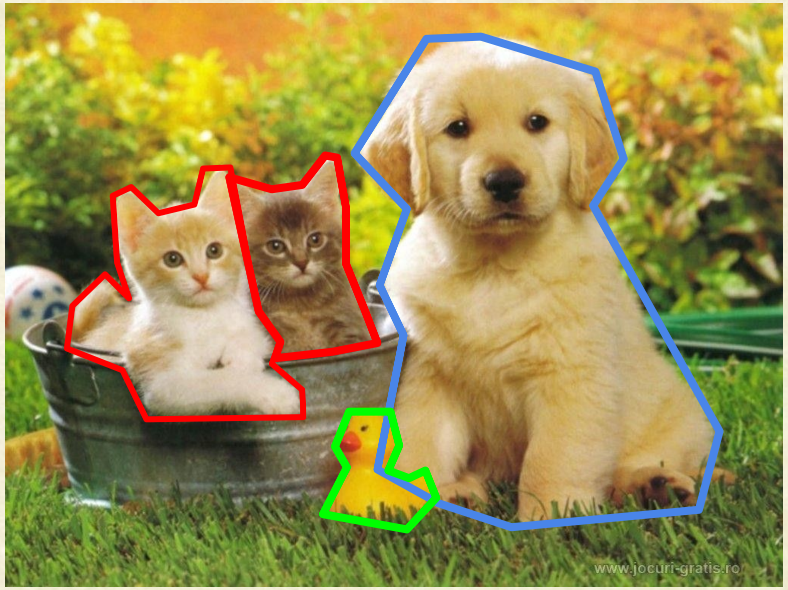
CS7.505: Computer Vision

Spring 2024: Segmentation & Dense Prediction



Makarand Tapaswi
Katha AI Lab, CVIT,
IIIT Hyderabad

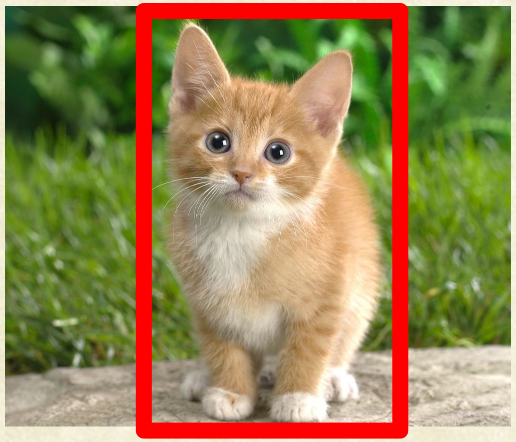
Anoop Namboodiri
Biometrics and Secure ID Lab, CVIT,
IIIT Hyderabad

