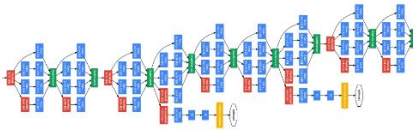
YOLO MobileNet

Vladislav Goncharenko
Fall 2023
MSAI

Recap

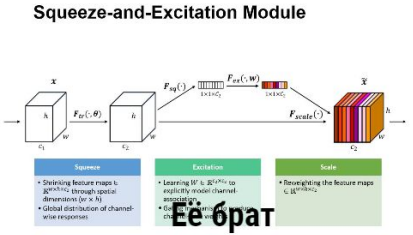
- Computer Vision tasks
  - Classification
  - Localisation
  - Detection
  - Semantic Segmentation
  - Instance Segmentation
- Metrics
  - IoU
  - mAP
- Datasets
  - PASCAL VOC
  - ImageNet
  - COCO
  - Open Images
- R-CNN
- Fast R-CNN
- Faster R-CNN

Девушка, которая мне нравится

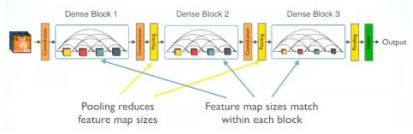


Её отец

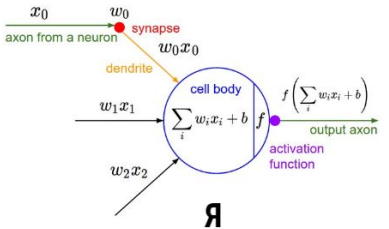
Её мать



Её брат



