Deep learning and Artificial Neural Network for Object recognition

Dr. Anthony Fleury - HDR
IMT Nord Europe

Dr. Sebastien Ambellouis
Ingénieur de recherche Université Gustave Eiffel
Associé à IMT Nord Europe

sebastien.ambellouis@univ-eiffel.fr
anthony.fleury@imt-nord-europe.fr

The perceptron

- Introduction
- What is classification ?
- The ImageNet large scale recognition challenge (ILSVRC)
- Advance up to now ...
- After ?

The image classification problem

$f($  $) = \text{“cat”}$

$f($  $) = \text{“horse”}$

The image classification problem

$$f(\ \ ) = \text{“cat”}$$

$$f : \mathbb{R}^{224 \times 224 \times 3} \rightarrow C$$

$$C = \{\text{dog}, \text{cat}, \text{horse}, \text{airplane}, \text{tree}\}$$

Multi-class classification

Assumption: classes are **complete** and **mutually exclusive.**

K classes. Want to predict **probability** of each class.

$$\mathbf{p} \in [0,1]^K$$

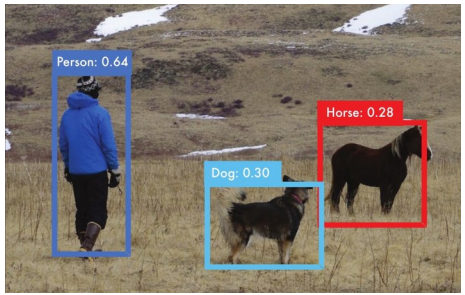
$$\sum_{i=1}^{K} \mathbf{p}_i = 1$$

(distinct from the **multi-label** classification task: classes are <u>not</u> mutually exclusive)

Classification vs ...

