What can(’t) we do with ConvNets?

Classification architectures + Semantic segmenation

Karel Zimmermann
Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Outline

- Architectures of classification networks
- Architectures of segmentation networks
- Architectures of regression networks
- Architectures of detection networks
- Architectures of feature matching networks

IMAGENET

Classification results

http://image-net.org/challenges/LSVRC/2017/index

Label: Steel drum

IMAGENET

Classification results

http://image-net.org/challenges/LSVRC/2017/index

Label: Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

IMAGENET

Classification results

http://image-net.org/challenges/LSVRC/2017/index

Label: Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

IMAGENET

Classification results

http://image-net.org/challenges/LSVRC/2017/index

Label: Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

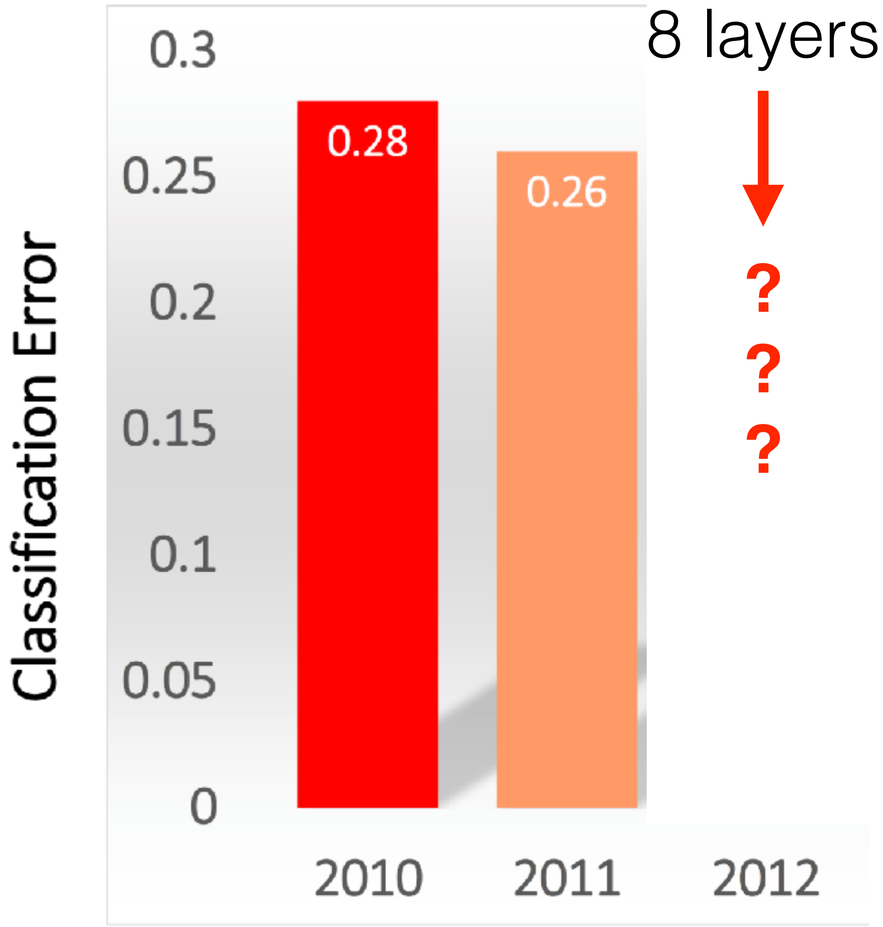
✗

$$\text{Error} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{incorrect on image i}]$$

IMAGENET

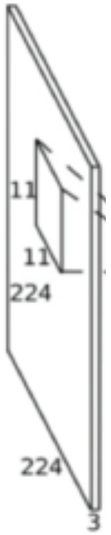
Classification results

AlexNet
8 layers



?
?
?

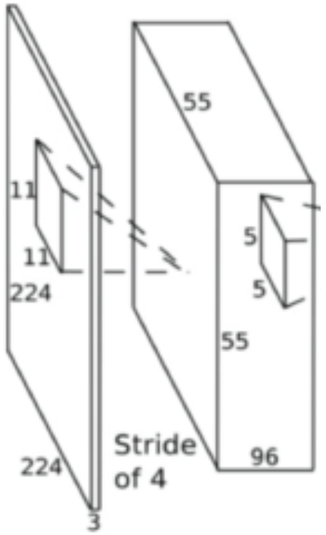
AlexNet on ImageNet 2012 (**over 27k citations !!!**)



- Param in layer1 (conv, 96 11x11 filters, stride=4, pad=0)?

Alex Krizhevsky et al, Imagenet classification with deep convolutional neural networks, NIPS, 2012
https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

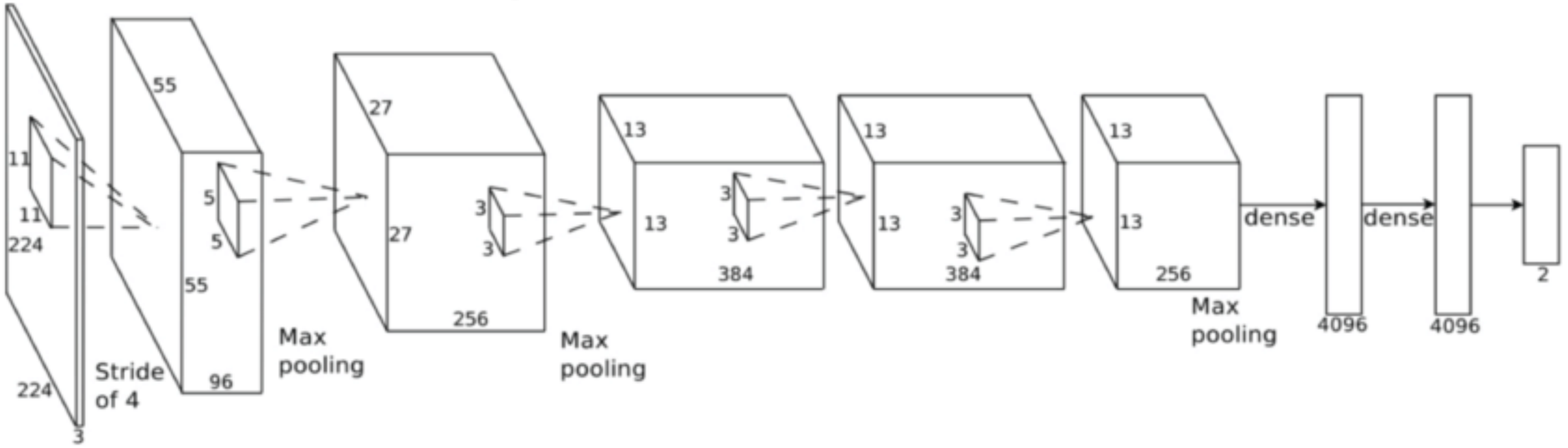
AlexNet on ImageNet 2012 (**over 27k citations !!!**)



- Param in layer1 (conv, 96 11x11 filters, stride=4, pad=0)?
- Param in layer2 (maxp,3x3 filters, stride=2, pad=0)?

Alex Krizhevsky et al, Imagenet classification with deep convolutional neural networks, NIPS, 2012
https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

AlexNet on ImageNet 2012 (**over 27k citations !!!**)



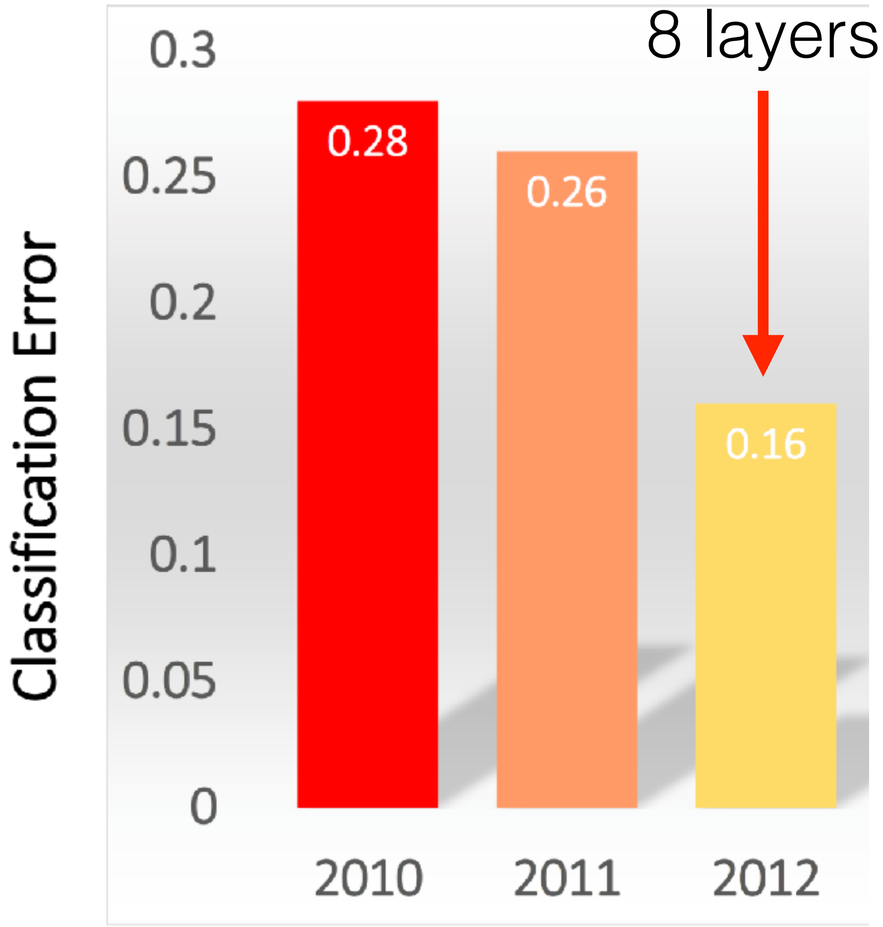
- Param in layer1 (conv, 96 11x11 filters, stride=4, pad=0)?
- Param in layer2 (maxp,3x3 filters, stride=2, pad=0)?
- Param in layer3 (conv, 256 5x5 filters, stride=1, pad=2?
- Parameters in total: 60M, Depth: 8 layers

Alex Krizhevsky et al, Imagenet classification with deep convolutional neural networks, NIPS, 2012
https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

IMAGENET

Classification results

AlexNet
8 layers

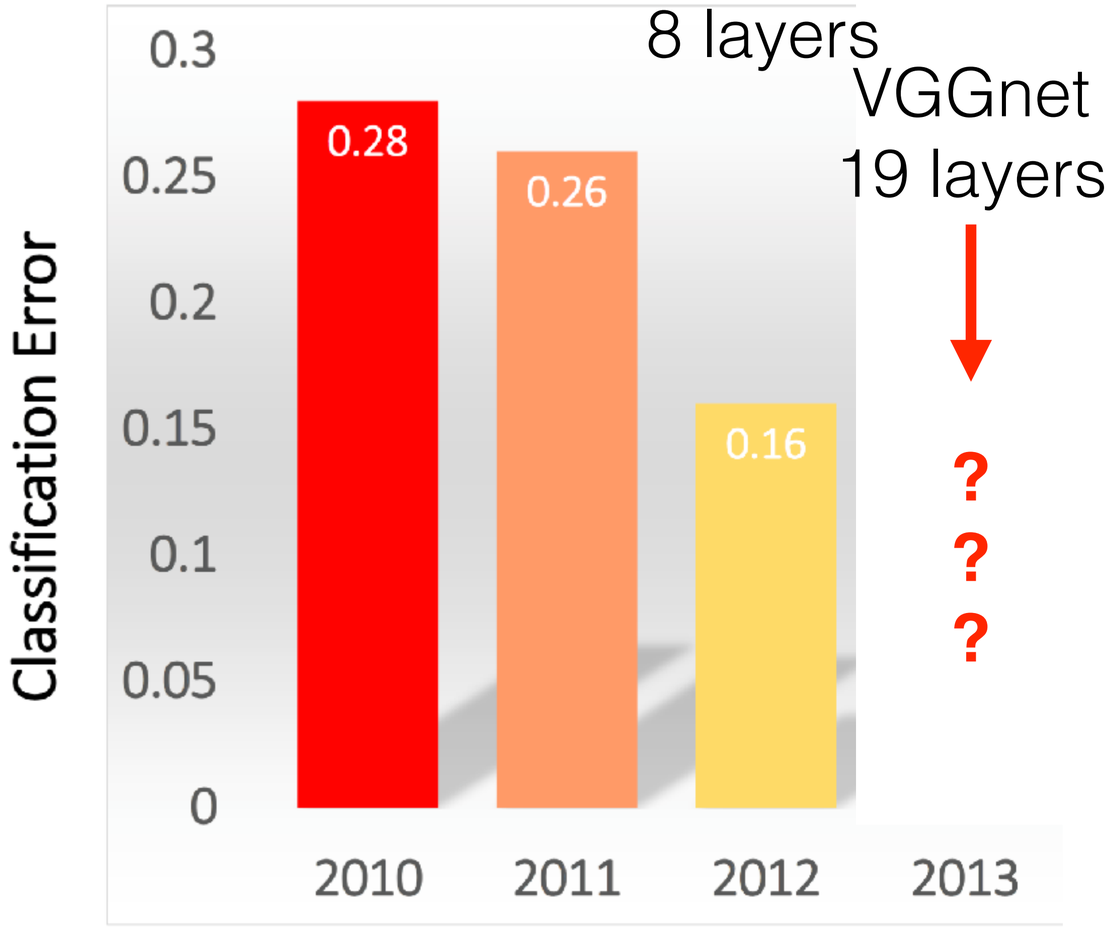


IMAGENET

Classification results

AlexNet
8 layers

VGGnet
19 layers



VGGNet vs AlexNet

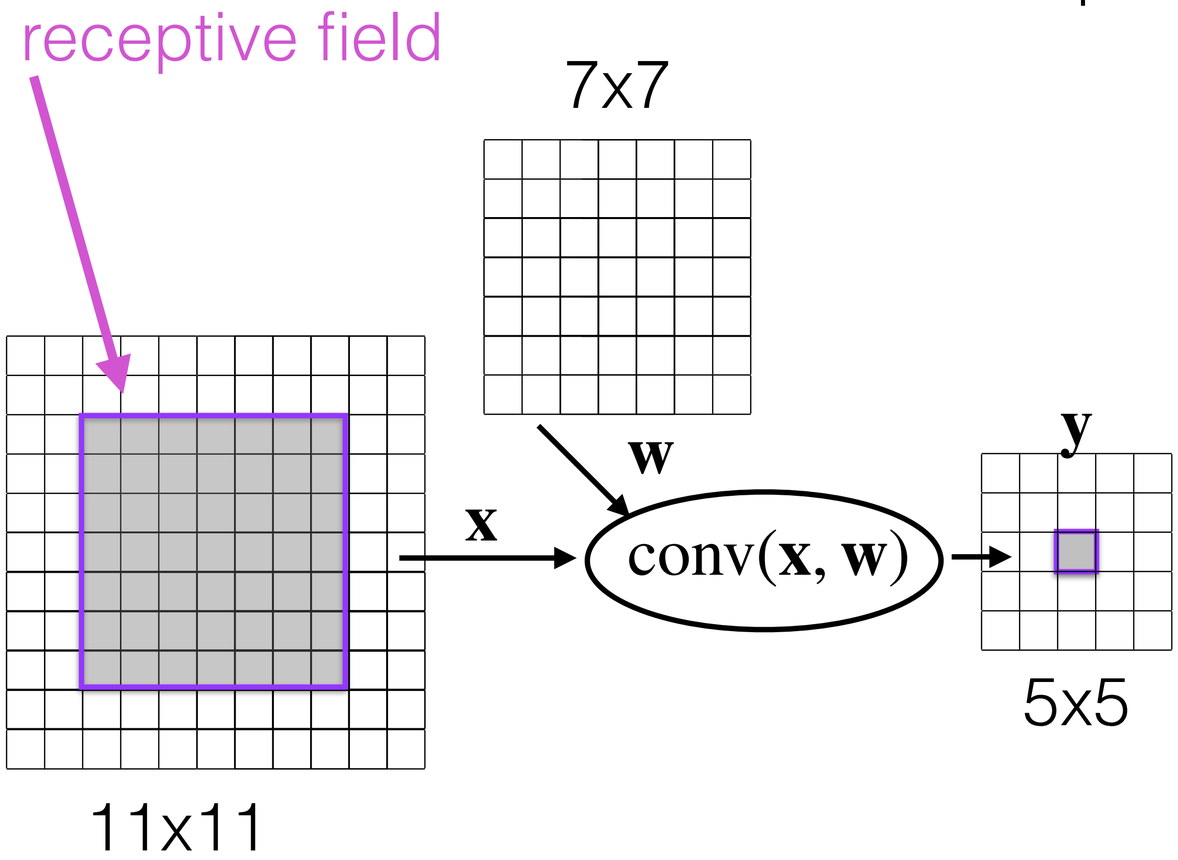
- large filters
- shallow (8 layers)

- small filters
- deeper (19 layers)

- Parameters in total: 138M, Depth: 19 layers

Simonyan and Zissermann,Very Deep Convolutional Networks for Large Scale Image Recognition, 2014
https://arxiv.org/abs/1409.1556

Receptive field

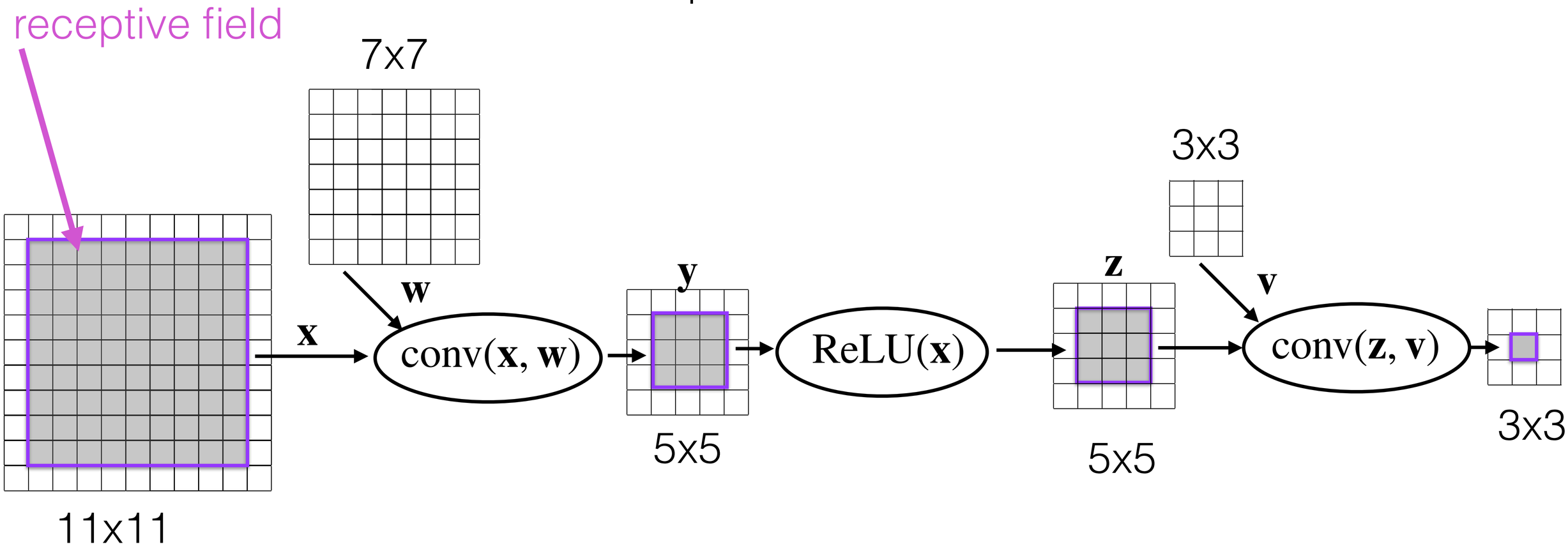


Receptive field: region in the input space that affect activation of a particular neuron.