Её бывший

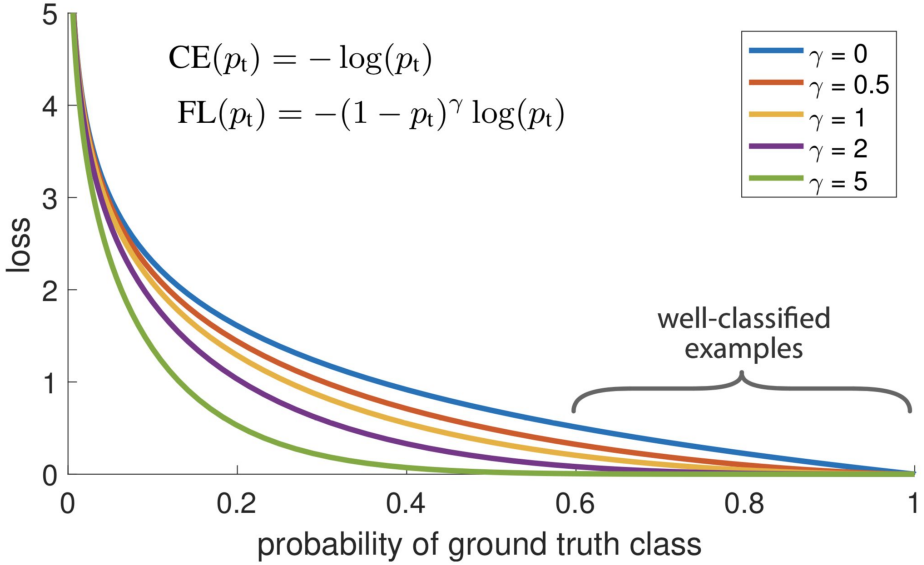


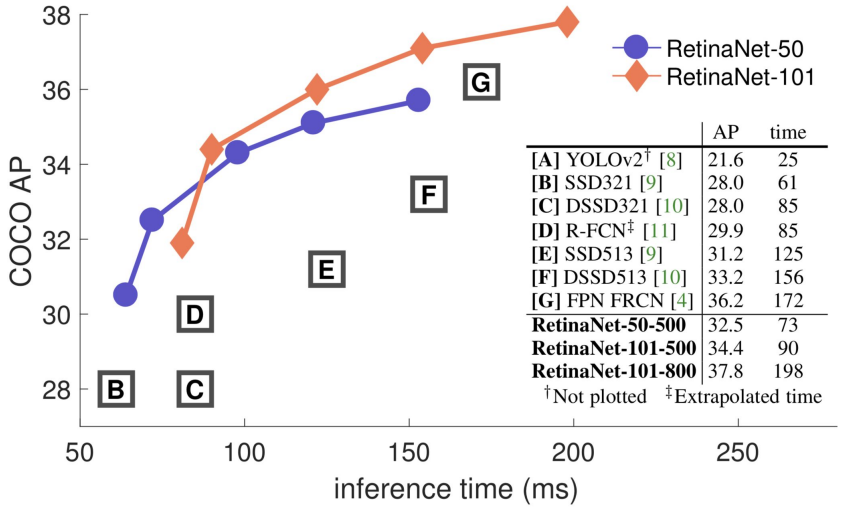
Я

Outline

- Focal Loss
- Non-Maximum Suppression
- YOLO
  - v1
  - v2
  - V3
- Separable convolutions
- MobileNet

Focal Loss



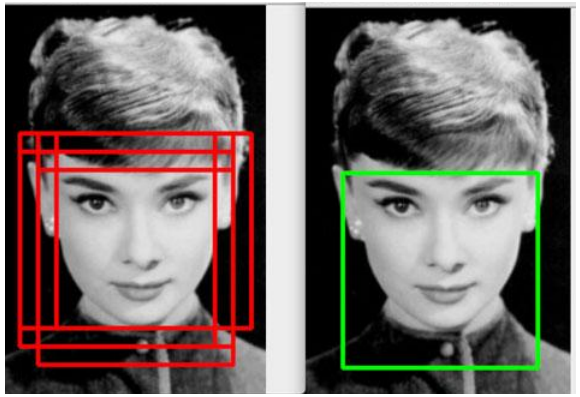


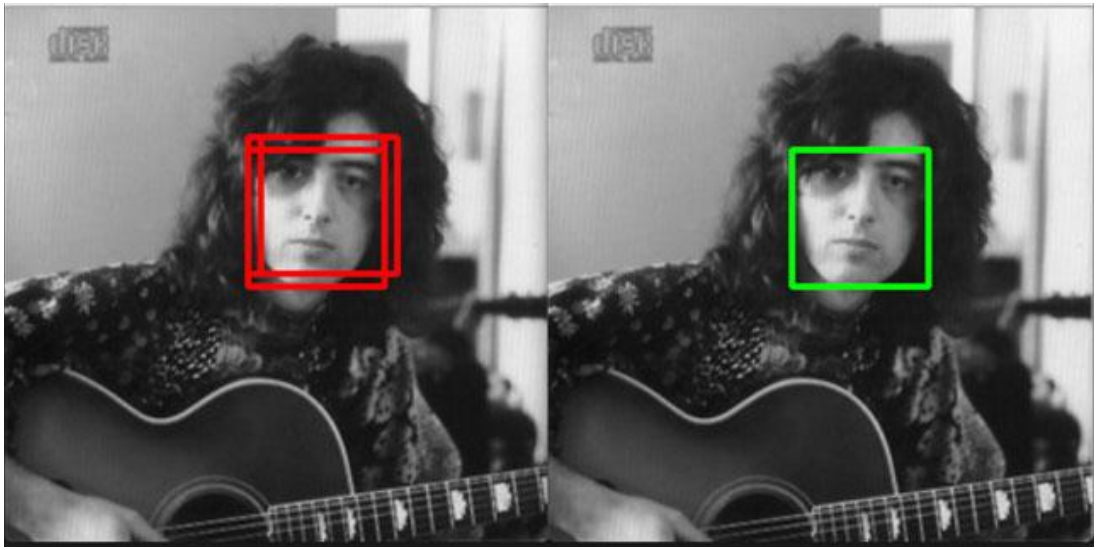
	AP	time
[A] YOLOv2† [8]	21.6	25
[B] SSD321 [9]	28.0	61
[C] DSSD321 [10]	28.0	85
[D] R-FCN‡ [11]	29.9	85
[E] SSD513 [9]	31.2	125
[F] DSSD513 [10]	33.2	156
[G] FPN FRCN [4]	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198

