MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam
Google Inc., 2017
Introduction

• MobileNets is a class of efficient models for mobile and embedded vision applications

• Use depthwise separable convolutions to build light weight deep neural networks

• Add two simple global hyperparameters: width multiplier and resolution multiplier that efficiently trade off between latency and accuracy
Introduction
Prior Work

• Compressing pretrained networks
  — Product quantization, Hashing, Pruning, Vector quantization, Huffman coding
  — Factorization

• Training small networks
  — Distillation
Depthwise Separable Convolution

Standard Convolution Filters

Depthwise Convolutional Filters

Pointwise Convolution

- \( D_K \): the spatial dimension of the kernel
- \( M \): the number of input channels (input depth)
- \( N \): the number of output channel (output depth)

-> Reduce computation and model size
Depthwise Separable Convolution

• Standard convolutional layer:
  — Input: $D_F \times D_F \times M$ feature map $F$
  — Output: $D_F \times D_F \times N$ feature map $G$
  — Convolution kernel: $D_K \times D_K \times M \times N$ kernel $K$

\[
G_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m}
\]

Cost: $D_K \times D_K \times M \times N \times D_F \times D_F$
Depthwise Separable Convolution

• Depthwise separable convolution:

  • Depthwise convolutions

  \[ \hat{G}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \]

  • Pointwise convolutions

Cost: \( D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \)
Depthwise Separable Convolution

• Reduction

\[
\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2}
\]
MobileNet structure

• 28 layers

```
3x3 Conv
  ↓
  BN
  ↓
  ReLU

3x3 Depthwise Conv
  ↓
  BN
  ↓
  ReLU
  ↓
  1x1 Conv
  ↓
  BN
  ↓
  ReLU
```
Width Multiplier: Thinner Models

- With width multiplier $\alpha$, the cost becomes:

$$D_K \cdot D_K \cdot \alpha M \cdot D_F \cdot D_F + \alpha M \cdot \alpha N \cdot D_F \cdot D_F$$

- $\alpha \in (0,1]$ with typical settings of 1, 0.75, 0.5 and 0.25
Resolution Multiplier: Reduced Representation

• With width multiplier $\rho$, the cost becomes:

$$DK \cdot DK \cdot \alpha M \cdot \rho DF + \alpha M \cdot \alpha N \cdot \rho DF$$

• $\rho \in (0,1]$ with typical settings $\rho$ to make input resolution of the network be 224, 192, 160 or 128
Depthwise Separable Convolutions v.s. Full Convolutions

Table 4. Depthwise Separable vs Full Convolution MobileNet

<table>
<thead>
<tr>
<th>Model</th>
<th>ImageNet Accuracy</th>
<th>Million Mult-Adds</th>
<th>Million Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv MobileNet</td>
<td>71.7%</td>
<td>4866</td>
<td>29.3</td>
</tr>
<tr>
<td>MobileNet</td>
<td>70.6%</td>
<td>569</td>
<td>4.2</td>
</tr>
</tbody>
</table>
Width Multiplier v.s. Shallow Model

<table>
<thead>
<tr>
<th>Model</th>
<th>ImageNet Accuracy</th>
<th>Million Mult-Adds</th>
<th>Million Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75 MobileNet</td>
<td>68.4%</td>
<td>325</td>
<td>2.6</td>
</tr>
<tr>
<td>Shallow MobileNet</td>
<td>65.3%</td>
<td>307</td>
<td>2.9</td>
</tr>
</tbody>
</table>
Trade off between computation (Mult-Adds) and Accuracy

Imagenet Accuracy vs Mult-Adds
Trade off between Number of and Accuracy

Imagenet Accuracy vs Million Parameters

[Graph showing the trade-off between ImageNet accuracy and million parameters for different resolutions (224, 192, 160, 128).]
Conclusion

• Propose a new model architecture called MobileNets

• Two Features:
  — Use depthwise separable convolutions
  — Use width multiplier and resolution multiplier