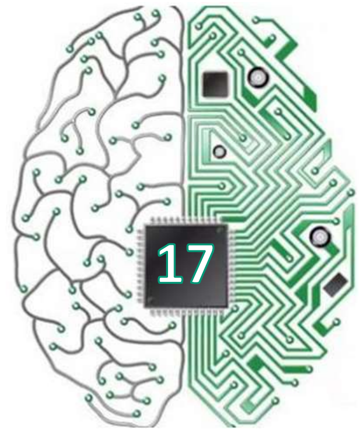
Open Elective Course [OE]
Course Code: CSO507
Winter 2023-24

Lecture#



Deep Learning

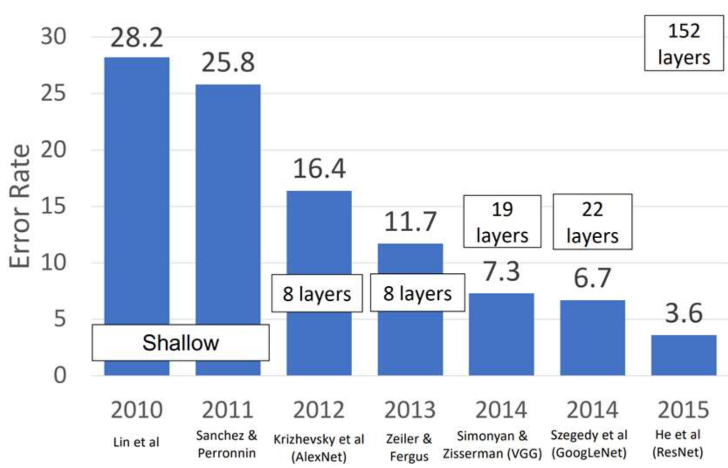
Unit-4: Convolutional Neural Networks (Part-V)

Course Instructor:

Dr. Monidipa Das
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology (Indian School of Mines) Dhanbad, Jharkhand 826004, India

ResNet

ImageNet Classification Challenge

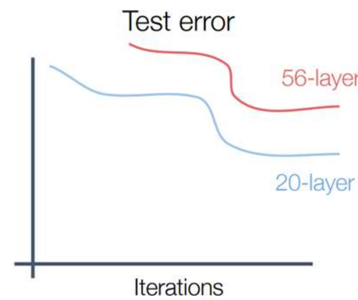


Residual Networks

3

Once we have Batch Normalization, we can train networks with 10+ layers. What happens as we go deeper?

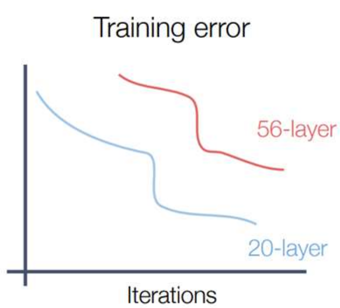
Deeper model does worse than shallow model!



Residual Networks

4

Once we have Batch Normalization, we can train networks with 10+ layers. What happens as we go deeper?

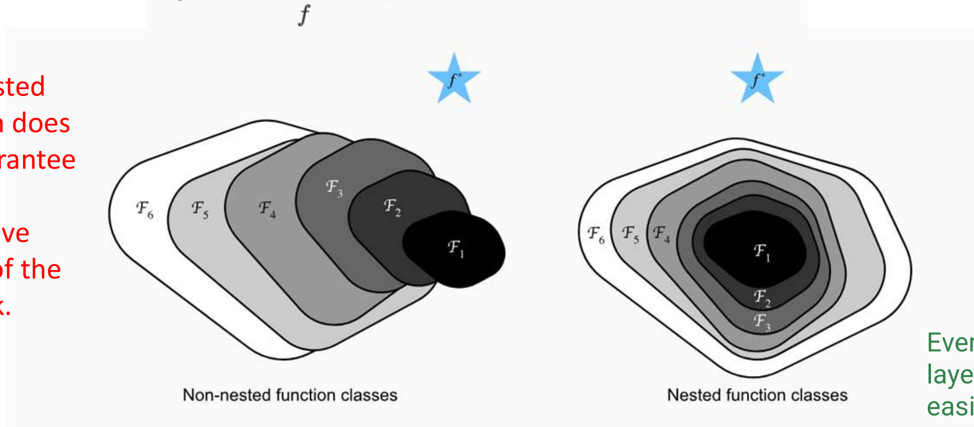


In fact the deep model seems to be **underfitting** since it also performs worse than the shallow model on the training set! It is actually **underfitting**

Why does this happen?