†Not plotted ‡Extrapolated time

Video talk

Non-Maximum Suppression





You Only Look Once (YOLO)

YOLOv1-v2

Original Joseph Redmon slides from conferences

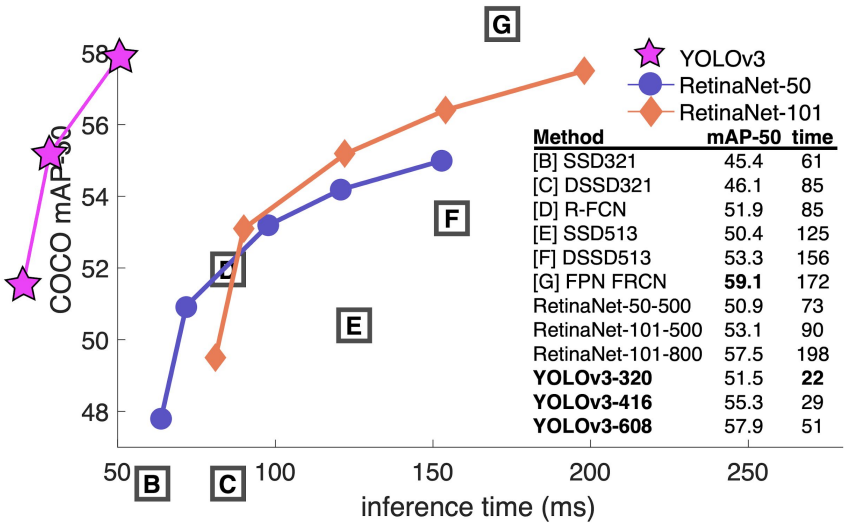
- YOLOv1
- YOLOv2

Also CVPR 2016 talk video

YOLOv3

- Bounding Box Prediction
- Class Prediction
- Predictions Across Scales
- Feature Extractor

All articles published on [Joseph's website]

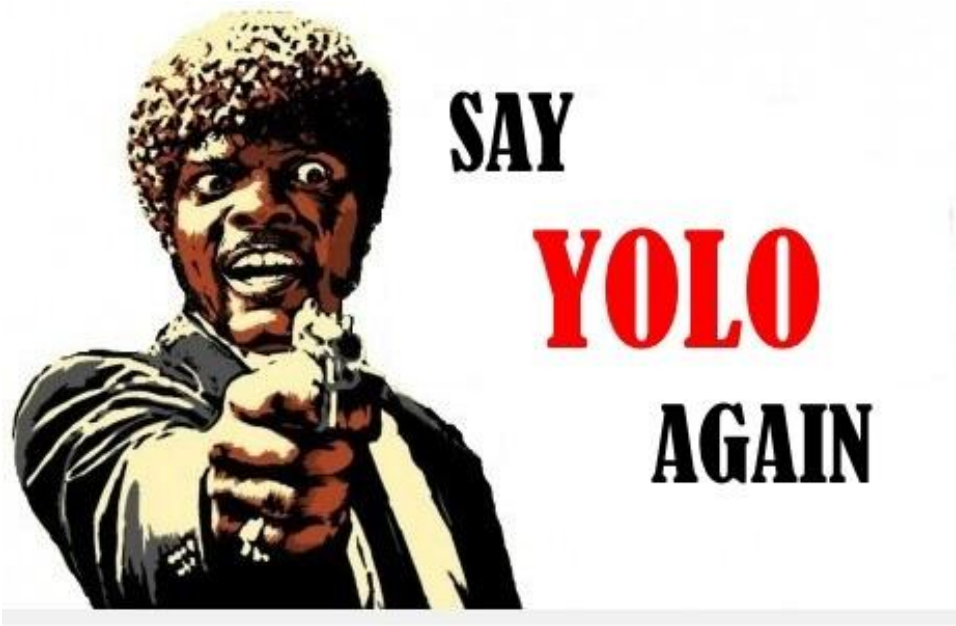


Practical notes

Write your own YOLO from scratch
https://blog.paperspace.com/how-to-implement-a-yolo-object-detector-in-pytorch/

