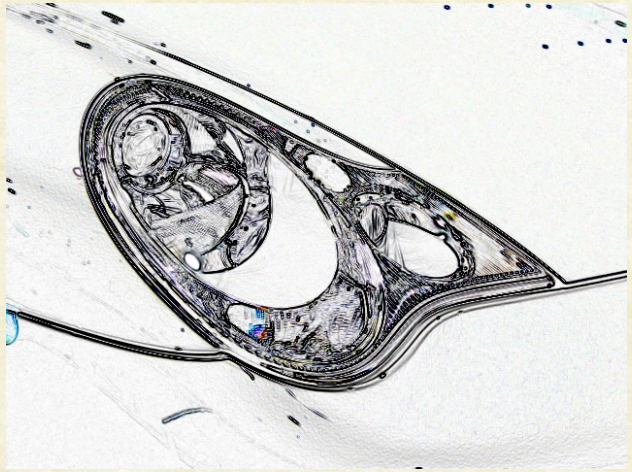
CS7.505: Computer Vision

Spring 2022: Semantic Segmentation with DL


Anoop M. Namboodiri

Biometrics and Secure ID Lab, CVIT,

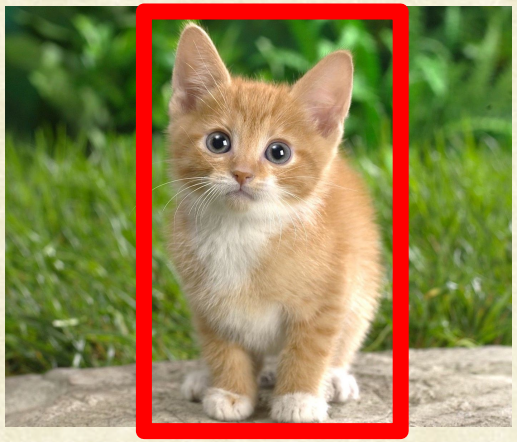
IIIT Hyderabad

Computer Vision Tasks

Classification

CAT

Classification + Localization

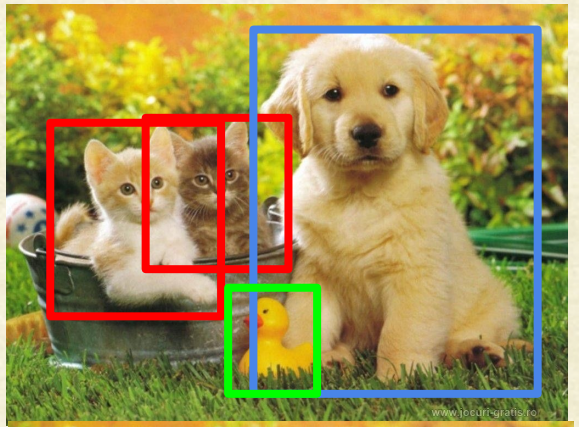


CAT

Single object

CAT, DOG, DUCK

Multiple objects



Object Detection

Semantic Segmentation

Instance Segmentation

Computer Vision Tasks

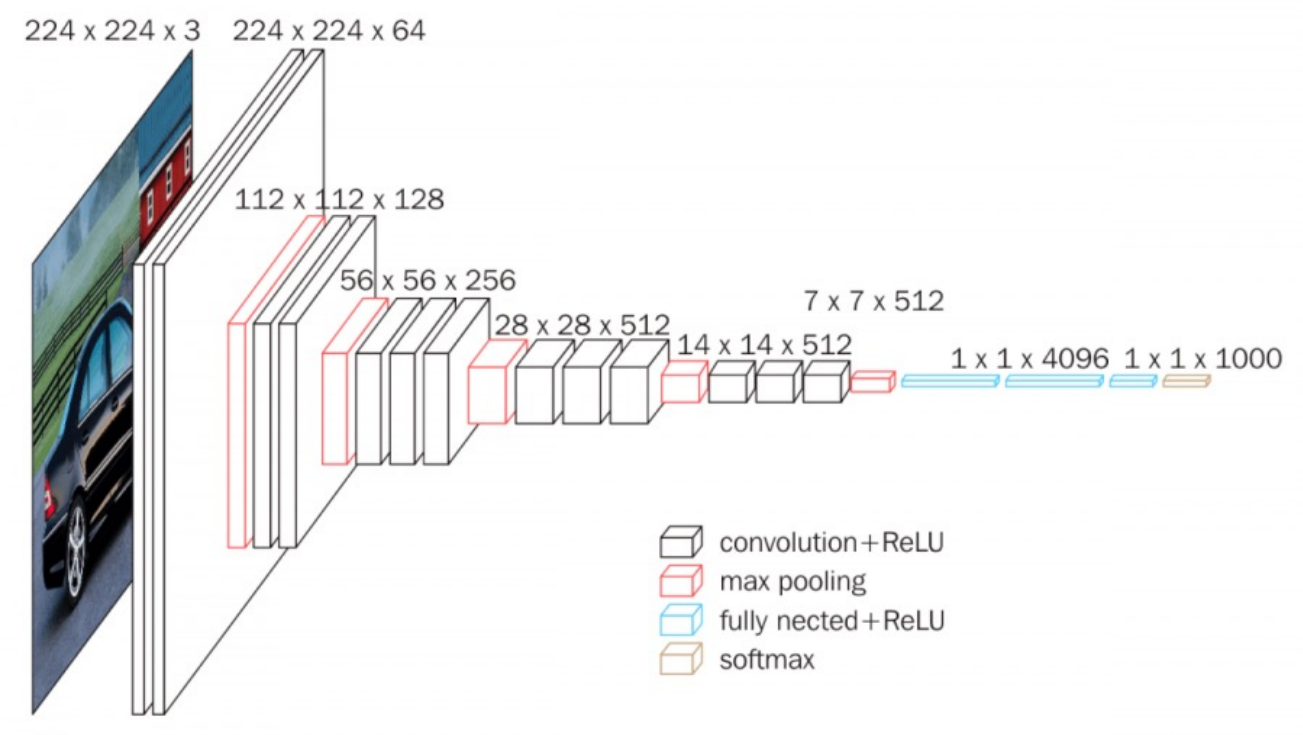
1. Classification
   - Classify (predict one label for) each image
2. Classification + Localization
   - Classify and predict one bounding box (regression) per image
3. Object Detection
   - Classify each window in an image
4. Semantic Segmentation
   - Classify each pixel of an image
5. Instance Segmentation
   - Object Detection + Semantic Segmentation

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

DL Solution: Speed

VGG16: Number of Parameters by Layers

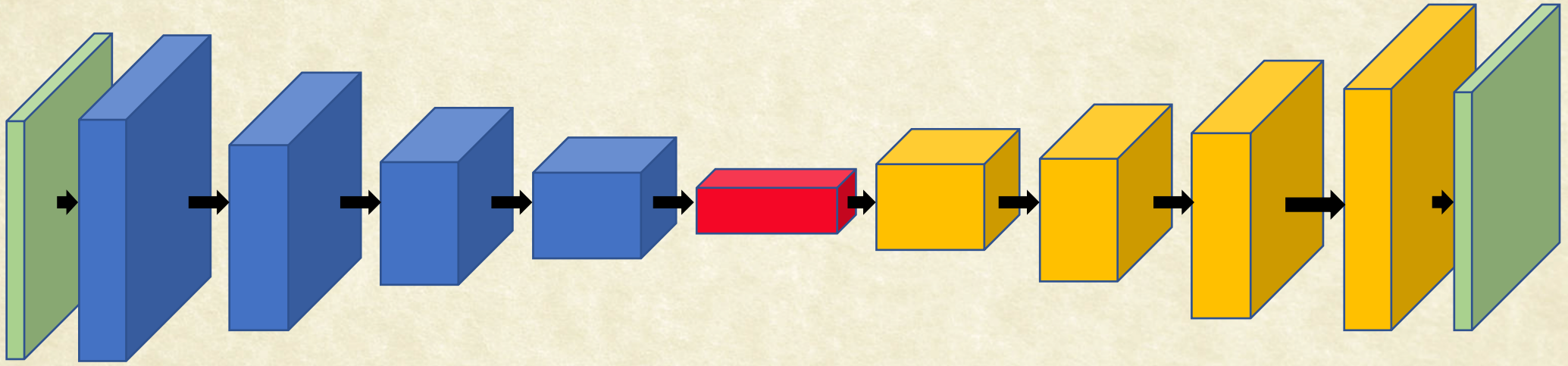


- Max-Pool layers are not shown in the table as they do not contain any learnable parameters
- Almost 90% of parameters are in the 3 FC layers !!

