A dip into the deep

Arnav Bhavsar

School of Computing and Electrical Engg.

IIT Mandi

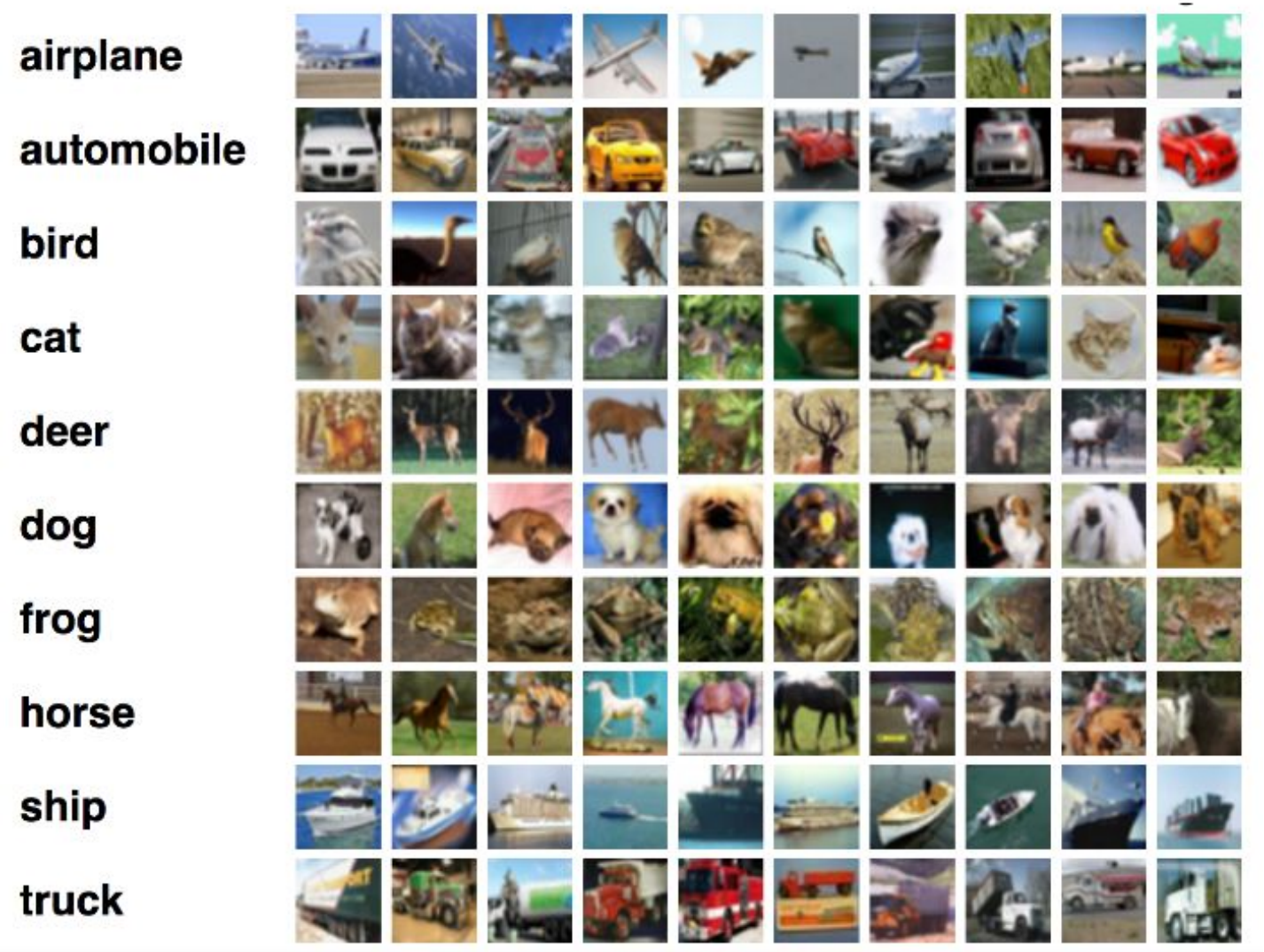
- Richness of (visual) information
- Hierarchical nature of information
- Variety of tasks
- (Sometimes) Sequencing / Decomposition of tasks
- …

The evolution

- Decentralization
- Evolutionary explosion
- Many unions and children
- Collective intelligence (Human and machine)

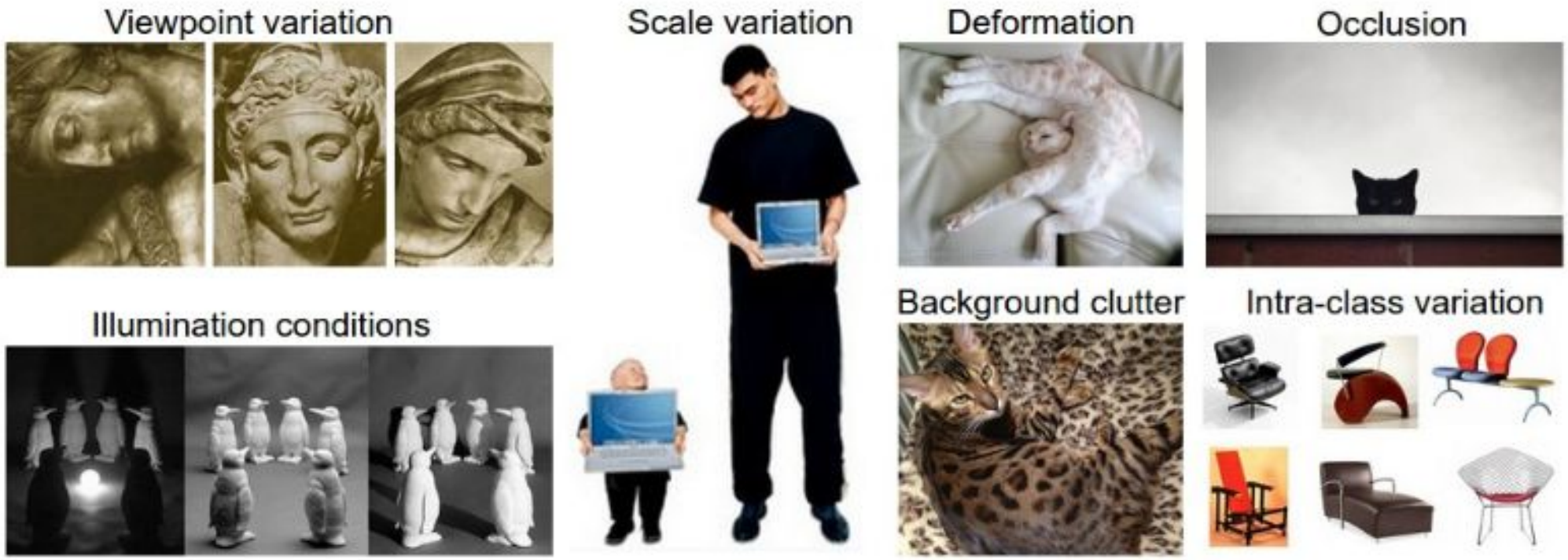
A glimpse of the variety of information and tasks

Classification



https://medium.com/@tifa2up/image-classification-using-deep-neural-networks-a-beginner-friendly-approach-using-tensorflow-94b0a090ccd4.

Classification



Source: http://cs231n.github.io/classification/

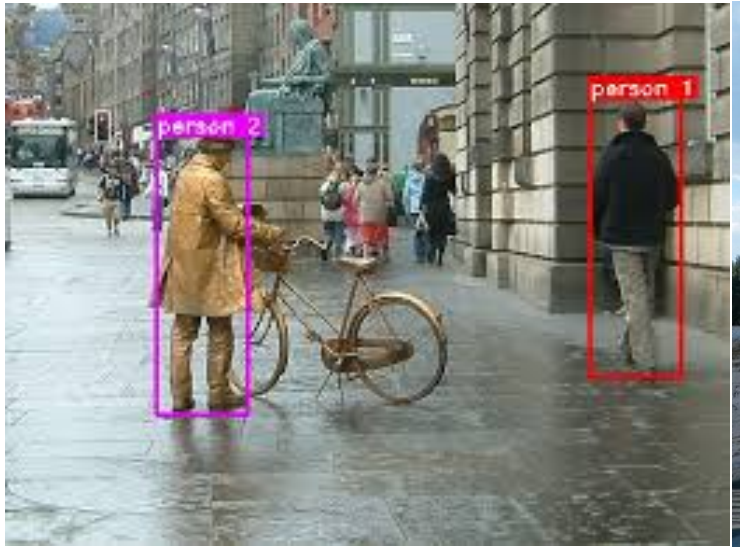
Annotation

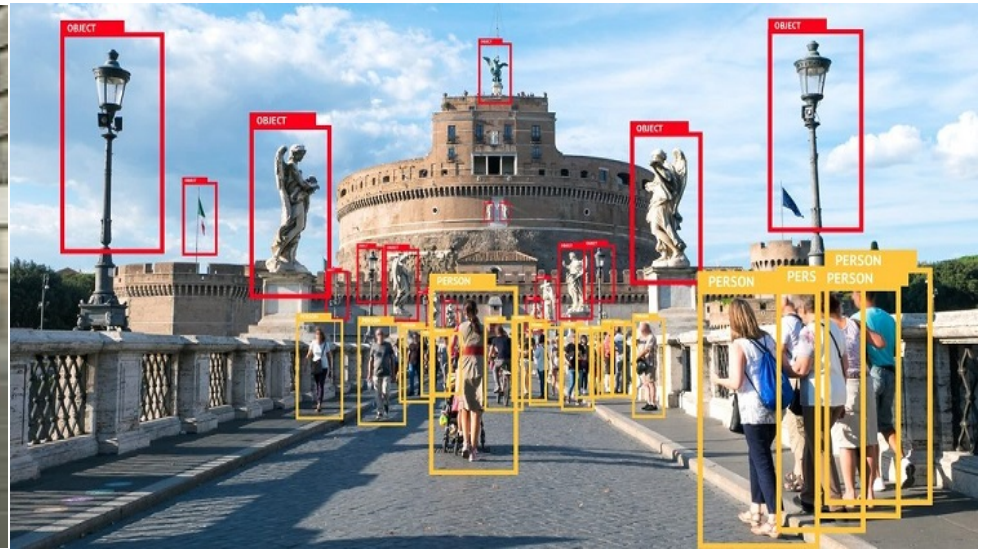
					
Predicted keywords	sky, jet, plane, smoke, formation	grass, rocks, sand, valley, canyon	sun, water, sea, waves, birds	water, tree, grass, deer, white-tailed	bear, snow, wood, deer, white-tailed
Human annotation	sky, jet, plane, smoke	rocks,sand, valley, canyon	sun, water, clouds, birds	tree, forest, deer, white-tailed	tree, snow, wood, fox

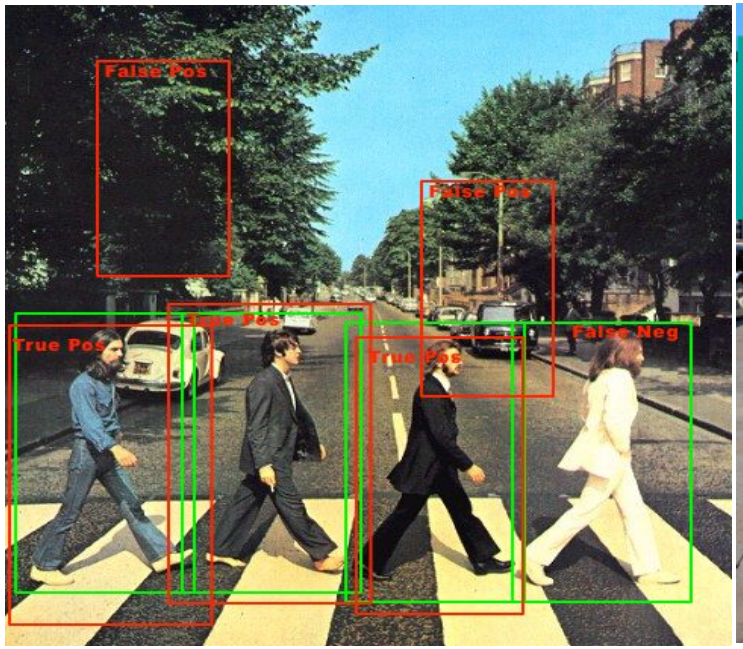
Prediction for queries: Sky (Row1), Street (Row2), Mare (Row3) and Train (Row4)

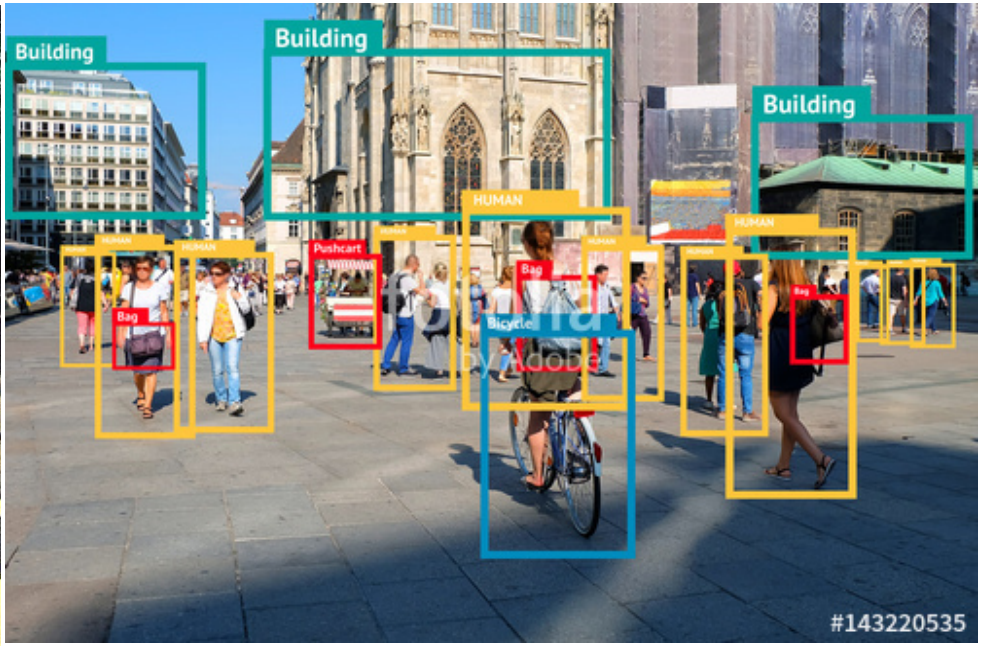
Ref: A. Makadia, V. Pavlovic, and S. Kumar. A New Baseline for Image Annotation. In Proceedings of ECCV 08.

Detection



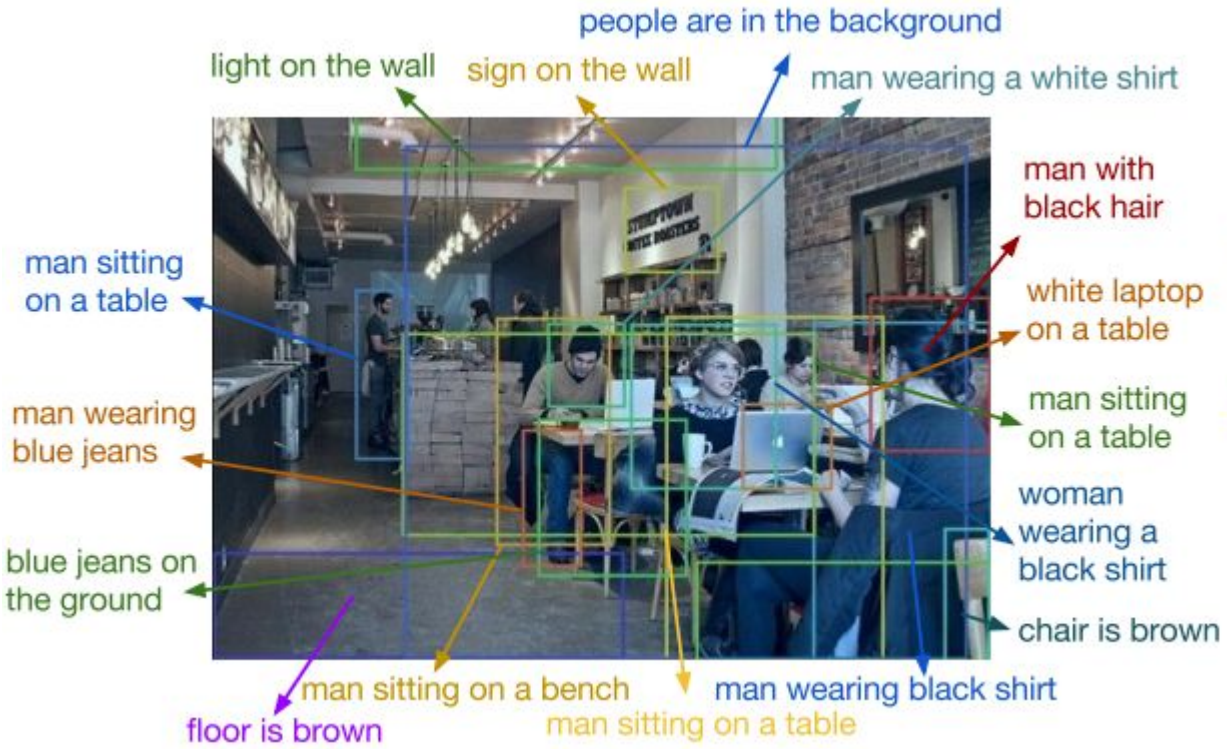


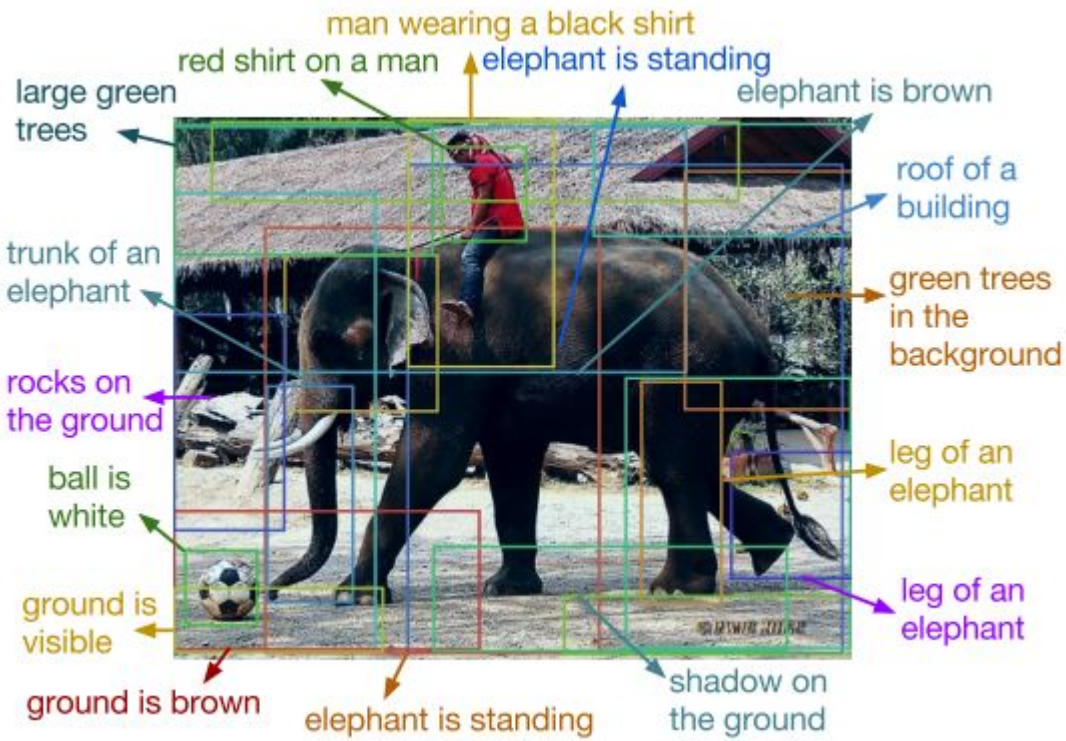




https://software.intel.com/en-us/articles/a-closer-look-at-object-detection-recognition-and-tracking, https://stackoverflow.com , https://www.alamy.com

Description



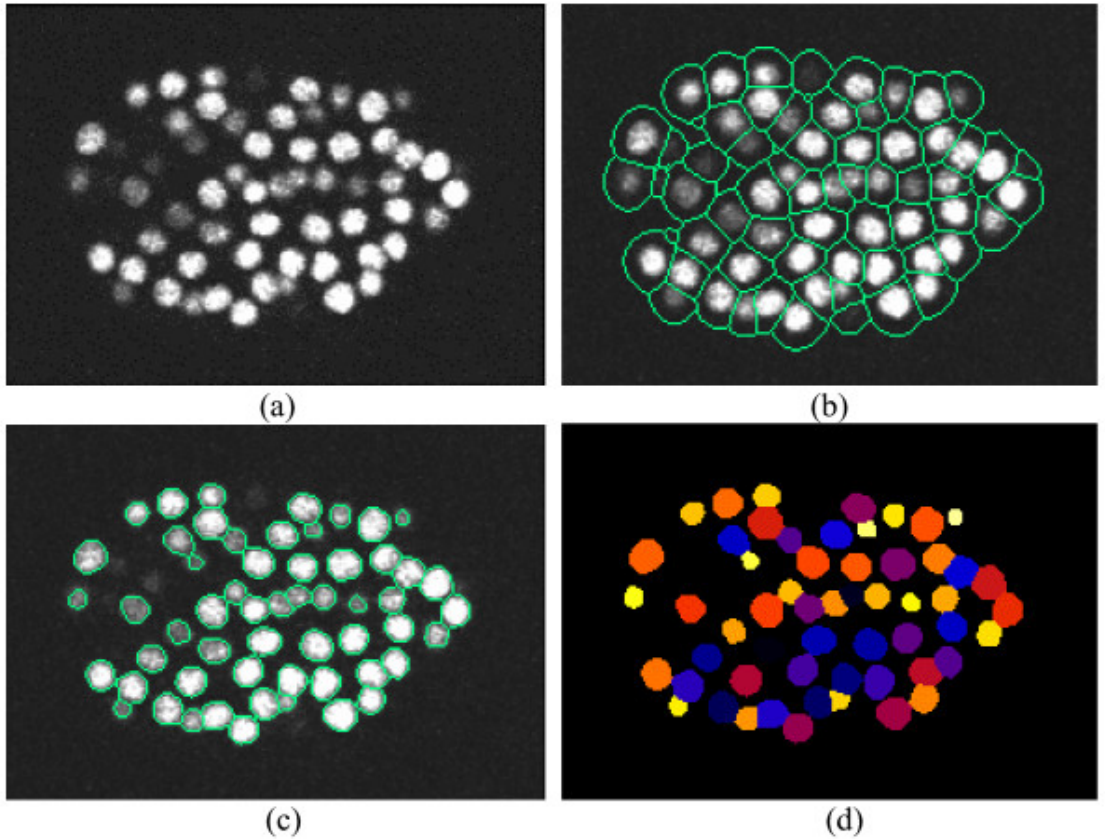


DenseCap: Fully Convolutional Localization Networks for Dense Captioning, Justin Johnson*, Andrej Karpathy*, Li Fei-Fei, (* equal contribution) in Proceedings of CVPR 2016.