Computer Vision Tasks

Classification

CAT

Classification + Localization

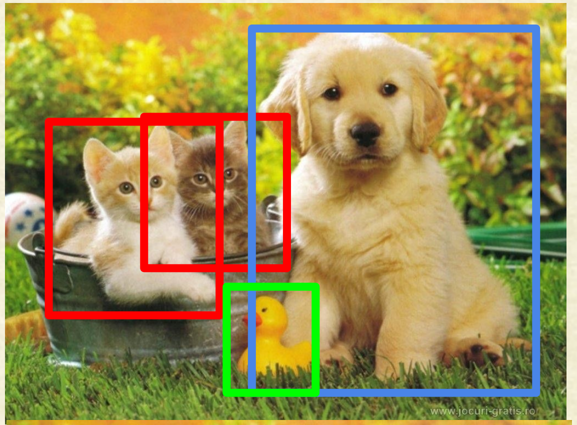


CAT

Single object

CAT, DOG, DUCK

Multiple objects

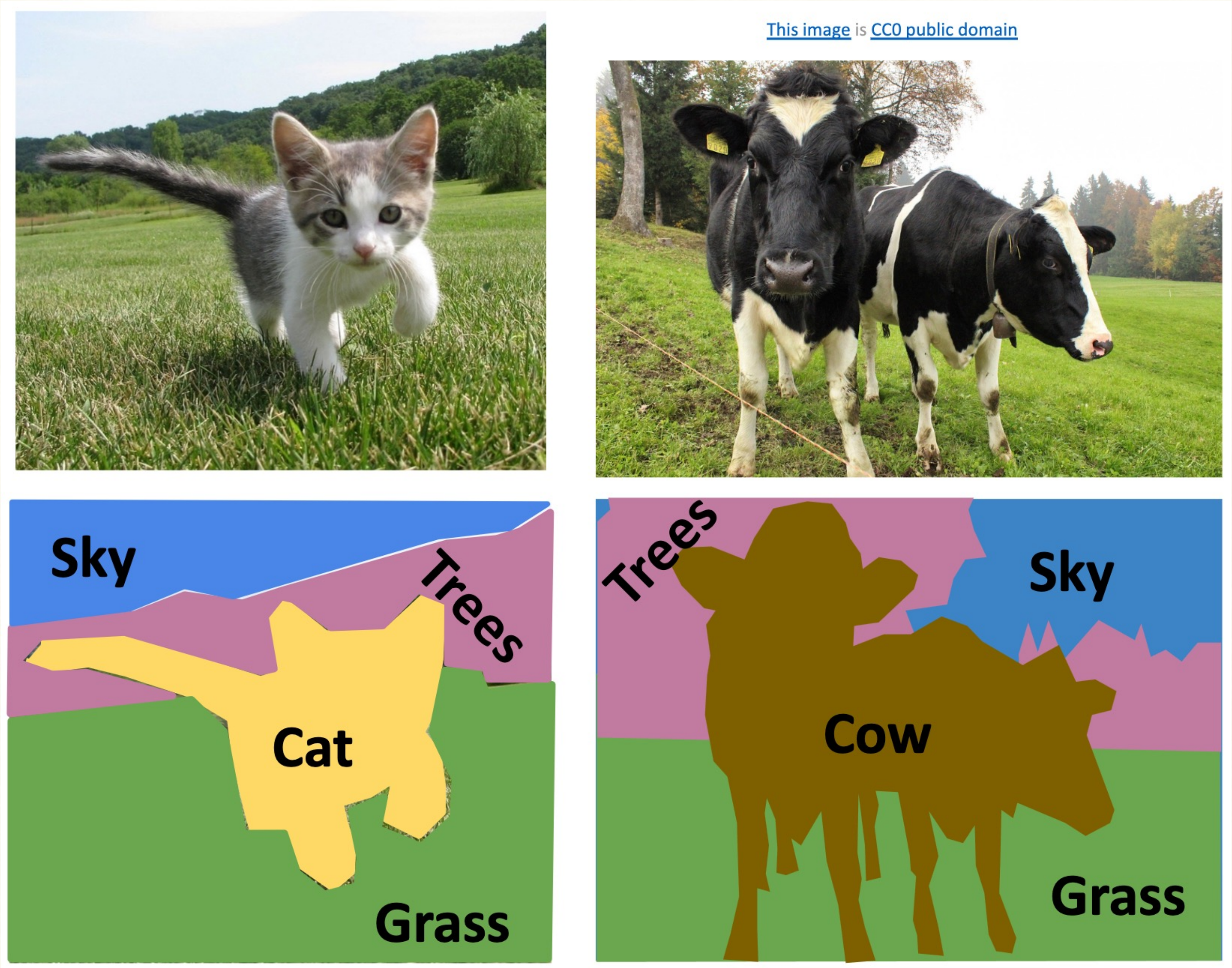


Object Detection

Semantic Segmentation

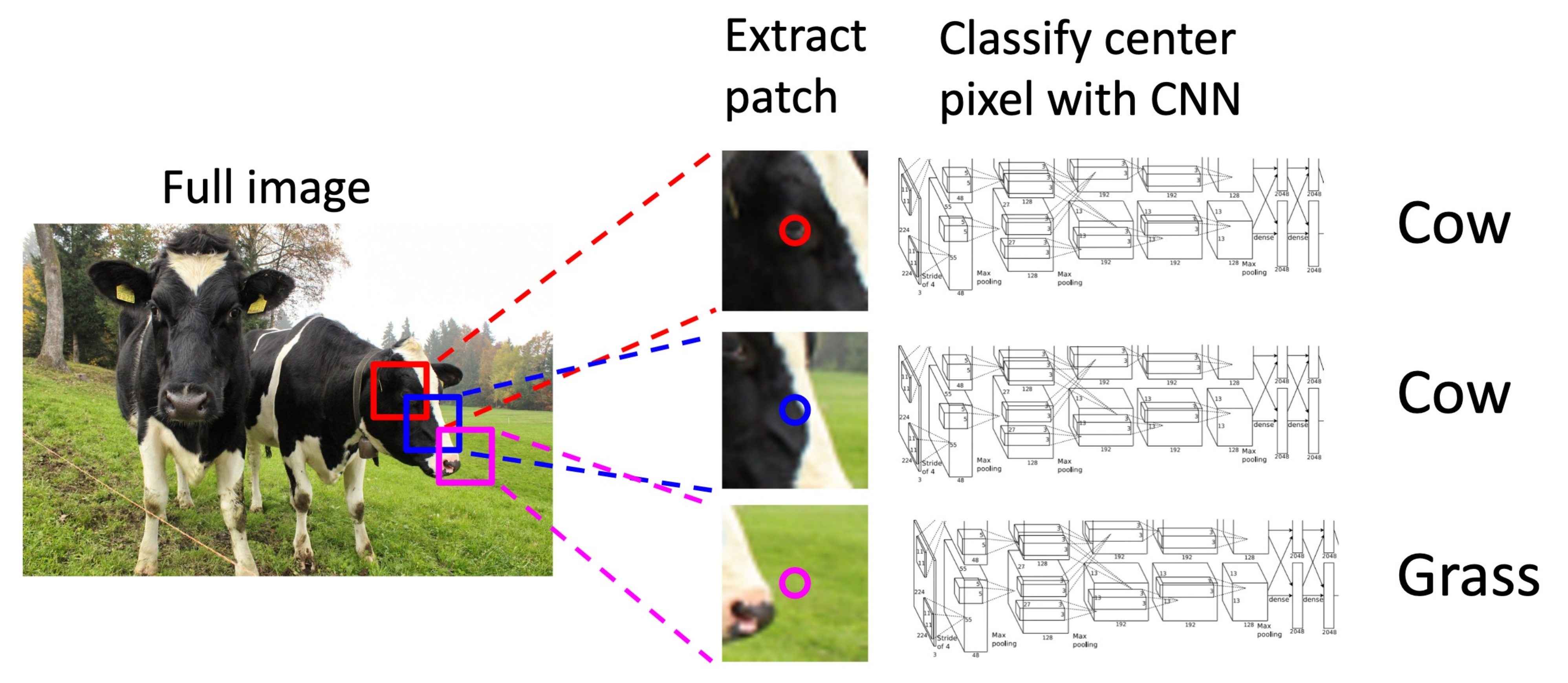
Instance Segmentation

Semantic Segmentation



Vision person’s first idea: sliding window!

