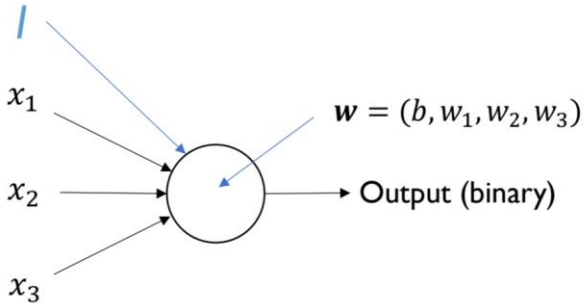
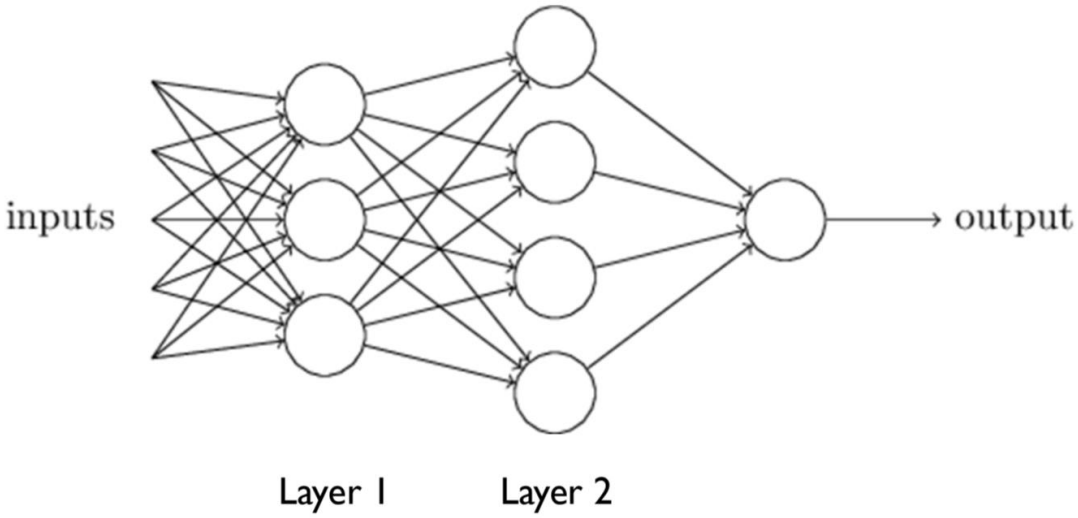
# Deep Learning for Computer Vision

## Perceptron - single neuron




$$\text{output} = \begin{cases} 0 & \text{if } w \cdot x \leq 0 \\ 1 & \text{if } w \cdot x > 0 \end{cases} \qquad w \cdot x \equiv \sum_j w_j x_j$$

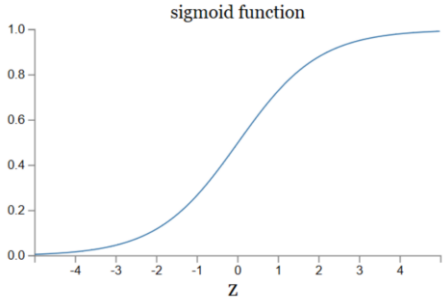
## Multi-Layer Perceptron



# Nonlinearity – activation function

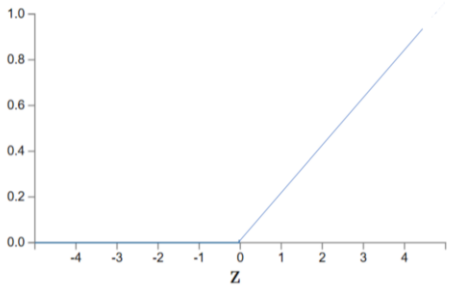
$$\sigma(w \cdot x + b)$$

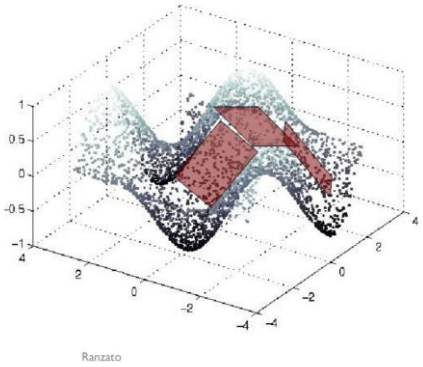
$$\sigma(z) \equiv \frac{1}{1+e^{-z}}.$$