Segmentation

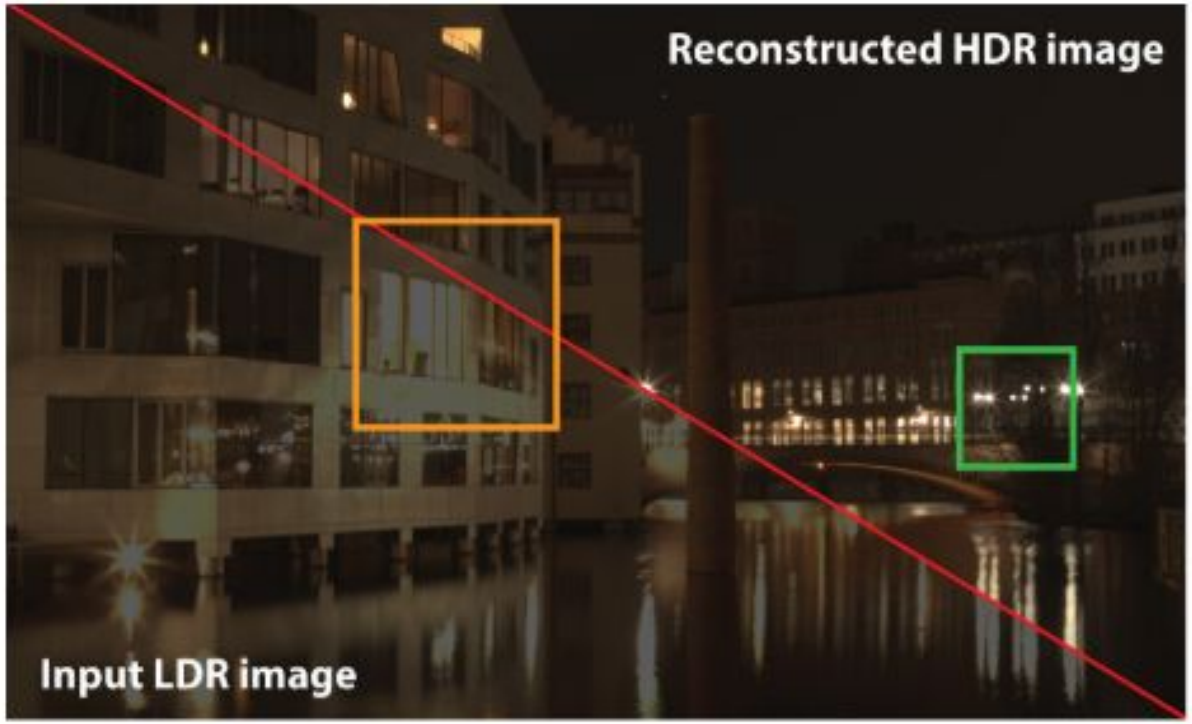
Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." PAMI, 2017.

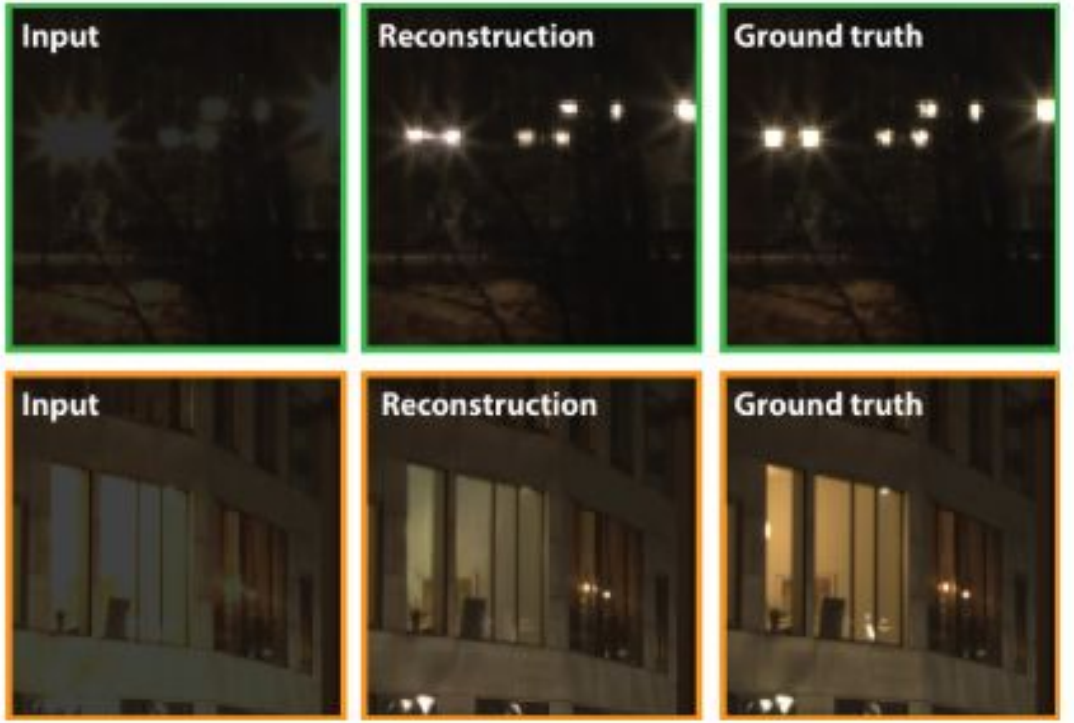
Segmentation



Li, Gang & Liu, Tianming & Tarokh, Ashley & Nie, Jingxin & Li, Kaiming & Mara, Andrew & Holley, Scott & Wong, Stephen. (2007). 3D cell nuclei segmentation based on gradient flow tracking. BMC cell biology. 8. 40. 10.1186/1471-2121-8-40.

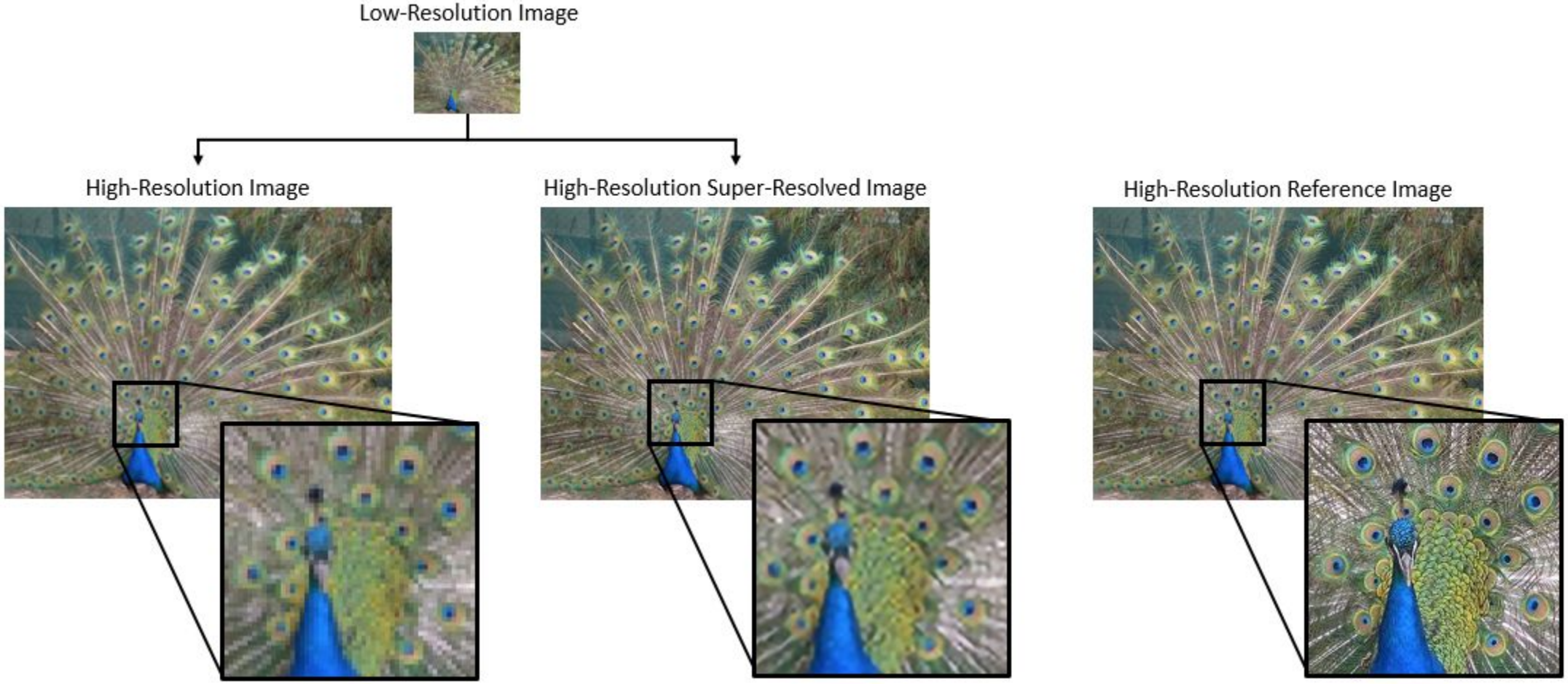
Reconstruction





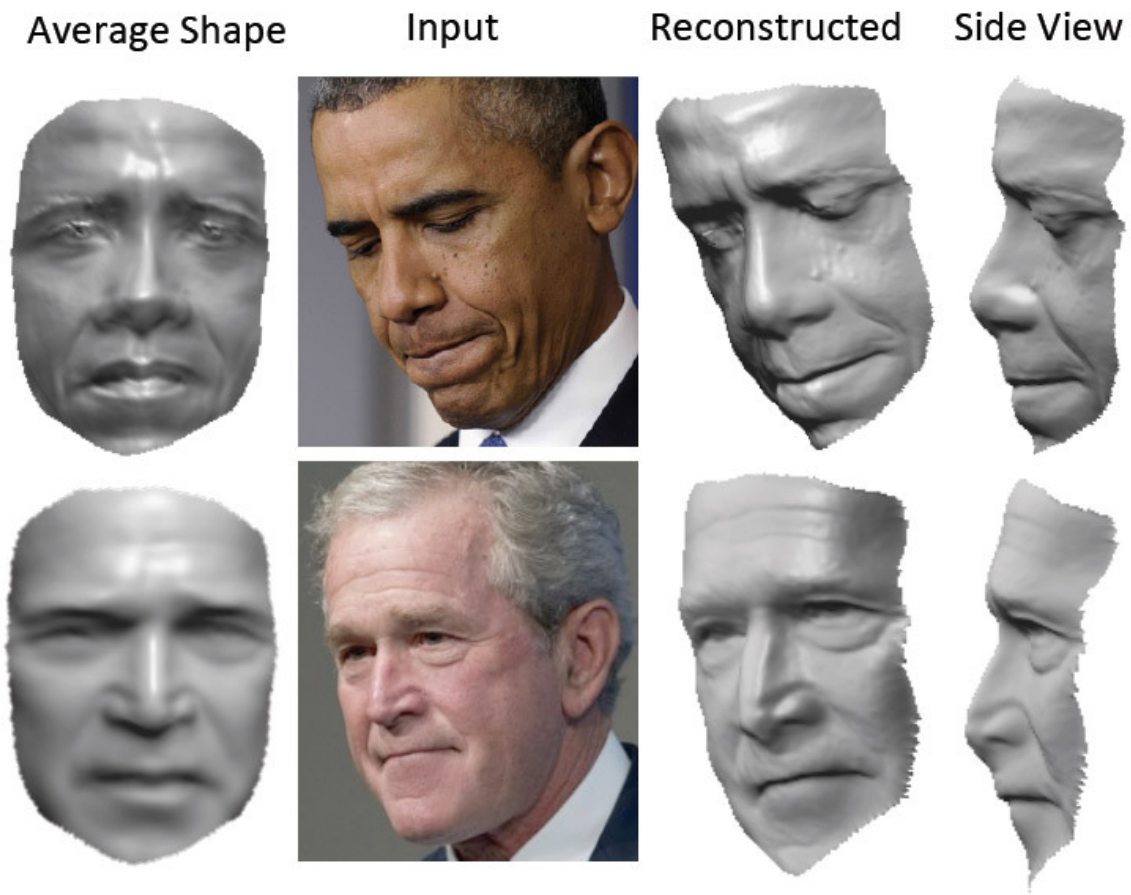
Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, Jonas Unger. *HDR image reconstruction from a single exposure using deep CNNs.* In: ACM Transactions on Graphics (Proc. of SIGGRAPH Asia 2017), 36(6), Article 178, 2017.

Restoration / Super-resolution



https://www.mathworks.com

3D computer vision



https://computer-vision-talks.com/post/computer-vision-digest-september-2014/ .
http://web.stanford.edu/class/cs231a/course_notes.html

Motion: Video classification




Frames from a video

Predictions: Image 1: Highjump = 0.37, Gymnastic = 0.14, Soccer = 0.09, Actual = Nunchucks, Result = fail.

Predictions: Image 2: Wallpushups = 0.67, Boxingbag = 0.09, Jugglingballs = 0.08, Actual = Wallpushups, Result = Top1 correct.

Predictions: Image3: Handstandwalk = 0.32, Nunchuks = 0.16, Jump-rope = 0.11, Actual = jump-rope, Result = Top5 correct.

Predictions: Image4: Drumming = 1.00, Actual = Drumming, Result = Top1 correct.

Web source: https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5.

Motion based detection/segmentation

Multiple object tracking: https://www.youtube.com/watch?v=0CbhfLSGrCM

Many applications…

- MedImage
- Biometrics
- Ecological studies, Agriculture
- Autonomous driving, Robotics
- Remote-sensing
- Image forensics
- Astronomy
- Entertainment and CGI
- Sports analysis
- Online shopping
- Search and social-networking
- …

Mapping or Transformation

Mapping as regression

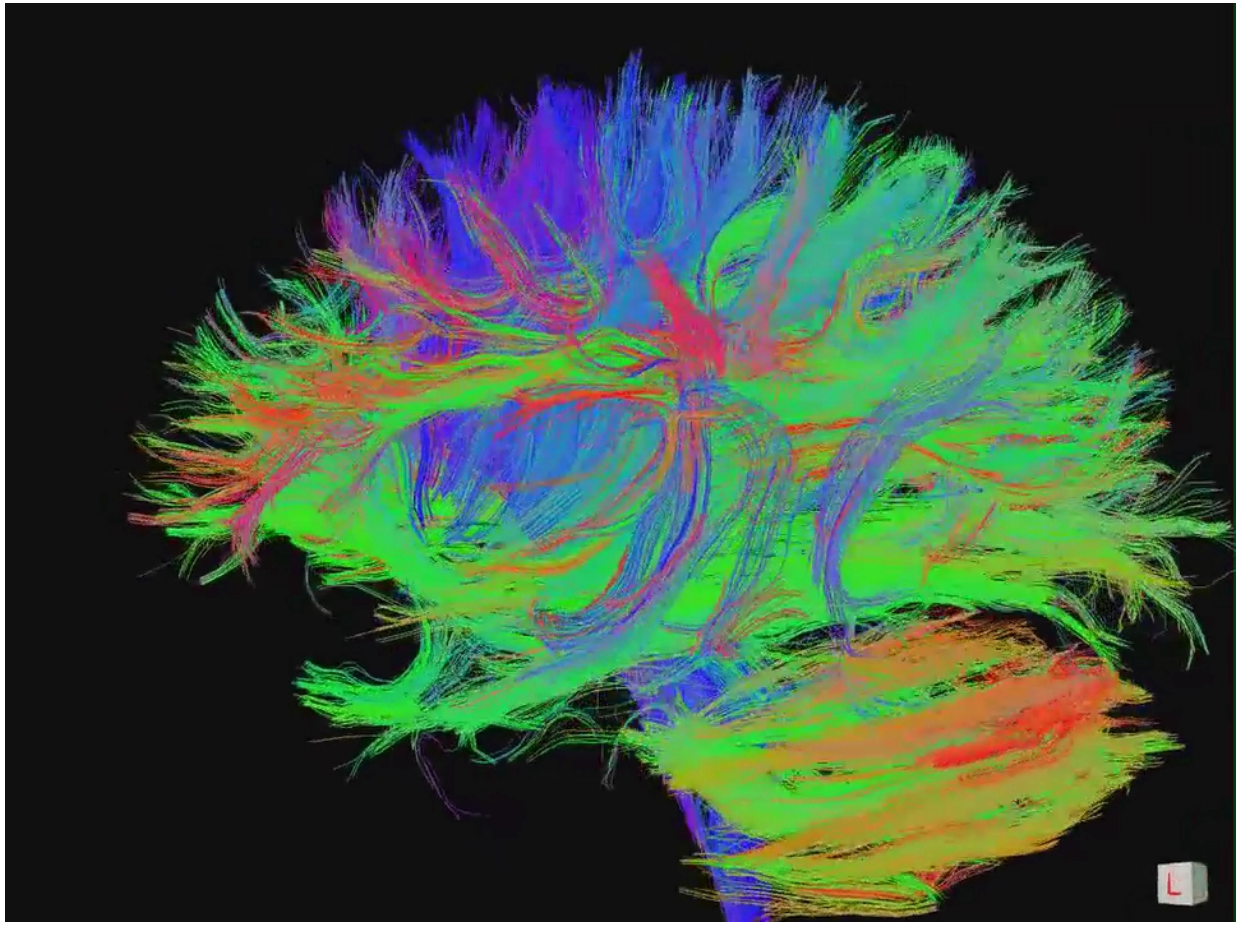
- Linear regression
- Logistic regression (classification)
- Non-linear regression

Mappings in deep neural networks

- Thousands and Tens-of-thousands of parameters
- Highly non-linear
- Data and computation

Despaired with a little complexity ?