- Don’t do this please …



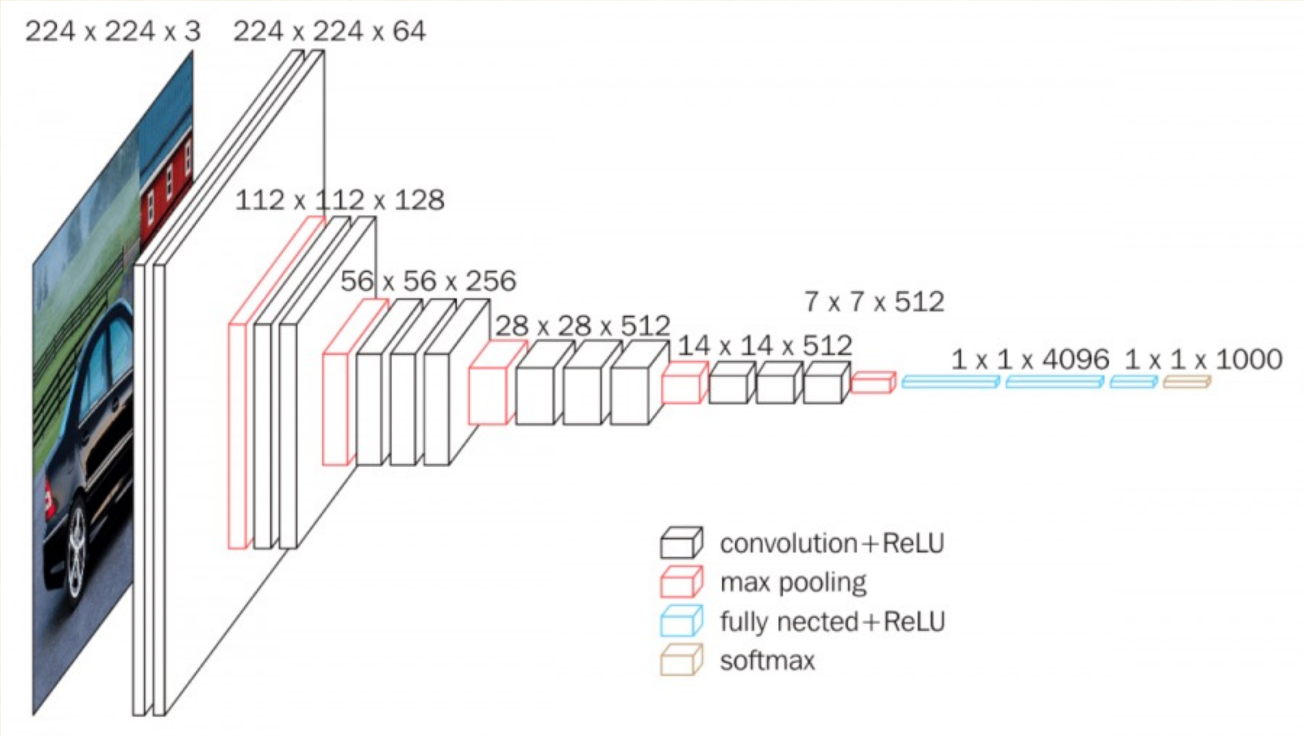
Slide credit: https://web.eecs.umich.edu/~justincj/slides/eecs498/WI2022/598_WI2022_lecture15.pdf

Semantic Segmentation: Challenges

- Number of pixels is very large
  - Still less than number of windows!!
- A pixel by itself does not contain enough information for the task
  - We need to use content information of pixels around
- The label of a pixel is highly correlated to labels of neighboring pixels
  - We need to use label predictions of neighboring pixels
- Objects tend to have highly irregular boundaries
- Porous objects and boundary pixels pose additional challenges

Recap CNNs (VGG)

VGG16: Number of Parameters by Layers

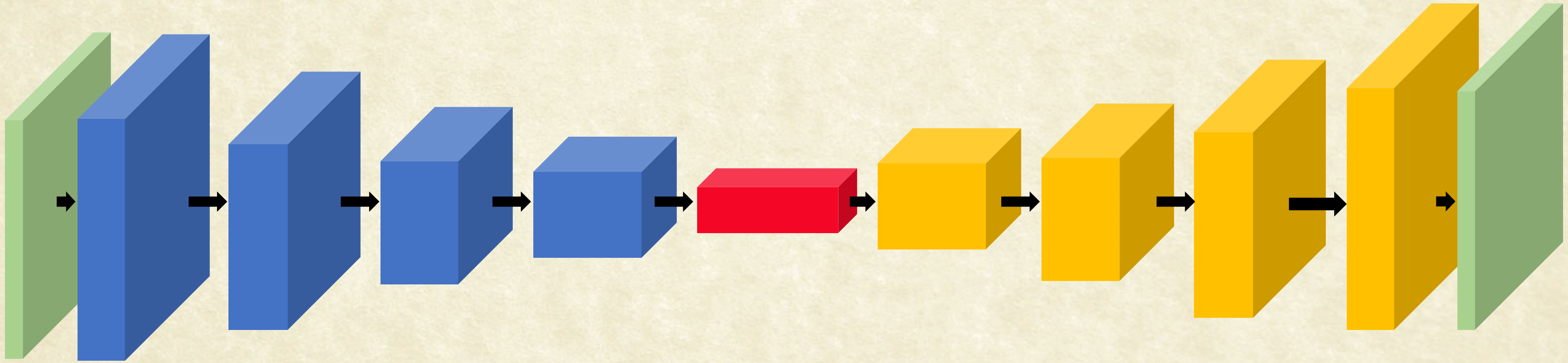


- Max-Pool layers are not shown in the table as they do not contain any learnable parameters
- Almost 90% of parameters are in the 3 FC layers !!

Layer	Type	Size	Chanls	Params (M)
1	Conv	3x3	64	0.002
2	Conv	3x3	64	0.037
3	Conv	3x3	128	0.074
4	Conv	3x3	128	0.148
5	Conv	3x3	256	0.295
6	Conv	3x3	256	0.590
7	Conv	3x3	256	0.590
8	Conv	3x3	512	1.180
9	Conv	3x3	512	2.360
10	Conv	3x3	512	2.360
11	Conv	3x3	512	2.360
12	Conv	3x3	512	2.360
13	Conv	3x3	512	2.360
14	FC	250888	4096	102.765
15	FC	4096	4096	16.781
16	FC	4096	1000	4.097
				138.423