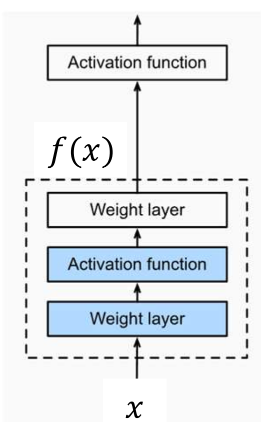
$$f^*_{\mathcal{F}} \stackrel{\text{def}}{=} \underset{f}{\operatorname{argmin}} L(\mathbf{X}, \mathbf{y}, f) \text{ subject to } f \in \mathcal{F}.$$

Non-nested function does not guarantee better expressive power of the network.



Every additional layer should more easily contain the identity function as one of its elements.

Residual Block

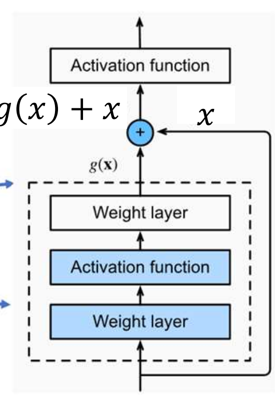


"Plain" block

If you set these to 0, the whole block will compute the identity function!

$f(x) = g(x) + x$

$g(x)$



x

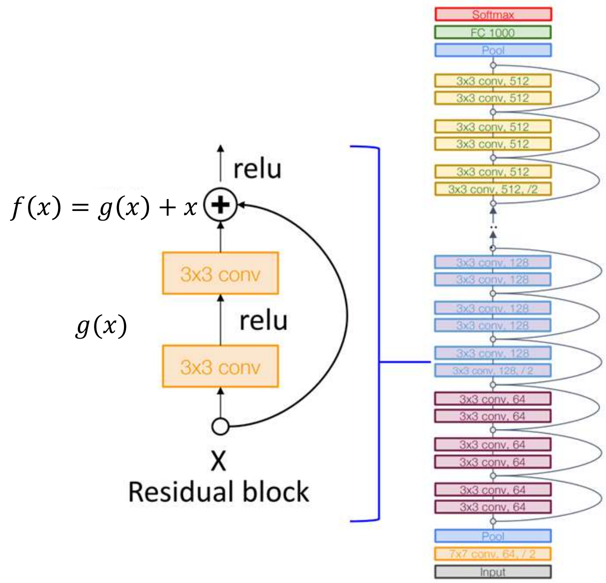
Additive "shortcut"

Residual Block

Residual Networks (ResNet)