Recognition

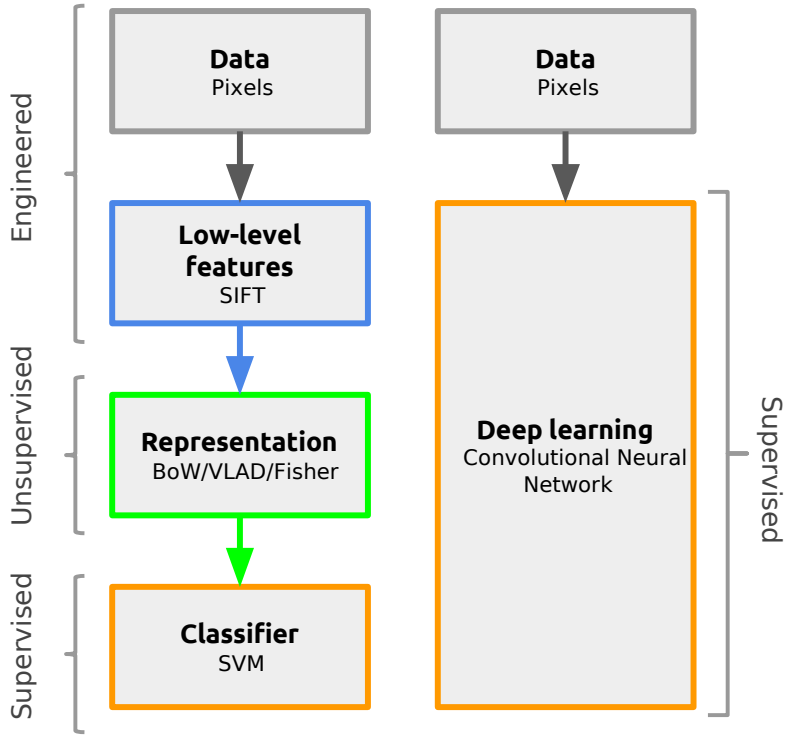
Detection



Semantic segmentation

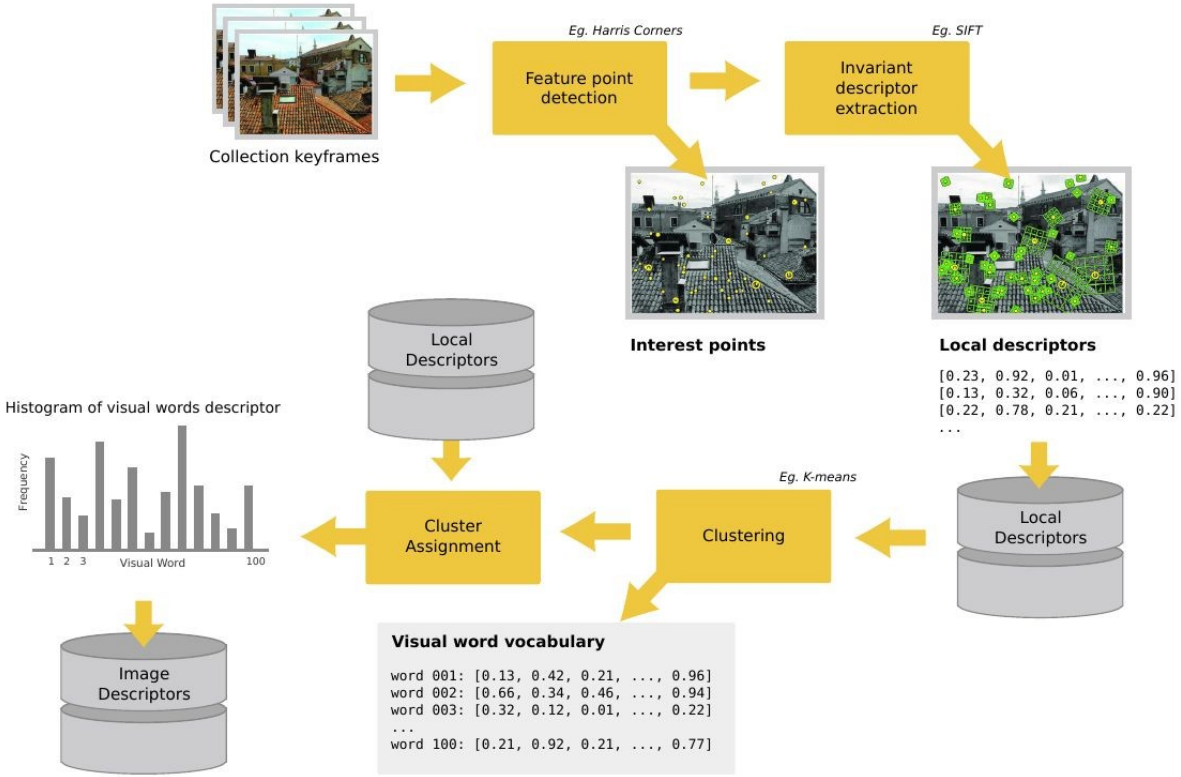
Deep learning vs shallow learning

- Old style machine learning:
  - Engineer features (by some unspecified method)
  - Create a representation (descriptor)
  - Train shallow classifier on representation
- Example:
  - SIFT features (engineered)
  - BoW representation (engineered + unsupervised learning)
  - SVM classifier (convex optimization)
- Deep learning
  - Learn layers of features, representation, and classifier in one go based on the data alone
  - Primary methodology: deep neural networks (non-convex)



