MML

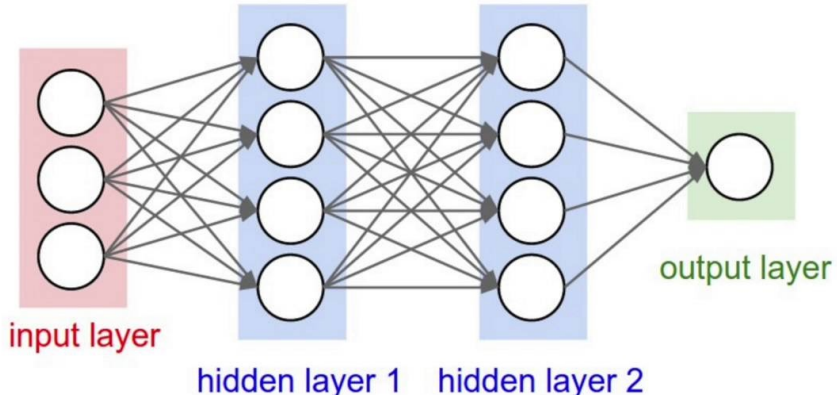
Convolutional Neural Networks Architectures

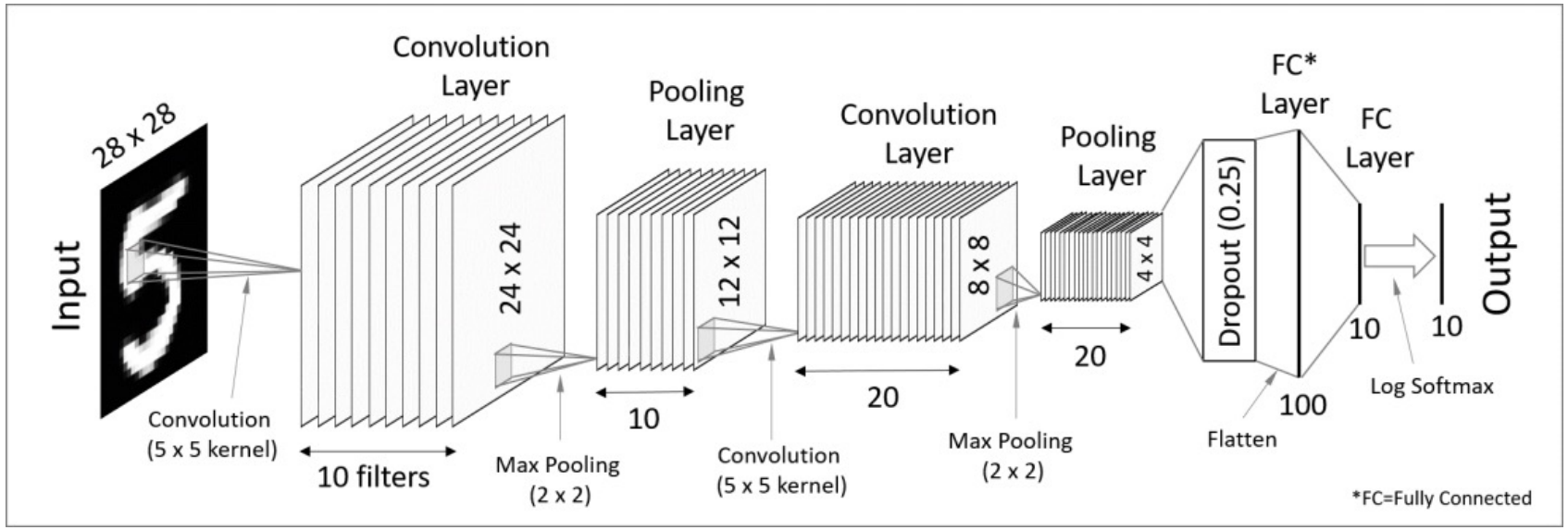
Executive Program on AI/ML

IIIT Hyderabad

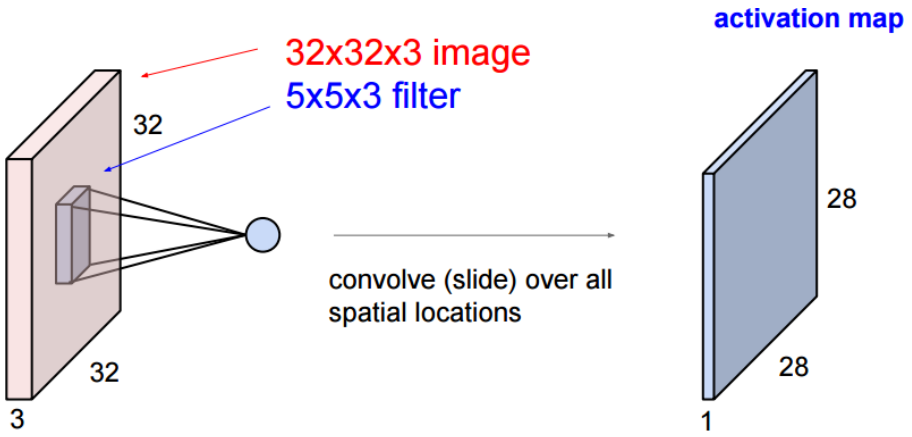
iHUB DATA
TECHNOLOGY INNOVATION HUB ON DATA DRIVEN TECHNOLOGIES

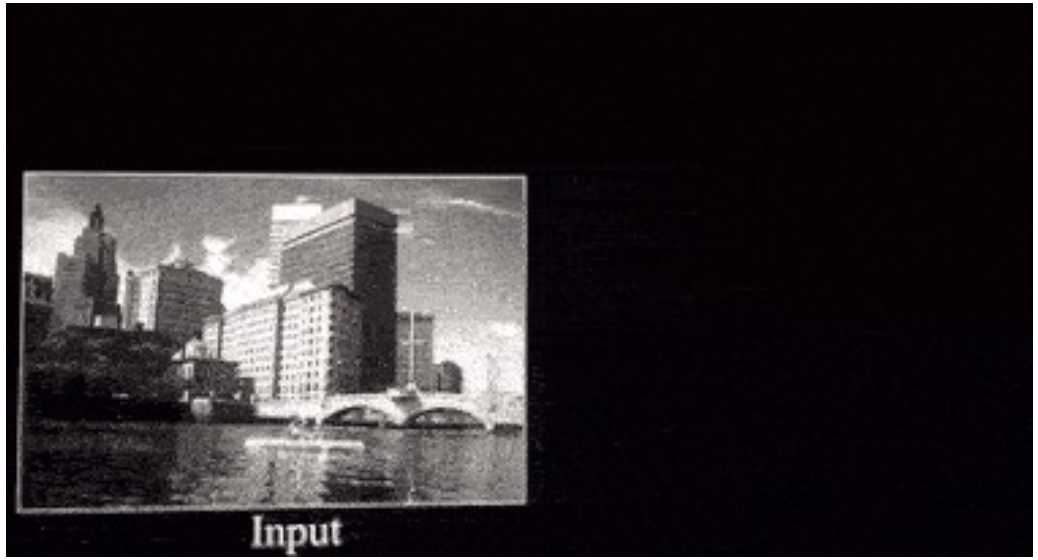
Convolutional Neural Network



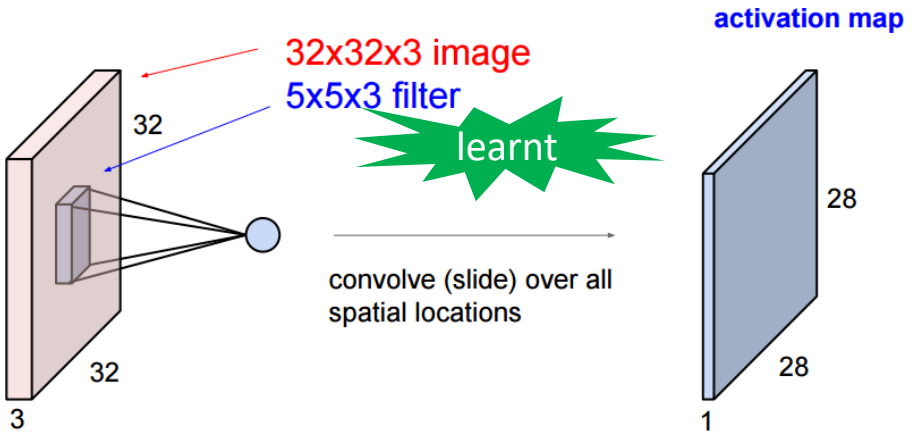


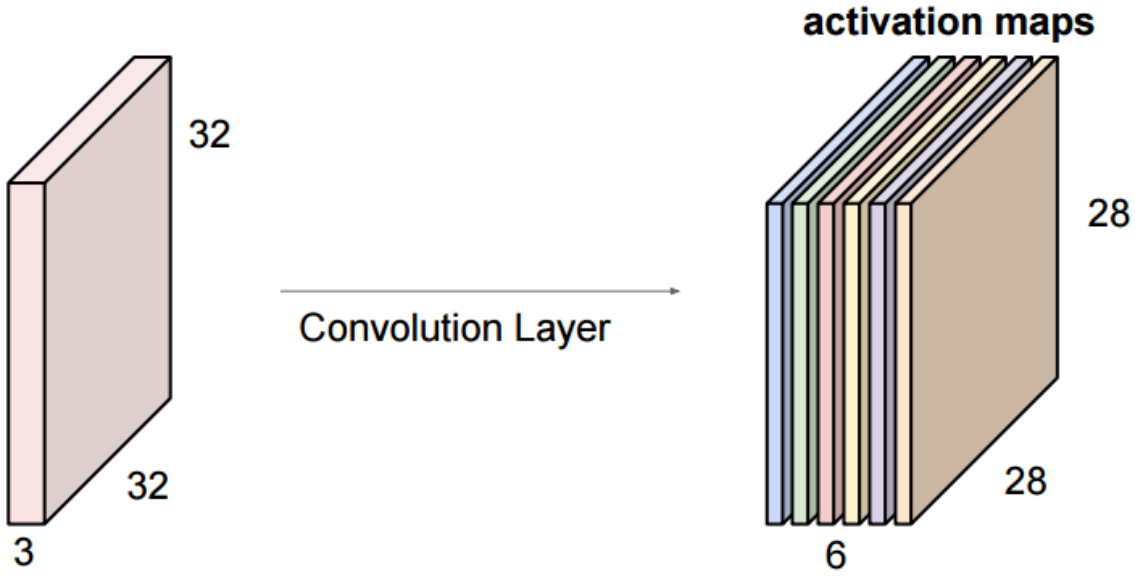
Convolutional Neural Network



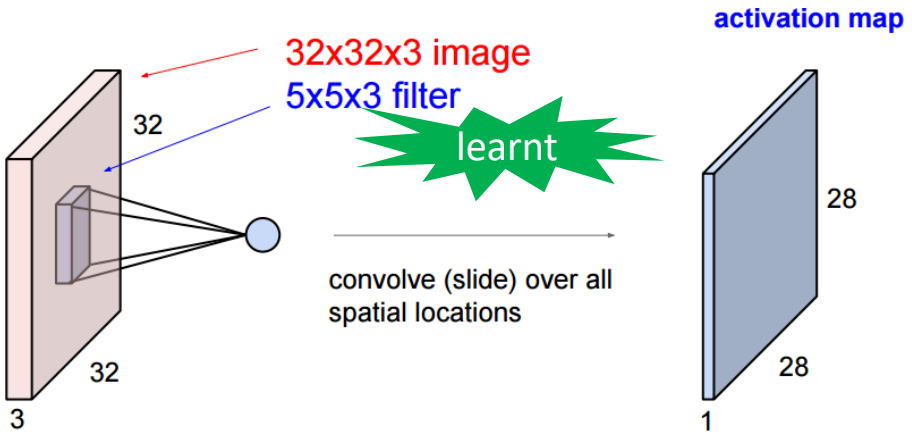


Convolution Layer

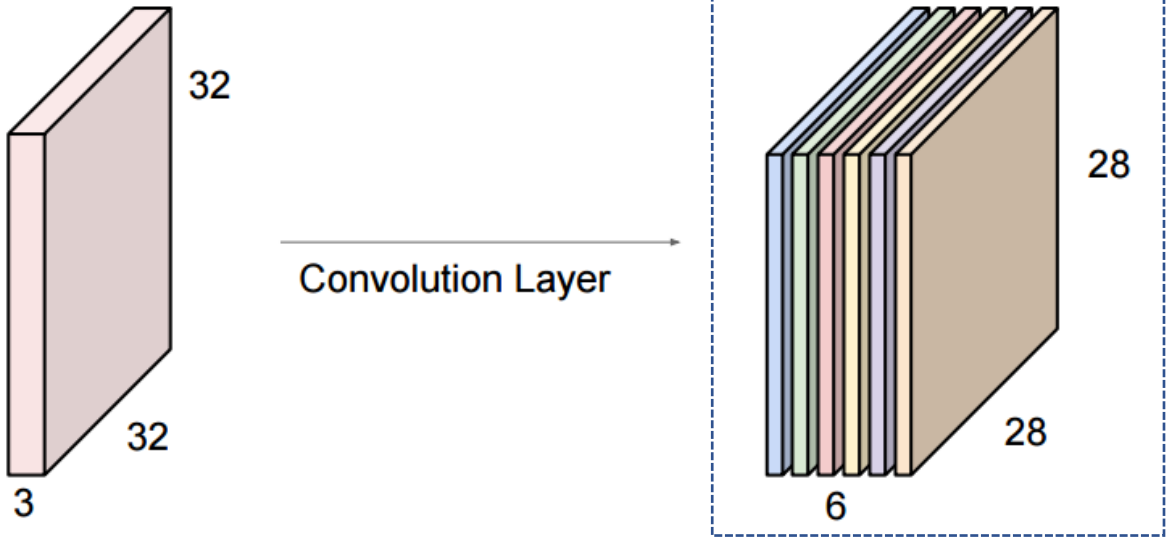




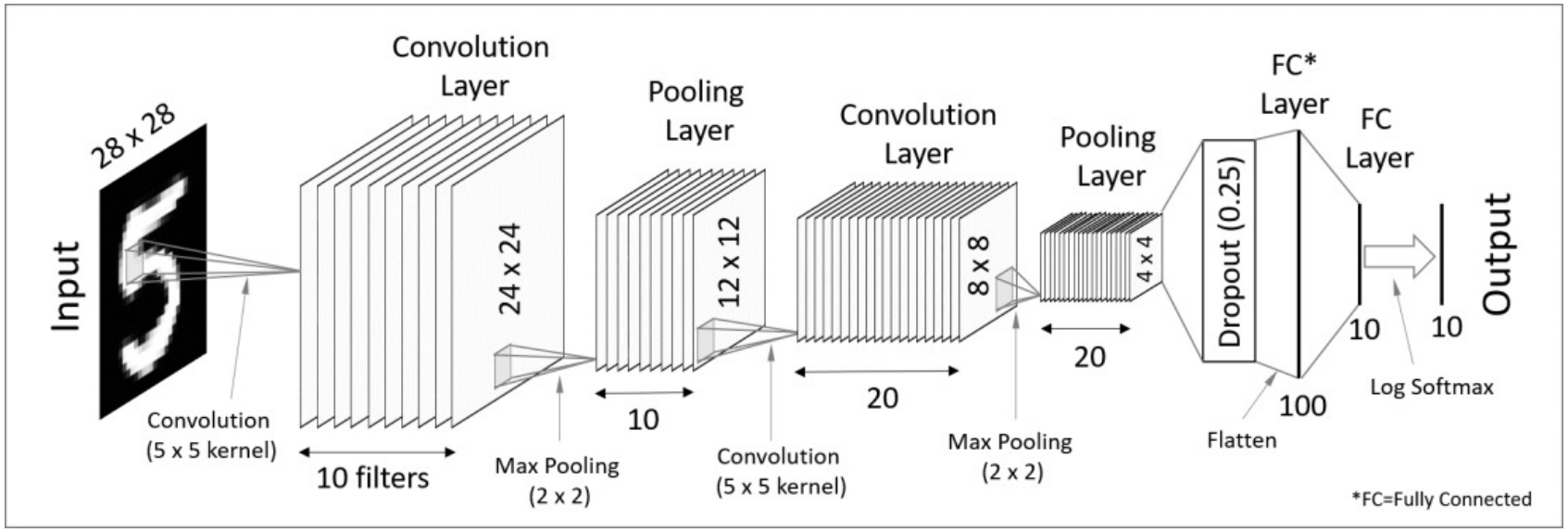
Convolution Layer



Input to next layer



Convolutional Neural Network



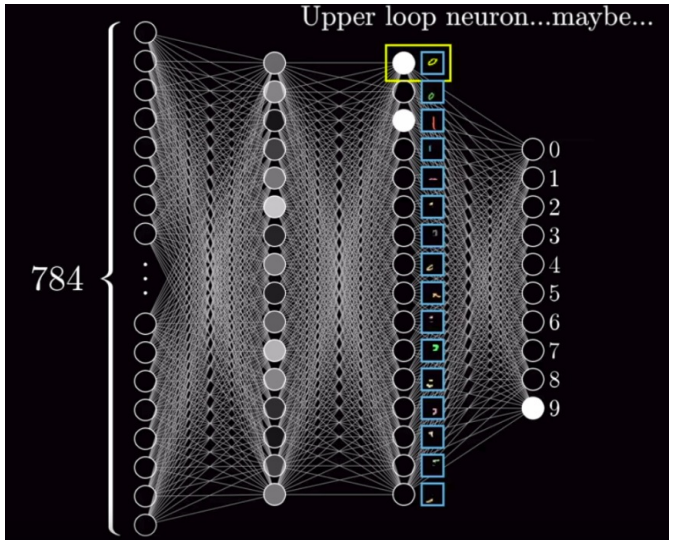
Pooling Layer

1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

max pool with 2x2 filters and stride 2 →

6	8
3	4

Fully-connected Layer



MML

Questions?

MML

ImageNet Challenge and AlexNet

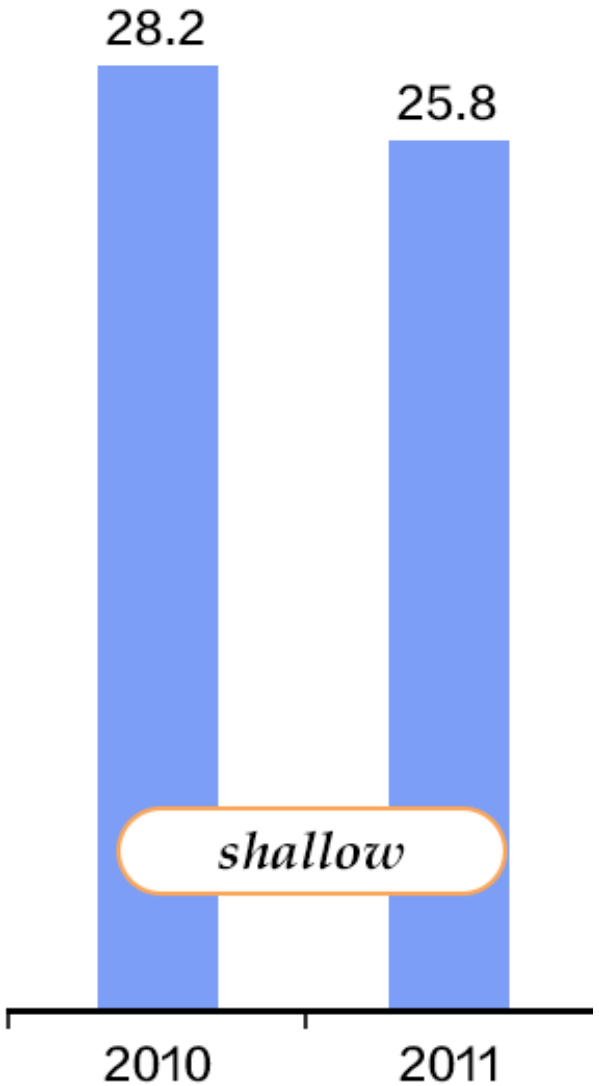
The net that triggered the deep learning revolution

ImageNet Challenge

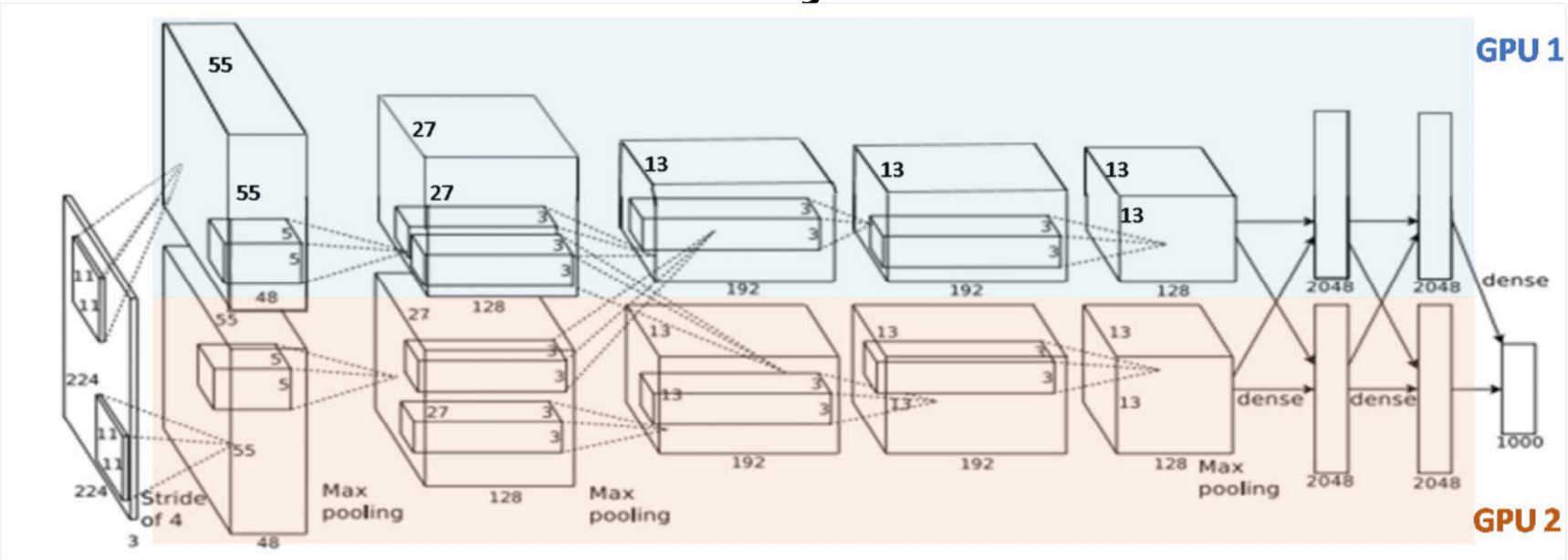
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