- Not really… We possess the Mother of all Neural Networks



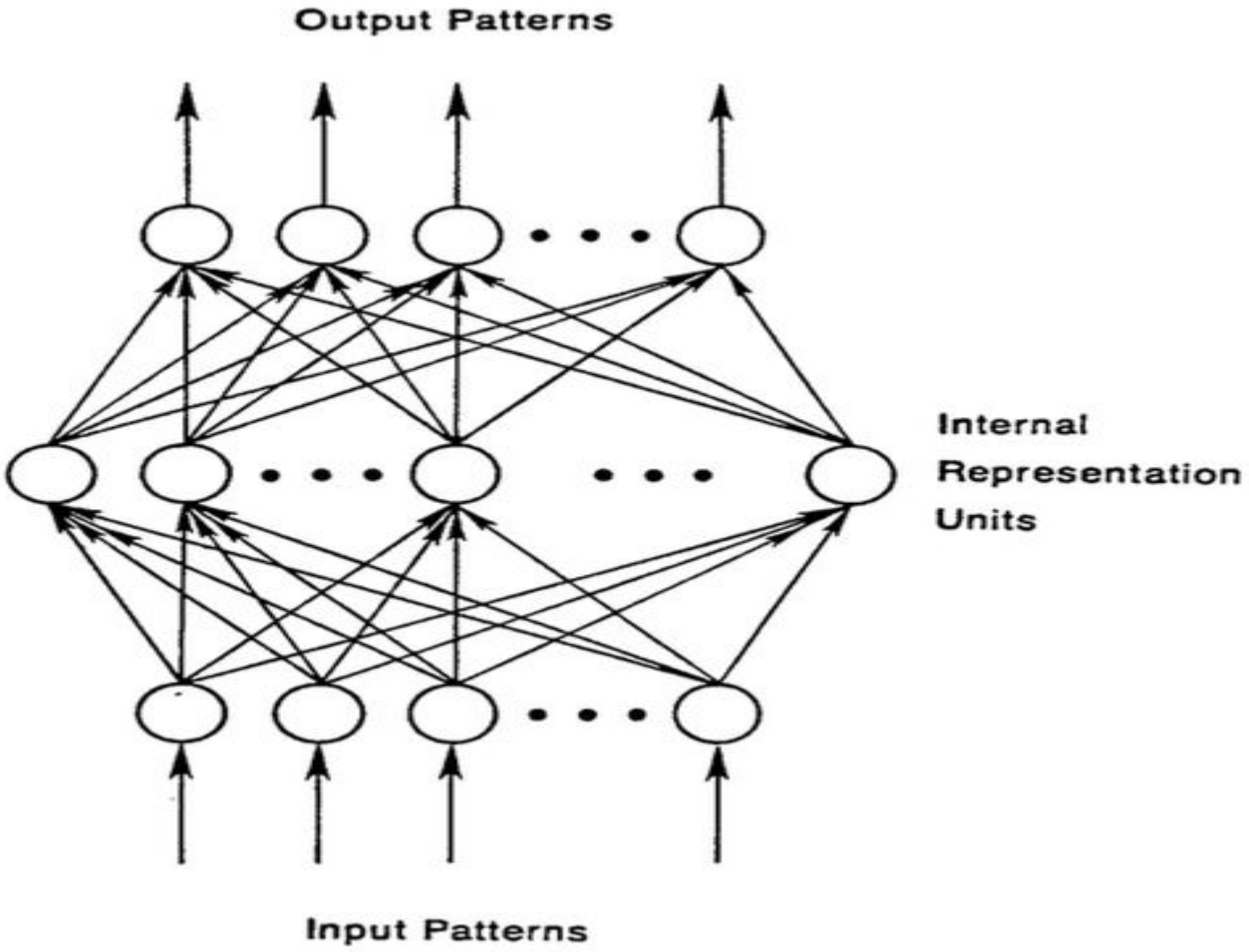
Wikimedia commons: Representation of the Human Connectome

A personal intuitive perspective

- Clever Representations:
  Hierarchy and dependencies in layer-wise mappings
  (structure within a network)
- Task based compositions
  (composite structures with multiple types of networks)
- Global learning objectives (cost function)
  (for individual networks, or for the composite networks)
- Elegant algorithms for cost optimization (or parameter estimation)
- Transfer learning

Some examples

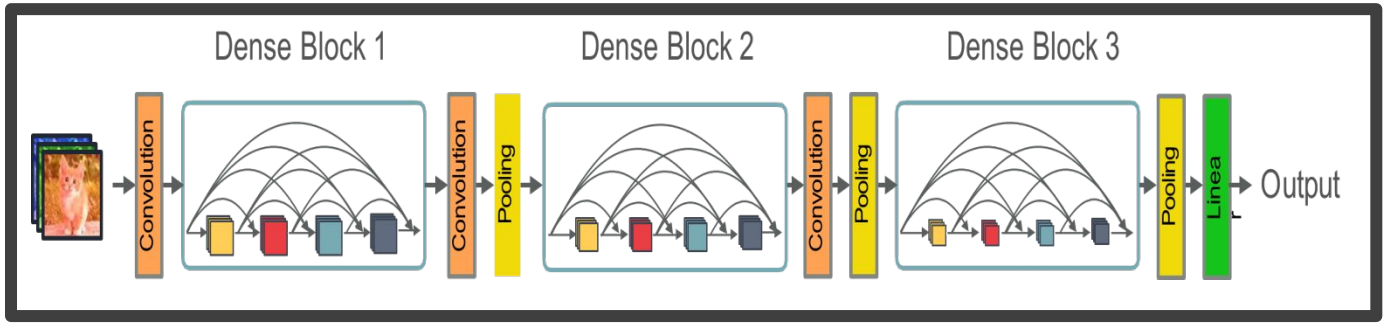
A fully connected deep neural network



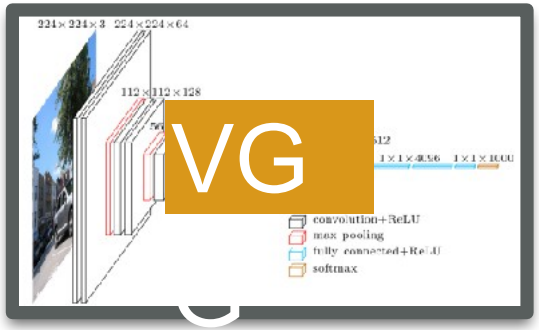
Convolutional networks

- Connections between very less no. of nodes across layers
- Weight sharing
- Notion of filters (and convolution)
- Pooling
- Extended to multiple dimensions (2, 3)

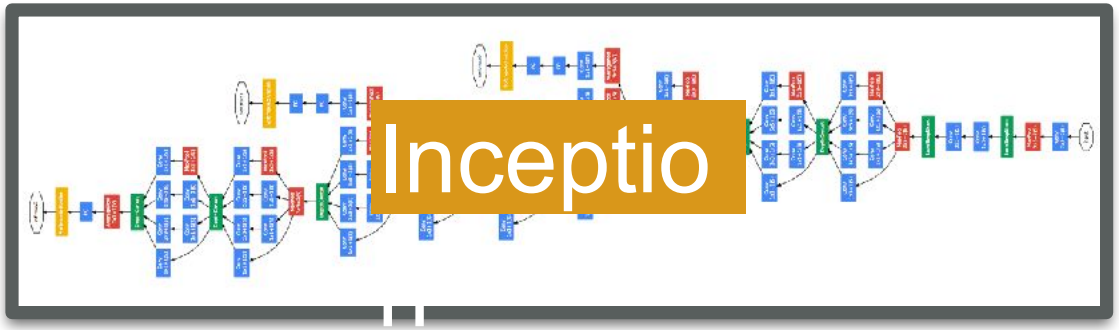
CONVOLUTIONAL ARCHITECTURES



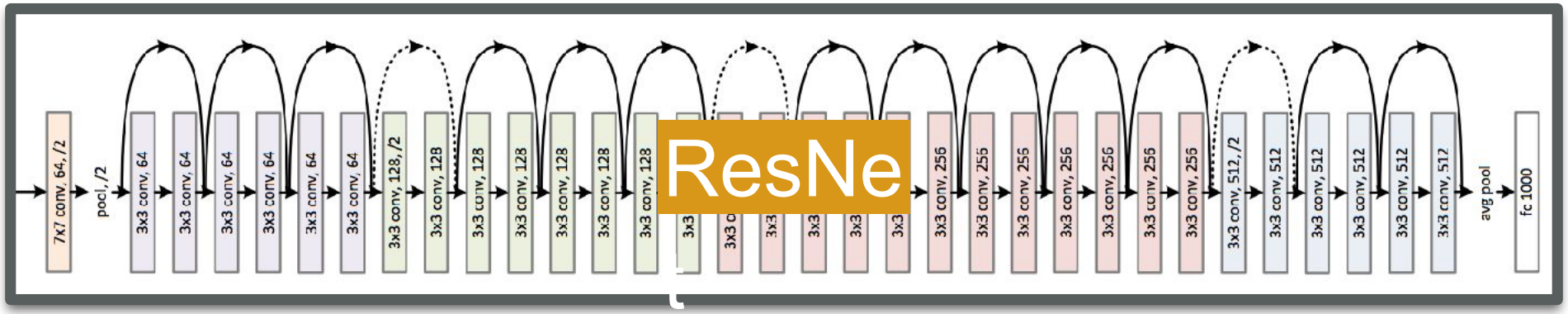
DenseNet



VG



Inceptio



ResNe

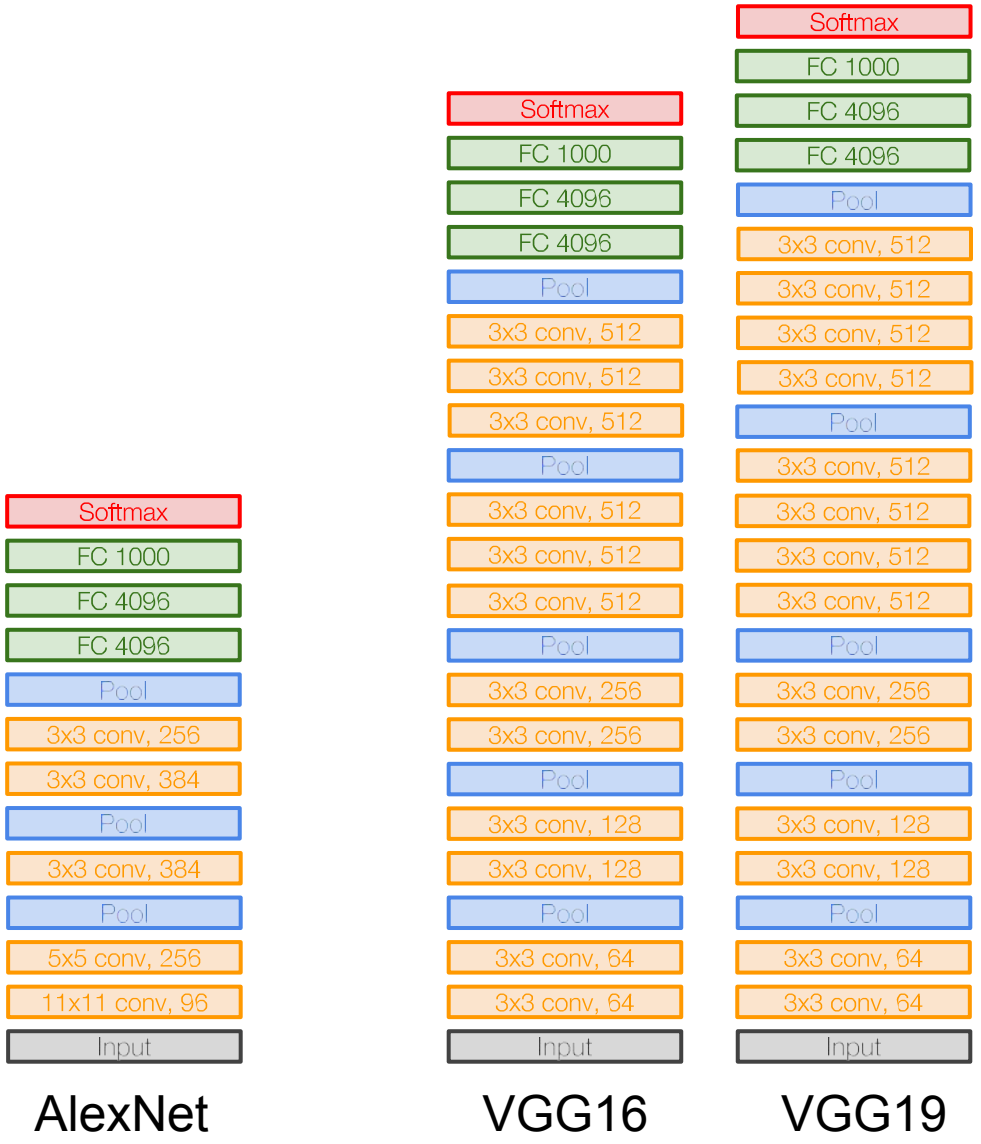
VGGNet

Small filters, Deeper networks

8 layers (AlexNet)
-> 16 - 19 layers (VGG16Net)

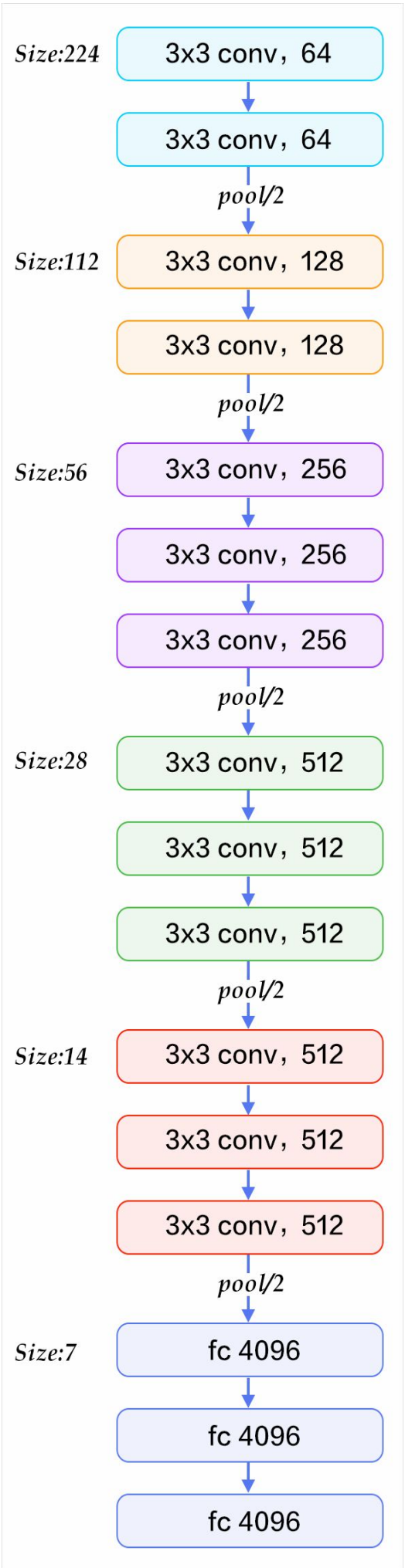
Only 3x3 CONV stride 1, pad 1
and 2x2 MAX POOL stride 2

11.7% top 5 error in ILSVRC'13
(ZFNet)
-> 7.3% top 5 error in ILSVRC'14



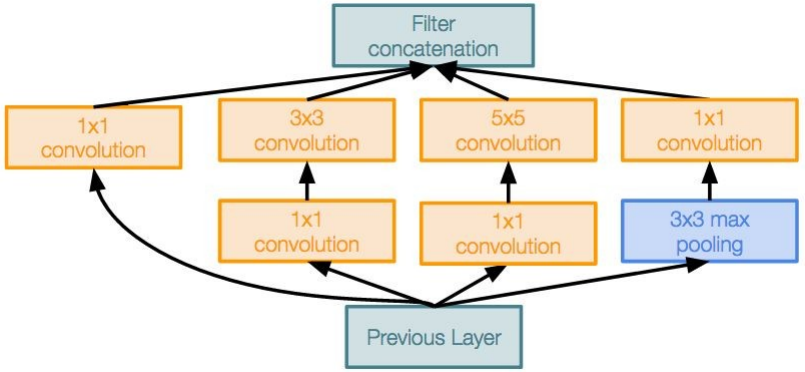
Simonyan and Zisserman, VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION,*2014*

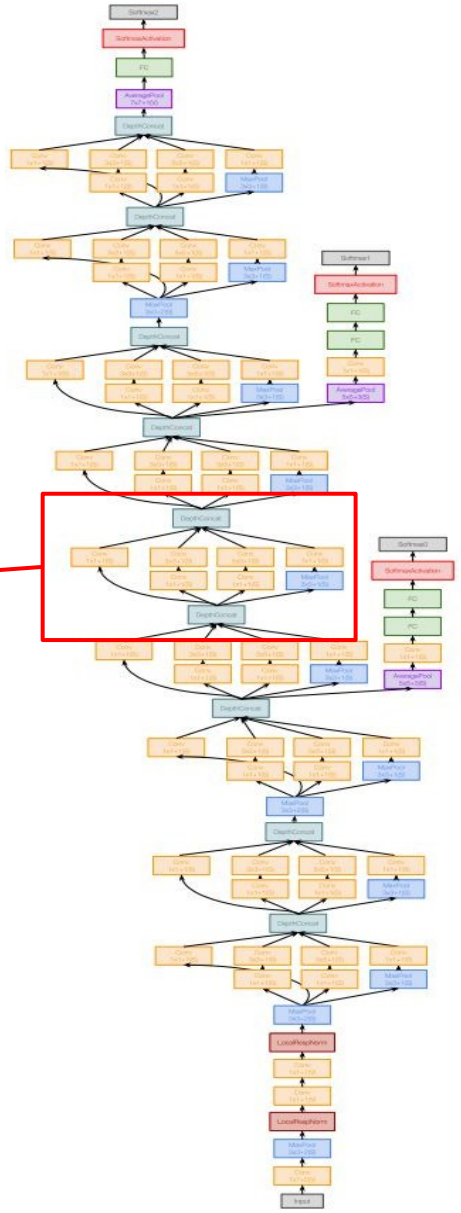
VGGNet


Size:224
3x3 conv，64
3x3 conv，64
pool/2
Size:112
3x3 conv，128
3x3 conv，128
pool/2
Size:56
3x3 conv，256
3x3 conv，256
3x3 conv，256
pool/2
Size:28
3x3 conv，512
3x3 conv，512
3x3 conv，512
pool/2
Size:14
3x3 conv，512
3x3 conv，512
3x3 conv，512
pool/2
Size:7
fc 4096
fc 4096
fc 4096

GoogLeNet: Inception module

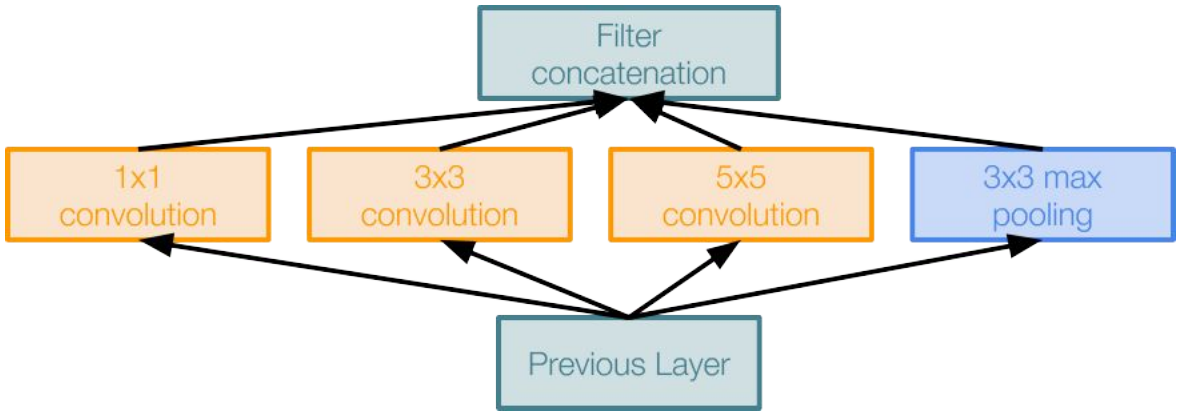
"Inception module": design a good local network topology (network within a network) and then stack these modules on top of each other