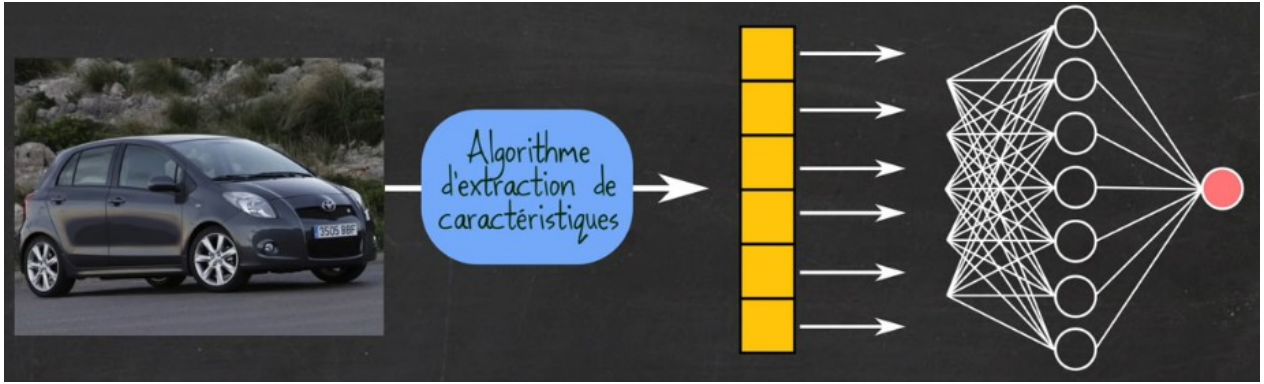
Example: feature engineering in computer vision

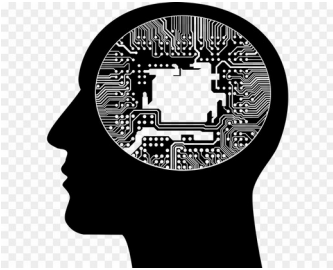
$\phi(x)$



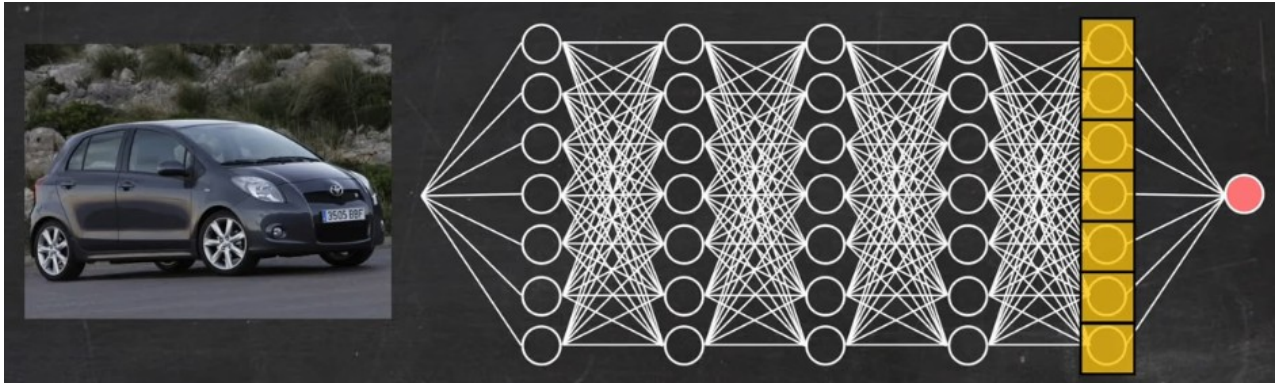
Feature based analysis

Features + Shallow network



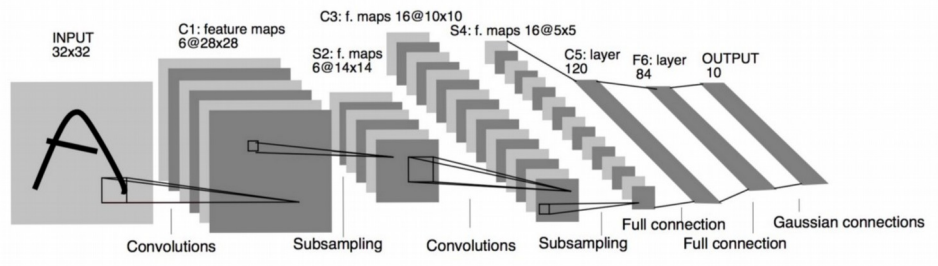


Always feature based analysis but ...