Top-5 misclassification error rate



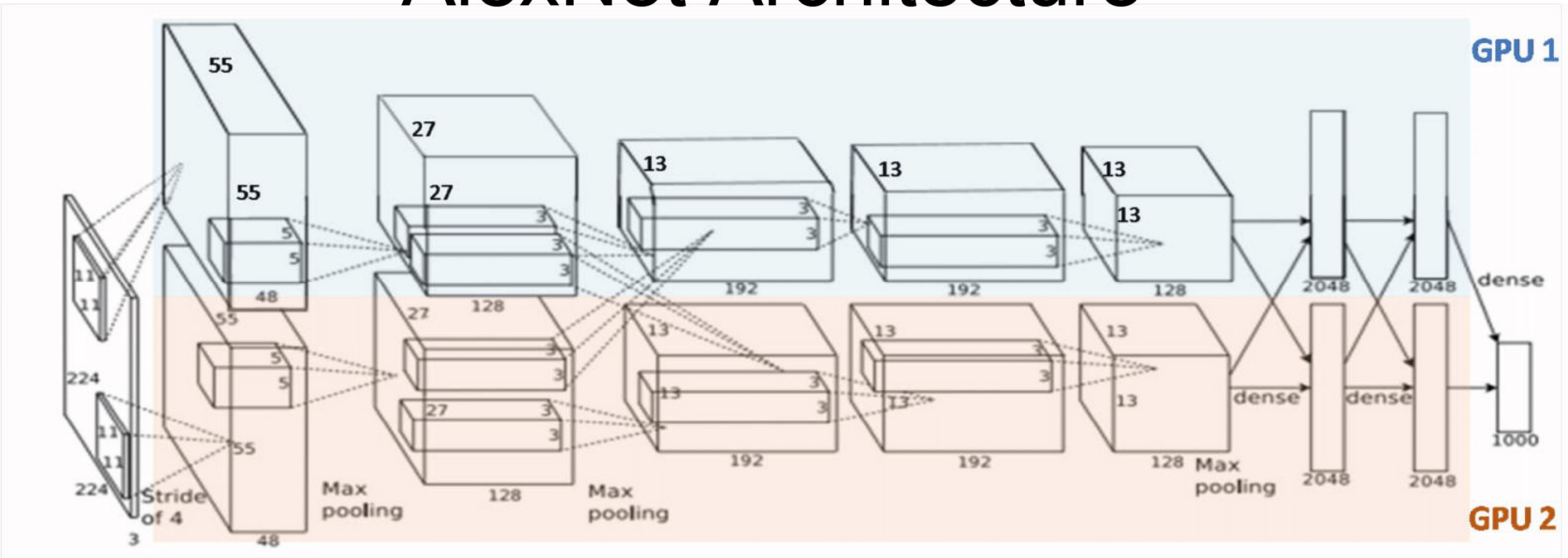
Case Study: AlexNet

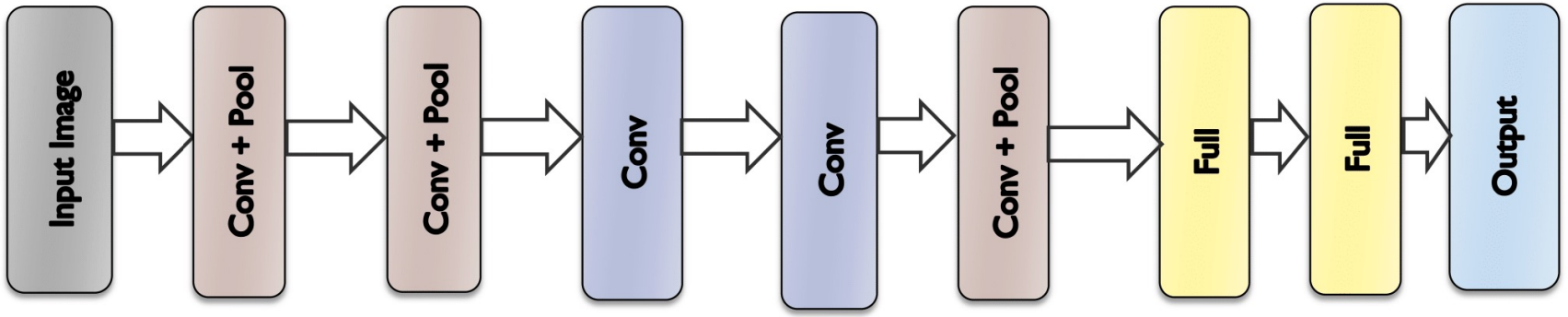


- Winner of ImageNet LSVRC-2010.
- Trained over 1.2M images using SGD with regularization.
- Deep architecture (60M parameters.)
- Optimized GPU implementation (cuda-convnet)

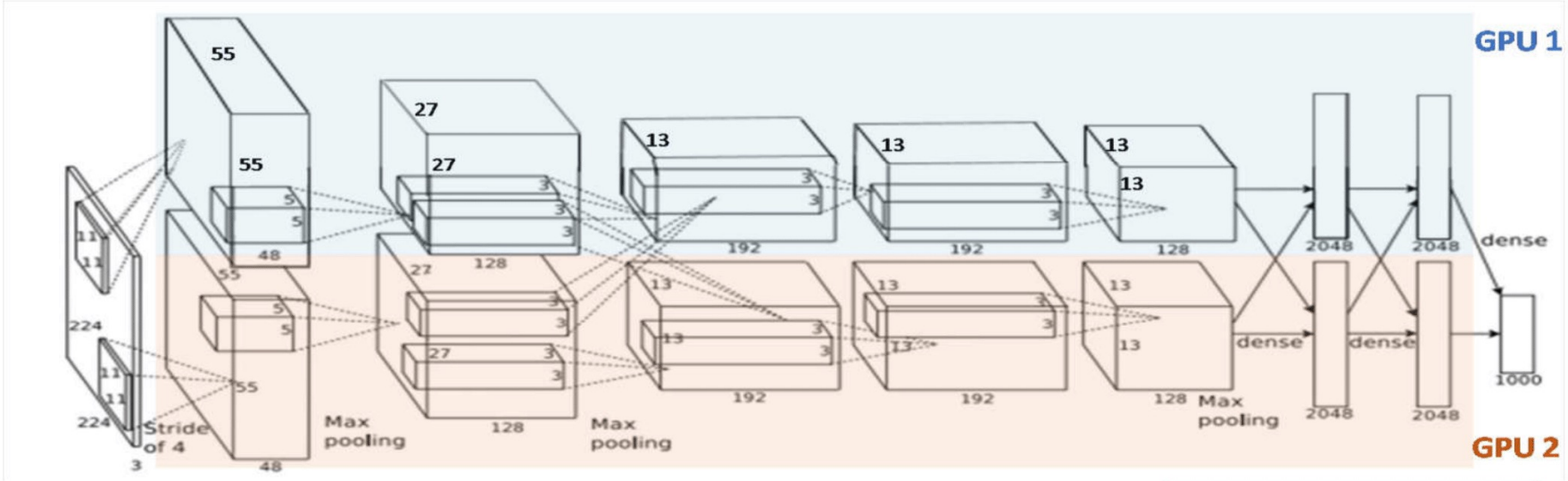
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *NIPS* 2012.

AlexNet Architecture





AlexNet Architecture

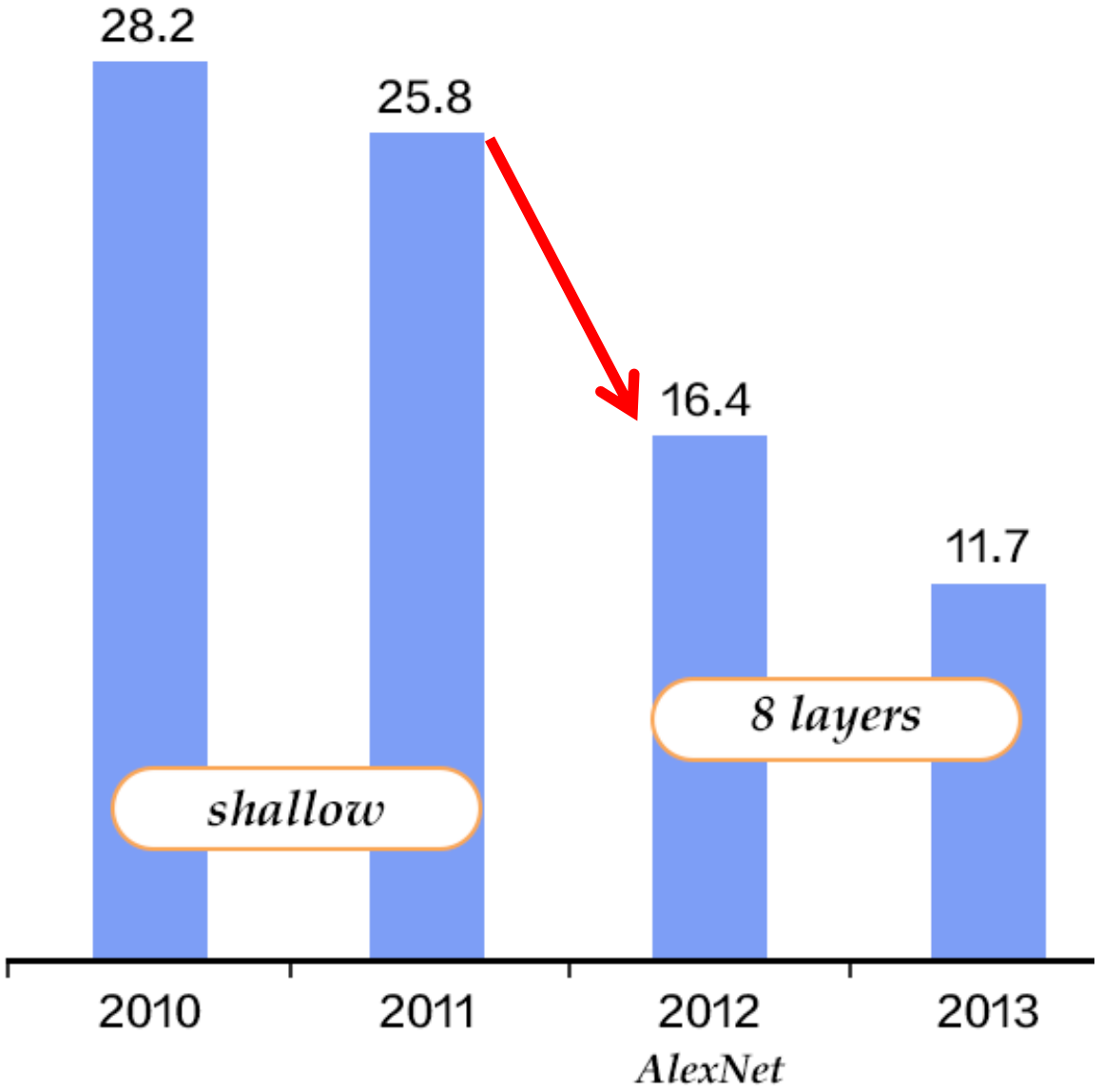


About 57 M parameters are in the fully connected layers

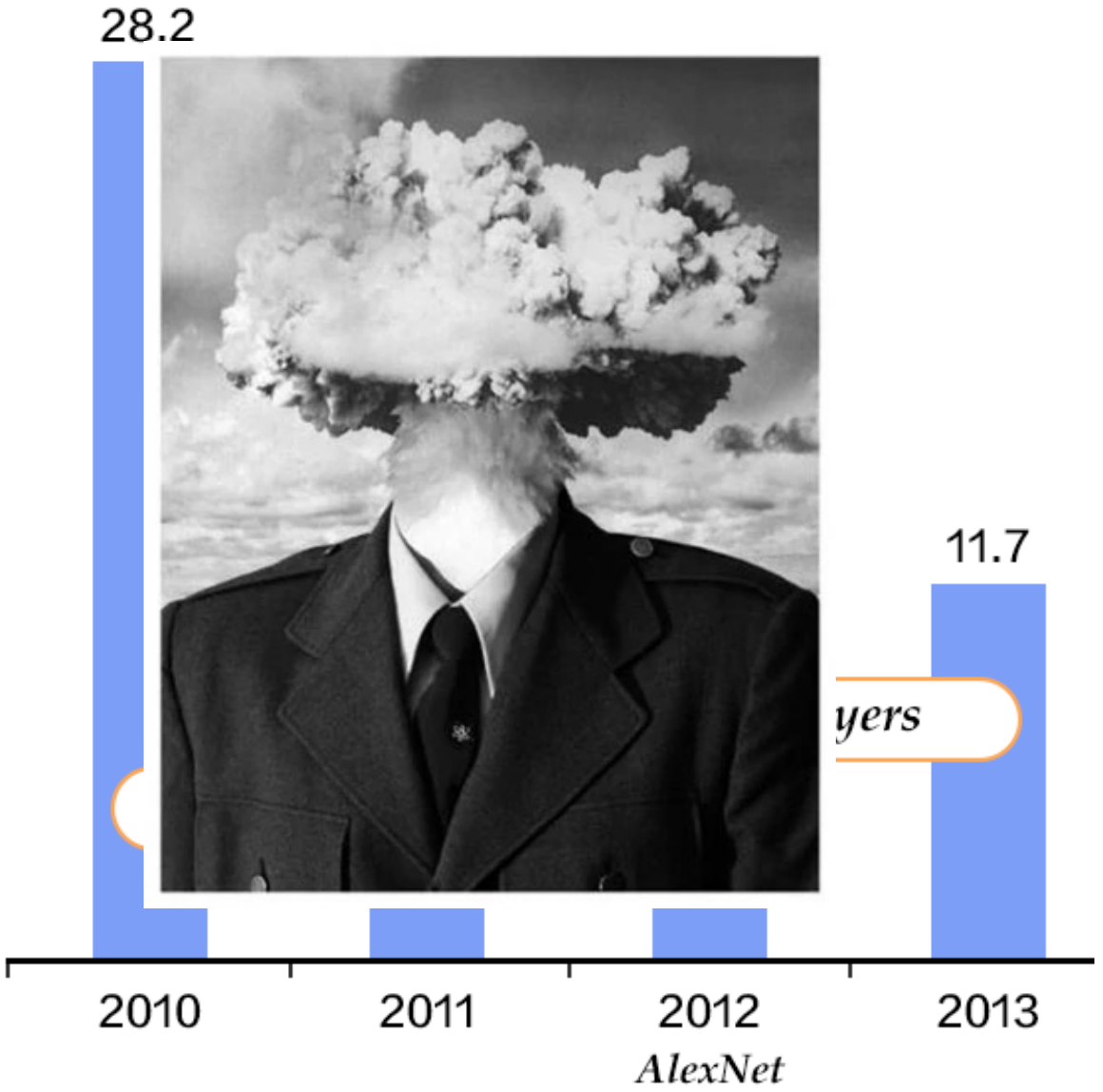
									Total
Parameters :	[(11 x 11 x 3) + 1] x 96 = 35 K	[5 x 5 x 48] x 256 = 307 K	[3 x 3 x 256] x 384 = 884 K	663 K	442 K	37 M	16 M	4 M	60 M
Neurons :	253,440	27 x 27 x 256 = 186,624	13 x 13 x 384 = 64,896	13 x 13 x 384 = 64,896	13 x 13 x 256 = 43,264	4096	4096	1000	0.63 M

- Convolutional layers cumulatively contain about 90-95% of computation, only about 5% of the parameters
- Fully-connected layers contain about 95% of parameters.

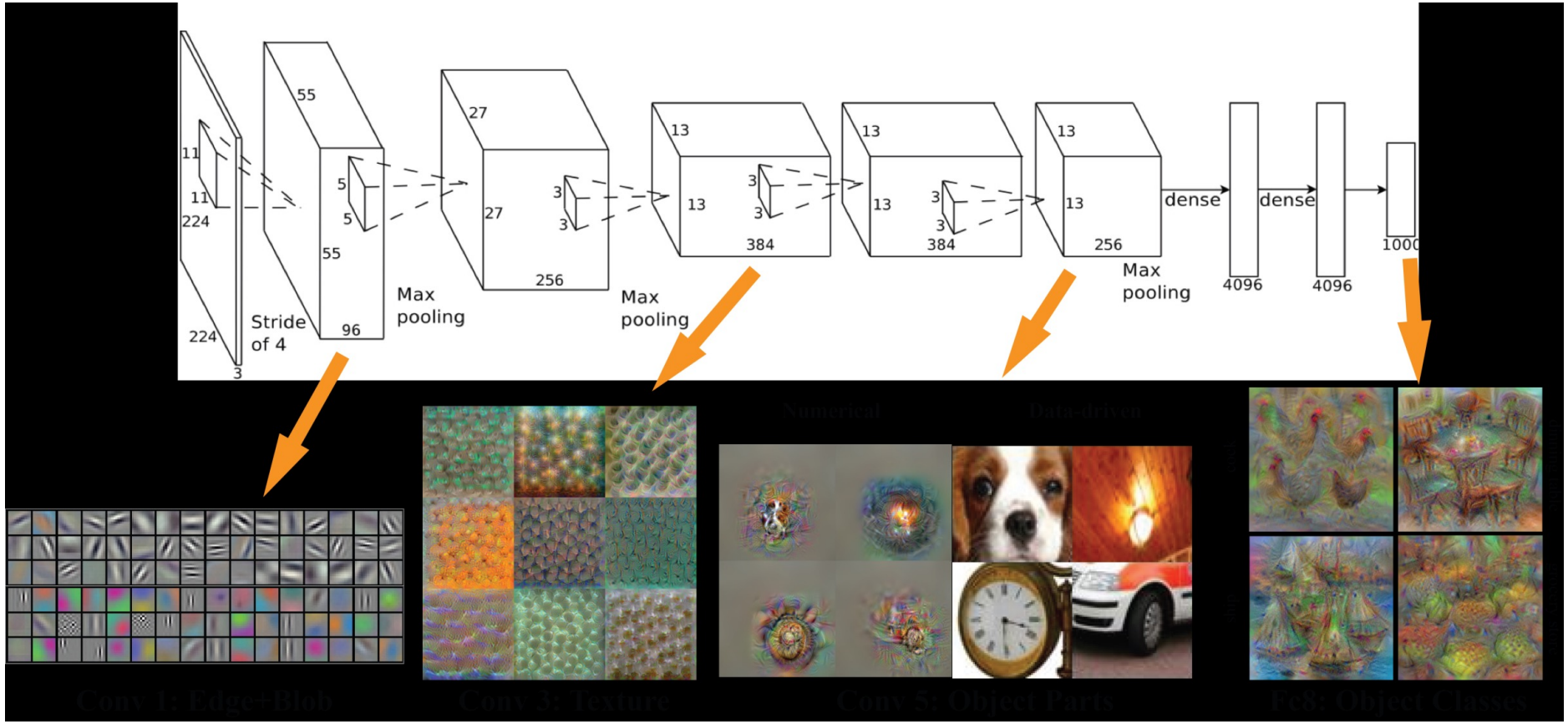
Top-5 misclassification error rate



Top-5 misclassification error rate



Filters learnt by AlexNet



https://www.cc.gatech.edu/~hays/compvision/proj6/deepNetVis.png

MML

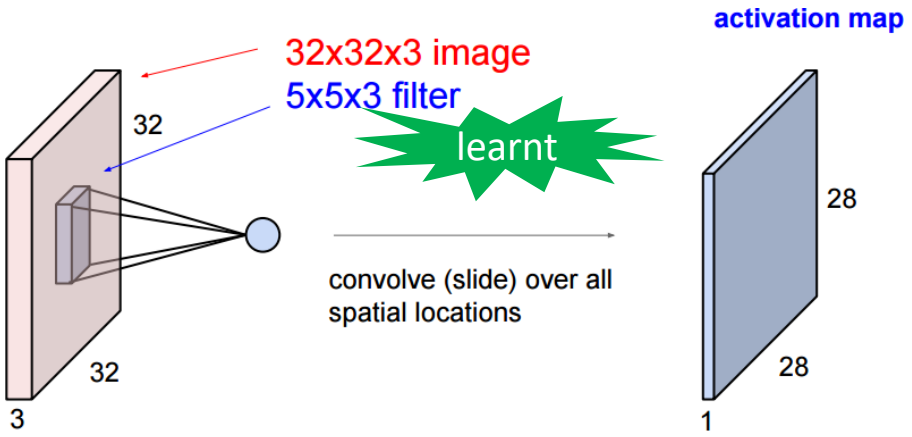
Questions?

MML

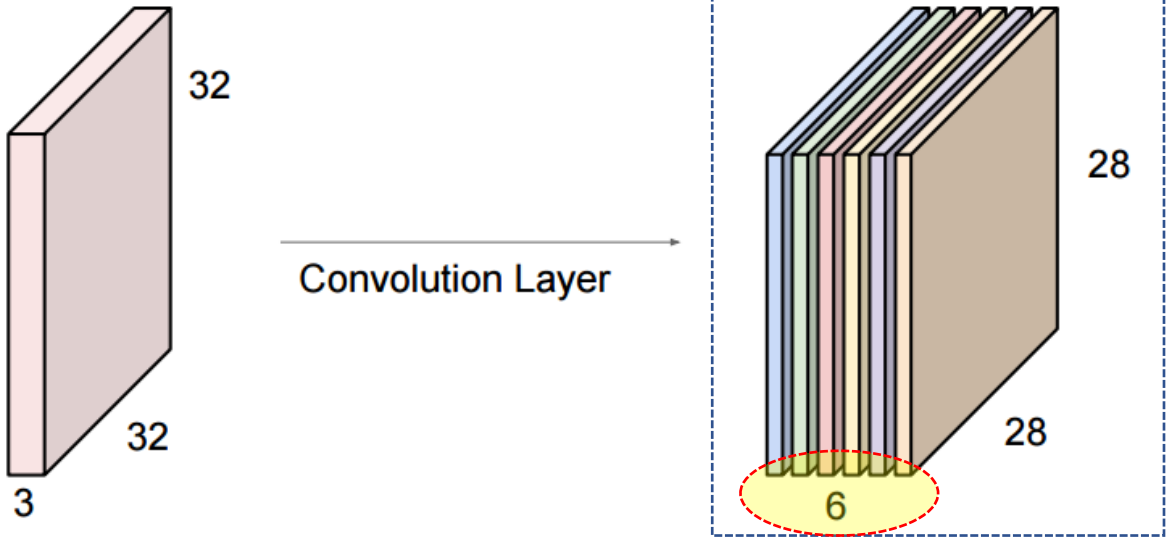
Popular CNN architectures

Improvements and Extensions

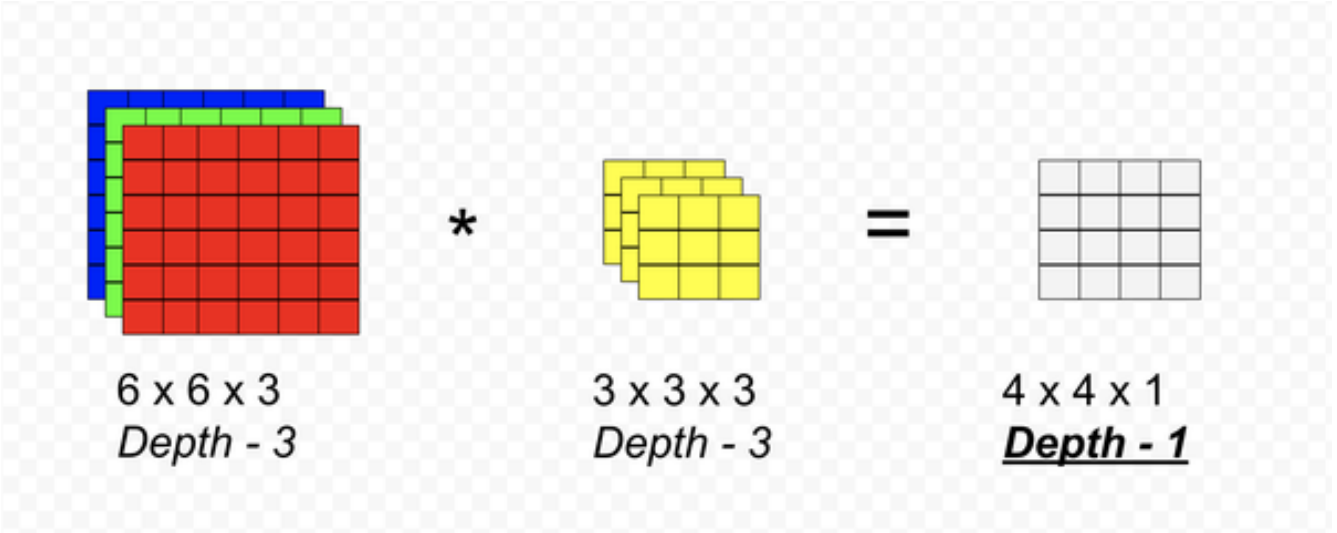
Convolution Layer



Input to next layer

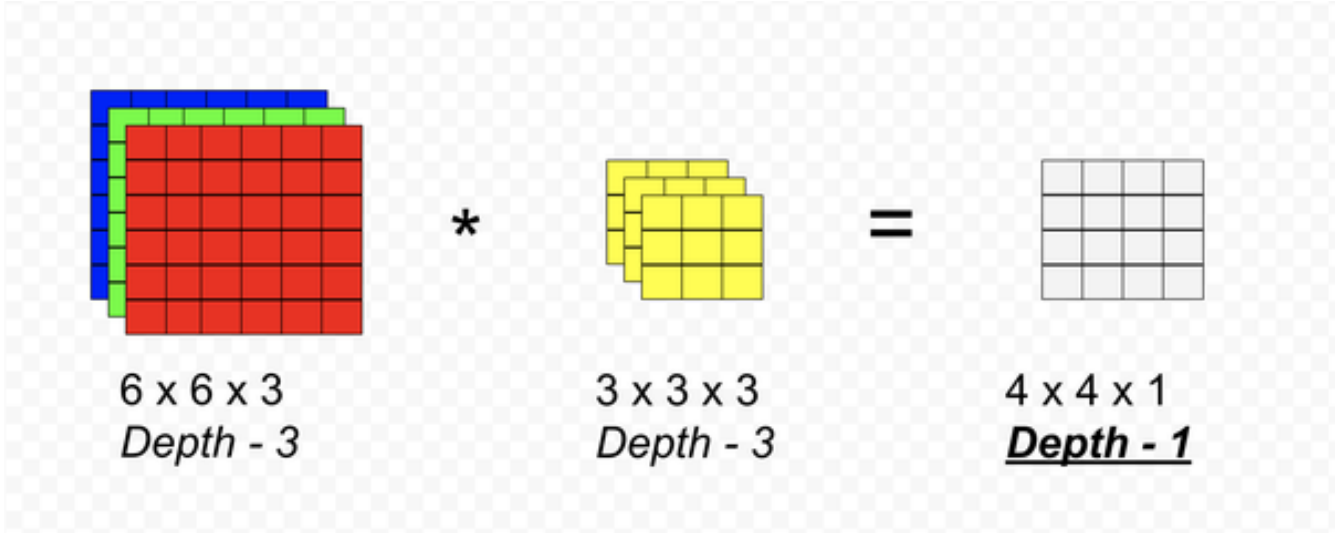


Convolution Layer



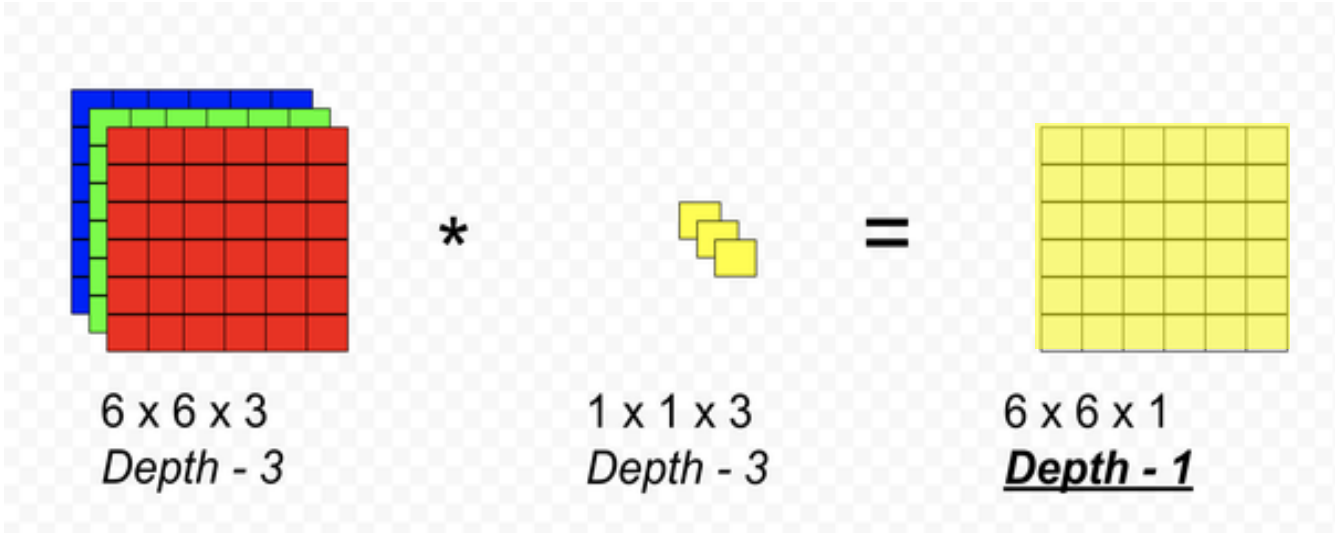
https://www.quora.com/What-is-a-1X1-convolution