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

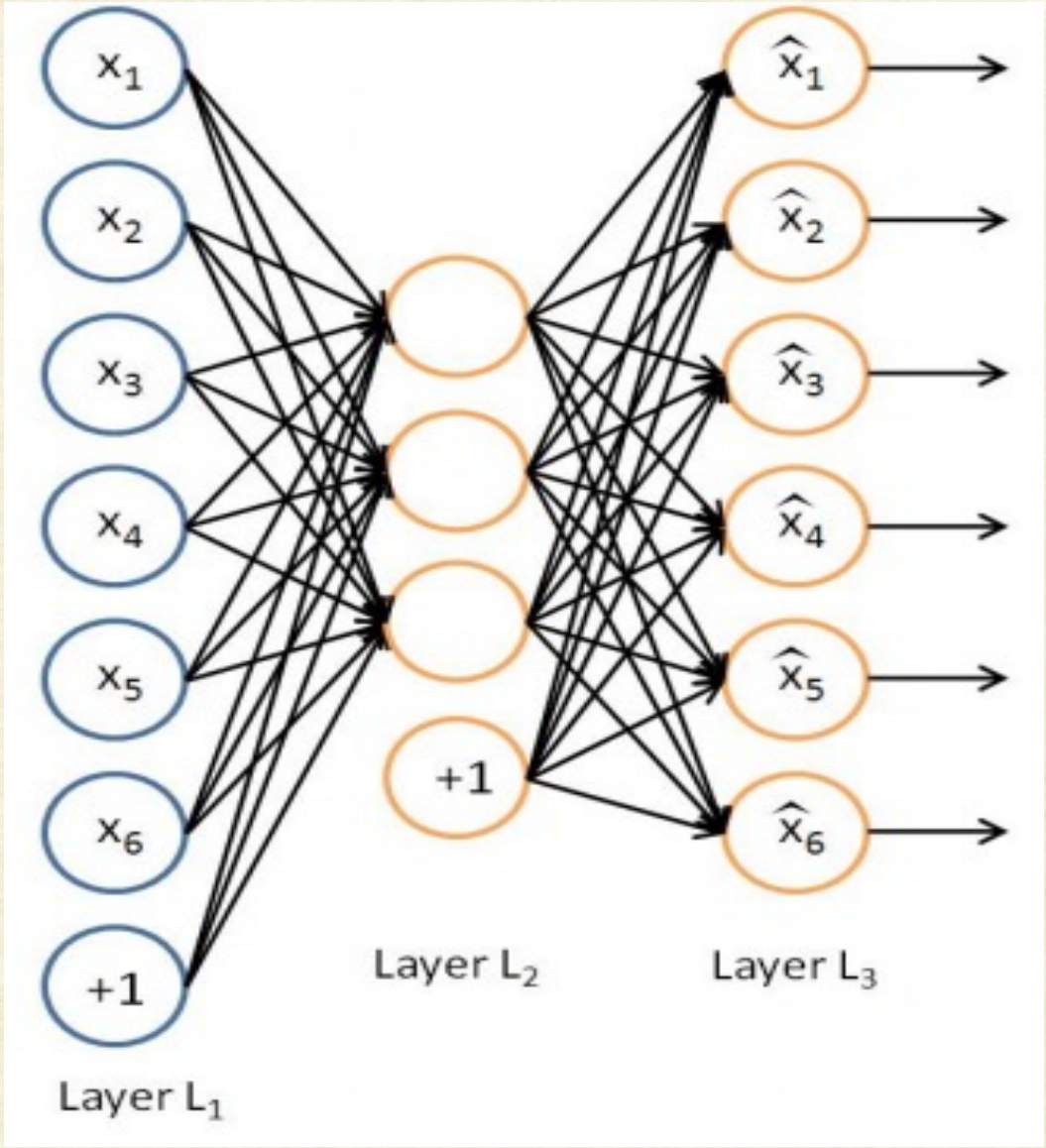
Deep Learning Solution: Fully Convolutional Network

- 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

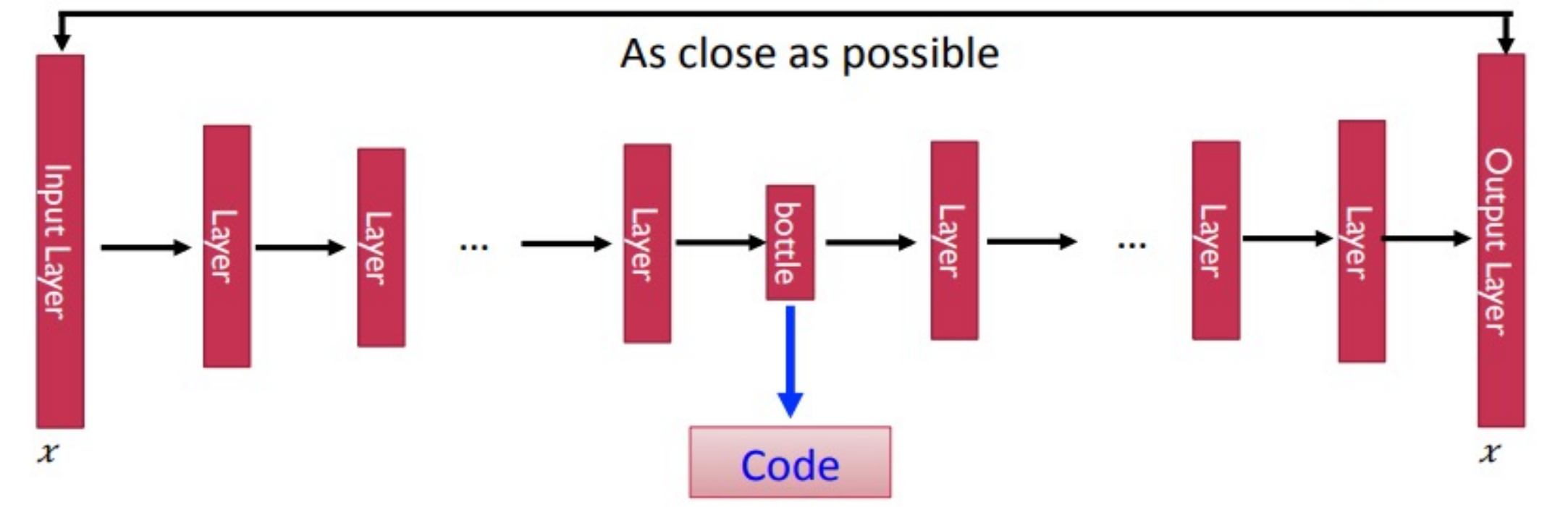
- How to train a network with only unlabeled data?
- Idea: Use the input itself as output.
- Similar to MLP
- Network learns to reconstruct
- “Bottleneck” layer learns a compact representation.



Deep Auto-encoder

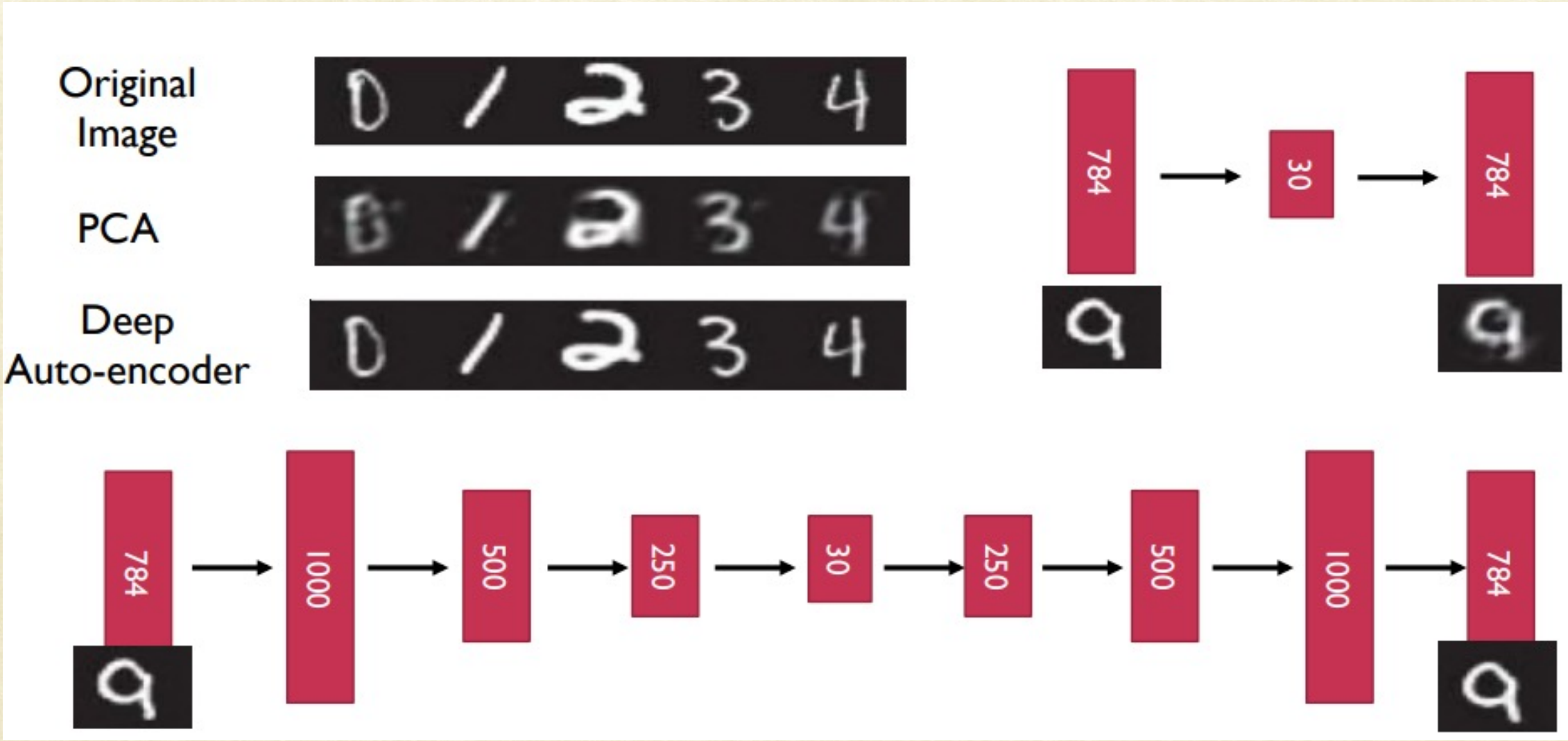
- Of course, the auto-encoder can be deep

