embedded VISION SUMMIT 2018

Building Efficient CNN Models for Mobile and Embedded Vision Applications

facebook

Peter Vajda

May 23, 2018

Challenges

Higher Accurac y

Faster Speed

Smaller Size

Smaller Memory

Outline

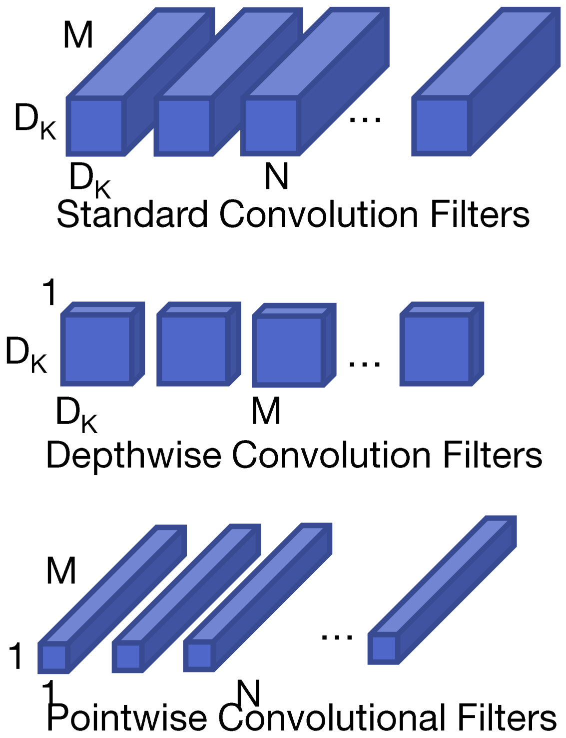
- Model Architecture
- Model Compression
- Realtime style transfer
- Realtime pose estimation

Copyright © 2018 Facebook

MobileNet

- MobileNet [Howard et al. 2017]
- Depthwise separable convolution
  - Depthwise convolution
  - Pointwise convolution
- Hyper-parameters to trade off between latency and accuracy
  - Width multiplier
  - Resolution multiplier

Convolution | MAC



MAC:

$D_K D_K M N D_F D_F$

$D_K D_K M D_F D_F$

$M N D_F D_F$

D_k – kernel size
D_f – featuremap size

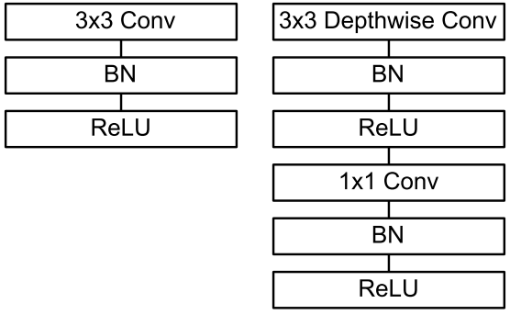
Copyright © 2018 Facebook

MobileNet

- MobileNet [Howard et al. 2017]

- Depthwise separable convolution
  - Depthwise convolution
  - Pointwise convolution

- Hyper-parameters to trade off between latency and accuracy
  - Width Multiplier
  - Resolution Multiplier