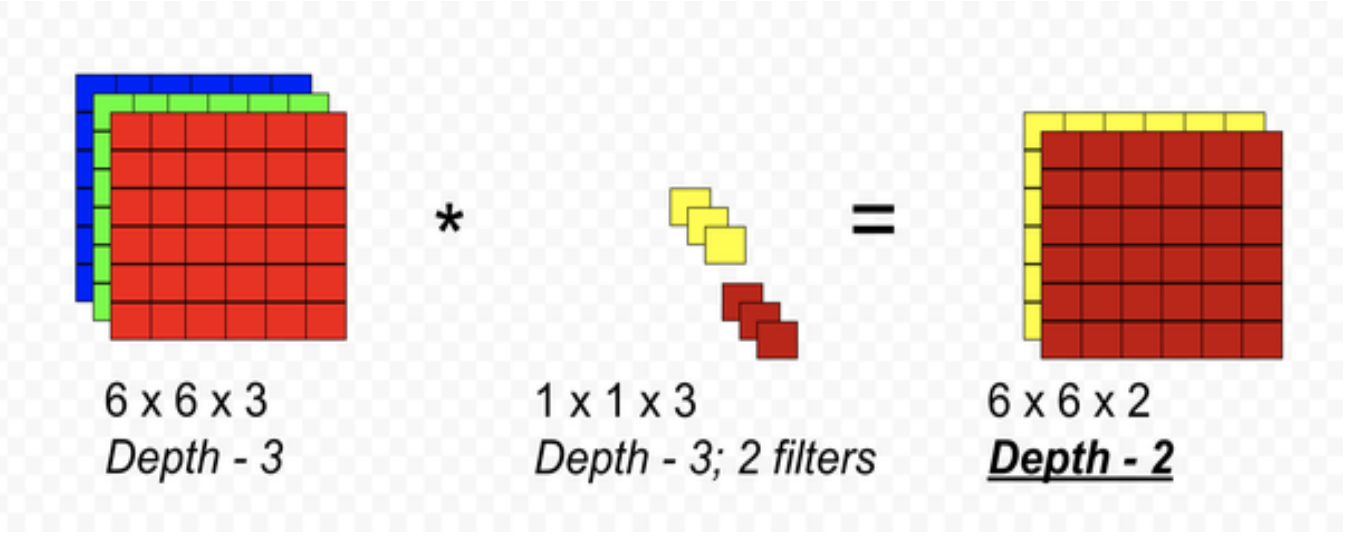
1x1 (Bottleneck) Convolution Layer



https://www.quora.com/What-is-a-1X1-convolution

1x1 (Bottleneck) Convolution Layer

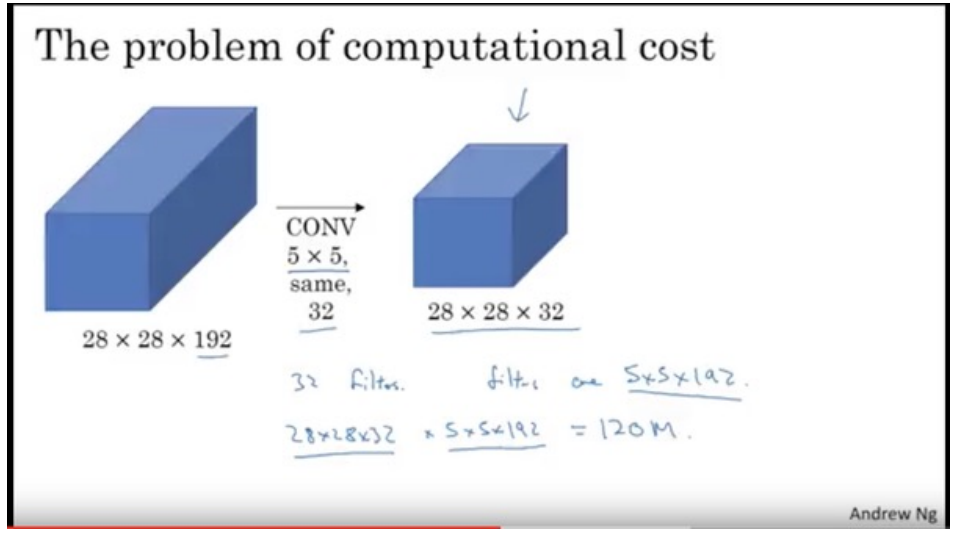




https://www.quora.com/What-is-a-1X1-convolution

1x1 (Bottleneck) Convolution Layer

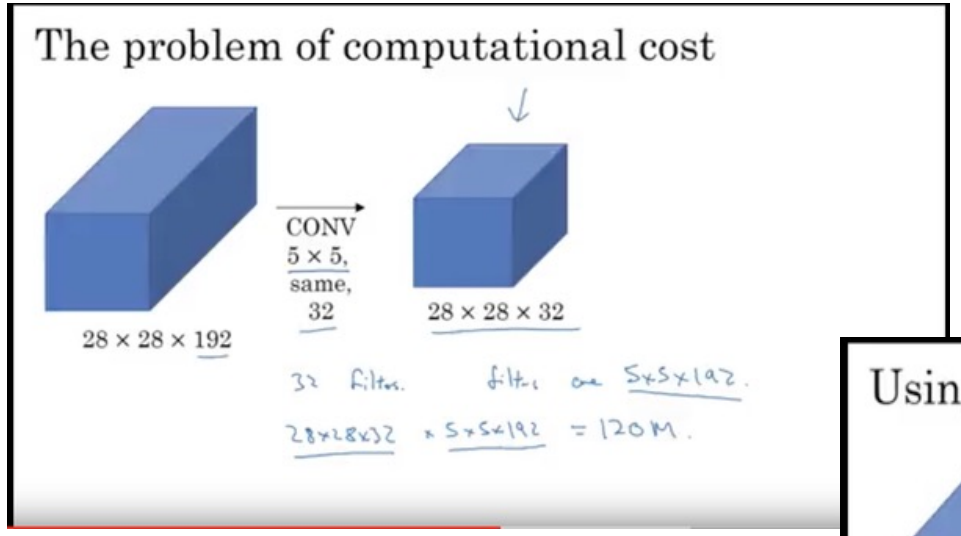
- Enable dimensionality reduction (computational savings)
- Add more non-linearity to representation space

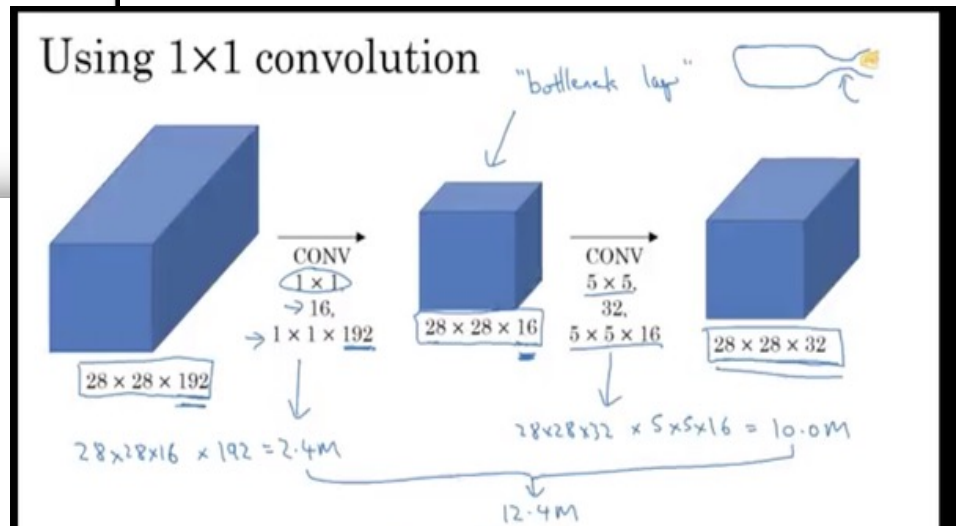


https://www.quora.com/What-is-a-1X1-convolution

1x1 (Bottleneck) Convolution Layer

- Enable dimensionality reduction (computational savings)
- Add more non-linearity to representation space





https://www.quora.com/What-is-a-1X1-convolution

Committing to a single filter size at each layer seems somewhat rigid !

3x3

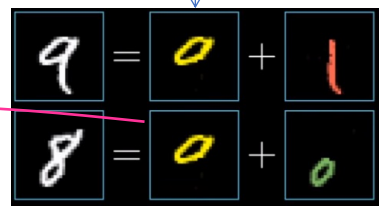
1	2	1
2	4	2
1	2	1

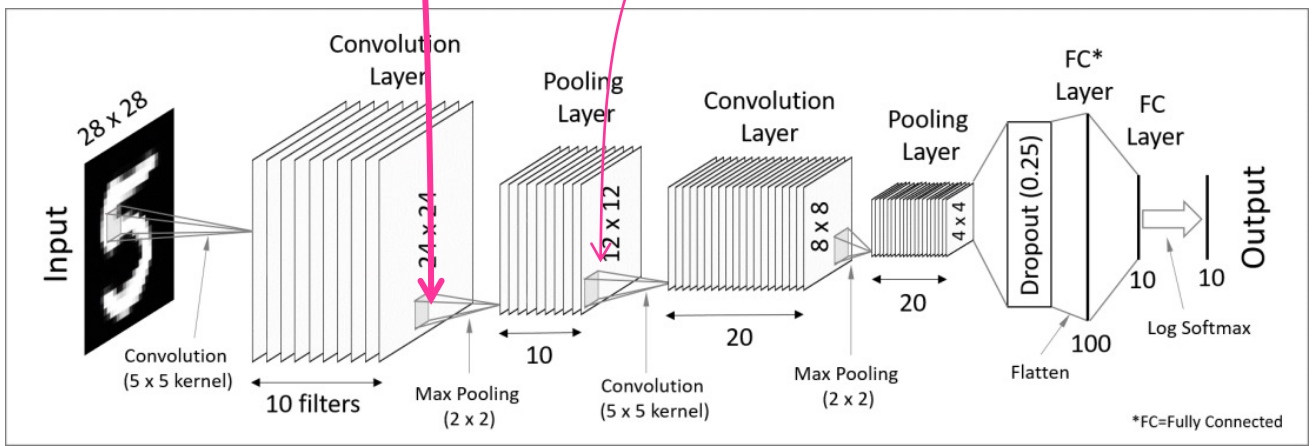
5x5

2	7	12	7	2
7	31	52	31	7
12	52	127	52	12
7	31	52	31	7
2	7	12	7	2

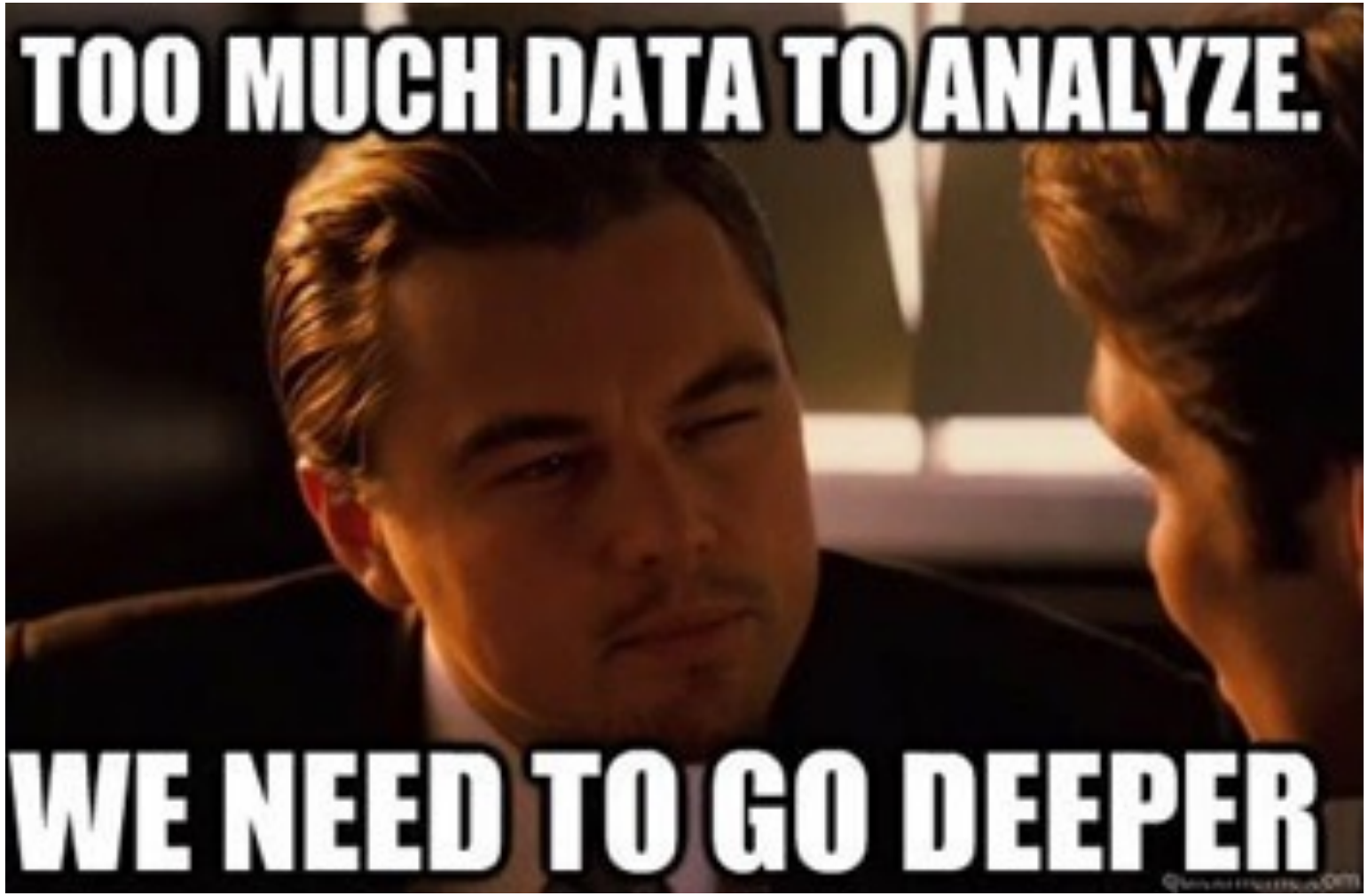
7x7

1	1	2	2	2	1	1
1	3	4	5	4	3	1
2	4	7	8	7	4	2
2	5	8	10	8	5	2
2	4	7	8	7	4	2
1	3	4	5	4	3	1
1	1	2	2	2	1	1

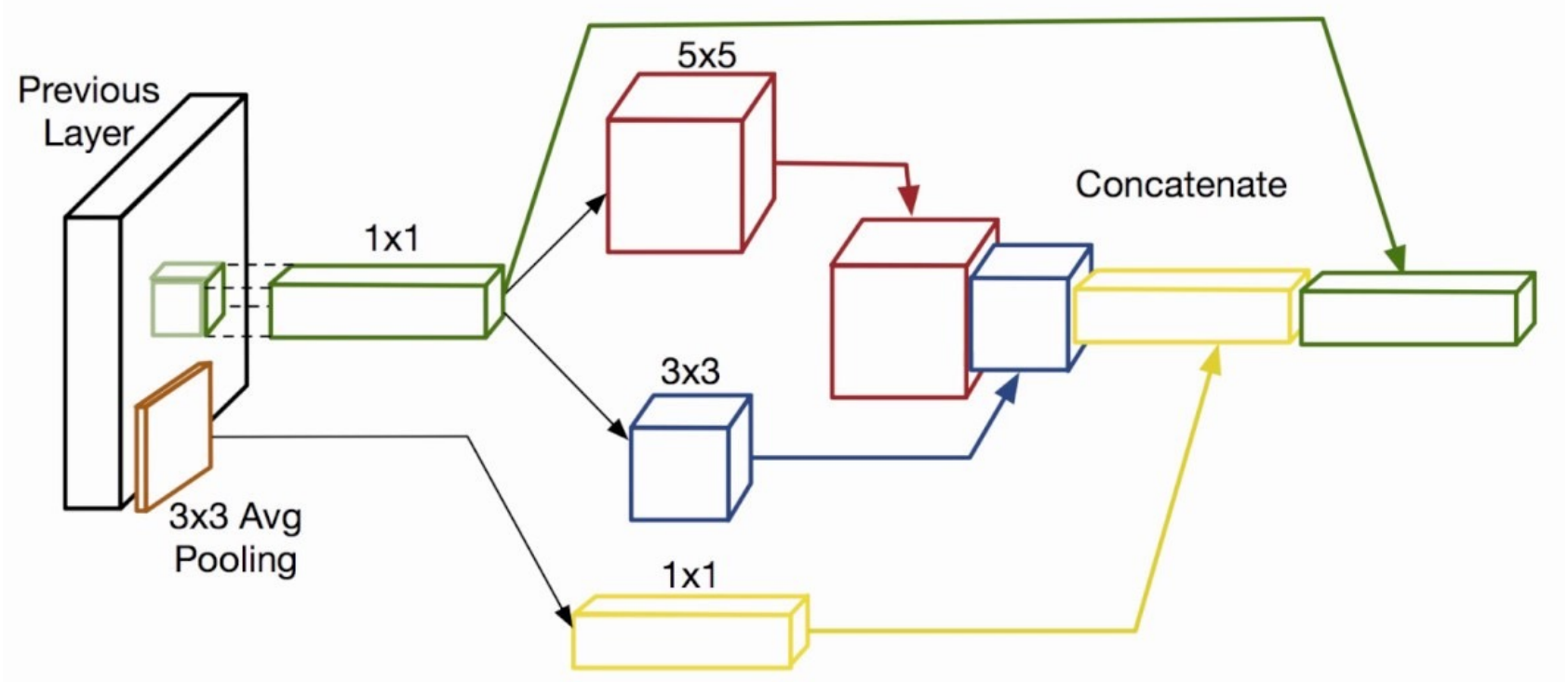




Why not have multiple levels (scales) of filtering ?



Why not have multiple levels (scales) of filtering ?

