ImageNet is the new MNIST

Chris Ying
Research SWE @ Google Brain

g.co/brain

on behalf of many people across Google
Goal: “Interactive ML supercomputing”

- **Hardware**
  - Cloud TPUs
  - TPU pods

- **Software**
  - TensorFlow Datasets, Layers, and Estimator APIs (open-source)
  - XLA compiler (open-source) with TPU backend

- **Research**
  - Understanding of generalization gap
  - Large-batch training advances
Motivation

(classical workflow)
Motivation
(classical workflow)
Motivation
(what's happening now)
Motivation

(our vision of the future)
ImageNet is the new MNIST

**MNIST**: 60,000 B&W images

**ImageNet**: 1,281,167 color images
## Motivating results

ResNet-50-v2 on ImageNet

<table>
<thead>
<tr>
<th># of TPU devices</th>
<th>Batch size</th>
<th>Time to 90 epochs</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>256</td>
<td>23 hours 22 minutes</td>
<td>76.6%</td>
</tr>
<tr>
<td>4</td>
<td>1024</td>
<td>5 hours 48 minutes</td>
<td>76.3%</td>
</tr>
<tr>
<td>16</td>
<td>4096</td>
<td>1 hour 30 minutes</td>
<td>76.5%</td>
</tr>
<tr>
<td>32</td>
<td>8192</td>
<td>45 minutes</td>
<td>76.1%</td>
</tr>
<tr>
<td>64</td>
<td>16384</td>
<td>22 minutes</td>
<td>75.0%</td>
</tr>
</tbody>
</table>

Only change between different runs is batch size (linearly scale LR) and hardware, no model changes or hyperparameter re-tuning!
Cloud TPU

180 TFLOPS of computation, 64 GB of HBM memory, 2400 GB/s mem BW
TPUv2 Chip

- 45 TFLOPS
- 16 GB of HBM
- 600 GB/s mem BW
- Vector unit: float32
- Scalar unit: float32
- Matrix unit (MXU): float32 input/output, reduced precision multiplication
Matrix Unit

- 128x128 systolic array
- float32 results*

- 16 GB of HBM
- 600 GB/s mem BW
- Scalar unit: 32b float
- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS

* reduced precision multiplication
Matrix Unit Systolic Array

Computing $y = Wx$

Toy example: 3x3 systolic array
$W = 3x3$ matrix
batch_size($x$) = 3
Matrix Unit Systolic Array

Computing $y = Wx$
with $W = 3 \times 3$, batch_size$(x) = 3$
Matrix Unit Systolic Array

Computing $y = Wx$

with $W = 3x3$, batch_size($x$) = 3
Matrix Unit Systolic Array

Computing $y = Wx$ with $W = 3x3$, batch_size($x$) = 3
Matrix Unit Systolic Array

Computing $y = Wx$
with $W = 3 \times 3$, batch_size($x$) = 3

Matrix Unit (MXU)

inputs

weights

accumulation

$Y_{11} = W_{11}X_{11} + W_{12}X_{12} + W_{13}X_{13}$

outputs
Matrix Unit Systolic Array

Computing $y = Wx$ with $W = 3 \times 3$, batch_size($x$) = 3
Computing $y = Wx$ with $W = 3\times 3$, batch_size($x$) = 3
Matrix Unit Systolic Array

Computing $y = Wx$ with $W = 3 \times 3$, $\text{batch\_size}(x) = 3$
Computing $y = Wx$
with $W = 3 \times 3$, \( \text{batch\_size}(x) = 3 \)
Cloud TPU Pod
64 Cloud TPUs in 2-D toroidal mesh
11.5 petaFLOPS
4 terabytes of HBM memory
Accelerated Linear Algebra (XLA)

- JIT / AOT compiler for linear algebra
- Targets multiple backends, e.g. CPUs, GPUs, and TPUs
- Compiler, runtime, and accelerator-specific optimizer

The life of a neural network:

TF Estimator code → TF Graph
Accelerated Linear Algebra (XLA)

- JIT / AOT compiler for linear algebra
- Targets multiple backends, e.g. CPUs, GPUs, and TPUs
- Compiler, runtime, and accelerator-specific optimizer

The life of a neural network:
Large batch training

- Understanding generalization gap (2016 N. Keskar et. al., 2017 E. Hoffer et. al.)
- Relationship of batch size and noise scale (2018 S. Smith et. al.)
- Learning rate scaling and schedule (2017 P. Goyal et. al.)
- New optimizers
  - Neumann*: approximate inverse Hessian (2018 S. Krishnan et. al.)
  - LARS: per-layer learning rate (2018 Y. You et. al.)

* stick around after this talk to hear more about these!
Experiments

ResNet-50 training on ImageNet

Validation accuracy

- 76.6%

45 min

Hours to 90 epochs

Batch size | # TPUs
---|---
256 | 1
1024 | 4
4096 | 16
8192 | 32
16384 | 64
Experiments
# Experiments

<table>
<thead>
<tr>
<th># of TPU devices</th>
<th>Batch size</th>
<th>Time to 90 epochs</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>8192</td>
<td>44.9 minutes</td>
<td>76.1%</td>
</tr>
<tr>
<td>64</td>
<td>8192</td>
<td>29.8 minutes</td>
<td>75.7%</td>
</tr>
<tr>
<td>64</td>
<td>16384</td>
<td>22.3 minutes</td>
<td>75.0%</td>
</tr>
<tr>
<td>64</td>
<td>65536</td>
<td>17.5 minutes</td>
<td>65.4%</td>
</tr>
<tr>
<td>64</td>
<td>8192 → 16384&lt;sup&gt;[1]&lt;/sup&gt;</td>
<td>29.5 minutes</td>
<td>76.1%</td>
</tr>
</tbody>
</table>

Only change between different runs is batch size (linearly scale LR) and hardware, no model changes or hyperparameter re-tuning!

<sup>[1]</sup> Don't Decay the Learning Rate, Increase the Batch Size (2018 S. Smith et. al)
More than just ImageNet

Transformer model from "Attention is All You Need" (2017 A. Vaswani et. al.)

WMT’14 English-German translation task

Adam optimizer - same learning rate schedule across configurations
Implications

- Faster training enables neural architecture search
  - Reinforcement learning architectures beat existing models in accuracy and cost [1]

Implications

- Faster training enables neural architecture search
  - Reinforcement learning architectures beat existing models in accuracy and cost [1]

- What's the "new ImageNet"?
  - Full ImageNet (14M), Open Images (9M), YouTube-8M
  - Performance increases logarithmically with data [2]

Thank you!

chrisying@google.com

g.co/brain
g.co/tpusignup