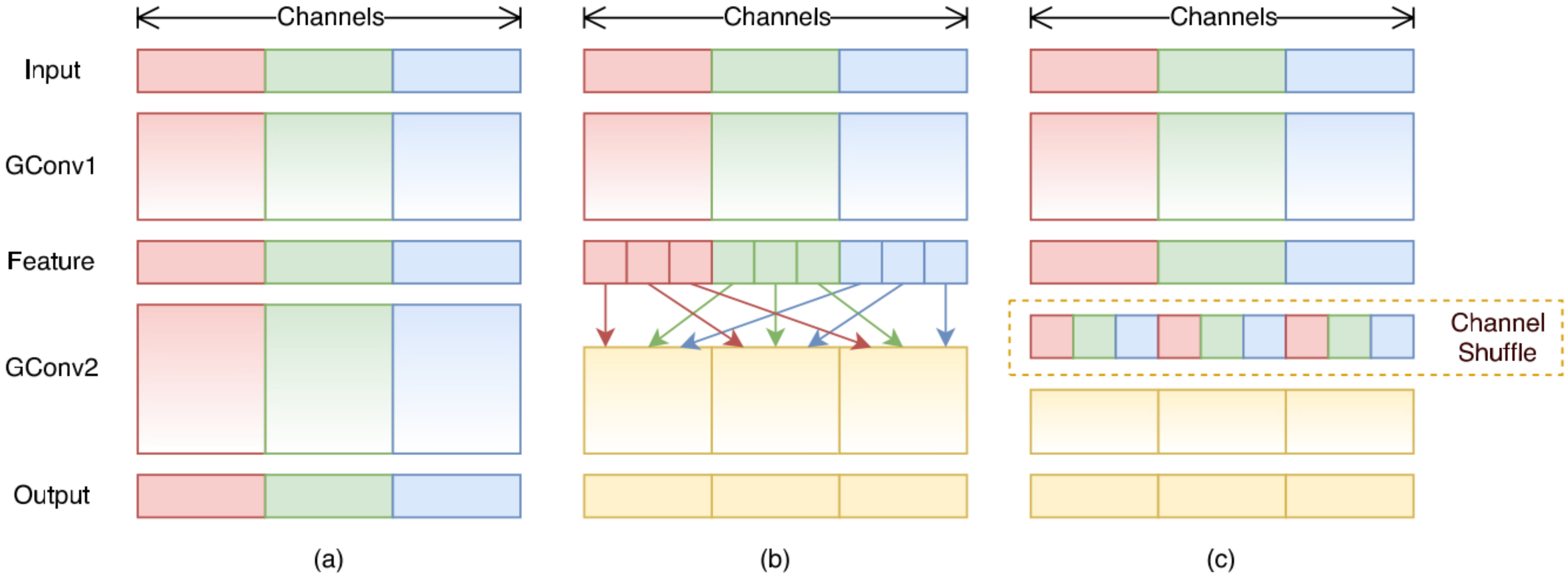


Type / Stride	Filter Shape
Conv / s2	$3 \times 3 \times 3 \times 32$
Conv dw / s1	$3 \times 3 \times 32$ dw
Conv / s1	$1 \times 1 \times 32 \times 64$
Conv dw / s2	$3 \times 3 \times 64$ dw
Conv / s1	$1 \times 1 \times 64 \times 128$
Conv dw / s1	$3 \times 3 \times 128$ dw
Conv / s1	$1 \times 1 \times 128 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw
Conv / s1	$1 \times 1 \times 128 \times 256$
Conv dw / s1	$3 \times 3 \times 256$ dw
Conv / s1	$1 \times 1 \times 256 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw
Conv / s1	$1 \times 1 \times 256 \times 512$
$5\times$ Conv dw / s1 Conv / s1	$3 \times 3 \times 512$ dw $1 \times 1 \times 512 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw
Conv / s1	$1 \times 1 \times 512 \times 1024$
Conv dw / s2	$3 \times 3 \times 1024$ dw
Conv / s1	$1 \times 1 \times 1024 \times 1024$
Avg Pool / s1	Pool 7×7
FC / s1	1024×1000
Softmax / s1	Classifier

Copyright © 2018 Facebook

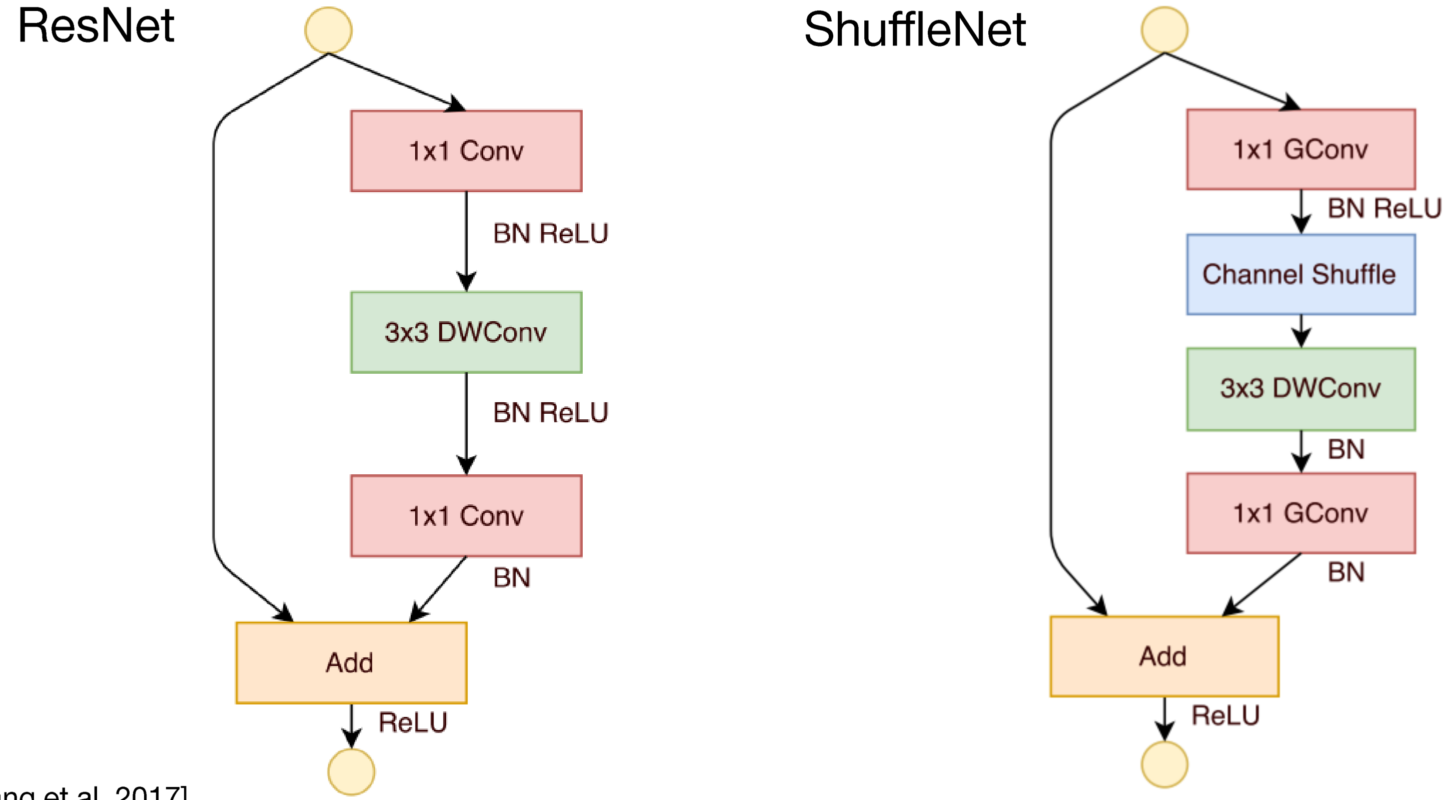
ShuffleNet

- Pointwise convolution is very expensive $M \cdot N \cdot D_F \cdot D_F \gg D_K \cdot D_K \cdot M \cdot D_F \cdot D_F$



ShuffletNet [Zhang et al. 2017]