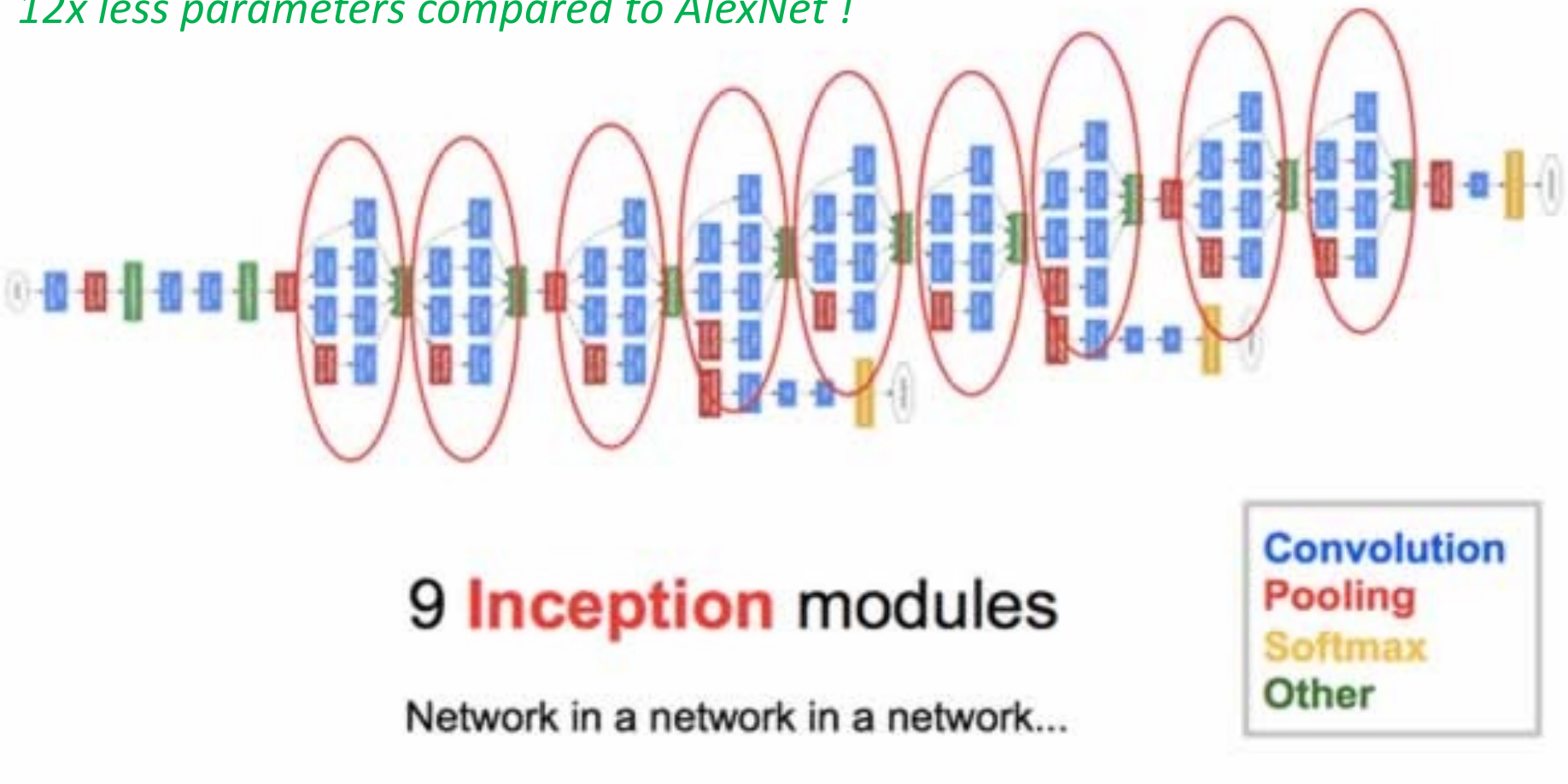


Inception Module

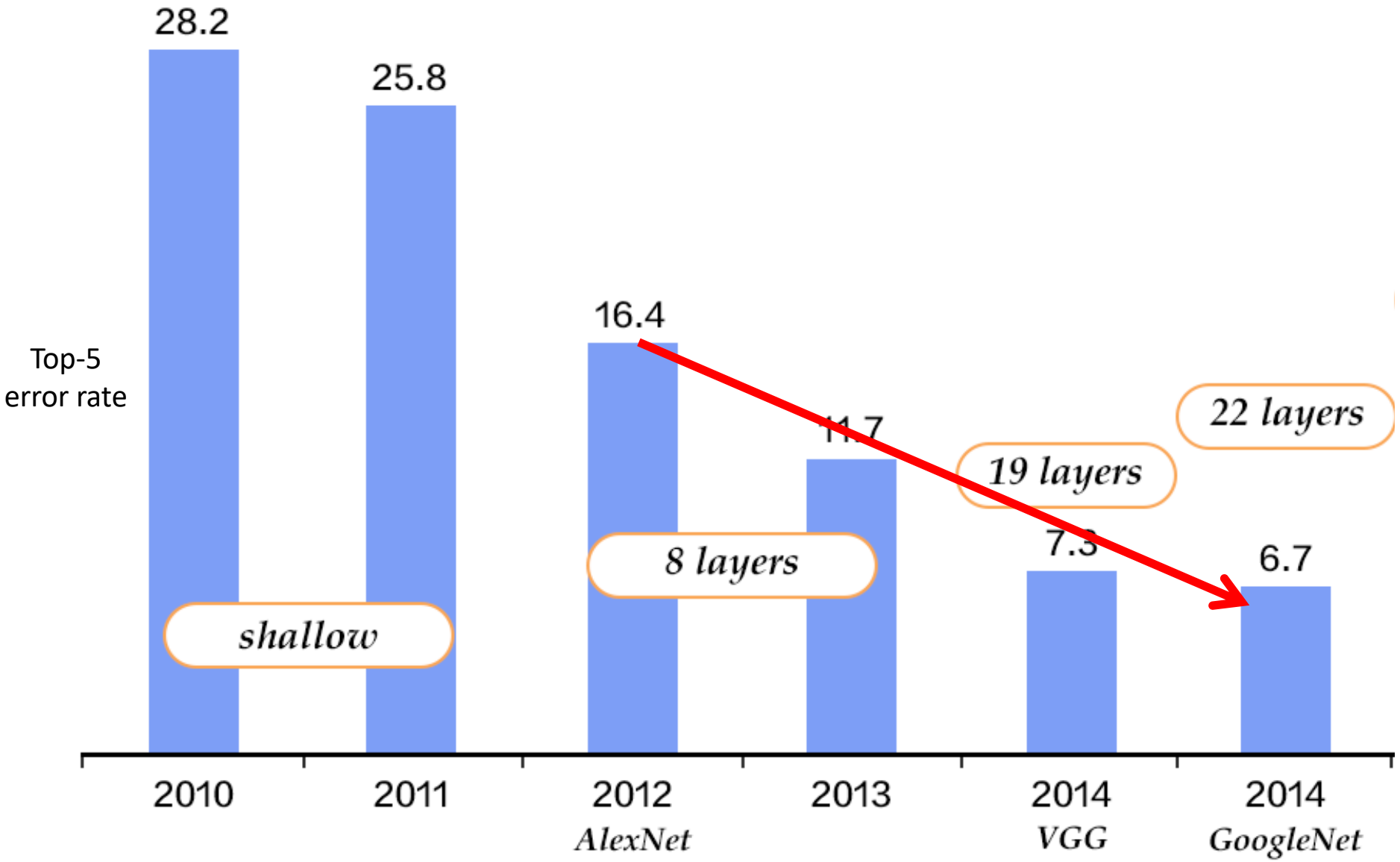
https://i.ytimg.com/vi/_XF7N6rp9Jw/maxresdefault.jpg

GoogLeNet

12x less parameters compared to AlexNet !



https://www.youtube.com/watch?v=LUaRa65J_Ms


Top-5
error rate
28.2
25.8
16.4
11.7
7.3
6.7
shallow
8 layers
19 layers
22 layers
2010
2011
2012
AlexNet
2013
2014
VGG
2014
GoogleNet

How deep can the network get ?

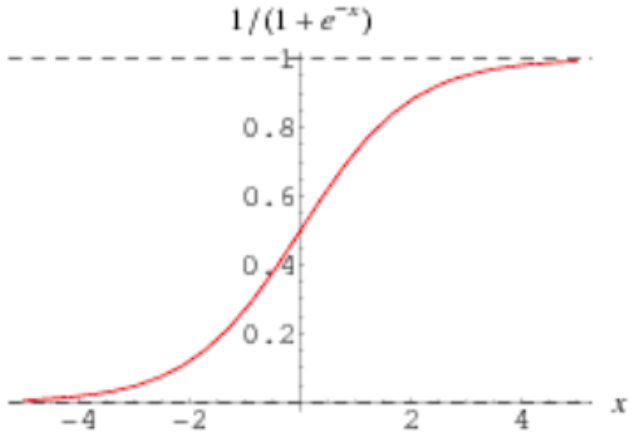
- 50 layers ? 100 ? 1000 ?

How deep can the network get ?

- 50 layers ? 100 ? 1000 ?
- Issue: Vanishing gradients

$$\sigma(x) = \frac{1}{(1 + e^{-x})}$$

Sigmoid



Plot of Sigmoid function

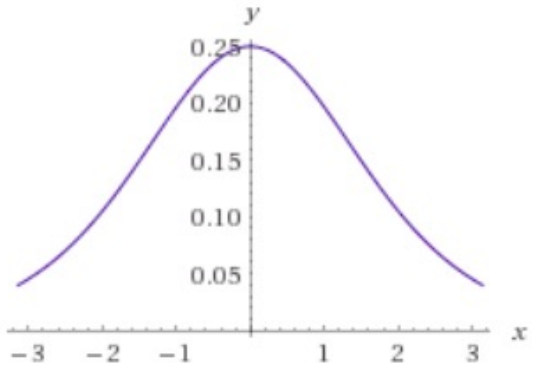
https://medium.com/@deepnarainsingh/vanishing-gradients-while-training-neural-network-42d6d5d83cdf

Vanishing gradients

$$\sigma'(x) = \frac{1}{(1+e^{-x})}(1 - \frac{1}{(1+e^{-x})})$$

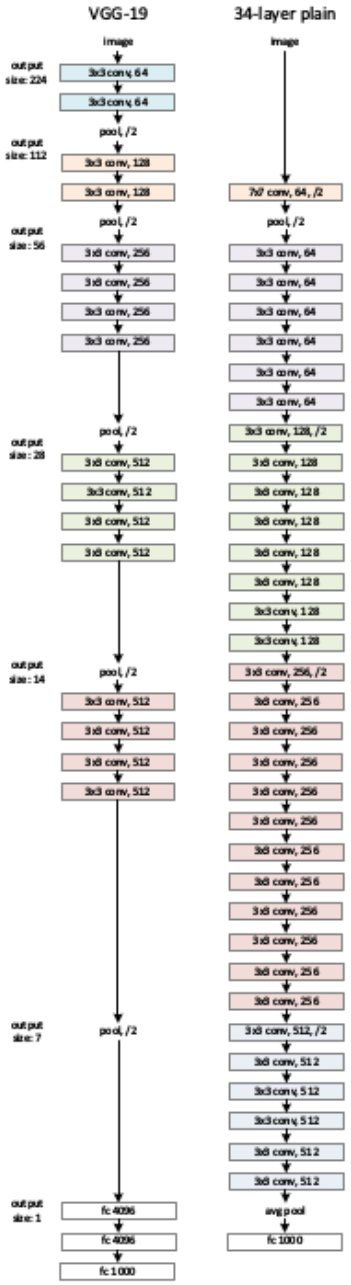
$$\sigma'(x) = \sigma(1-\sigma)$$

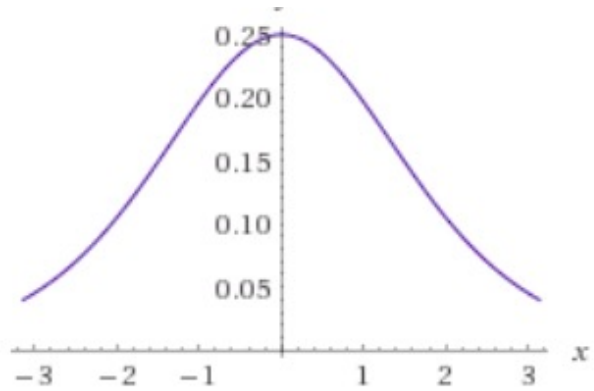
Derivative of sigmoid can be represented as



(*x* from −3.1 to 3.1)

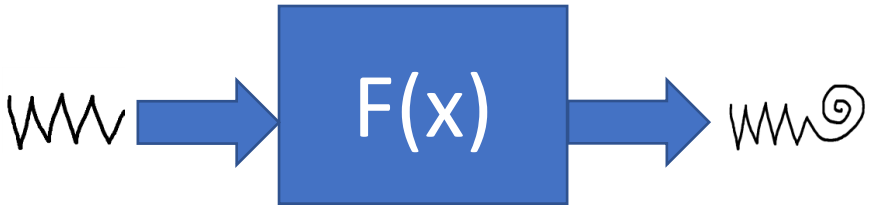
https://medium.com/@deepnarainsingh/vanishing-gradients-while-training-neural-network-42d6d5d83cdf

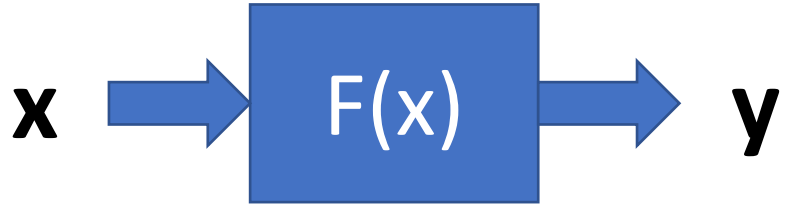




https://medium.com/@14prakash/understanding-and-implementing-architectures-of-resnet-and-resnext-for-state-of-the-art-image-cf51669e1624

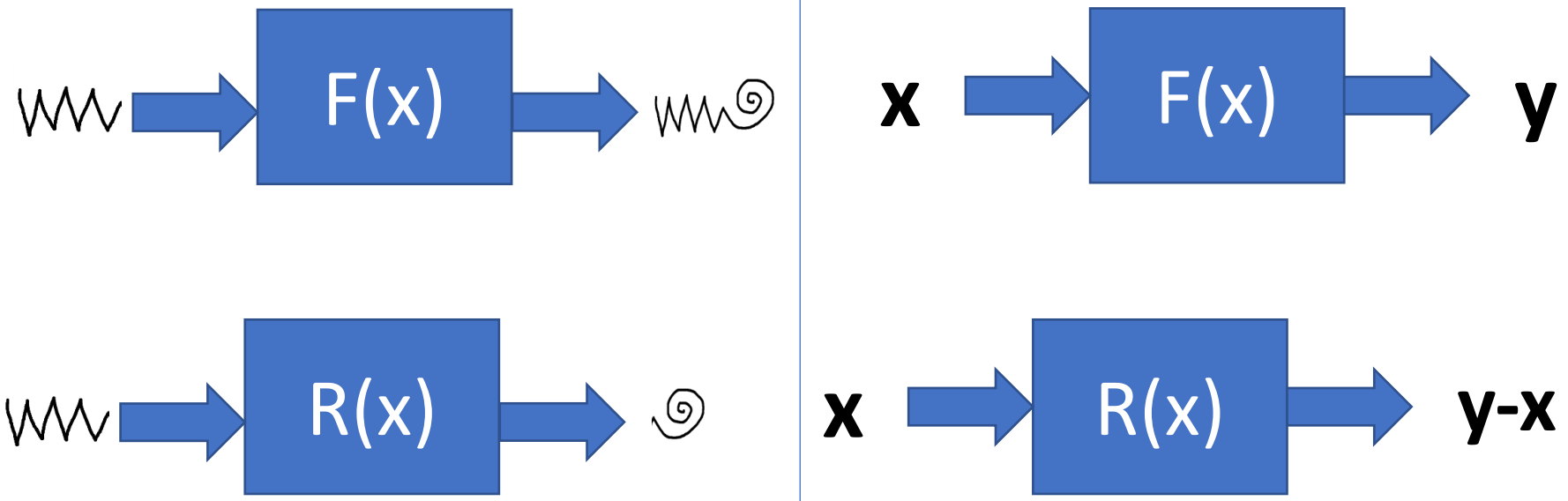
Residual Networks – A crayon analogy





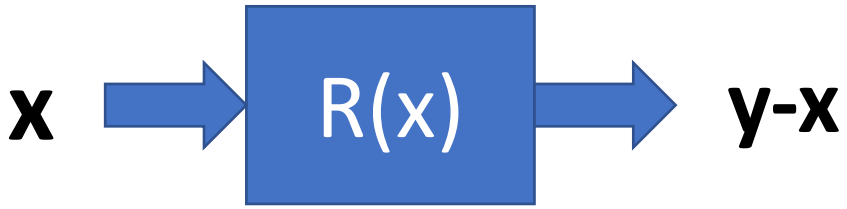
https://mc.ai/explaining-resnet-with-crayons/

Residual Networks – A crayon analogy



https://mc.ai/explaining-resnet-with-crayons/

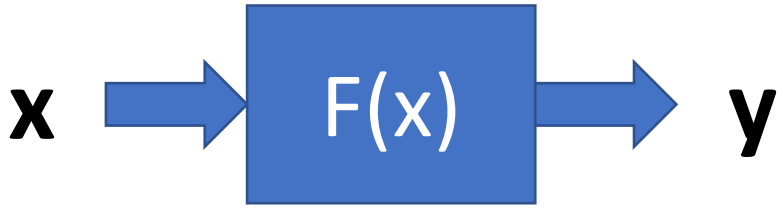
Residual Networks – A crayon analogy

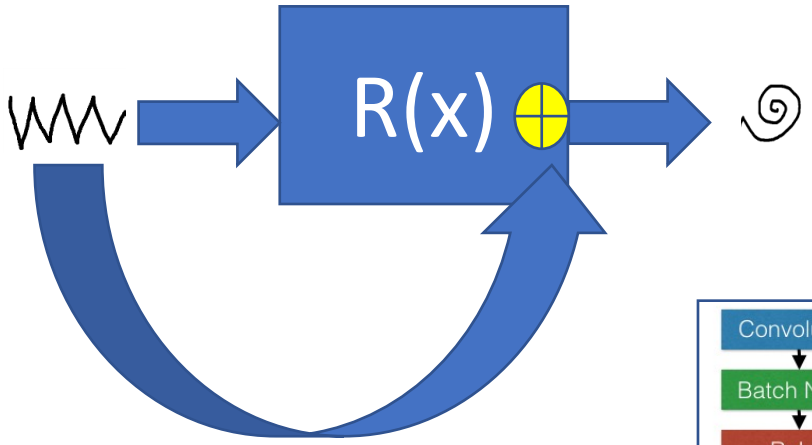


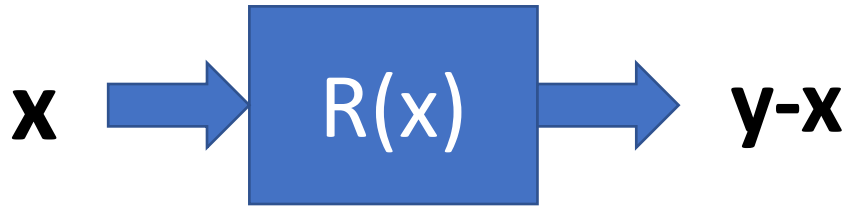
$y = F(x) = x + R(x)$

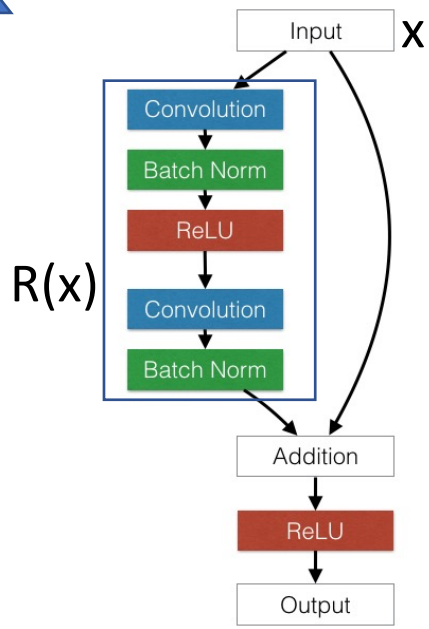
https://mc.ai/explaining-resnet-with-crayons/

Residual Networks – A crayon analogy







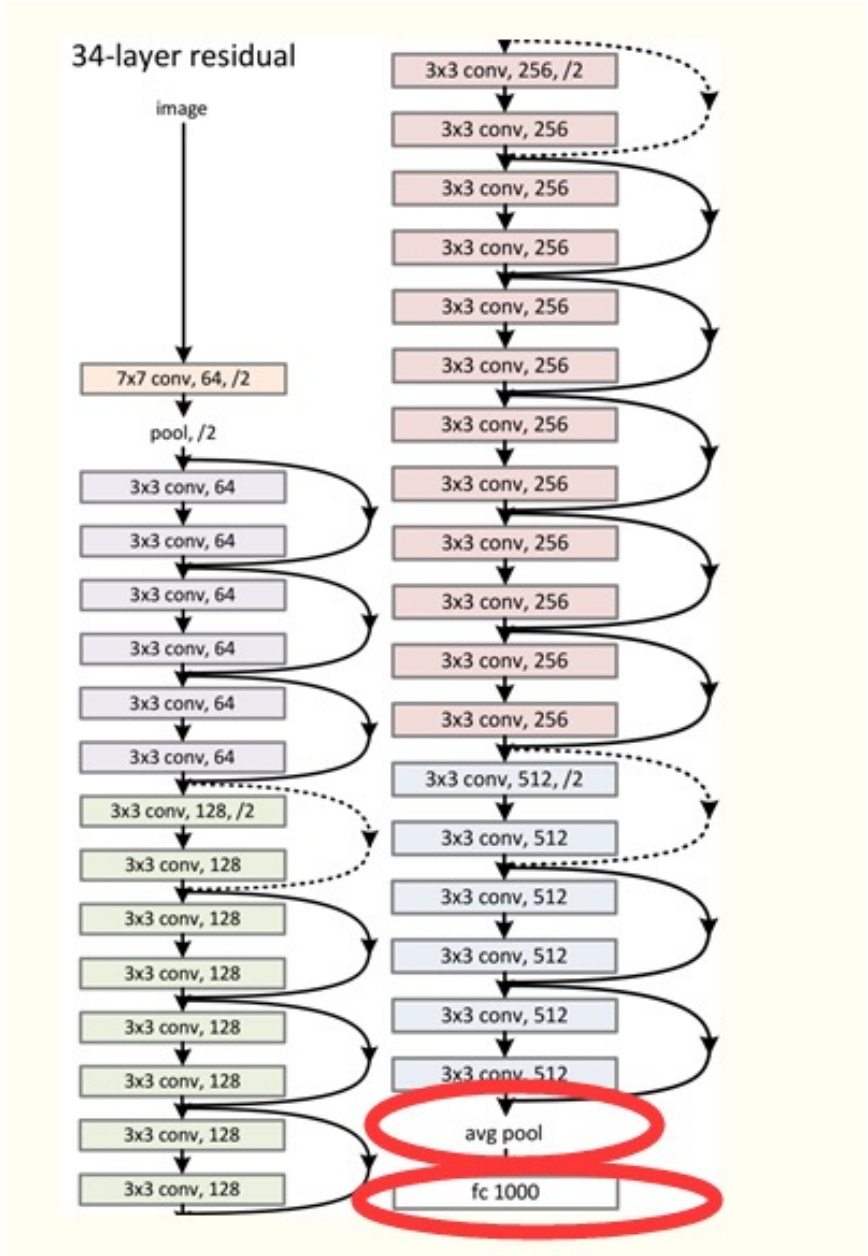


Architecture of a residual block

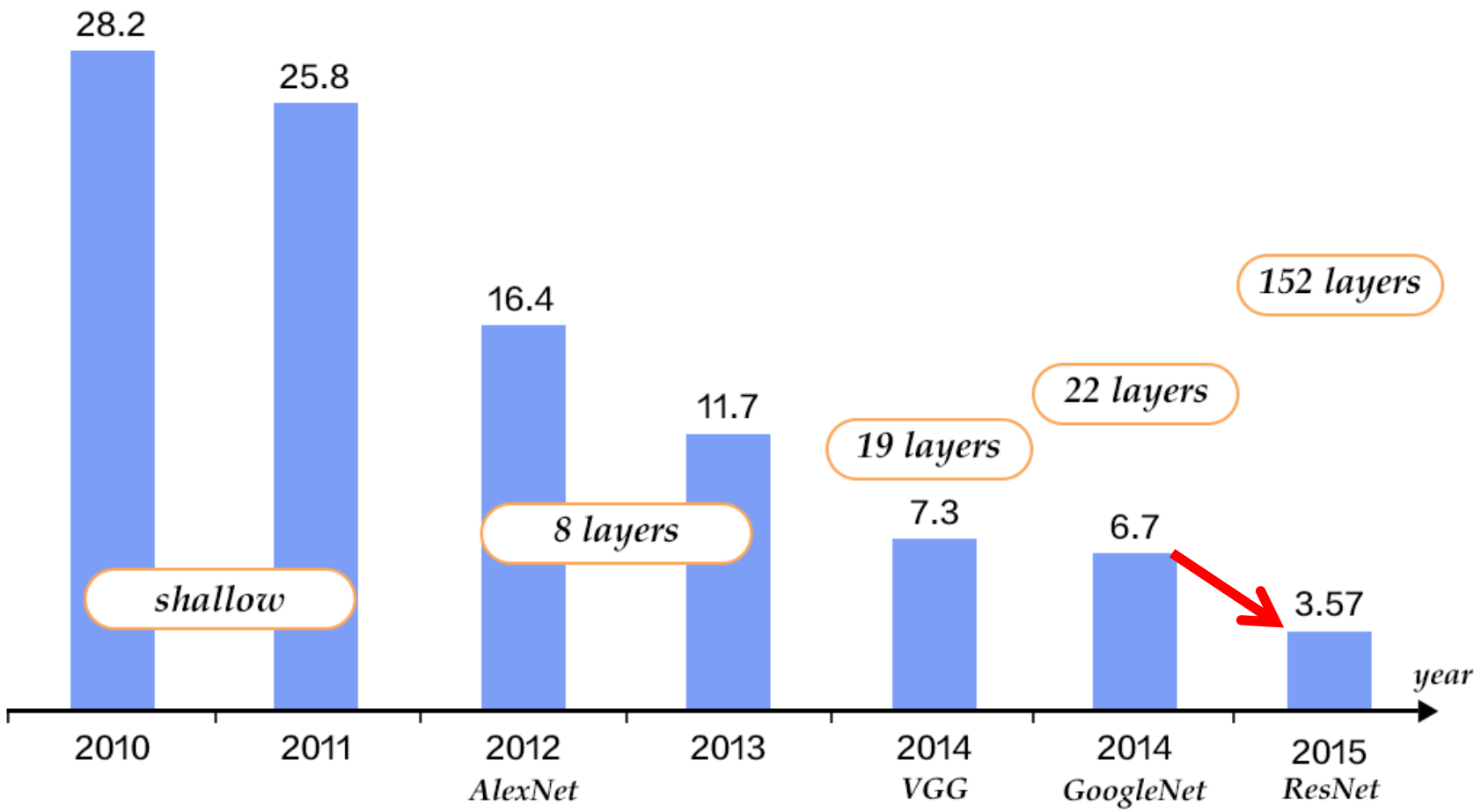
$F(x) = x + R(x)$

https://mc.ai/explaining-resnet-with-crayons/

Residual Networks



Gradients can be supplied to shallower layers directly if needed (via skips)


28.2
25.8
16.4
11.7
7.3
6.7
3.57
shallow
8 layers
19 layers
22 layers
152 layers
year
2010
2011
2012
AlexNet
2013
2014
VGG
2014
GoogleNet
2015
ResNet

