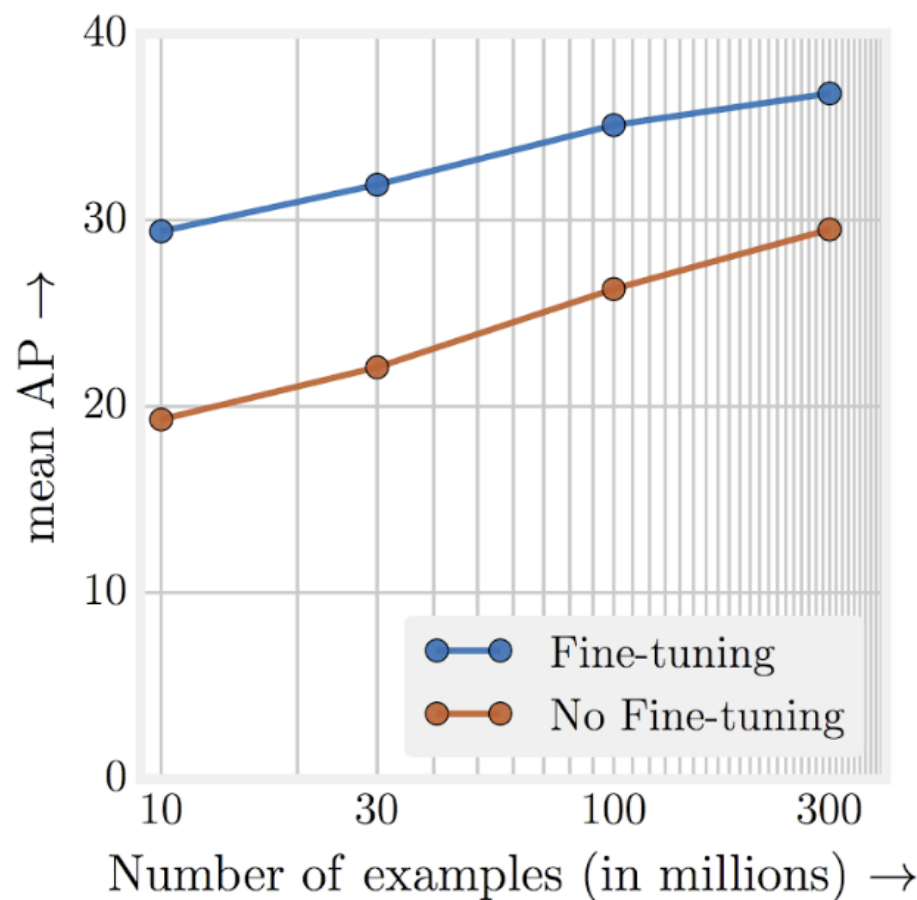# Big Data for AI

Thomas Breuel

Alex Aizman

NVIDIA
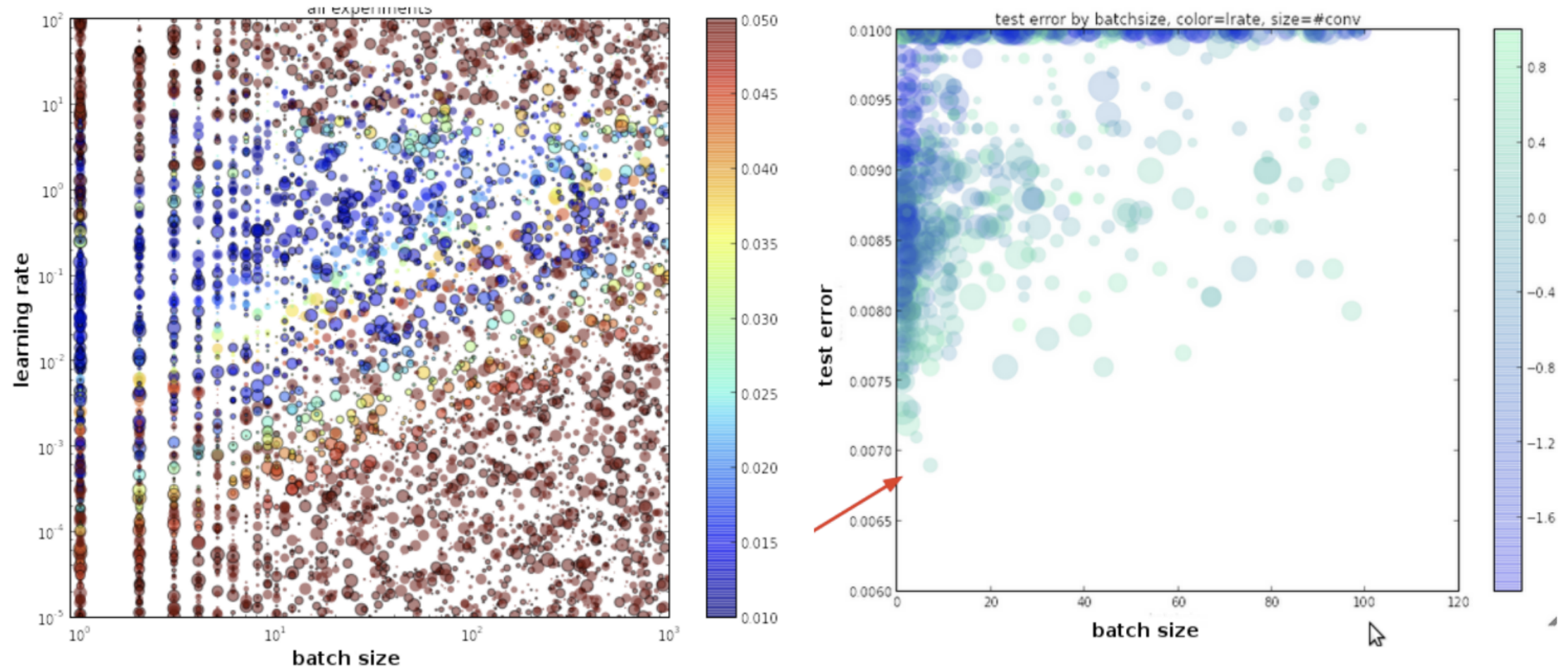
Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