What is the size of the receptive field??? 7x7

https://arxiv.org/abs/1409.1556

Receptive field

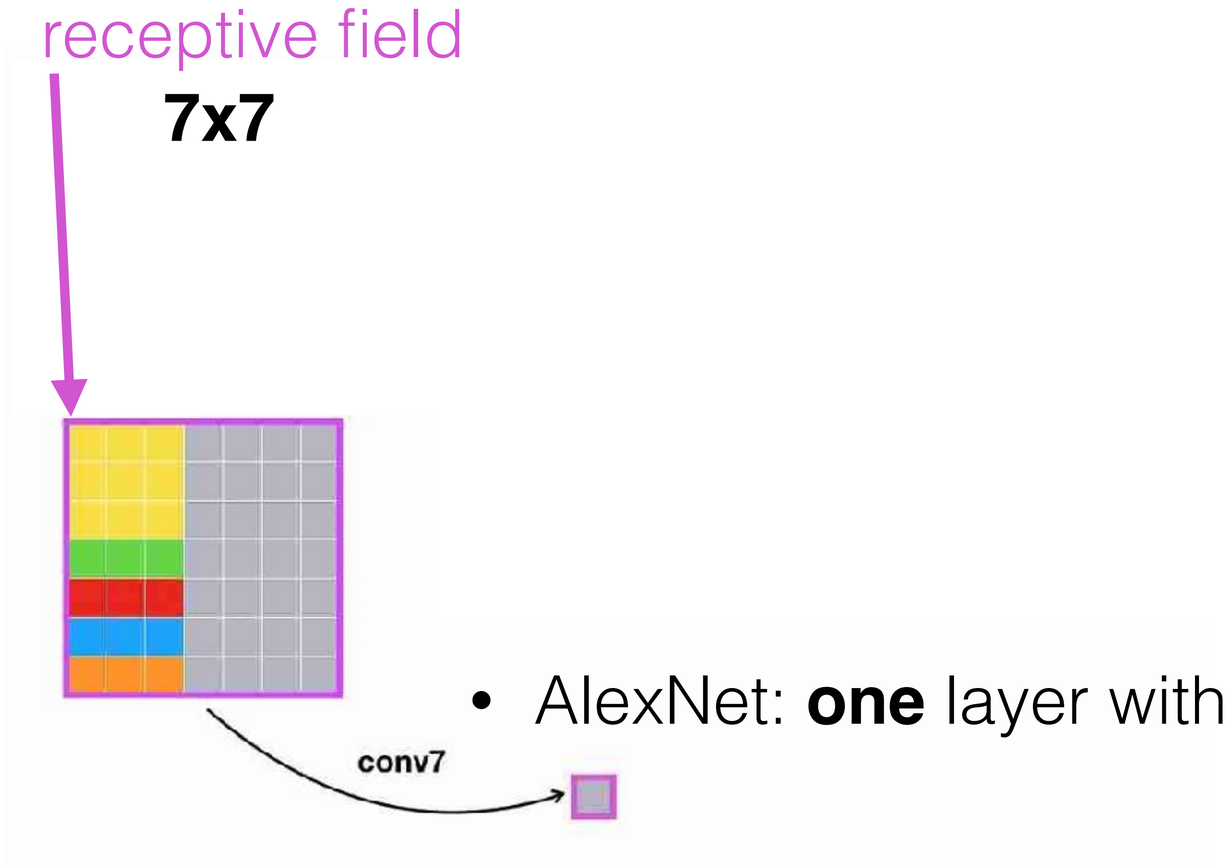


Receptive field: region in the input space that affect activation of a particular neuron.

What is the size of the receptive field??? 9x9

https://arxiv.org/abs/1409.1556

VGGNet vs AlexNet

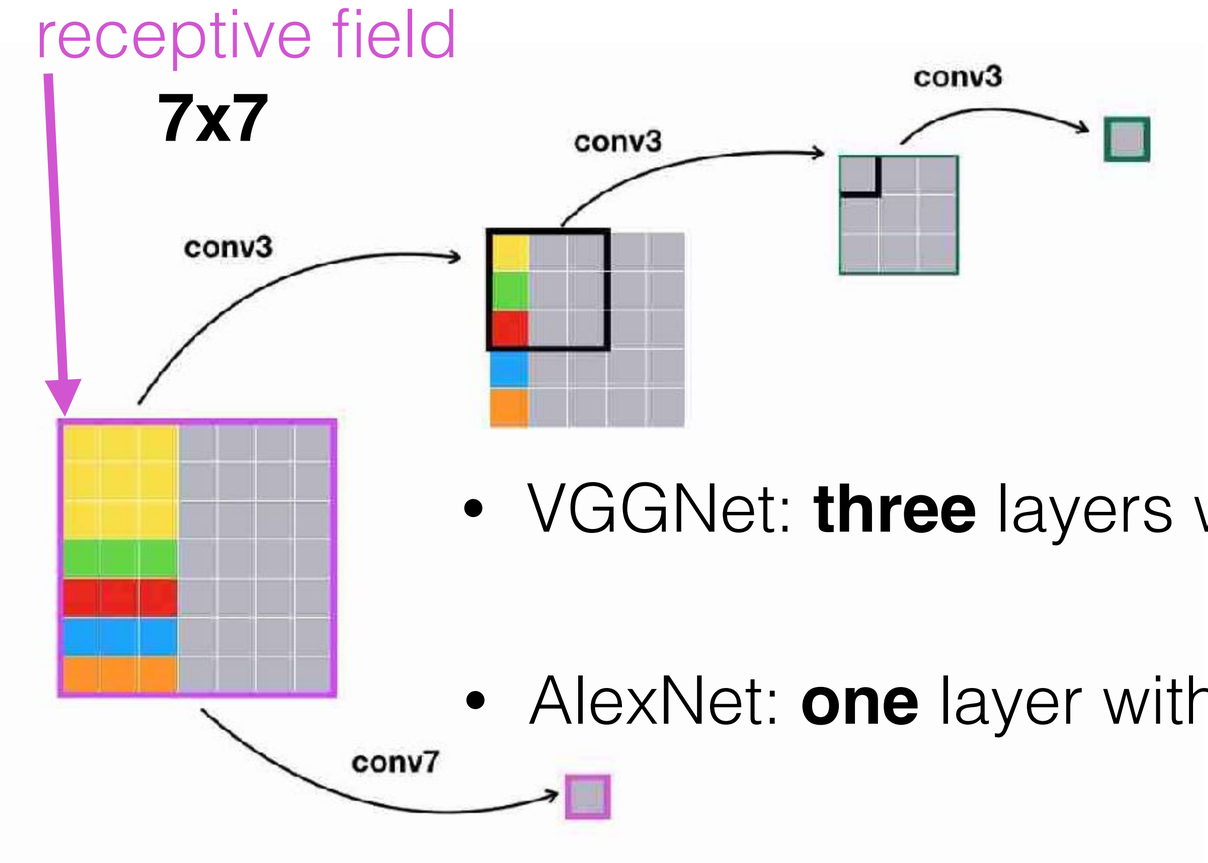


- AlexNet: **one** layer with **7x7** filter (49+1 params)

Receptive field: region in the input space that affect activation of a particular neuron.

https://arxiv.org/abs/1409.1556

VGGNet vs AlexNet



- VGGNet: **three** layers with **3x3** filters (3x9+3 params)
- AlexNet: **one** layer with **7x7** filter (49+1 params)

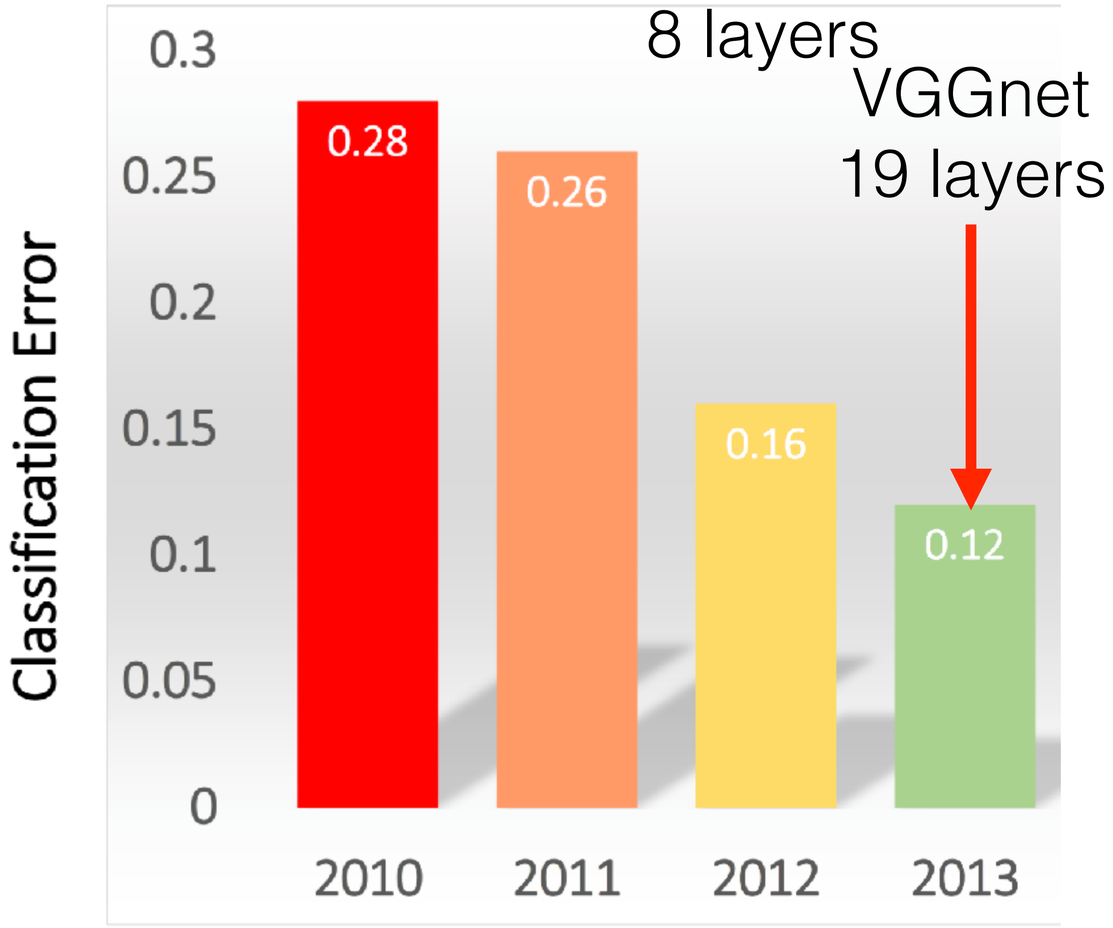
Receptive field: region in the input space that affect activation of a particular neuron.

