Mobile-Ready Deep Learning Networks

- SqueezeNet
  - Same accuracy as AlexNet with 50x fewer parameters
  - Replaces 3x3 filters with 1x1 filters
  - Squeeze layer
- MobileNet
  - Depthwise separable convolutions
  - Probably the best first choice for your deep learning applications

Programming Support for Deep Learning on Mobiles

- Core ML for iOS
- Caffe2 for iOS and Android
- TensorFlow Lite for Android and iOS
- PyTorch Mobile for Android and iOS
- Other players:
  - Fritz AI
  - Snapdragon SDK
  - …

ML Kit and Firebase ML

- Firebase – a framework supporting mobile app development
  - Messaging, authentication, database, monitoring, etc.
- Firebase ML
  - Support for distributing models to mobile devices
- ML Kit – on-device ML
  - Prebuilt models for text recognition, face detection, object detection and tracking, barcode scanning, etc.

TensorFlow Lite

- TensorFlow – a framework for NN programming
  - Build and train your NN (on a powerful computer)
  - Validate/test your NN
  - Keras – higher-level API for building and training NNs
- TensorFlow Lite – a mobile NN support library
  - Interpret a TF NN model on a mobile device
  - Firebase ML and ML Kit models use TensorFlow Lite under the hood

TensorFlow Lite

- Supported platforms
  - Android, iOS, Raspberry Pi, microcontrollers
- Means of operation
  - Pre-train a model in TensorFlow (Keras)
  - Convert the model to TensorFlow Lite, save to a file, and ship with your app
  - At runtime, an Interpreter runs a model on device

You can also dynamically change the model remotely via the Firebase ML console!

TensorFlow Lite – Achieving Speedup

- Running platform optimization
  - The model can be ran on CPU or GPU
- Faster loading
  - Memory mapped files in Android
- Quantization
  - To 16b floats or integers
  - To 8b dynamic range
  - Weights only, or weights and activations
- Pruning
- Weight sharing/virtualisation

Don't compress the model file!

TensorFlow Lite – Bootstrapping

- A number of models are already available:
  - https://www.tensorflow.org/lite/models
  - https://github.com/tensorflow/models
- Transfer learning
  - The higher the layer is in the hierarchy, the more specific its inference is
  - Take the first *N-k* layers of the pre-built model and re-train the last *k* layers with your data

This is what you do in next week's lab!

TensorFlow Lite – Bootstrapping

- A shortcut:
  - AutoML model re-trained(?) with your dataset
  - Deployed to Android or pulled from the server on demand
  - To start go to: https://cloud.google.com/automl
  - Prepare your dataset with labels
  - Train the model and load the file in your app or provide a link through the AutoML API

TODO

- Read "*Mental Health and Behavior of College Students During the Early Phases of the COVID-19 Pandemic: Longitudinal Smartphone and Ecological Momentary Assessment Study* " for next week!
- Keep working on your projects!