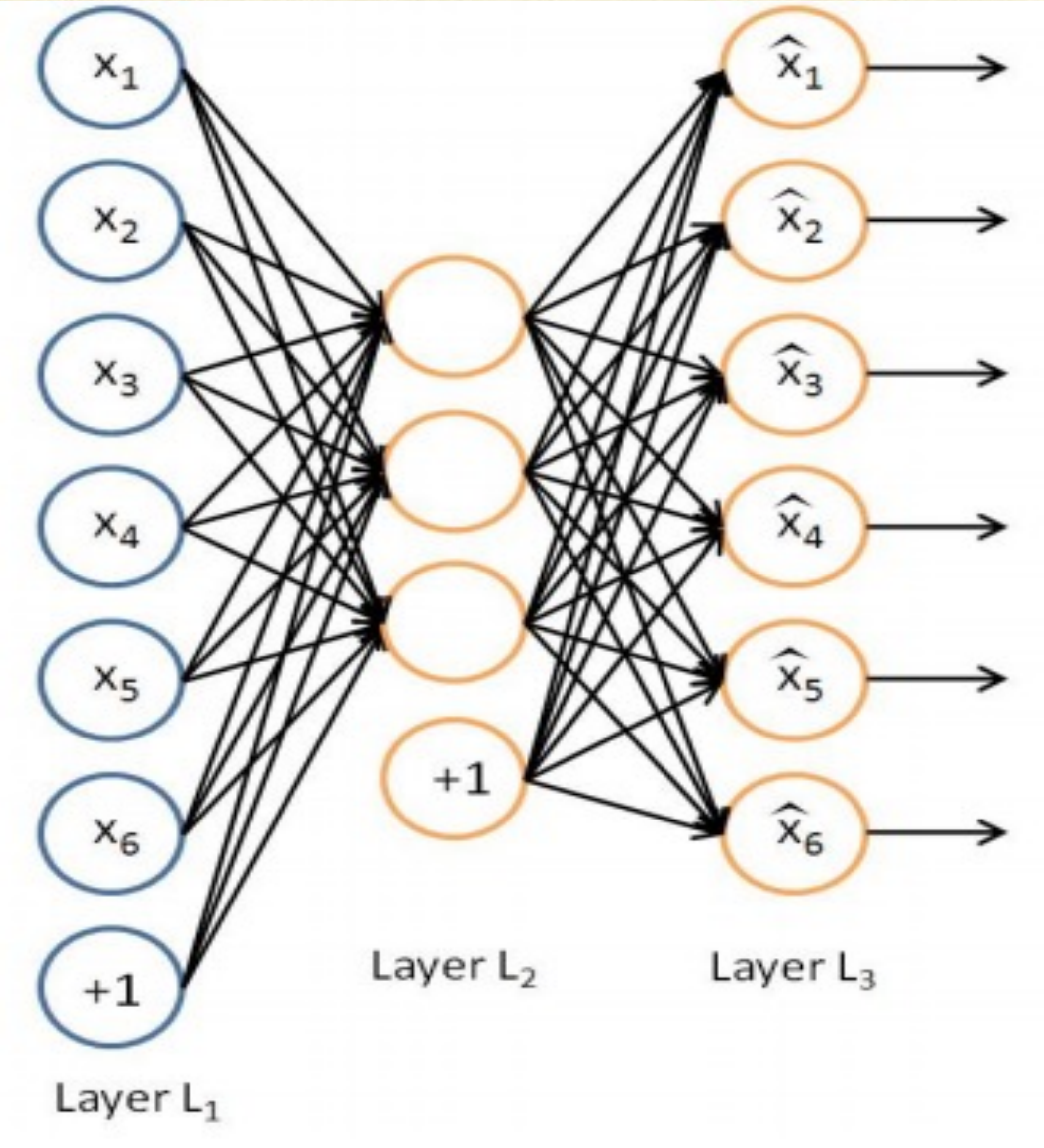
DL Solution: Fully Convolutional Network

- A series of Conv layers (+ MP, Norm, ReLU); avoid FC Layers
- Inputs, intermediate outputs and final outputs are 2D (or 3D)
- Often has a bottleneck in middle (Encoder-Decoder)



Auto Encoders

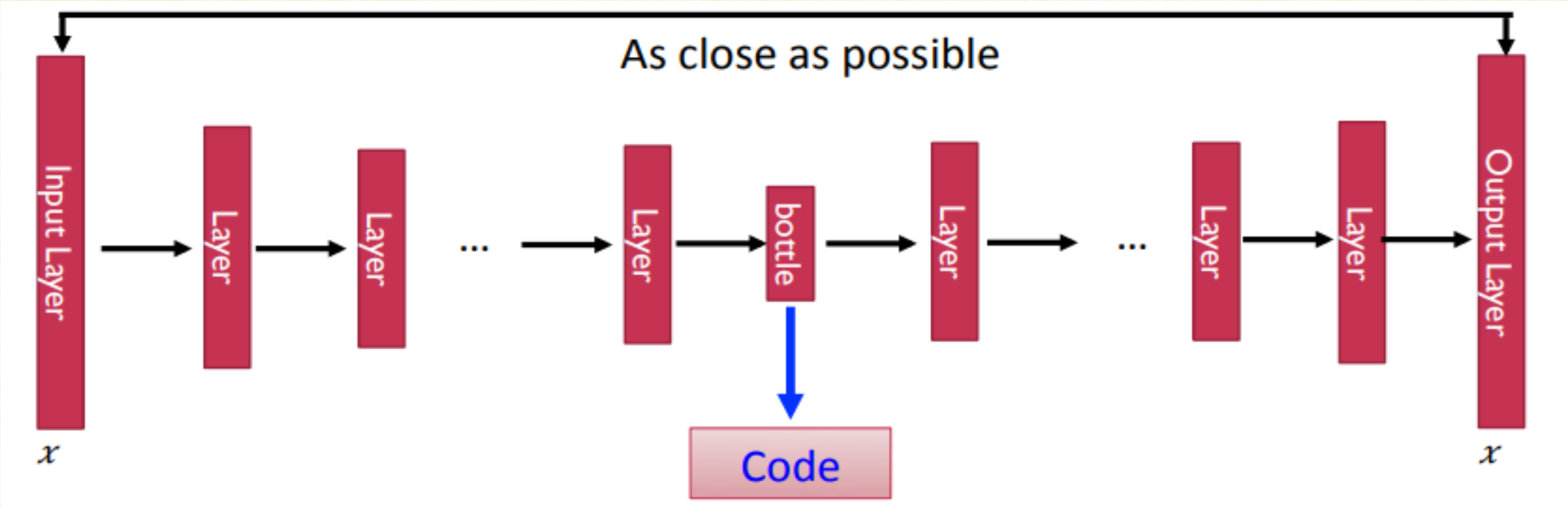
- How to train a network with only unlabeled data?
- Idea: Use the input itself as output.
- Network learns to reconstruct
- “Bottleneck” layer learns a compact representation.