VGGNet has **the same receptive field** with **less parameters**

https://arxiv.org/abs/1409.1556

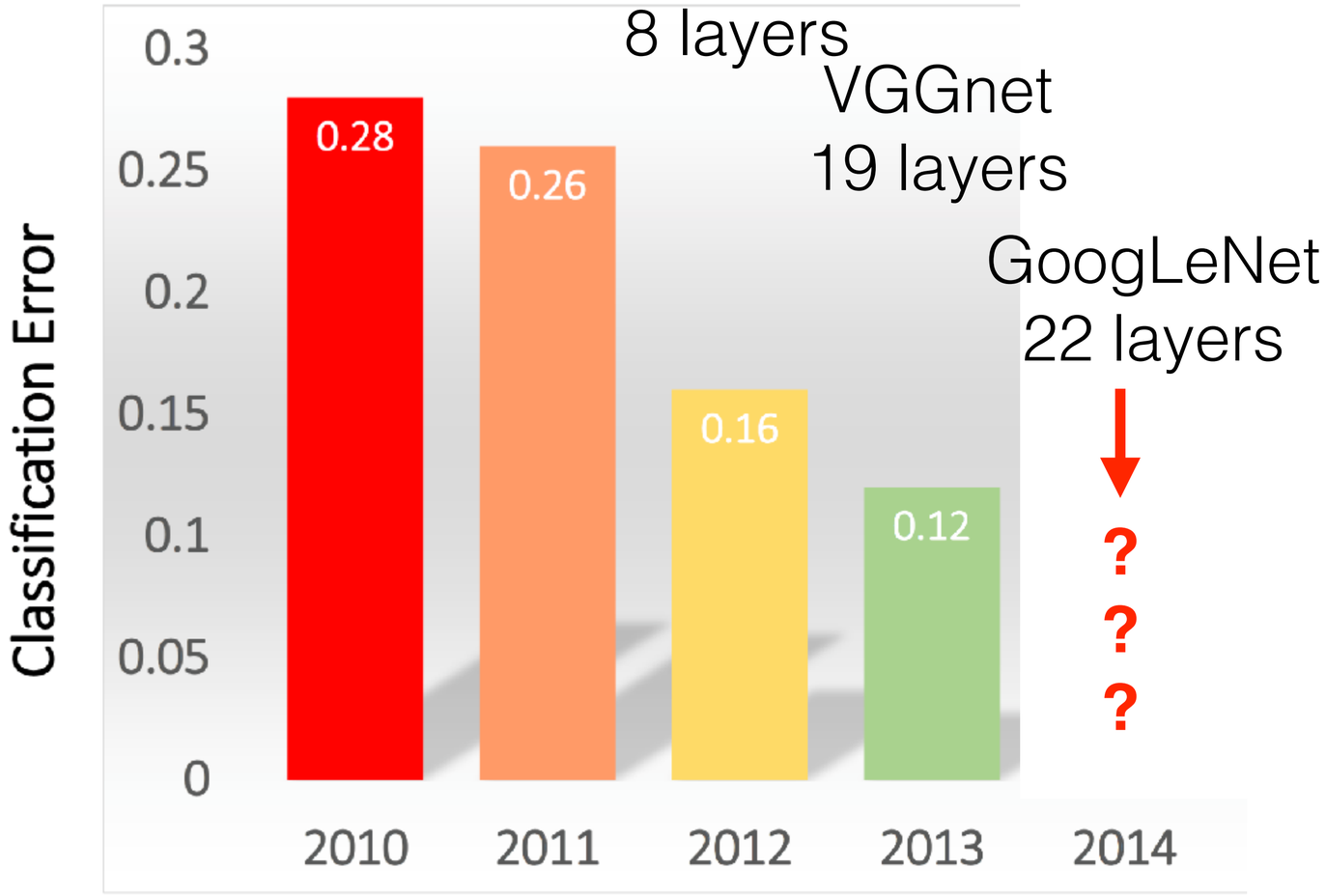
IMAGENET

Classification results

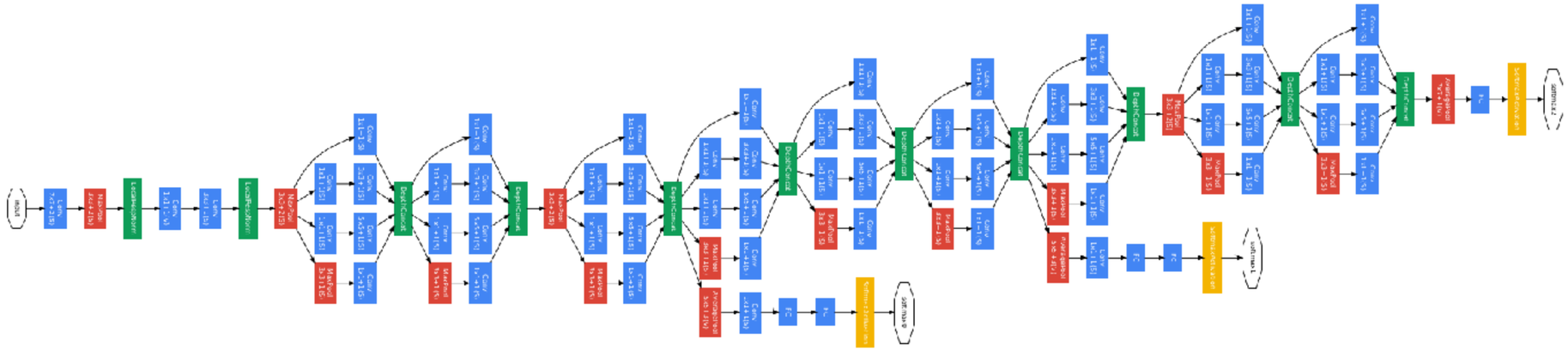


IMAGENET

Classification results

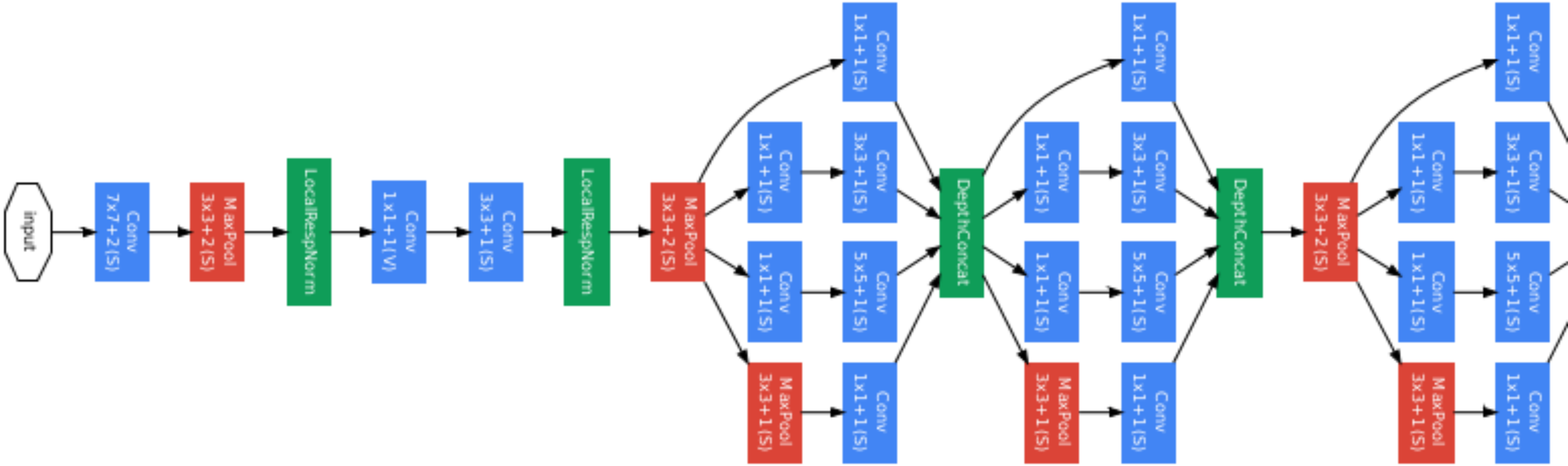


GoogLeNet



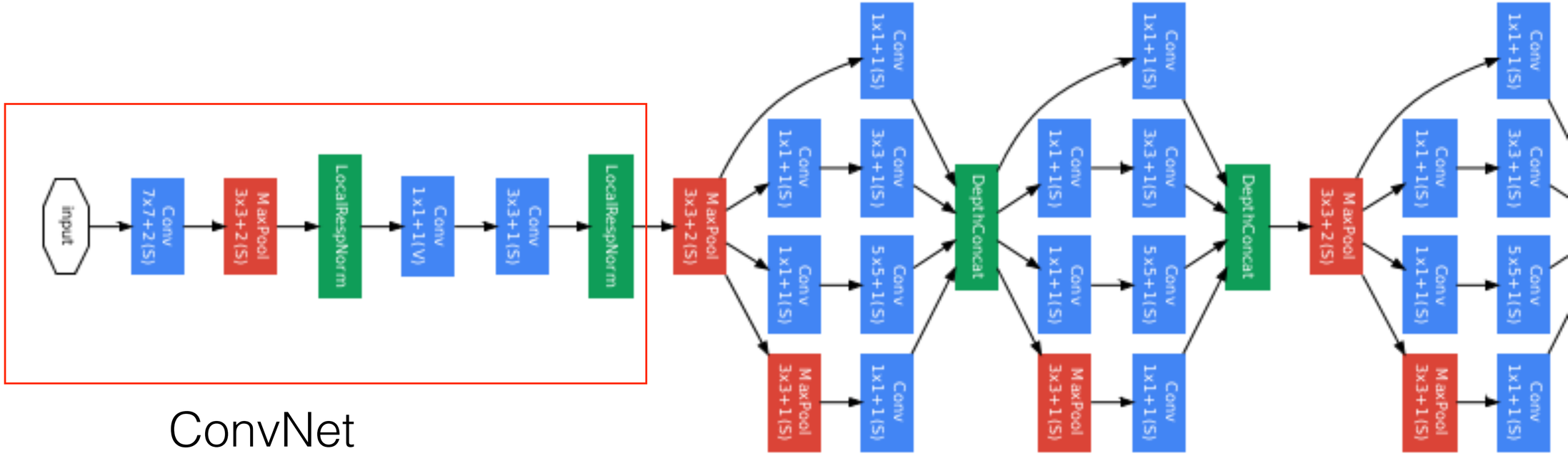
Szegedy et al. Going Deeper with Convolutions, CVPR, 2014
https://arxiv.org/abs/1409.4842

GoogLeNet



Szegedy et al. Going Deeper with Convolutions, CVPR, 2014
https://arxiv.org/abs/1409.4842

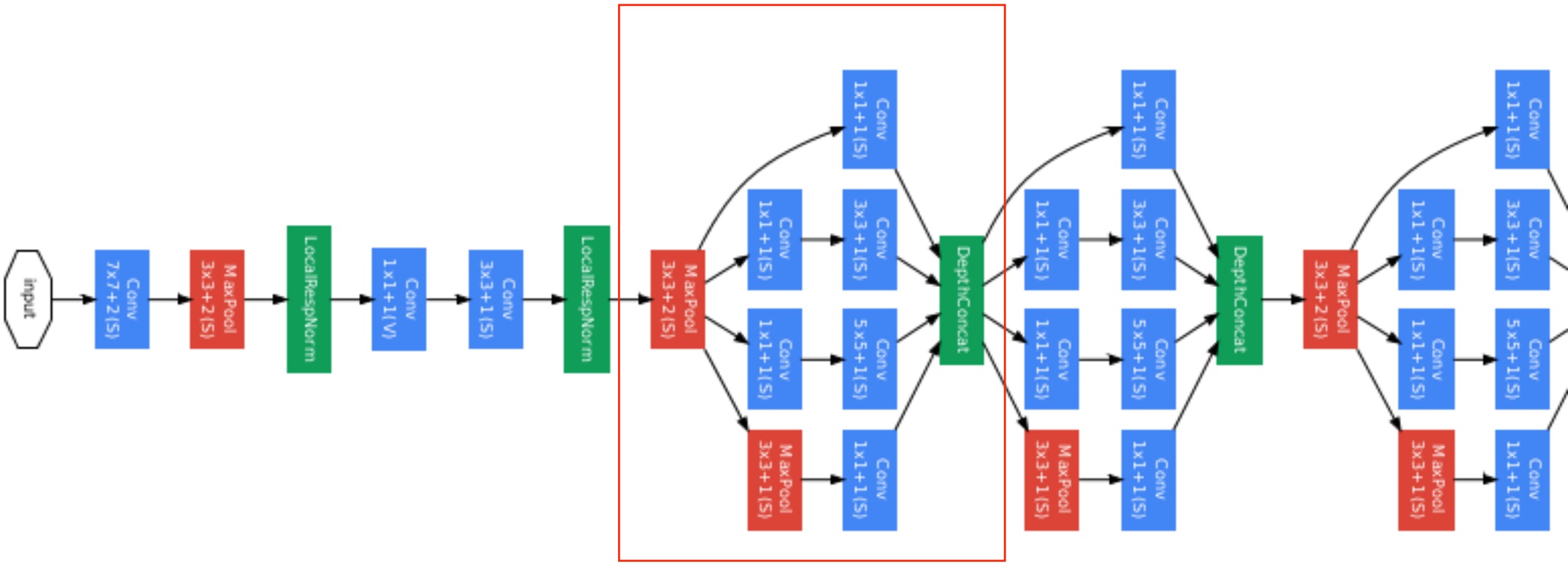
GoogLeNet



ConvNet

Szegedy et al. Going Deeper with Convolutions, CVPR, 2014
https://arxiv.org/abs/1409.4842

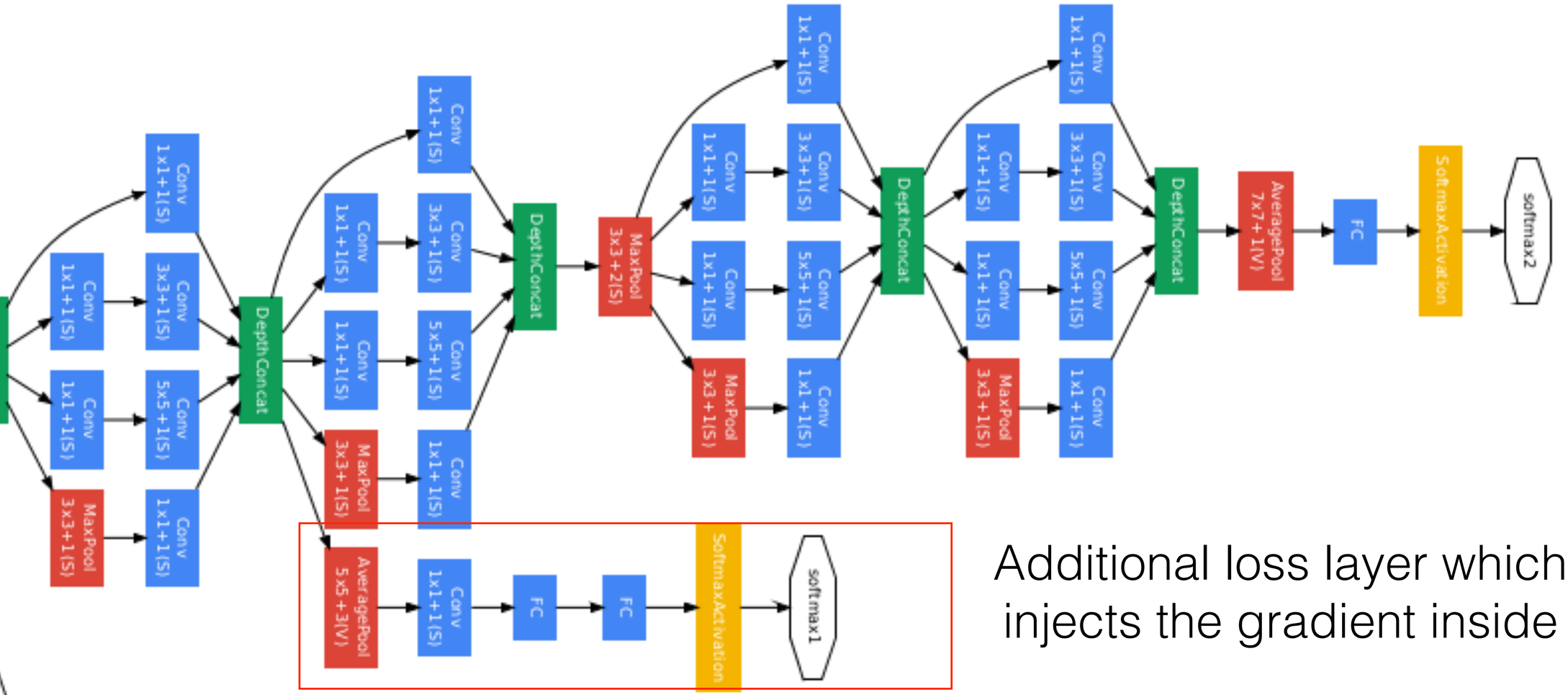
GoogLeNet



Inception module

Szegedy et al. Going Deeper with Convolutions, CVPR, 2014
https://arxiv.org/abs/1409.4842

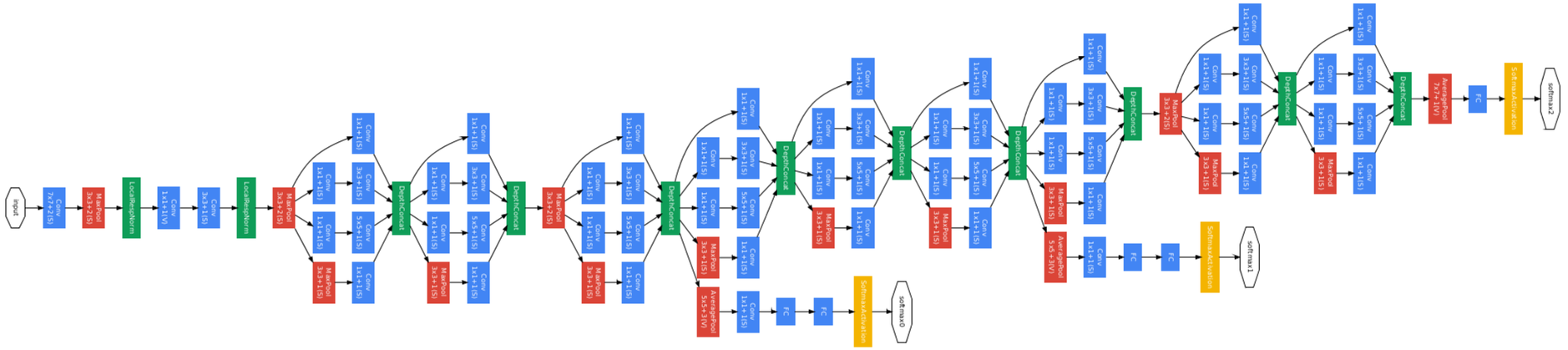
GoogLeNet



Additional loss layer which injects the gradient inside

Szegedy et al. Going Deeper with Convolutions, CVPR, 2014
https://arxiv.org/abs/1409.4842

GoogLeNet

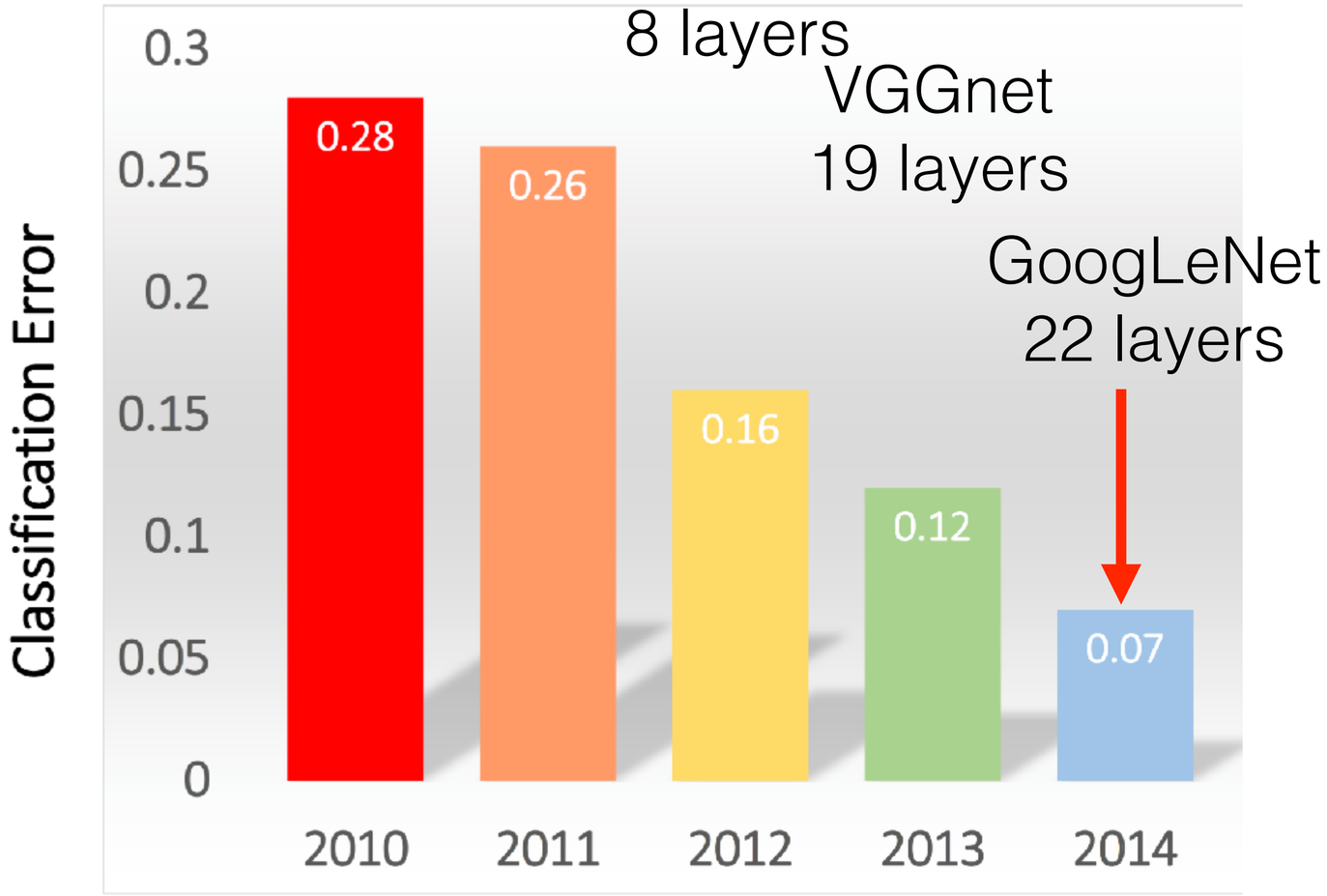


- 12x fewer parameters than AlexNet
- depth 22 layers
- training: few high-end GPU about a week

Szegedy et al. Going Deeper with Convolutions, CVPR, 2014
https://arxiv.org/abs/1409.4842

