Intro to TensorFlow and Machine Learning

October 18, 2017

Edward Doan
Google Cloud Customer Engineering
@edwardd
Agenda

Deep Learning

What is TensorFlow

Under the hood

Resources
Deep Learning
Deep Learning

Input

Out

Google Cloud

@edwardd
goo.gl/stbR1K
## Results with Deep Learning

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixels: <img src="lion.png" alt="Image of a lion" /></td>
<td>“lion”</td>
</tr>
<tr>
<td>Audio: <img src="audio.png" alt="Audio visualization" /></td>
<td>“see at tuhl res taur aun ts”</td>
</tr>
<tr>
<td><code>&lt;query, doc&gt;</code></td>
<td>$P(\text{click on doc})$</td>
</tr>
<tr>
<td>“Hello, how are you?”</td>
<td>“Bonjour, comment allez-vous?”</td>
</tr>
<tr>
<td>Pixels: <img src="child.png" alt="Image of a child" /></td>
<td>“A close up of a small child holding a stuffed animal”</td>
</tr>
</tbody>
</table>
Growing Use of Deep Learning at Google

![Graph showing the number of directories containing model description files over time. The graph indicates a significant increase in unique project directories from Q1 2012 to Q3 2015.](image)

Google Cloud
Important Property of Neural Networks

Results get better with

More data +

Bigger models +

More computation

(Better algorithms, new insights, and improved techniques always help, too!)
What is TensorFlow?
Like so many good things, it started as a research paper...

TensorFlow: Largest-Scale Machine Learning on Heterogeneous Distributed Systems

(Preliminary White Paper, November 9, 2015)


Google Research*

Abstract

TensorFlow [1] is an interface for expressing machine learning algorithms, and an implementation for executing such algorithms. A computation expressed using TensorFlow can be executed with little or no change on a wide variety of heterogeneous systems, ranging from mobile devices such as phones and tablets up to large-scale distributed systems of hundreds of machines and thousands of computational devices such as GPU cards. This system is flexible and can be used to express a wide variety of algorithms, including training and inference algorithms for deep neural network models, and it has been used for conducting research and for deploying machine learning systems into production across more than a dozen areas of computer science and other fields, including speech recognition, computer vision, robotics, information retrieval, natural language processing, geographic information extraction, and computational drug discovery. This paper describes the TensorFlow interface and an implementation of that interface that we have built at Google. The TensorFlow API and a reference implementation were released as an open-source package under the Apache 2.0 license in November 2015 and are available at www.tensorflow.org.

1 Introduction

The Google Brain project started in 2011 to explore the use of very-large-scale deep neural networks, both for research and for use in Google’s products. As part of the early work in this project, we built DistBelief, our first-generation scalable distributed training and inference system [14], and this system has served as well. We and others at Google have performed a wide variety of research using DistBelief including work on unsupervised sequence prediction [47], move selection for Go [34], poled action detectors [2], reinforcement learning [38], and other areas [17, 3]. In addition, often in close collaboration with the Google Brain team, more than 50 teams at Google and other Alphabet companies have deployed deep neural networks using DistBelief in a wide variety of products, including Google Search [11], our advertising products, our speech recognition systems [30, 6, 46], Google Photos [43], Google Maps and StreetView [19], Google Translate [39], YouTube, and many others.

Based on our experience with DistBelief and a more complete understanding of the decoupled system propensities and requirements for training and using neural networks, we have built TensorFlow, our second-generation system for the implementation and deployment of large-scale machine learning models. TensorFlow takes computations described using a dataflow-like model and maps them onto a wide variety of different hardware platforms, ranging from running inference on mobile device platforms such as Android and iOS to modernized training and inference systems using single machines containing one or many GPU cards to large-scale training systems running on hundreds of specialized machines with thousands of CPUs. Having a single system that can span such a broad range of platforms significantly simplifies the real-world use of machine learning system, as we have found that having separate systems for large-scale training and small-scale deployment leads to significant maintenance burdens and leaky abstractions. TensorFlow computations are expressed as a directed dataflow graph (discussed in more detail in Section 2), and we have focused on making the system both flexible enough for quick experimenting with new models for research purposes and sufficiently high performance and...
#1 repository for "machine learning" category on GitHub
What is TensorFlow?