Deep Auto-encoder

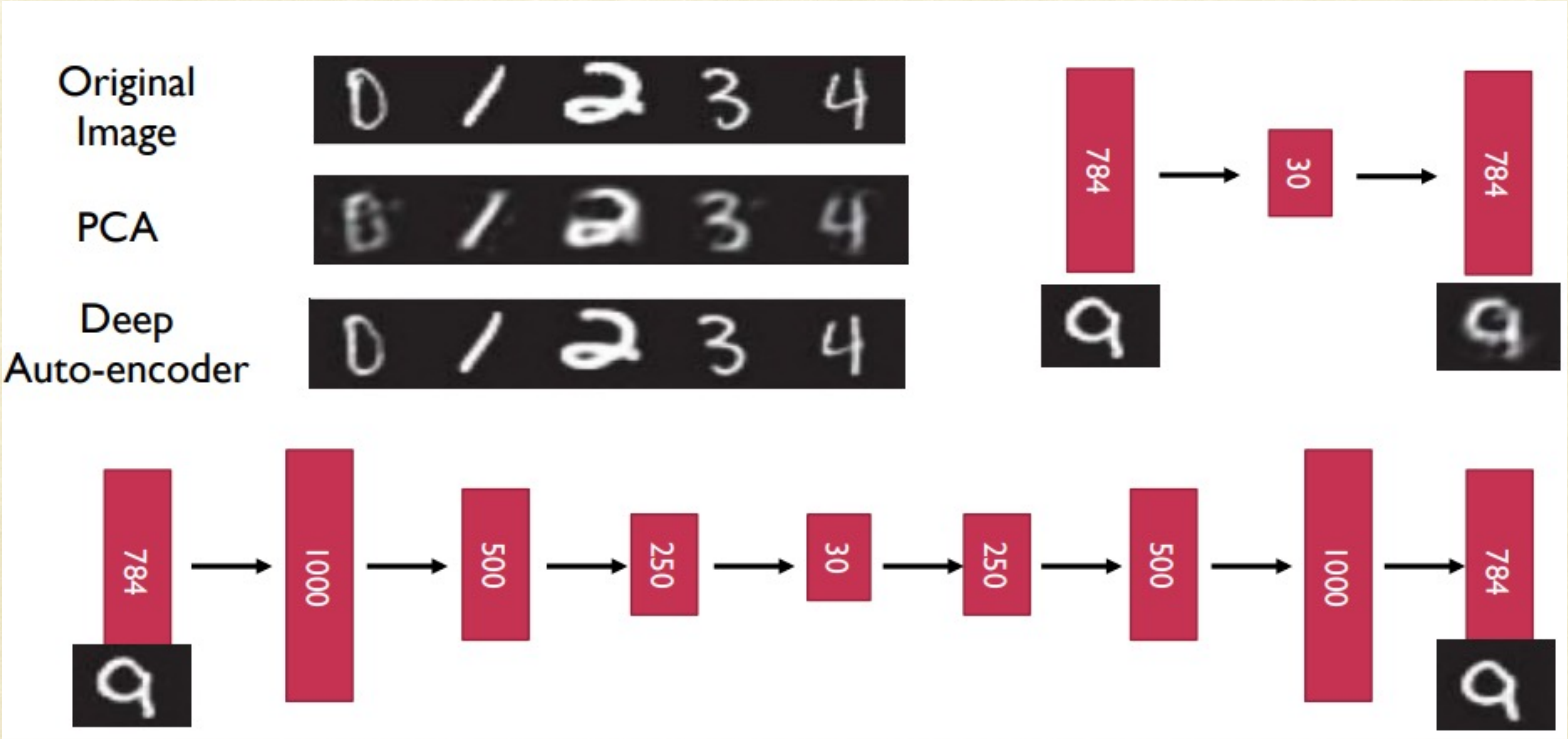
- Of course, the auto-encoder can be deep

