Image Classification Networks: classical architectures and common design patterns

Antonia Marcu & Jonathon Hare

Vision, Learning and Control
University of Southampton

Motivation: Image Classification

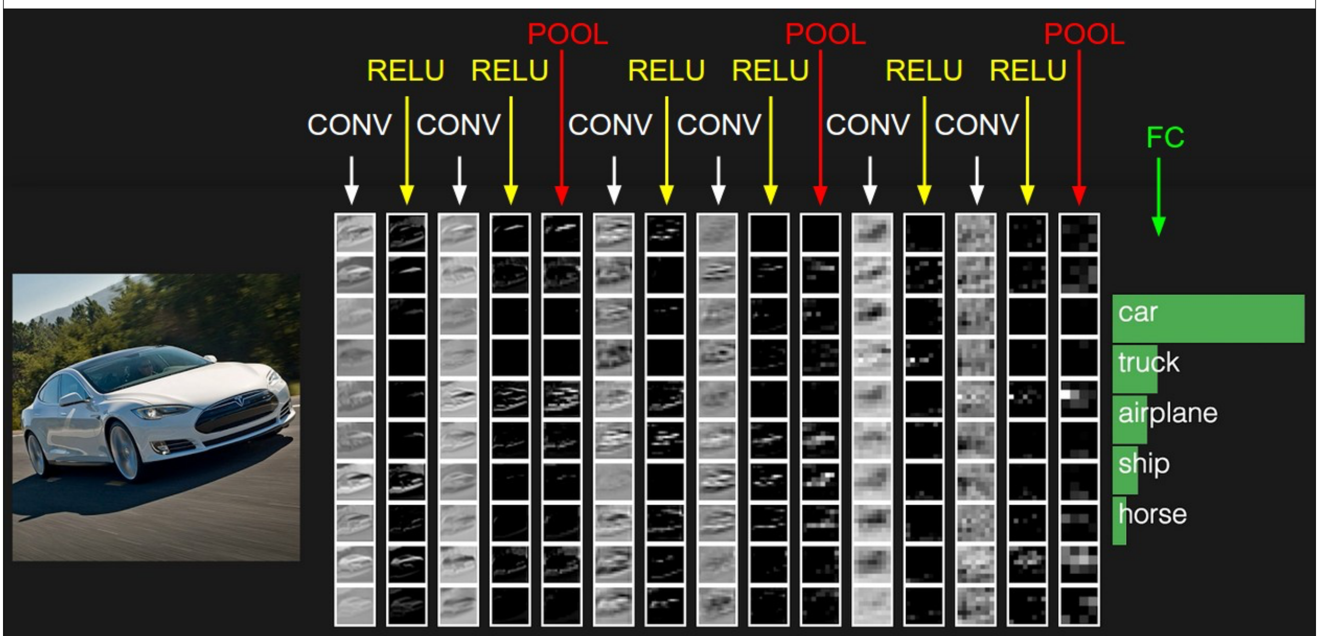
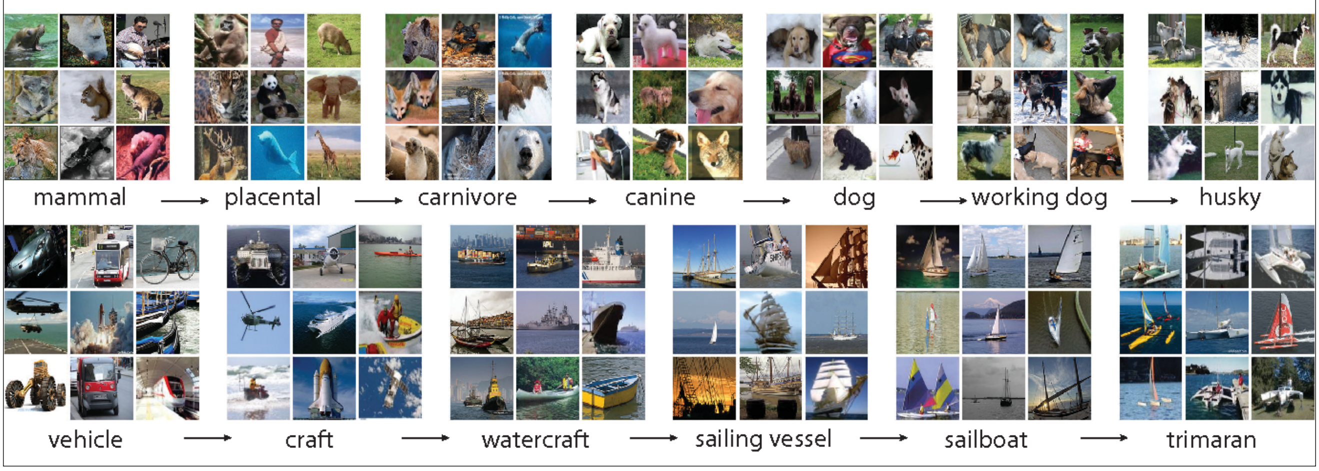