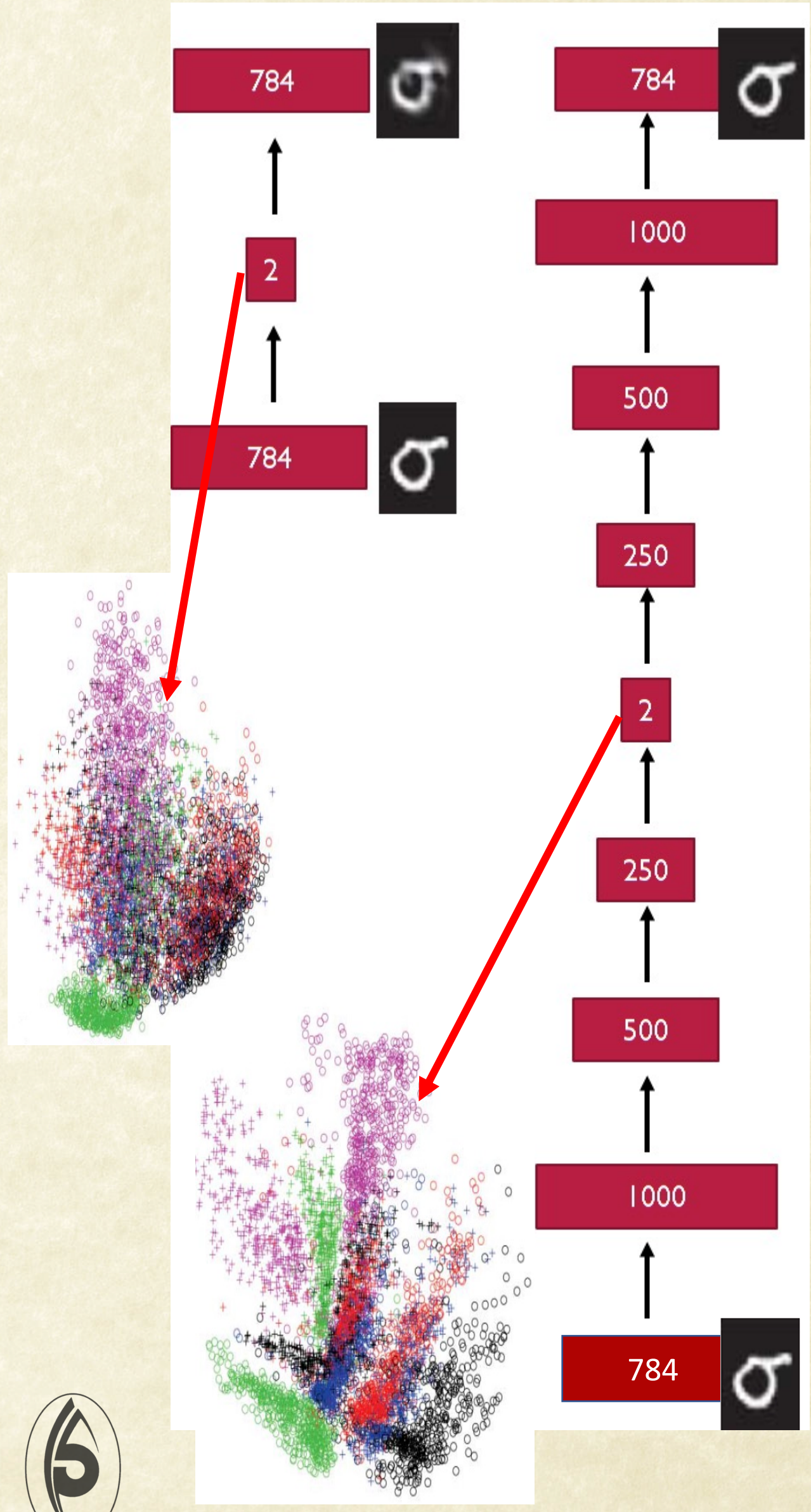
Symmetry is not necessary



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

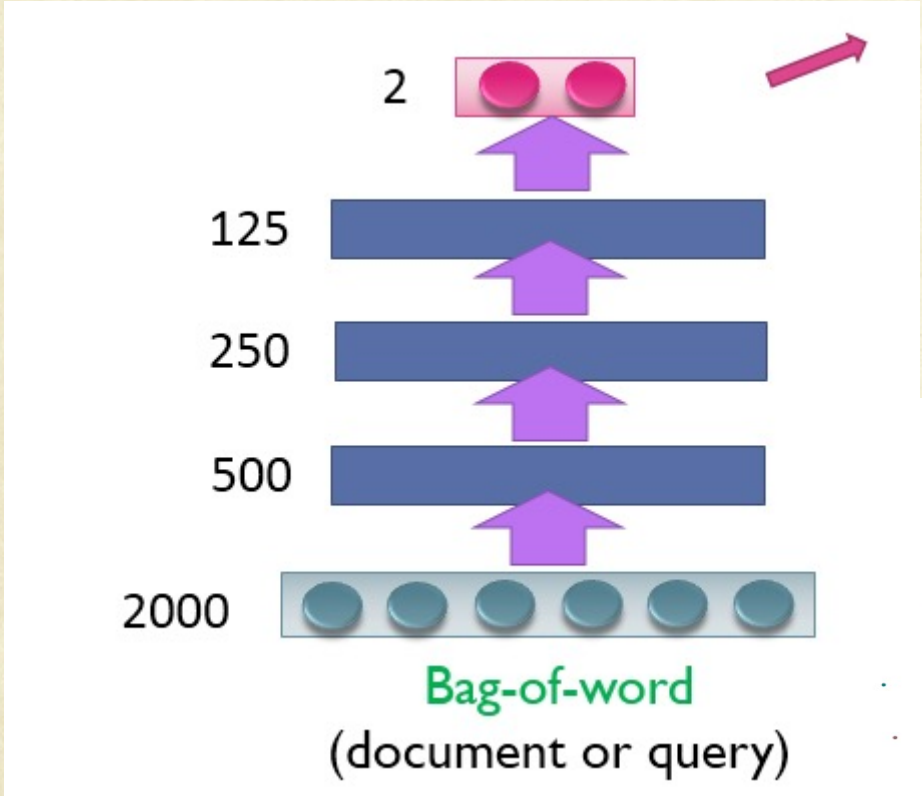
Deep Auto-encoder

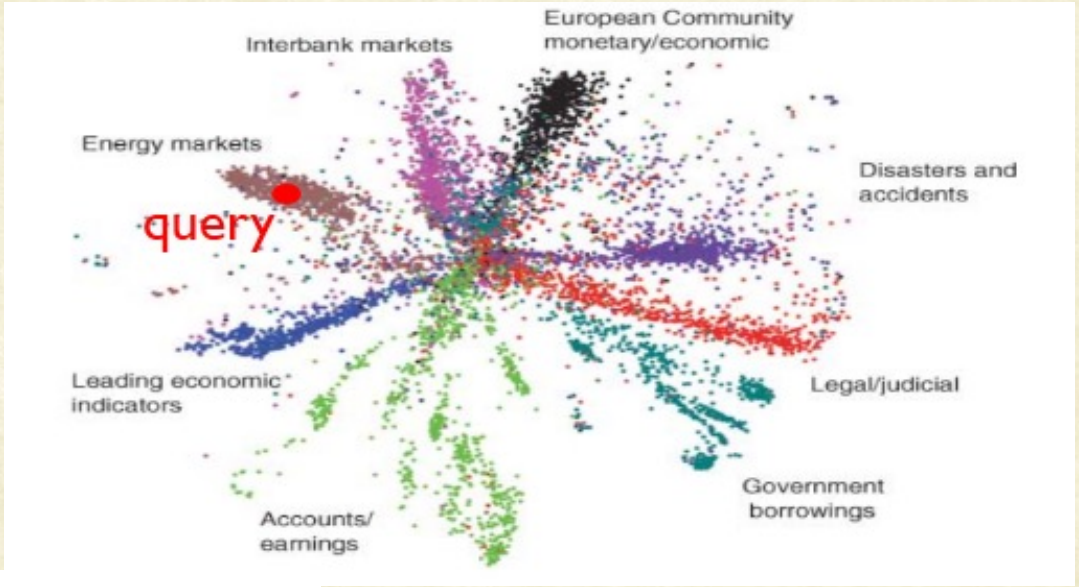


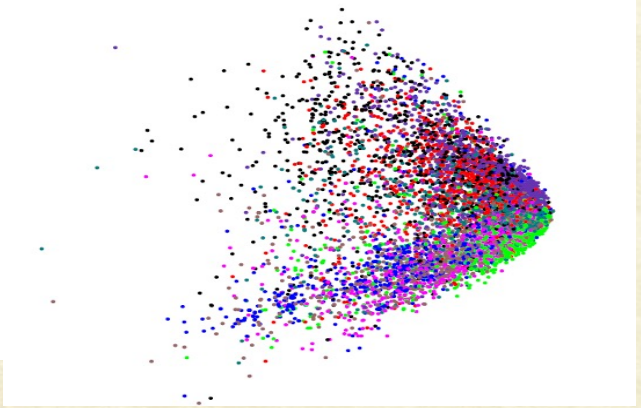

784
2
784
784
1000
500
250
2
250
500
1000
784

Auto-encoder – Text Retrieval

The documents talking about the same thing will have close code.

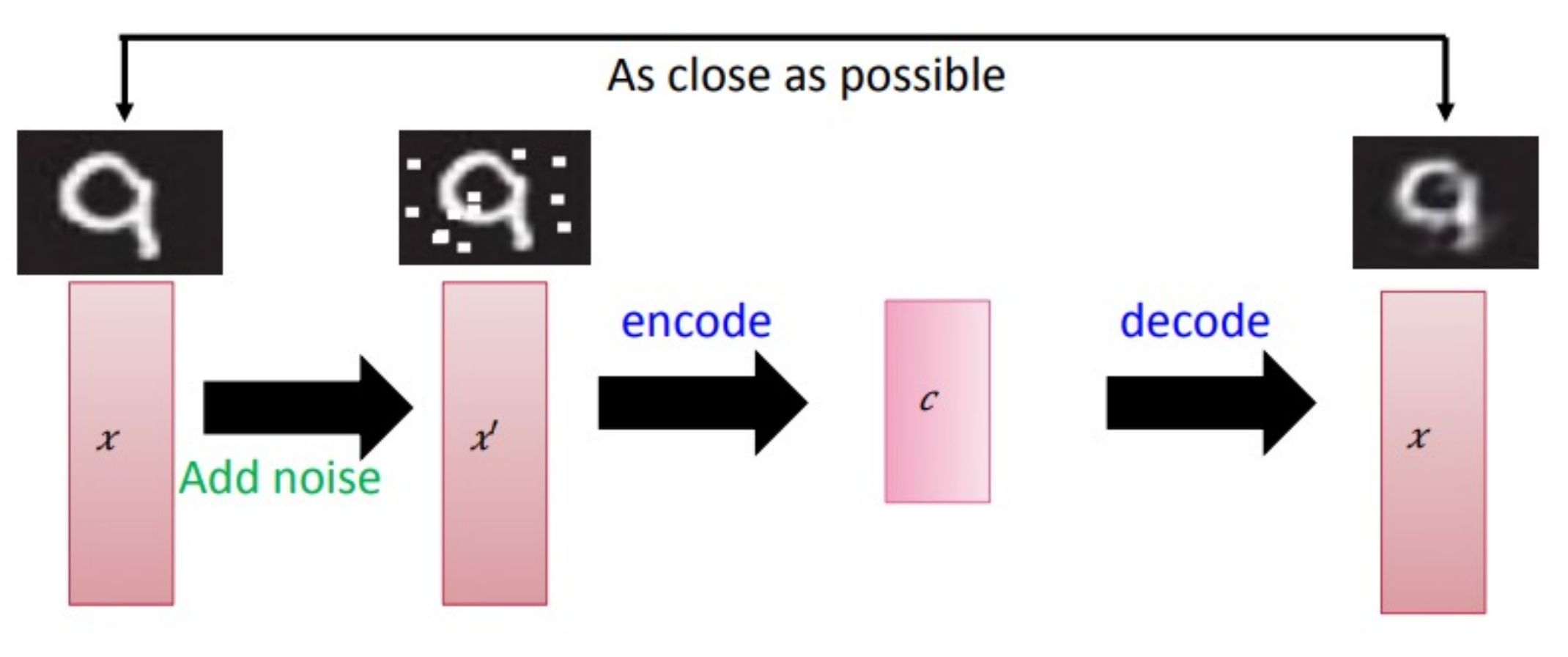






LSA: project documents to 2 latent topics

Denoising Auto-encoder



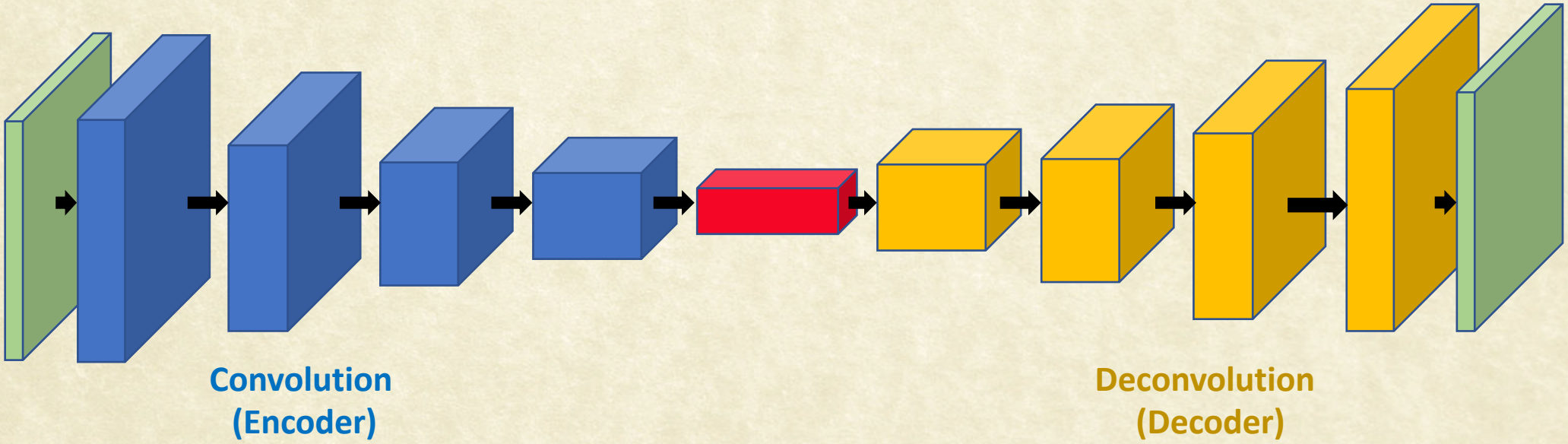
Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." *ICML,* 2008.

Autoencoder: Comments

- Nonlinear PCA (Dimensionality Reduction)
- Data compression
- Unsupervised Learning (Self-Supervised)
  - “We expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object.” - LeCun, Bengio, Hinton, Nature 2015
  - As I've said in previous statements: most of human and animal learning is unsupervised learning. If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake. We know how to make the icing and the cherry, but we don't know how to make the cake. - Yann LeCun, March 14, 2016 (Facebook)