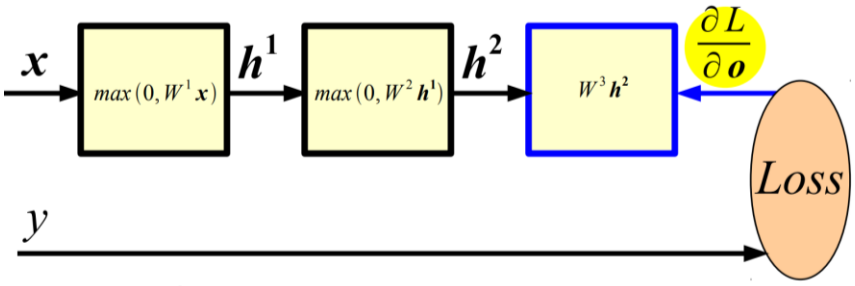

# ReLU

$$f(x) = \max(0, x)$$




# Learning – Backpropagation Algorithm

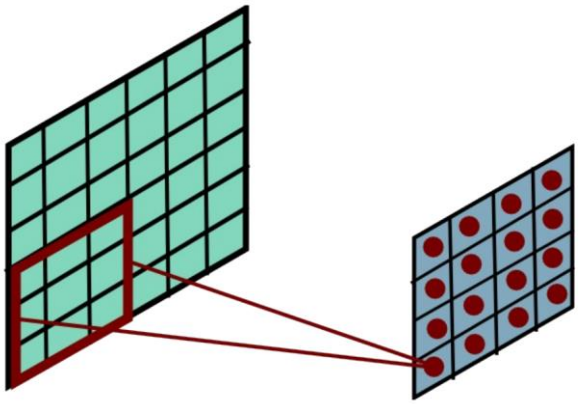



Given $\partial L / \partial \boldsymbol{o}$ and assuming we can easily compute the Jacobian of each module, we have:

$$\frac{\partial L}{\partial W^3} = \frac{\partial L}{\partial \boldsymbol{o}} \frac{\partial \boldsymbol{o}}{\partial W^3} \qquad \frac{\partial L}{\partial \boldsymbol{h}^2} = \frac{\partial L}{\partial \boldsymbol{o}} \frac{\partial \boldsymbol{o}}{\partial \boldsymbol{h}^2}$$

$$\frac{\partial L}{\partial W^3} = (p(c|\boldsymbol{x}) - \boldsymbol{y})\ \boldsymbol{h}^{2\,T} \qquad \frac{\partial L}{\partial \boldsymbol{h}^2} = W^{3\,T}(p(c|\boldsymbol{x}) - \boldsymbol{y})$$

# Convolutional Neural Nets (CNNs)

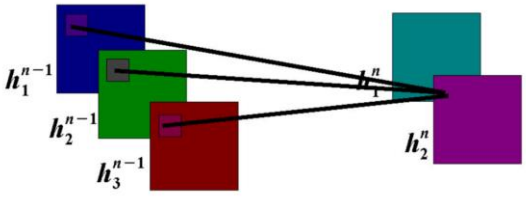


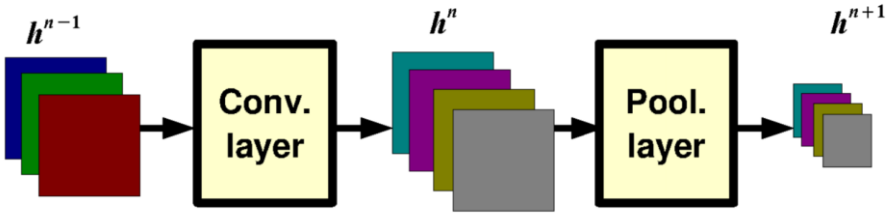
$$h_j^n = max\left(0, \sum_{k=1}^{K} h_k^{n-1} * w_{kj}^n\right)$$