C. Szegedy et al., Going deeper with convolutions, CVPR 2015

GoogLeNet: Inception module



Naive Inception module

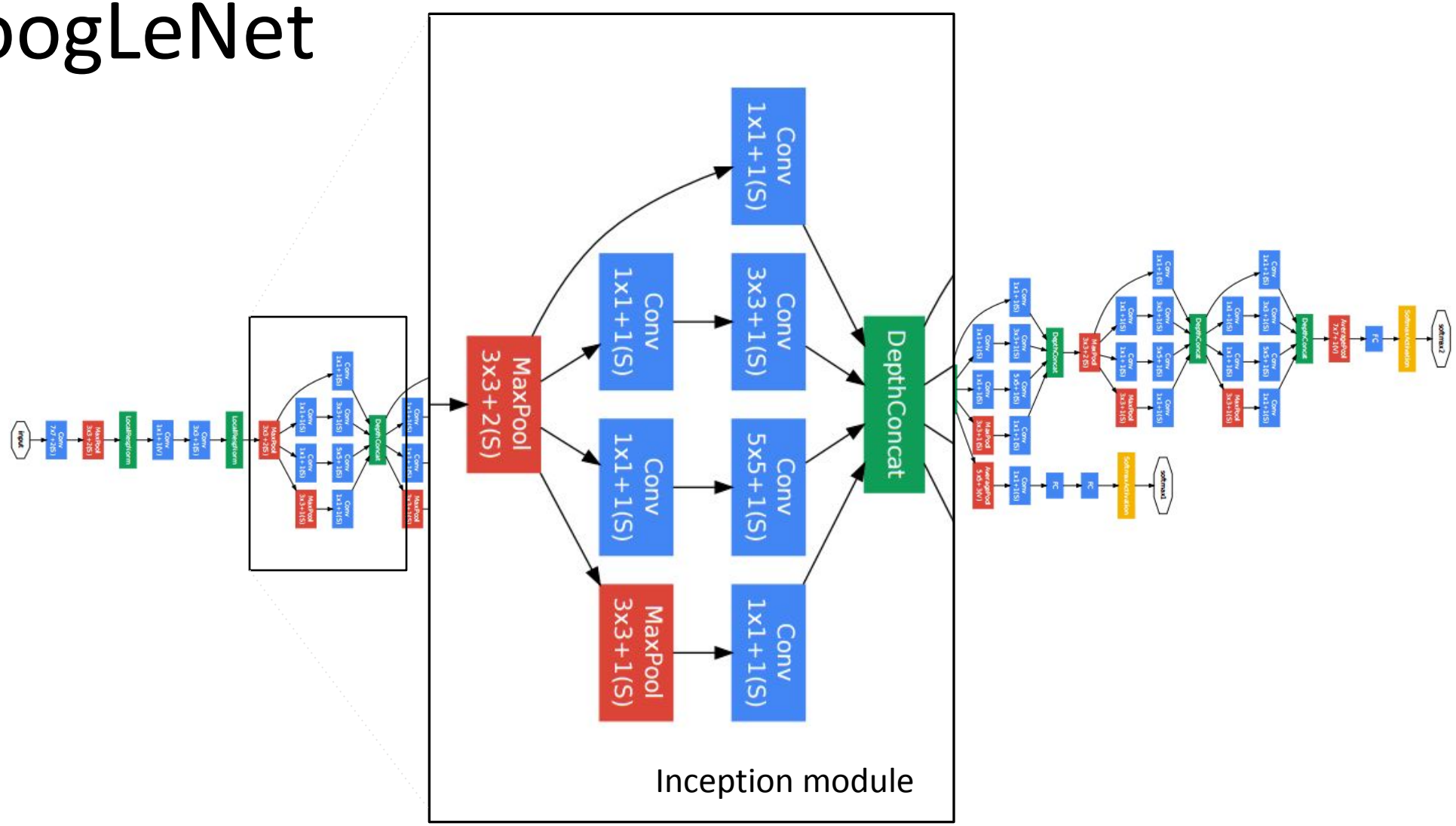
Apply parallel filter operations on the input from previous layer:

- Multiple receptive field sizes for convolution
- Pooling operation (3x3)

Concatenate all filter outputs together depth-wise

C. Szegedy et al., Going deeper with convolutions, CVPR 2015

GoogLeNet



C. Szegedy et al., Going deeper with convolutions, CVPR 2015

ResNet

- Higher training error in deeper models: Vanishing gradients as an important factor

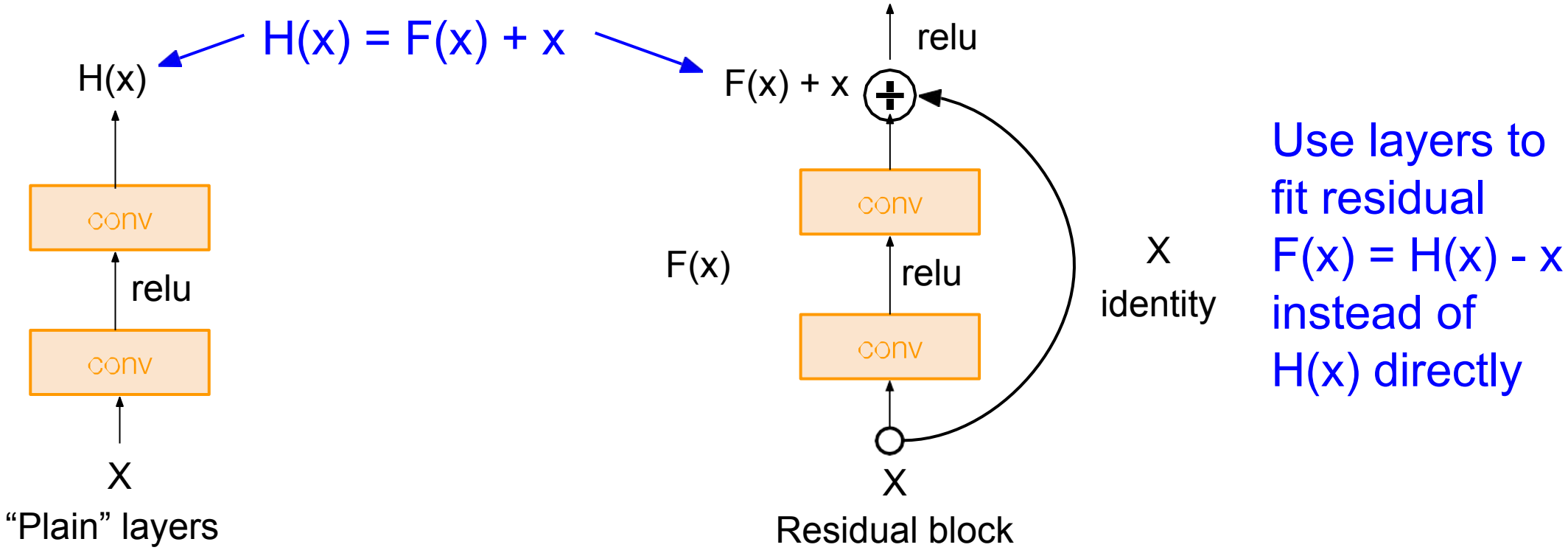
- What if the mappings are identity ?
- The deeper model should be able to perform at least as well as the shallower model.
- The solvers may not approximate (close to) identity mappings well.

- A solution by construction is copying the learned layers from the shallower model.
- Adding identity mappings

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep Residual Learning for Image Recognition, CVPR 2016 (Best Paper)

ResNet

Solution: Use network layers to fit a residual mapping instead of directly trying to fit a desired underlying mapping

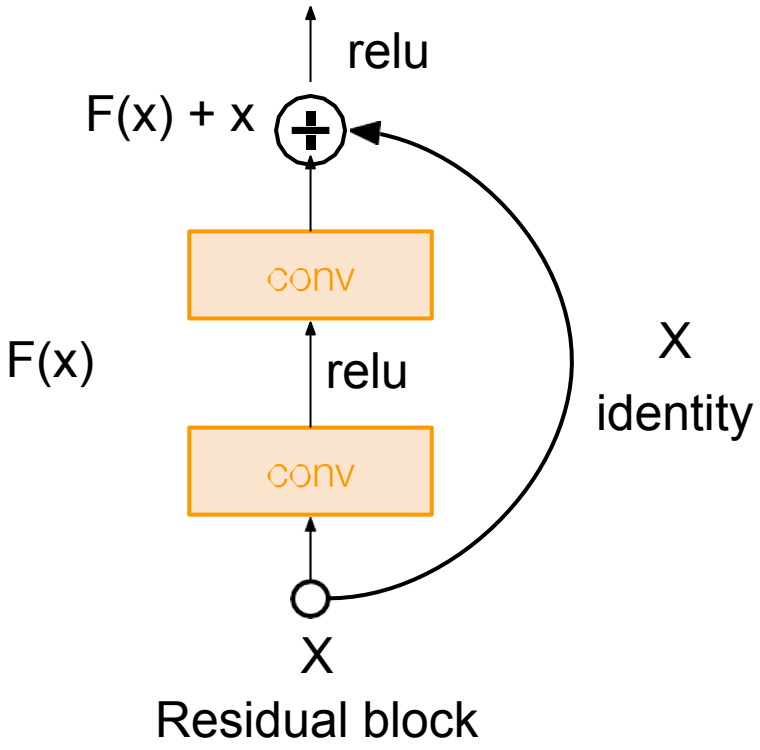


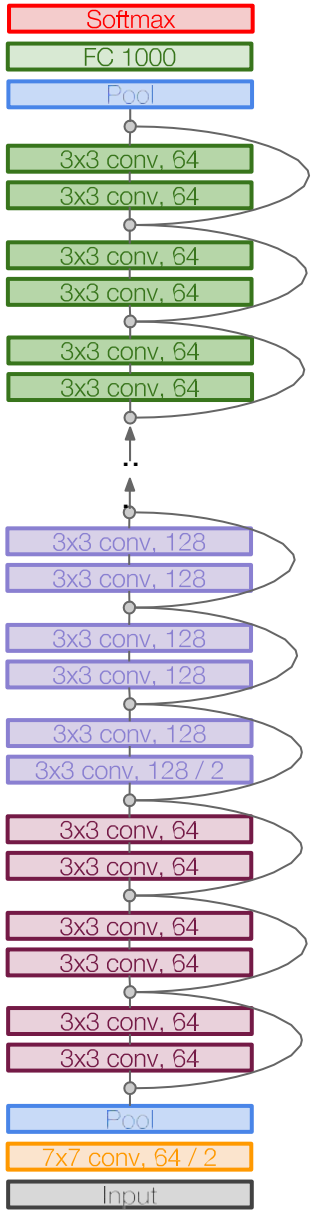
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep Residual Learning for Image Recognition, CVPR 2016 (Best Paper)

ResNet

Very deep networks using residual connections

- 152-layer model for ImageNet
- ILSVRC'15 classification winner (3.57% top 5 error)
- Swept all classification and detection competitions in ILSVRC'15 and COCO'15!

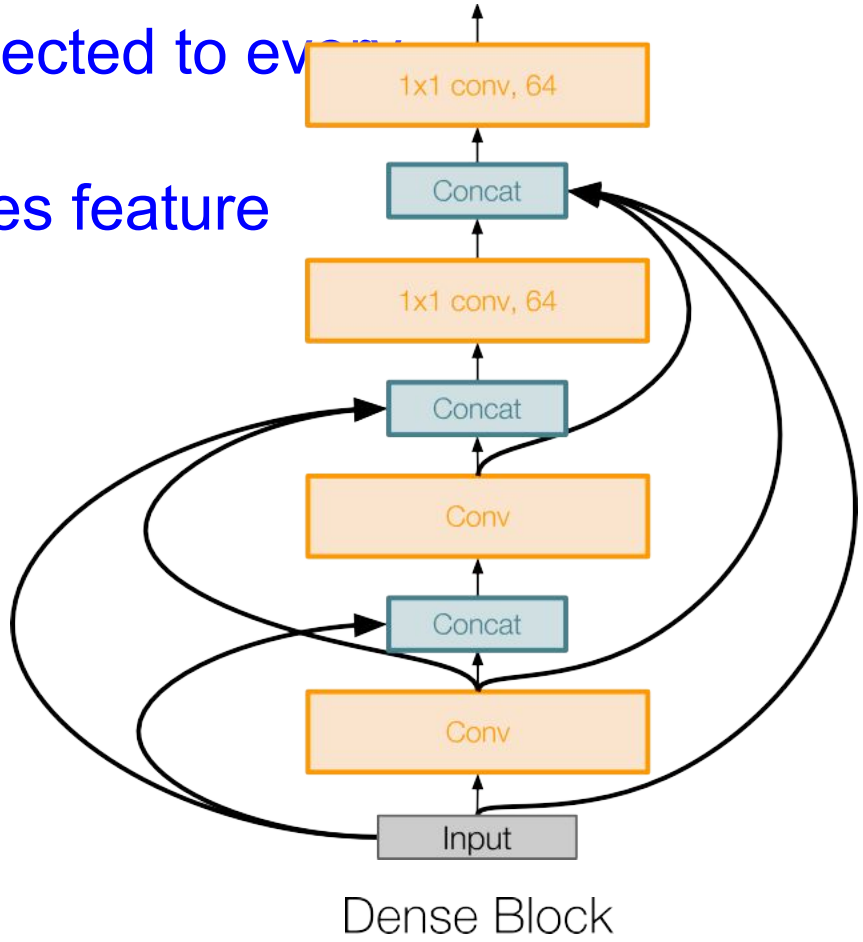




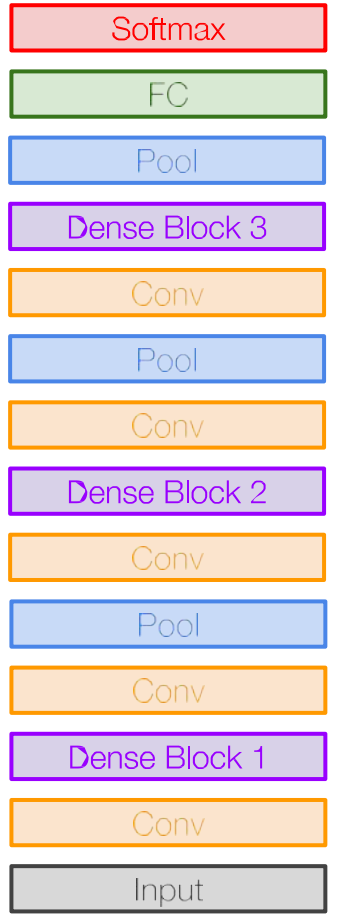
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep Residual Learning for Image Recognition, CVPR 2016

DenseNet

- Dense blocks where each layer is connected to ev… other layer in feedforward fashion
- Alleviates vanishing gradient, encourages feature reuse

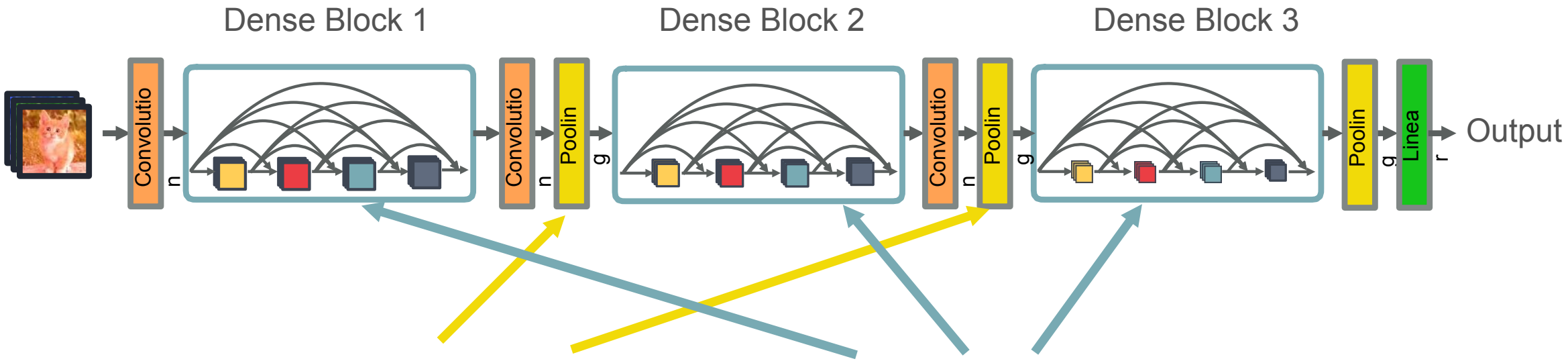


Dense Block



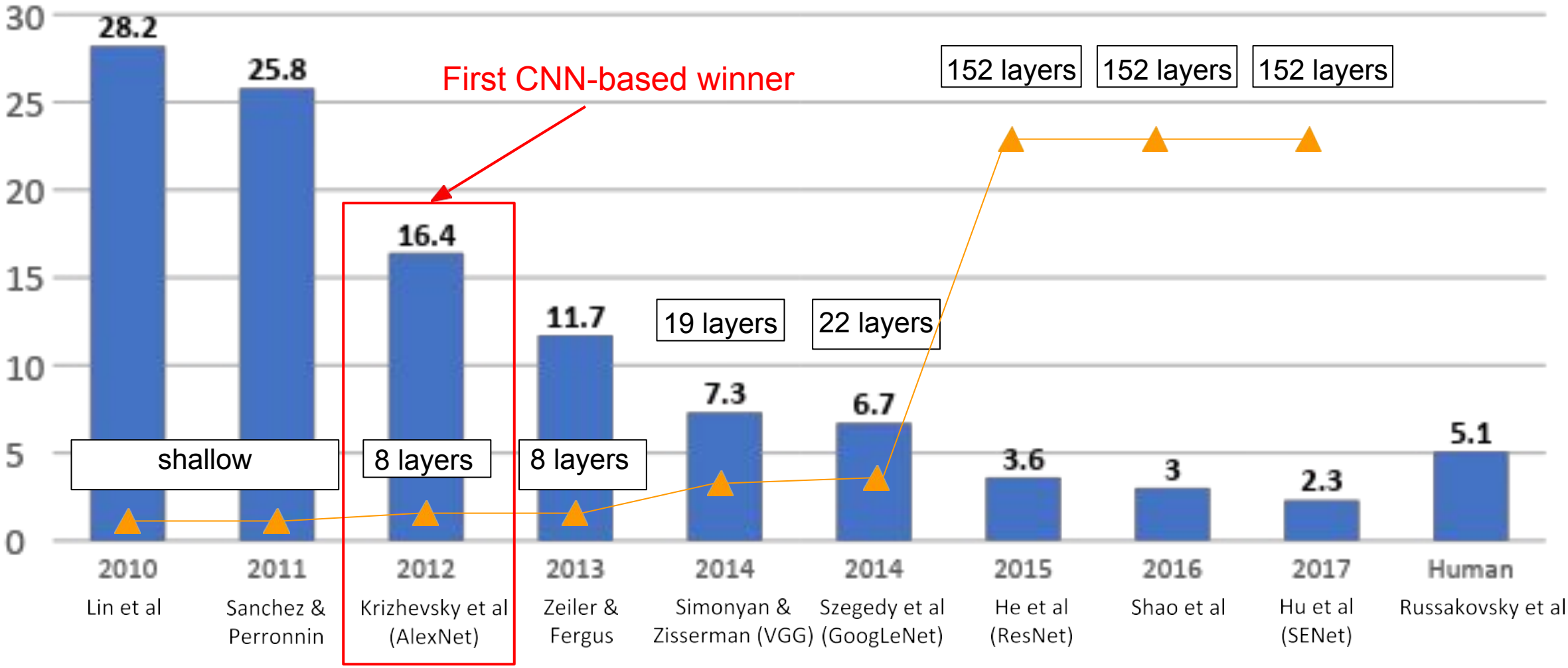
Gao Huang*, Zhuang Liu*, Laurens van der Maaten, Kilian Q.Weinberger , DENSELY CONNECTED CONVOLUTIONAL NETWORKS , CVPR 2017