ShuffleNet

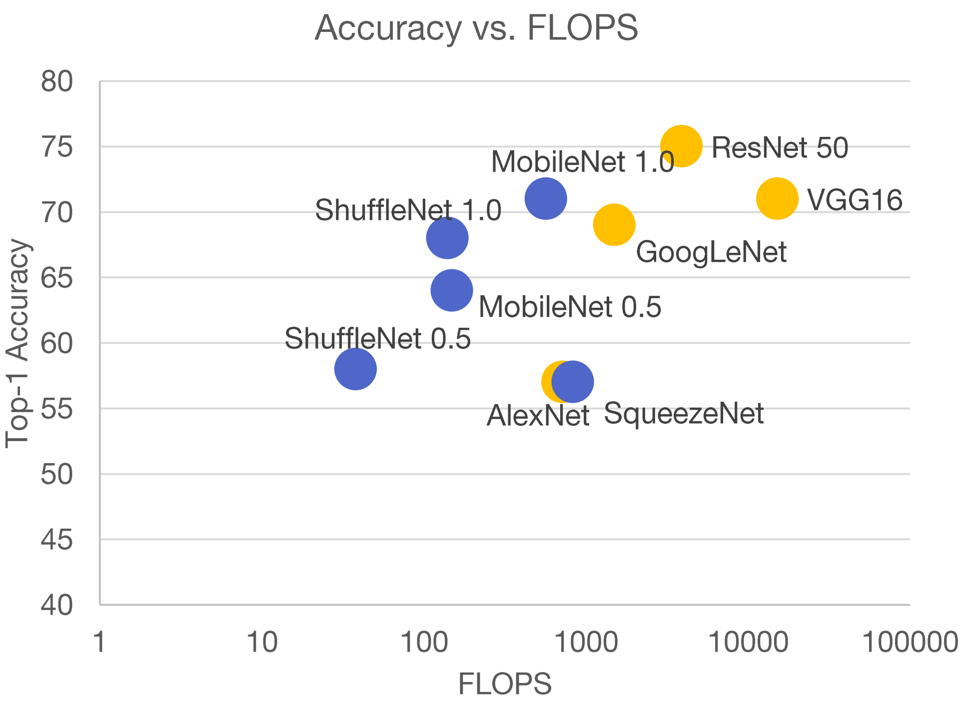


ShuffletNet [Zhang et al. 2017]

Copyright © 2018 Facebook

embedded VISION SUMMIT 2018

Accuracy vs. FLOPS



Model	FLOPS	Top-1 Accuracy [%]
AlexNet	720	57
VGG 16	15300	71
GoogLeNet	1500	69
Resnet 50	3900	75
SqueezeNet	833	57
MobileNet 1.0 - 224	569	71
MobileNet 0.5 - 224	149	64
ShuffleNet 1.0	140	68
ShuffleNet 0.5	38	58