MML

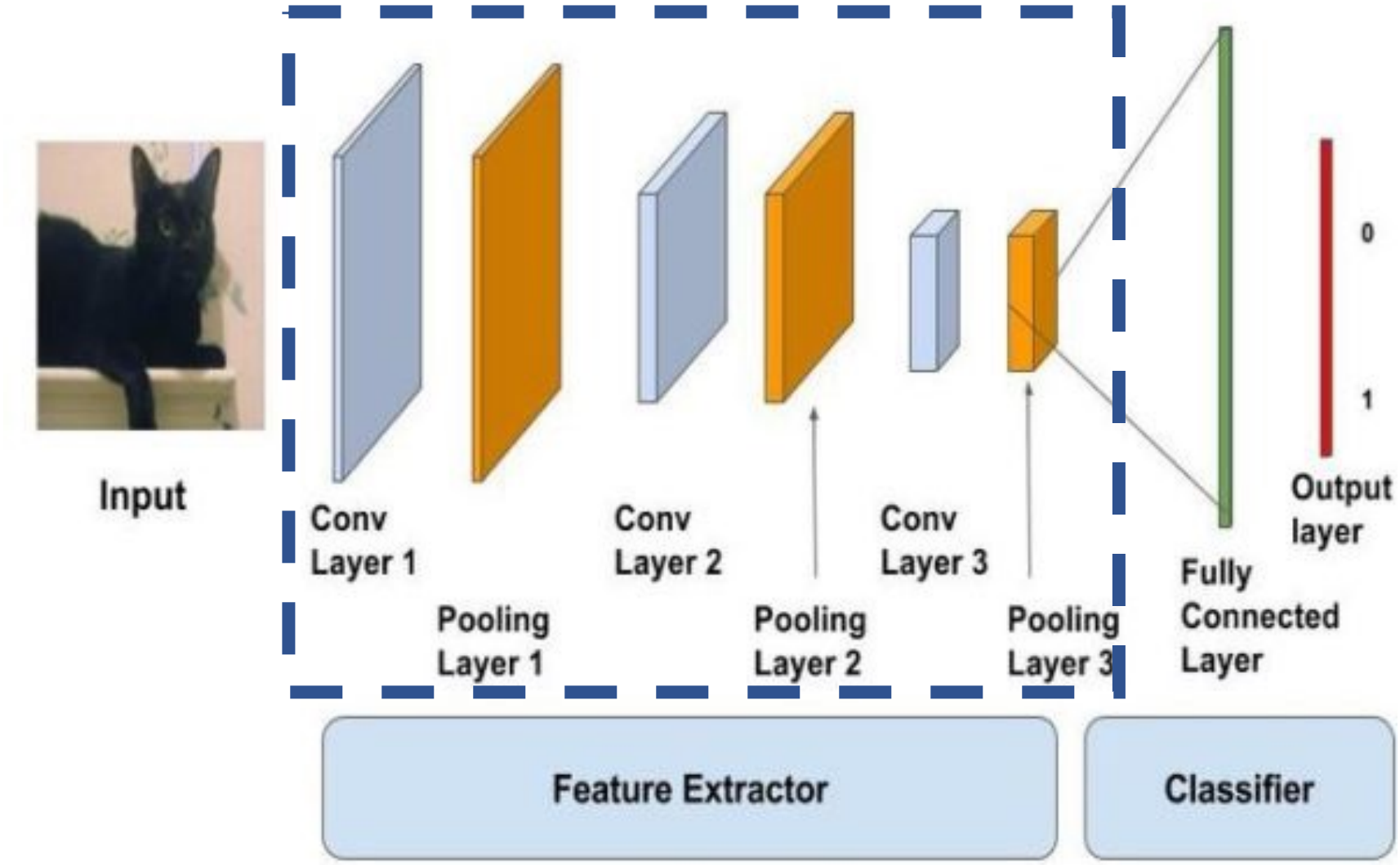
Questions?

MML

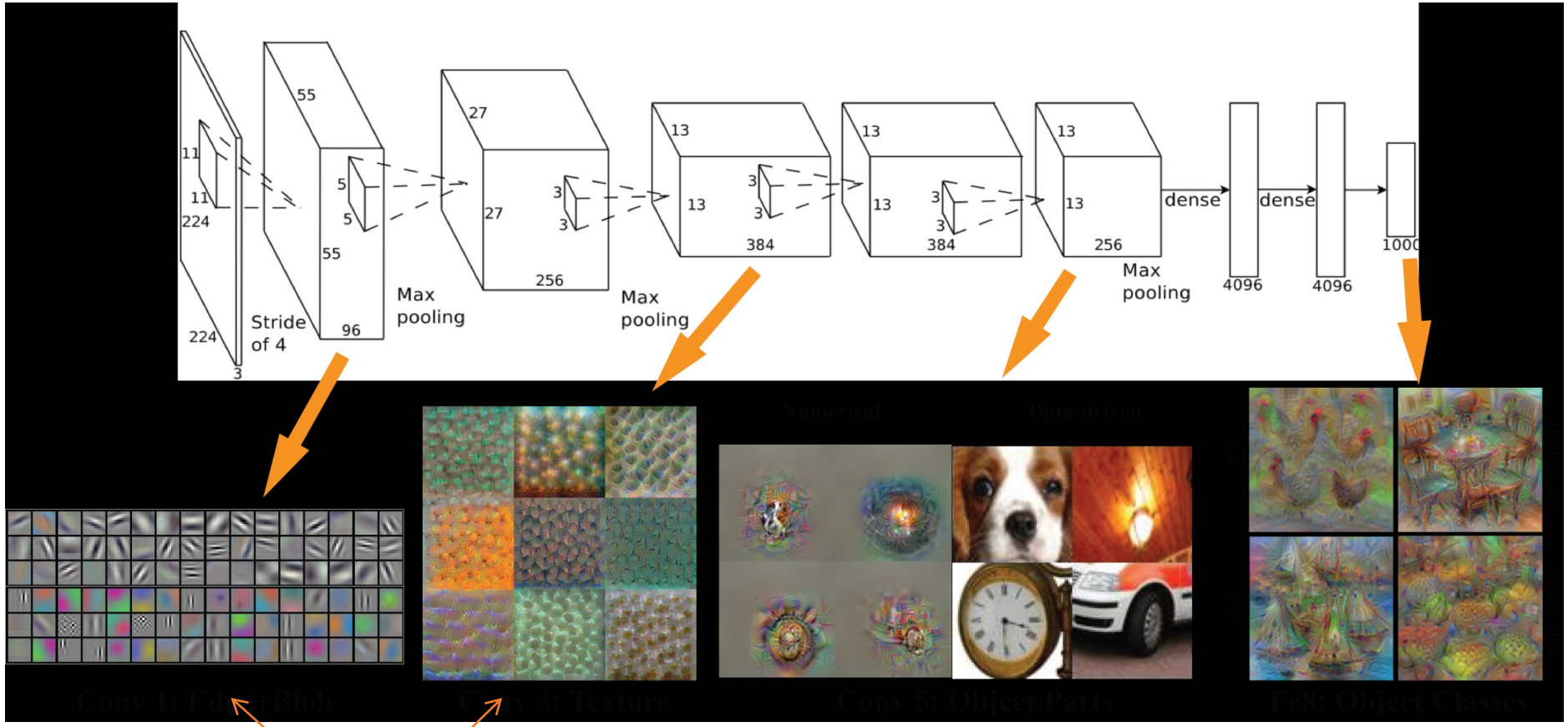
Landscape of CNNs

Applications and Architectures

Transfer Learning



Filters learnt by AlexNet

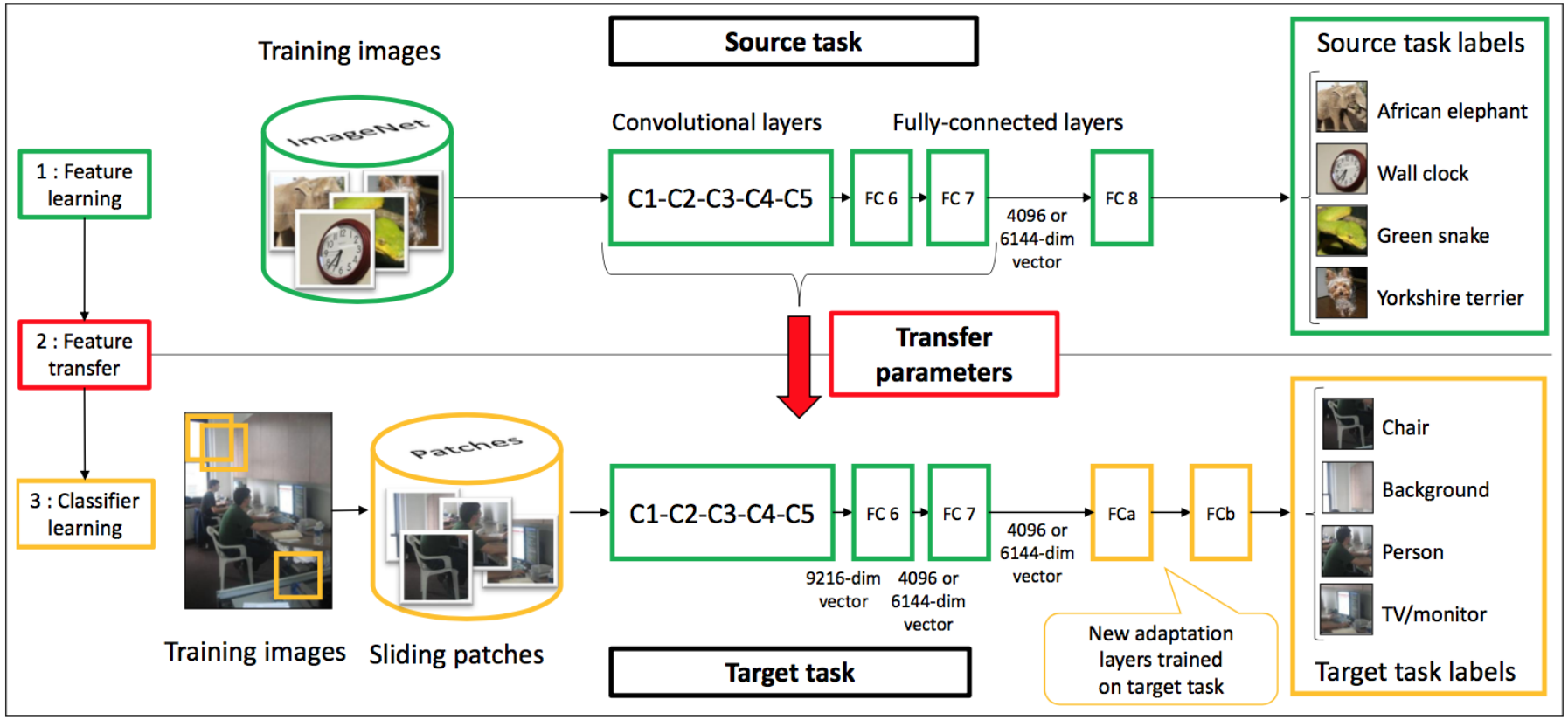


These features seem generic enough. Can we re-use them ?

https://www.cc.gatech.edu/~hays/compvision/proj6/deepNetVis.png

Transfer Learning

Transfer Learning



Examples

Pre-trained features are quite general

ImageNet pre-training

Object classification (PASCAL VOC)

- [Chatifield et al. 2014, Razavian et al. 2014, Zeiler et al. 2014]

Object detection (PASCAL VOC)

- R-CNN [Girshick et al. 2014]

Fine-grained classification (UCSD birds)

- Part-R-CNN [Zhang et al. 2014]

MIT 67 scene classification

- [Razavin et al. 2014]

…

Semantic segmentation

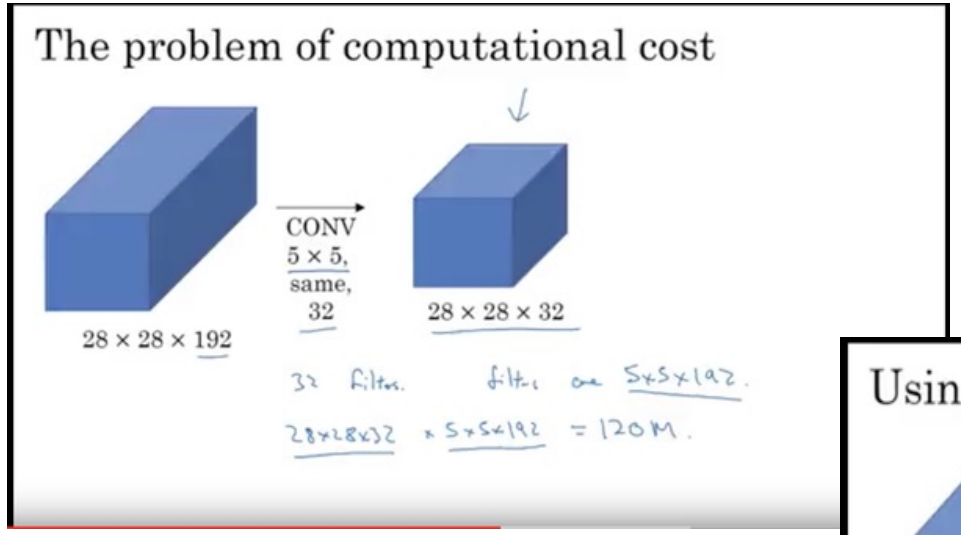
Input

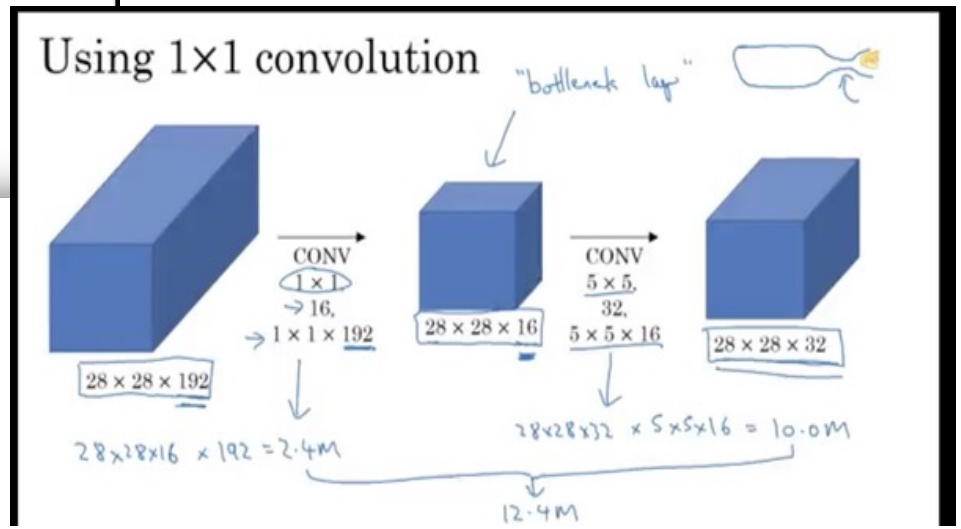
Segmentation [9]

Semantic segmentation

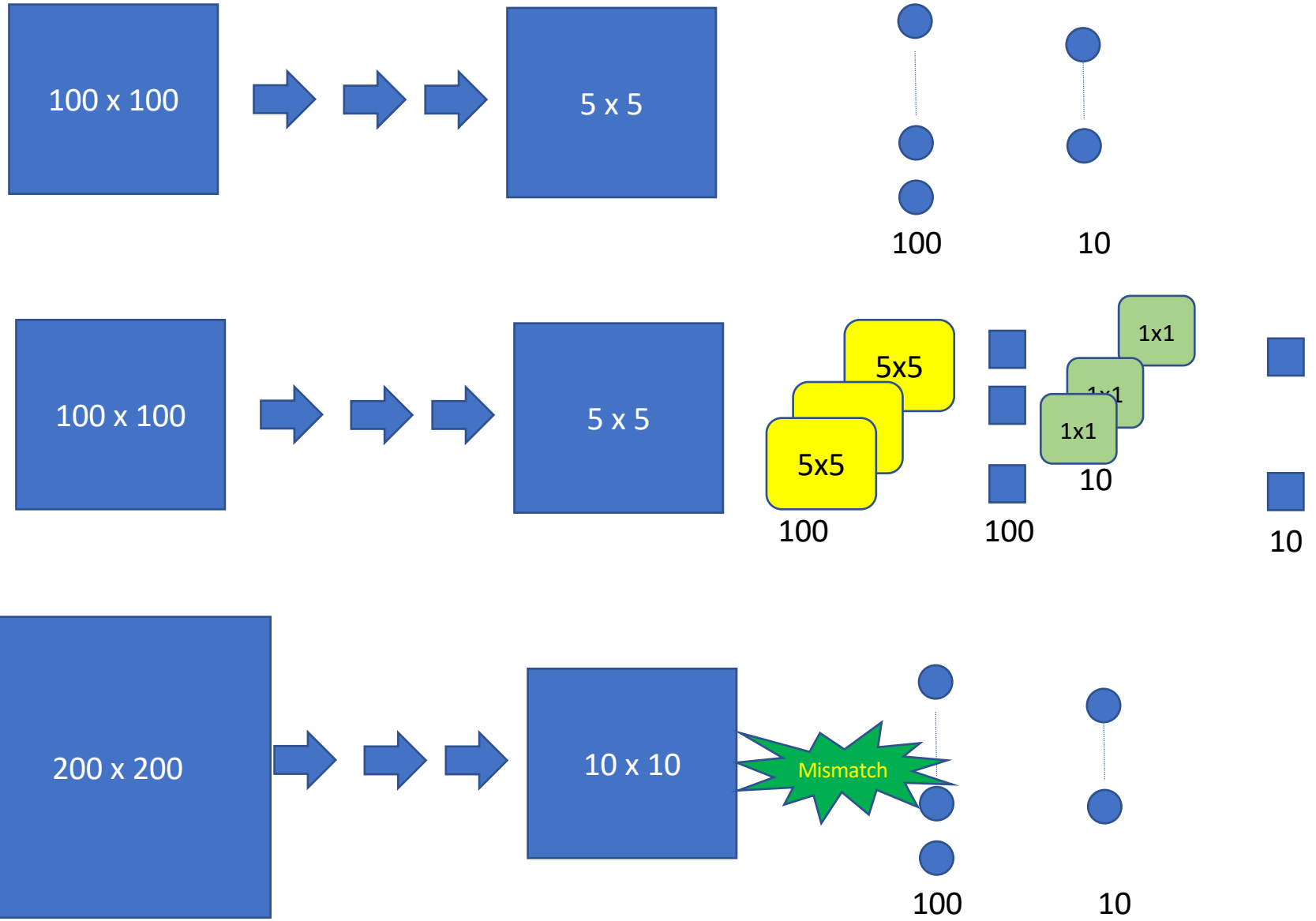
1x1 (Bottleneck) Convolution Layer

- Enable dimensionality reduction (computational savings)
- Add more non-linearity to representation space