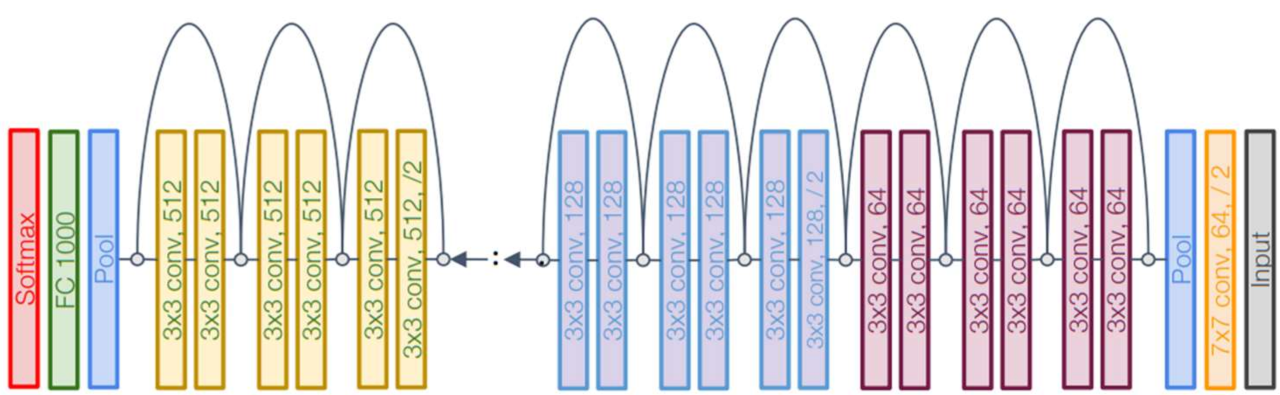
7

- Residual Networks
  - A residual network is a stack of many residual blocks
  - Regular design, like VGG: each residual block has two 3x3 conv
  - Network is divided into stages: the first block of each stage halves the resolution (with stride-2 conv) and doubles the number of channels



Residual Networks (ResNet)

8

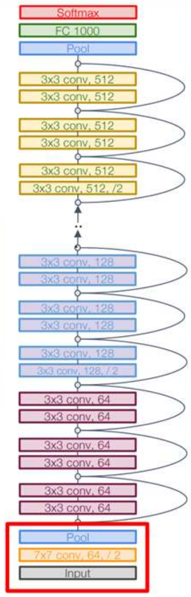


Residual Networks

9

Uses the same aggressive **stem** as GoogleNet to downsample the input 4x before applying residual blocks:

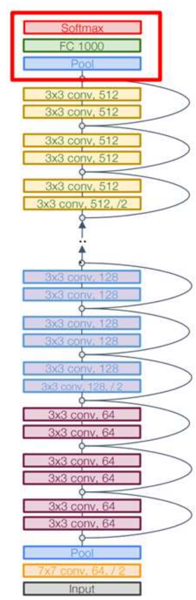
	Input size		Layer				Output size				
Layer	C	H/W	filters	kernel	stride	pad	C	H/W	memory (KB)	params (k)	flop (M)
conv	3	224	64	7	2	3	64	112	3136	9	118
max-pool	64	112		3	2	1	64	56	784	0	2



Residual Networks

10

Like GoogLeNet, no big fully-connected-layers: instead use **global average pooling** and a single linear layer at the end



Residual Networks

11

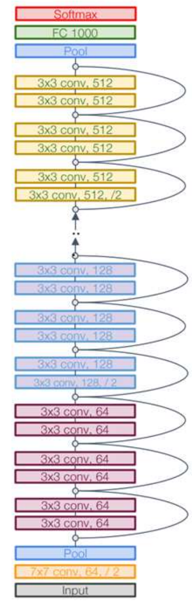
ResNet-18:
Stem: 1 conv layer
Stage 1 (C=64): 2 res. block = 4 conv
Stage 2 (C=128): 2 res. block = 4 conv
Stage 3 (C=256): 2 res. block = 4 conv
Stage 4 (C=512): 2 res. block = 4 conv
Linear

ImageNet top-5 error: 10.92
GFLOP: 1.8

ResNet-34:
Stem: 1 conv layer
Stage 1: 3 res. block = 6 conv
Stage 2: 4 res. block = 8 conv
Stage 3: 6 res. block = 12 conv
Stage 4: 3 res. block = 6 conv
Linear

ImageNet top-5 error: 8.58
GFLOP: 3.6

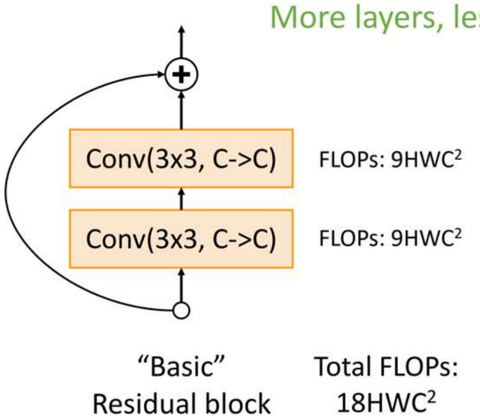
VGG-16:
ImageNet top-5 error: 9.62
GFLOP: 13.6