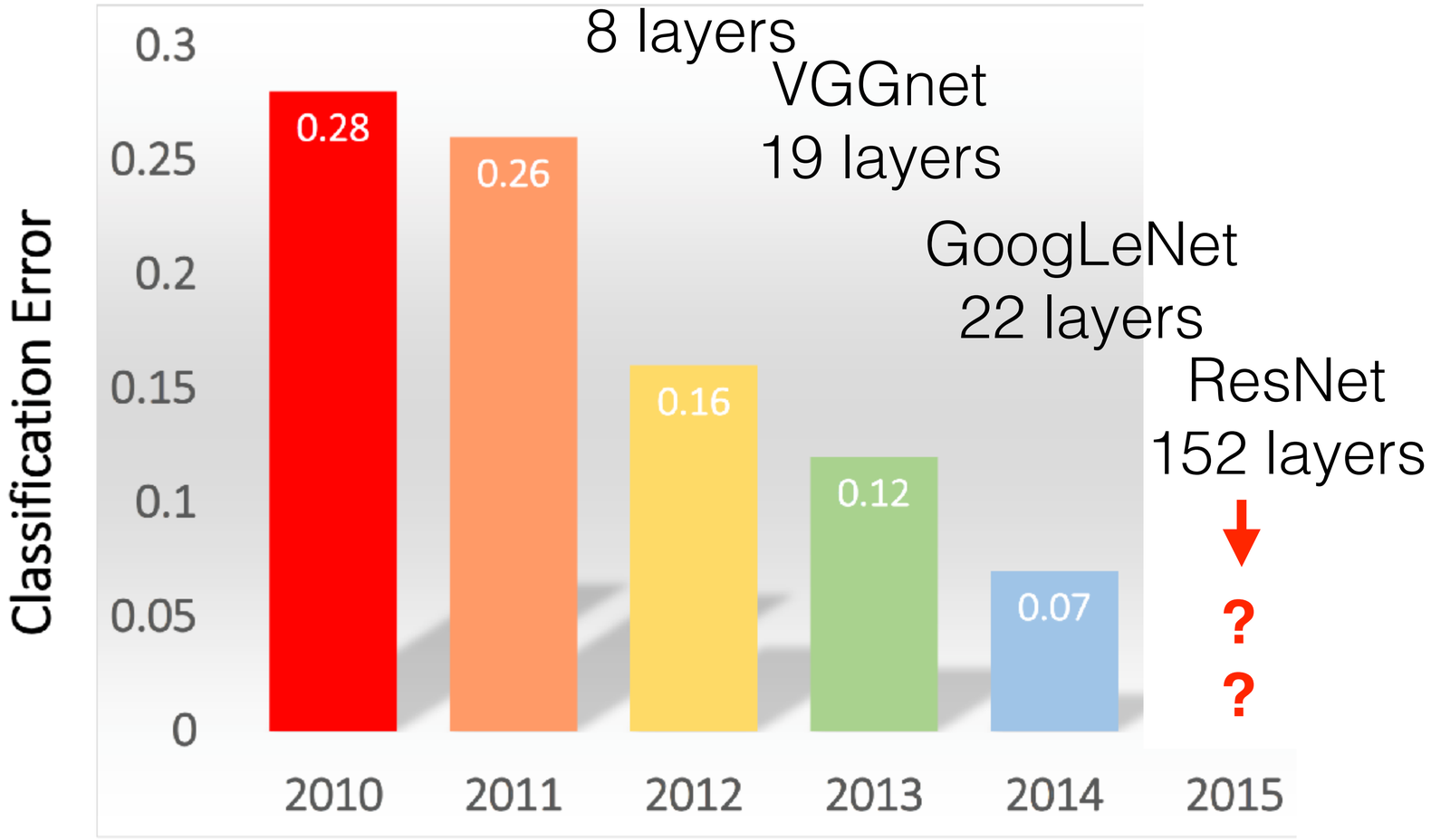
IMAGENET

Classification results



IMAGENET

Classification results



ResNet
Better results with smaller kernels + deeper architectures

Well said Leo, well said

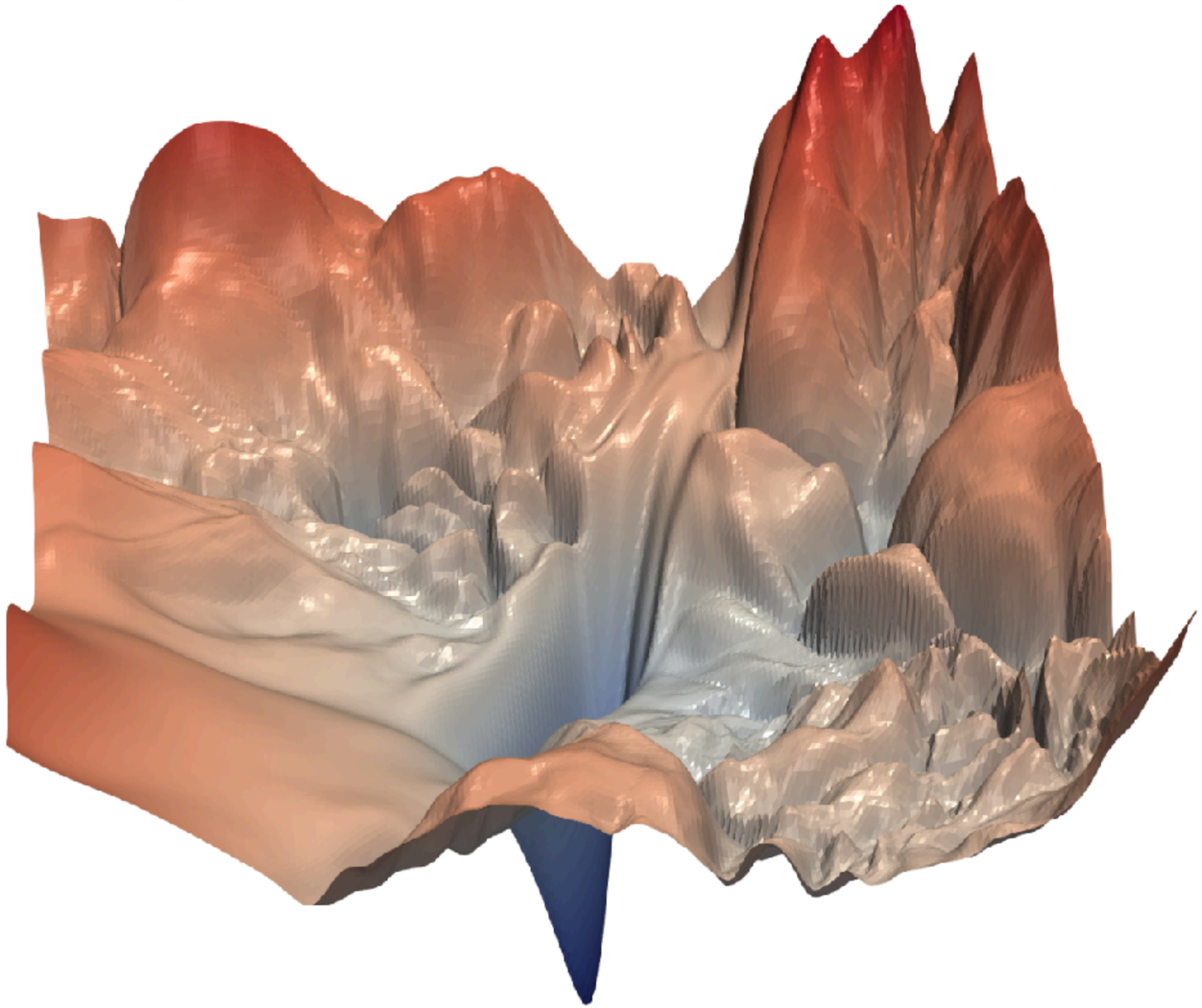
- deeper ConvNet architectures yielded **higher training errors**.=> **is it overfitting?**
- => no overfitting, but vanishing gradient !

He et al. Going Deeper with Convolutions, CVPR, 2015
https://arxiv.org/abs/1512.03385

Visualizing Loss Landscape of Neural Nets

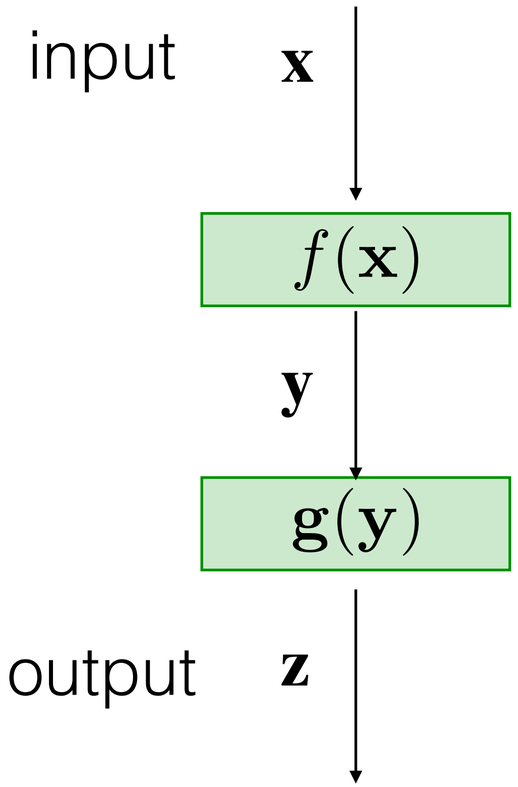
$$f(\alpha, \beta) = \mathcal{L}(\mathbf{w}^* + \alpha\mathbf{u} + \beta\mathbf{v})$$

for randomly chosen (and normalized) directions $\mathbf{u}, \mathbf{v}$



[Li et al, NIPS, 2018] https://arxiv.org/pdf/1712.09913.pdf

forward pass:



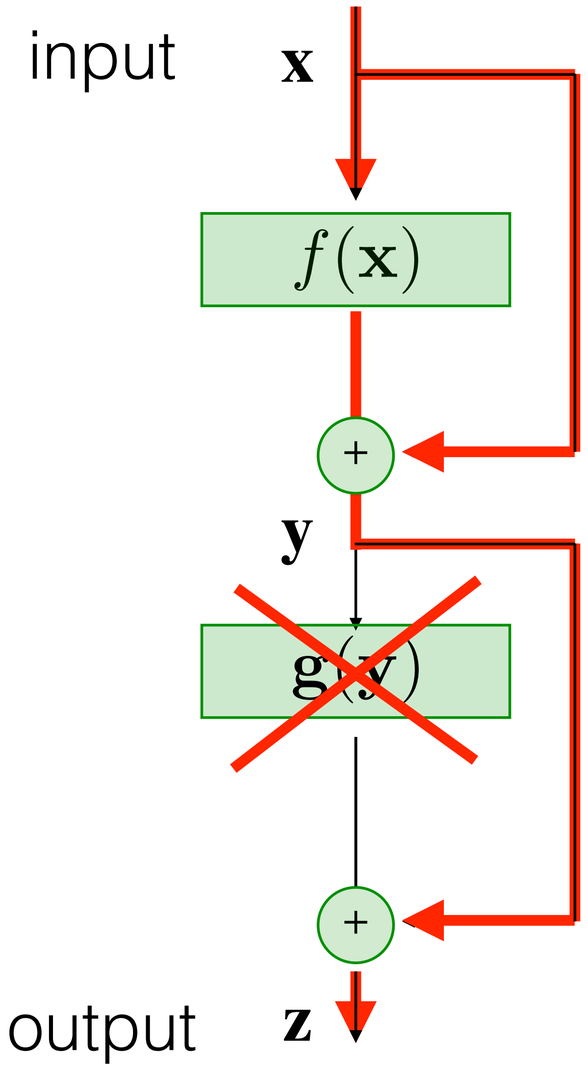
backward pass:

gradient $\dfrac{\partial \mathbf{z}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{z}}{\partial \mathbf{y}} \dfrac{\partial \mathbf{y}}{\partial \mathbf{x}} \approx \mathbf{0}$ then gradient is always zero

$\approx \mathbf{0}$ $\approx \mathbf{0}$

if any local gradient is zero

forward pass:



backward pass:

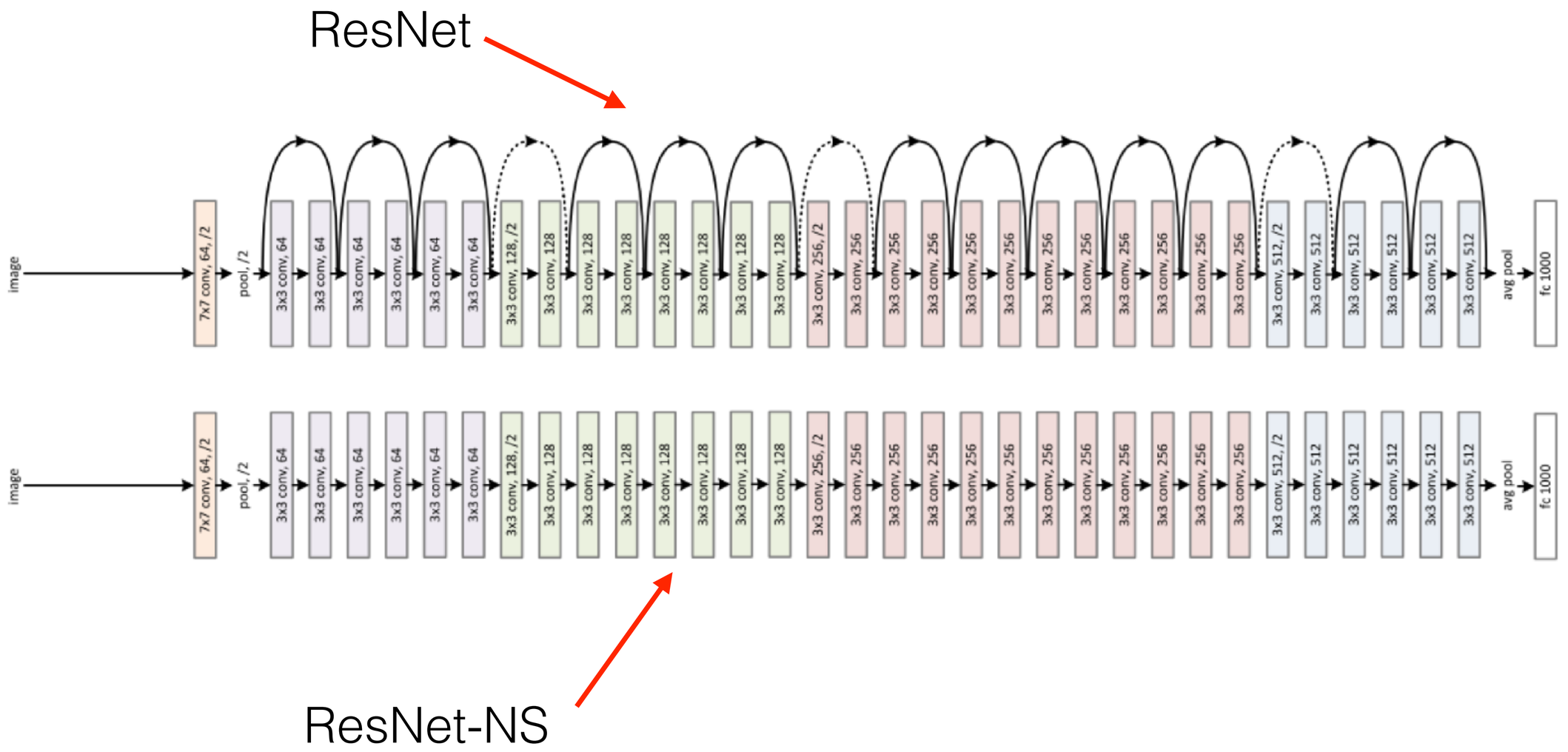
gradient $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}} + 1\right)\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} + 1\right) \neq \mathbf{0}$

$\approx \mathbf{0}$ $\approx \mathbf{0}$

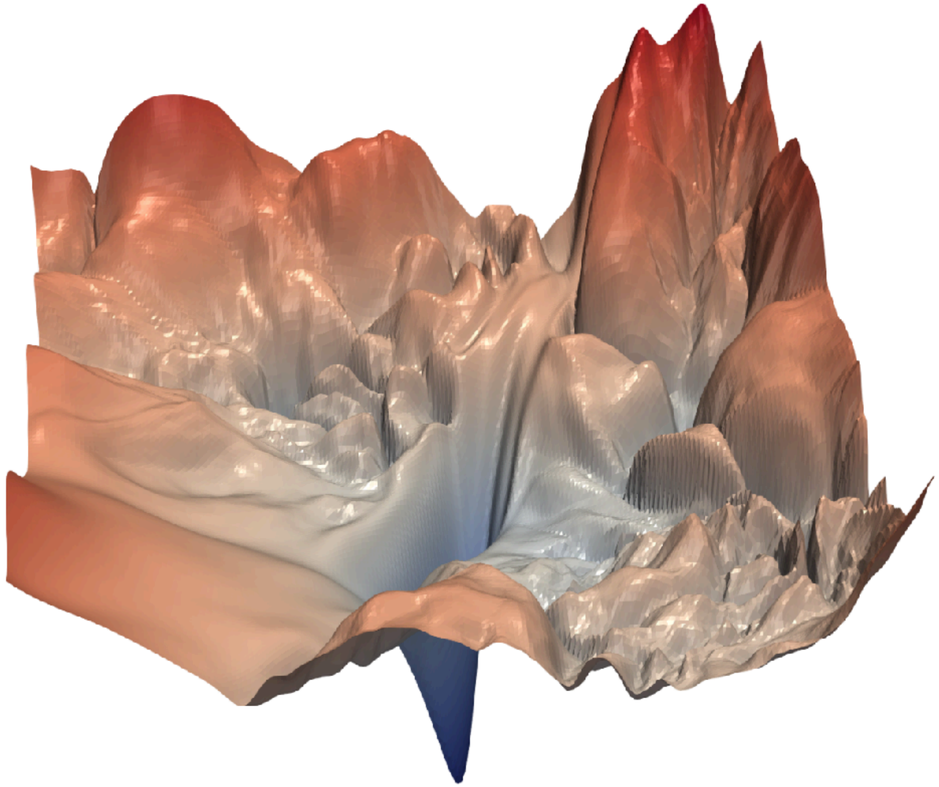
if any local gradient is zero

the gradient can still flow through another path

ResNet: deep ConvNet with skip connections

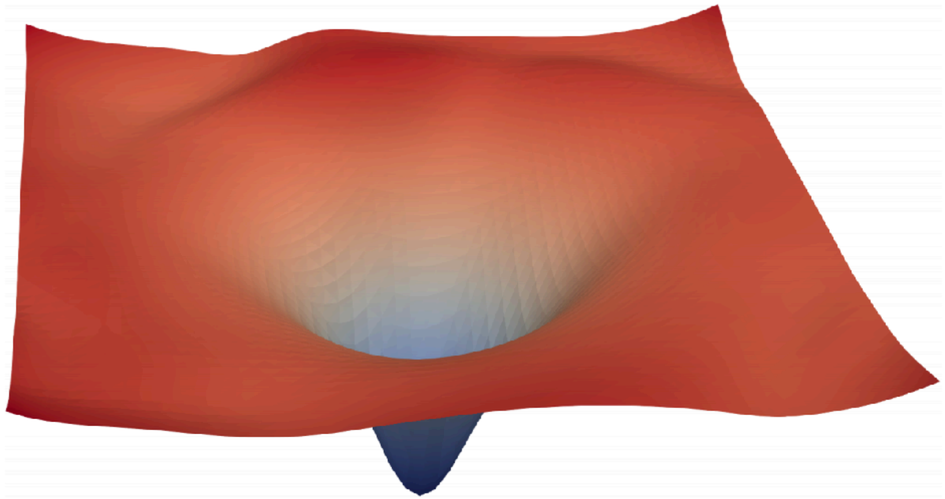


Visualizing Loss Landscape of Neural Nets
[Li et al, NIPS, 2018] https://arxiv.org/pdf/1712.09913.pdf



(a) without skip connections

ResNet-NS

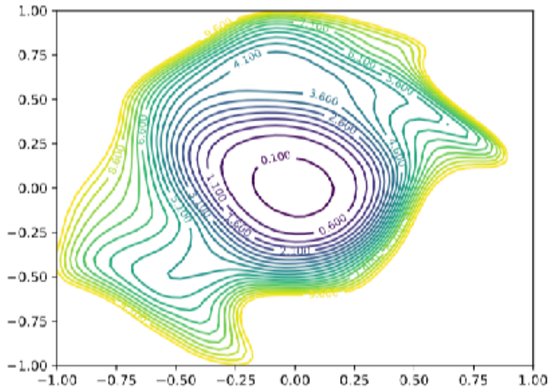


(b) with skip connections

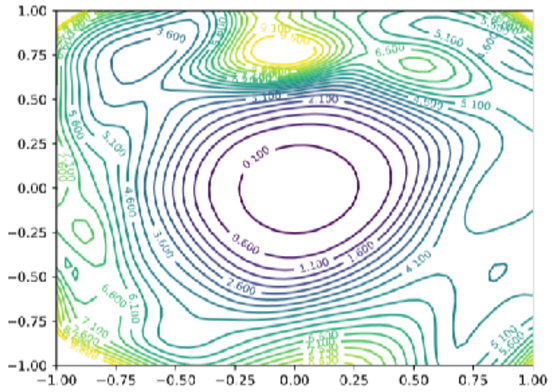
ResNet

ResNet: deep ConvNet with skip connections
[Li et al, NIPS, 2018] https://arxiv.org/pdf/1712.09913.pdf