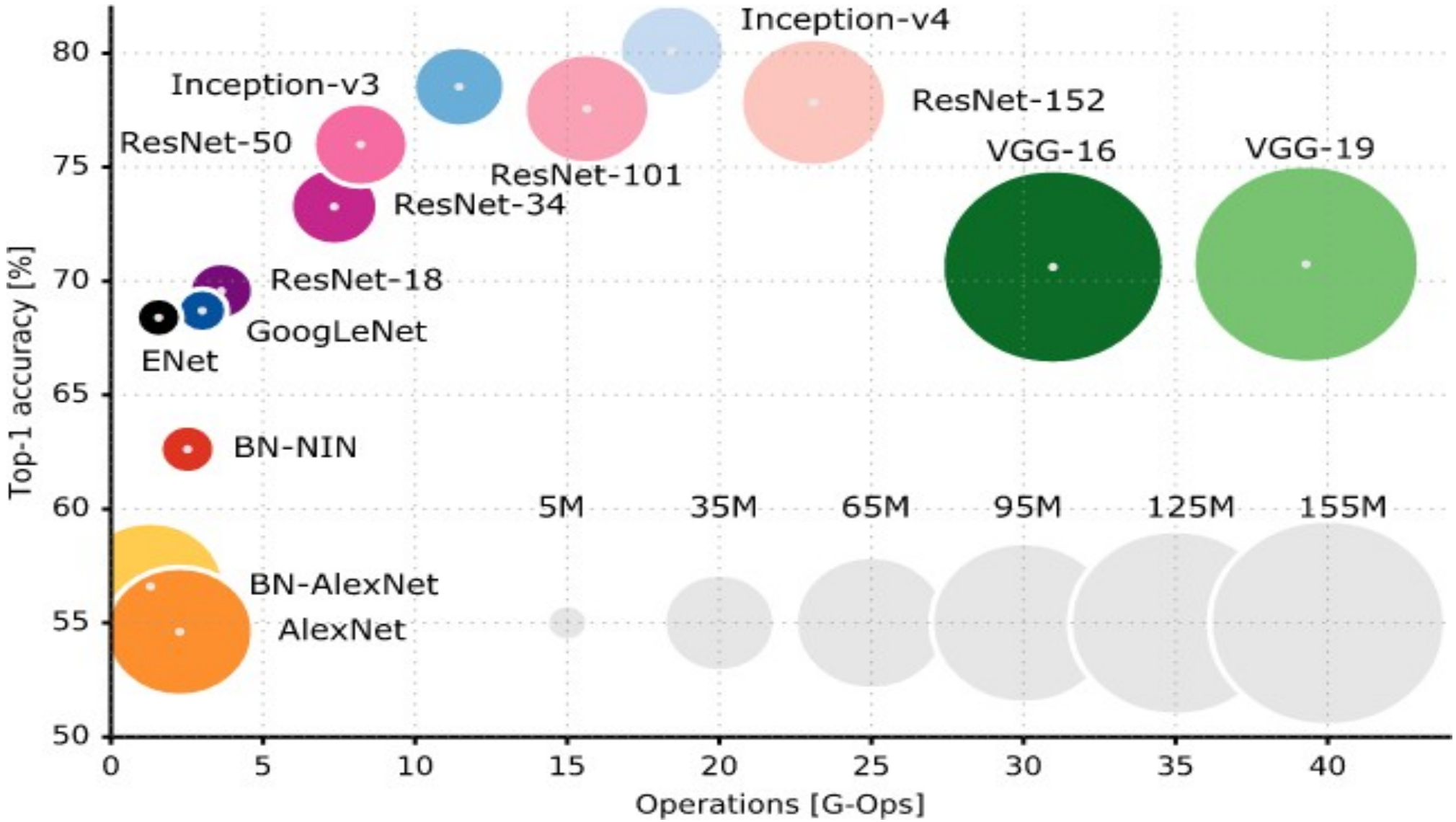
DenseNet



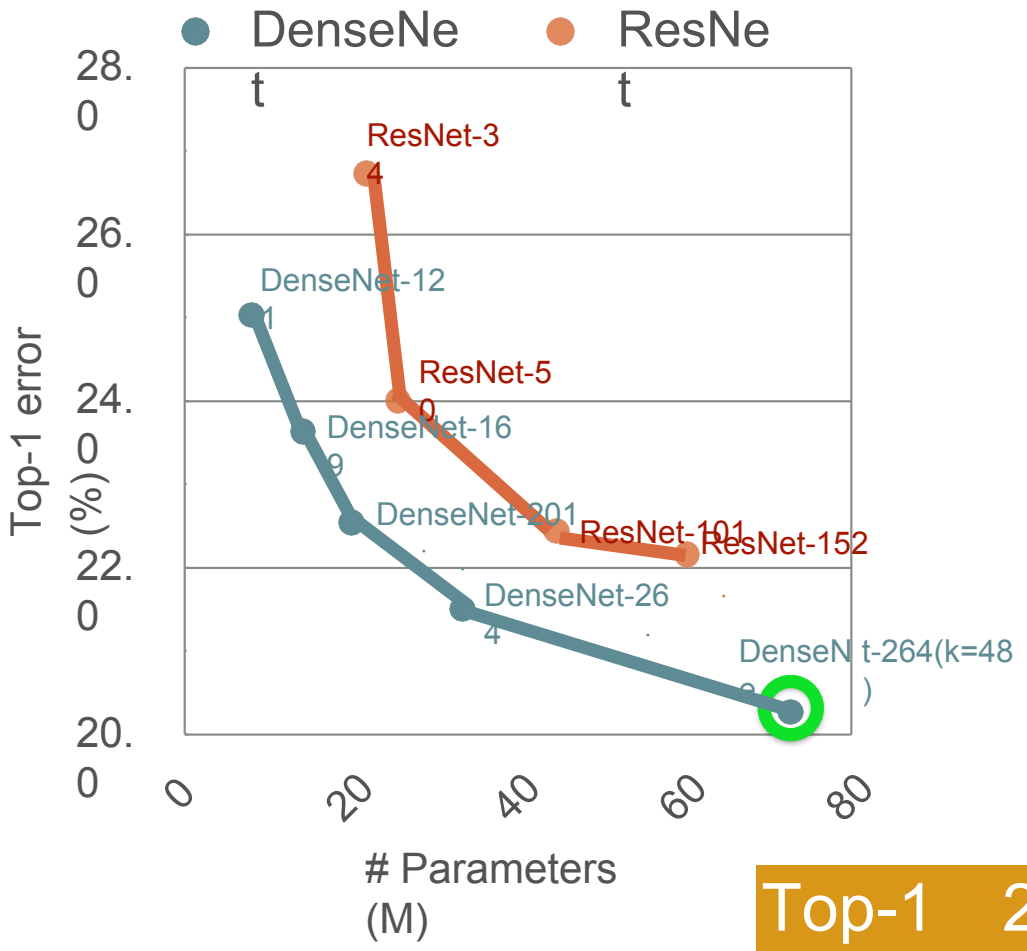
Gao Huang*, Zhuang Liu*, Laurens van der Maaten, Kilian Q.Weinberger , DENSELY CONNECTED CONVOLUTIONAL NETWORKS , CVPR 2017 (Best Paper)

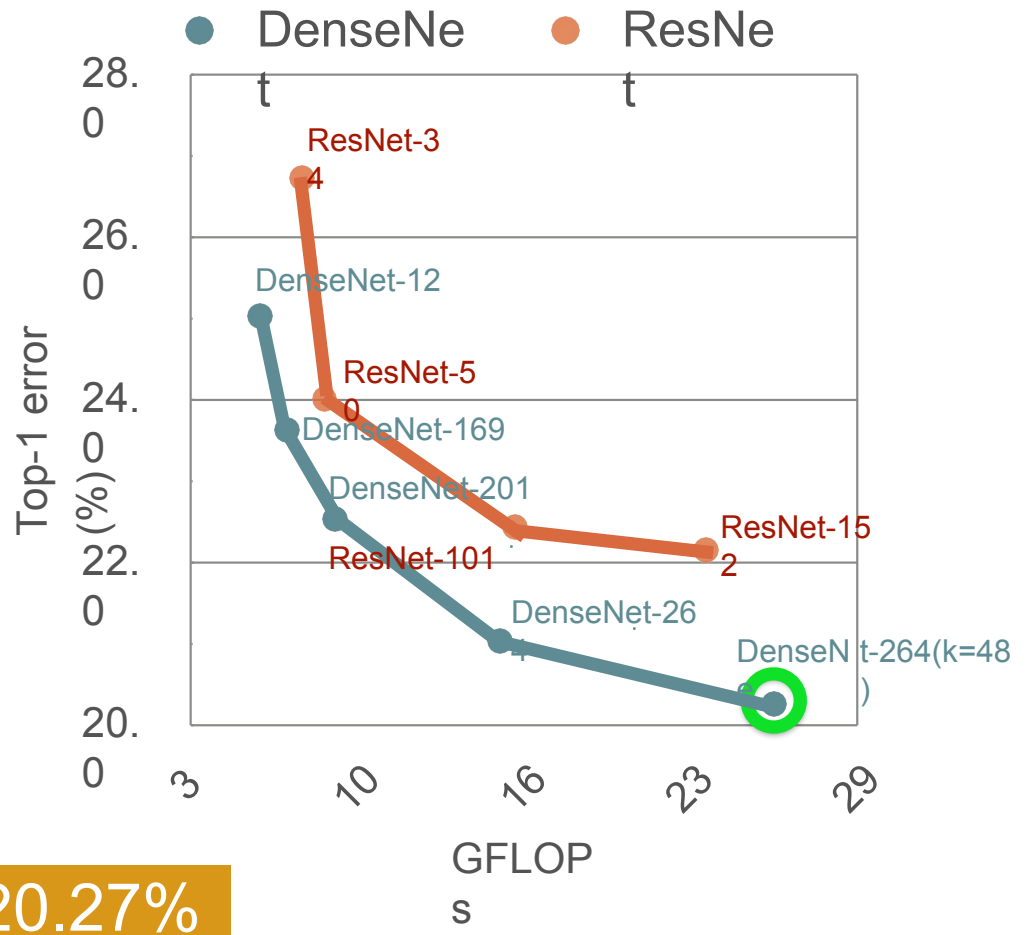
ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners




Inception-v4
Inception-v3
ResNet-152
ResNet-50
VGG-16
VGG-19
ResNet-101
ResNet-34
ResNet-18
GoogLeNet
ENet
BN-NIN
5M
35M
65M
95M
125M
155M
BN-AlexNet
AlexNet
Top-1 accuracy [%]
80
75
70
65
60
55
50
0
5
10
15
20
25
30
35
40
Operations [G-Ops]

RESULTS ON **IMAGENET**





Top-1 :	20.27%
	5.17%

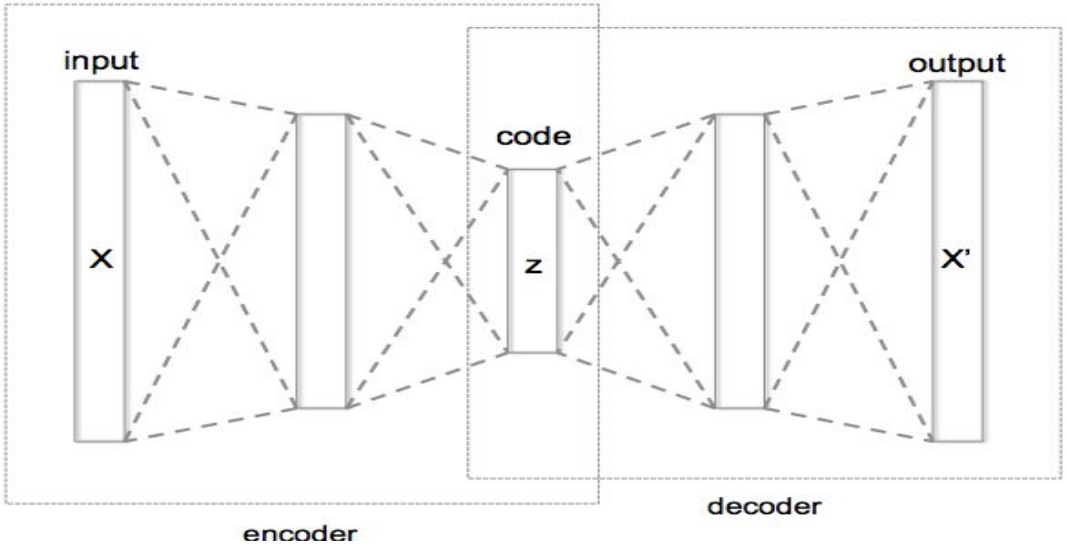
Autoencoders neural networks

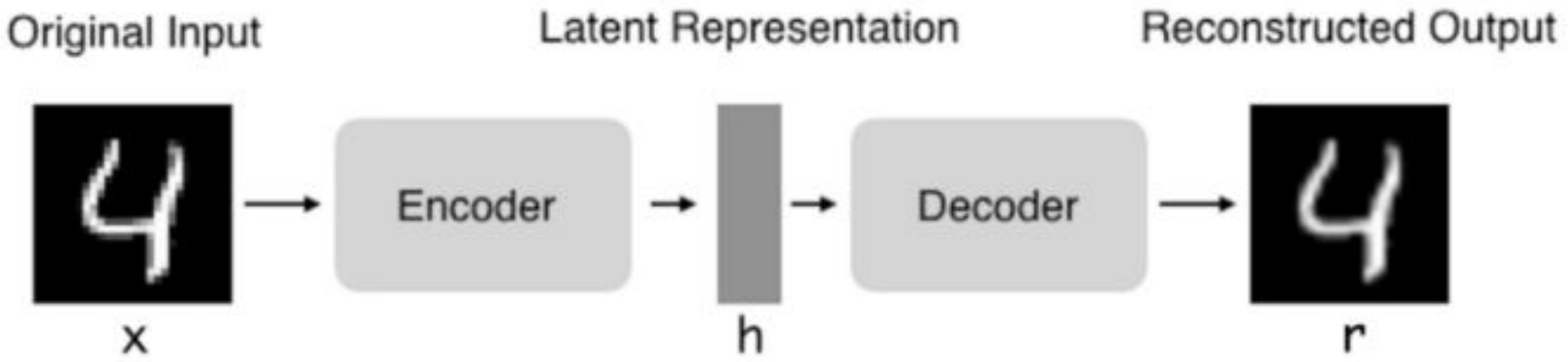
- An unsupervised machine learning algorithm.

- Learn a representation (encoding) for a set of data, typically for the purpose of dimensionality reduction, image recognition and information retrieval.

- **Encoder:** compresses the input into a latent-space representation. It can be represented by an encoding function $h=f(x)$.
- **Decoder:** to reconstruct the input from the latent space representation. It can be represented by a decoding function $r=g(h)$.

Autoencoders neural network





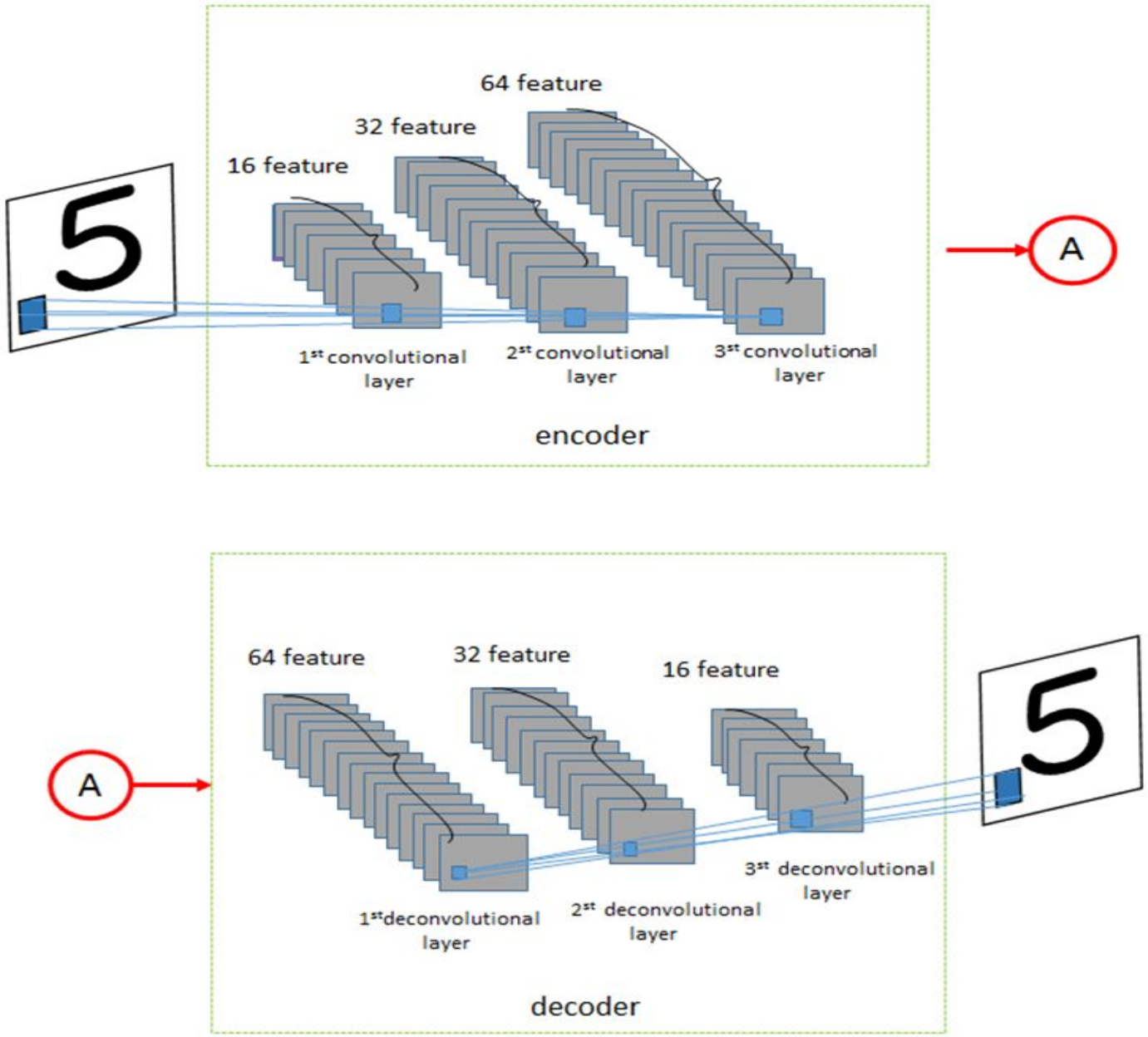
Ref: Pierre Baldi. 2011. Autoencoders, unsupervised learning and deep architectures. In Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop - Volume 27 (UTLW'11), Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver (Eds.), Vol. 27. JMLR.org 37-50.

Convolutional autoencoders

- Convolutional Autoencoders (CAE's) use the convolution operator to encode the input in a set of simple signals and then reconstruct the input from them.
- Learn the optimal filters that minimize the reconstruction error.
- Auto-encoders are models that learn the *non-trivial identity* function by learning a manifold on which the data lies on and can be used to **generate** the samples from the learned manifold.
- Convolution auto-encoders carry the same features except the fact that convolutional layers are present in the model.