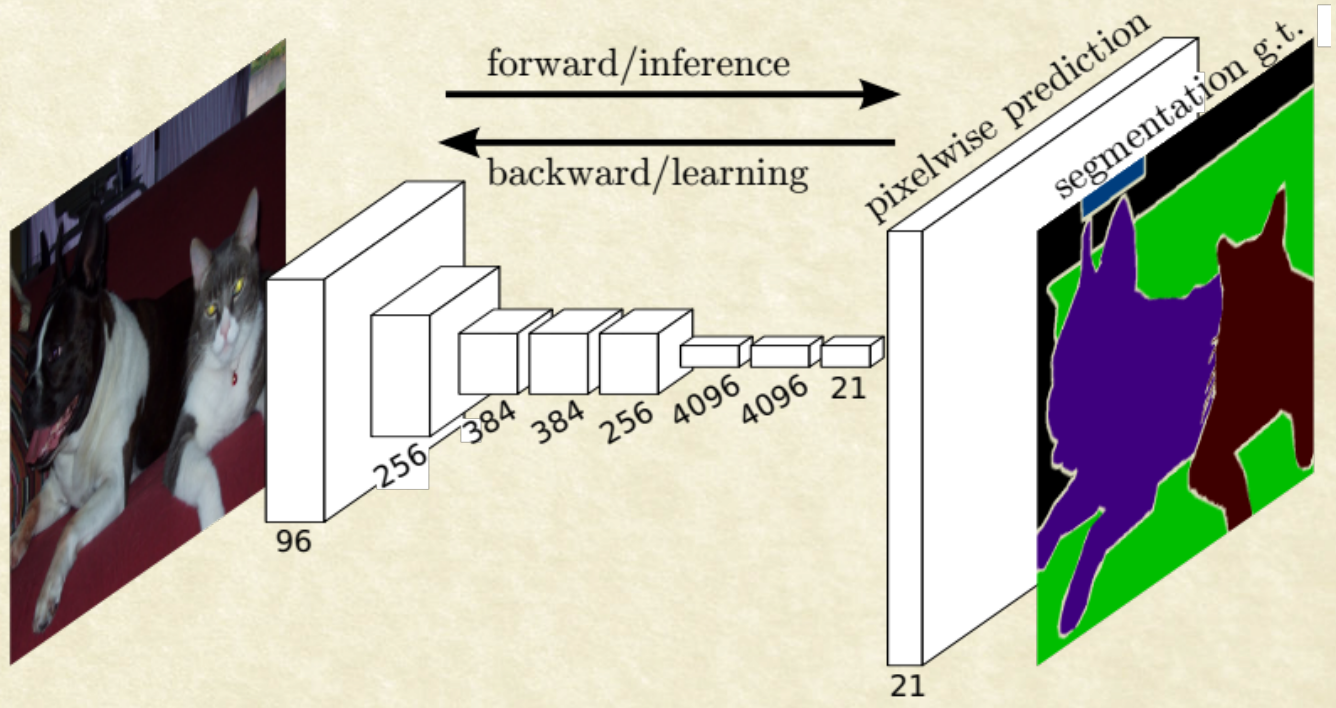
Fully Convolutional Autoencoders

- 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 de-convolution



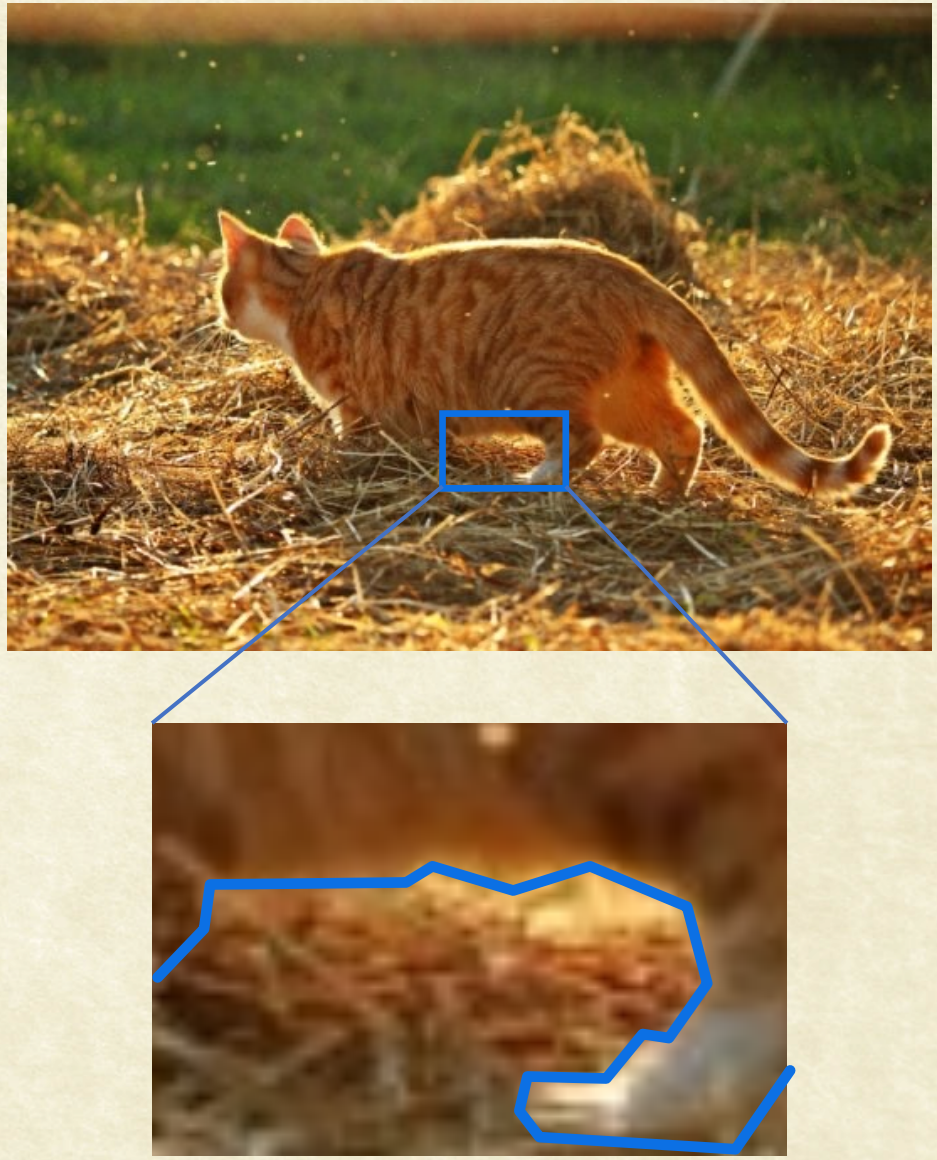
Fully Convolutional Network for Segmentation