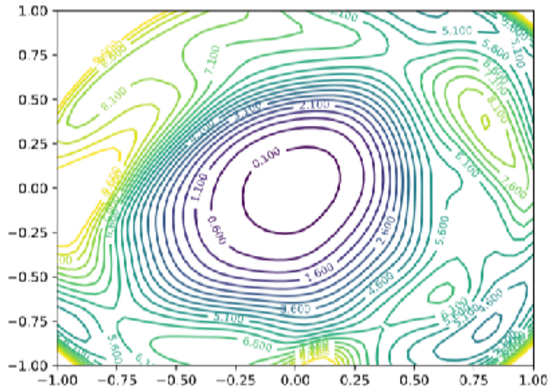
ResNet



(a) ResNet-20, 7.37%

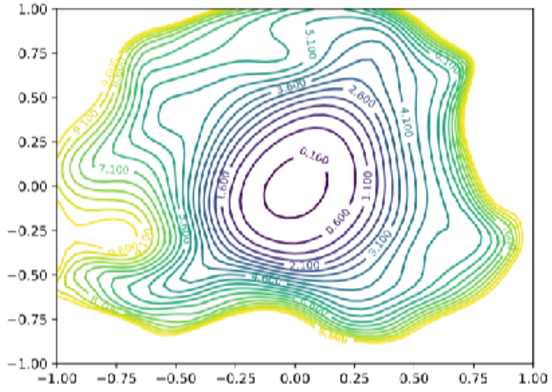


(b) ResNet-56, 5.89%

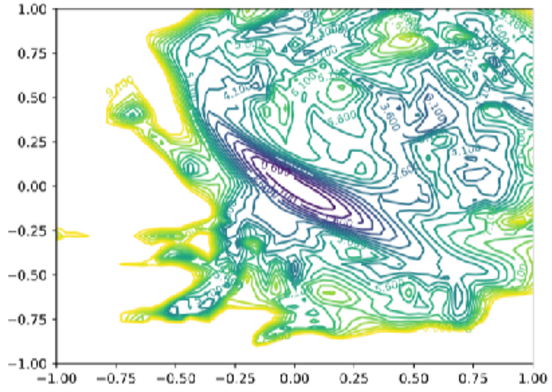


(c) ResNet-110, 5.79%

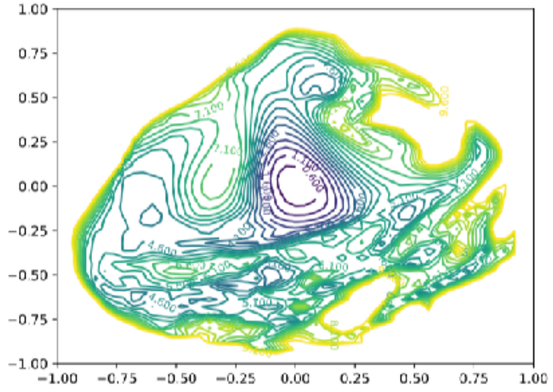
ResNet-NS



(d) ResNet-20-NS, 8.18%

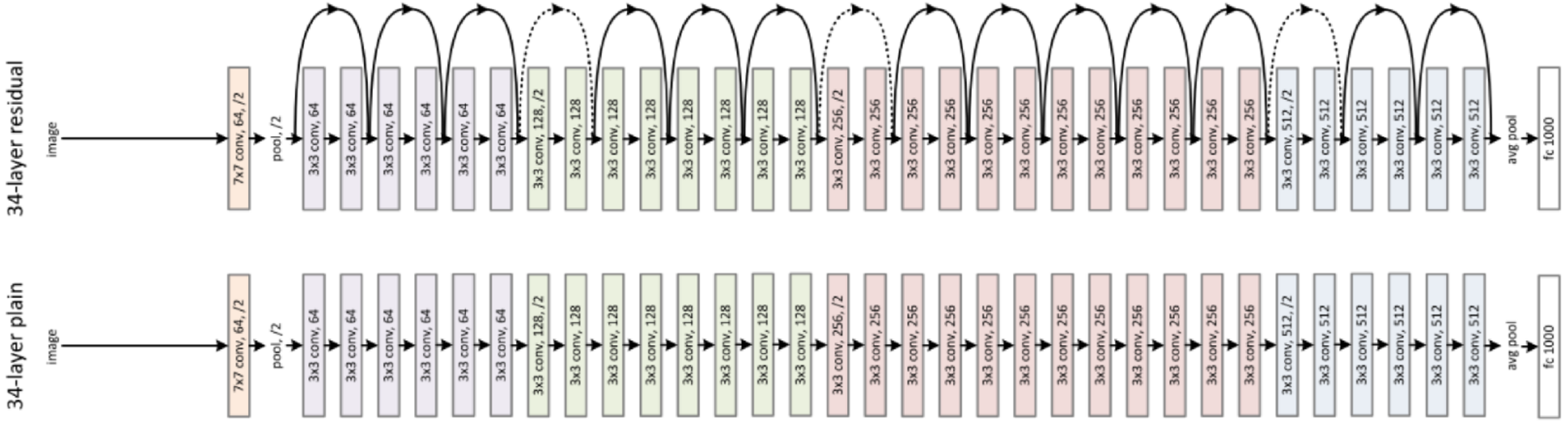


(e) ResNet-56-NS, 13.31%



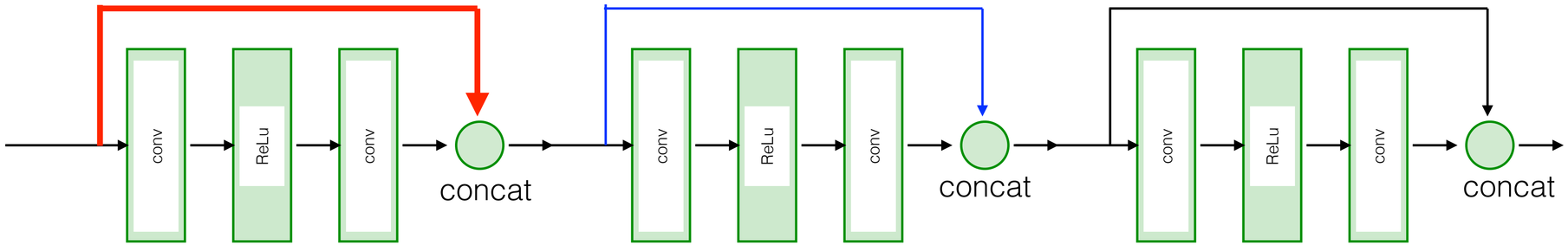
(f) ResNet-110-NS, 16.44%

ResNet: deep ConvNet with skip connections



- **1k+ layers** possible
- **Initialization** with **zero** weights is meaningful
- **Better gradient flow:** many independent paths, opt-friendly landscape
- **Robustness** wrt **noise** and layer removal (if some important feature is not detected in a particular layer it can be detected in following layers)

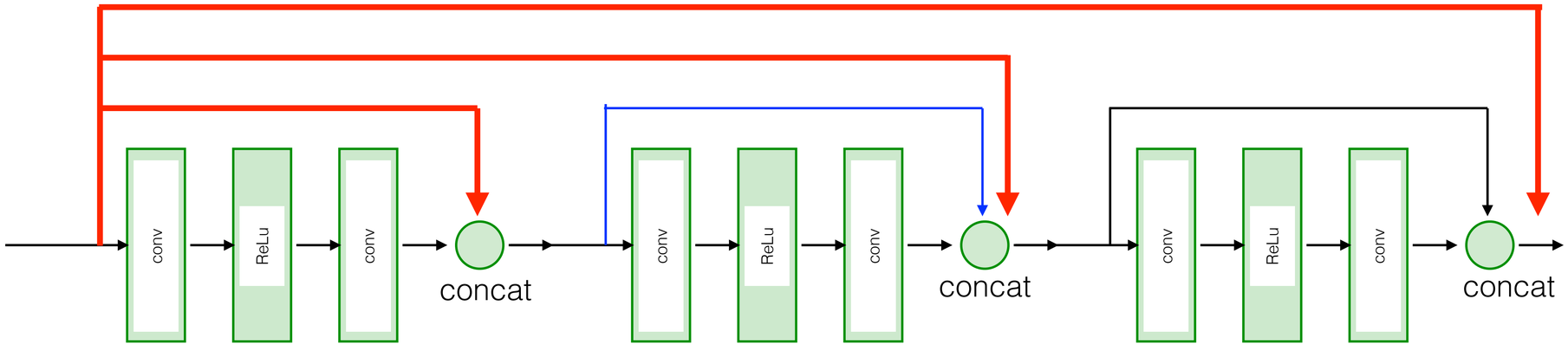
ResNet=>DenseNet



- Directly propagate each feature map to all following layers

Huang, Densely Connected Convolutional Networks, CVPR 2017. https://arxiv.org/abs/1608.06993

DenseNet

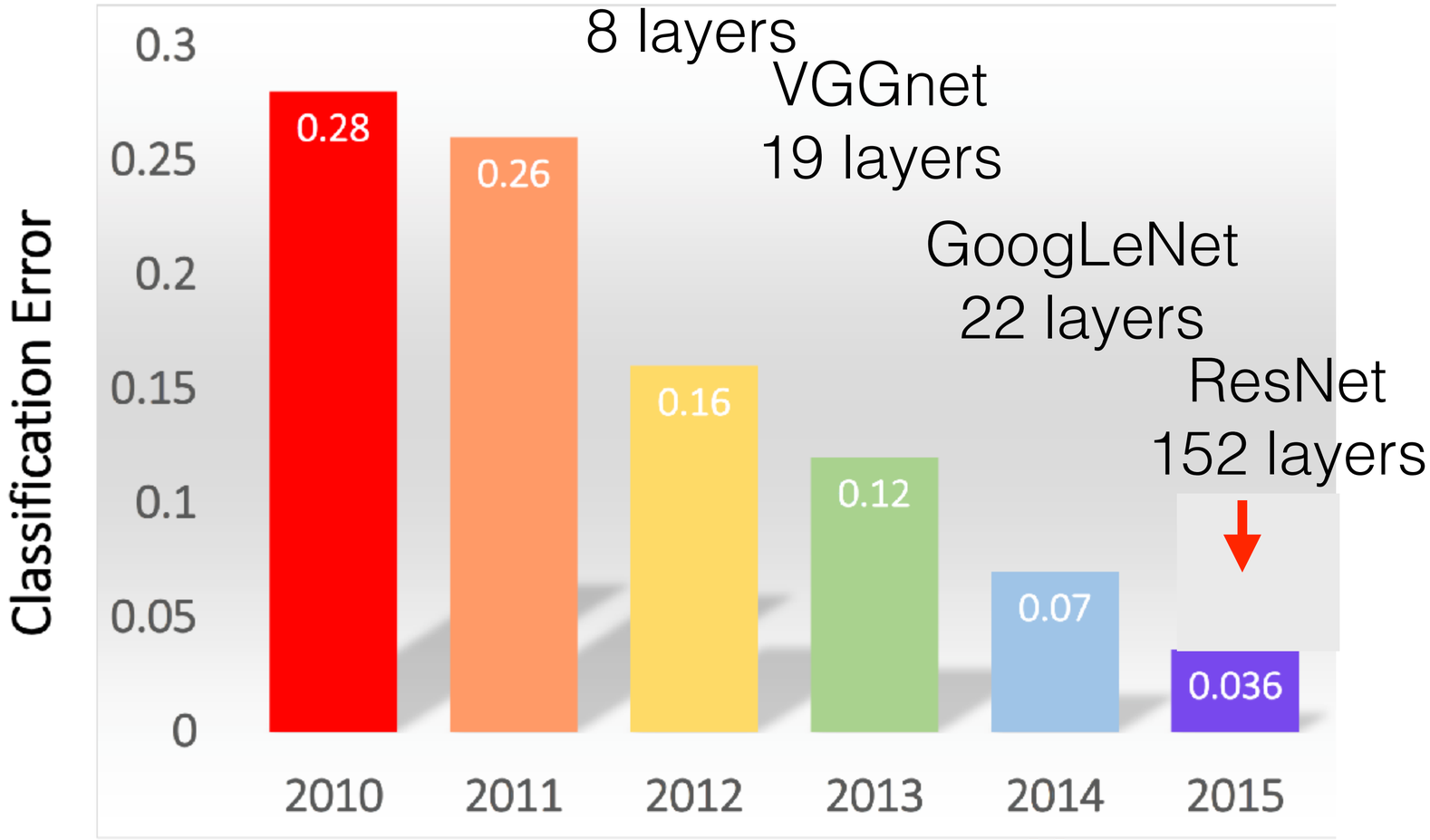


- Directly propagate each feature map to all following layers

Huang, Densely Connected Convolutional Networks, CVPR 2017.
https://arxiv.org/abs/1608.06993

IMAGENET

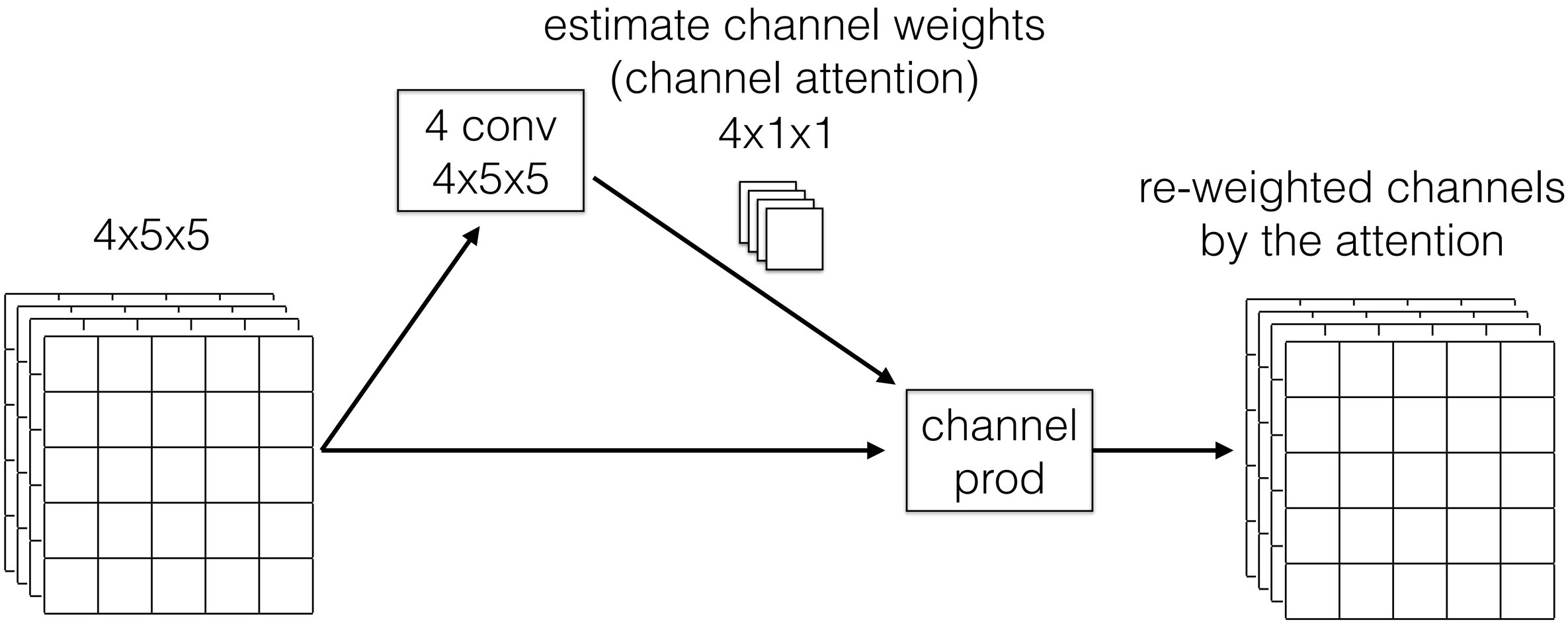
Classification results



Human error around 5%

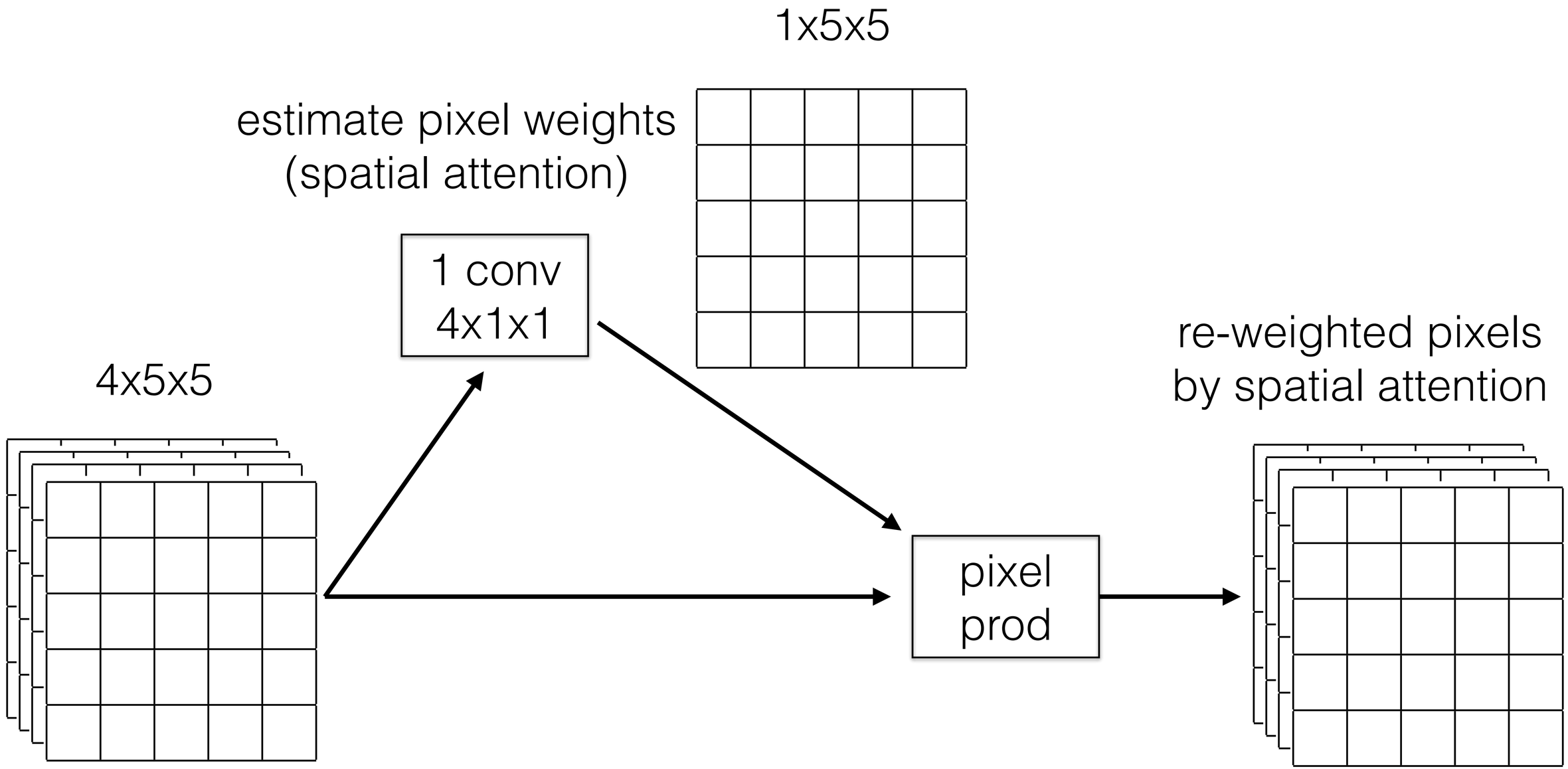
Squeeze and Excitation Networks [Hu et al, CVPR oral, 2017]
https://arxiv.org/pdf/1709.01507.pdf