output feature map

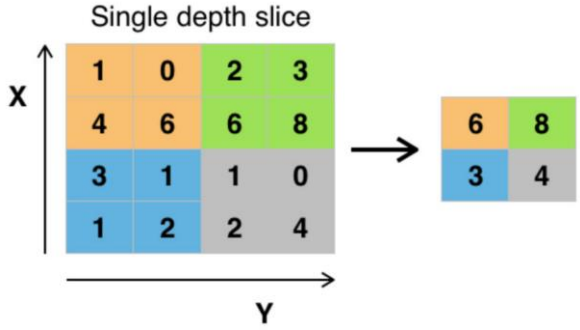
input feature map

kernel



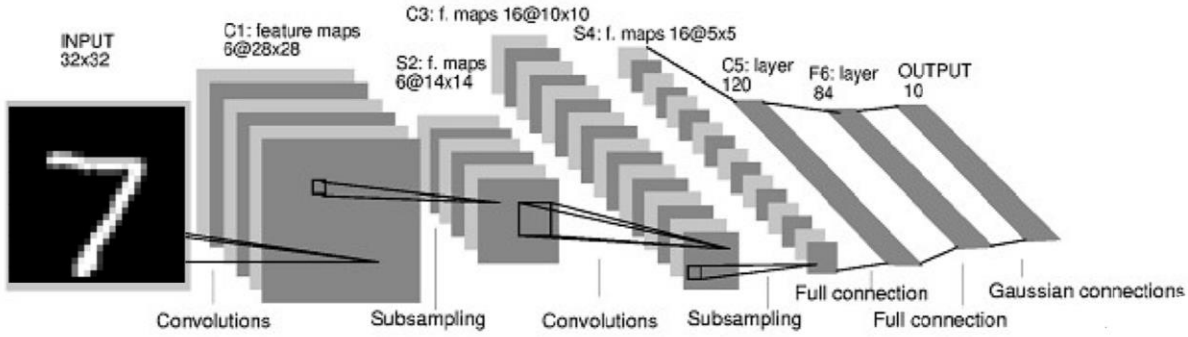




# Pooling Layers - downsampling

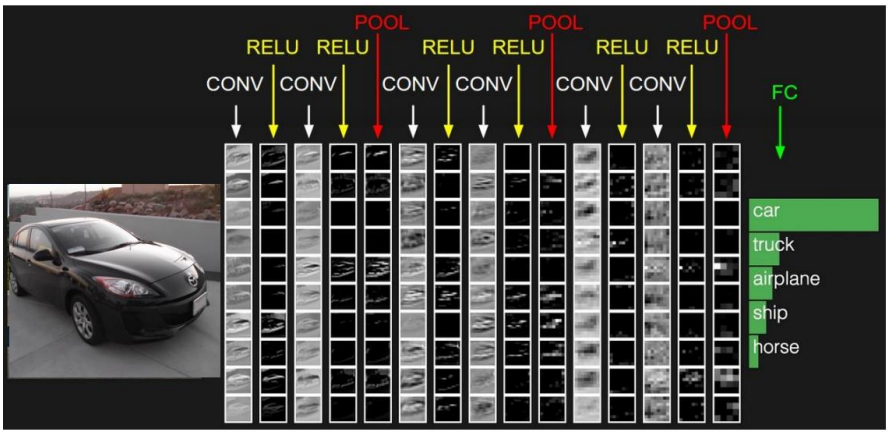



(max pooling)