Original YOLO written on Darknet - custom NN framework

Weights import to TF via
https://github.com/thtrieu/darkflow



Current state

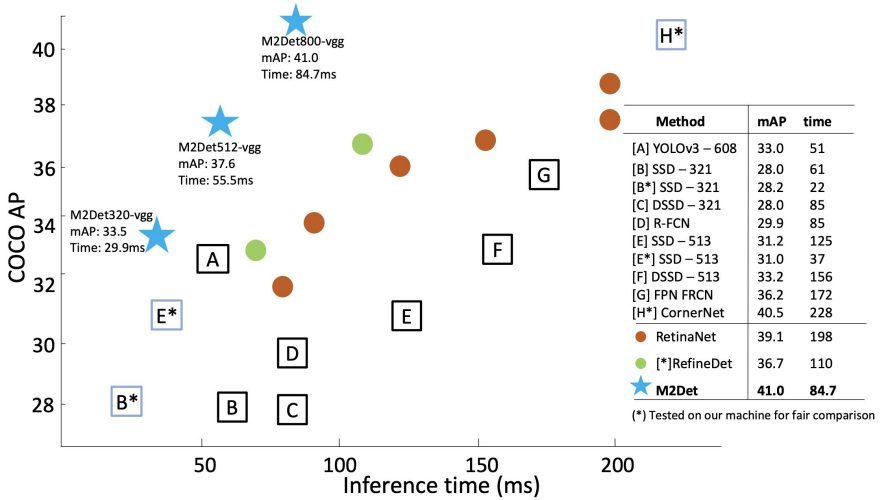


Figure 5: Speed (ms) vs. accuracy (mAP) on COCO *test-dev*.

M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network

https://arxiv.org/pdf/1811.04533.pdf

Who is responsibile for all these YOLOs

https://docs.google.com/spreadsheets/d/1BxI0SkyFGNvfzLLM_46xZ3i5crTst7jsUKKuZZDWrA0/edit#gid=0