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

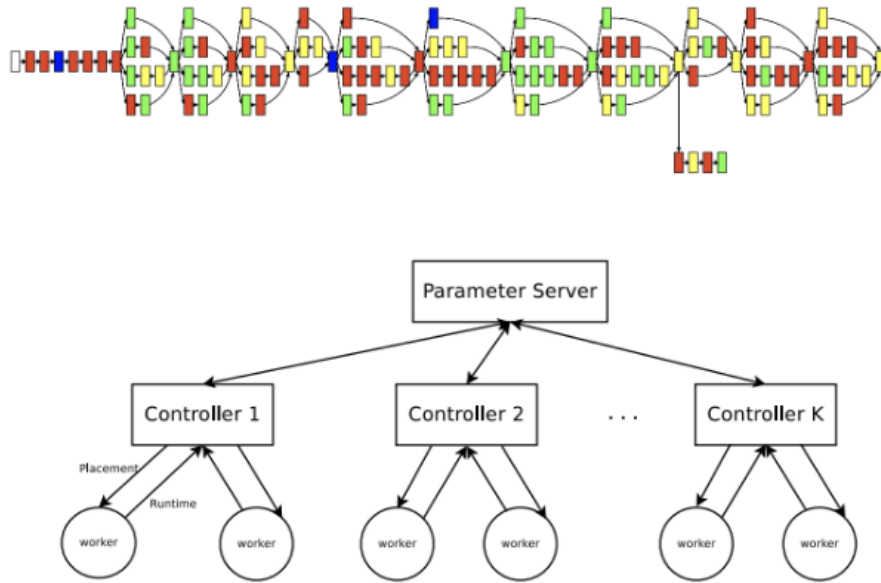# The Effects of Hyperparameters on SGD Training of Neural Networks