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



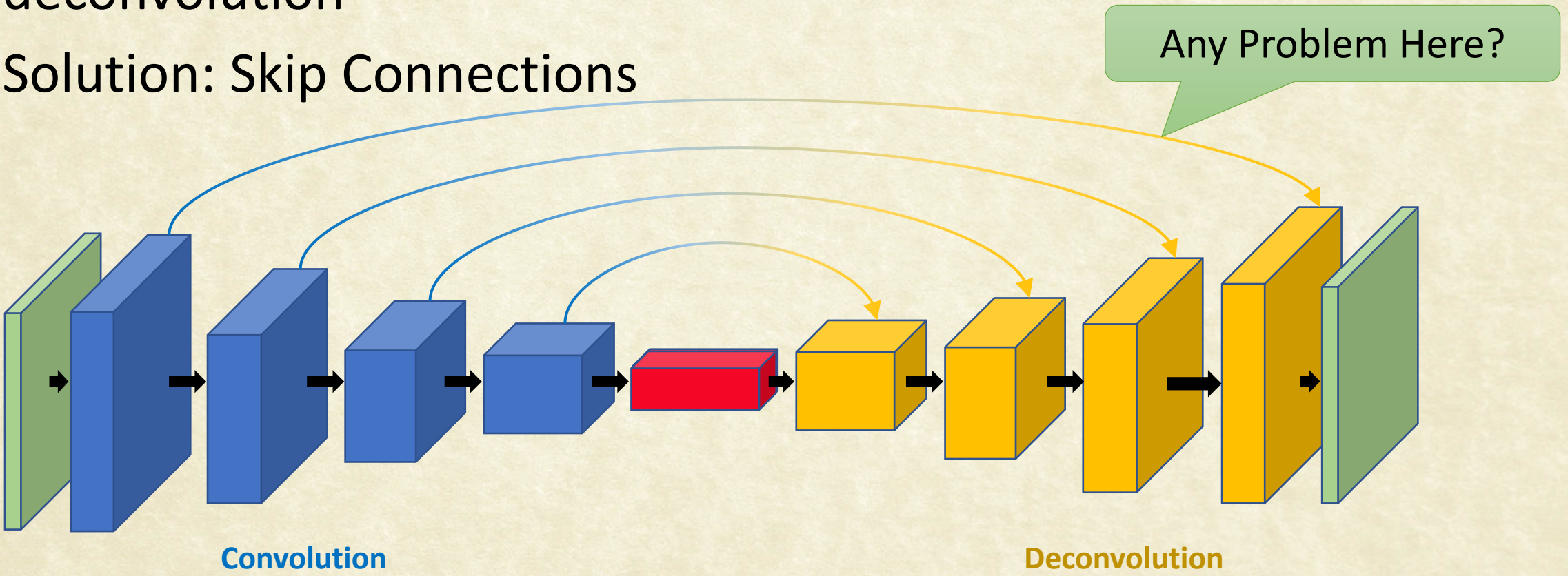
The Dilemma: Local or Global

- 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



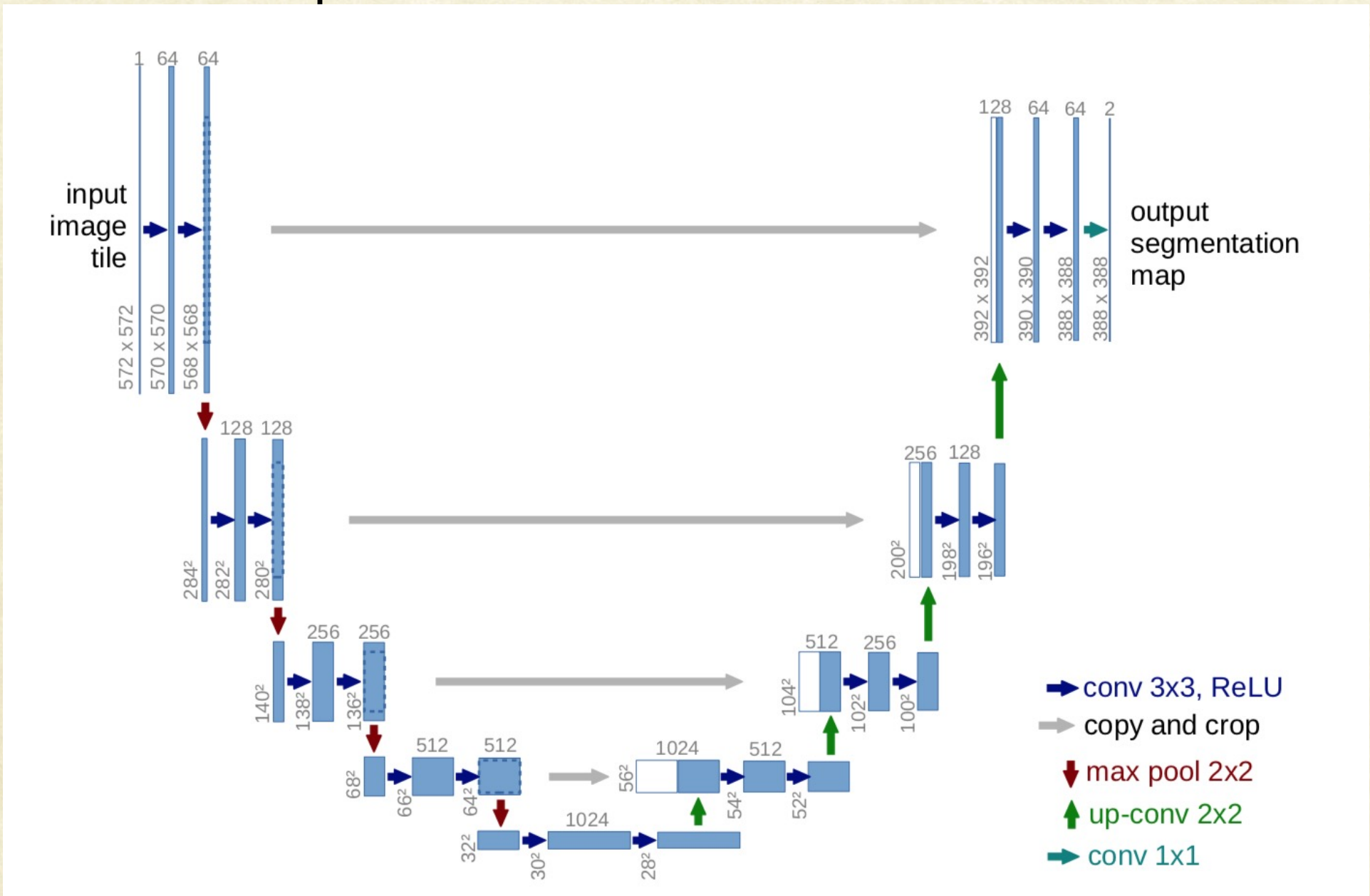
Improving Output Resolution

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



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

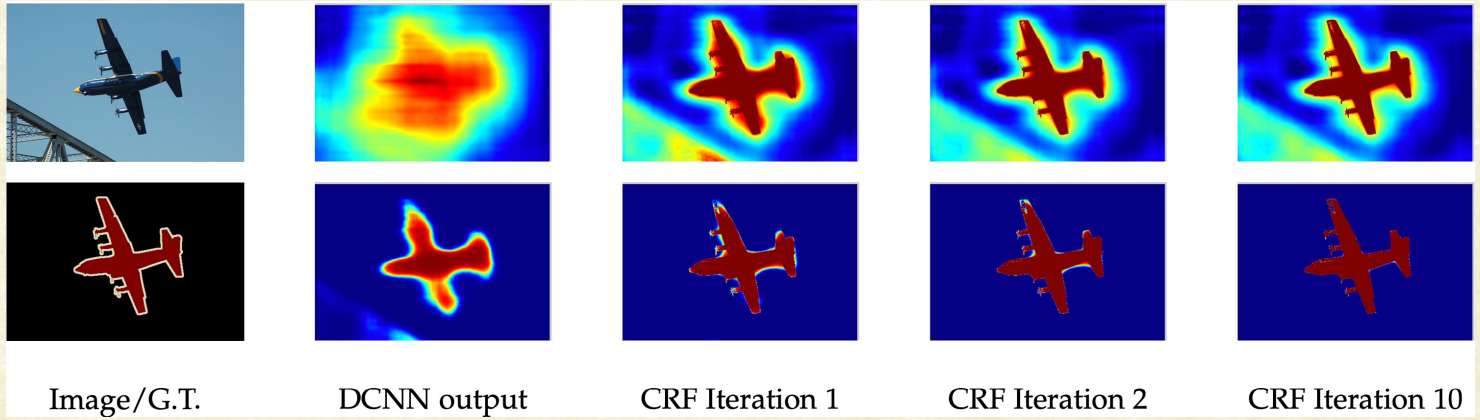
UNet: Skip Connections

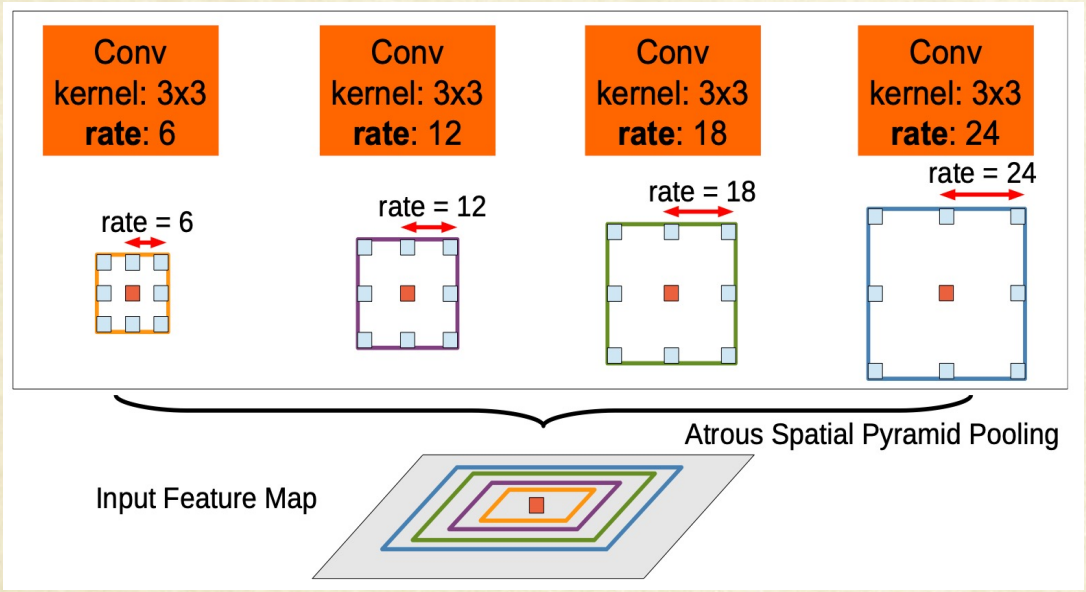


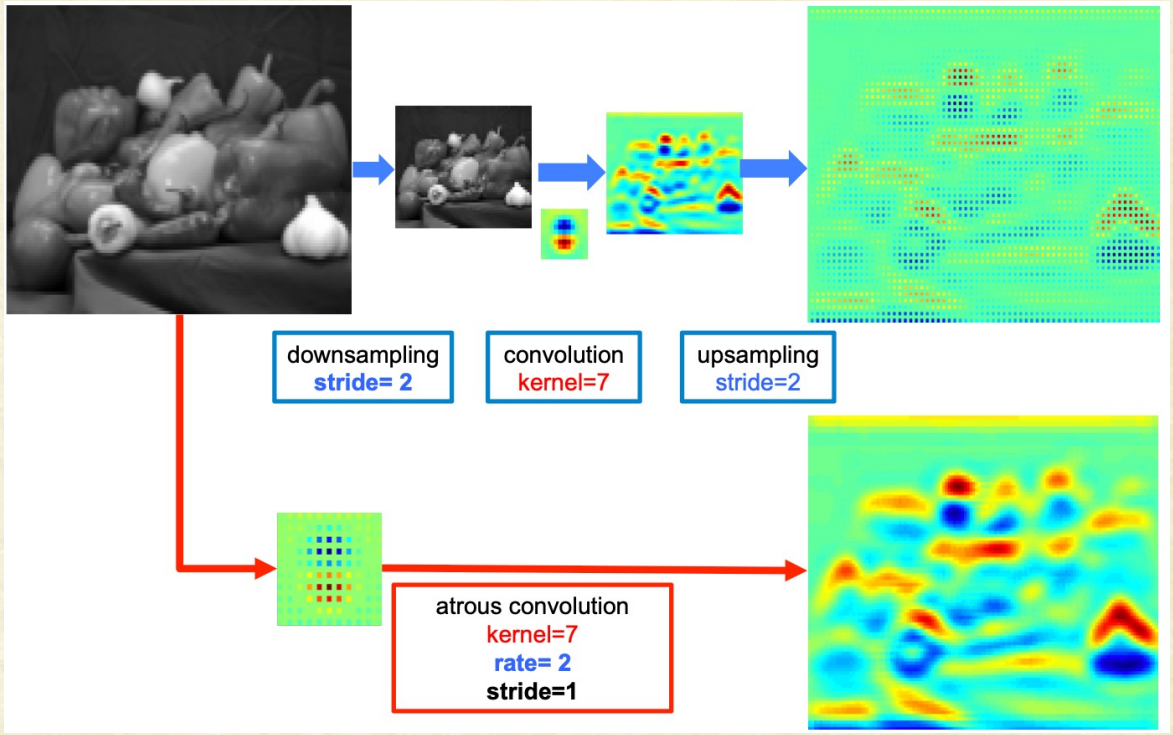