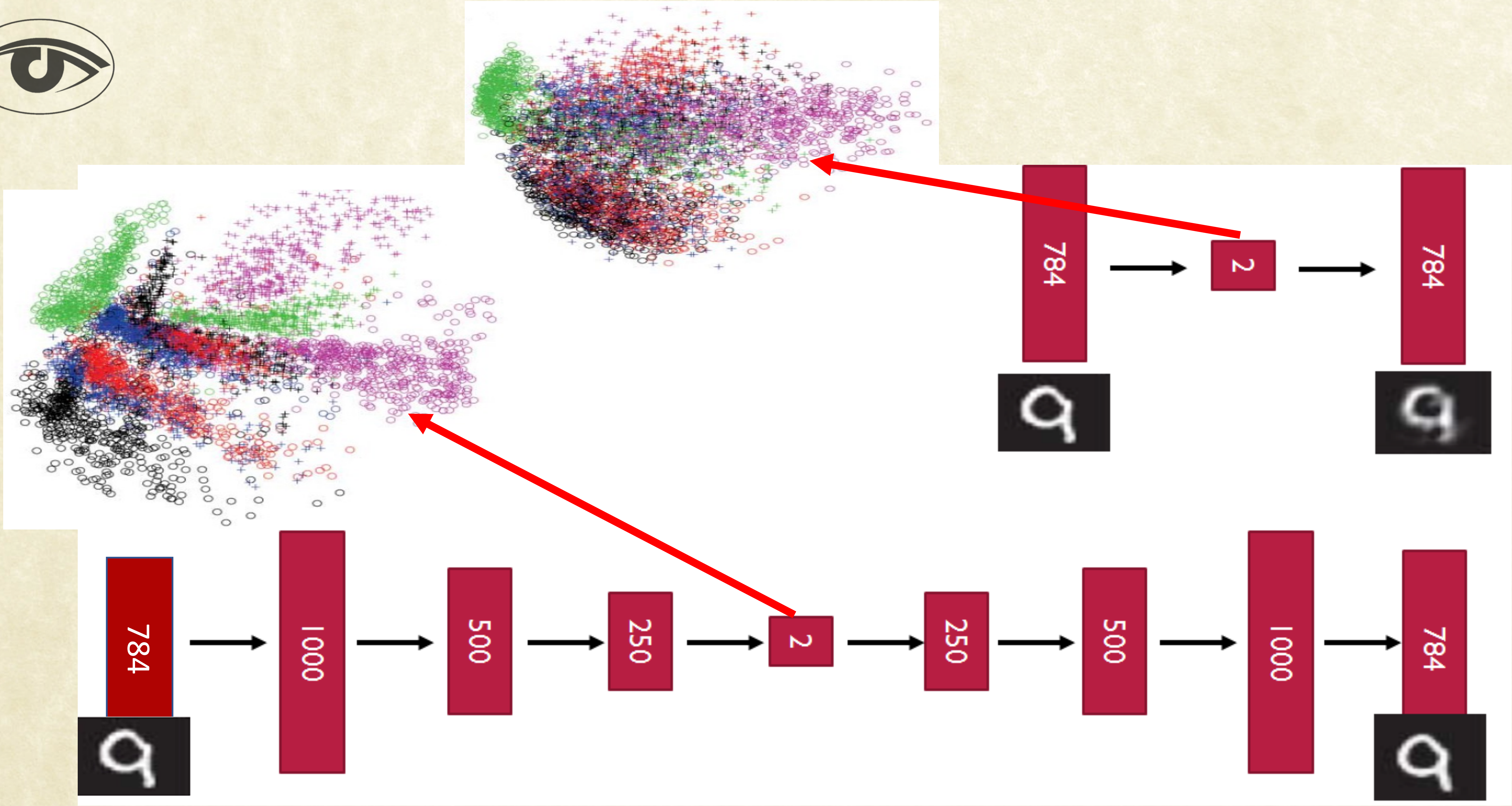
Symmetry is not necessary



Reference: Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

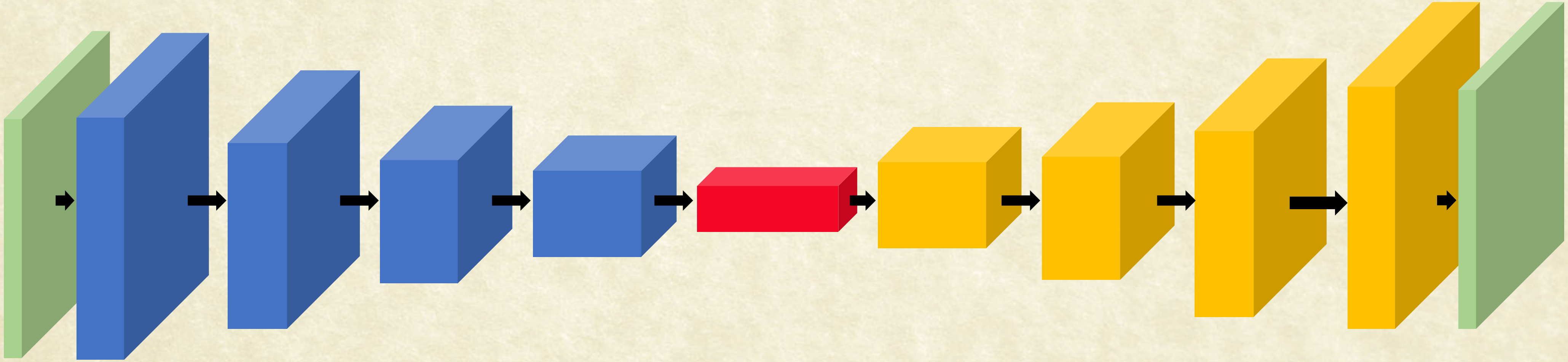
Deep Auto-encoder




784
2
784
784
1000
500
250
2
250
500
1000
784

Fully Convolutional Autoencoders

- All layers of the autoencoder are convolutional
- Output size reduces till bottleneck; channels increase.
- The encoder is often taken from a pre-trained network
- The network is independent of input image size !
- The layers after bottleneck does transposed convolution

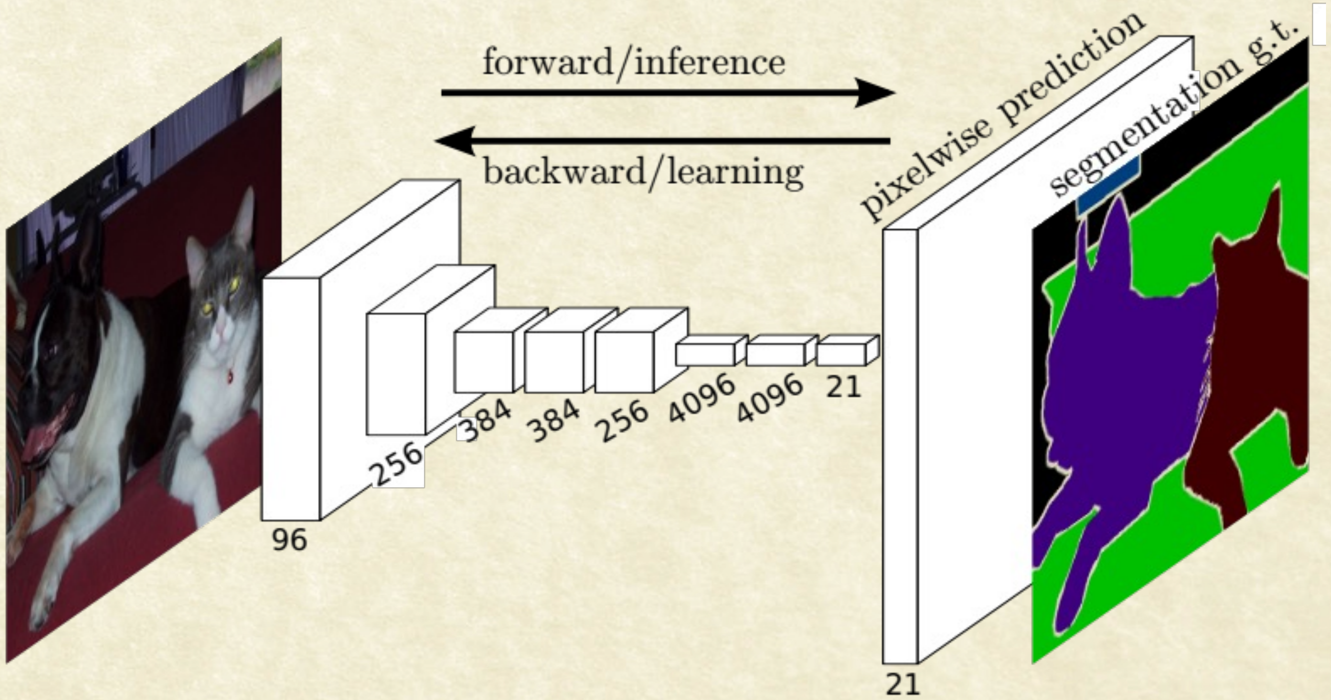


Convolution
(Encoder)