Channel attention

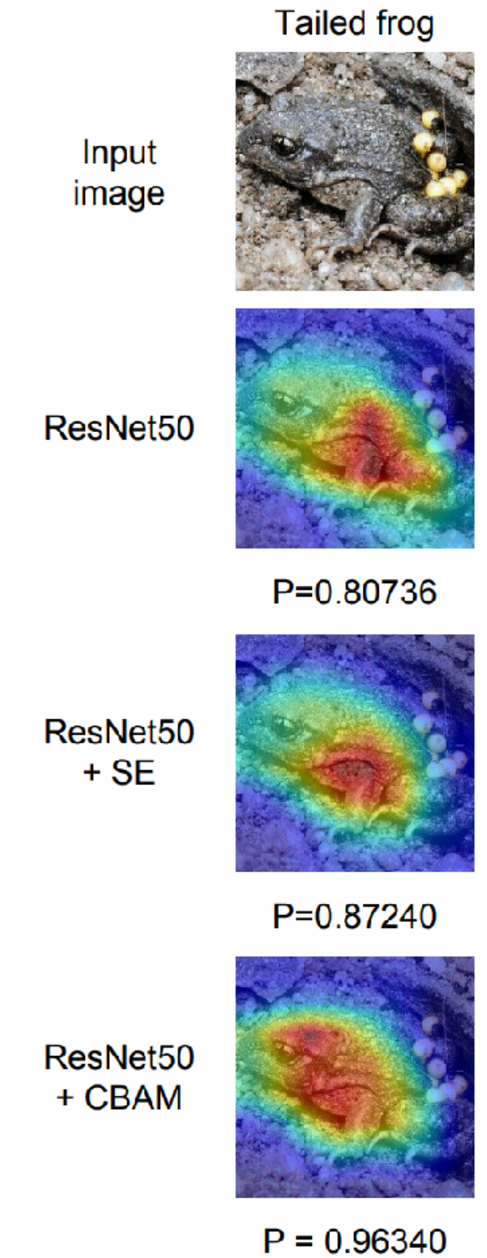


GradCAM [Selvaraju et al,ICCV, 2018]

Spatial attention



Attention modules [Woo et al,ECCV, 2018]
https://arxiv.org/pdf/1807.06521v2.pdf



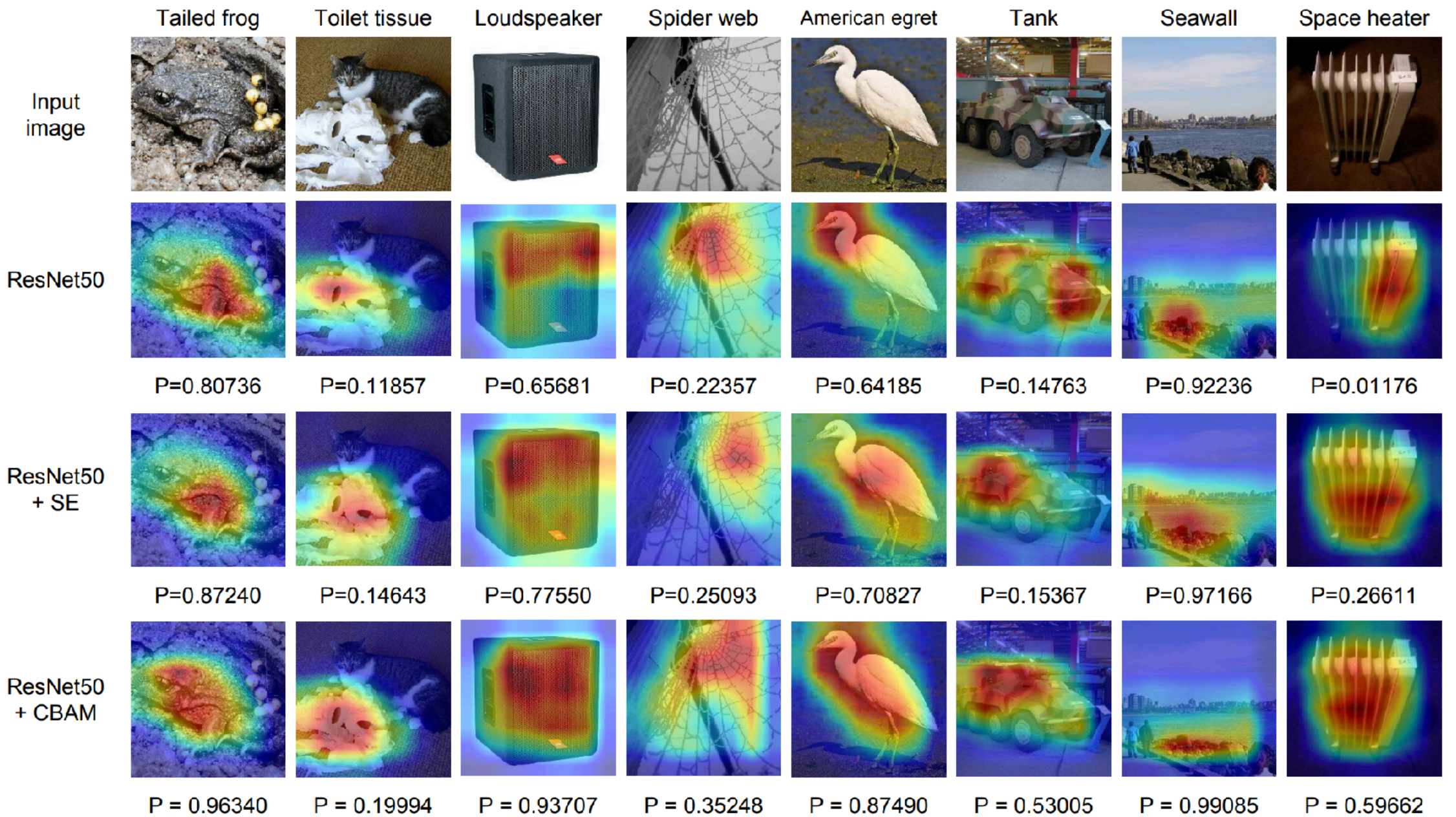
The attention map of classified object is better localized

w/o attention

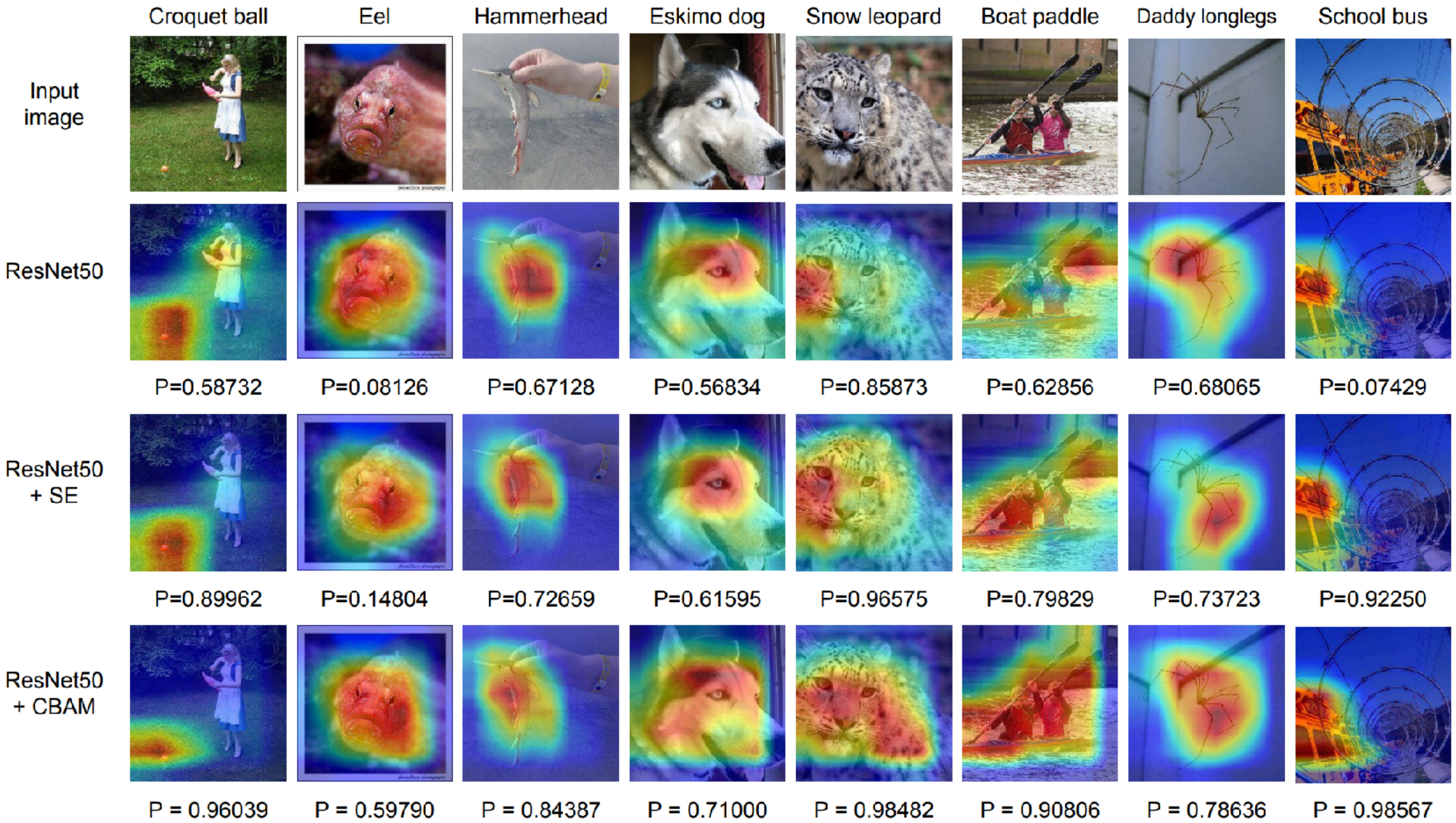
with channel attention

with spatial attention

Attention modules [Woo et al,ECCV, 2018]
https://arxiv.org/pdf/1807.06521v2.pdf



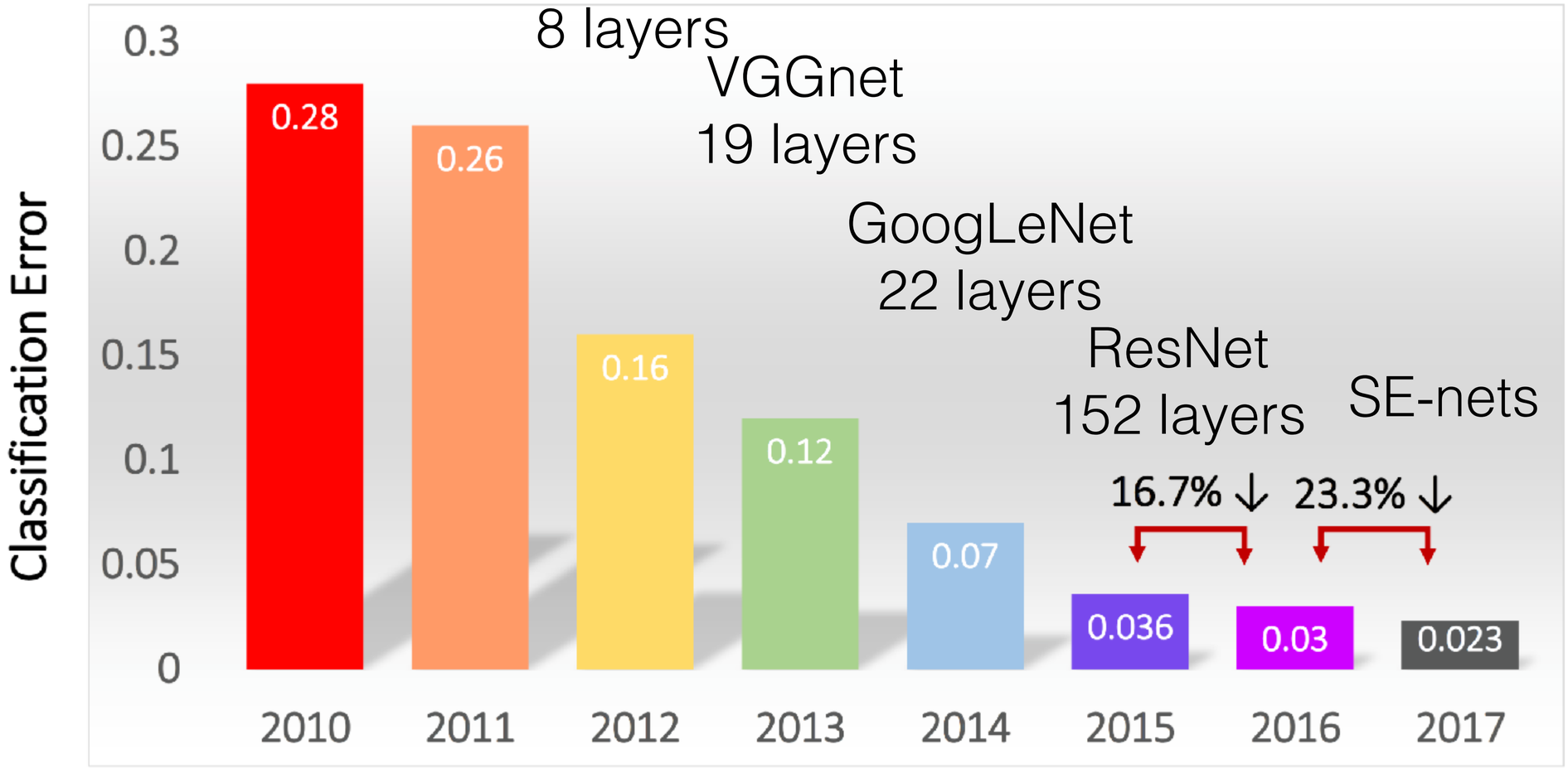
Attention modules [Woo et al,ECCV, 2018]
https://arxiv.org/pdf/1807.06521v2.pdf



GradCAM [Selvaraju et al,ICCV, 2018]

IMAGENET

Classification results



Summary classification

- Injecting gradient
- Skip connections
- Receptive field (“one 7x7” vs “three 3x3 convs”)
- Spatial and Channel Attention

Outline

- Architectures of classification networks
- Architectures of segmentation networks
- Architectures of regression networks
- Architectures of detection networks
- Architectures of regression networks
- Architectures of feature matching networks

Semantic segmentation

road

sideway

pedestrian

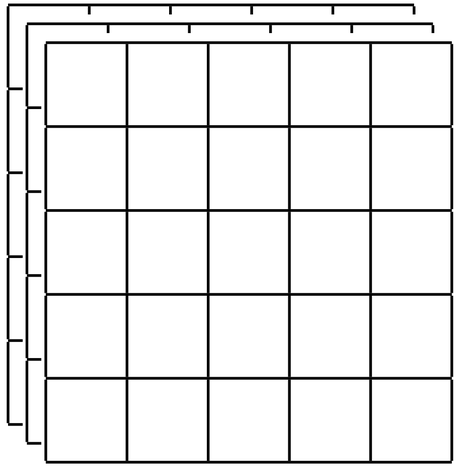
traffic sign

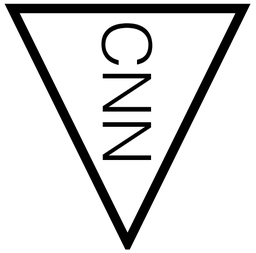
trees

sky

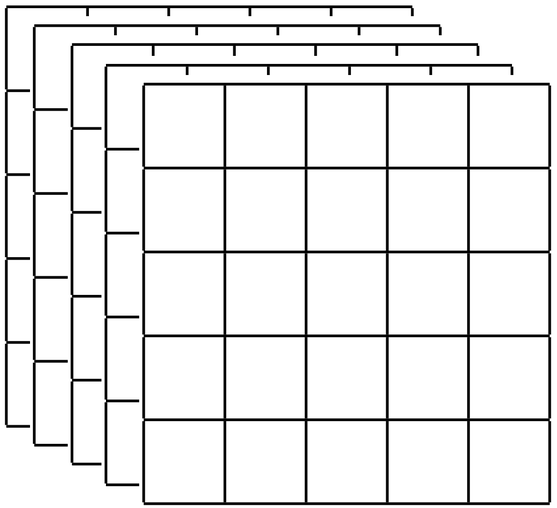
Semantic segmentation

RGB image
(HxWx3)





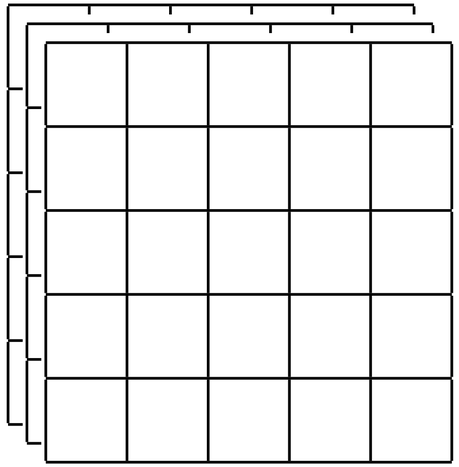
pixel-wise probs
(HxWxN)

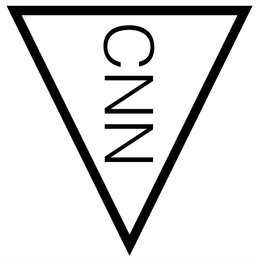


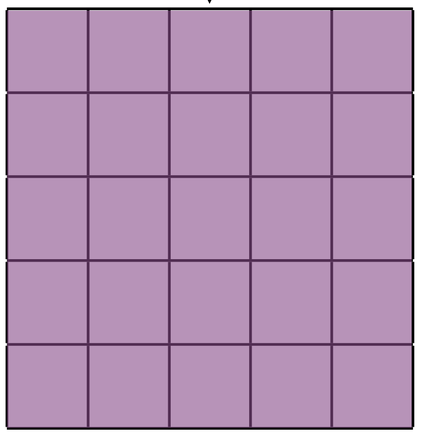
- road
- sideway
- pedestrian
- traffic sign
- trees
- sky

Semantic segmentation

RGB image
(HxWx3)







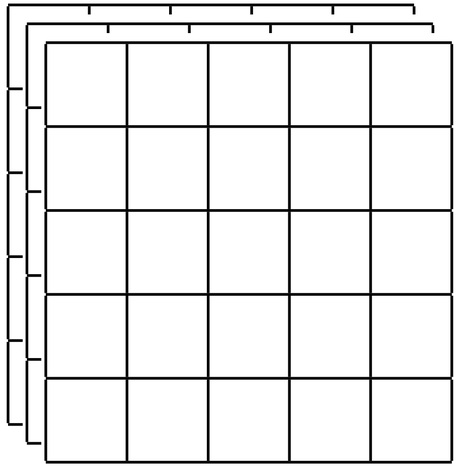
road
sideway
pedestrian
traffic sign
trees
sky