DeepLab

Key Ideas:

- Atrous Convolution
- Spatial Pyramid Pooling
- Fully Connected CRF

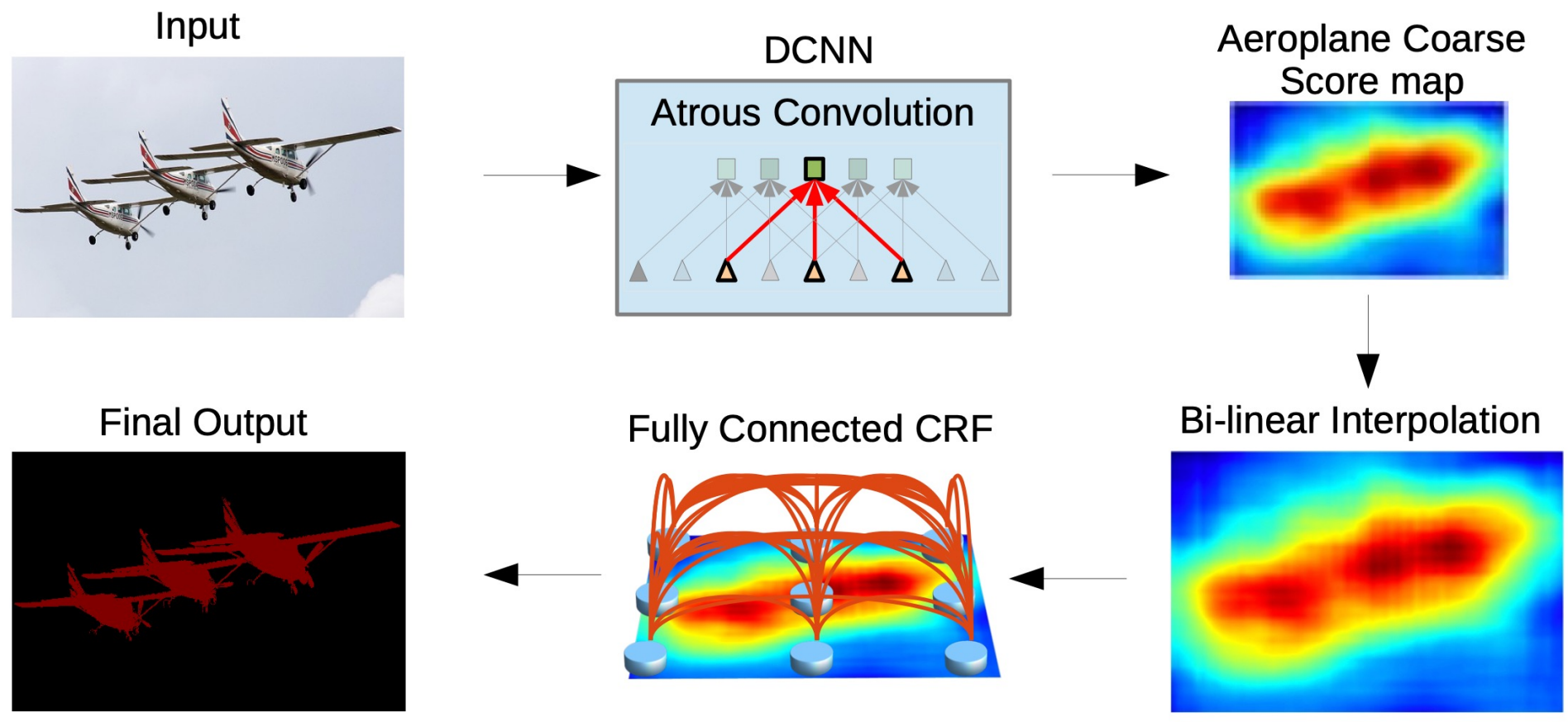






Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”, CVPR 2016

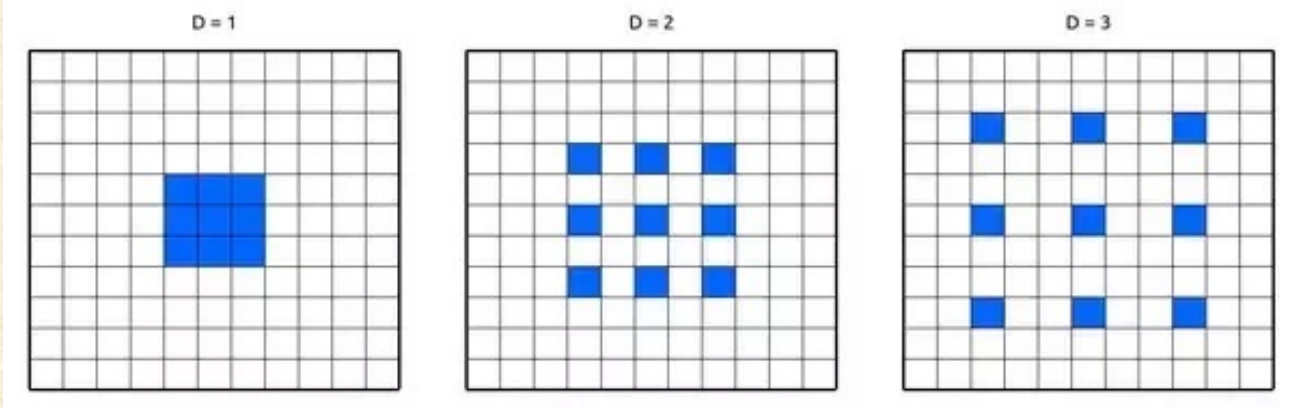
DeepLab: Pipeline



- Atrous convolution: Reduces downsampling from 32x to 8x
- Bilinear Interpolation: Restores to original resolution
- Fully Connected CRF: Refine segmentation boundaries

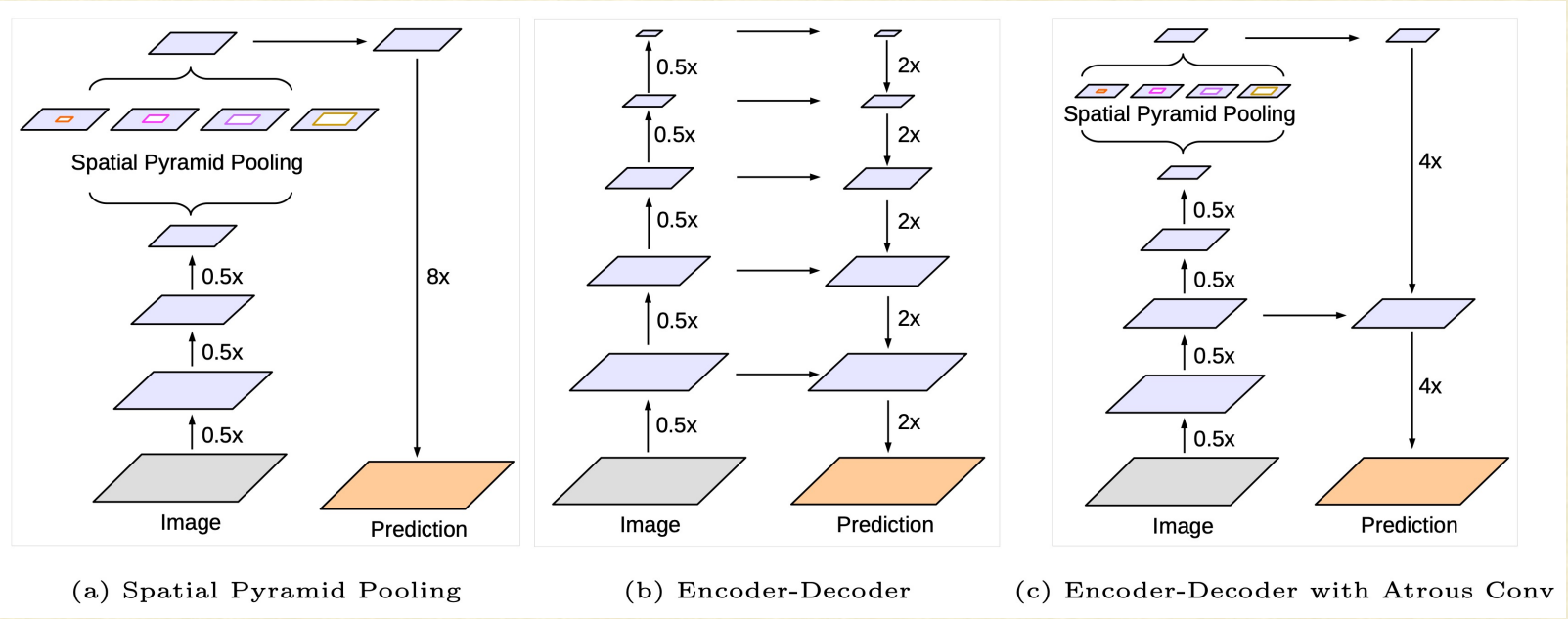
DeepLab V3: DeepLab + Depthwise Sep. Conv.