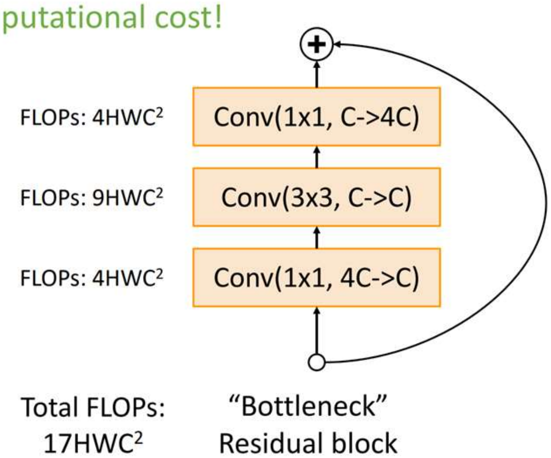


Residual Networks

12

More layers, less computational cost!

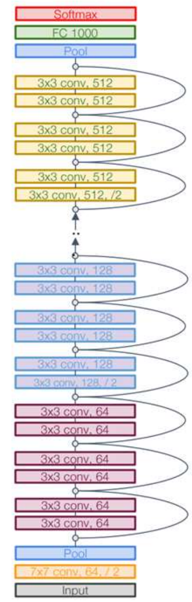




Residual Networks

13

	Block type	Stem layers	Stage 1 Blocks	Stage 1 Layers	Stage 2 Blocks	Stage 2 Layers	Stage 3 Blocks	Stage 3 Layers	Stage 4 Blocks	Stage 4 Layers	FC layers	GFLOP	ImageNet top-5 error
ResNet-18	Basic	1	2	4	2	4	2	4	2	4	1	1.8	10.92
ResNet-34	Basic	1	3	6	4	8	6	12	3	6	1	3.6	8.58
ResNet-50	Bottle	1	3	9	4	12	6	18	3	9	1	3.8	7.13
ResNet-101	Bottle	1	3	9	4	12	23	69	3	9	1	7.6	6.44
ResNet-152	Bottle	1	3	9	8	24	36	108	3	9	1	11.3	5.94



Prof. Monidipa Das, Department of CSE, IIT (ISM) Dhanbad

Residual Networks

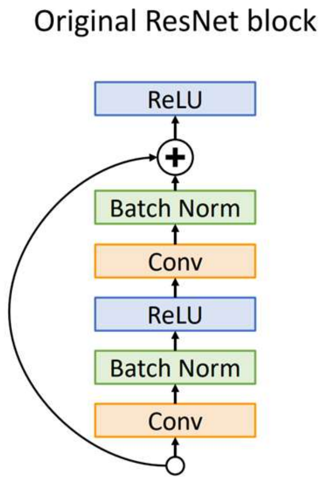
14

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2 $\begin{bmatrix}3\times3, 64\\ 3\times3, 64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 64\\ 3\times3, 64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\ 3\times3, 64\\ 1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\ 3\times3, 64\\ 1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\ 3\times3, 64\\ 1\times1, 256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3, 128\\ 3\times3, 128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 128\\ 3\times3, 128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\ 3\times3, 128\\ 1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\ 3\times3, 128\\ 1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\ 3\times3, 128\\ 1\times1, 512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3, 256\\ 3\times3, 256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 256\\ 3\times3, 256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\ 3\times3, 256\\ 1\times1, 1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\ 3\times3, 256\\ 1\times1, 1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1, 256\\ 3\times3, 256\\ 1\times1, 1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3, 512\\ 3\times3, 512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 512\\ 3\times3, 512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\ 3\times3, 512\\ 1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\ 3\times3, 512\\ 1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\ 3\times3, 512\\ 1\times1, 2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Prof. Monidipa Das, Department of CSE, IIT (ISM) Dhanbad