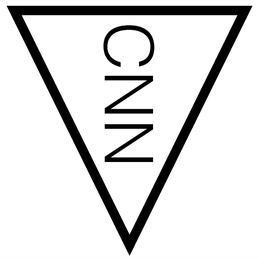
pixel-wise probability
of being **road**

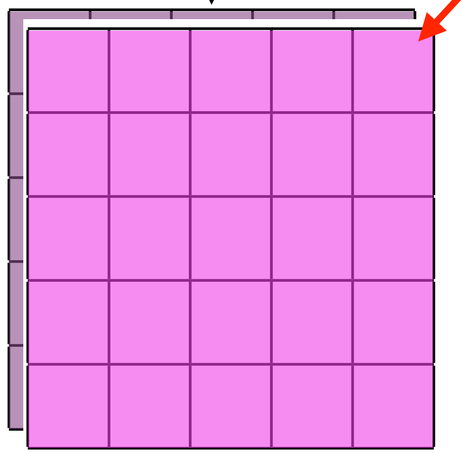
channel 1

Semantic segmentation

RGB image
(HxWx3)







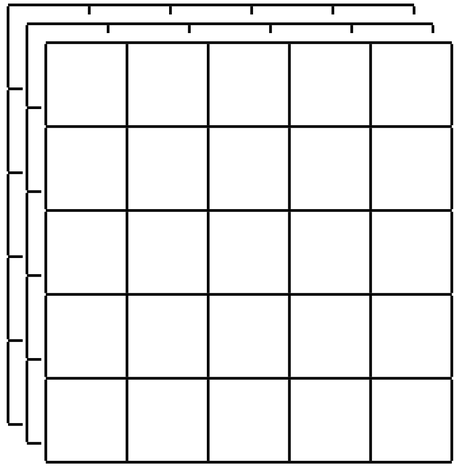
road
sideway
pedestrian
traffic sign
trees
sky

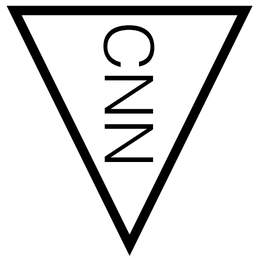
pixel-wise probability
of being **sideway**

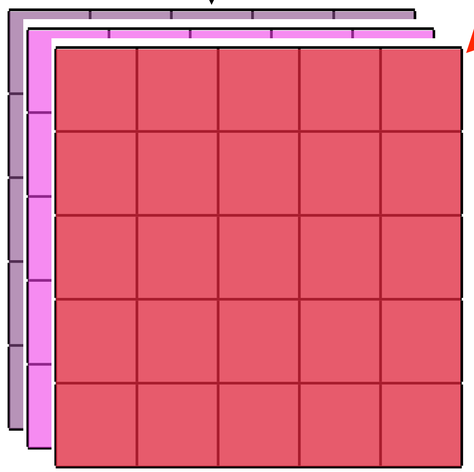
channel 2

Semantic segmentation

RGB image
(HxWx3)







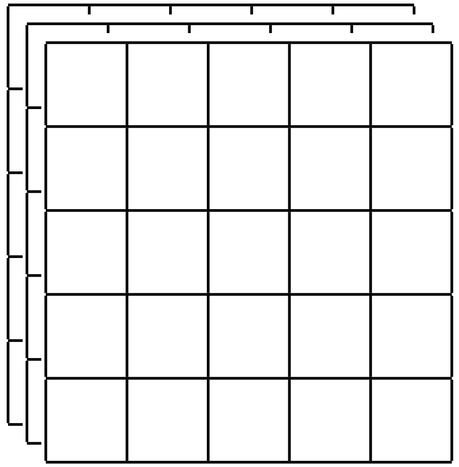
road
sideway
pedestrian
traffic sign
trees
sky

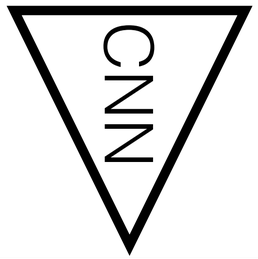
pixel-wise probability
of being **pedestrian**

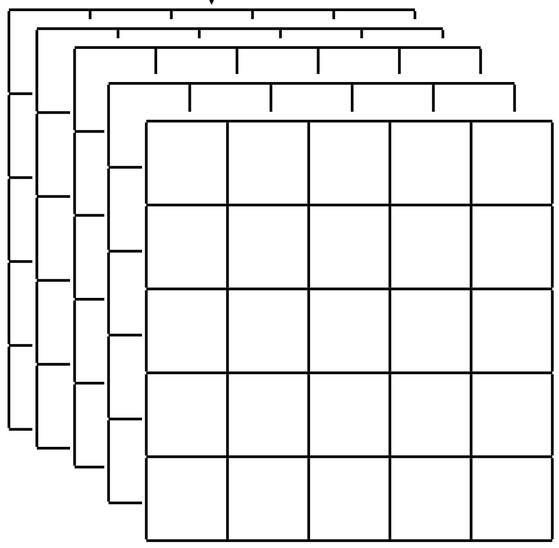
channel 3

Semantic segmentation

RGB image
(HxWx3)







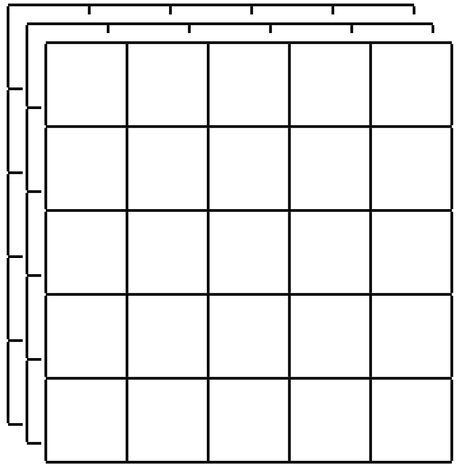
pixel-wise probs
(HxWxN)

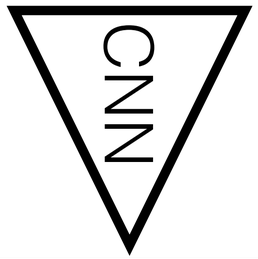
- road
- sideway
- pedestrian
- traffic sign
- trees
- sky

as many output channels
as semantic labels

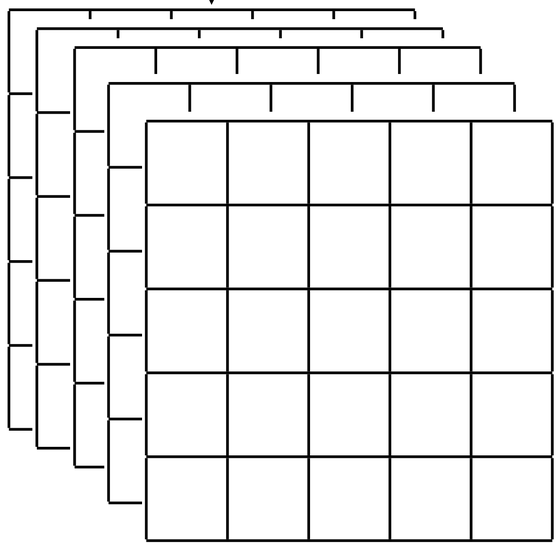
Semantic segmentation

RGB image
(HxWx3)

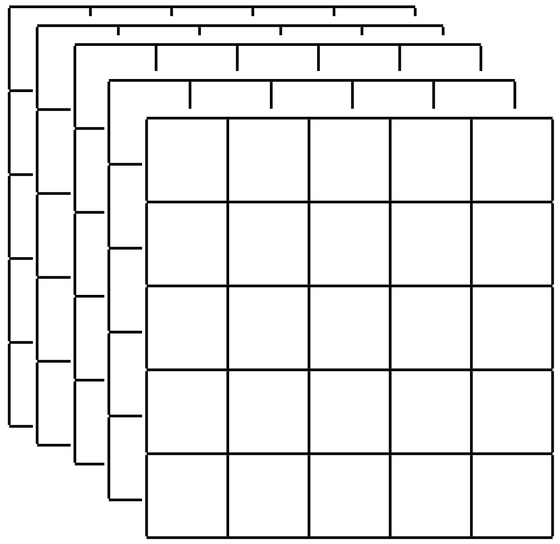




pixel-wise probs
(HxWxN)



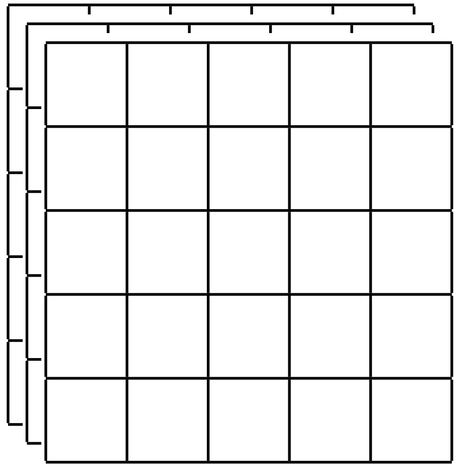
ground truth (1-hot encoding)



annotations
(HxWxN)

Semantic segmentation

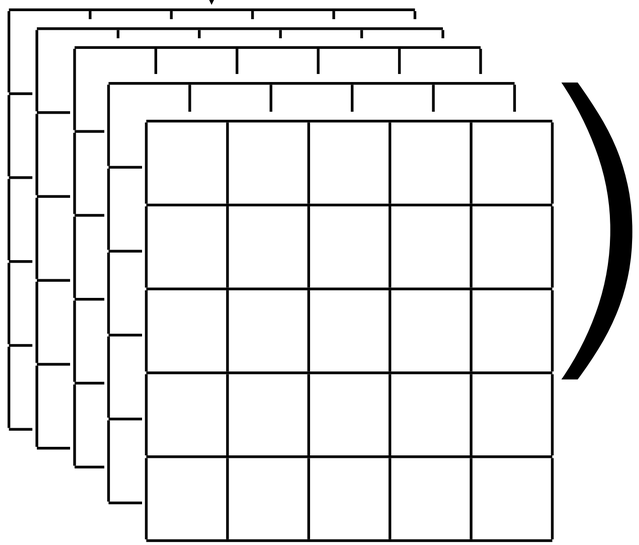
RGB image
(HxWx3)



CNN

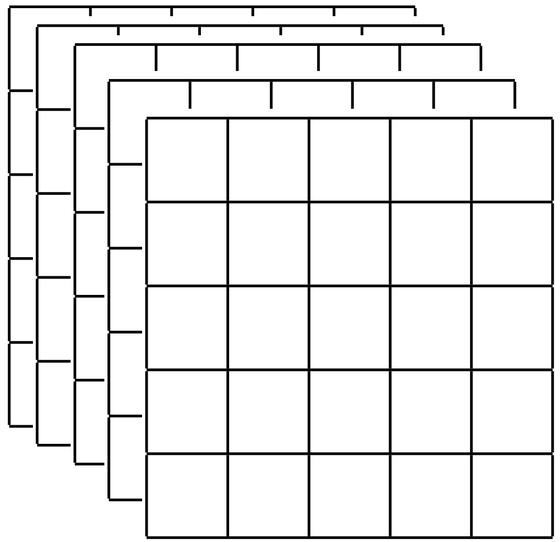
pixel-wise CE-loss
(HxWxN)

$$-\log\Big(\quad\Big)$$



$\odot$

ground truth (1-hot encoding)

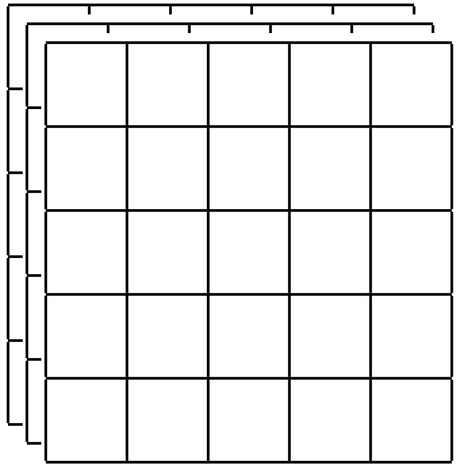


annotations
(HxWxN)

Semantic segmentation

Pixel-wise classifier with the loss summed over all pixels

RGB image
(HxWx3)

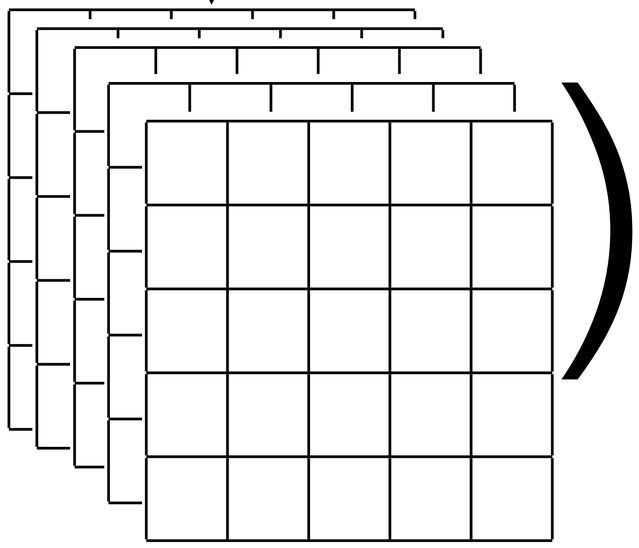


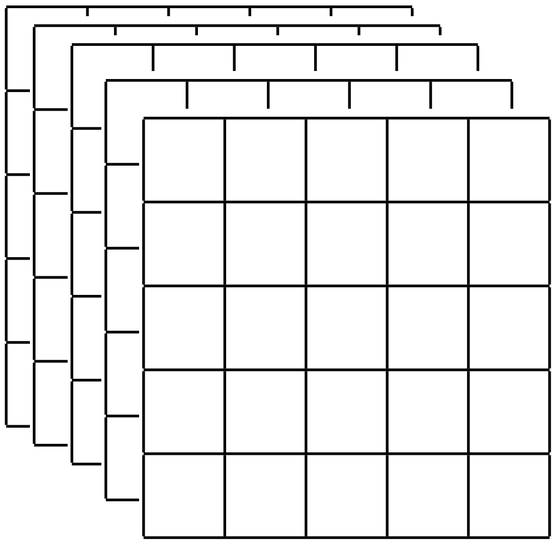
CNN

ground truth (1-hot encoding)

pixel-wise CE-loss
(HxWxN)

$$\sum_{\substack{\text{pixels} \\ \text{channels}}} -\log\Big(\Big) \odot$$





annotations
(HxWxN)

U-net architecture

Mirror the usual classification network

[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

U-net architecture

[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

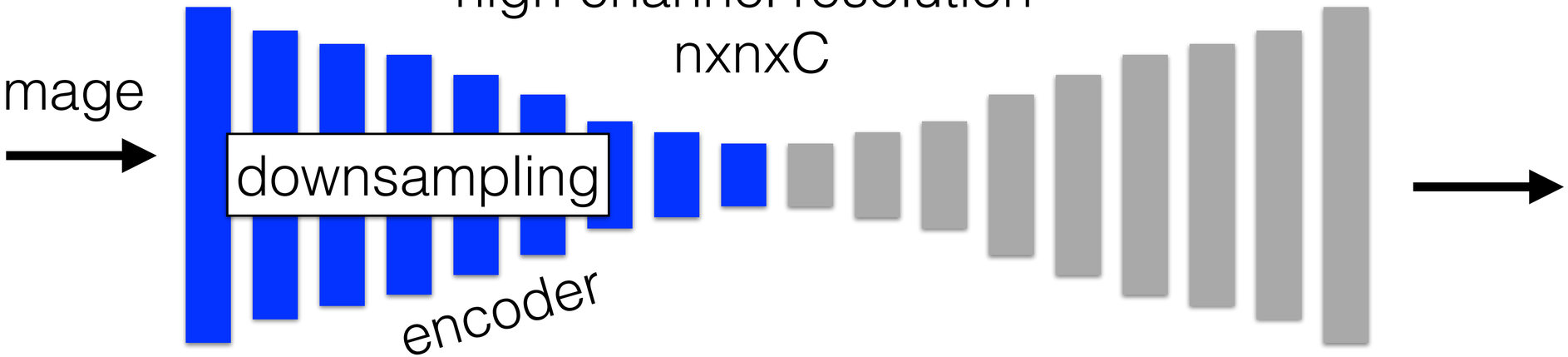
U-net architecture

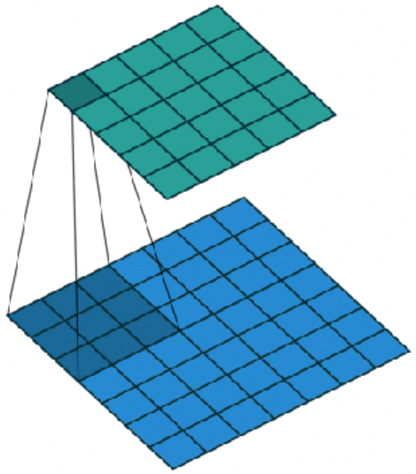
high spatial resolution
NxNxc

image embedding

high channel resolution
nxnxC

image





convolution layers

[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

U-net architecture

high spatial resolution
NxNxc

image embedding

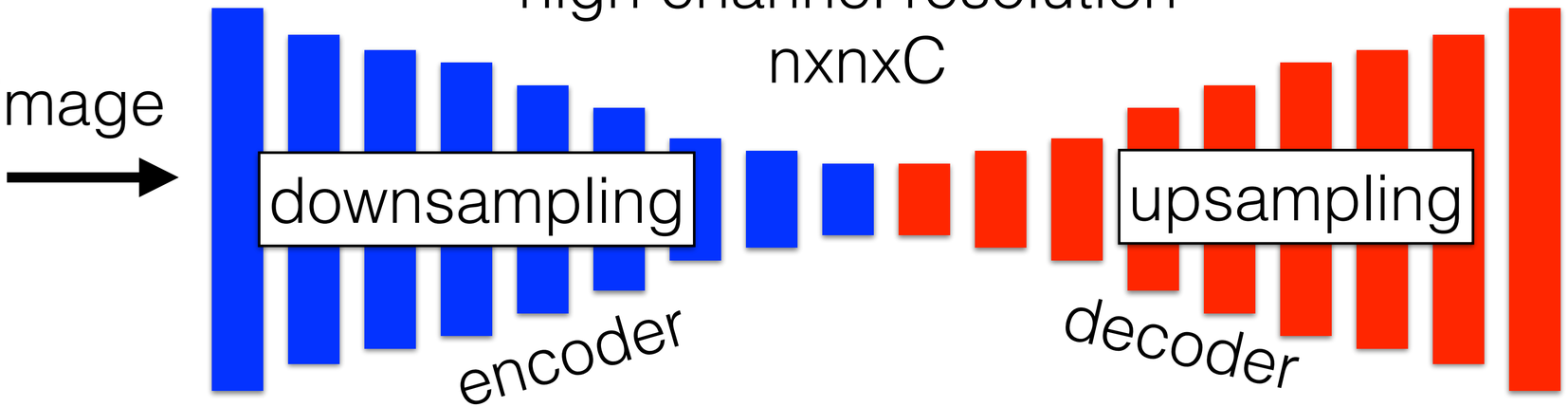
high channel resolution
nxnxC

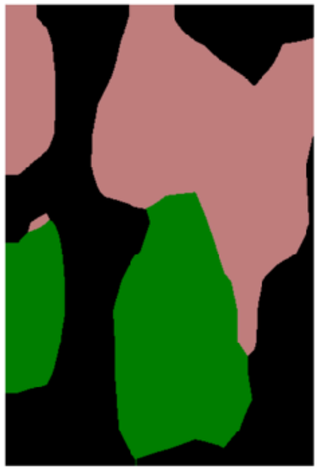
high spatial resolution
NxNxc

Why???