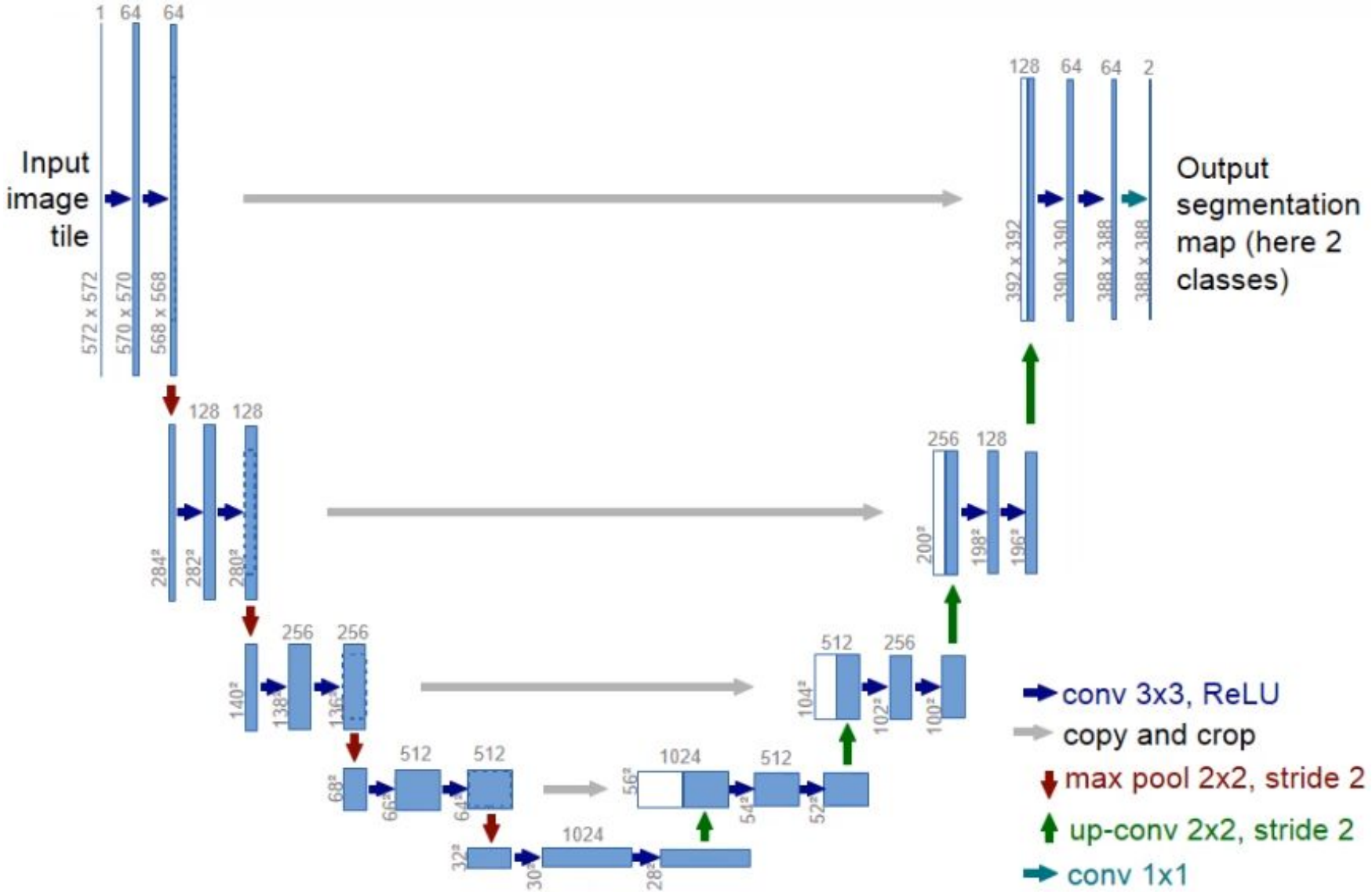
B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang and D. Tao, "Stacked Convolutional Denoising Auto-Encoders for Feature Representation," in *IEEE Transactions on Cybernetics* , vol. 47, no. 4, pp. 1017-1027, April 2017.

Convolutional autoencoders

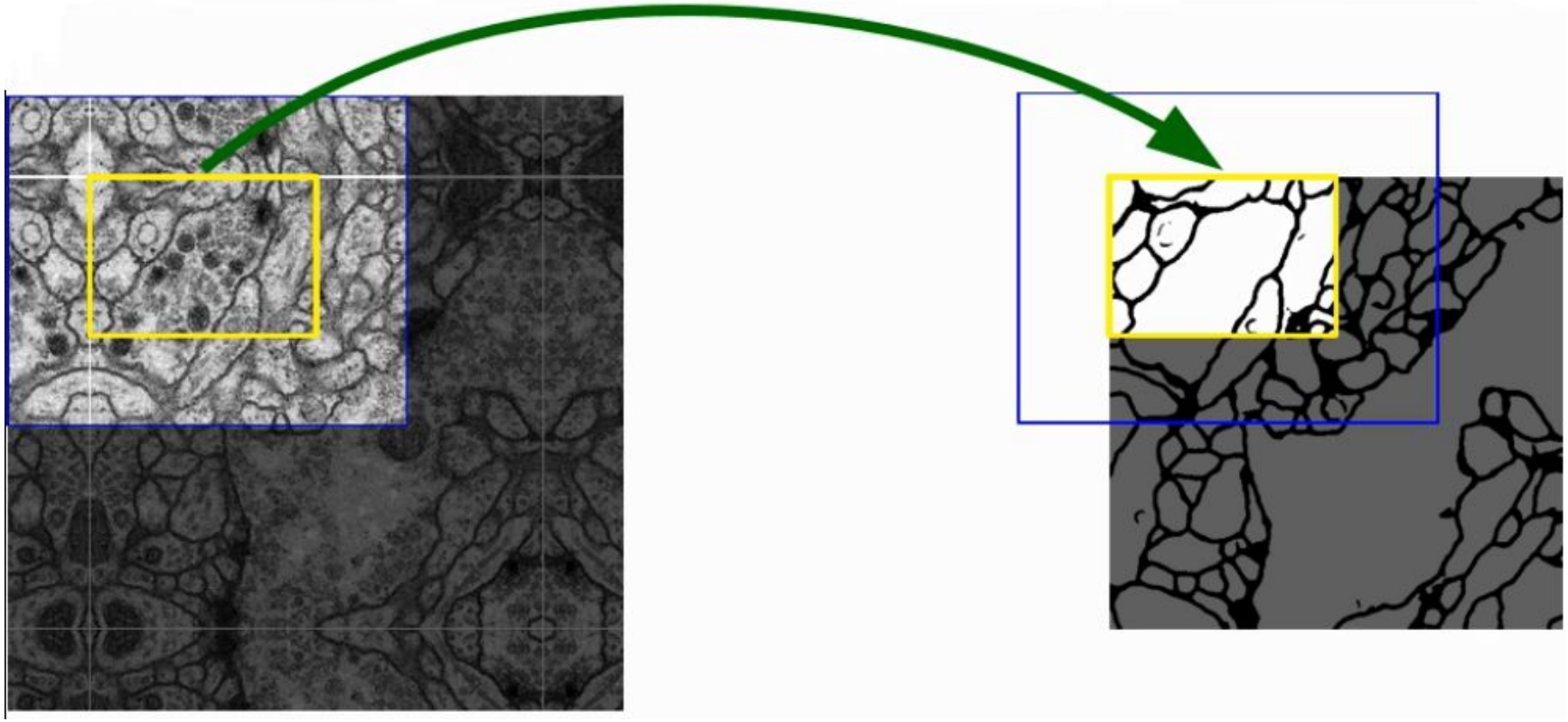


Segmentation: U-Net and Seg-Net

U-Net architecture

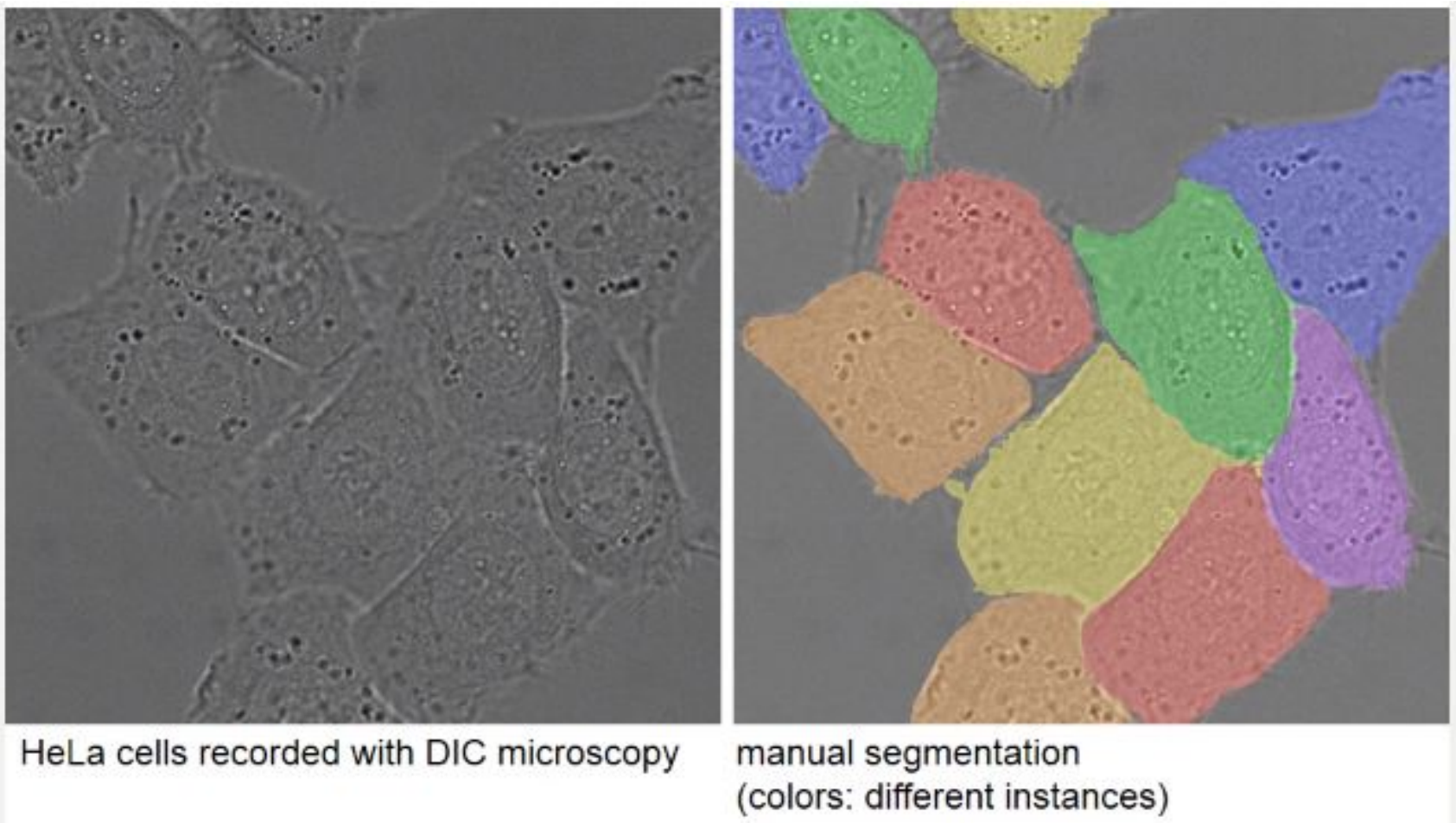


Example result



- Segmentation of yellow area uses input image of blue area.

Results: Segmentation of touching objects

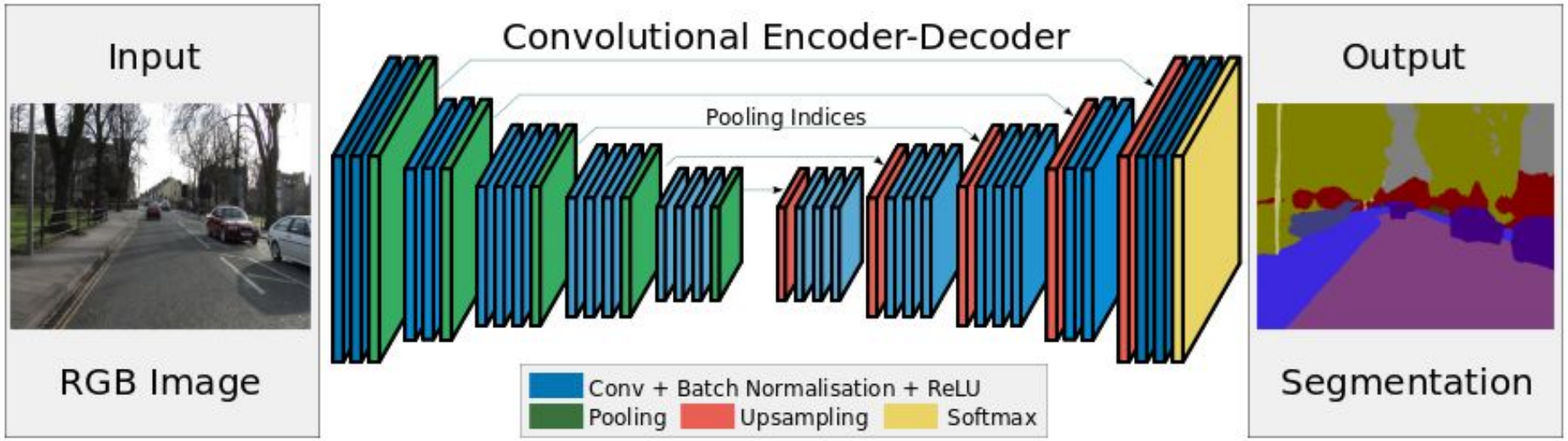


HeLa cells recorded with DIC microscopy

manual segmentation
(colors: different instances)

Source: Data provided by Dr. Gert Van Cappellen, Erasmus medical center, Rotterdam The Netherlands.

SegNet



Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." PAMI, 2017.

SegNet

- Semantic pixel-wise segmentation.
- Topologically identical to the convolutional layers in VGG16.
- The key component: decoder network, consisting of a hierarchy of decoders.
- The decoders use the max-pooling indices received from the corresponding encoder to perform non-linear upsampling of their input feature maps.
- The upsampled maps are sparse and are then convolved with trainable filters to produce dense feature maps.
- Efficient both in terms of memory and computational time during inference.

Results: Scene Image Segmentation

Outdoor scenes

Indoor scenes

Object detection networks

(F-RCNN and YOLO)

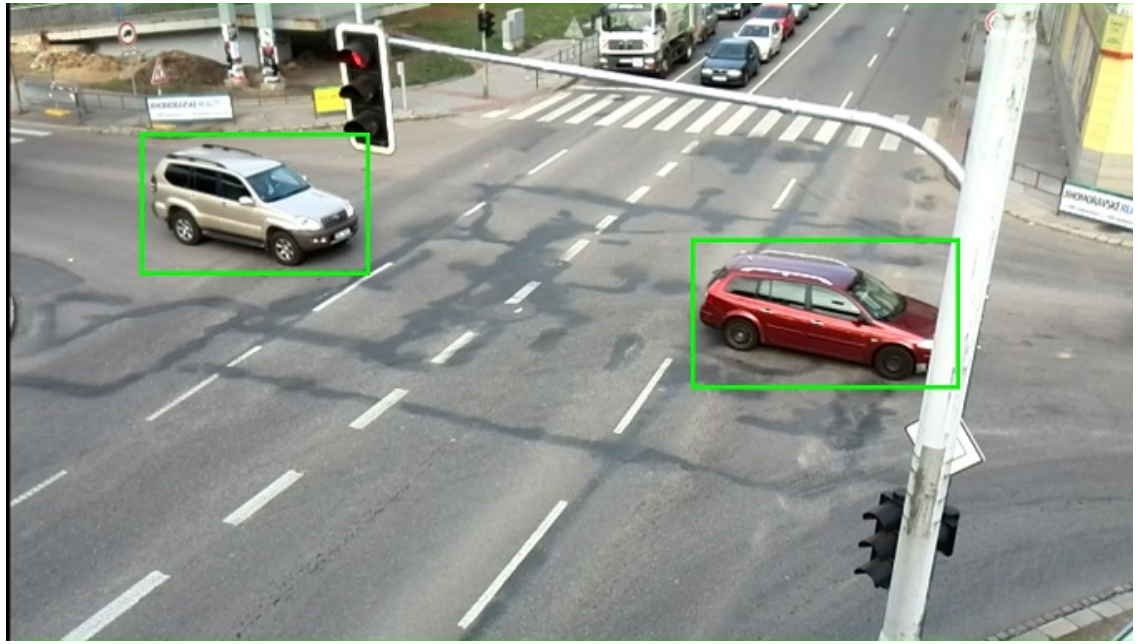
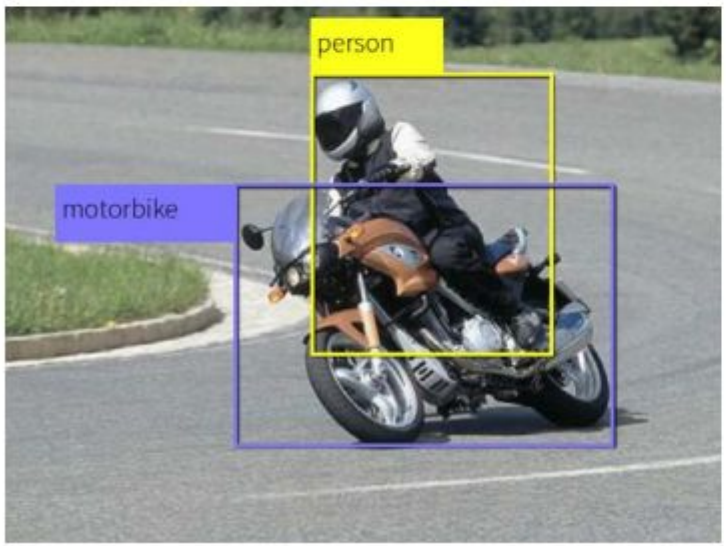


Image source: http://answers.opencv.org/question/87261/extract-parts-from-image/.

Detection and Segmentation

input image

object detection



segmentation

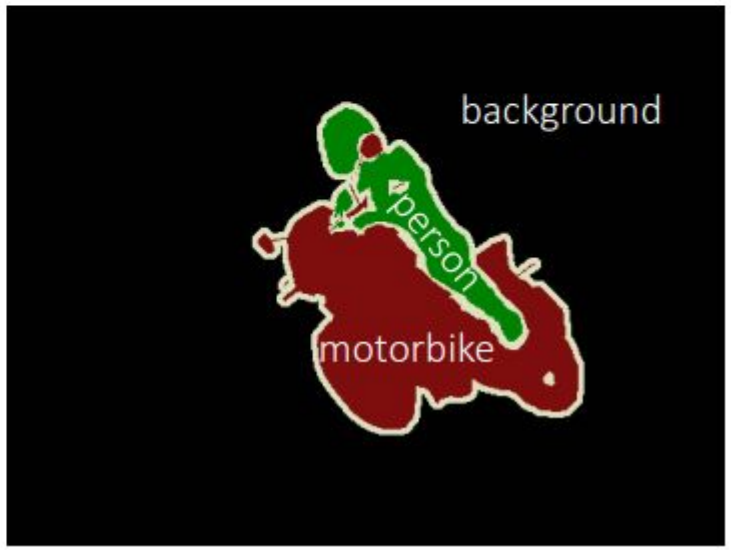


Image source: https://courses.cs.washington.edu/courses/cse590v/14au.