Residual Networks

15

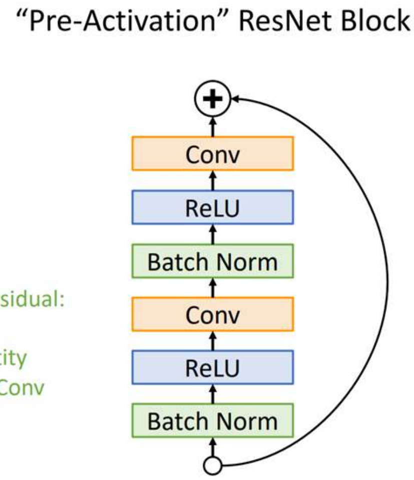


Note ReLU **after** residual:

Cannot actually learn identity function since outputs are nonnegative!

Note ReLU **inside** residual:

Can learn true identity function by setting Conv weights to zero!



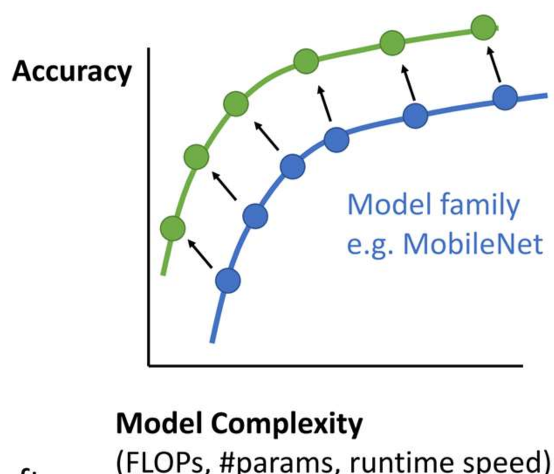
Tiny Neural Networks for Mobile Devices

16

Instead of pushing for the largest network with biggest accuracy, consider tiny networks and accuracy / complexity tradeoff

Compare **families of models**:

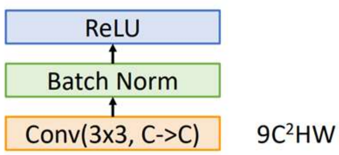
One family is better than another if it moves the whole curve up and to the left



MobileNets: Tiny Networks (For Mobile Devices)