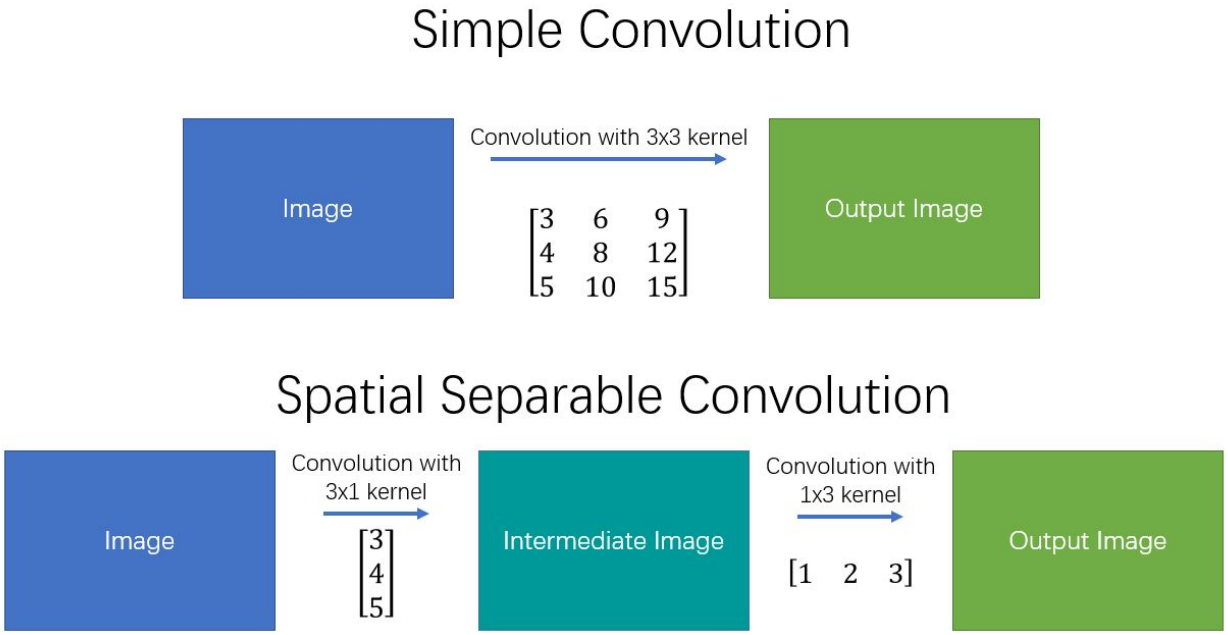
Separable convs
MobileNet

Spatial Separable Convolution

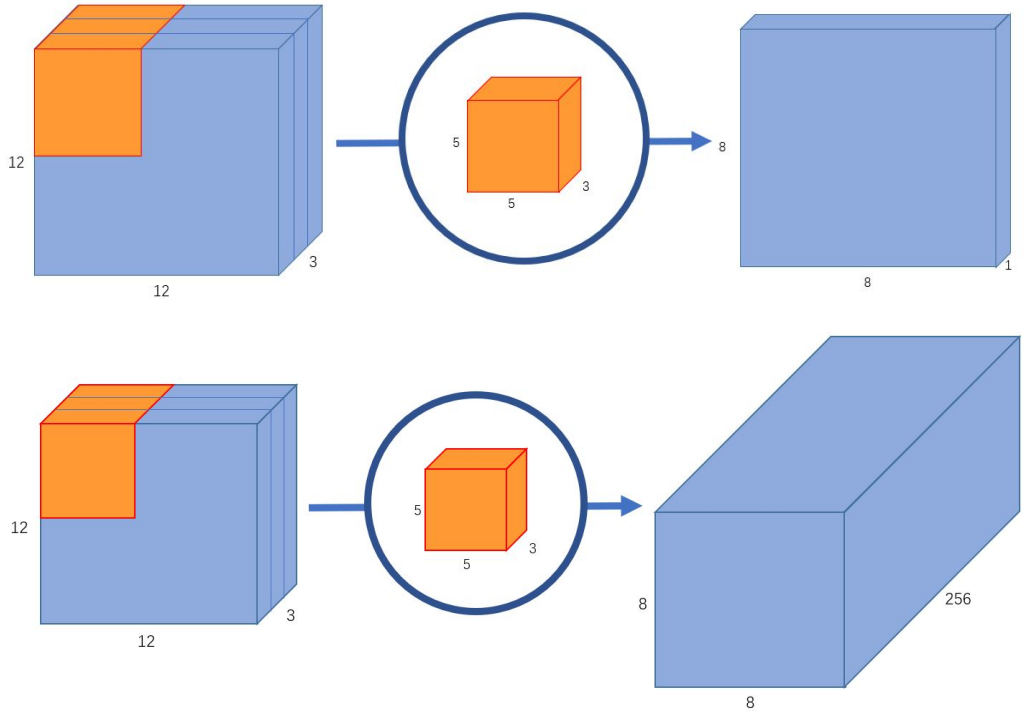
$$\begin{bmatrix} 3 & 6 & 9 \\ 4 & 8 & 12 \\ 5 & 10 & 15 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

