Machine Intelligence at Google Scale: Vision/Speech API, TensorFlow and Cloud ML
Kaz Sato

Staff Developer Advocate
Tech Lead for Data & Analytics
Cloud Platform, Google Inc.
What we’ll cover

Deep learning and distributed training

Large scale neural network on Google Cloud

Cloud Vision API and Speech API

TensorFlow and Cloud Machine Learning
Deep Learning and Distributed Training
DNN = a large matrix ops
a few GPUs >> CPU
(but it still takes days to train)
a supercomputer >> a few GPUs
(but you don't have a supercomputer)
You need Distributed Training on the cloud
Google Brain.
Large scale neural network on Google Cloud
Google Cloud is

The Datacenter as a Computer
Jupiter network

10 GbE x 100 K = 1 Pbps

Consolidates servers with microsec latency
Borg

No VMs, pure containers
10K - 20K nodes per Cell
DC-scale job scheduling
CPUs, mem, disks and IO
Google Cloud + Neural Network = Google Brain
The Inception model (GoogLeNet, 2015)
What's the scalability of Google Brain?

"Large Scale Distributed Systems for Training Neural Networks", NIPS 2015

- Inception / ImageNet: 40x with 50 GPUs
- RankBrain: 300x with 500 nodes
Large-scale neural network for everyone
Growing use of deep learning at Google

Number of directories containing model description files

- 1500
- 1000
- 500

Across many areas:
- AlphaGo
- Apps
- Maps
- Photos
- Gmail
- Speech
- Android
- YouTube
- Translation
- Robotics Research
- Image Understanding
- Natural Language Understanding
- Drug Discovery

2012 2013 2014 2015
Cloud Vision API

Pre-trained models. No ML skill required

REST API: receives images and returns a JSON

$2.5 or $5 / 1,000 units (free to try)

Public Beta - cloud.google.com/vision
Demo
Cloud Speech API

- **Pre-trained** models. No ML skill required
- **REST** API: receives audio and returns texts
- Supports **80+** languages
- **Streaming** or **non-streaming**
- **Limited Preview** - cloud.google.com/speech
Demo Video
TensorFlow
The Machine Learning Spectrum

- TensorFlow
- Cloud Machine Learning
- Machine Learning APIs

Academic / research

Industry / applications
What is TensorFlow?

Google's open source library for machine intelligence

tensorflow.org launched in Nov 2015
The second generation
Used by many production ML projects
What is TensorFlow?

**Tensor**: N-dimensional array

**Flow**: data flow computation framework (like MapReduce)

For Machine Learning and Deep Learning

Or any **HPC** (High Performance Computing) applications
```python
# define the network
import tensorflow as tf
x = tf.placeholder(tf.float32, [None, 784])
W = tf.Variable(tf.zeros([784, 10]))
b = tf.Variable(tf.zeros([10]))
y = tf.nn.softmax(tf.matmul(x, W) + b)

# define a training step
y_ = tf.placeholder(tf.float32, [None, 10])
xent = -tf.reduce_sum(y_*tf.log(y))
step = tf.train.GradientDescentOptimizer(0.01).minimize(xent)
```
# initialize session
init = tf.initialize_all_variables()

# training
for i in range(1000):
    batch_xs, batch_ys = mnist.train.next_batch(100)
    sess.run(step, feed_dict={x: batch_xs, y_: batch_ys})
<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element-wise mathematical operations</td>
<td>Add, Sub, Mul, Div, Exp, Log, Greater, Less, Equal, ...</td>
</tr>
<tr>
<td>Array operations</td>
<td>Concat, Slice, Split, Constant, Rank, Shape, Shuffle, ...</td>
</tr>
<tr>
<td>Matrix operations</td>
<td>MatMul, MatrixInverse, MatrixDeterminant, ...</td>
</tr>
<tr>
<td>Stateful operations</td>
<td>Variable, Assign, AssignAdd, ...</td>
</tr>
<tr>
<td>Neural-net building blocks</td>
<td>SoftMax, Sigmoid, ReLU, Convolution2D, MaxPool, ...</td>
</tr>
<tr>
<td>Checkpointing operations</td>
<td>Save, Restore</td>
</tr>
<tr>
<td>Queue and synchronization operations</td>
<td>Enqueue, Dequeue, MutexAcquire, MutexRelease, ...</td>
</tr>
<tr>
<td>Control flow operations</td>
<td>Merge, Switch, Enter, Leave, NextIteration</td>
</tr>
</tbody>
</table>

Table 1: Example TensorFlow operation types
Portable

- Training on:
  - Data Center
  - CPUs, GPUs and etc

- Running on:
  - Mobile phones
  - IoT devices
TensorBoard: visualization tool
Cloud Machine Learning
Cloud Machine Learning (Cloud ML)

Fully managed, distributed training and prediction for custom TensorFlow graph

Supports Regression and Classification initially

Integrated with Cloud Dataflow and Cloud Datalab

Limited Preview - cloud.google.com/ml
Use your own data to train models

Cloud Dataflow (Apache Beam)
Cloud Machine Learning
Cloud Storage
Google BigQuery
Cloud Datalab

Develop / Model / Test
Distributed Training with TensorFlow
Distributed Training with TensorFlow

- CPU/GPU scheduling
- Communications
  - Local, RPC, RDMA
  - 32/16/8 bit quantization
- Cost-based optimization
- Fault tolerance
Data Parallelism
= split data, share model

(but ordinary network is 1,000x slower than GPU and doesn't scale)
Cloud ML demo video
Cloud ML demo

Jeff Dean's keynote: [YouTube video](#)

Define a custom TensorFlow graph

Training at local: **8.3 hours** w/ 1 node

Training at cloud: **32 min** w/ **20 nodes** (15x faster)

Prediction at cloud at **300 reqs / sec**
Summary
Ready to use Machine Learning models

Use your own data to train models

Cloud Vision API
Cloud Speech API
Cloud Translate API

Cloud Dataflow (Apache Beam)
Cloud Machine Learning
Cloud Storage
Google BigQuery
Cloud Datalab

Develop - Model - Test

Stay Tuned...
Large Scale Distributed Systems for Training Neural Networks, Jeff Dean and Oriol Vinals

Cloud Vision API: cloud.google.com/vision

Cloud Speech API: cloud.google.com/speech

TensorFlow: tensorflow.org

Cloud Machine Learning: cloud.google.com/ml

Cloud Machine Learning: demo video
Thank you!