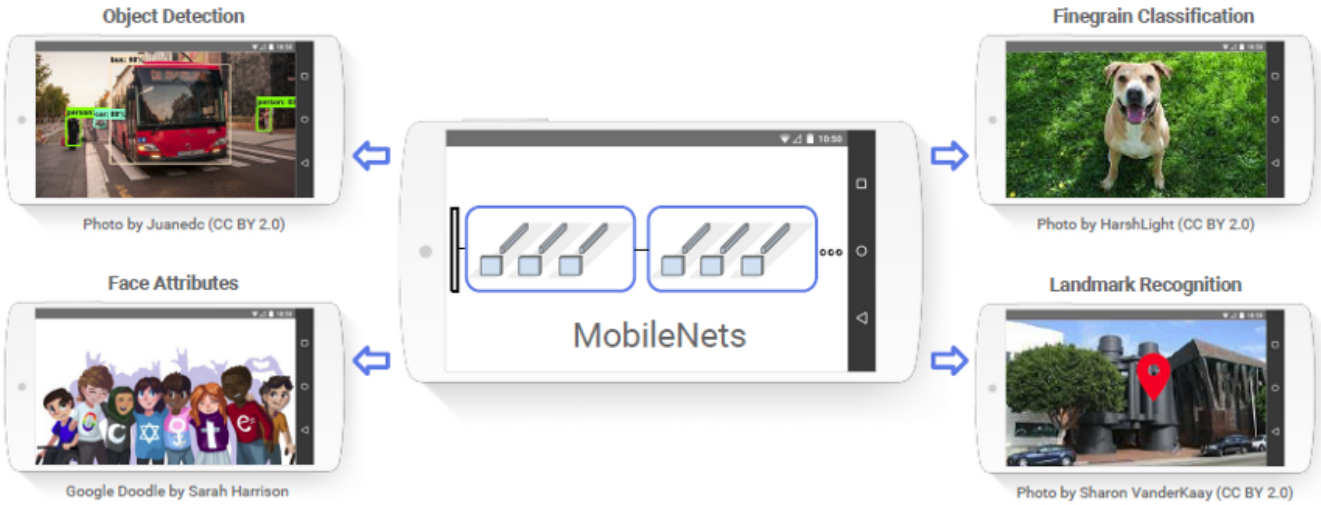
Review: MobileNetV1 — Depthwise Separable Convolution (Light Weight Model)

Sik-Ho Tsang Oct 14, 2018 · 5 min read

In this story, **MobileNetV1** from Google is reviewed. **Depthwise Separable Convolution** is used to reduce the model size and complexity. It is **particularly useful for mobile and embedded vision applications**.

- **Smaller model size**: Fewer number of parameters
- **Smaller complexity**: Fewer Multiplications and Additions (Multi-Adds)



When MobileNets Applied to Real Life

more than 600 citations when I was writing this paper. (Sik-Ho Tsang @ Medium)

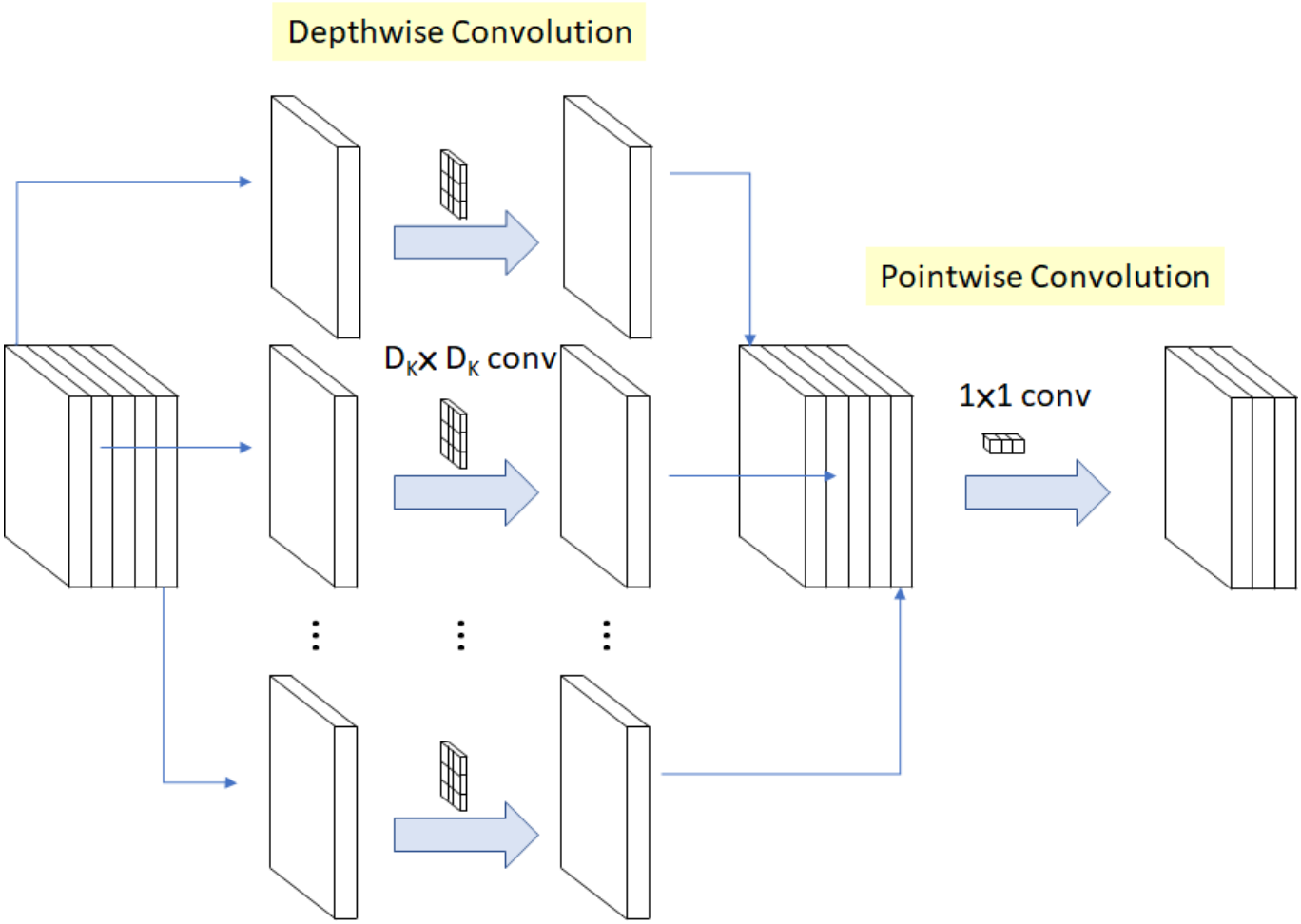
The above object detection example is the MobileNet which is actually amazing because it can achieve around 25 fps with such large amount of objects detected at the same time.