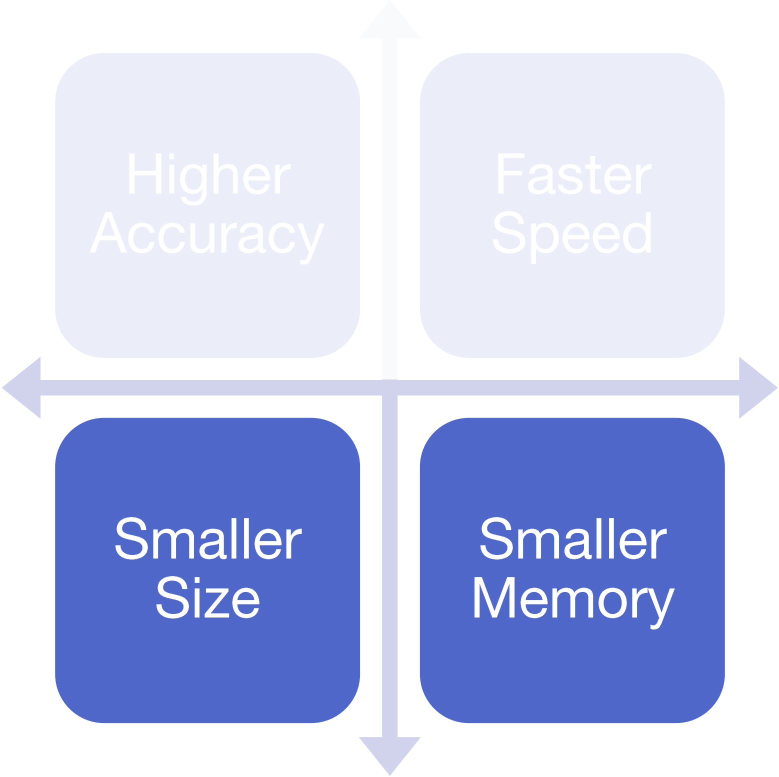
Full models

Mobile models

facebook

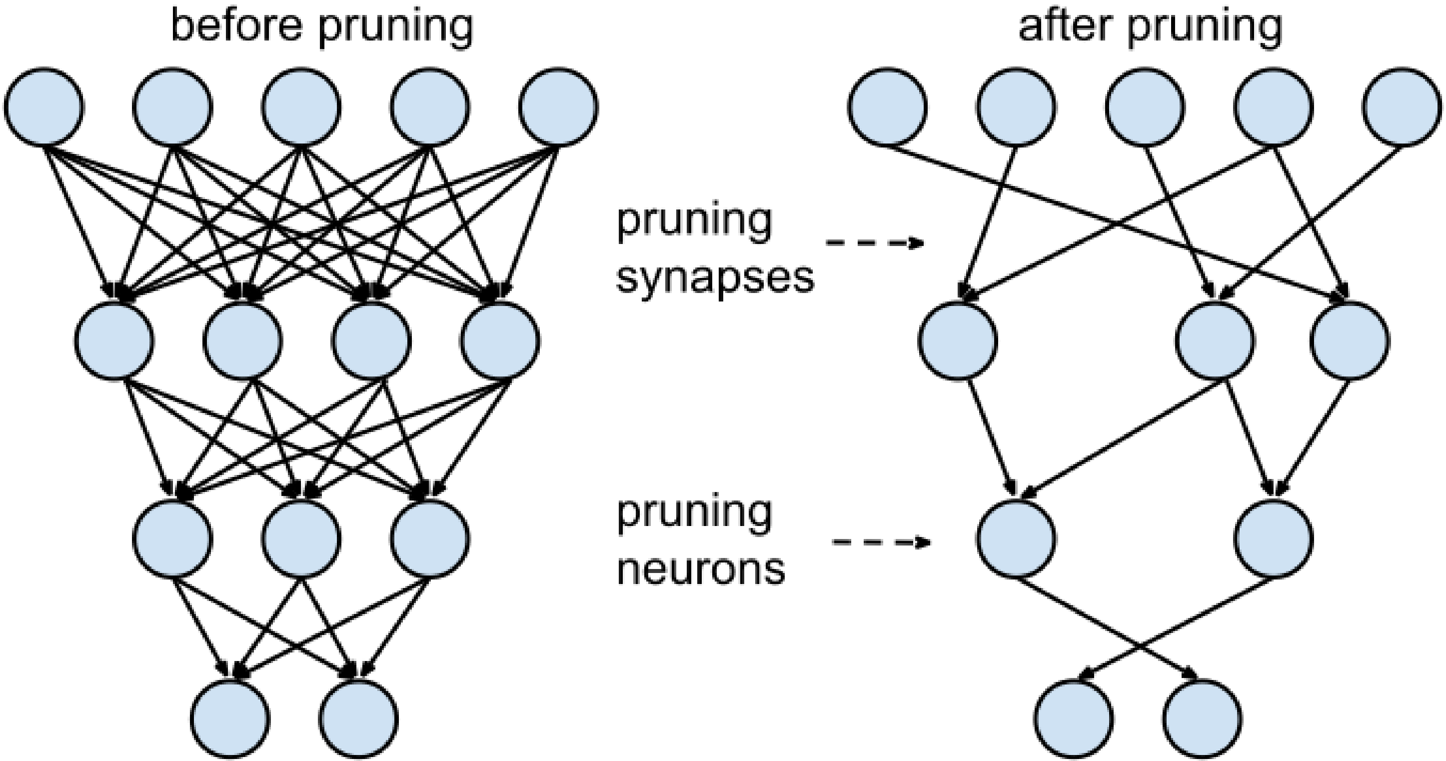
Copyright © 2018 Facebook

Model Compression



Copyright © 2018 Facebook

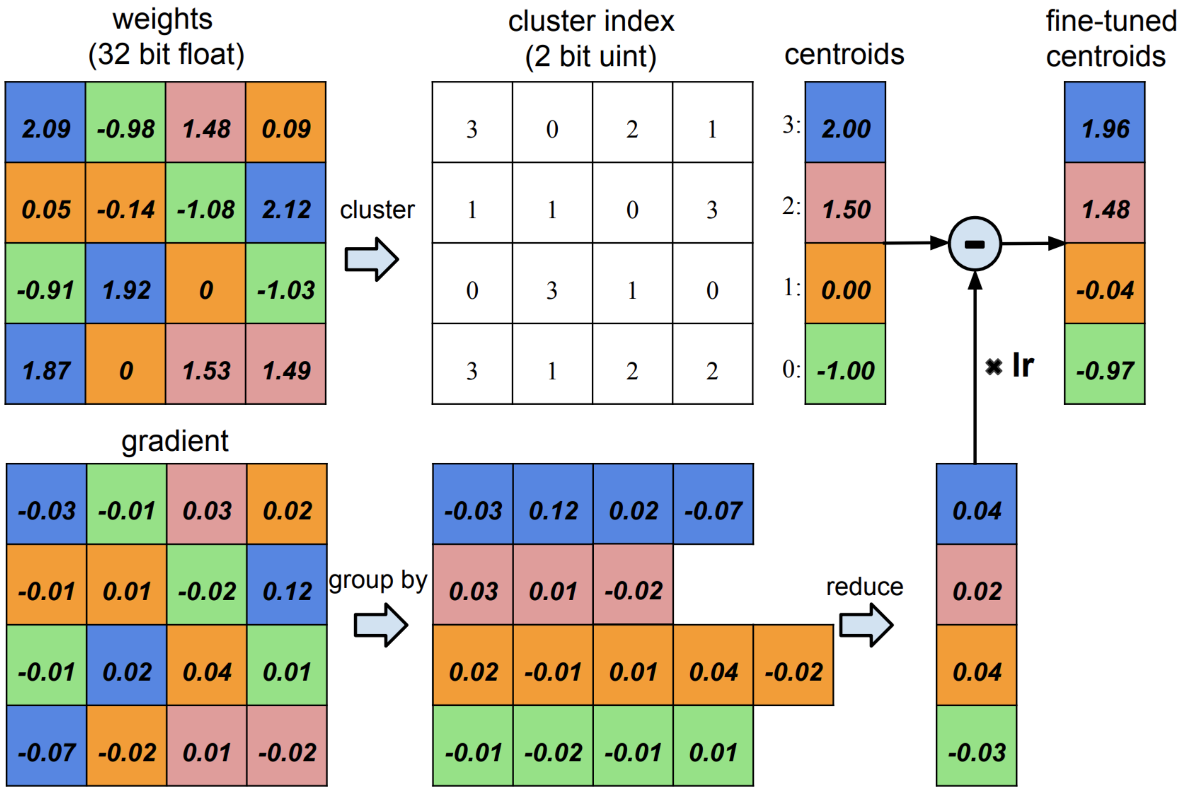
Pruning



[Han et al. NIPS 2015]