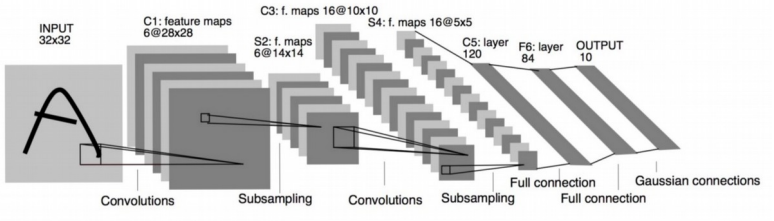
Learning → MIRACLE !

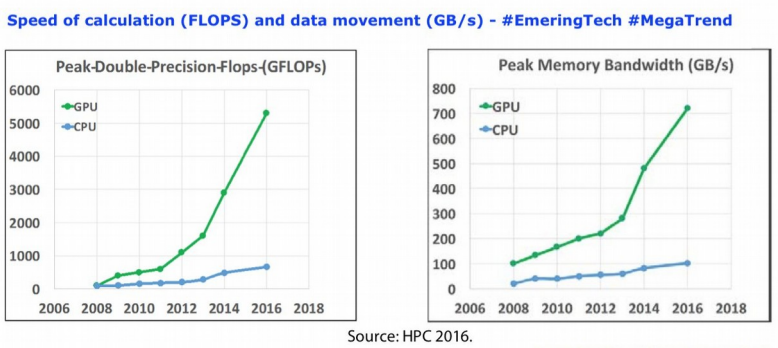
1988 … 1994 - Y. Le Cun - LeNet5



Why deep learning grows up ?

1988 … 1994 - Y. Le Cun - LeNet5



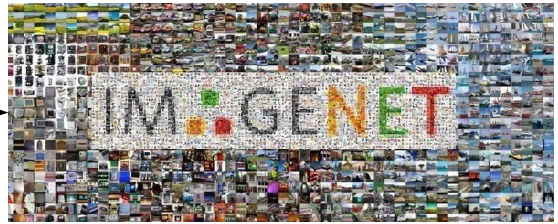


Algorithms have been improved

Adapted architecture have been proposed

High computing power

Data availability



The ImageNet large scale visual recognition challenge (ILSVRC)

IMAGENET **Large Scale Visual Recognition Challenge (ILSVRC)**

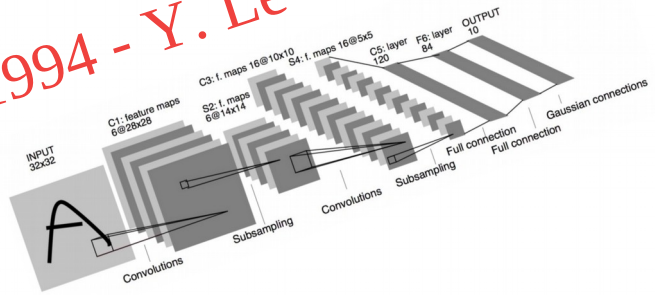
Competition

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale. One high level motivation is to allow researchers to compare progress in detection across a wider variety of objects -- taking advantage of the quite expensive labeling effort. Another motivation is to measure the progress of computer vision for large scale image indexing for retrieval and annotation.

For details about each challenge please refer to the corresponding page.

- ILSVRC 2017
- ILSVRC 2016
- ILSVRC 2015
- ILSVRC 2014
- ILSVRC 2013
- ILSVRC 2012
- ILSVRC 2011
- ILSVRC 2010

1988 … 1994 - Y. Le Cun - LeNet5



The ImageNet large scale visual recognition challenge (ILSVRC)