What Are Covered

1. **Depthwise Separable Convolution**
2. **Whole Network Architecture**
3. **Width Multiplier α for Thinner Models**
4. **Resolution Multiplier ρ for Reduced Representation**
5. **Comparison With State-of-the-art Approaches**

1. Depthwise Separable Convolution

Depthwise separable convolution is **a depthwise convolution followed by a pointwise convolution** as follows:



1. **Depthwise convolution** is the **channel-wise DK×DK spatial convolution**. Suppose in the figure above, we have 5 channels, then we will have 5 DK×DK spatial convolution.

2. **Pointwise convolution** actually is the **1×1 convolution** to change the dimension.

With above operation, the operation cost is:

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

where M: Number of input channels, N: Number of output channels, DK: Kernel size, and DF: Feature map size.

For standard convolution, it is:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$$

Standard Convolution Cost

Thus, the computation reduction is:

$$\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2}$$

Depthwise Separable Convolution Cost / Standard Convolution Cost

When DK×DK is 3×3, 8 to 9 times less computation can be achieved, but with only small reduction in accuracy.

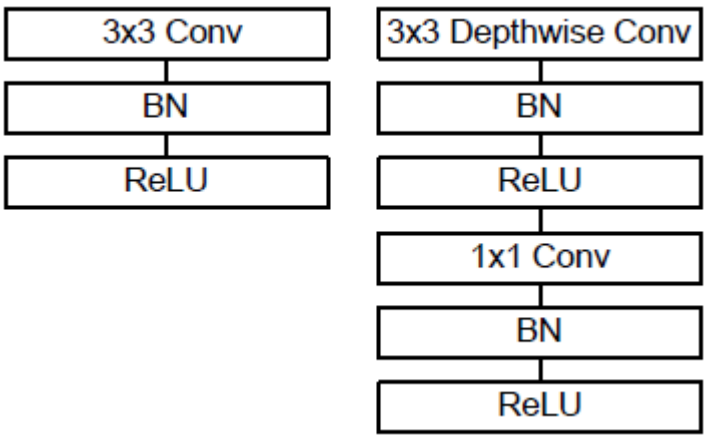
2. Whole Network Architecture

Below is the MobileNet Architecture:

Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Whole Network Architecture for MobileNet

It is noted that Batch Normalization (BN) and ReLU are applied after each convolution:



Standard Convolution (Left), Depthwise separable convolution (Right) With BN and ReLU

If Standard Convolution vs Depthwise Separable Convolution for ImageNet dataset:

	Accuracy	Mult-Adds	Parameters
Conv MobileNet	71.7%	4866	29.3
MobileNet	70.6%	569	4.2

Standard Convolution vs Depthwise Separable Convolution (ImageNet dataset)

MobileNet only got 1% loss in accuracy, but the Mult-Adds and parameters are reduced tremendously.

3. Width Multiplier α for Thinner Models

Width Multiplier α is introduced to **control the number of channels or channel depth**, which makes M become αM. And the depthwise separable convolution cost become:

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

Depthwise Separable Convolution Cost with Width Multiplier α

where α is between 0 to 1, with typical settings of 1, 0.75, 0.5 and 0.25. When α=1, it is the baseline MobileNet. And the computational cost and the number of parameters can be reduced quadratically by roughly α^2.