Copyright © 2018 Facebook

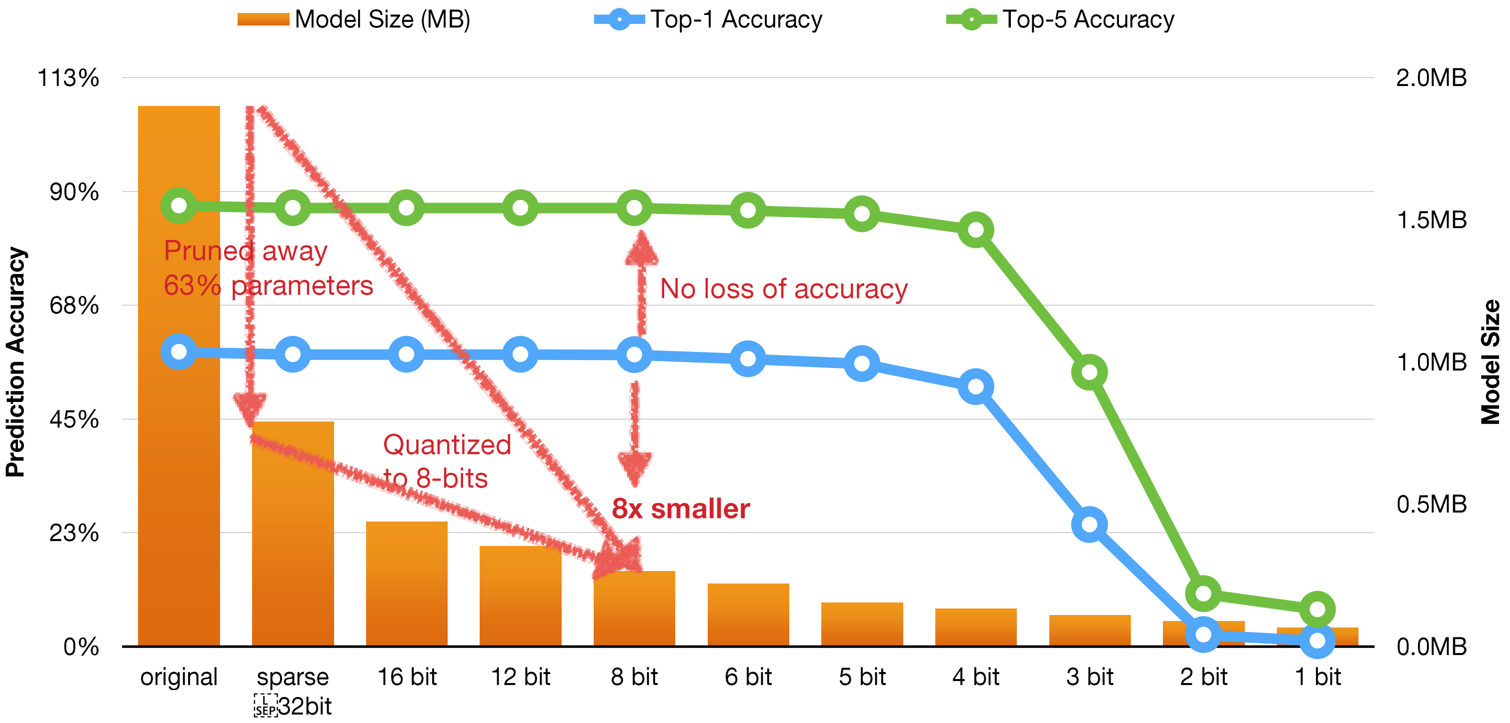
Quantization



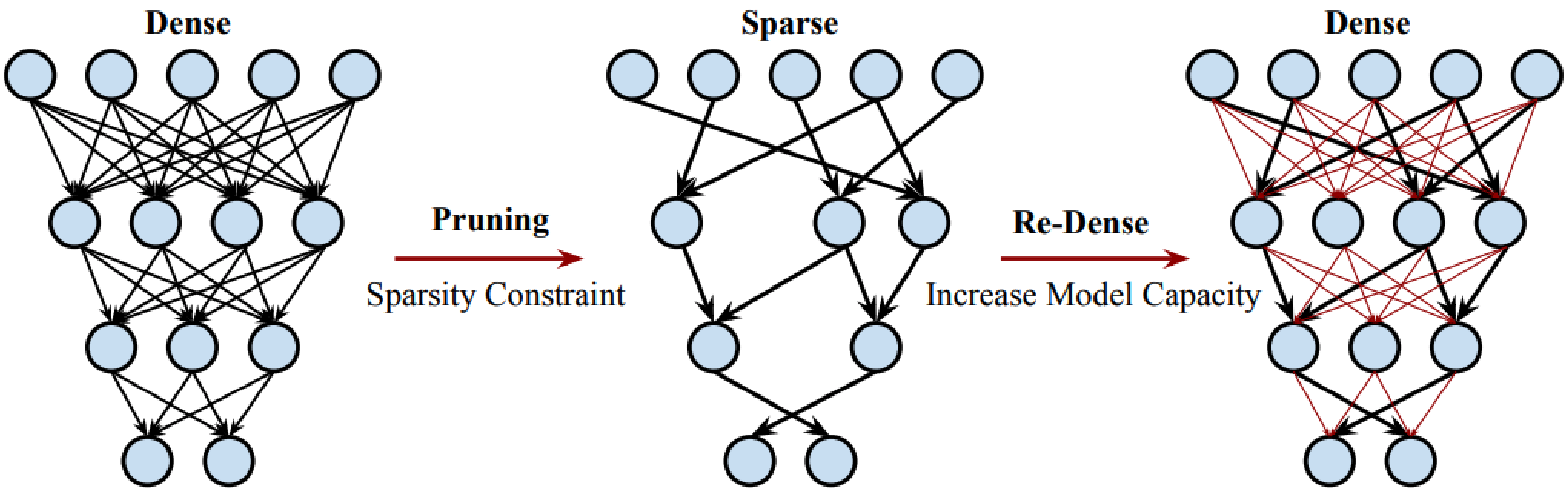
[Han et al. ICLR 2016]

Copyright © 2018 Facebook

Pruning + Quantization



Dense-Sparse-Dense



Copyright © 2018 Facebook

Dense-Sparse-Dense

Neural Network	Top-1 error	DSD training – Top-1 error
GoogLeNet	31.1%	**30.0%**
VGG-16	31.5%	**27.2%**
ResNet-18	30.4%	**29.2%**
ResNet-50	24.0%	**22.8%**

Copyright © 2018 Facebook

Summary of Model Optimization Approaches

Model

Faster Speed

- Latency-aware model design
- Automatic architecture search
- Parallelize graph branches
- Low precision network