Transposed convolution
(Decoder)

Fully Convolutional Network for Segmentation

- Encoder captures semantic information; Decoder projects it into the pixel space
- Bottleneck layer results in low resolution; fuzzy boundary
- Can handle arbitrary image size



Long, et al. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015. (47k citations!)

Diving into details

- “Upsampling” or “transposed convolutions”
- https://web.eecs.umich.edu/~justincj/slides/eecs498/WI2022/598_WI2022_lecture15.pdf
- Slides 43 – 70

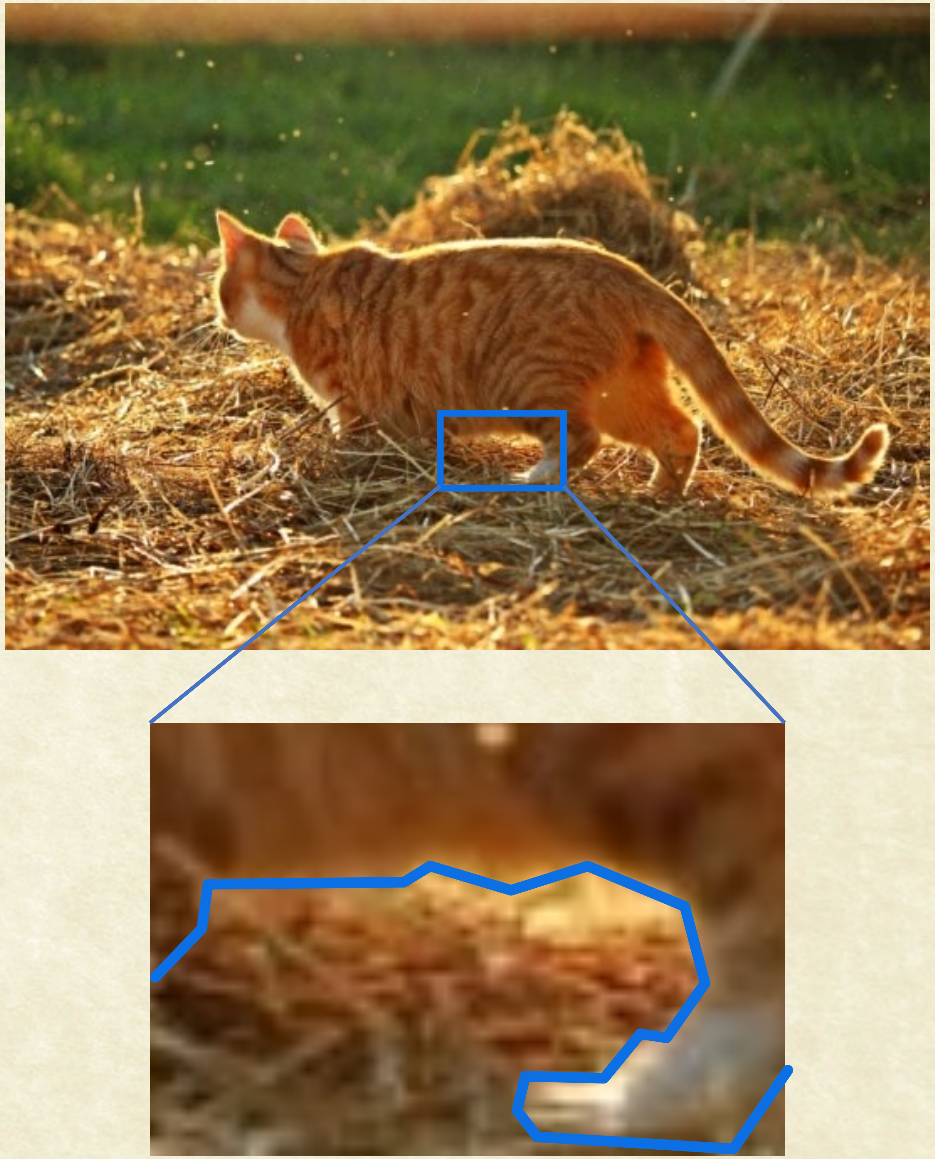
It is NOT deconvolution

- Very nice read: https://medium.com/@marsxiang/convolutions-transposed-and-deconvolution-6430c358a5b6

- Deconvolution is the process of reversing convolution effects
- Transposed convolution is learnable upsampling!