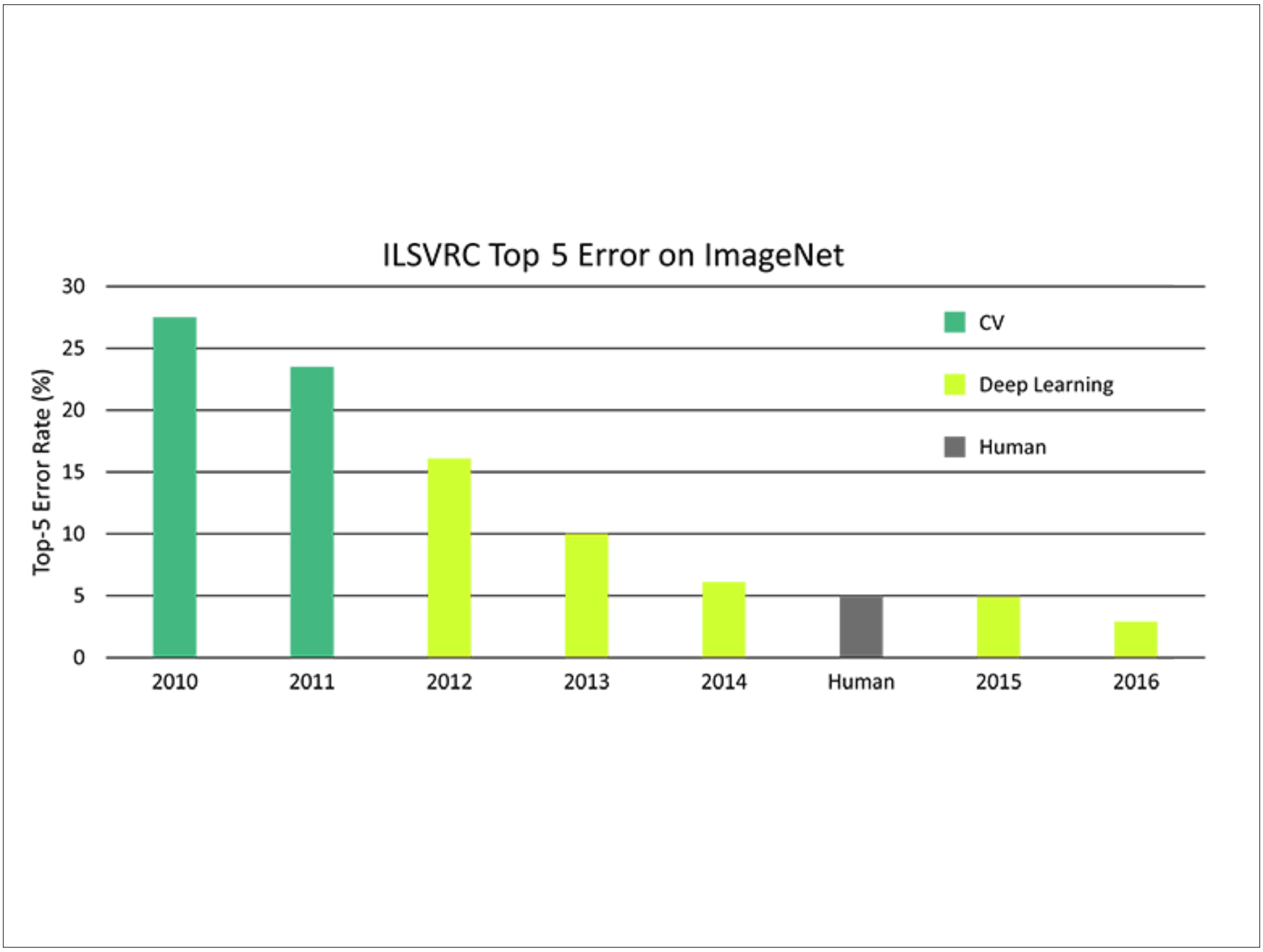
Image Classification Competitions

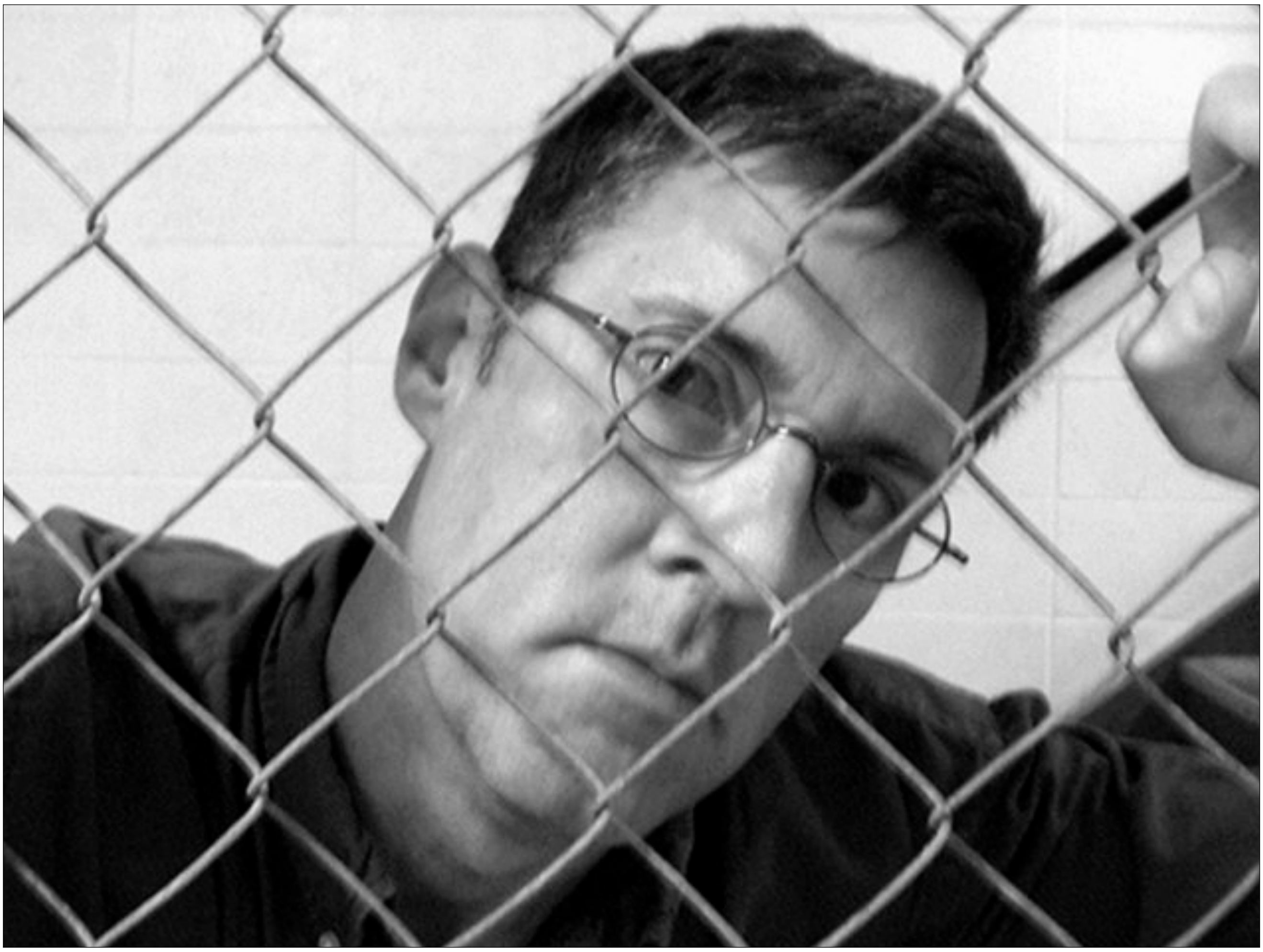
- Corel Dataset, early 2000's
    - Annotation/multi-label classification
    - ~4500/500 small images
- PASCAL VOC Challenges ~2007
    - Object detection and classification

The ImageNet Challenge

- Circa 2009/2010
- ILSVRC Challenge Dataset: 1.3 Million Images in 1000 classes from a larger superset




ILSVRC Top 5 Error on ImageNet
Top-5 Error Rate (%)
30
25
20
15
10
5
0
2010
2011
2012
2013
2014
Human
2015
2016
CV
Deep Learning
Human