Smaller Size

- Pruning: remove 60%-70% redundant weights
- Weight Sharing: quantize to 8-bits

Higher Accuracy

- Dense-Sparse-Dense Regularization
- Distillation

Copyright © 2018 Facebook

Realtime Style Transfer

Copyright © 2018 Facebook

Realtime Style Transfer

- Gatys et al. 2015

c

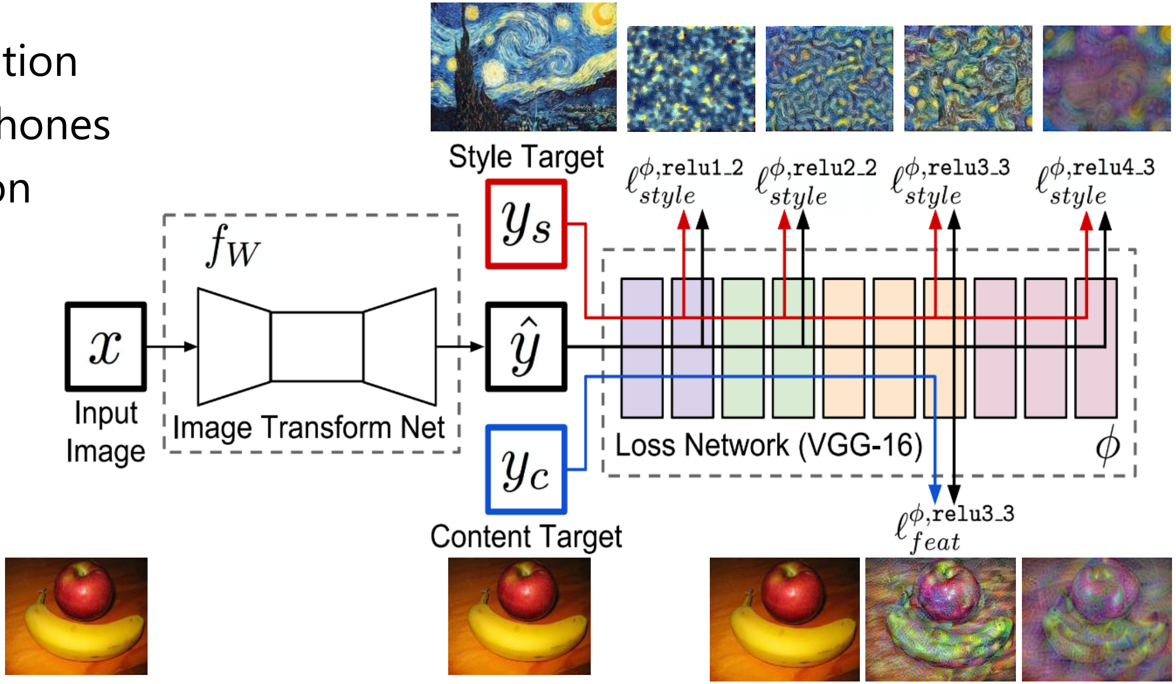
$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left(\alpha \mathcal{L}_{\text{content}}(\mathbf{c}, \mathbf{x}) + \beta \mathcal{L}_{\text{style}}(\mathbf{s}, \mathbf{x}) \right)$$

facebook

Copyright © 2018 Facebook

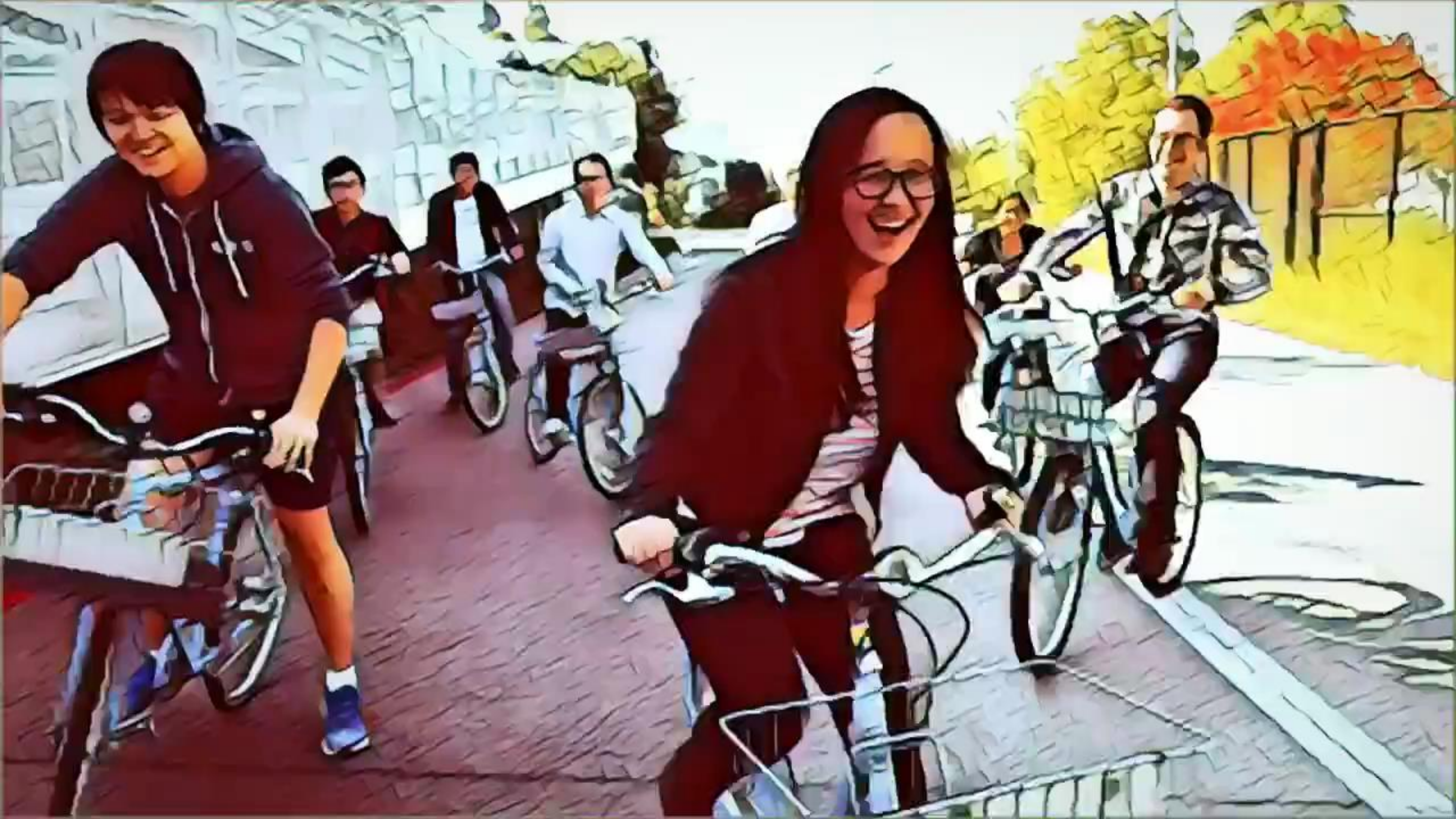
Realtime Style Transfer

- Caffe2Go
  - NNPack, NEON optimization
  - Real time on high-end phones
  - Specific layer optimization
- Model
  - Reduce layer number
  - Reduce channel number
  - Model compression
  - Pruning
- Quality
  - A/B test



[Justin et al. ECCV 2016]

