Deep Neural Networks for Mobile Platforms

Oleksandr Baiev

PhD, Sr. Engineer
Samsung R&D Institute Ukraine

AI Ukraine, 2016
Outline

Problems overview
  - Key problems
  - Deeper problem overview

Topology
  - Overview

Quantization
  - Overview

Pruning
  - Overview
  - Examples
Outline

Problems overview
  • Key problems
  • Deeper problem overview

Topology
  • Overview

Quantization
  • Overview

Pruning
  • Overview
  • Examples
Neural networks

- Cutting edge results for CV, NLP, Signal Processing, Recommendations.
- Unified solution for different problems.
Neural networks

Convolution neural network
Deployment stage problem

Models quality

Models top-5 accuracy

Deployment stage problem

Models size

Application more than 100MB requires WiFi for downloading via app stores
Deployment stage problem
Models execution

Test with https://github.com/sh1r0/caffe-android-lib
Outline

Problems overview
  • Key problems
  • Deeper problem overview

Topology
  • Overview

Quantization
  • Overview

Pruning
  • Overview
  • Examples
Problem overview

Approximate layers distribution

- Fully Connected layers are bigger than Convolution layers in terms of MB.
- Convolution takes much more time for forward pass.
- Target device have to store layer’s feature maps in RAM for at least one layer.
Problem overview

Execution time and size by layer

- Exec time of the layer decreases with NN depth
- Size of the layer's weights grows with NN depth
Problem overview

Parameters importance

- Feature map’s **width**, **height** influence on execution time
- Feature map’s **depth** influences on model size
Outline

Problems overview
  - Key problems
  - Deeper problem overview

Topology
  - Overview

Quantization
  - Overview

Pruning
  - Overview
  - Examples
Bottlenecks

Inception module

Compose different kernel sizes

- Christian Szegedy, et. al., Rethinking the Inception Architecture for Computer Vision, 2015
Outline

Problems overview
  - Key problems
  - Deeper problem overview

Topology
  - Overview

Quantization
  - Overview

Pruning
  - Overview
  - Examples
Quantization

- Pete Warden, How to Quantize Neural Networks with TensorFlow, 2016
- Matthieu Courbariaux, et. al., BinaryConnect: Training Deep Neural Networks with binary weights during propagations, 2015
Quantization Schema

- All operations use precalculated Min and Max values which are used for rescaling.
- Min and Max values selected from function behavior and real number values.
Outline

Problems overview
  - Key problems
  - Deeper problem overview

Topology
  - Overview

Quantization
  - Overview

Pruning
  - Overview
  - Examples
The idea of pruning is removing unimportant weights. The one question is how to define “unimportant”.
Pruning basics

Unimportant criteria

- Song Han, et al., Learning both Weights and Connections for Efficient Neural Networks, 2015
Pruning basics

Iterative Process
Outline

Problems overview
  - Key problems
  - Deeper problem overview

Topology
  - Overview

Quantization
  - Overview

Pruning
  - Overview
  - Examples
Pruning example VGG

![Graph showing accuracy over part of pruned weights.](image)

- **Baseline**
- **VGG iter #1** (base: original)
- **VGG iter #2** (base: 87% removed and trained)
- **VGG iter #3** (base: 88% removed and trained)
- **Final** (83% weights removed, 88.99% final accuracy)

---

Song Han, DSD: Regularizing Deep Neural Networks with Dense-Sparse-Dense Training Flow, 2016
Pruning example VGG

Pruned weights distribution. VGG-16, Layer conv5_2
Pruning example GoogLeNet

Hao Li, et.al., Pruning Filters for Efficient ConvNets, 2016
Pruning example ResNet-152

- Baseline
- ResNet-152. iter #1 (base: original)
- ResNet-152. iter #2 (base: 70% removed and trained)
- Final (42% weights removed, 91.66% final accuracy)
Summary

- Deep neural networks provides excellent quality, but requires powerful computation instances
- There are several simple and useful approaches for reducing required memory size and execution time without increasing hardware cost