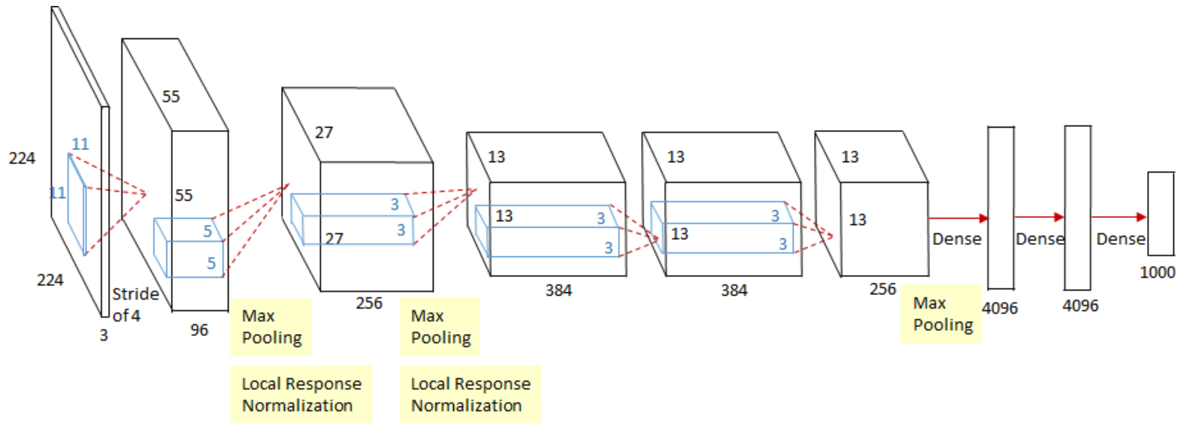


Classic Architectures

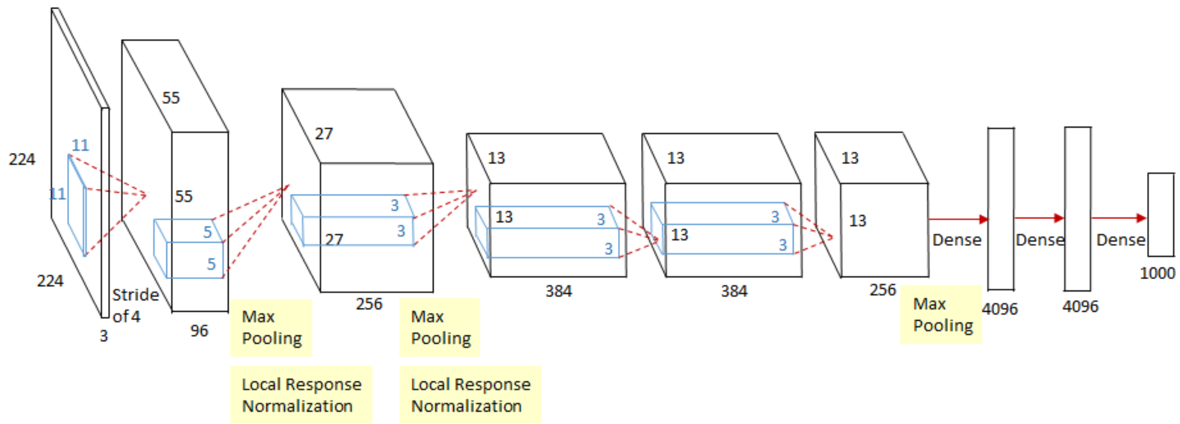
AlexNet

ImageNet Classification with Deep Convolutional Neural Networks. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