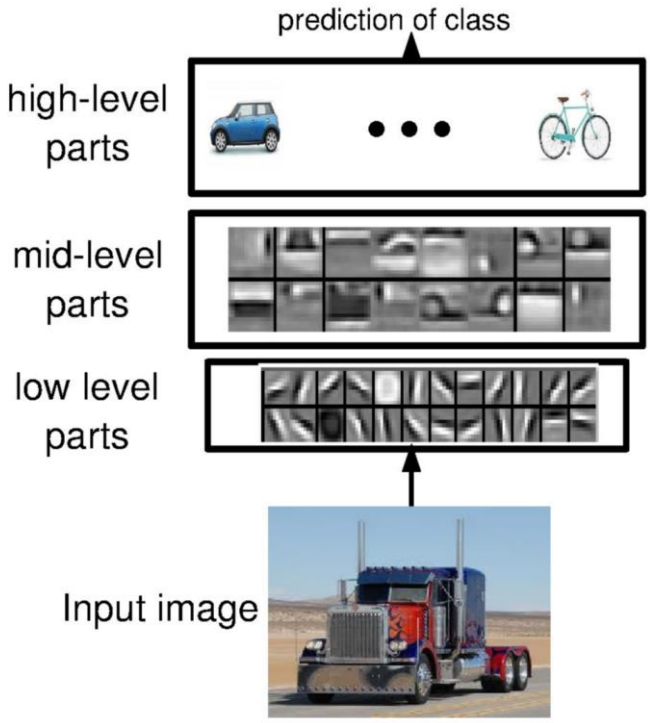
# Yann LeCun's MNIST Architecture



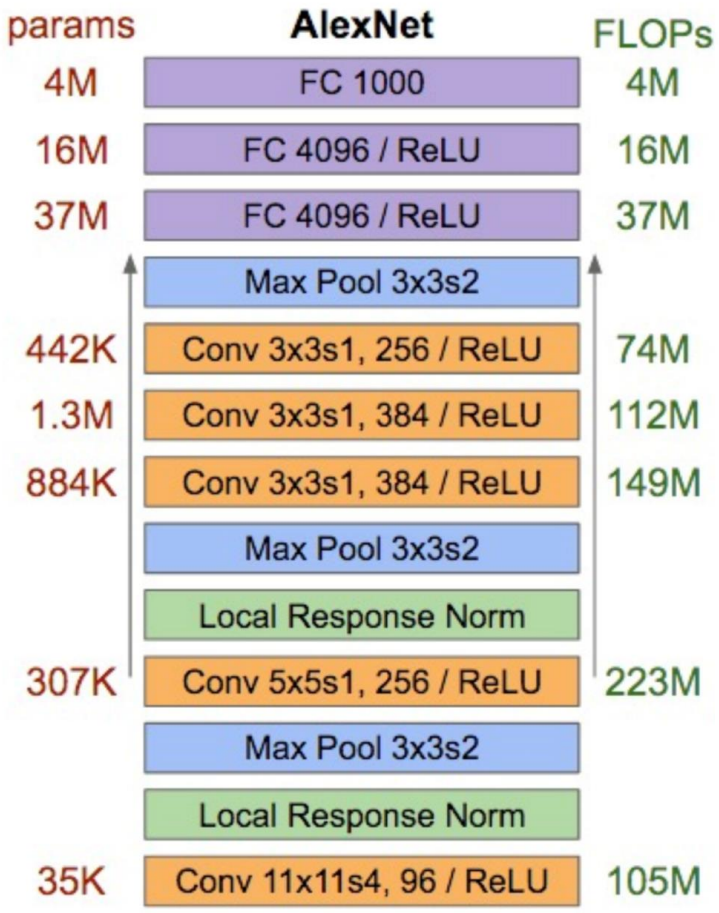

# Layering

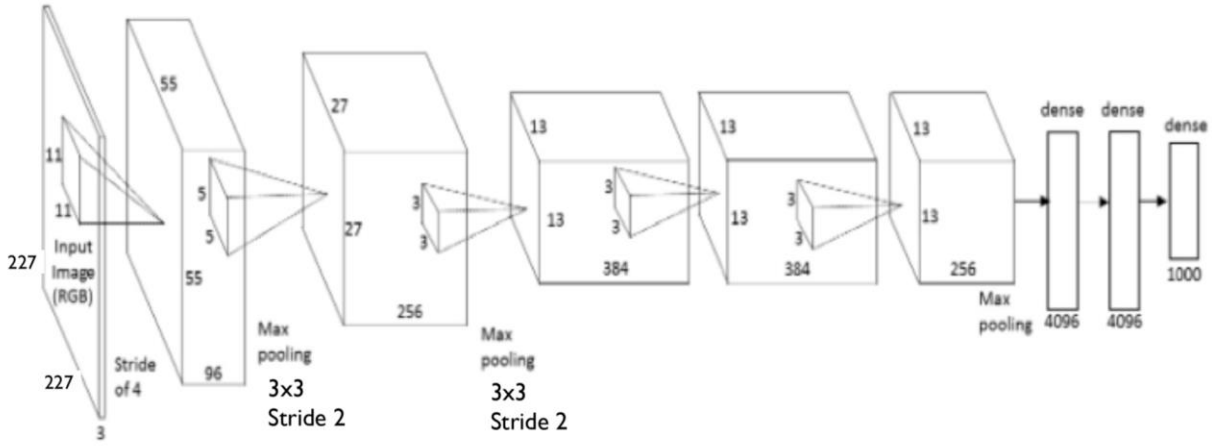



# Hierarchical Feature Learning



# AlexNet (Krizhevsky et al. 2012)




**Conv 1**
11 x 11 x 3
Stride 4
96 filters

**Conv 2**
5 x 5 x 96
Stride 1
256 filters

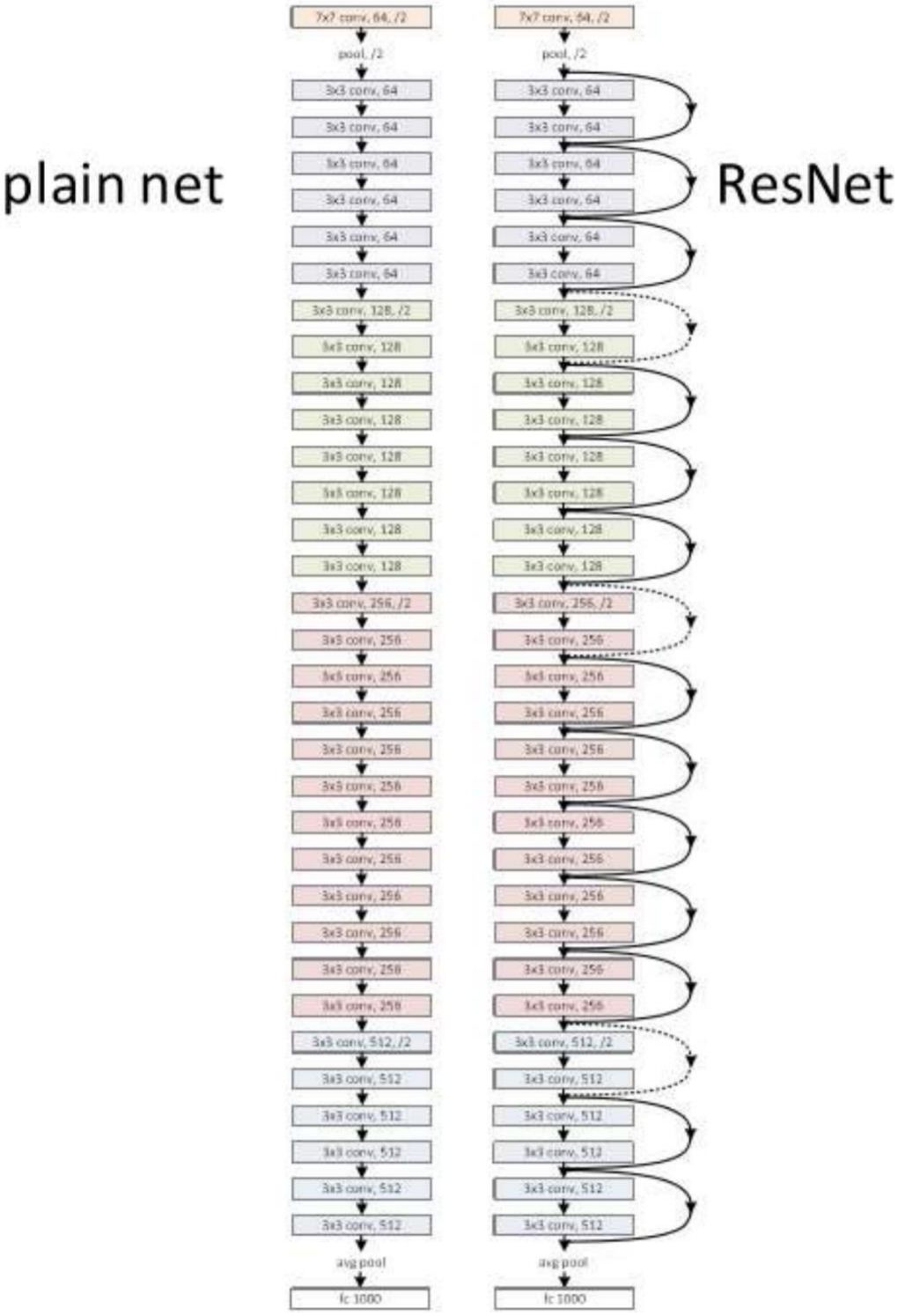
**Conv 3**
3 x 3 x 256
Stride 1
384 filters