Table 6. MobileNet Width Multiplier

Width Multiplier	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
0.75 MobileNet-224	68.4%	325	2.6
0.5 MobileNet-224	63.7%	149	1.3
0.25 MobileNet-224	50.6%	41	0.5

Different Values of Width Multiplier α

Accuracy drops off smoothly from α=1 to 0.5 until α=0.25 which is too small.

network, with Resolution Multiplier ρ, the cost become:

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

Depthwise Separable Convolution Cost with Both Width Multiplier α and Resolution Multiplier ρ

where ρ is between 0 to 1. And the input resolution is 224, 192, 160, and 128. When ρ=1, it is the baseline MobileNet.

Table 7. MobileNet Resolution

Resolution	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
1.0 MobileNet-192	69.1%	418	4.2
1.0 MobileNet-160	67.2%	290	4.2
1.0 MobileNet-128	64.4%	186	4.2

Different Values of Resolution Multiplier ρ

Accuracy drops off smoothly across resolution from 224 to 128.

5. Comparison With State-of-the-art Approaches

When 1.0 MobileNet-224 is used, it outperforms GoogLeNet (Winner of ILSVRC 2014) and VGGNet (1st Runner Up of ILSVRC 2014) while the multi-adds and parameters are much fewer:

	Accuracy	Mult-Adds	Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

ImageNet Dataset

When smaller network, 0.50 MobileNet-160, is used, it outperforms Squeezenet and AlexNet (Winner of ILSVRC 2012) while the multi-adds and parameters are much fewer:

Table 9. Smaller MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
0.50 MobileNet-160	60.2%	76	1.32
Squeezenet	57.5%	1700	1.25
AlexNet	57.2%	720	60

ImageNet Dataset

It is also competitive with Inception-v3 (1st Runner Up in ILSVRC 2015) while the multi-adds and parameters are much fewer:

Table 10. MobileNet for Stanford Dogs

Model	Top-1 Accuracy	Million Mult-Adds	Million Parameters
Inception V3 [18]	84%	5000	23.2
1.0 MobileNet-224	83.3%	569	3.3
0.75 MobileNet-224	81.9%	325	1.9
1.0 MobileNet-192	81.9%	418	3.3
0.75 MobileNet-192	80.5%	239	1.9

Stanford Dogs Dataset

Many other datasets are also tried to prove the effectiveness of MobileNet:

Table 11. Performance of PlaNet using the MobileNet architecture. Percentages are the fraction of the Im2GPS test dataset that were localized within a certain distance from the ground truth. The numbers for the original PlaNet model are based on an updated version that has an improved architecture and training dataset.

Scale	Im2GPS [7]	PlaNet [35]	PlaNet MobileNet
Continent (2500 km)	51.9%	77.6%	79.3%
Country (750 km)	35.4%	64.0%	60.3%
Region (200 km)	32.1%	51.1%	45.2%
City (25 km)	21.9%	31.7%	31.7%
Street (1 km)	2.5%	11.0%	11.4%

GPS Localization Via Photos

Table 12. Face attribute classification using the MobileNet architecture. Each row corresponds to a different hyper-parameter setting (width multiplier α and image resolution).

Width Multiplier / Resolution	Mean AP	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	88.7%	568	3.2
0.5 MobileNet-224	88.1%	149	0.8
0.25 MobileNet-224	87.2%	45	0.2
1.0 MobileNet-128	88.1%	185	3.2
0.5 MobileNet-128	87.7%	48	0.8
0.25 MobileNet-128	86.4%	15	0.2
Baseline	86.9%	1600	7.5

Face Attribute Classification

COCO primary challenge metric (AP at IoU=0.50:0.05:0.95)

Framework Resolution	Model	mAP	Billion Mult-Adds	Million Parameters
SSD 300	deeplab-VGG	21.1%	34.9	33.1
	Inception V2	22.0%	3.8	13.7
	MobileNet	19.3%	1.2	6.8
Faster-RCNN 300	VGG	22.9%	64.3	138.5
	Inception V2	15.4%	118.2	13.3
	MobileNet	16.4%	25.2	6.1
Faster-RCNN 600	VGG	25.7%	149.6	138.5
	Inception V2	21.9%	129.6	13.3
	Mobilenet	19.8%	30.5	6.1

Microsoft COCO Object Detection Dataset

Figure 6. Example objection detection results using MobileNet SSD.

MobileNet SSD

	Accuracy	Mult-Adds	Parameters
FaceNet [25]	83%	1600	7.5
1.0 MobileNet-160	79.4%	286	4.9
1.0 MobileNet-128	78.3%	185	5.5
0.75 MobileNet-128	75.2%	166	3.4
0.75 MobileNet-128	72.5%	108	3.8

Face Recognition

To conclude, similar performance with state-of-the-art approaches but with much smaller network is achieved using MobileNet, favored by Depthwise Separable Convolution.

Indeed, there are still many applications I haven't mentioned above, like GPS Localization Via Photos, Face Attribute Classification and Face Recognition. Hope I can cover them in the coming future. :)

References

1. [2017 arXiv] [MobileNetV1]
   MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

My Reviews

[Inception-v3] [BatchNorm] [GoogLeNet] [VGGNet] [AlexNet]