F-RCNN: Fast Region based CNN

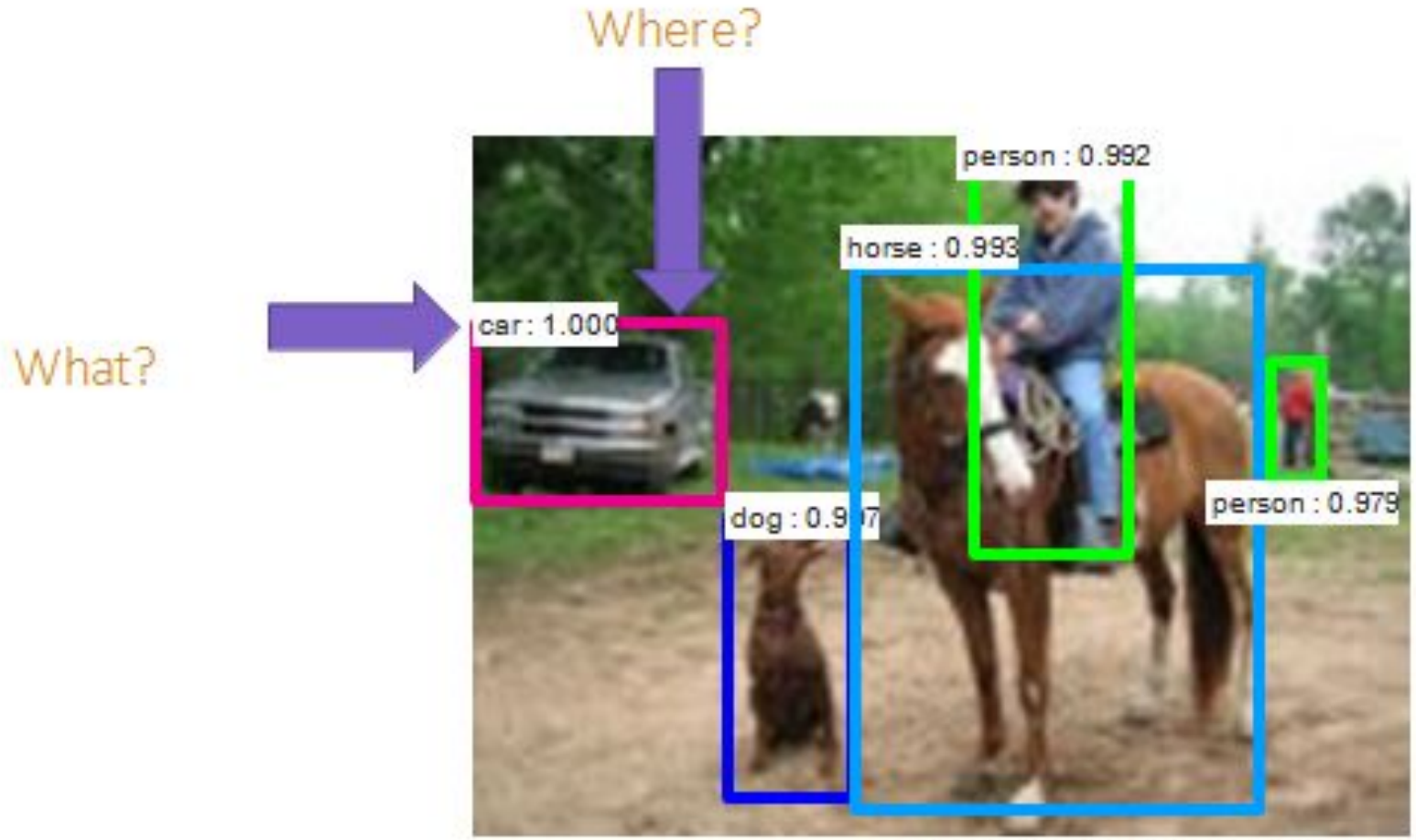
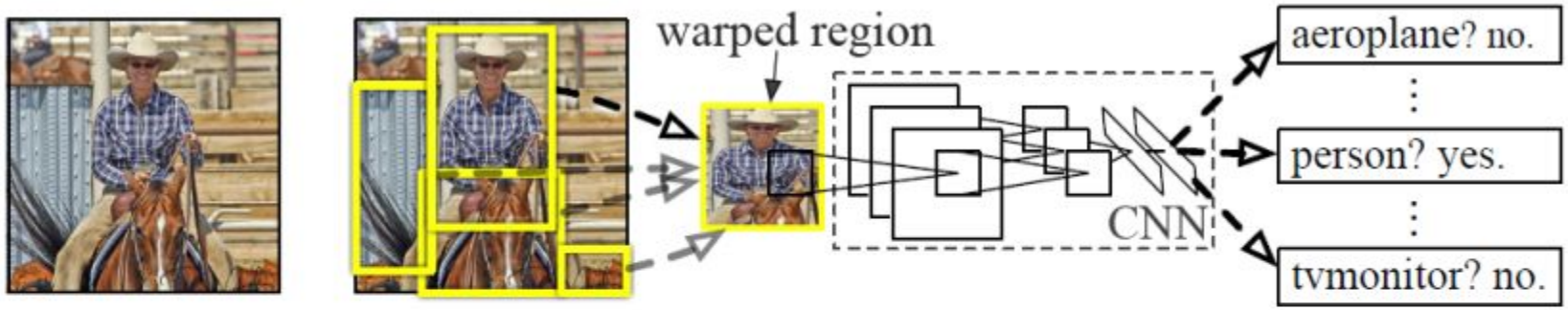
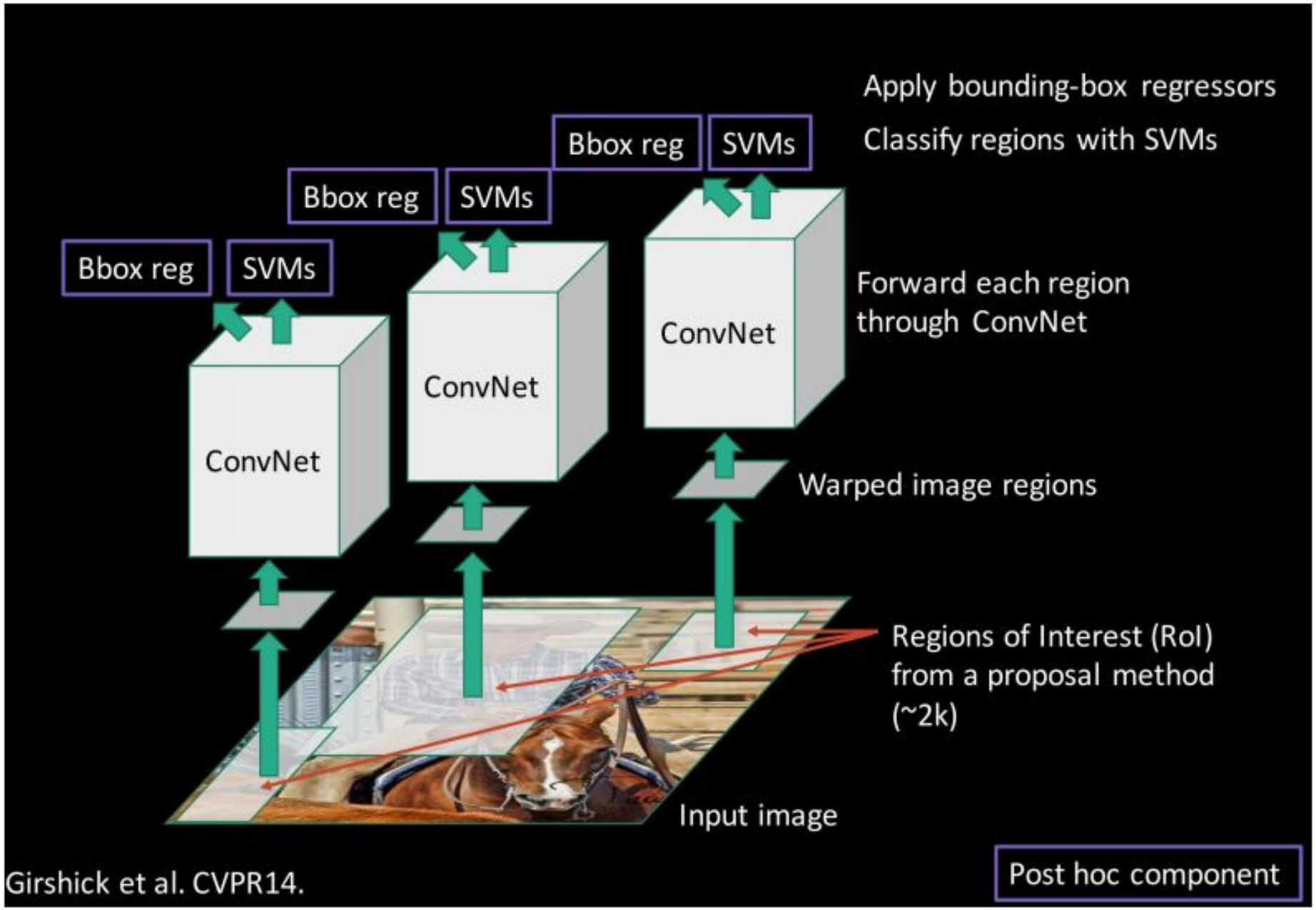


Figure adapted from Kaiming He

RCNN: Region proposals + CNN

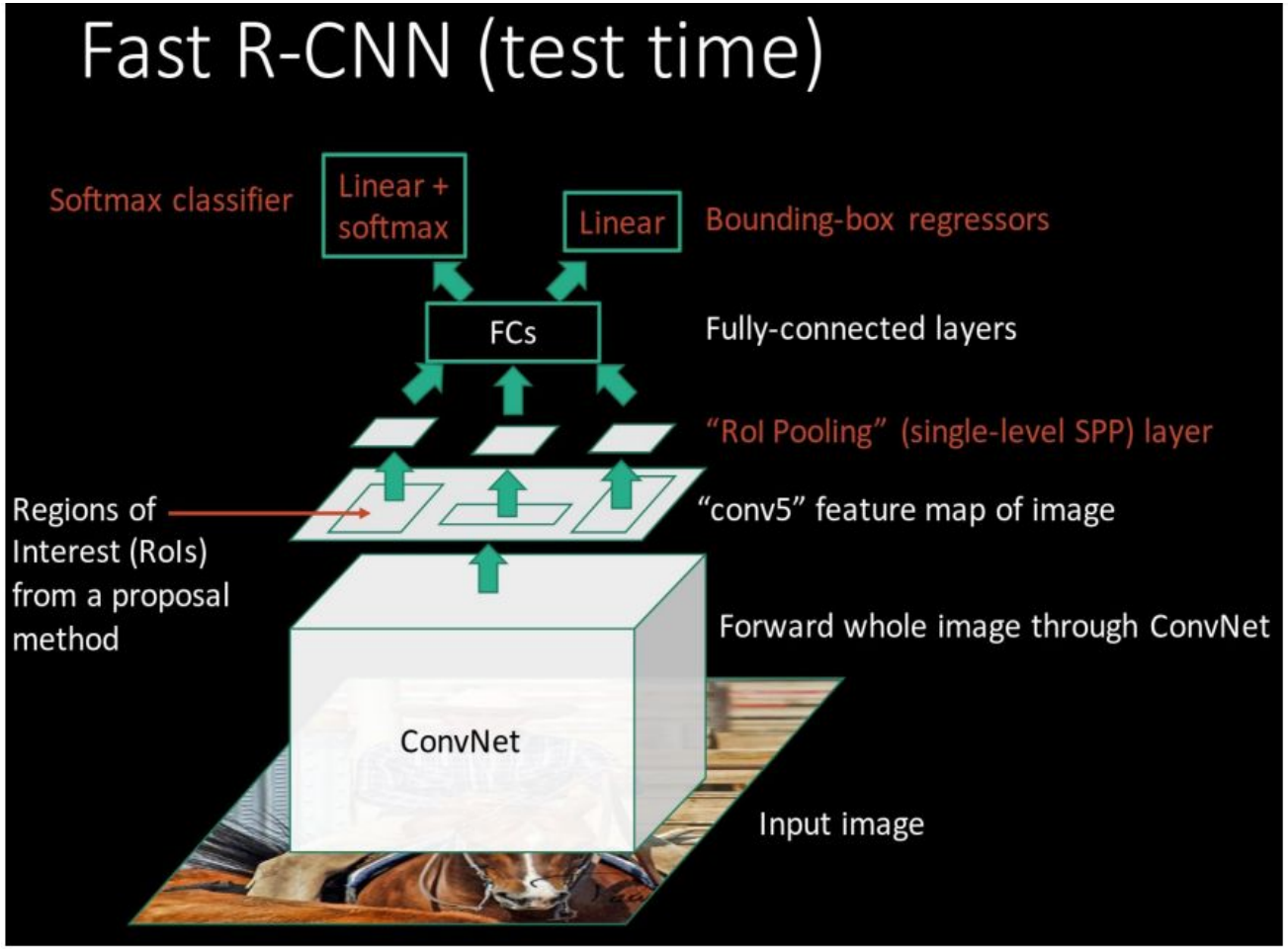


	localization	feature extraction	classification
	selective search	deep learning CNN	binary linear SVM
alternatives:	objectness, constrained parametric min-cuts, sliding window …	HOG, SIFT, LBP, BoW, DPM …	SVM, Neural networks, Logistic regression …



Girschick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
Slide credit: Ross Girschick.

Fast RCNN

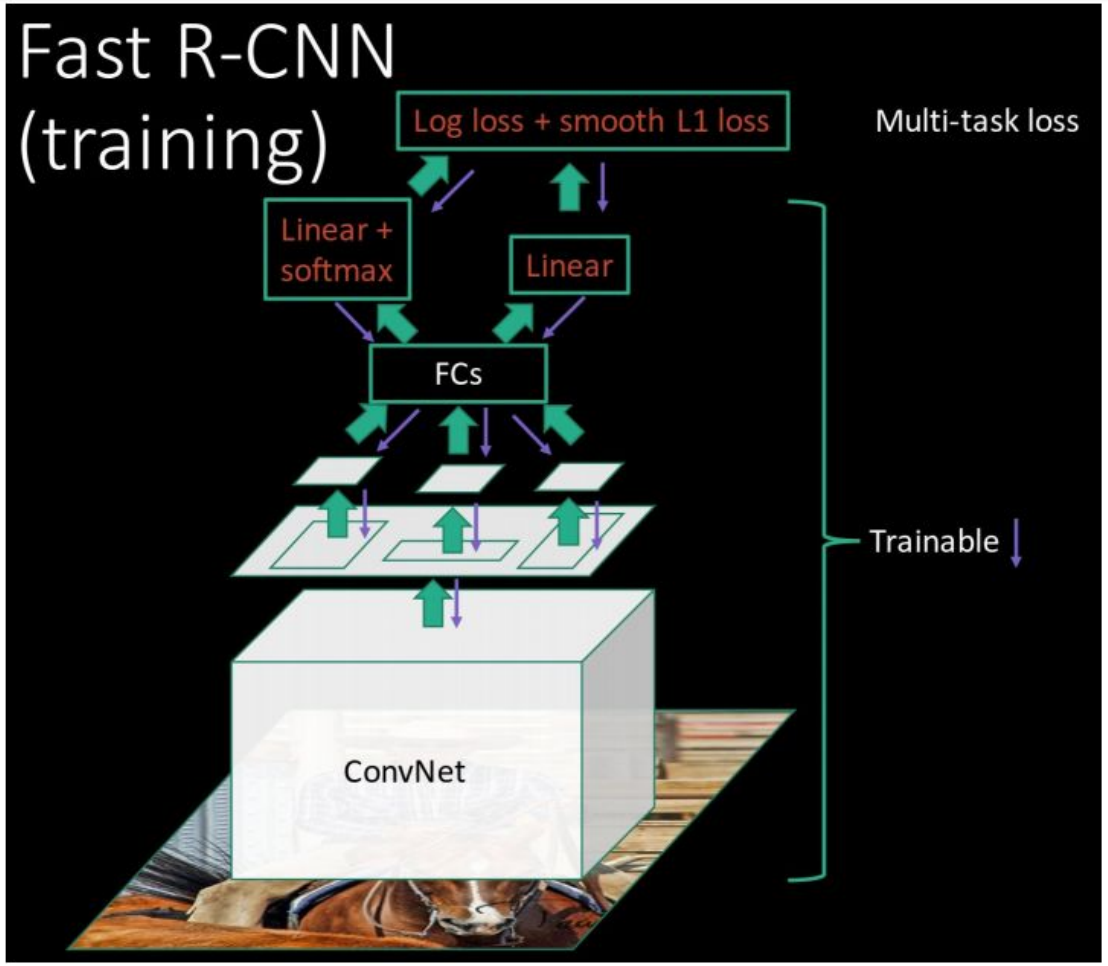


R-CNN Problem #1:
Slow at test-time due to independent forward passes of the CNN

Solution:
Share computation of convolutional layers between proposals for an image

Source: Lecture slides by Fei Fei li, Andrew Karpathy and Justin Johnson.
Ref: R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015.

Fast RCNN



R-CNN Problem #2:
Post-hoc training: CNN not updated in response to final classifiers and regressors

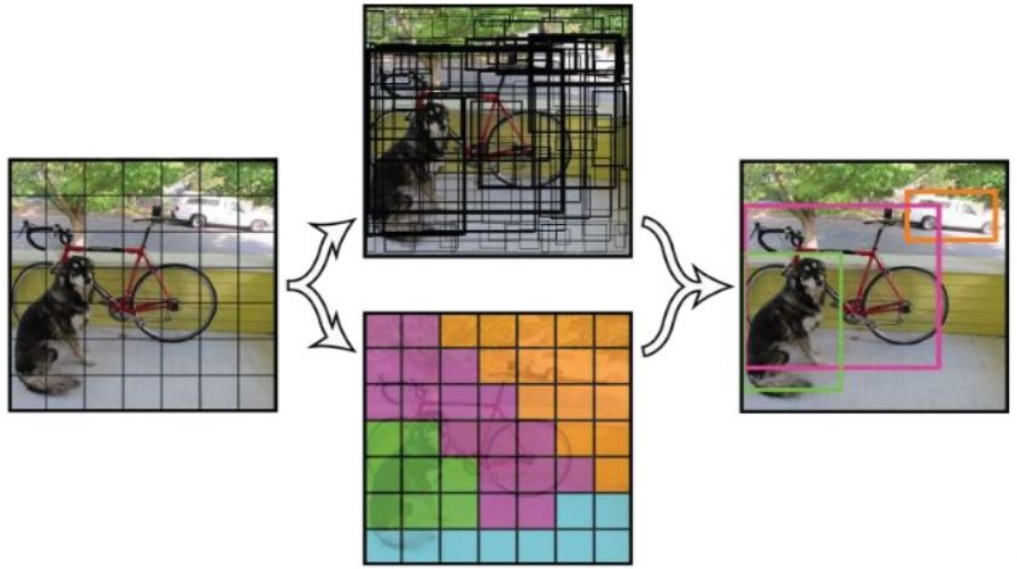
R-CNN Problem #3:
Complex training pipeline

Solution:
Just train the whole system end-to-end all at once!

Source: Lecture slides by Fei Fei li, Andrew Karpathy and Justin Johnson.
Ref: R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015.

YOLO: Real time object detection

- Very fast and accurate system.
- Single regression problem with single convolutional network.
- Prior detection systems repurpose classifiers or localizers to perform detection.
- Applying the model to an image at multiple locations and scales. High scoring regions of the image are considered detections.



J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , Las Vegas, NV, 2016, pp. 779-788.

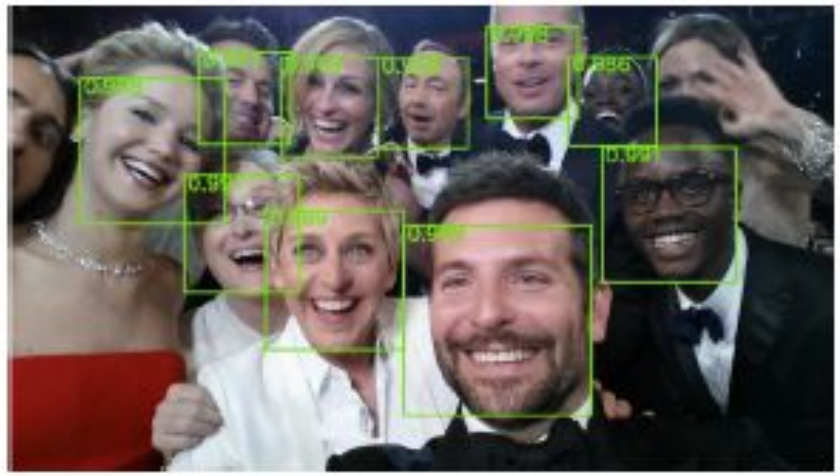
Similarity learning with CNN

Siamese network

Several applications of Similarity Measures exist in today's world.

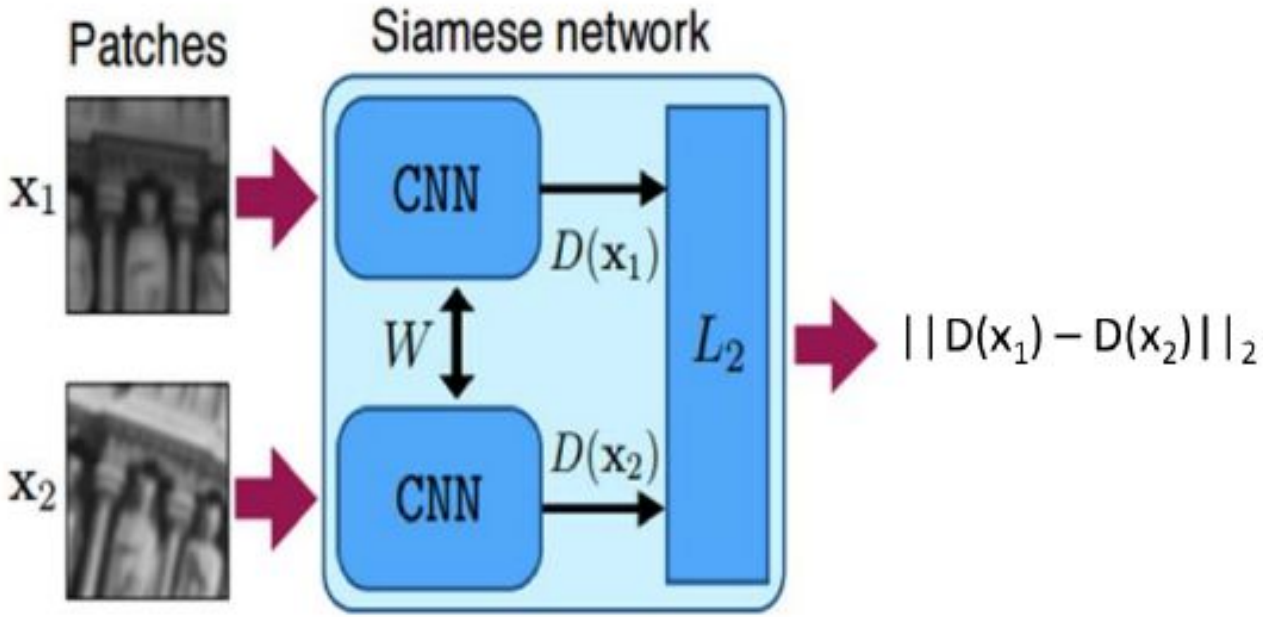
- Handwriting recognition.