**Conv 4**
3 x 3 x 192
Stride 1
384 filters

**Conv 4**
3 x 3 x 192
Stride 1
256 filters

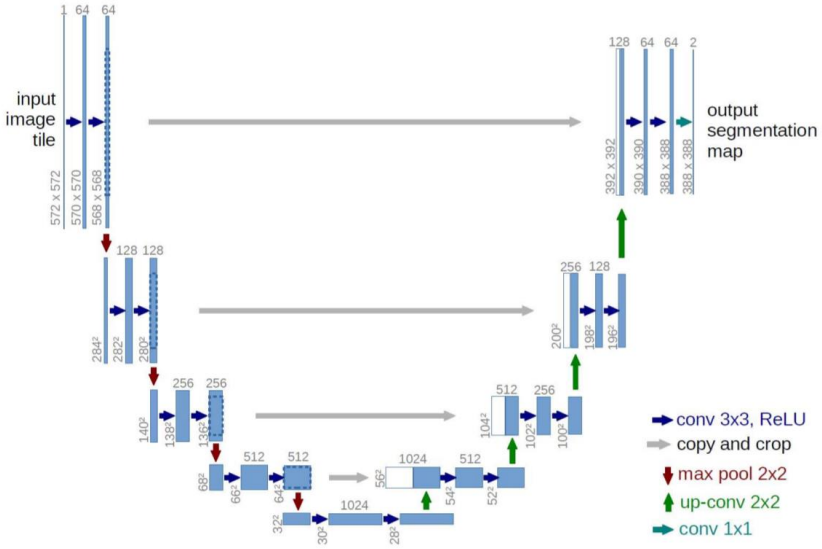
[Krizhevsky et al. 2012]

