TensorFlow
On Embedded Devices
Who am I?

Pete Warden (petewarden@google.com)
Tech Lead of the TensorFlow Mobile/Embedded team.
Why am I here?
Why is deep learning so important?
Where does TensorFlow fit in?

DistBelief (1st system) was great for scalability, and production training of basic kinds of models.

Better understanding of problem space allowed us to make some dramatic simplifications.

Google brings years of production experience, and a large team with a long-term commitment.
TensorFlow: Expressing High-Level ML Computations

Core in C++

Very low overhead
Core in C++

Very low overhead

Different front ends for specifying/driving the computation

Python and C++ today, easy to add more
TensorFlow: Expressing High-Level ML Computations

Core in C++

Very low overhead

Different front ends for specifying/driving the computation

Python and C++ today, easy to add more

Core TensorFlow Execution System

C++ front end  Python front end
Computation is a dataflow graph

Graph of *Nodes*, also called *Operations* or *ops.*
Computation is a dataflow graph with tensors

Edges are N-dimensional arrays: Tensors

- biases
- weights
- examples
- labels

MatMul → Add → Relu → Xent
TensorFlow: Expressing High-Level ML Computations

Automatically runs models on range of platforms:

from phones ...

to single machines (CPU and/or GPUs) ...

to distributed systems of many 100s of GPU cards
What about the cloud?
What does this mean in practice?

Start with file format:


Then deeper integration.

Eight-bit is enough!
Eight-bit resources

http://github.com/google/gemmlowp - 60 GOPs/s on Nexus 5!

GoogLeNet v1 is 7MB after just quantization.

BNNM API in Android

More example code and models to come.
Next steps

Ask me how - petewarden@google.com

Thanks!