- TensorFlow is a machine learning library enabling researchers and developers to build the next generation of intelligent applications.
- Provides distributed, parallel machine learning based on general-purpose dataflow graphs
- Targets heterogeneous devices:
  - single PC with CPU or GPU(s)
  - mobile device
  - clusters of 100s or 1000s of CPUs and GPUs
Flexible

- General computational infrastructure
  - Works well for Deep Learning
  - Deep Learning is a set of libraries on top of the core
  - Also useful for other machine learning algorithms, maybe even more traditional high performance computing (HPC) work
  - Abstracts away the underlying devices
A “TPU pod” built with 64 second-generation TPUs delivers up to 11.5 petaflops of machine learning acceleration.
TensorFlow powered Cucumber Sorter

Arduino Micro
Controls Conveyor
Servo Motors + Sorts
Using Cloud Data

Raspberry Pi 3
TensorFlow/OpenCV Takes & Sends Cucumber Photos to Server + Controls Servo Motors

Google Cloud
TensorFlow/Django Sort Cucumber Photos into 9 Classes + Stores & Manages Data

From: http://workpiles.com/2016/02/tensorflow-cnn-cucumber/

@edwardd goo.gl/stbR1K
Under the hood
TensorFlow

C++ front end

Python front end

Core TensorFlow Execution System

Ops

Kernels

add
mul
print
reshape
...

CPU
GPU
Android
iOS
...

Bindings + Compound Ops

Google Cloud

@edwardd
goo.gl/stbR1K
Computation is a dataflow graph

Graph of Nodes, also called Operations or ops.

- biases
- weights
- examples
- labels

MatMul → Add → Relu → Xent
Dataflow graph

Edges are N-dimensional arrays: *Tensors*
Dataflow graph with state

'Biases' is a variable

Some ops compute gradients

-= updates biases

biases

learning rate

Add

Mul

-=

...
Parallelism

- Parallel op implementations
Task Parallelism in the Dataflow Graph

... → MatMul → ...

... → MatMul → ...

... → MatMul → ...
Data parallelism by replicating the graph
Model parallelism by splitting one op into many

Diagram:
- Matrix
  - Split
    - MatMul
    - MatMul
  - Matrix
  - Concat
input = ...
biases = tf.get_variable('biases', ...)
weights = tf.get_variable('weights', ...)
out = tf.matmul(input, weights)
out = tf.add(out, biases)
out = tf.nn.relu(out)
input = ...
output = tf.layers.fully_connected(input, ...)
TensorFlow contains complete algorithms

Linear{Classifier,Regressor}

DNN{Classifier,Regressor}

DNNLinearCombined{Classifier,Regressor}

SVM

KMeansClustering

GMM

...
Simple machine learning

classifier = learn.LinearClassifier(feature_columns=feature_columns,
                                 n_classes=10)

classifier.fit(data, labels, batch_size=100, steps=1000)

classifier.evaluate(eval_data, eval_labels)

Tooling provided for distributed training and evaluation, graphical debugging, and export to production server (tensorflow/serving).
Visualizing learning
LIVE DEMO IN FRONT OF AN AUDIENCE
Let’s play!

Visit [https://jhub.edwarddoan.com](https://jhub.edwarddoan.com) (sorry for the SSL cert error 🙃)

Click this button -> Sign in with Google

Log in as i2-x@glusd.org / gogoogle (where x is unique to you)

Start “My Server”

If you’re new to Tensorflow, play with 2_getting_started.ipynb

If you’re kinda familiar with Tensorflow, play with 3_mnist_from_scratch.ipynb

Click the Play button to go through the markdown and code cell-by-cell
Resources

Website: https://www.tensorflow.org

Code: https://github.com/tensorflow/tensorflow