# ResNet (He et al., 2015)

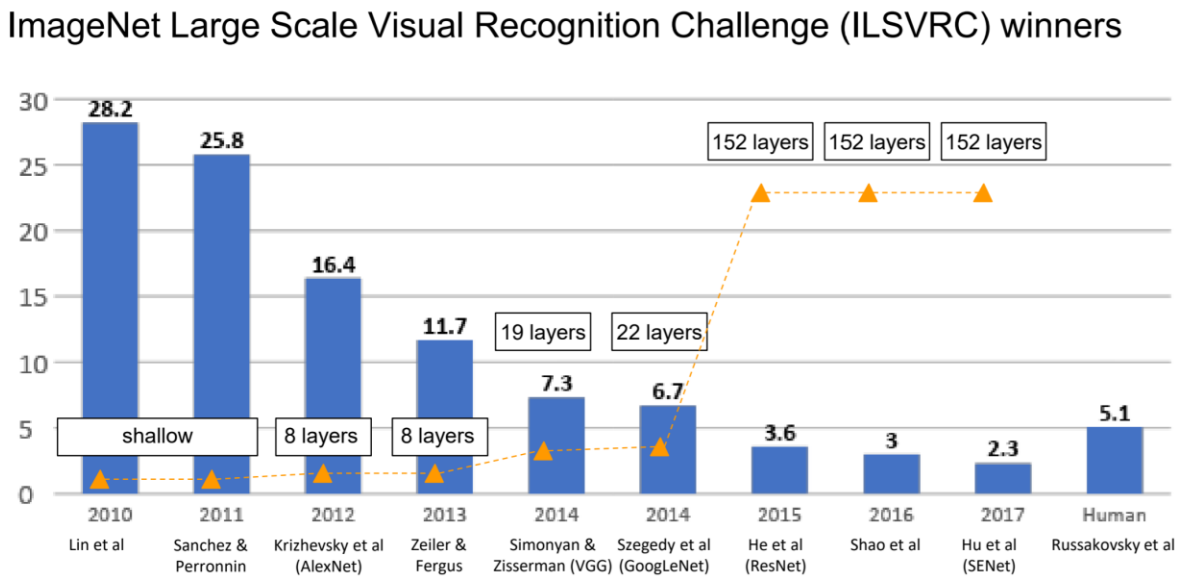


# VGG Net (Simonyan and Zisserman, 2014)

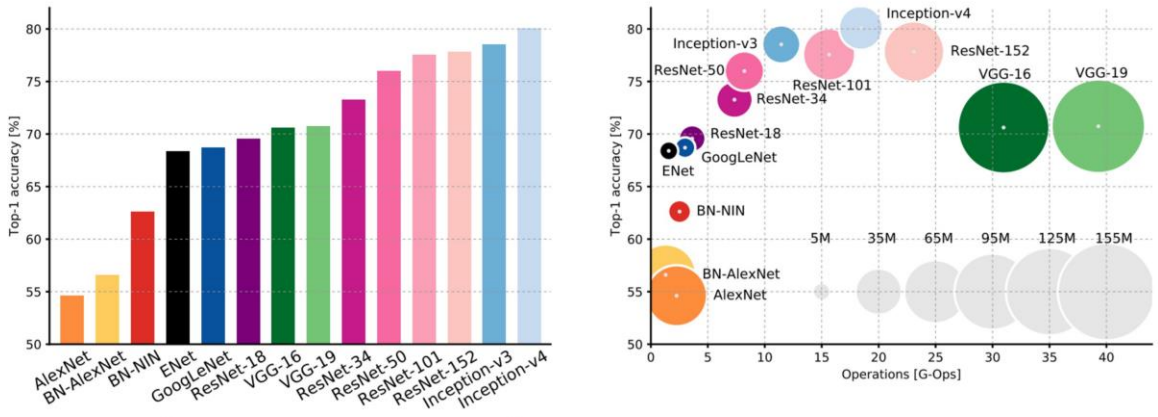

# UNets



# Performance and Model Complexity




## Comparing complexity...




An Analysis of Deep Neural Network Models for Practical Applications, 2017.

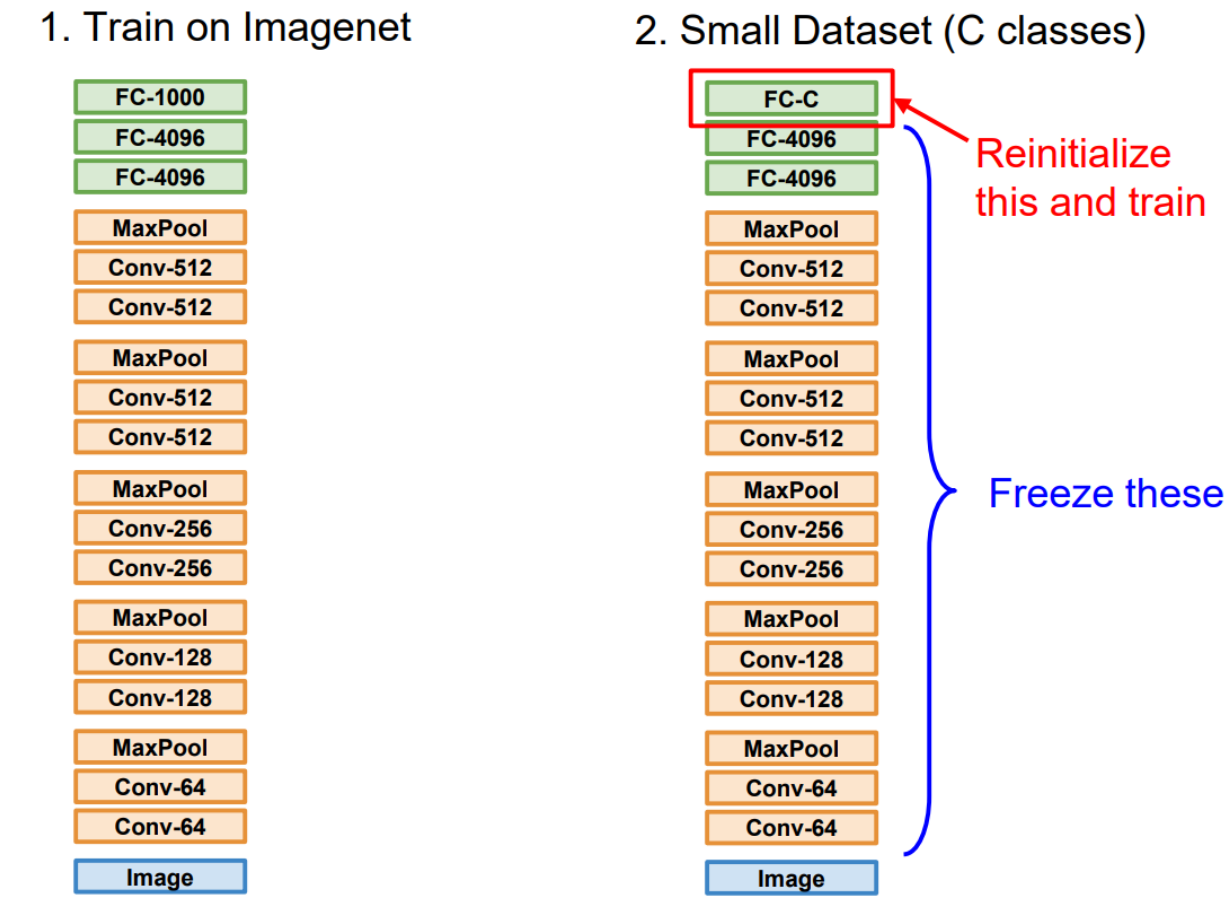
# ImageNet Dataset

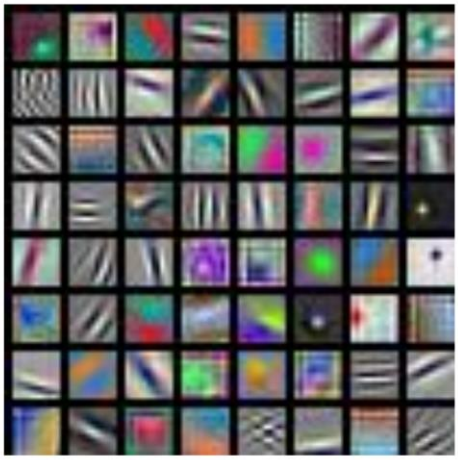
ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection
1-K has 1000 object classes, 1,281,167 training, 50,000 validation, 100,000 test images.
22-K has 21841 classes, 14,197,122 images