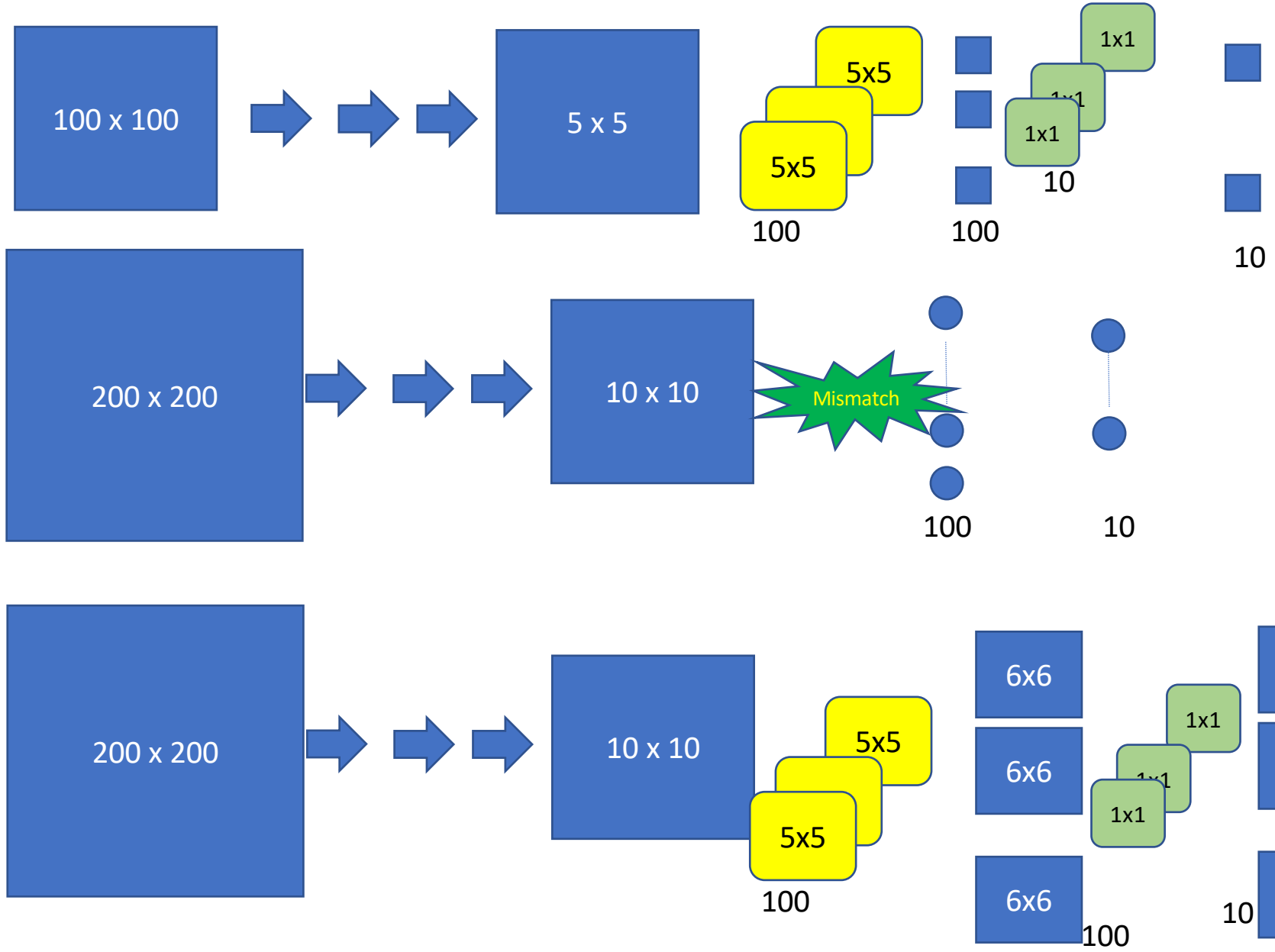




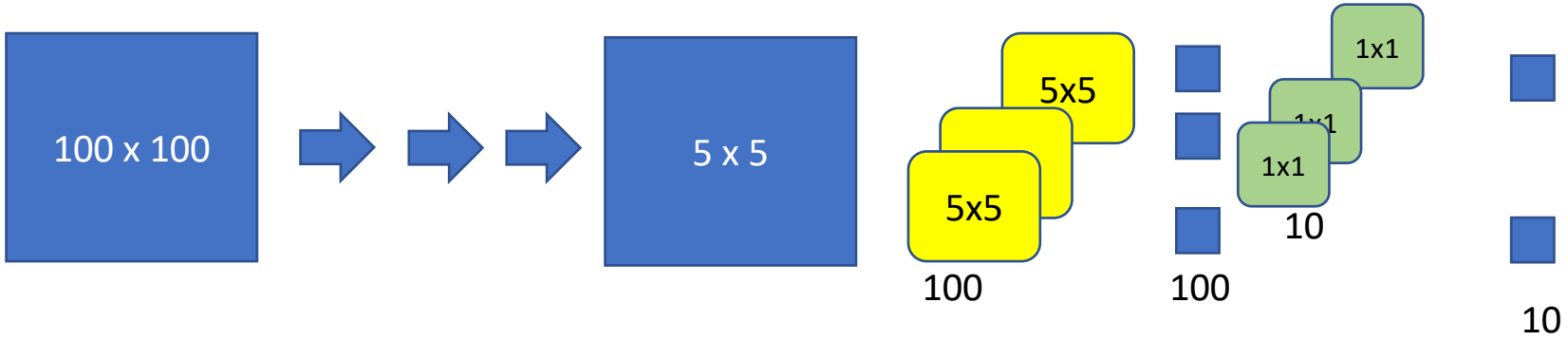
Converting FC Layers to Conv Layers

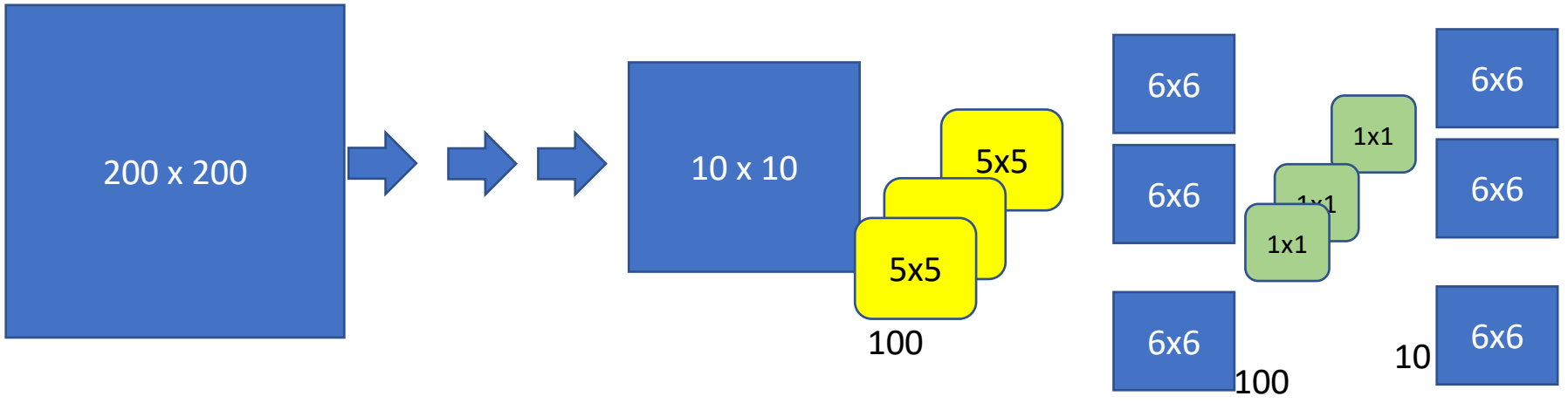


Converting FC Layers to Conv Layers



Converting FC Layers to Conv Layers

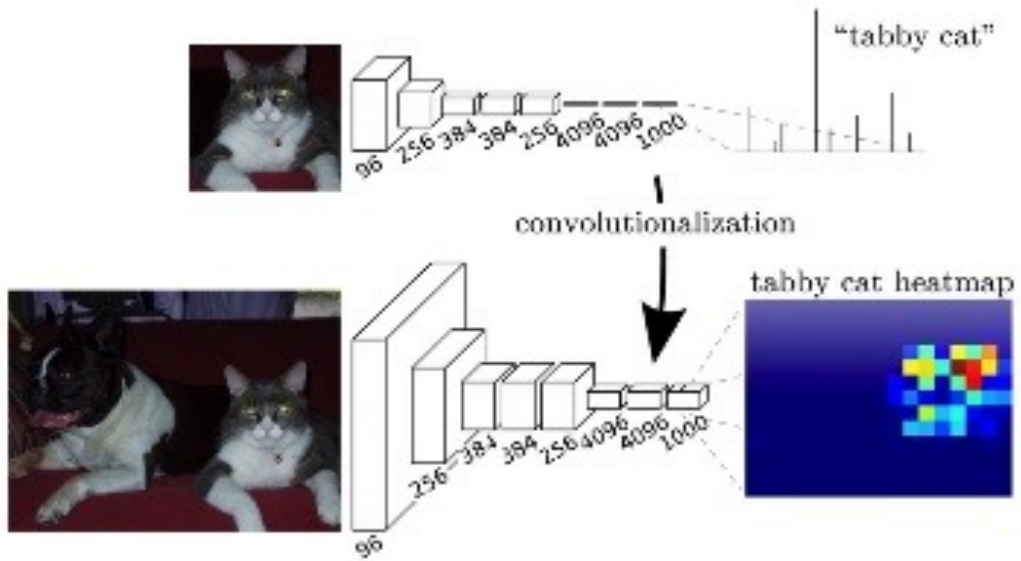




Scores for local image patches

Semantic segmentation

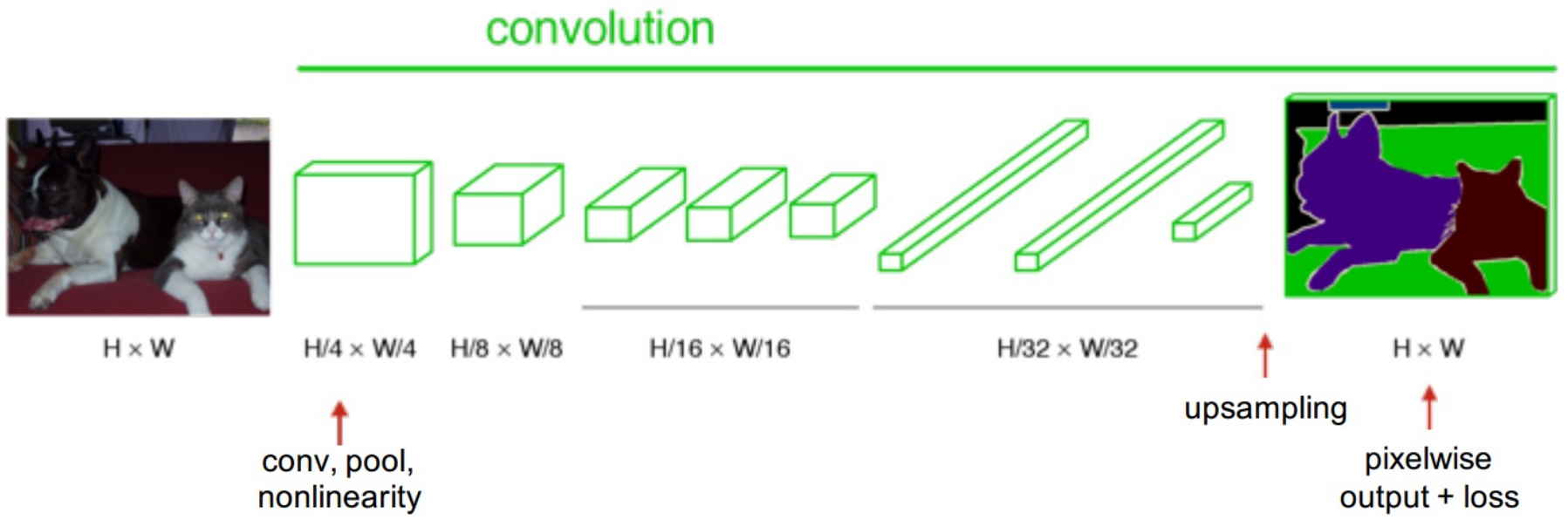
Reinterpret classification as coarse prediction



Shelhamer, Long, Darrell, Fully Convolutional Networks for Semantic Segmentation, 2014-2016

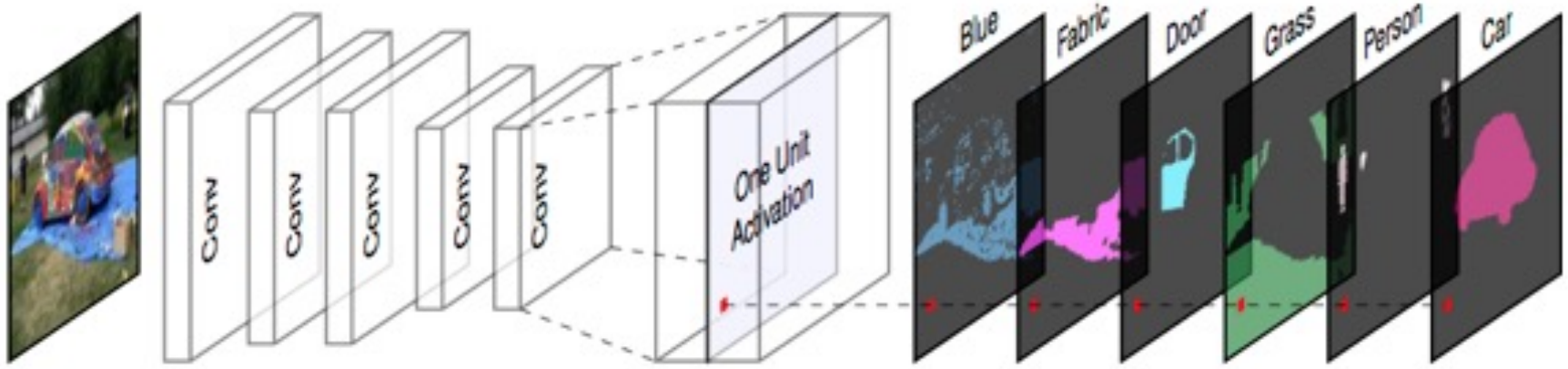
Semantic segmentation

Reinterpret classification as coarse prediction



https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html

Semantic segmentation

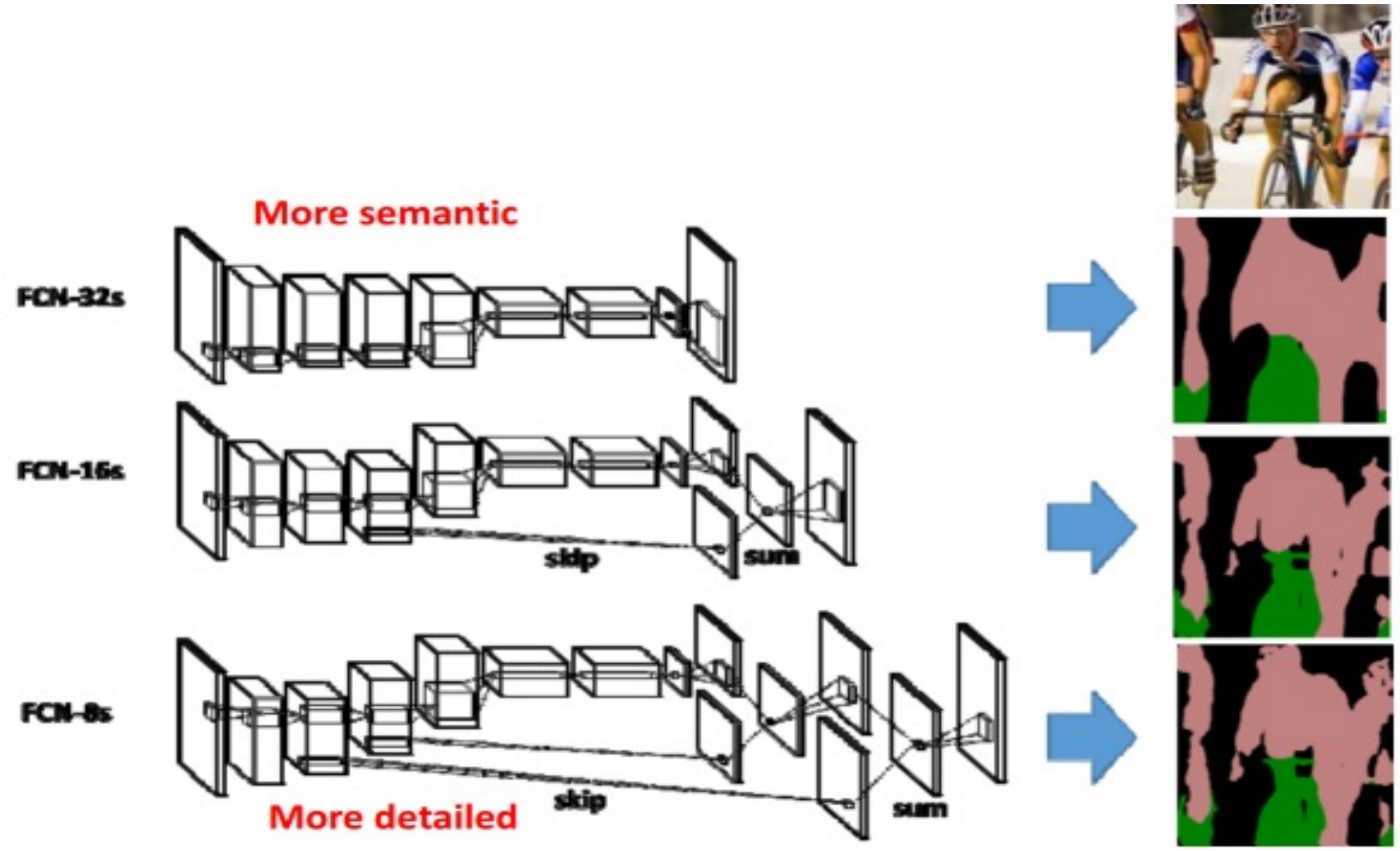


http://netdissect.csail.mit.edu/thumb/slides-thumbnail.png

https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html

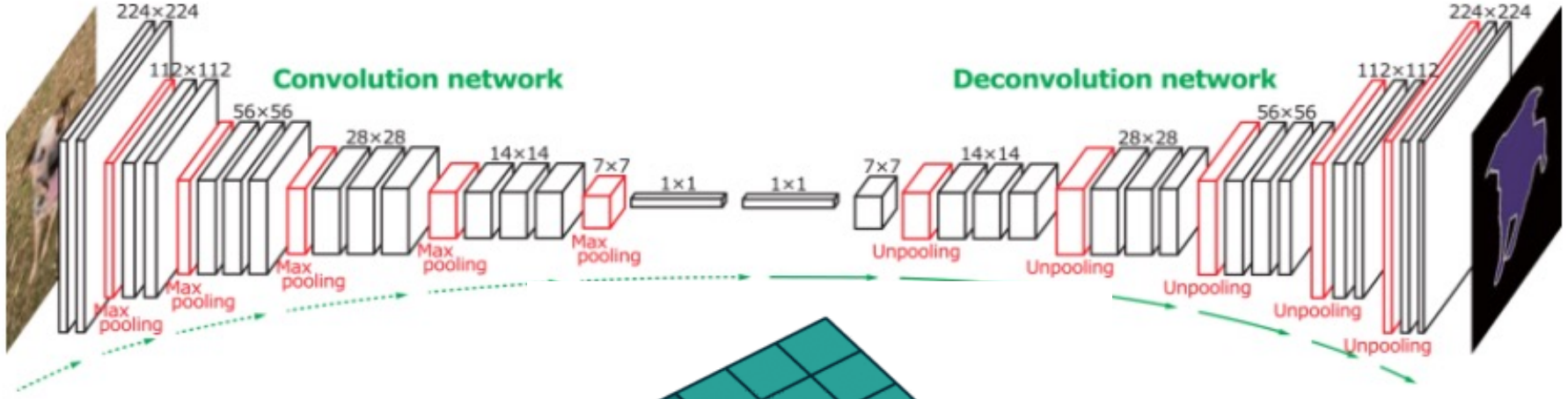
Semantic segmentation

Skip connections

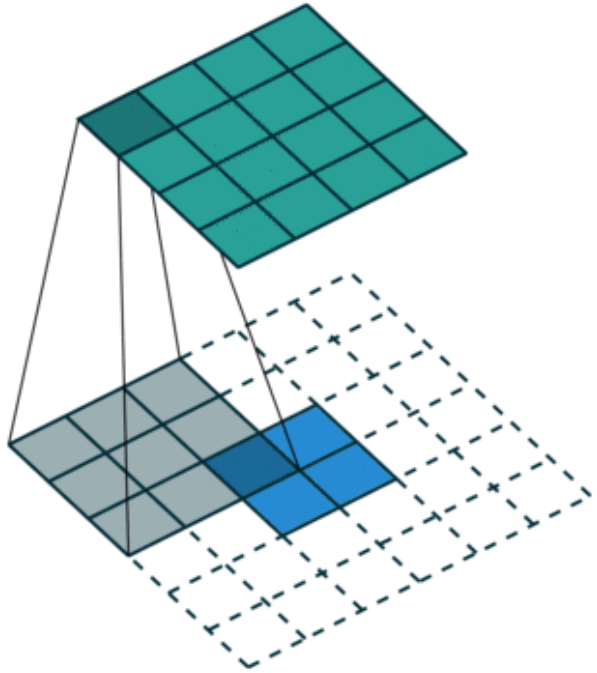


https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html

(Extreme) Semantic

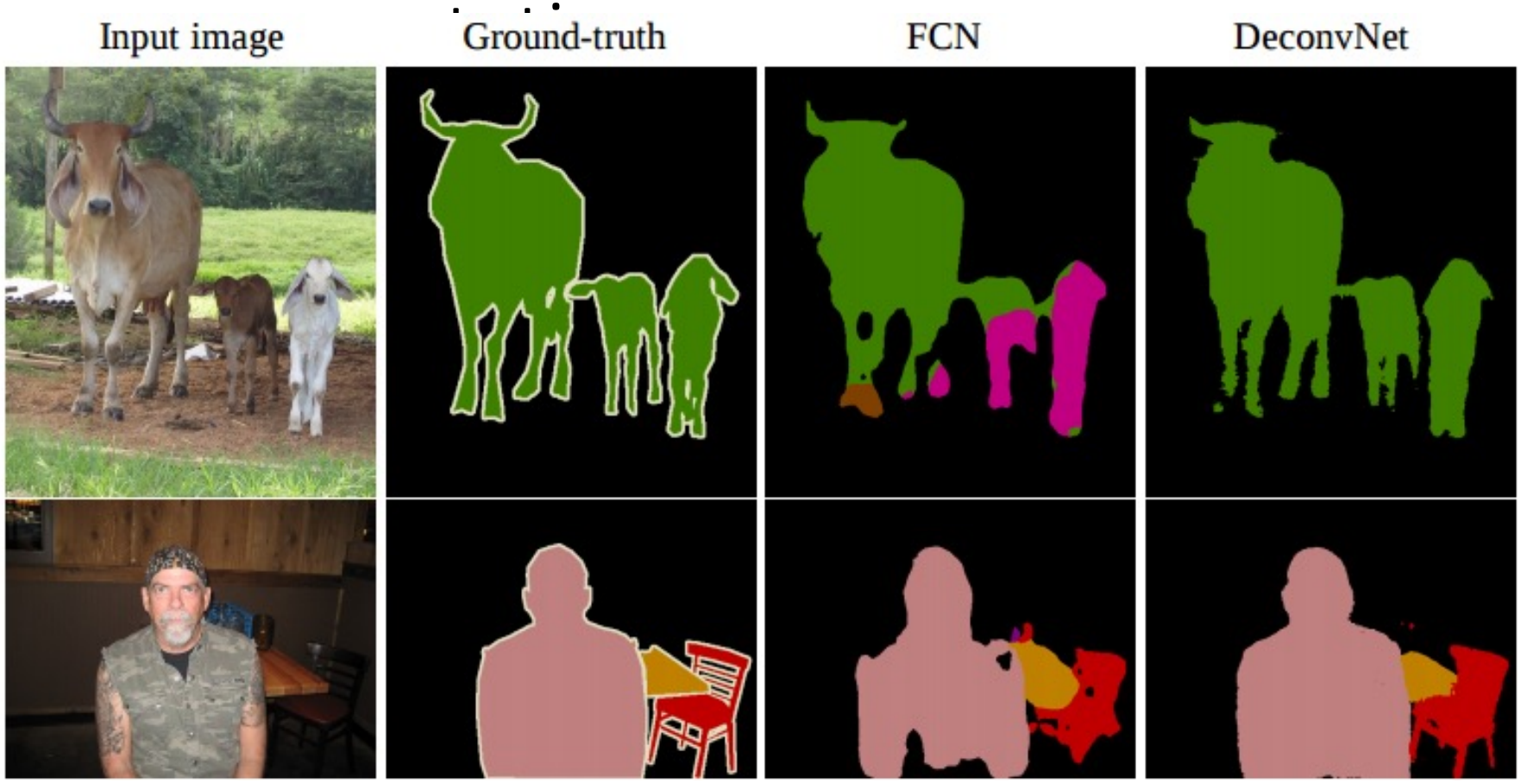


DeconvNet (2015)