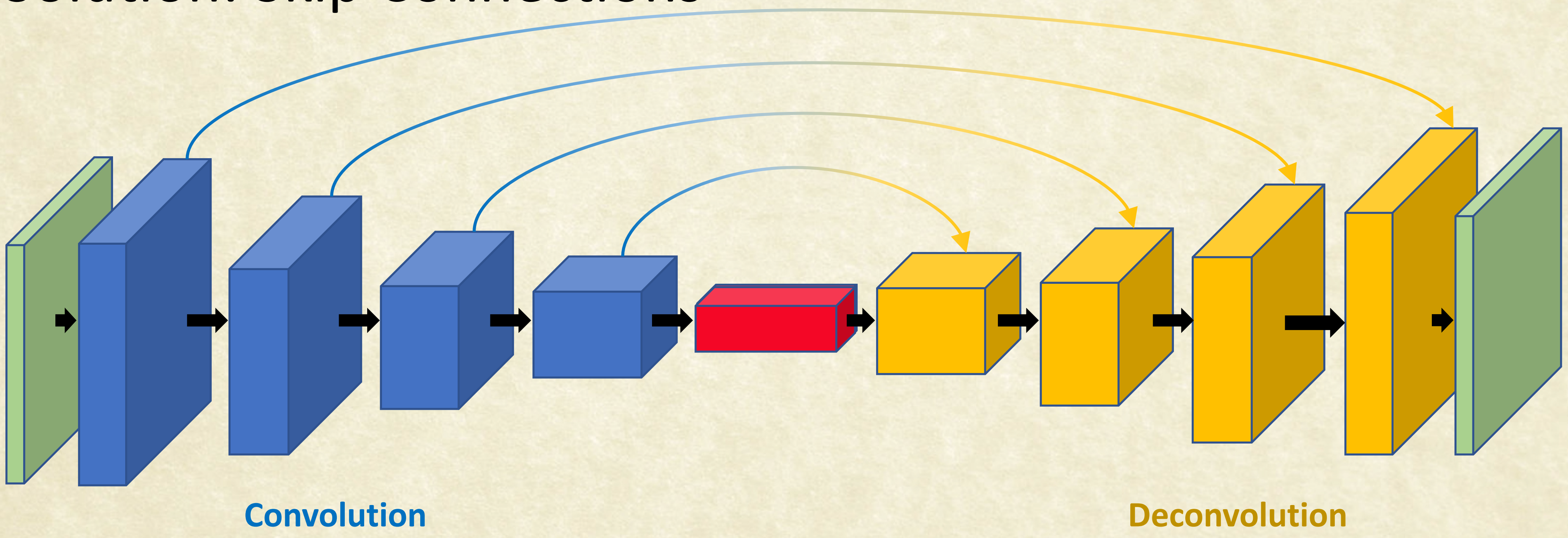
The Dilemma: Local or Global

- Focusing on Global information (context) is essential for robust classification (local invariance)
- Focusing on Local information is essential for localization or fine semantic boundaries (location sensitivity)
- Autoencoder output tends to focus on overall information due to bottleneck layer



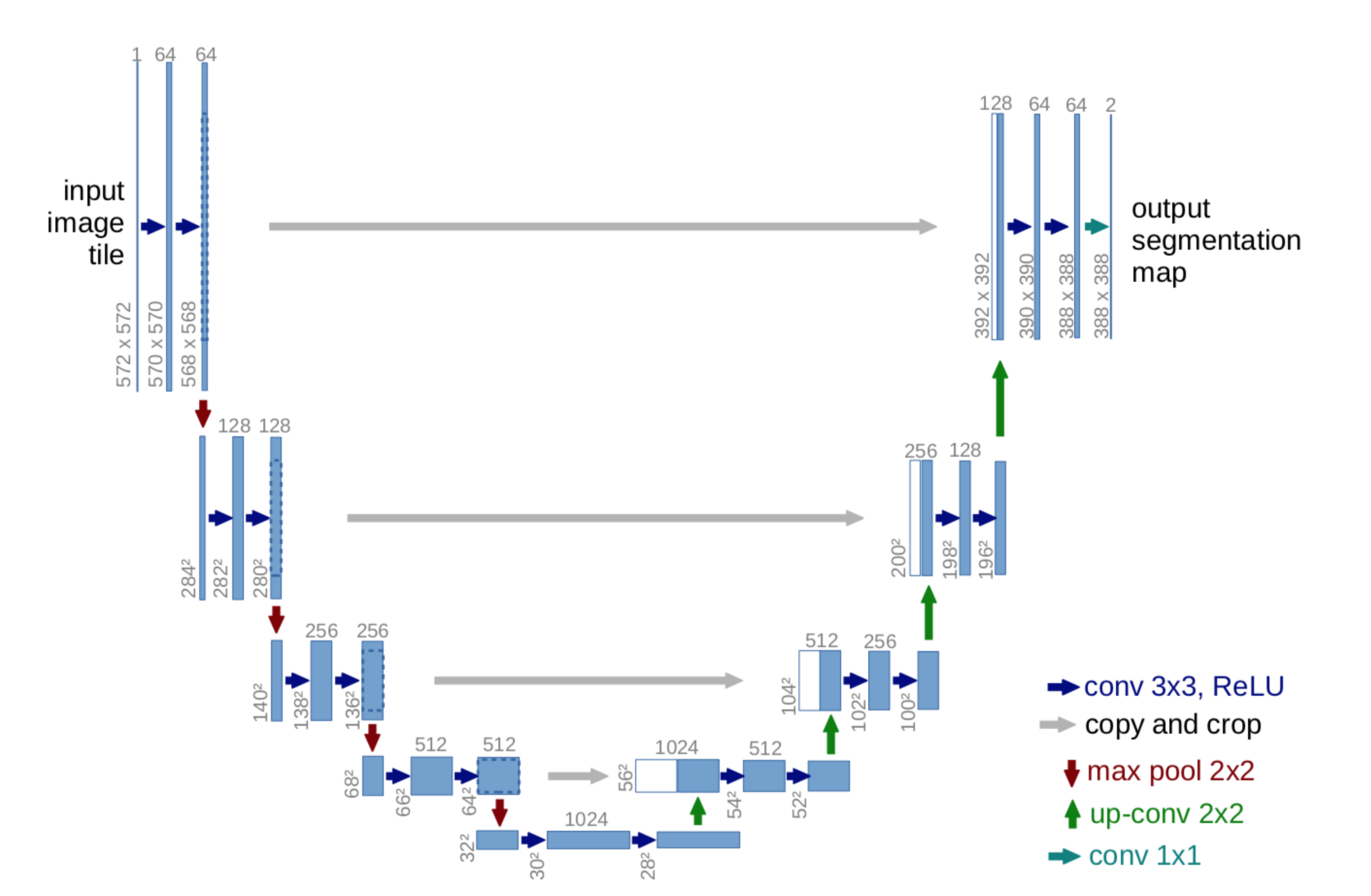
Improving Output Resolution

- The bottleneck layer is of low resolution
- Recovering detailed information is difficulty during deconvolution
- Solution: Skip Connections



Source:- https://arxiv.org/abs/1505.04597

UNet: Skip Connections



Atrous Convolutions