# CIFAR-10

60,000 32x32 color images, 10 classes

# Transfer Learning

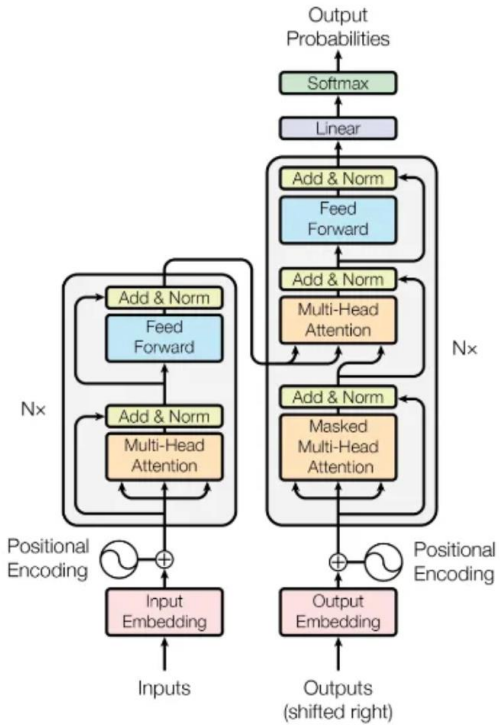


AlexNet:
64 x 3 x 11 x 11

(learned features, first stage)

# The Transformer Architecture




Transformer Architecture

Attention Is All You Need, Vaswani et al. (2017)

# ViTs – Vision Transformers

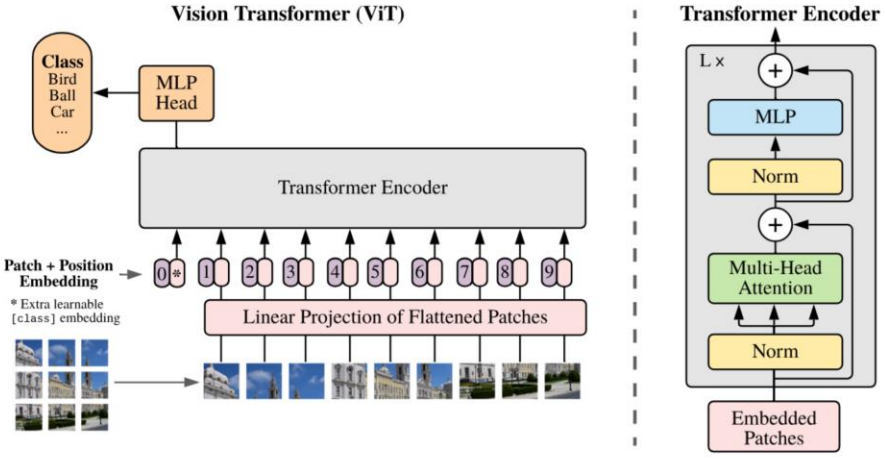
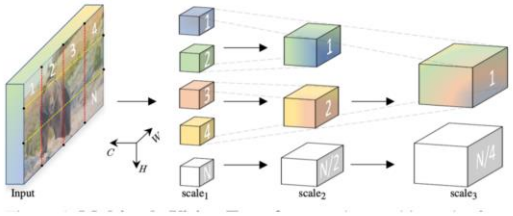
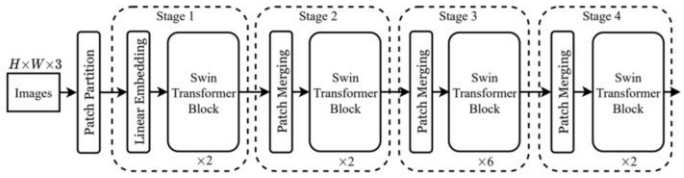



Figure from:
Dosovitskiy et al, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ArXiv 2020