https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html

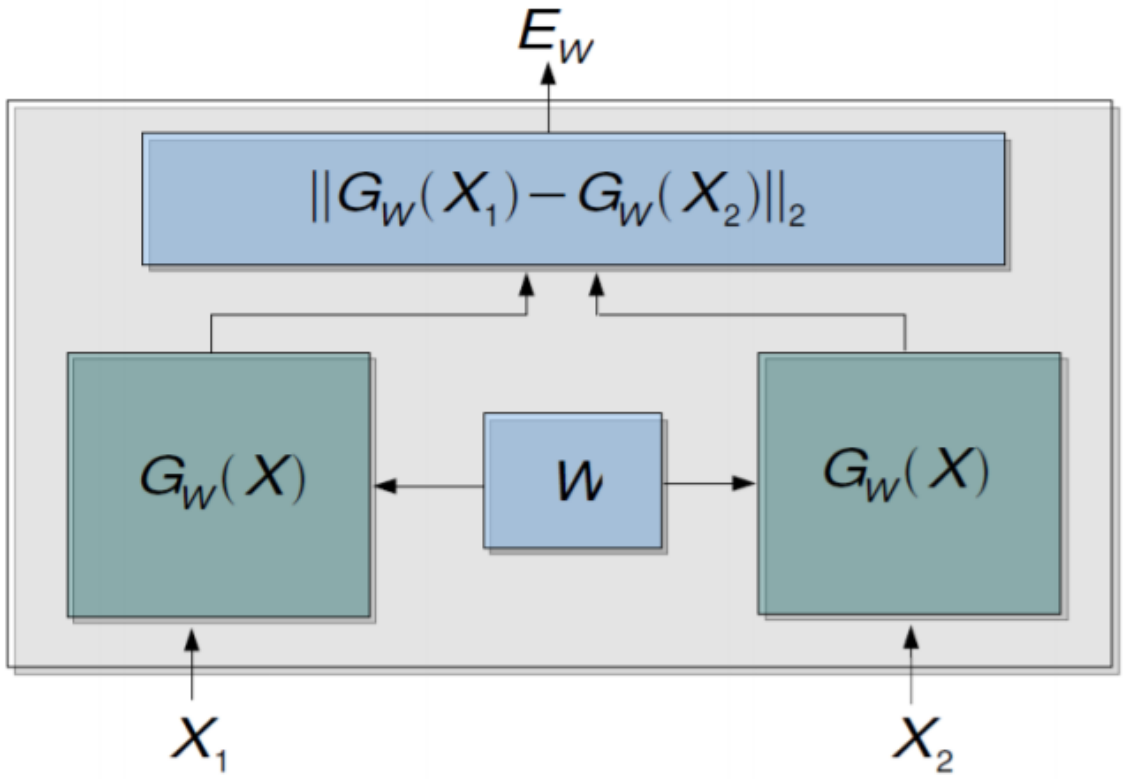
(Extreme) Semantic



https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html

Siamese Network

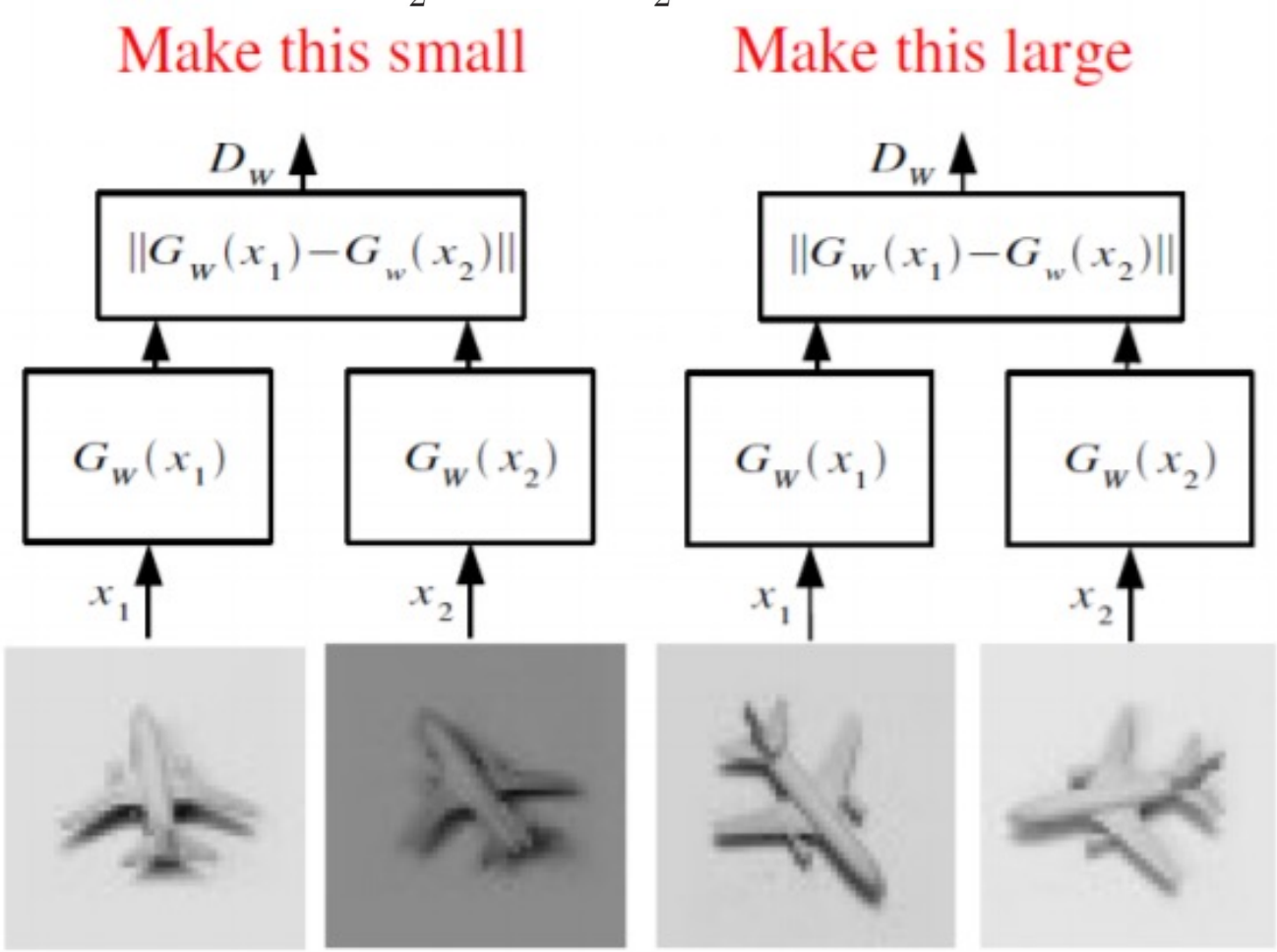
Siamese Network



Learning Hierarchies of Invariant Features. Yann LeCun

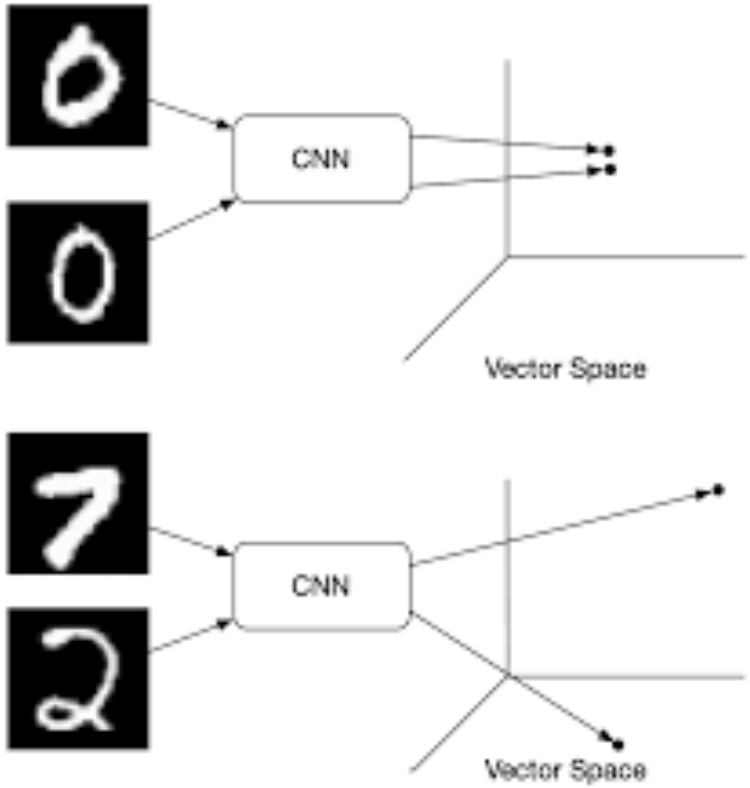
Siamese Network

$$(1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$



Similar Images (Y=0)

Dissimilar Images (Y=1)

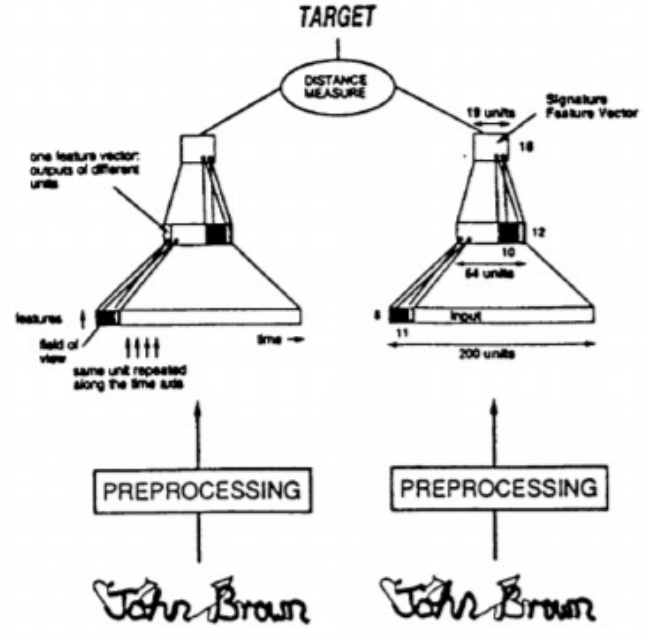


Learning Hierarchies of Invariant Features. Yann LeCun

Siamese Network

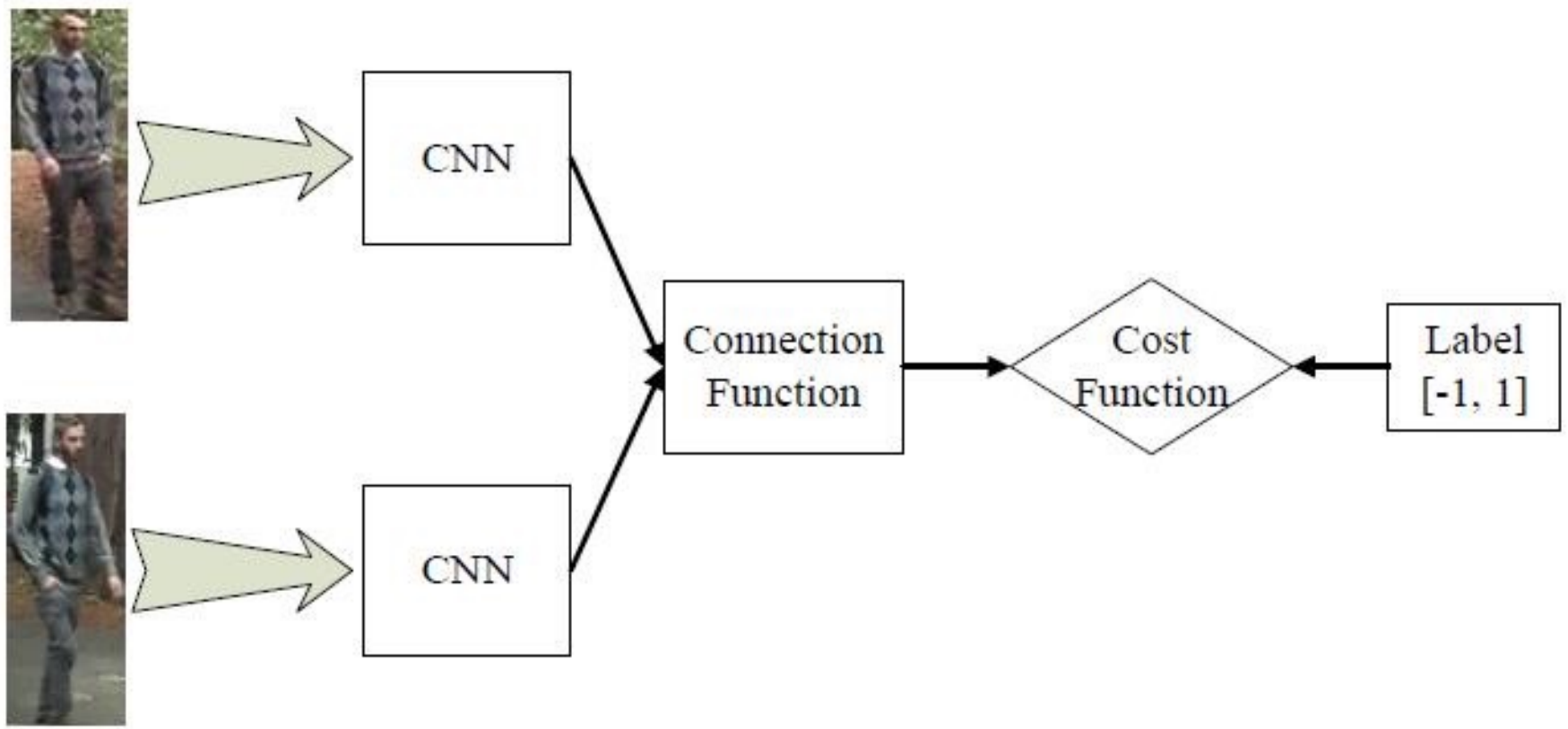
Application in Signature Verification

- The input is 8(feature) x 200(time) units.
- The cosine distance was used, (1 for genuine pairs, -1 for forgery pairs)



Bromley J, Guyon I, Lecun Y, et al. Signature Verification using a" Siamese" Time Delay Neural Network, NIPS Proc. 1994

Siamese Network (Person re-id)



http://www.fubin.org/research/Person_ReID/Person_ReID.html

MML

Questions?

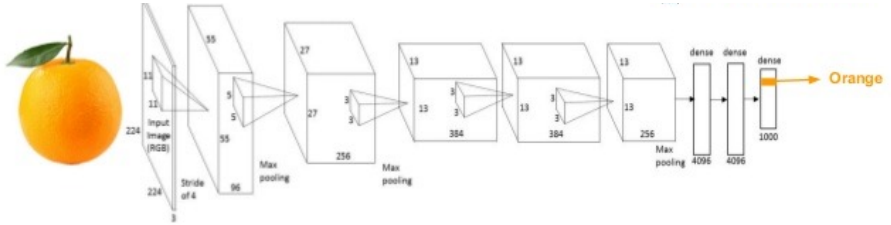
MML

Landscape of CNNs

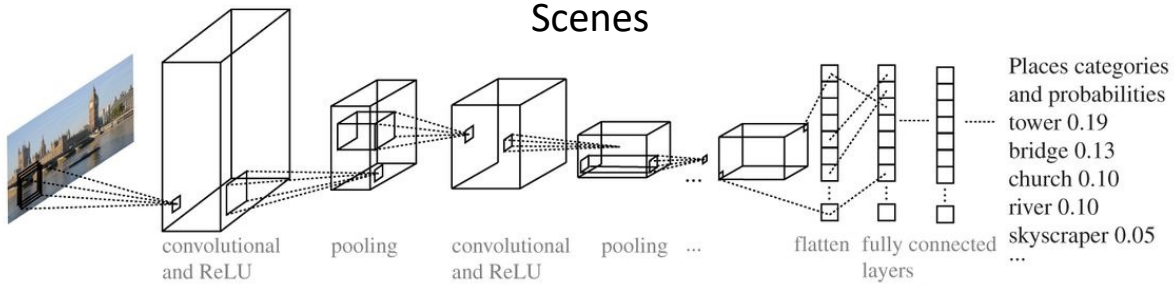
Applications

Image → Label

Objects



Scenes



Faces

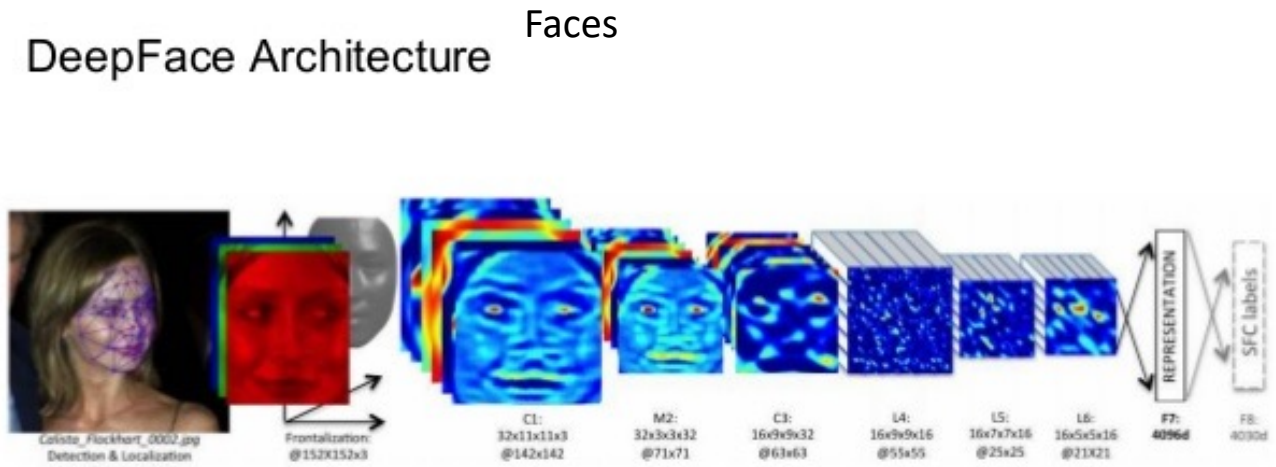
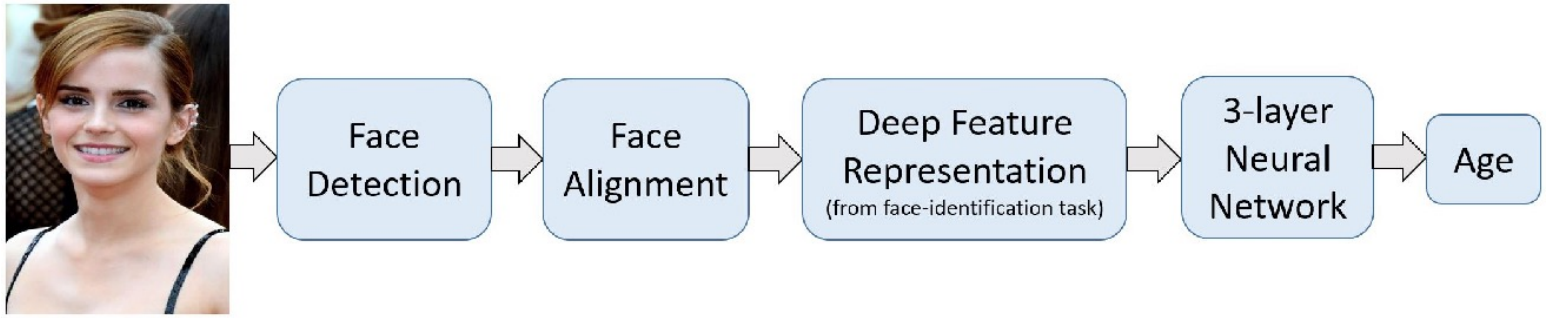
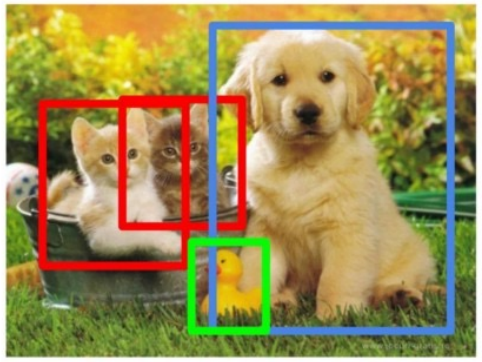


Image → Number



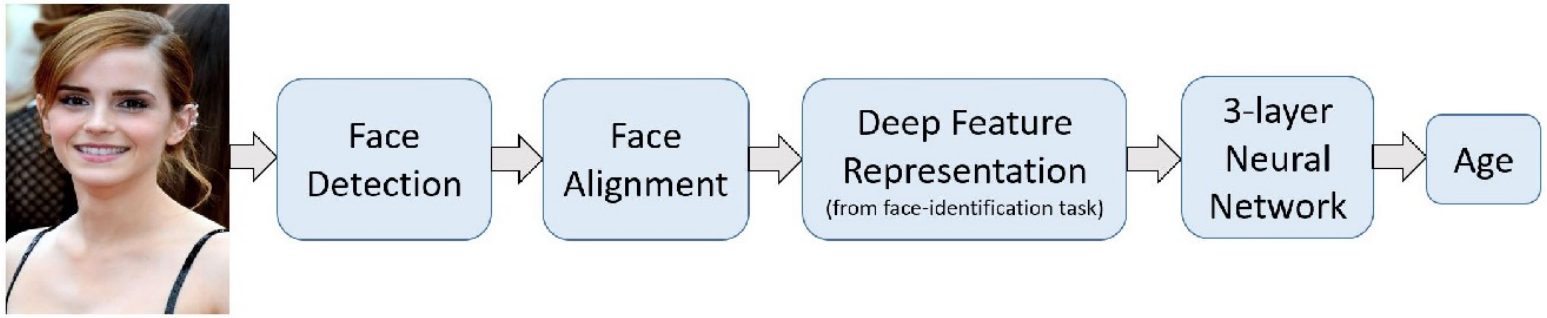
Object Detection



CAT, DOG, DUCK

Image → Number

Age Estimation



Object detection

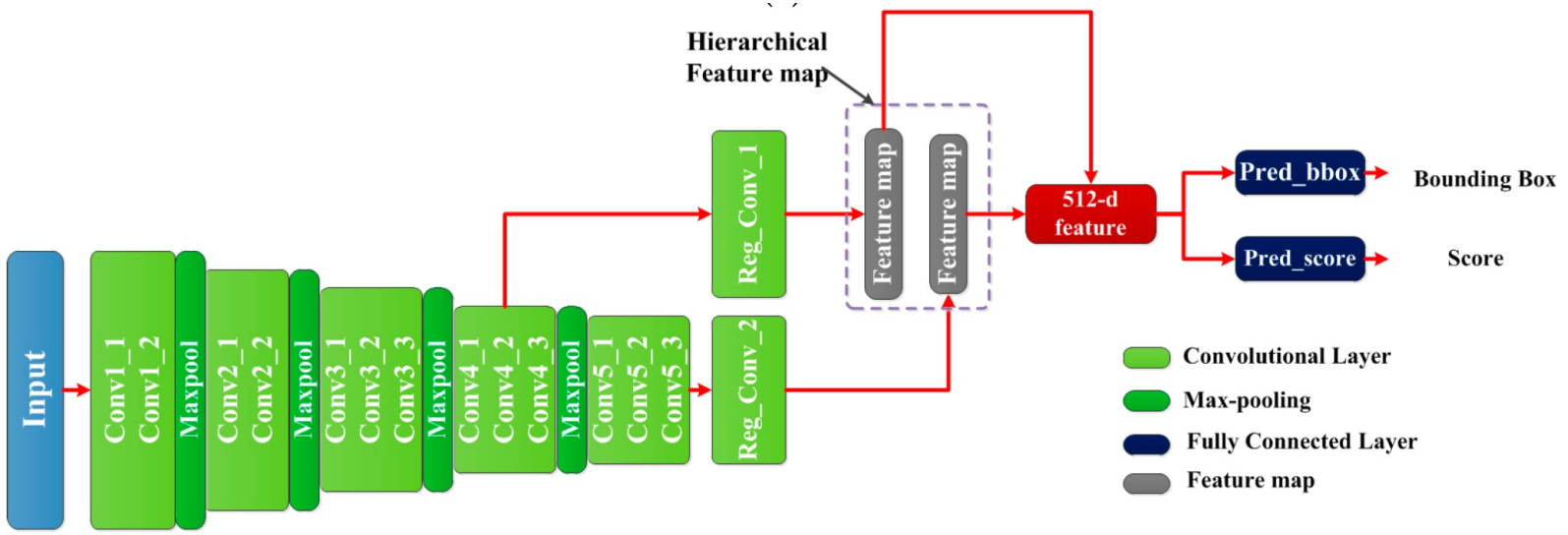
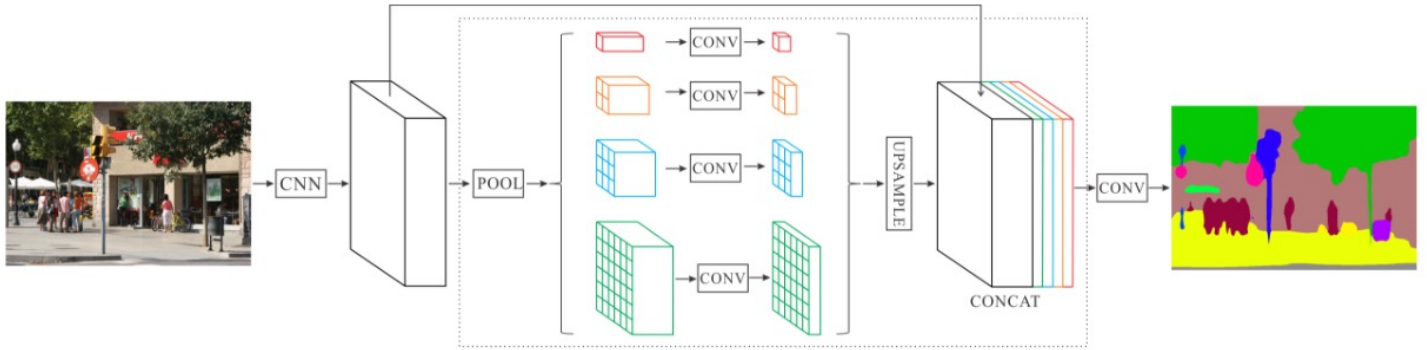