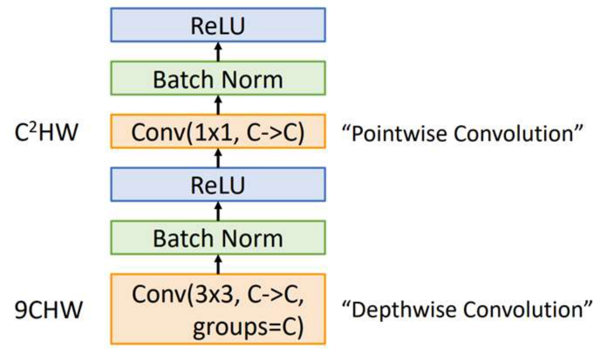
17

Standard Convolution Block
Total cost: $9C^2HW$



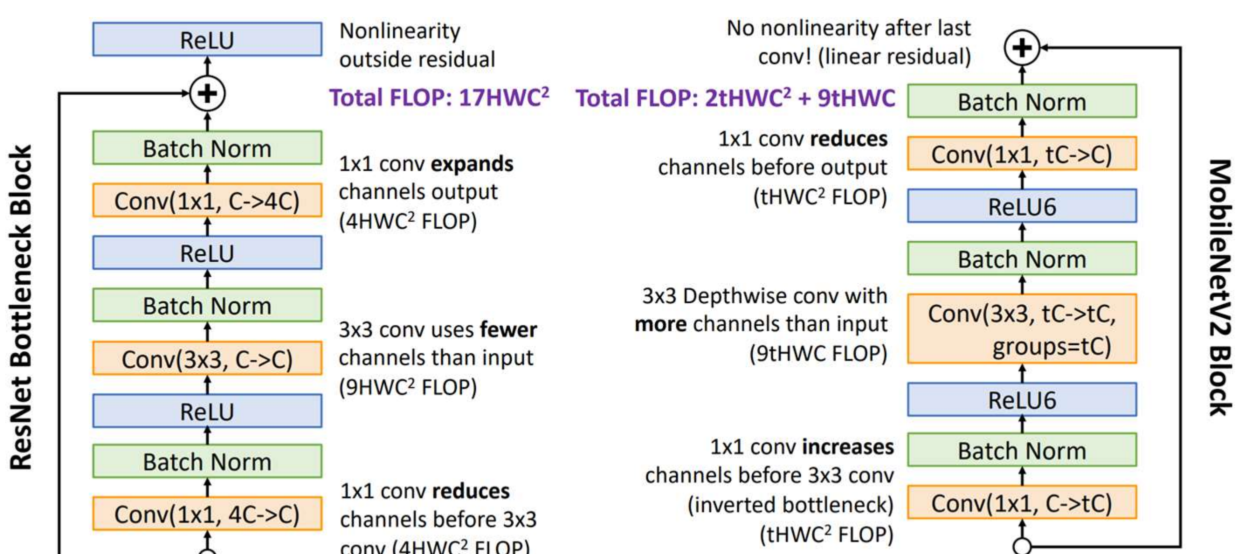
Speedup = $9C^2/(9C+C^2)$
= $9C/(9+C)$
=> 9 (as C->inf)

Depthwise Separable Convolution
Total cost: $(9C + C^2)HW$



MobileNetV2: Inverted Bottleneck, Linear Residual

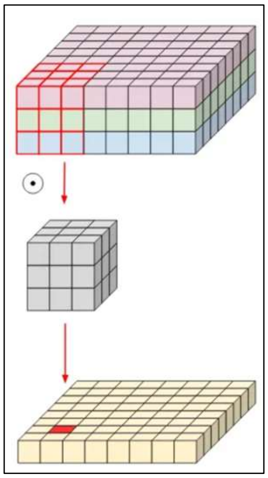
18



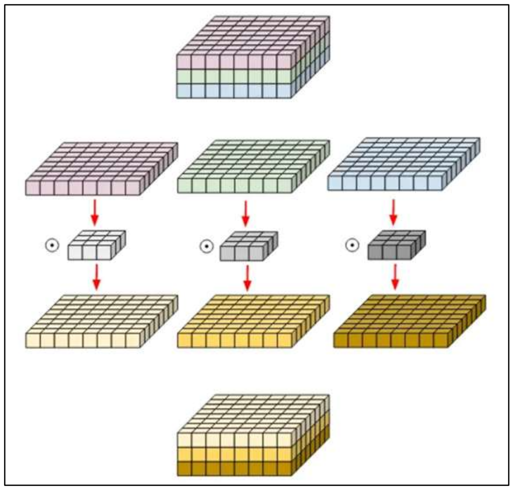
Sandler et al, "MobileNetV2: Inverted Residuals and Linear Bottlenecks", CVPR 2018

Illustrations for Standard, Depth-wise, and Depth-wise Separable Convolution

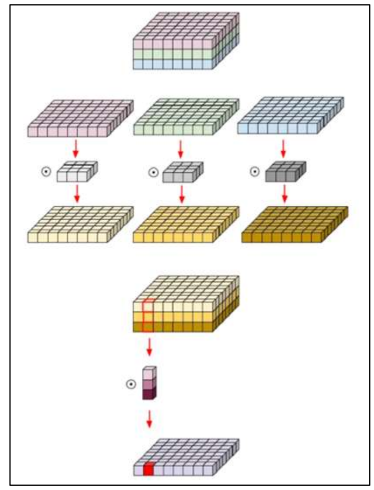
19



Standard convolution



Depth-wise convolution



Depth-wise separable convolution

Source: Internet

20

Questions?

Acknowledgement: Prof. W. L. Hamilton, Fei-Fei Li, Ranjay Krishna, Danfei Xu, Justin Johnson, Xavier Bresson