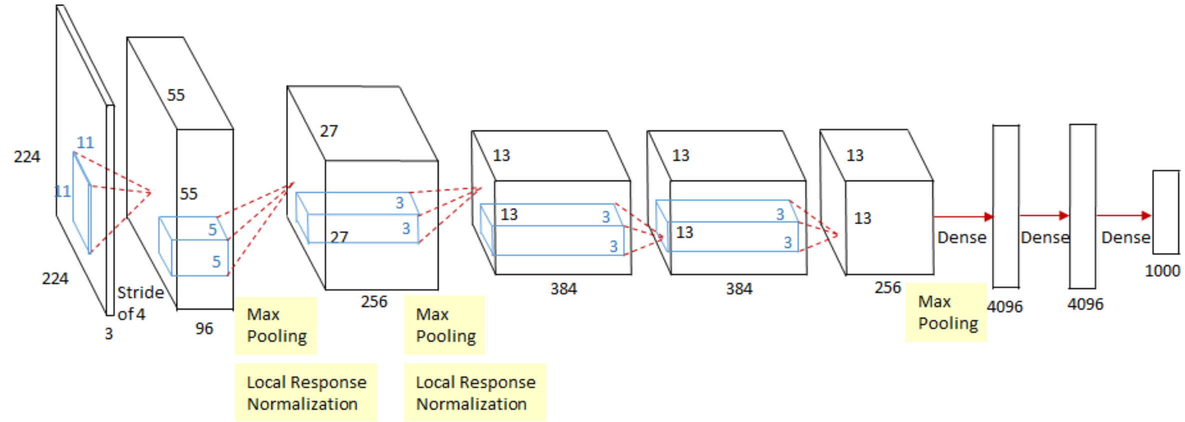
CaffeNet



CaffeNet - How many parameters?



CaffeNet - Effective Receptive Field Sizes



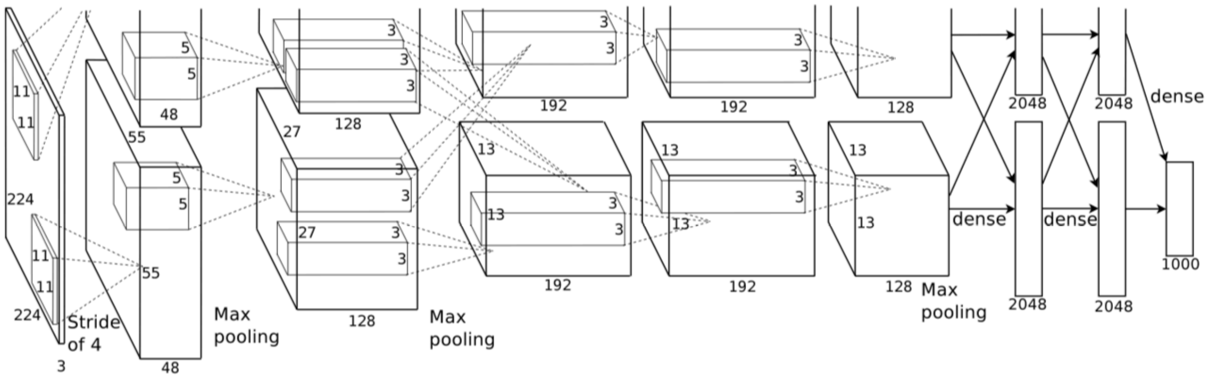


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

LRN Layers

- The original AlexNet (and the VGG & GoogleLetNet) contained networks “Local Response Normalisation” layers
  - The motivation was to provide locally higher contrast in feature maps

The All CNN

Striving for Simplicity: The All Convolutional Net. https://arxiv.org/pdf/1412.6806.pdf

Model

A	B	C
Input 32×32 RGB image		
5×5 conv. 96 ReLU	5×5 conv. 96 ReLU 1×1 conv. 96 ReLU	3×3 conv. 96 ReLU 3×3 conv. 96 ReLU
3×3 max-pooling stride 2		
5×5 conv. 192 ReLU	5×5 conv. 192 ReLU 1×1 conv. 192 ReLU	3×3 conv. 192 ReLU 3×3 conv. 192 ReLU
3×3 max-pooling stride 2		
3×3 conv. 192 ReLU		
1×1 conv. 192 ReLU		
1×1 conv. 10 ReLU		
global averaging over 6×6 spatial dimensions		
10 or 100-way softmax		

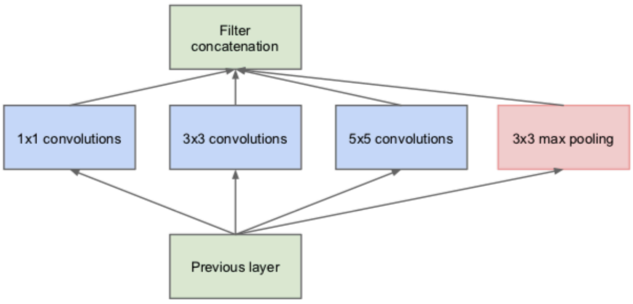
The VGG Networks

Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/pdf/1409.1556.pdf

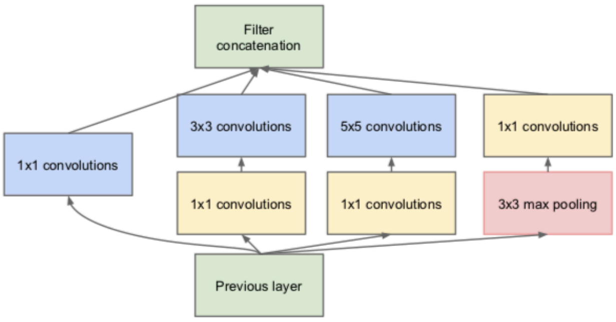
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 $\times$ 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