Atrous Convolutions



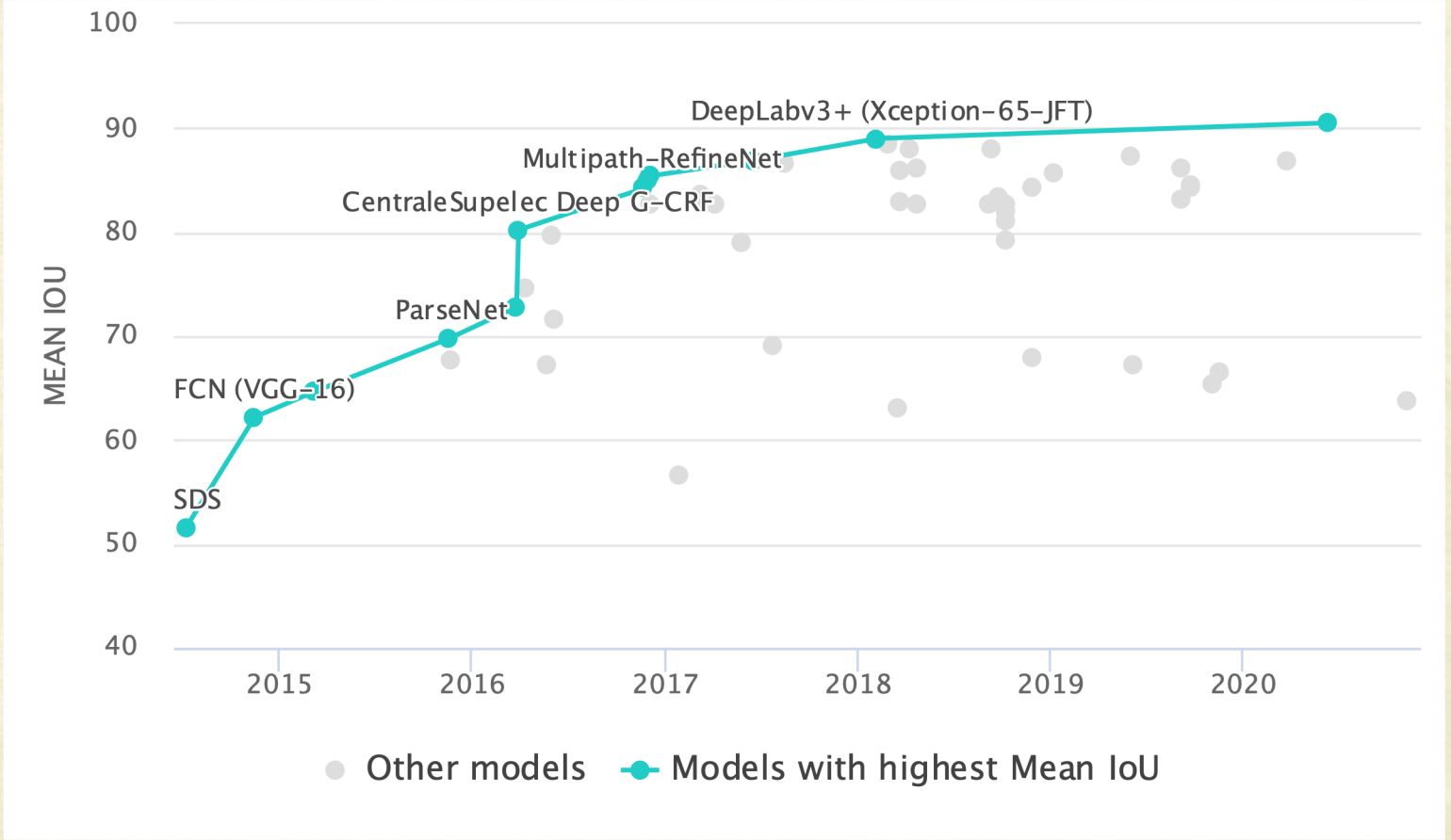
Spatial Pyramids + Enc-Dec + Atrous Conv.

Depth-wise Separable Conv. (Xception Model)



Evolution of Segmentation

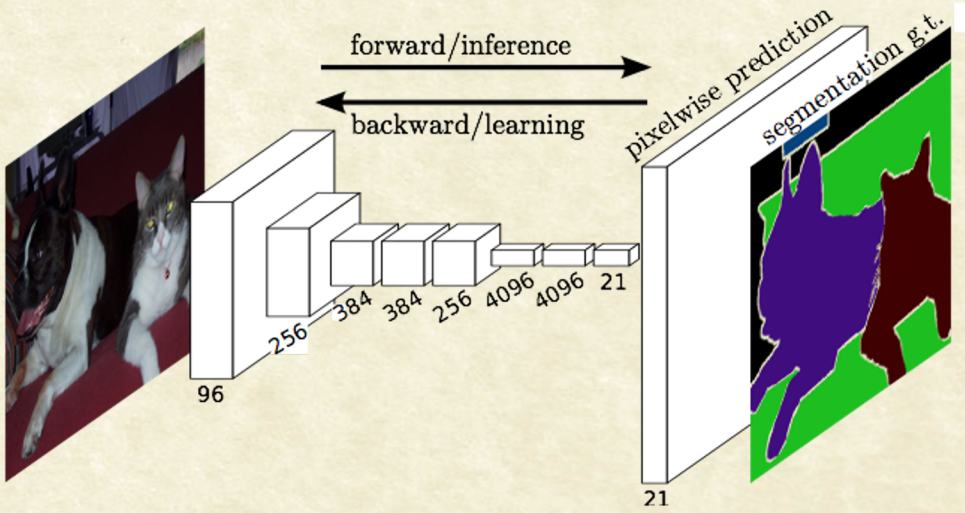
- GCN: Larger Kernels, Boundary Refinement
- CRF for regularization
- Better Training
  - Self Training
- Use of LSTM for video segmentation

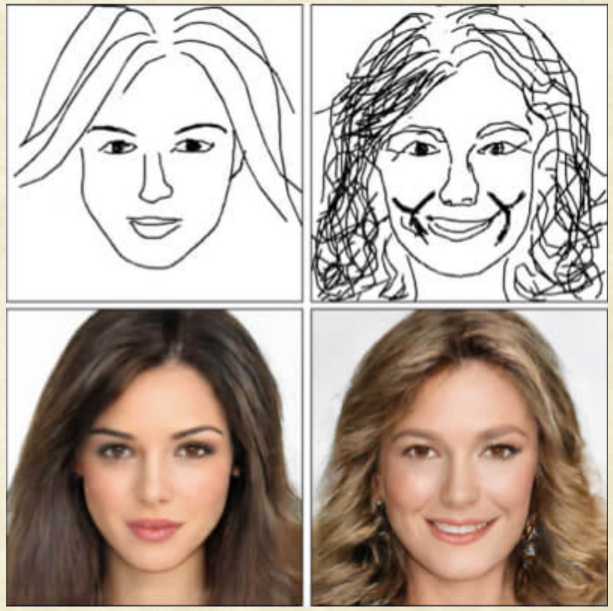


Source: https://paperswithcode.com/sota/semantic-segmentation-on-pascal-voc-2012

Applications of Image-to-Image Networks

- Segmentation
- Background Removal; Portrait Mode
- Image Generation from Sketches
- Monocular Depth Estimation
  - Autonomous Navigation
  - Map Generation
- Image Coloring





Questions?