- Detection of faces in camera image.



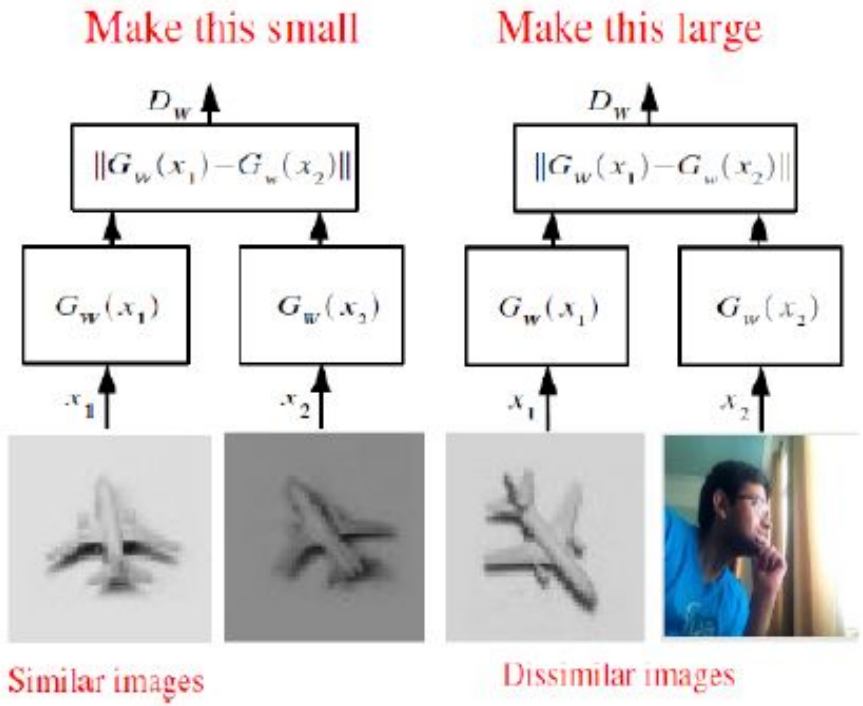
Siamese network

- **Input:** A pair of input images.
- **Output:** A label, 0 for **similar**, 1 **else**.



Slide credits: Mr. Daksh Thapar, IIT Mandi.

Siamese Network: Loss function

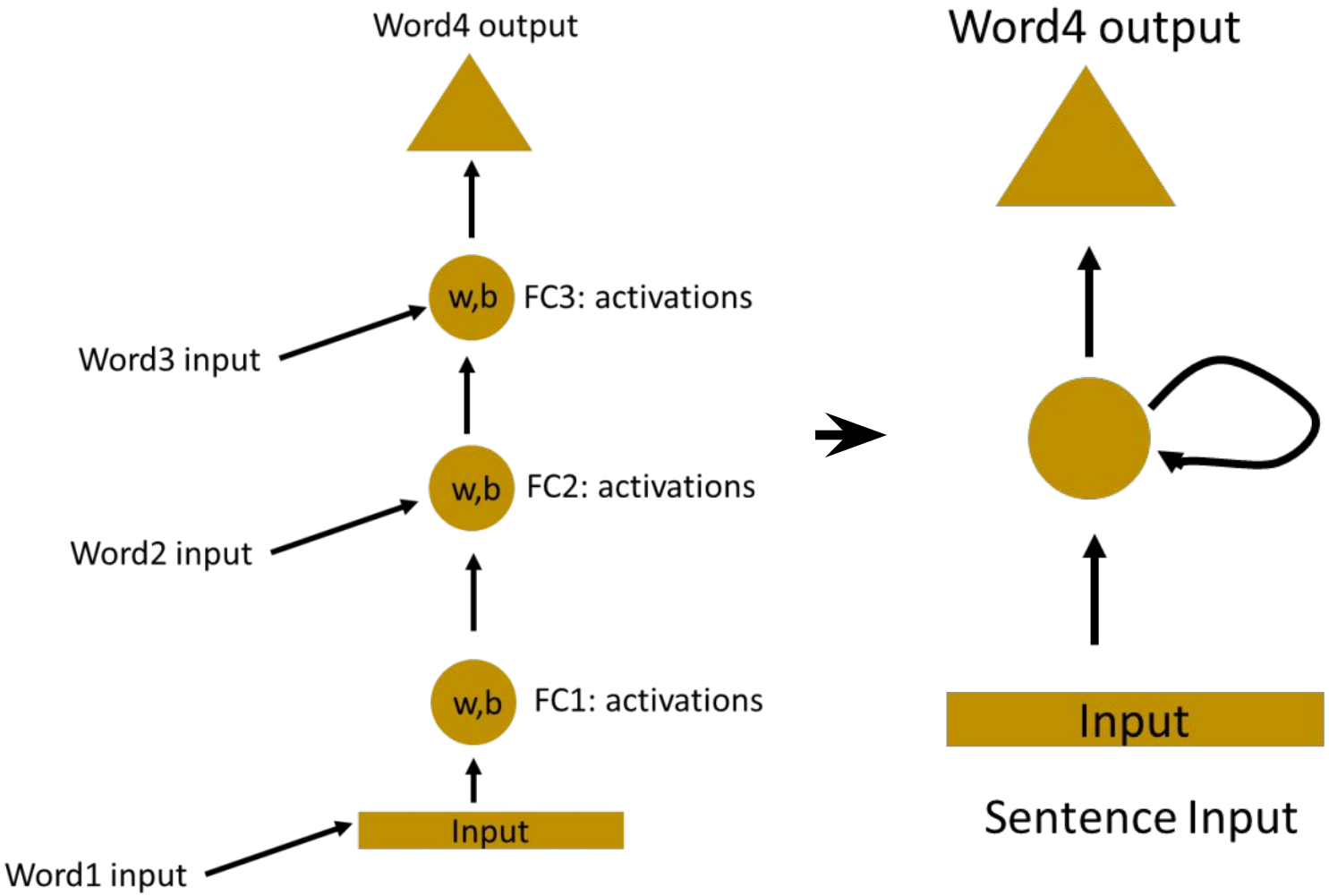


The final loss is defined as:

$$L = \Sigma loss\ of\ \ positive\ pairs + \Sigma loss\ of\ \ negative\ pairs$$

Sequence prediction networks
LSTM and RNN

RNN (Recurrent Neural Network)



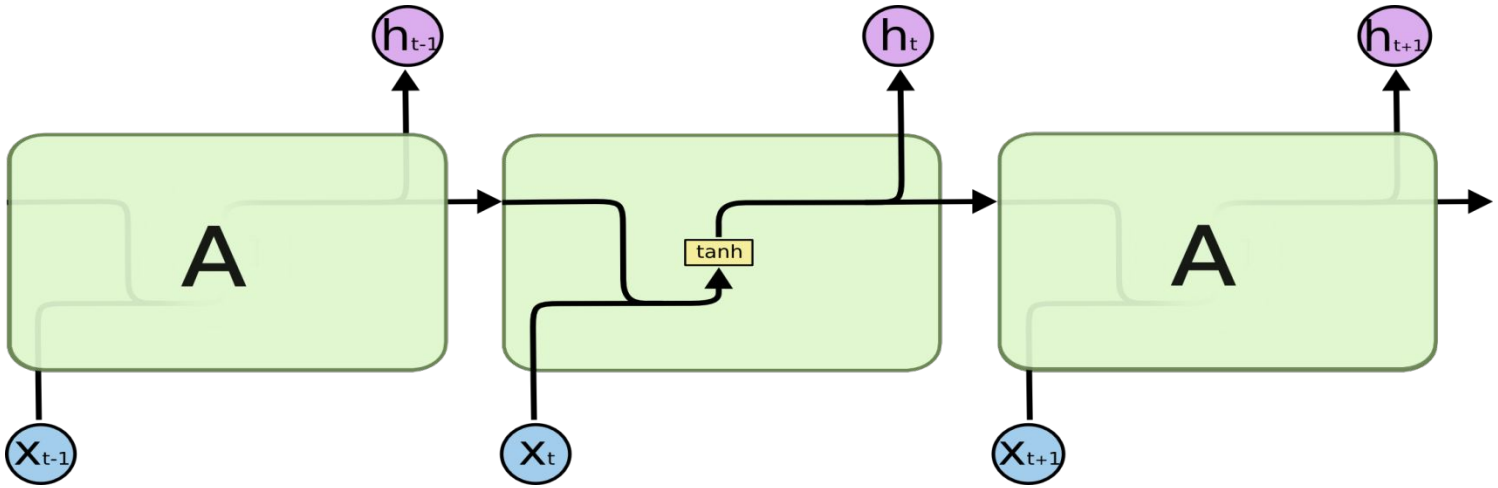
LSTM (Long Short Term Memory)

- RNNs can learn the recent past information but for long term past memory, LSTMs are useful.

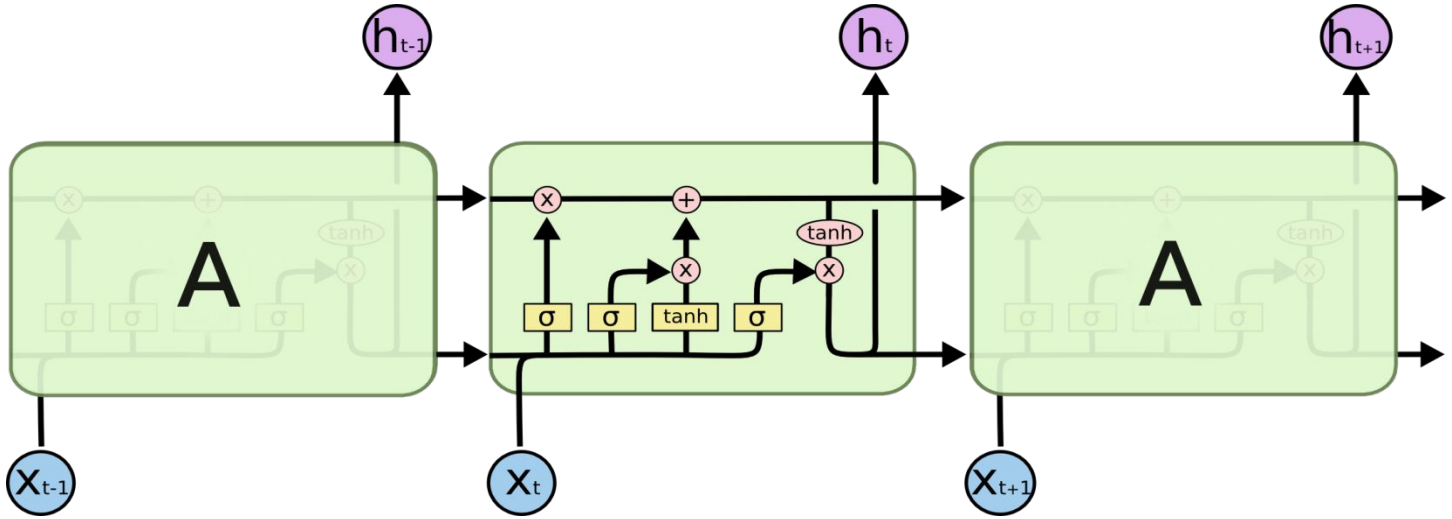
- “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.

- In RNN, the repeating modules of chain have similar structures while in LSTM, the repeating modules have different structures.

J. Li, A. Mohamed, G. Zweig and Y. Gong, "LSTM time and frequency recurrence for automatic speech recognition," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* , Scottsdale, AZ, 2015, pp. 187-191.



RNN architecture



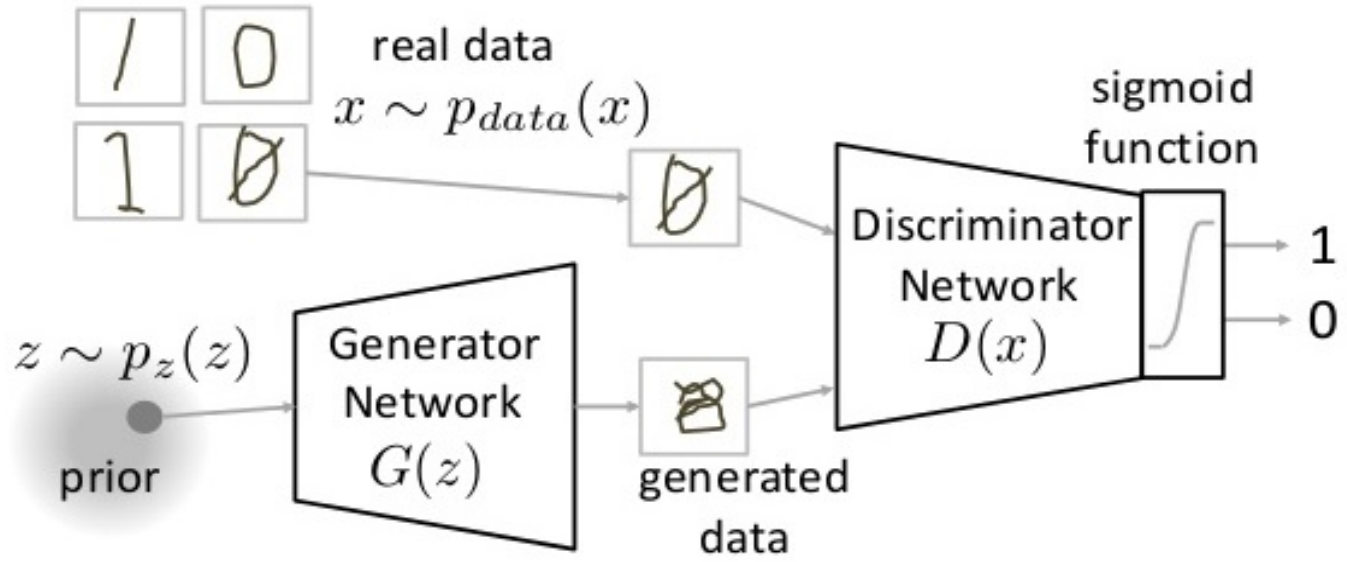
LSTM architecture

Generative networks
(GAN and Conditional GAN)

(GANs), and the variations that are now being proposed is the most interesting idea in the last 10 years in ML, in my opinion..

- Yann Le Cunn

- Generative adversarial networks (GAN) take a collection of points and infer a function that describes the distribution that generated them.



Generative Adversarial Networks (2014) — Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio.

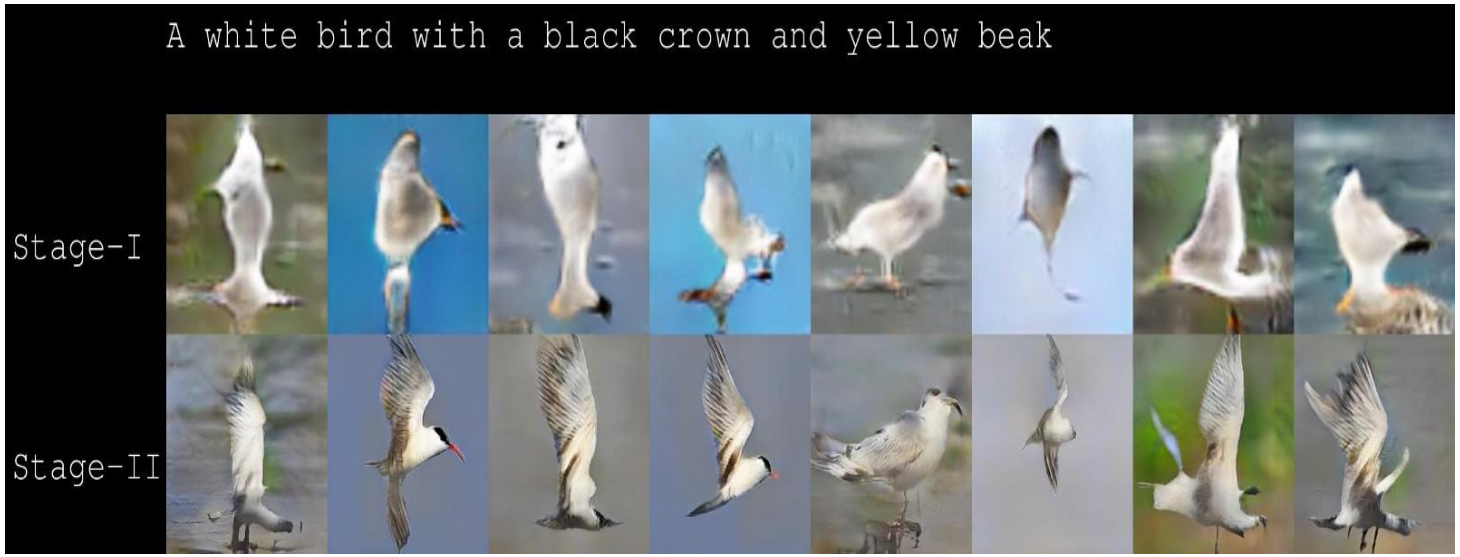
Conditional GANs

- Conditional version of Generative Adversarial Nets (GAN) where both generator and discriminator are conditioned on some data **y** (class label or data from some other modality).
- Feed **y** into both the generator and discriminator as additional input layers such that **y** and input are combined in a joint hidden representation.
- the additional information is constraining the generator (and discriminator) to generate a certain type of output.

Mirza, Mehdi and Simon Osindero. “Conditional Generative Adversarial Nets.” *CoRR*abs/1411.1784 (2014).

GAN (Generative Adversarial Networks)

- GANs are neural networks composed up of two networks competing with each other: generator — to generate data set and discriminator — to validate the data set.
- The goal is to generate data points that are similar to some of the data points in the training set.

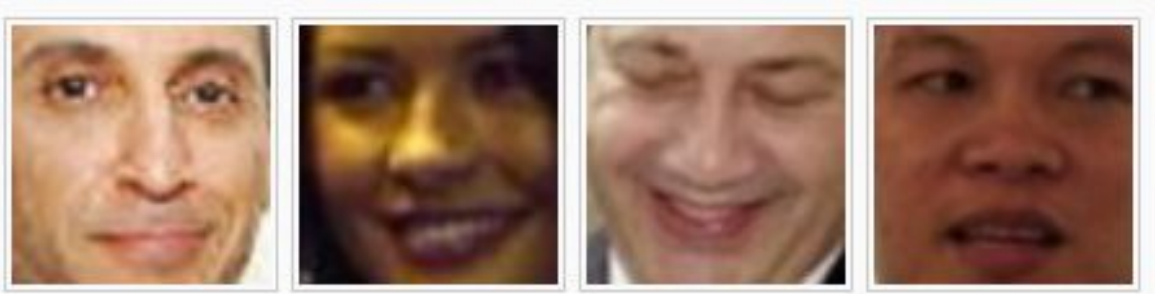


The algorithm was given a text “A white bird with a black crown and yellow beak”. And the GAN was able to generate the image by itself based on the text given.

https://hackernoon.com/generative-adversarial-networks-a-deep-learning-architecture-4253b6d12347

Conditional GANs: An example

Training images:



we build a density model p(x | y), where y is the “conditional information”.

Now the question for predictor is:

1. Is this an x given y? (where y can be the age, emotion or race)
2. If we’re looking for faces of type , should we accept x as a good example or not?

Samples of some generated faces, with a condition of smiling faces.

Ref: Gauthier, J. (2015). Conditional generative adversarial nets for convolutional face generation.

A refresher…

- Clever Representations:
  Hierarchy and dependencies of non-linear mappings
  (structure within a network)
- Task based compositions
  (composite structures with multiple types of networks)
- Global learning objectives (cost function)
  (for individual networks, or for the composite networks)
- Elegant algorithms for cost optimization (or parameter estimation)
- Transfer learning

So, what more ?…. A lot more…

- Multi-task learning, Multi-modal learning
  (Can a machine recognize objects, translate languages, play games, write blogs and generate music ?)

- Unsupervised frameworks, Very less data,
  Active learning, Relevance feedback learning, Reinforcement learning

- Adversarial examples
  (Really stupid mistakes)

- **What are the networks learning ?**
  (A deep question)

A deeper and a more important question

What are WE learning ?