Thomas M. Breuel (Google Research)
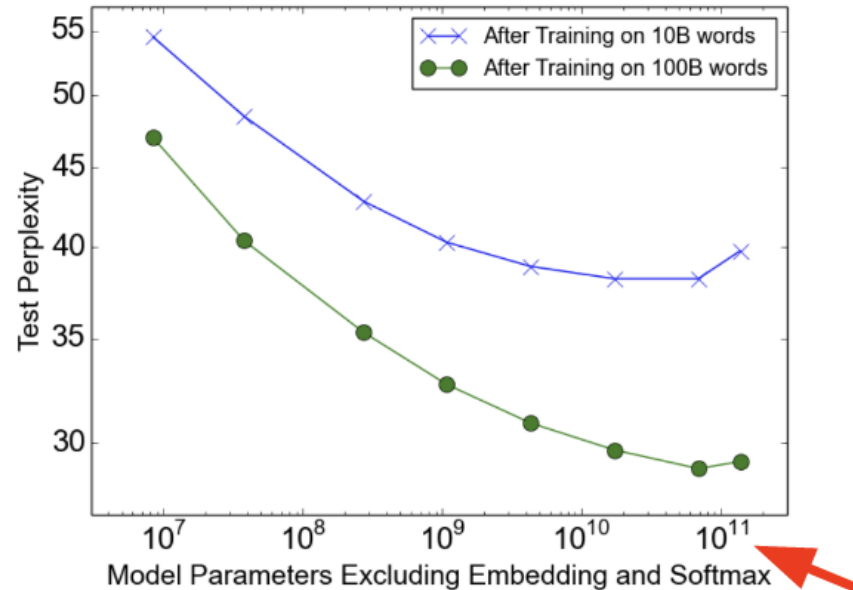

1024 GPUs for a few months

New Architectures

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

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

https://arxiv.org/abs/1706.04972

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.
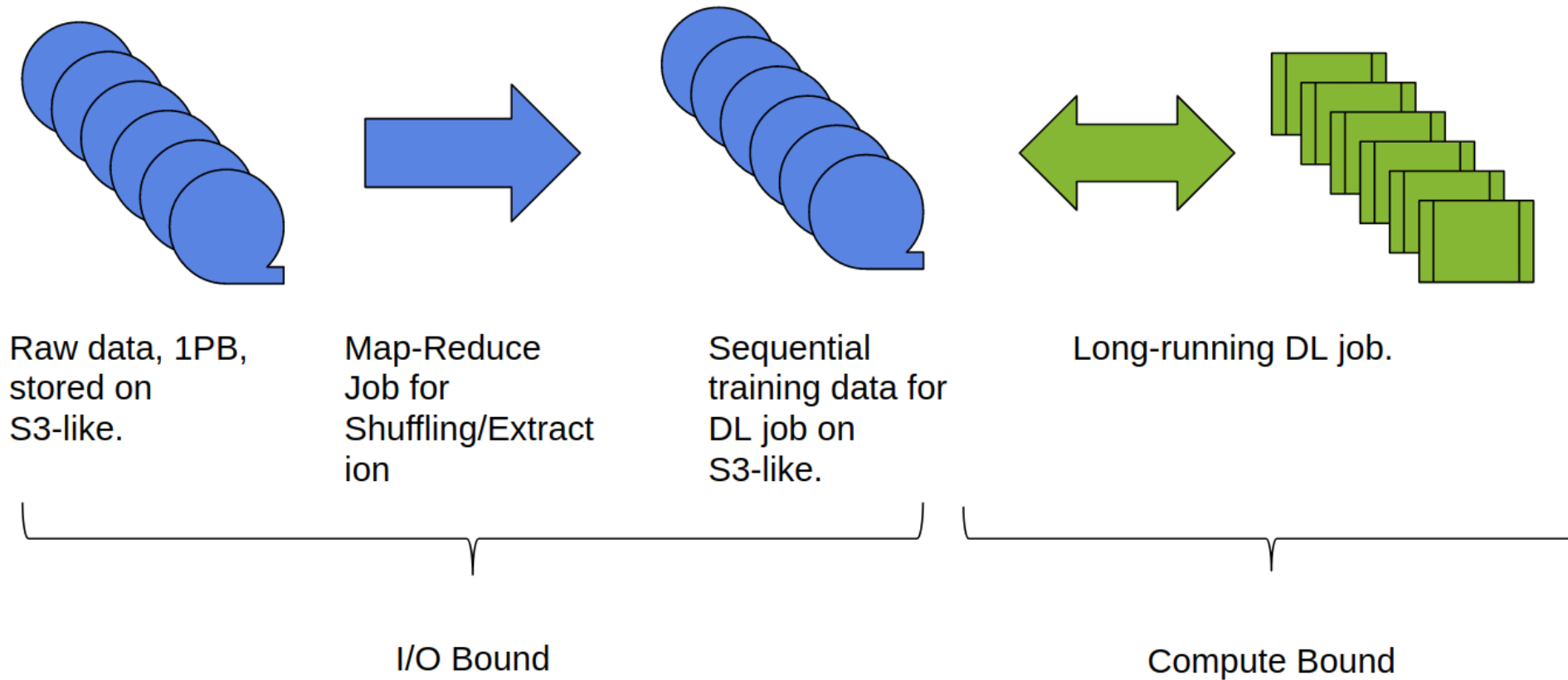
Sample Problem: YT8m

- 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

Raw data, 1PB, stored on S3-like.

Map-Reduce Job for Shuffling/Extraction

Sequential training data for DL job on S3-like.

Long-running DL job.

I/O Bound

Compute Bound

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

Available in about a week on Github under:

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

- slide files: Markdown slides in MARP format

- Notebooks in Jupyter format

- Ansible and shell scripts for installing/Docker containers

- some notebooks run in Docker, others run require K8s access

- run locally, on MicroK8s, and/or any cloud Kubernetes