Copyright © 2018 Facebook



Realtime Pose Estimation

Copyright © 2018 Facebook

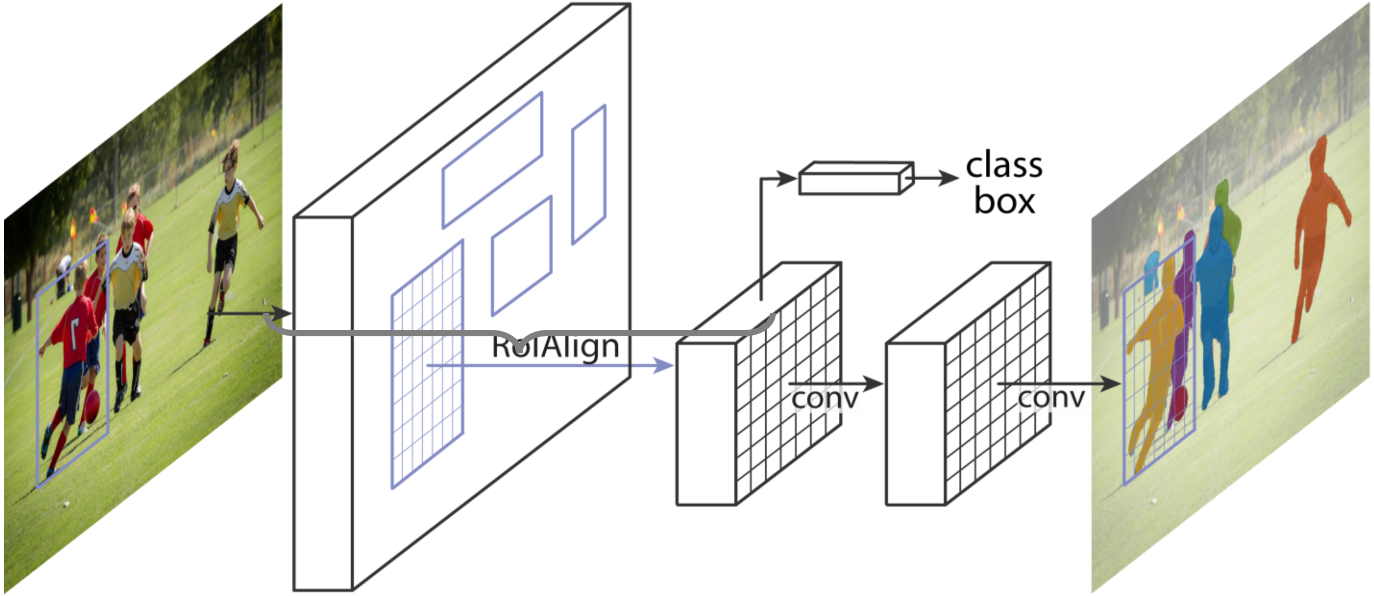
Realtime pose estimation

- Open-sourced optimized layers:
  - GenerateProposalsOp
  - BBoxTransformOp
  - BoxWithNMSLimit
  - RoIAlignOp
- Use-cases
  - Gesture recognition
  - AR effects
  - Avatar for AR/VR
- Caffe2Go
  - Mobile CPU, GPU, DSP

Images and annotations from COCO keypoint challenge

[Kaiming et al. ICCV 2017]

Realtime pose estimation



CNN applied to
entire image

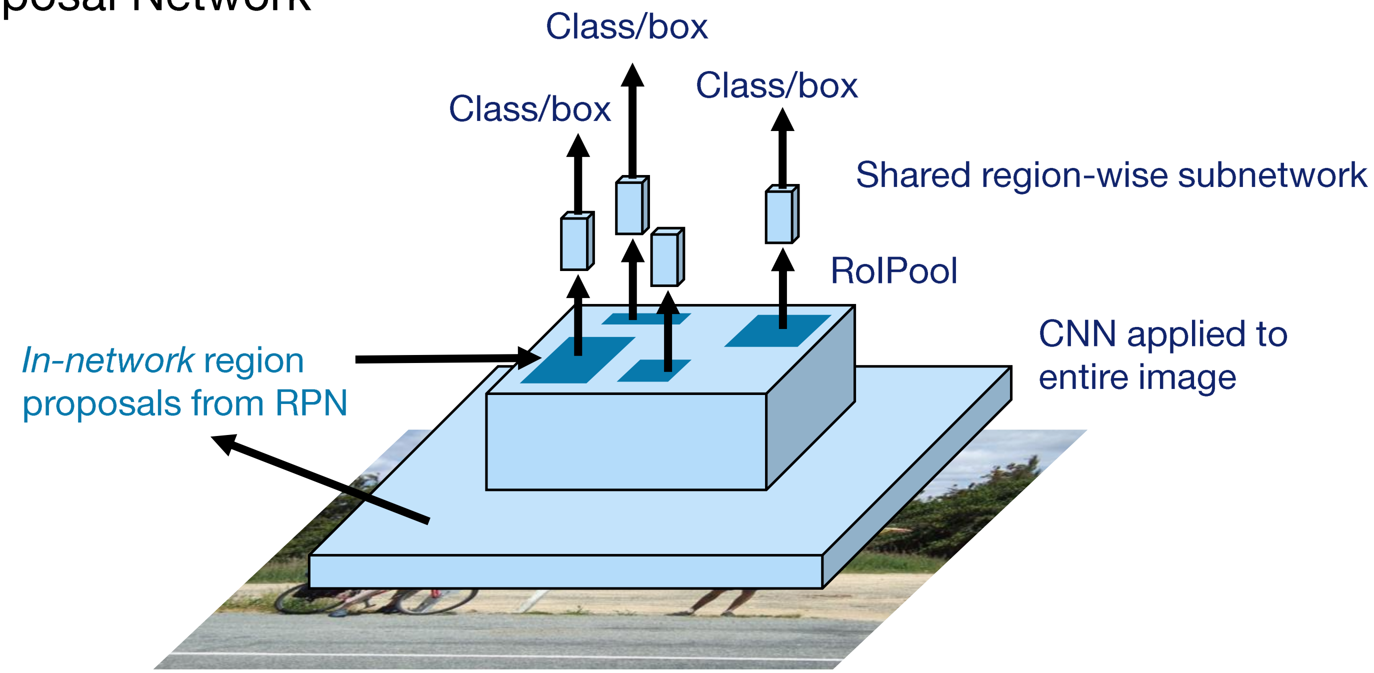
Mask “head”