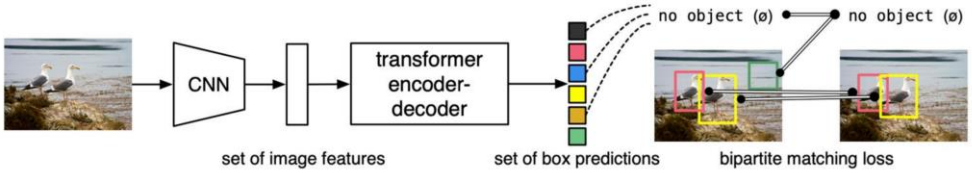
# Vision Transformers




Fan et al, “Multiscale Vision Transformers”, ICCV 2021



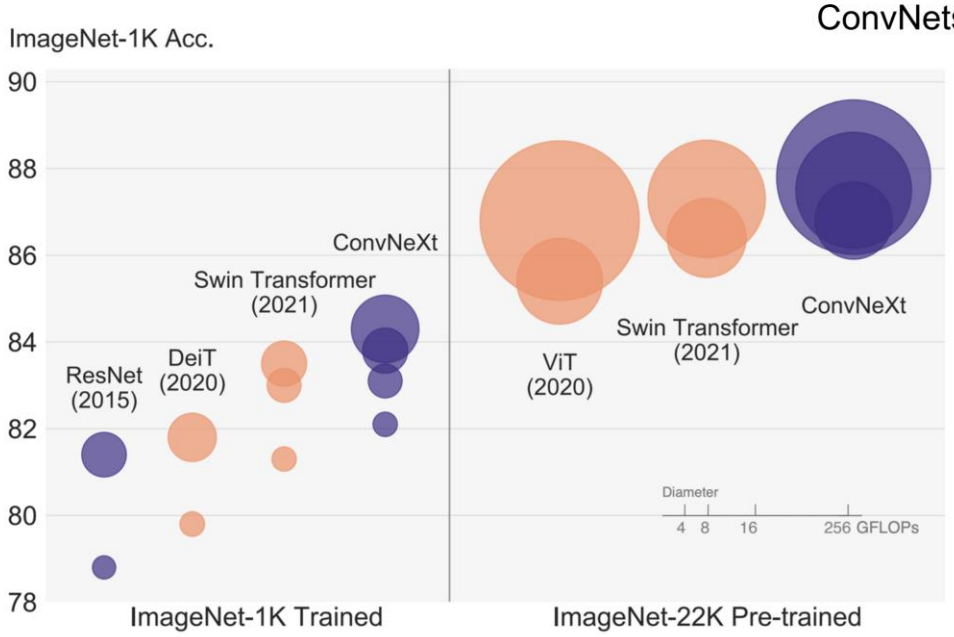

Liu et al, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, CVPR 2021



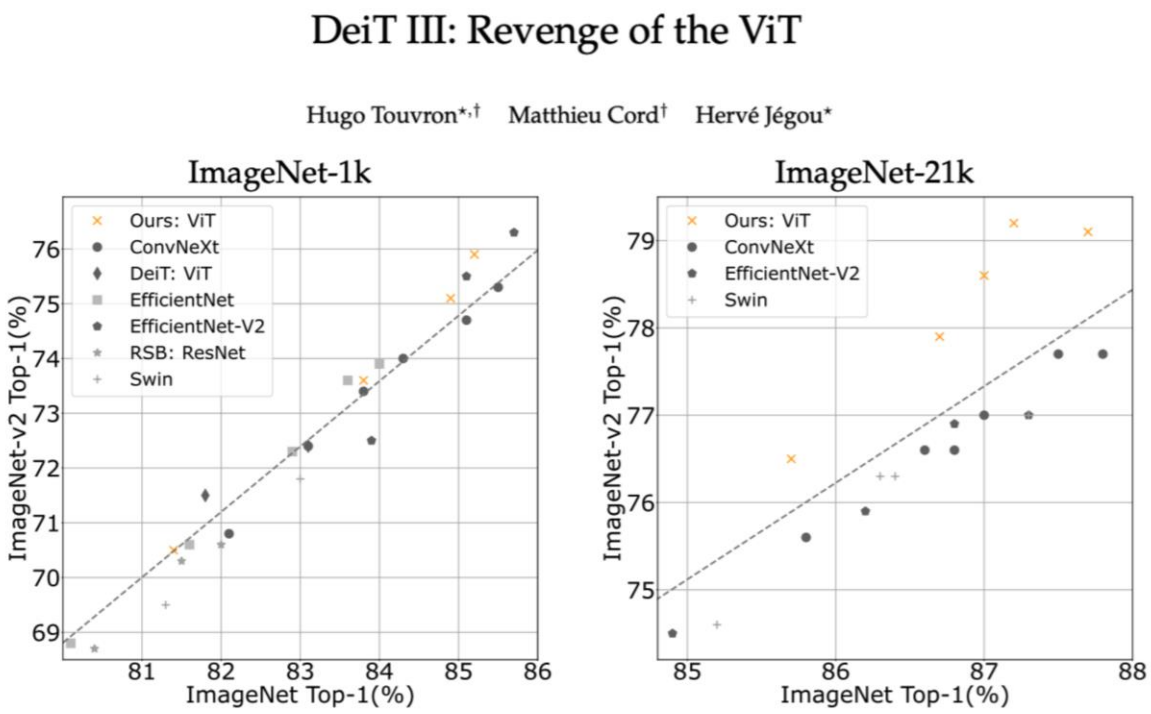

Carion et al, “End-to-End Object Detection with Transformers”, ECCV 2020

# Performance

ConvNets strike back!




A ConvNet for the 2020s. Liu et al. CVPR 2022




# Resources

Wikipedia

Deep Learning for Vision Lecture, Marc'Aurelio Ranzato – DeepMind

Deep Learning for CV Course Notes, Fei-Fei Li & Ehsan Adeli -– Stanford Univ.

Attention Is All You Need, Vaswani et al., Advances in Neural Information Processing Systems (2017).

Computer Vision Course Notes, Srinath Sridhar – Brown Univ.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) website.

Imagenet classification with deep convolutional neural networks, A Krizhevsky, I Sutskever, GE Hinton, Advances in neural information processing systems (2012).

Deep Residual Learning for Image Recognition, Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015).