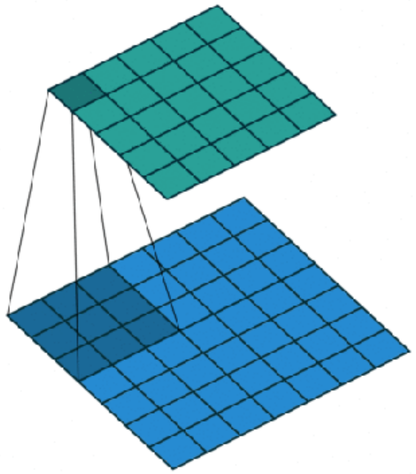
output

image

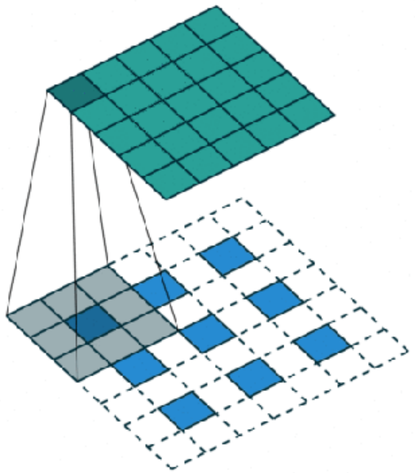




ground truth



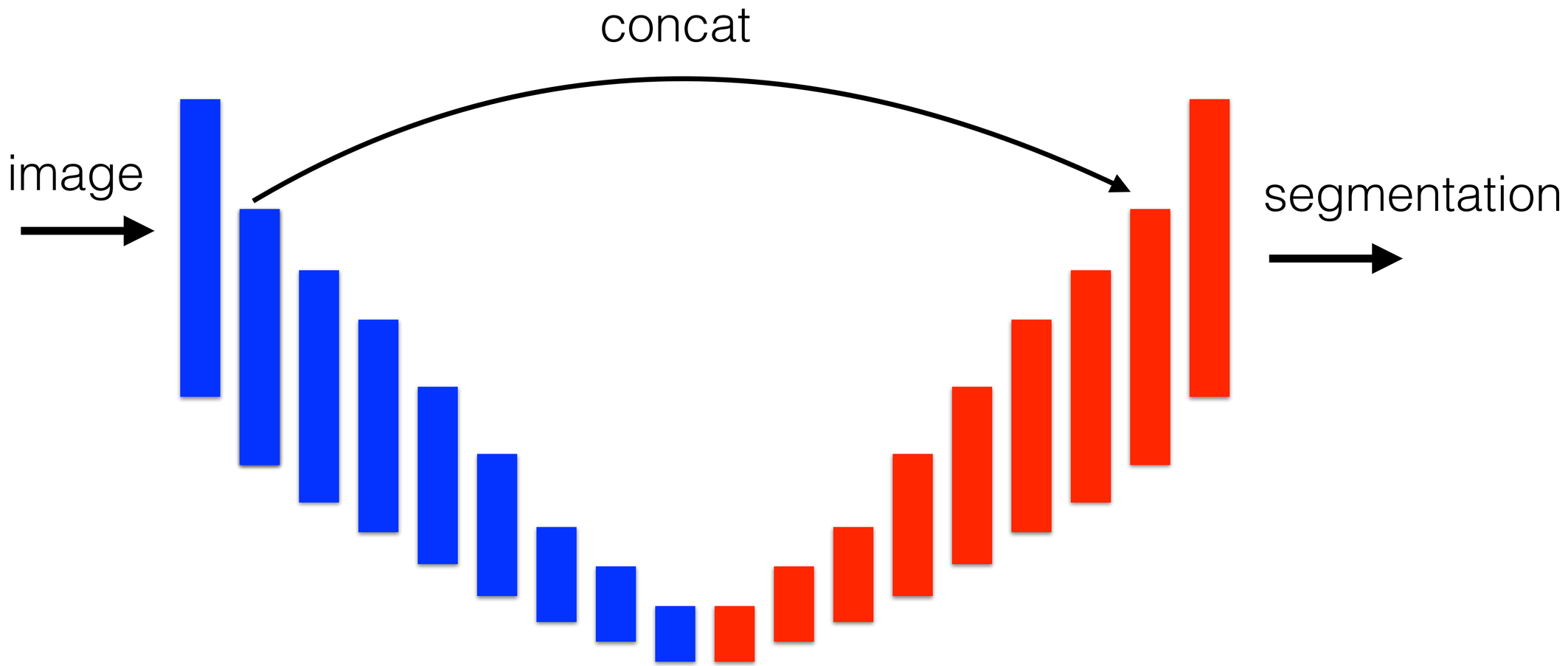
convolution layers



deconvolution layers

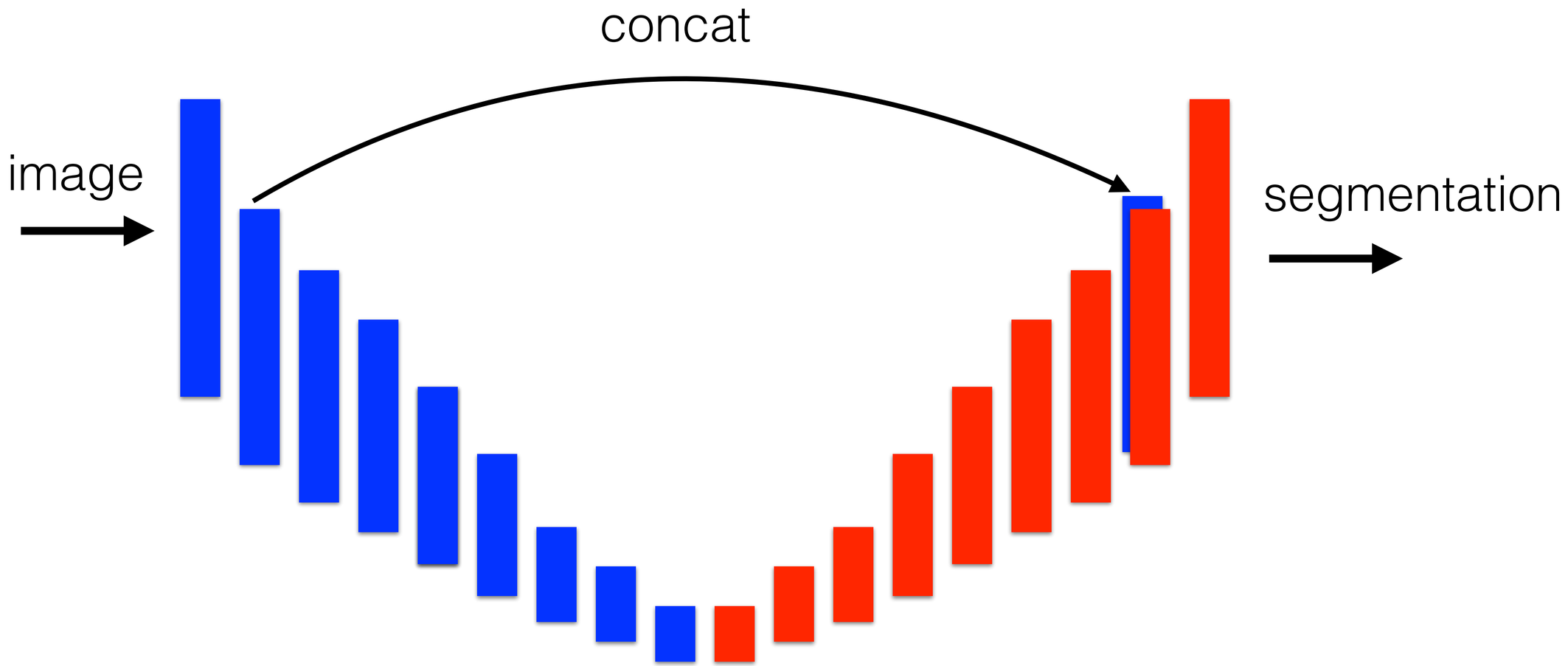
[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

U-net architecture



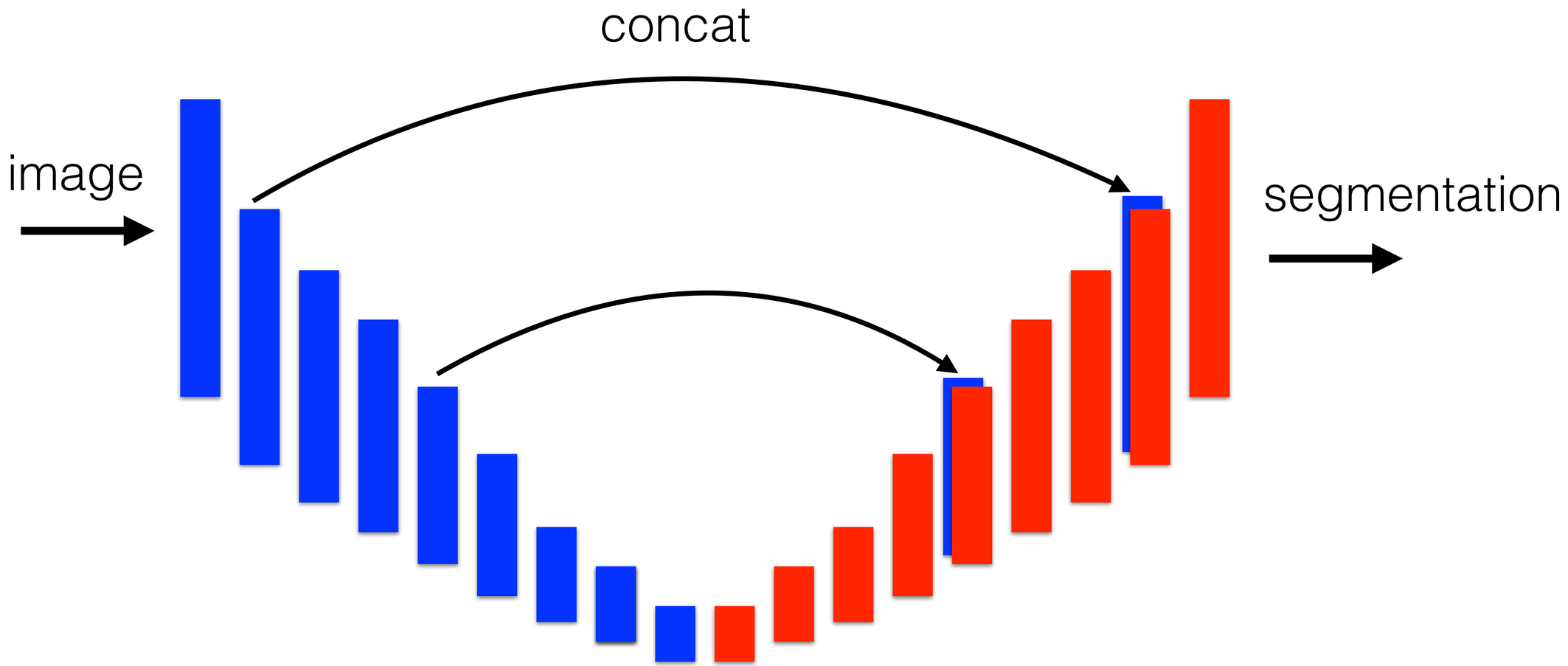
[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

U-net architecture



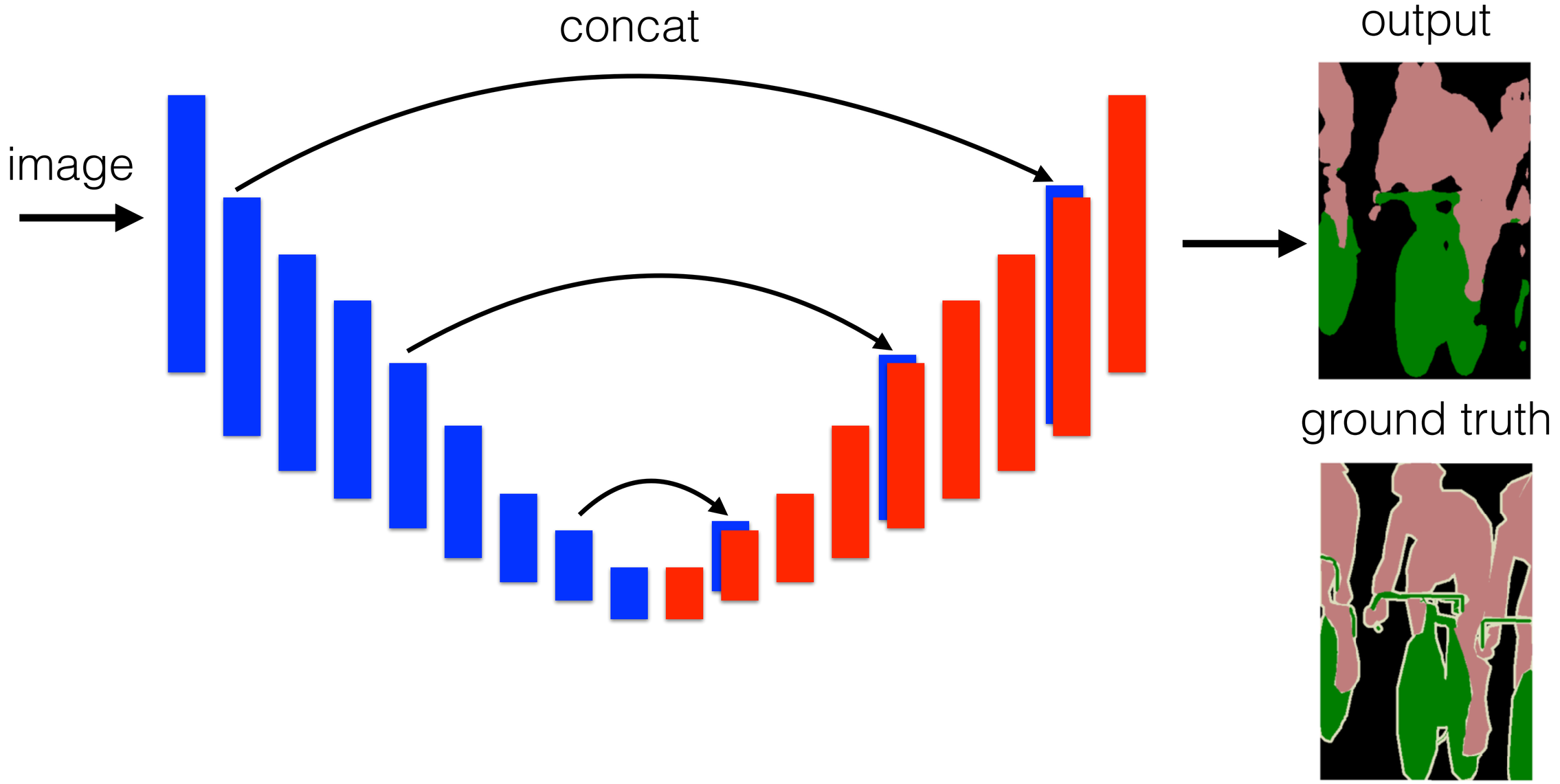
[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

U-net architecture



[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

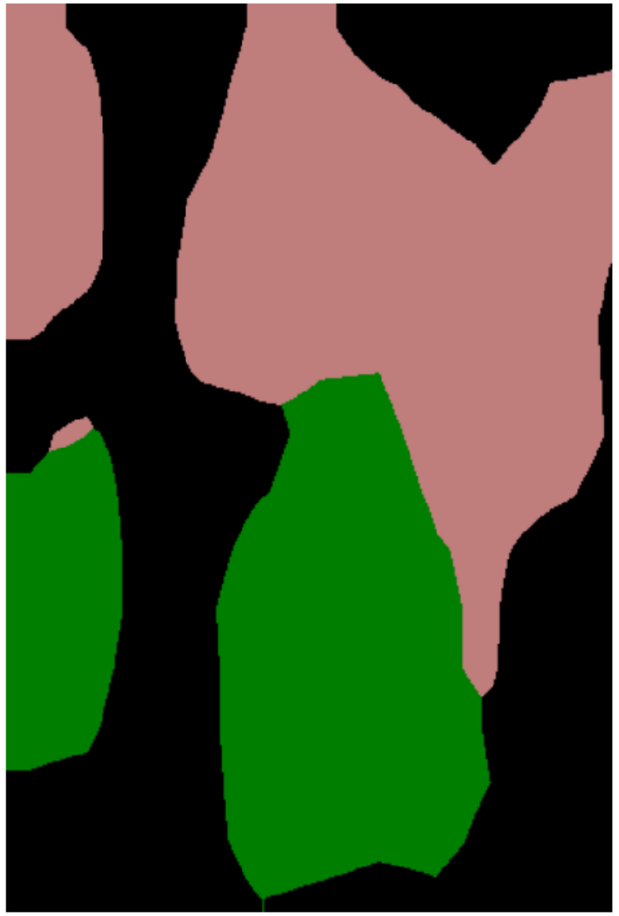
U-net architecture



[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

U-net architecture

no skip connections



with skip connections

ground truth

[Noh et al ICCV 2015] https://arxiv.org/pdf/1505.04366.pdf

Segmentation architectures

- **FCN** (Fully Convolutional Network): LongCVPR 2015, FCN was one of the pioneering architectures for end-to-end pixel-wise prediction.
- **U-Net**: symmetric architecture and skip connections, which helps in capturing both global and local features.
- **DeepLab**: DeepLab uses dilated convolutions to enlarge the field of view and capture multi-scale information. DeepLabv3 is an improved version.
- **MobileNet**: lightweight architecture designed for real-time semantic segmentation. It is known for its efficiency and speed.
- **LinkNet**: LinkNet employs a skip connection structure with a series of encoder and decoder blocks for segmentation tasks.
- **HRNet** (High-Resolution Network): HRNet focuses on maintaining high-resolution representations throughout the network, which can be beneficial for capturing fine details.
- **ViTS**: Vision tranformer exptended for segmentation task
- **SAM**: Segment anything architecture from Facebook 2023