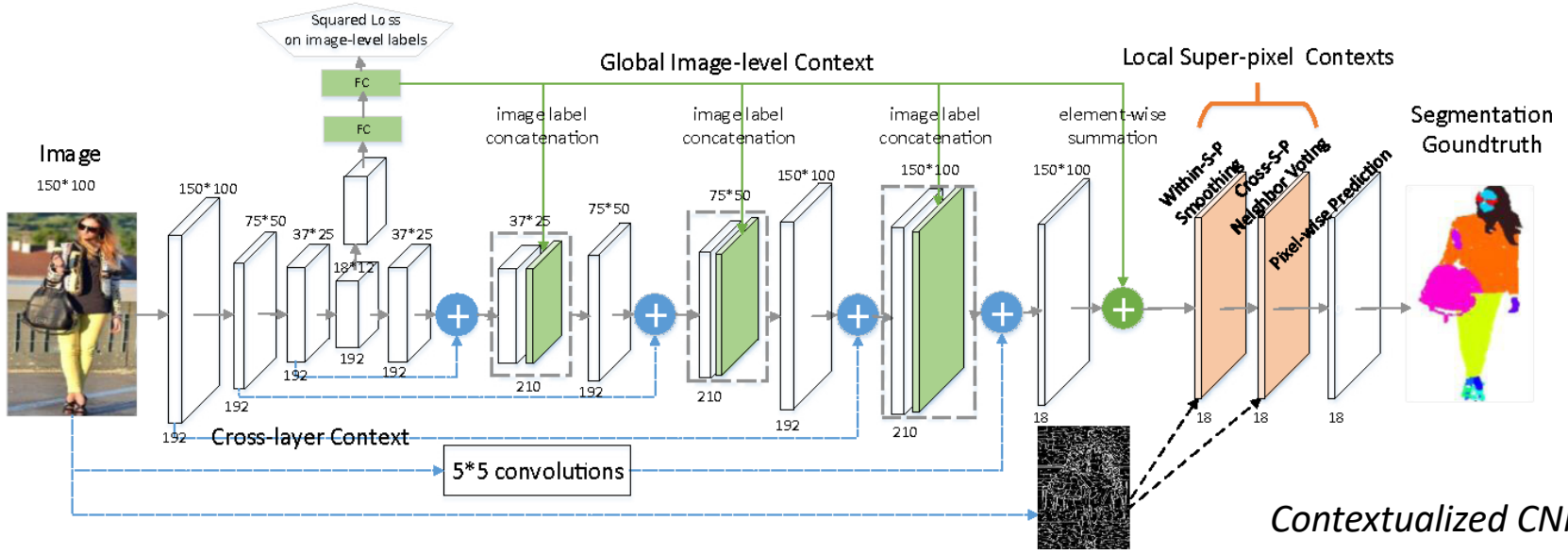
Image → Label Image

Scene Parsing



PSPNet

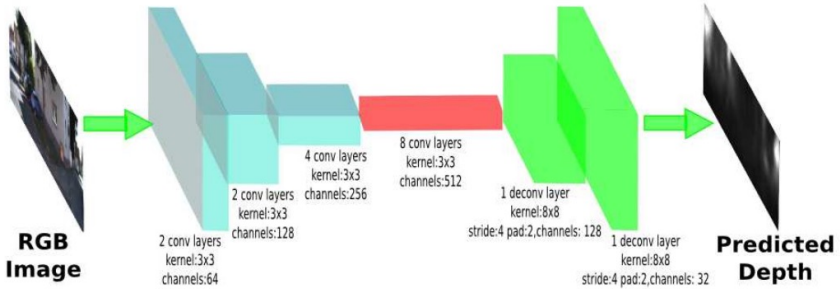
Object Parsing



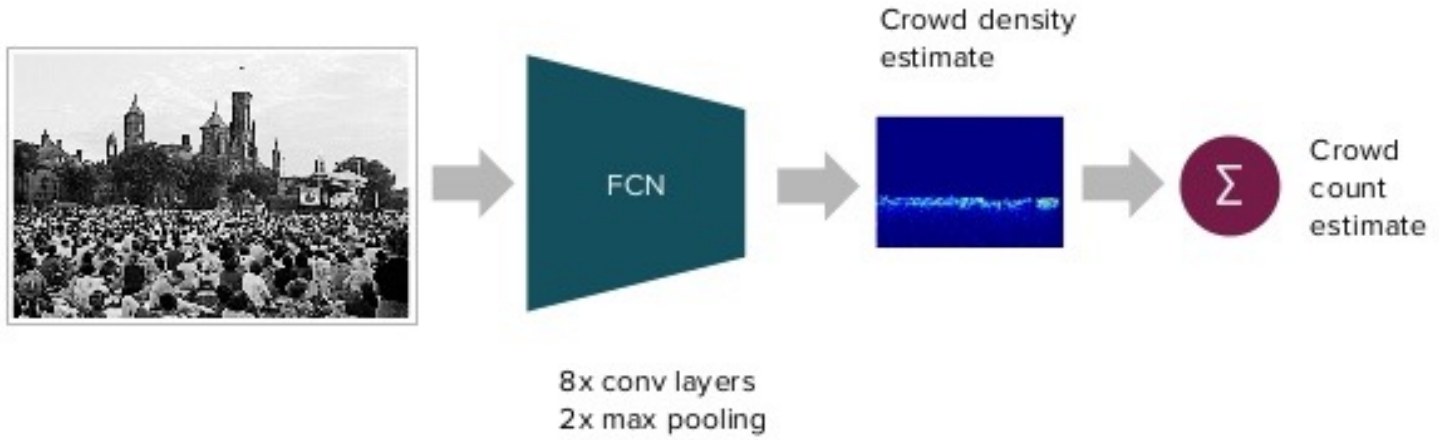
Contextualized CNN

Image → Image

Depth Estimation



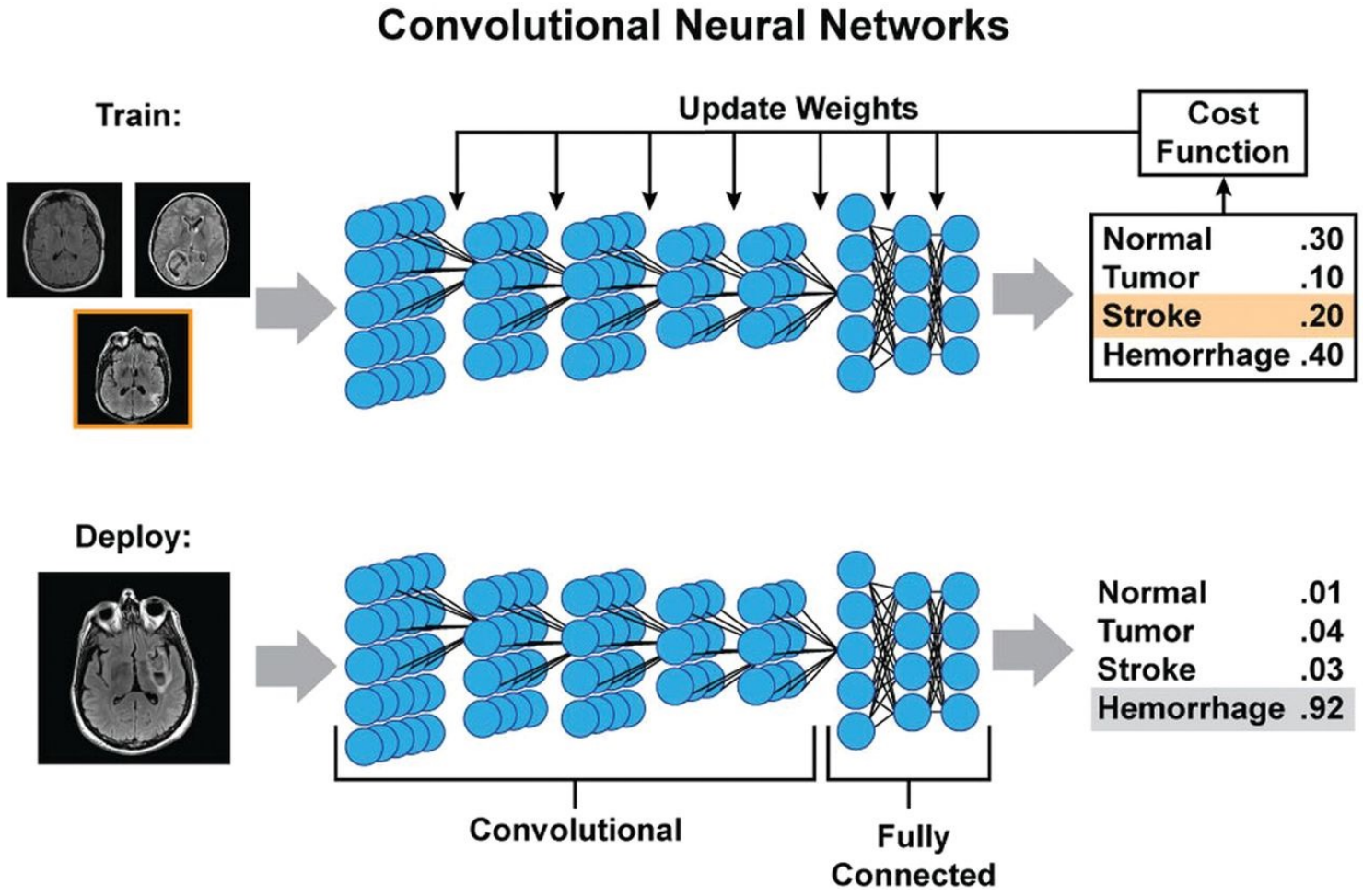
Crowd Counting



Marsden et al. **Fully convolutional crowd counting on highly congested scenes.** VISAPP 2017
https://arxiv.org/abs/1612.00220

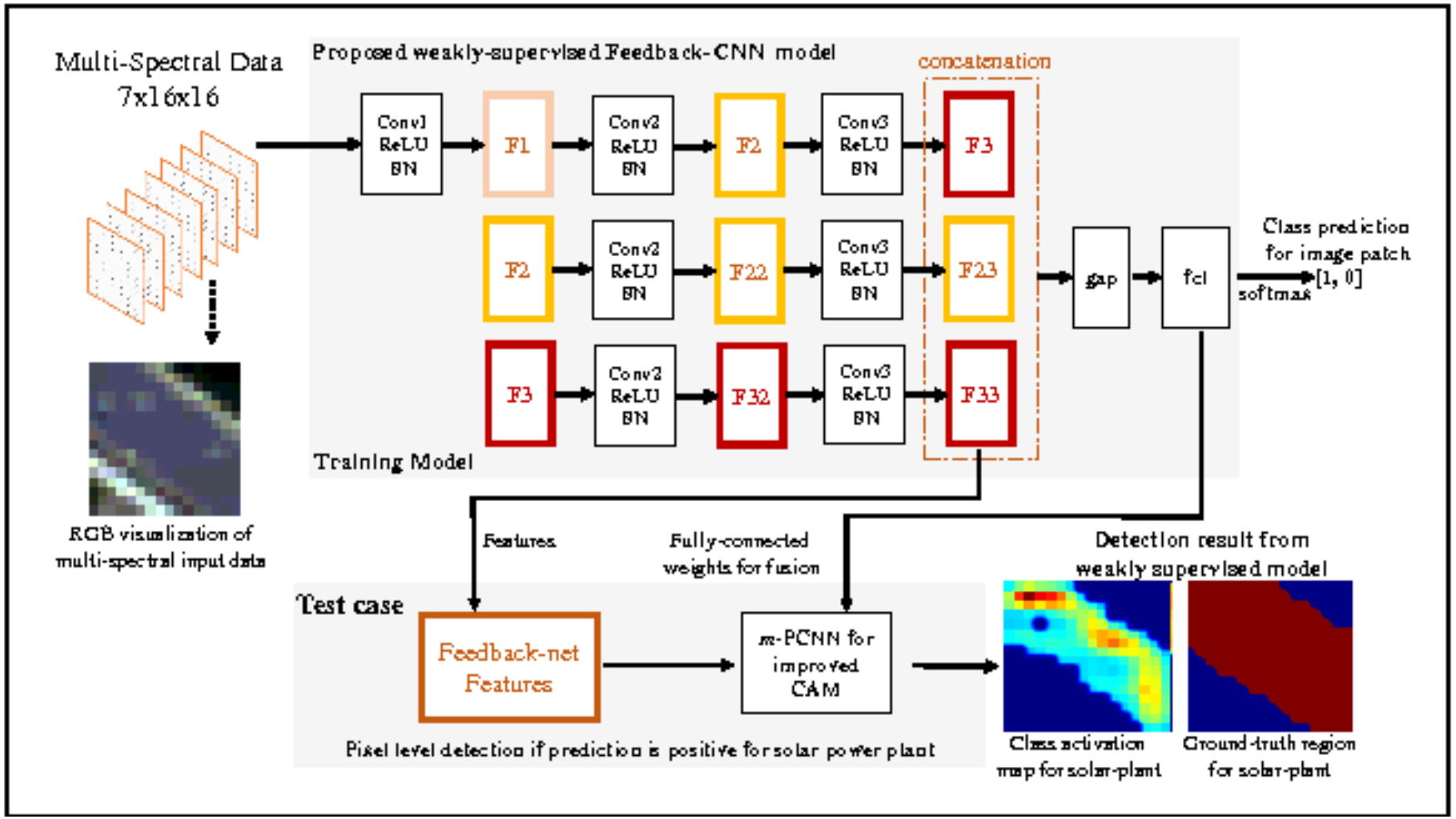
Multi-channel input

NeuroRadiology (fMRI)



Multi-channel input

Solar-power plant detection from multi-spectral data



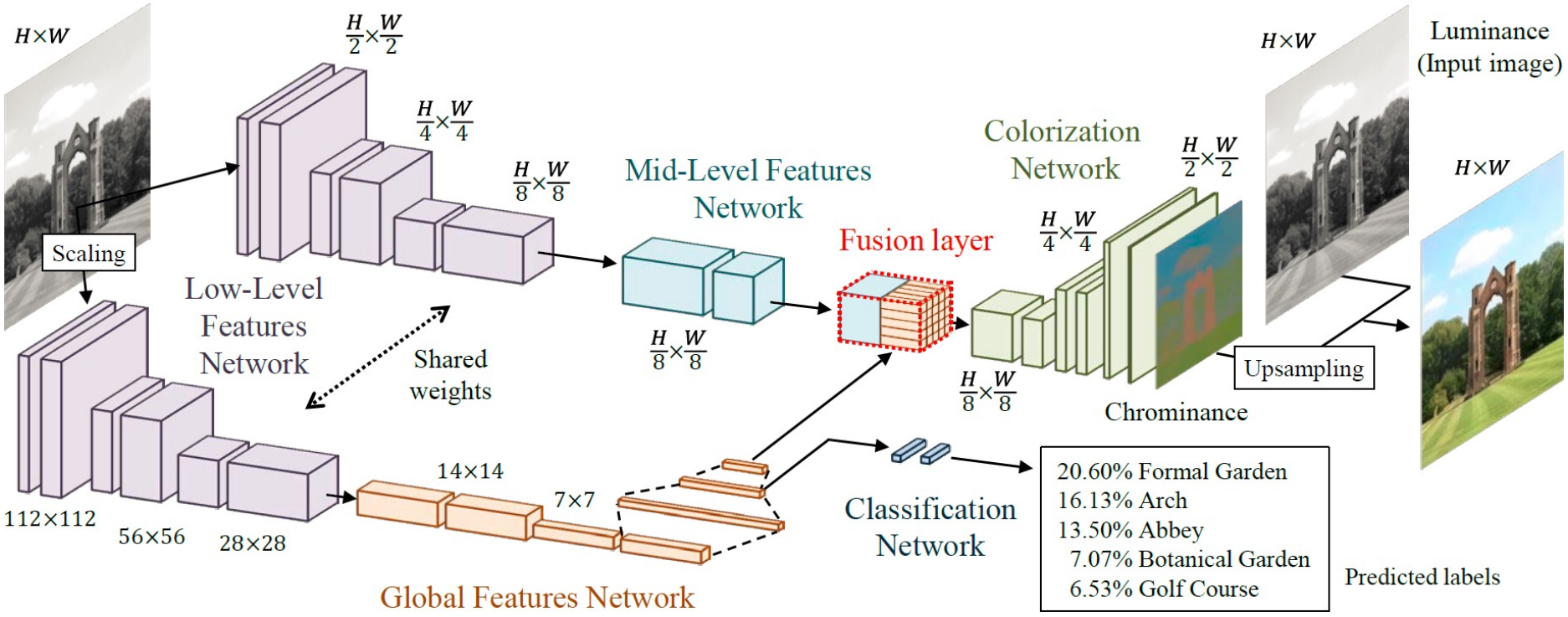
Multi-branch input

Image Colorization

https://algorithmia.com/algorithms/deeplearning/ColorfulImageColorization

Multi-branch input

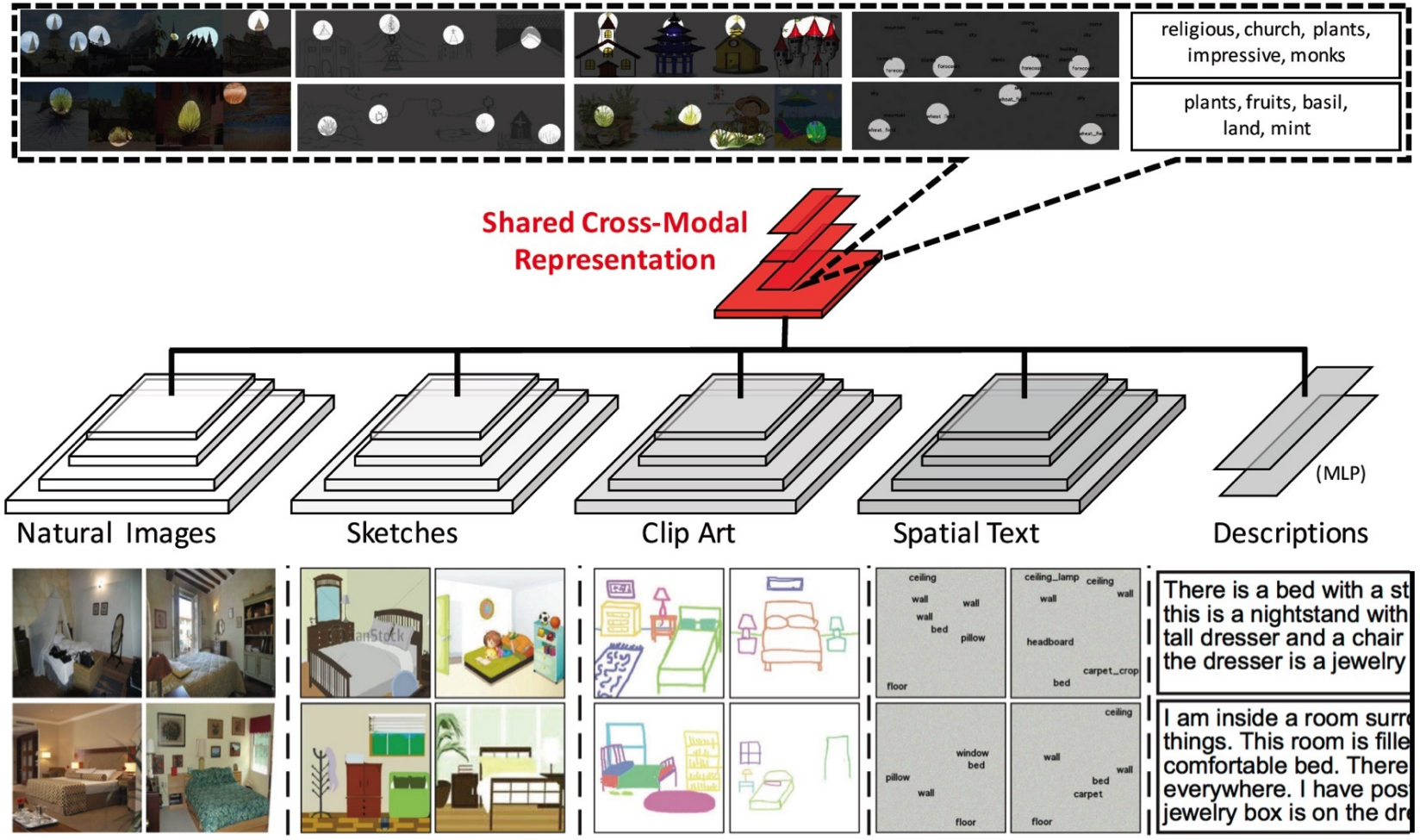
Image Colorization



http://hi.cs.waseda.ac.jp/~iizuka/projects/colorization/en/

Multi-branch input

Cross-modal Scene Classification



http://projects.csail.mit.edu/cmplaces/

Multi-branch input

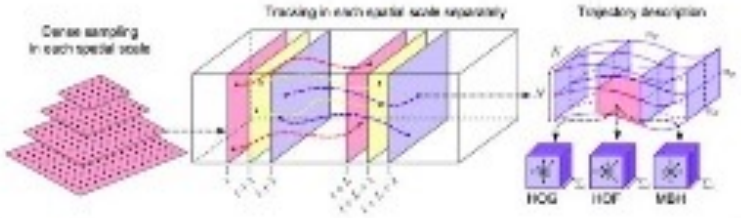
Activity Recognition

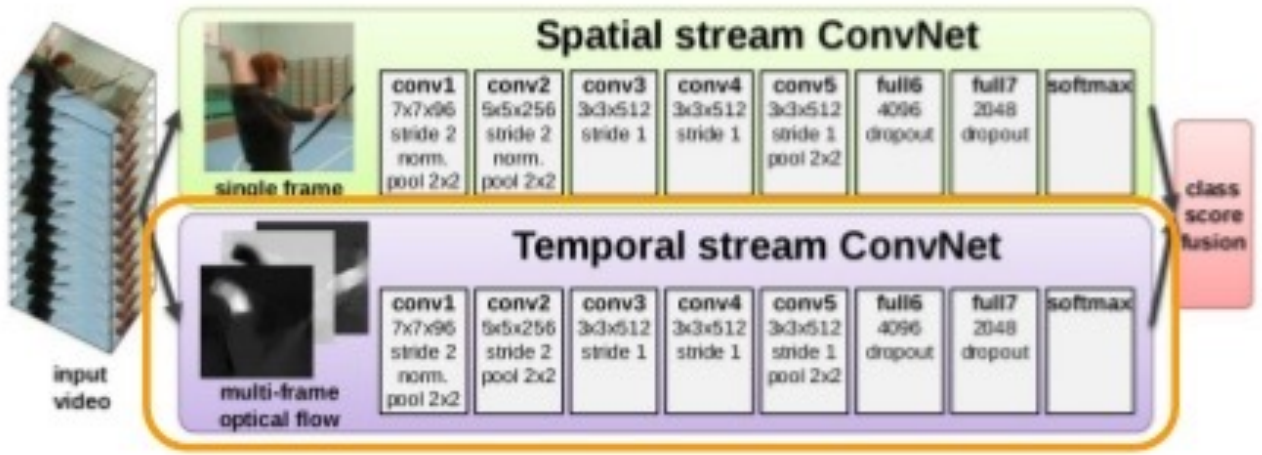
Recognition: Two stream

Two CNNs in paralel:

- One for RGB images
- One for Optical flow (hand-crafted features)

Fusion after the softmax layer

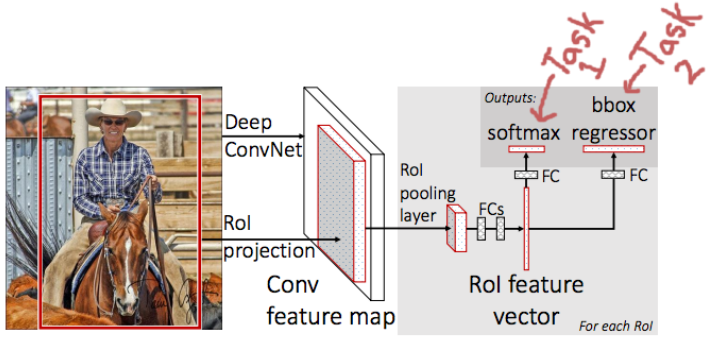




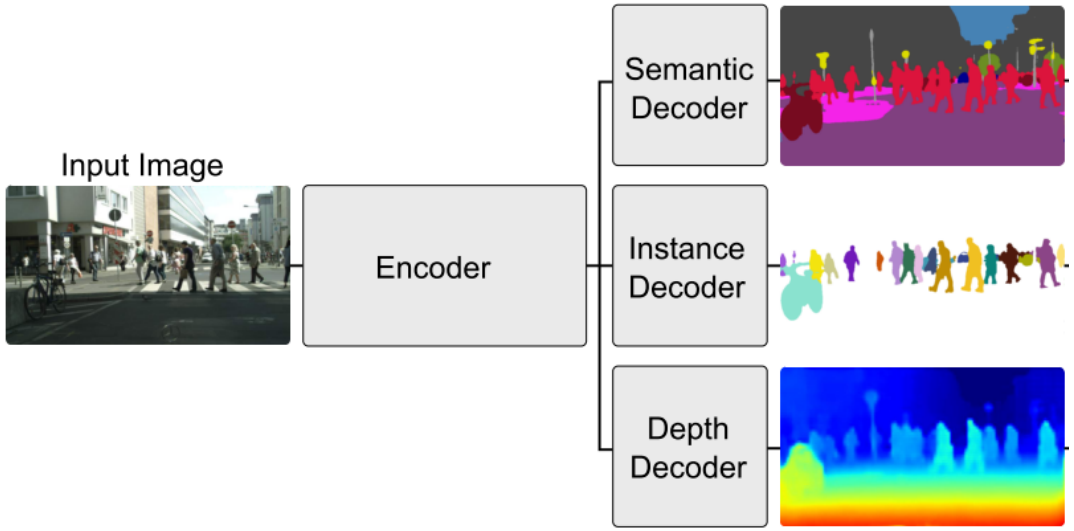
Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." NIPS 2014.

Multi-branch output

Object Detection, Classification



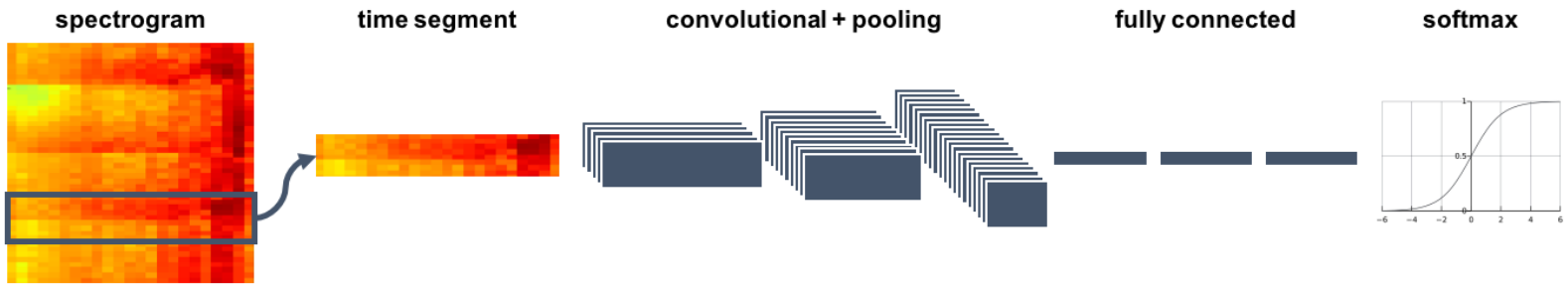
Scene Parsing



http://ruder.io/multi-task/

Image CNNs for non-image data

Audio Beat Detection



http://tommymullaney.com/projects/rhythm-games-neural-networks

MML

Questions?