GoogLeNet and the Inception Module

Going Deeper with Convolutions. https://arxiv.org/pdf/1409.4842



(a) Inception module, naïve version

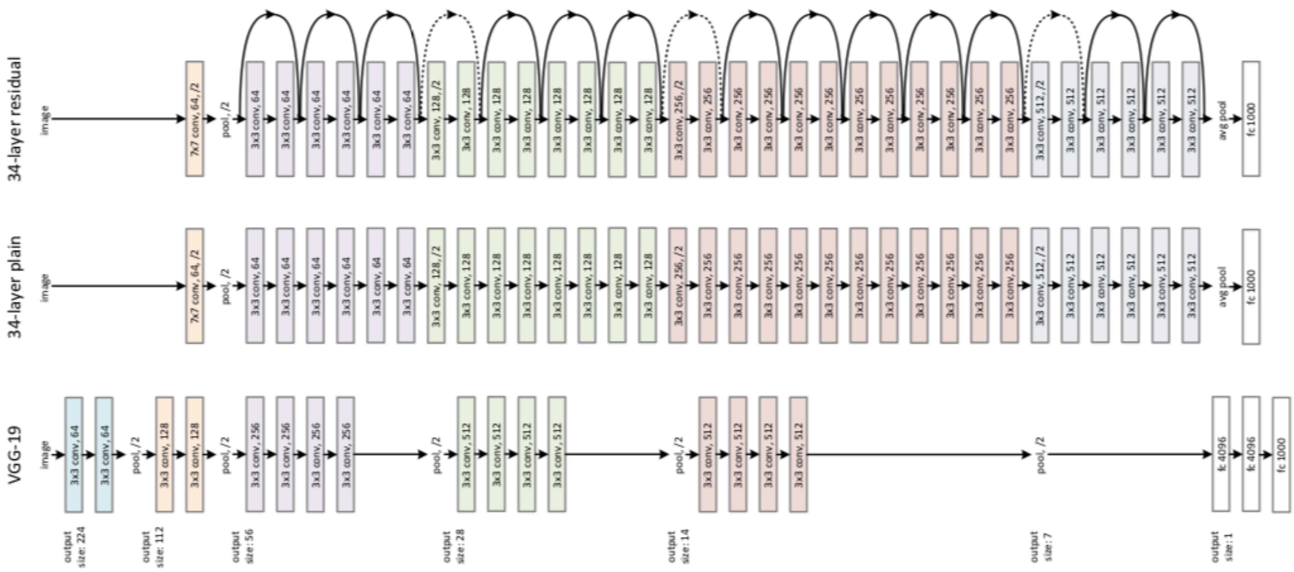


(b) Inception module with dimension reductions

58.9M params

Deep Residual Networks / ResNet

Deep Residual Learning for Image Recognition. https://arxiv.org/pdf/1512.03385.pdf



Do deeper ResNets get you better performance?

method	# layers	# params	error (%)
Maxout [10]			9.38
NIN [25]			8.81
DSN [24]			8.22
FitNet [35]	19	2.5M	8.39
Highway [42, 43]	19	2.3M	7.54 (7.72±0.16)
Highway [42, 43]	32	1.25M	8.80
ResNet	20	0.27M	8.75
ResNet	32	0.46M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	**6.43** (6.61±0.16)
ResNet	1202	19.4M	7.93

Table 6. Classification error on the **CIFAR-10** test set. All methods are with data augmentation. For ResNet-110, we run it 5 times and show "best (mean±std)" as in [43].

Are ResNets really deep?

Training Classification Networks

Overfitting is a serious concern

- Early nets used dropout extensively
  - BatchNorm has replaced this in more recent architectures
- Significant amounts of data augmentation (the original AlexNet had 2048 augmentations for each training image!)

Competing in ImageNet

- Almost all the winners use a form of test-time augmentation
    - Take multiple views of the input image (e.g. AlexNet took 10 augmentations) and average over the classifications.