Images credentials

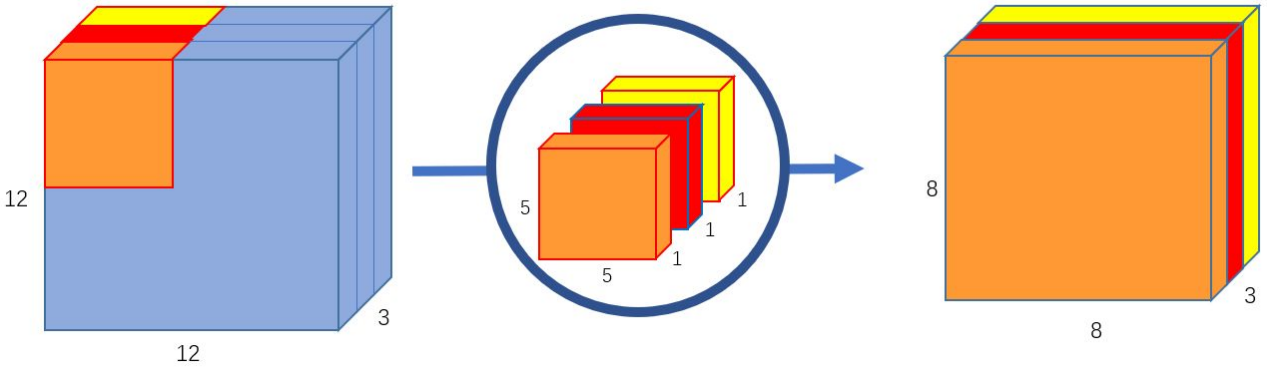
Spatial Separable Convolution



Normal Convolution

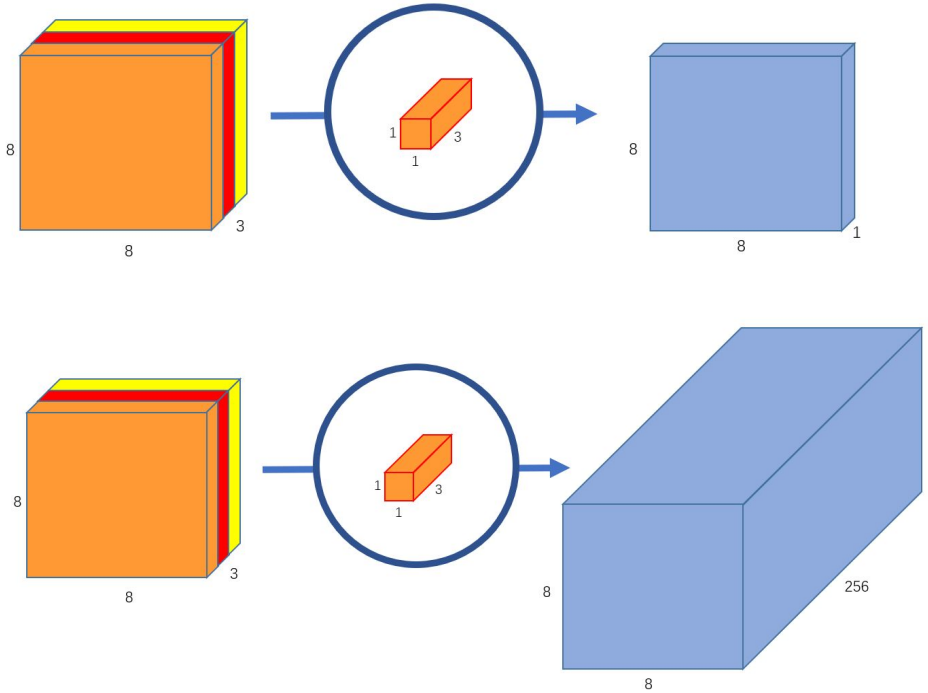


Depthwise Convolution



Don’t change channels number

Pointwise Convolution



MobileNet v1

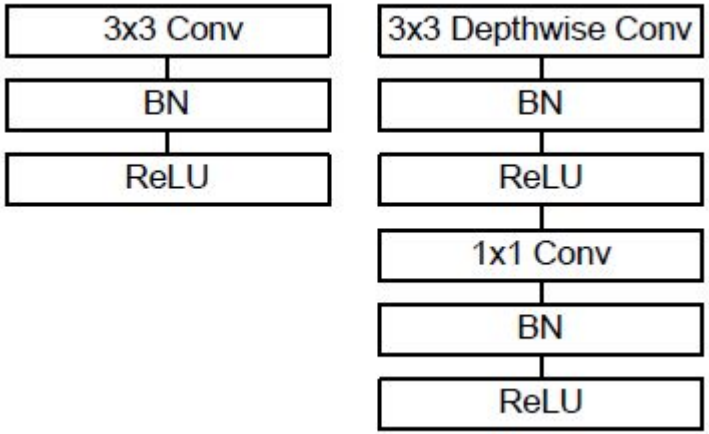
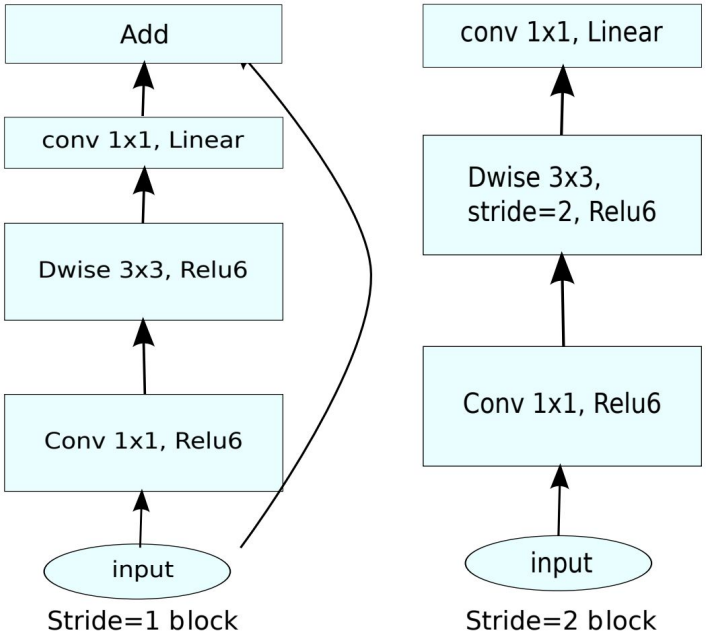
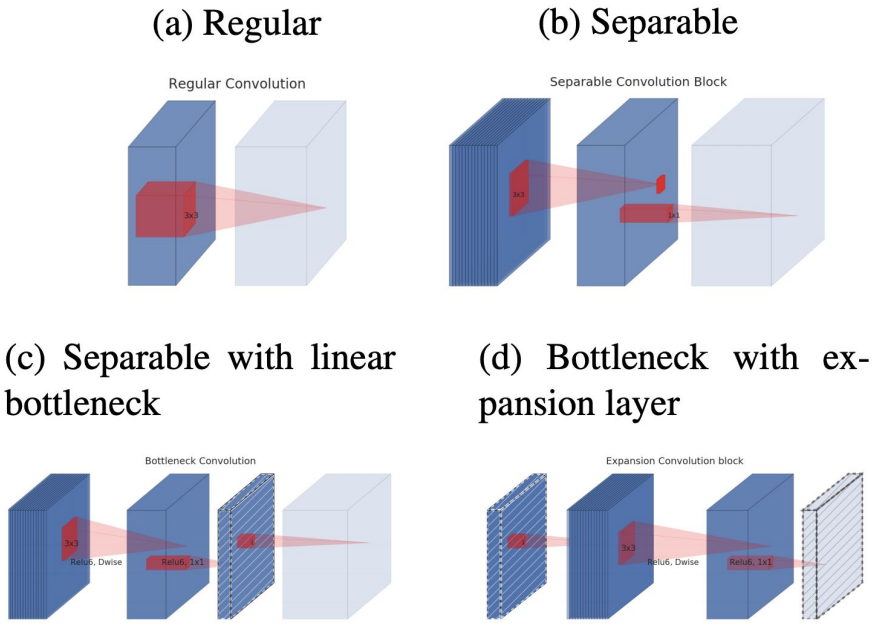


Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
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	$1 \times 1 \times 1000$

MobileNet v2





Results

Архитектура сети	Количество параметров	Top-1 accuracy	Top-5 accuracy
Xception	22.91M	0.790	0.945
VGG16	138.35M	0.715	0.901
MobileNetV1 (alpha=1, rho=1)	4.20M	0.709	0.899
MobileNetV1 (alpha=0.75, rho=0.85)	2.59M	0.672	0.873
MobileNetV1 (alpha=0.25, rho=0.57)	0.47M	0.415	0.663
MobileNetV2 (alpha=1.4, rho=1)	6.06M	**0.750**	**0.925**
MobileNetV2 (alpha=1, rho=1)	3.47M	0.718	0.910
MobileNetV2 (alpha=0.35, rho=0.43)	1.66M	0.455	0.704

Revise

- Recap
- Focal Loss
- Non-Maximum Suppression
- YOLO
- Separable convolutions
- MobileNet

Next time

- Segmentation

See [Mask R-CNN ICCV17 talk] to prepare yourself