[Kaiming et al. ICCV 2017]

Copyright © 2018 Facebook

Realtime pose estimation

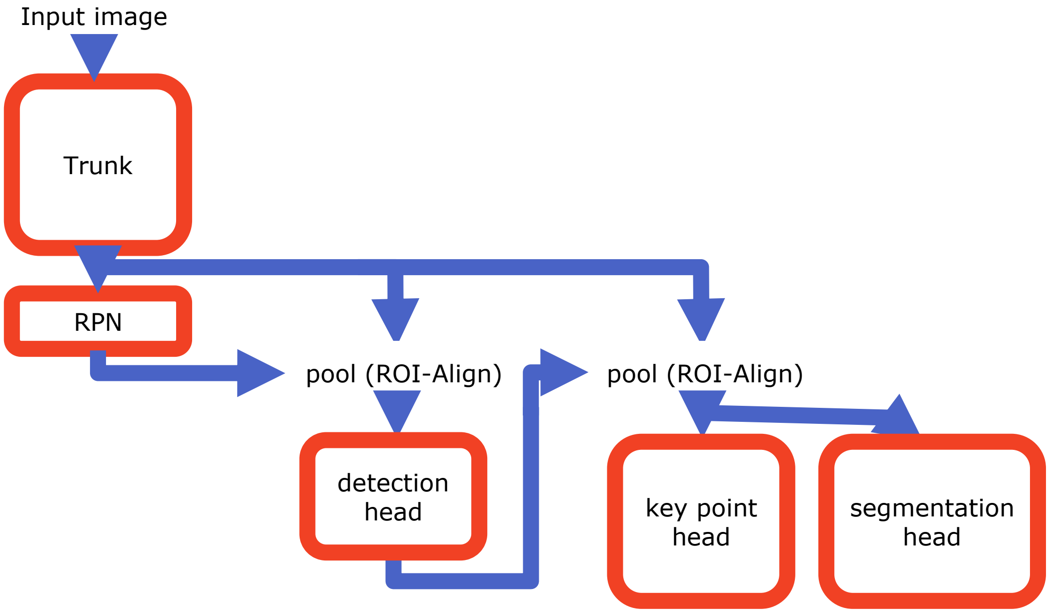
- Region Proposal Network



Copyright © 2018 Facebook

Realtime pose estimation

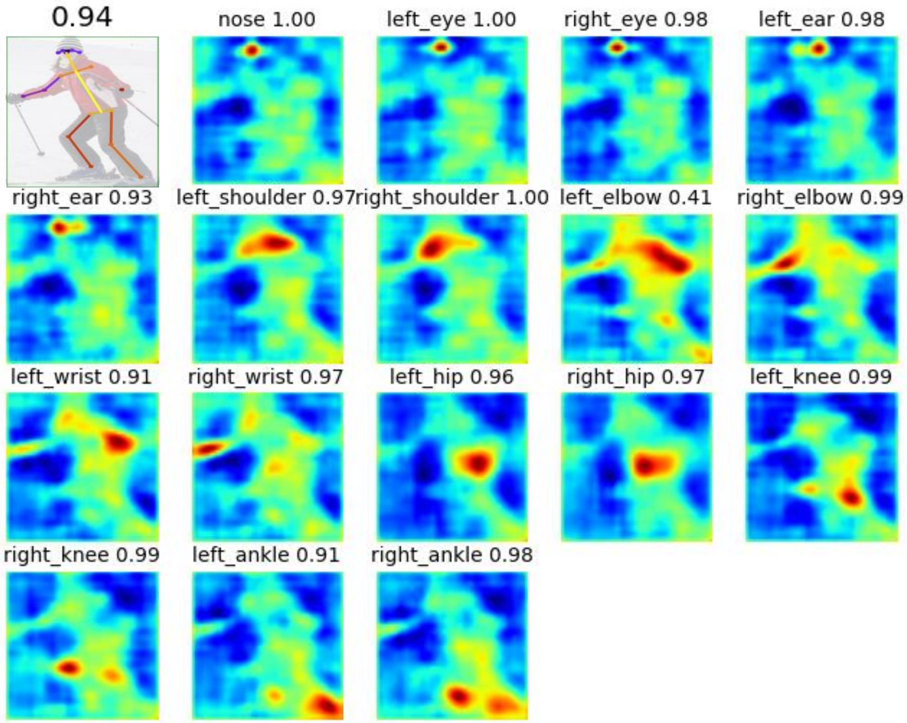
- Mask R-CNN2Go open-sourced layers
- Real-time on high-end phones
- Model
  - Reduce layer number
  - Reduce channel number
  - Model compression
  - Pruning
  - Split prev. nets to trunk, head
- Balanced blocks
  - Small detection head
  - Small number of proposals
  - High input resolution



Copyright © 2018 Facebook

Realtime pose estimation

- Challenges
  - Data, data, data…
  - Parameter search and optimization
  - Corner cases
  - Real-time benchmark ~~MAC~~
- Keypoint = 1-hot mask
- Represent pose as 17 masks



Mask R-CNN2Go

Future work

- More efficient model architectures
- High-performance inference
- Leveraging CPU, GPU and DSP on the phone

Copyright © 2018 Facebook

Take - away

- **Data, Data, Data, …**
- Diverse hardware support for high/low end iOS/Android. Open sourced Caffe2/Pytorch 1.0 efficient implementation on CPU/GPU/DSP on phone.
- Efficient model architecture balanced optimization on hardware
  - Accuracy
  - Model/Memory size
  - Speeeeeed
- Download models from cloud and run on mobile

+ Realtime interaction
+ Low/No bandwidth needs
+ Works offline
+ Privacy

- Battery
- Low-end phone support
- Limited accuracy and data capacity

Copyright © 2018 Facebook

Resource Slide

- Peter Vajda: vajdap@fb.com
  - https://research.fb.com/people/vajda-peter/
- https://www.facebook.com/

- https://code.facebook.com/posts/196146247499076/delivering-real-time-ai-in-the-palm-of-your-hand/ google: caffe2go
- https://research.fb.com/enabling-full-body-ar-with-mask-r-cnn2go/ google: maskrcnn2go

Copyright © 2018 Facebook