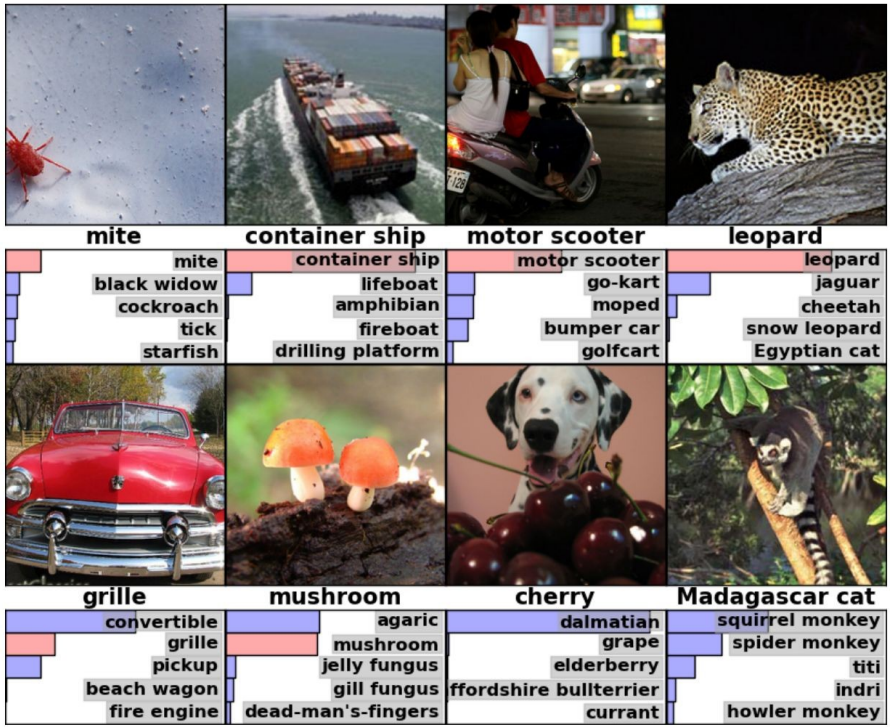
Dataset:

- 1.2 million training images
- 100 K test images
- 1,000 object classes (categories)
- Balanced dataset

Categories are leaves of the ImageNet hierarchy

- **No overlap**: presence of one category implies absence of another
- Suitable for **softmax** classification

Peculiarities: the 1000 classes contains 120 breeds of dogs!



ILSVRC 2012

Pre-2012 approaches based on:

- SIFT, CSIFT, GIST, LBP, color stats
- Fisher vector encoding
- Linear SVMs
- Performance starting to saturate

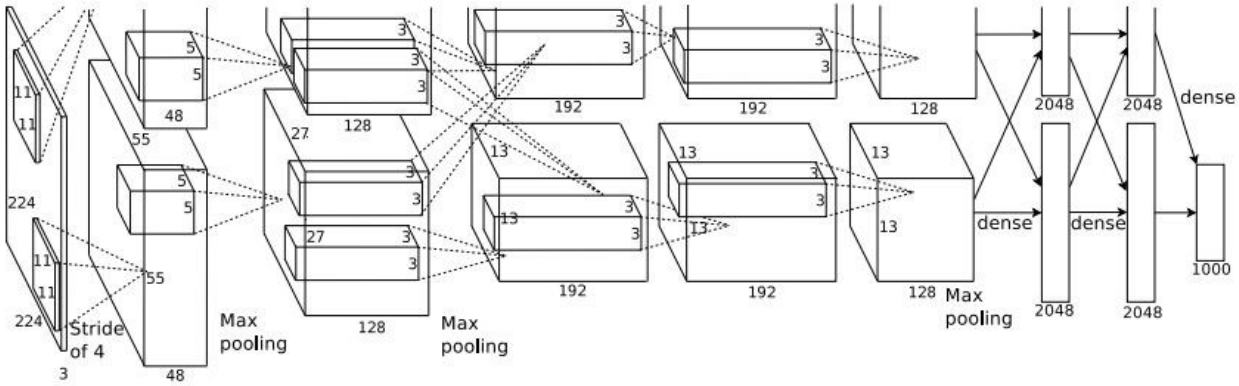
2012 winner

- Supervision (Krizhevsky, Hinton, ...)
- "**Alexnet**" convnet
- 10% margin on other approaches
- Revolution in computer vision
- Top-5 error: **15.315%**

Krizhevsky et al. ImageNet classification with deep convolutional neural networks. NIPS, 2012.

Alexnet

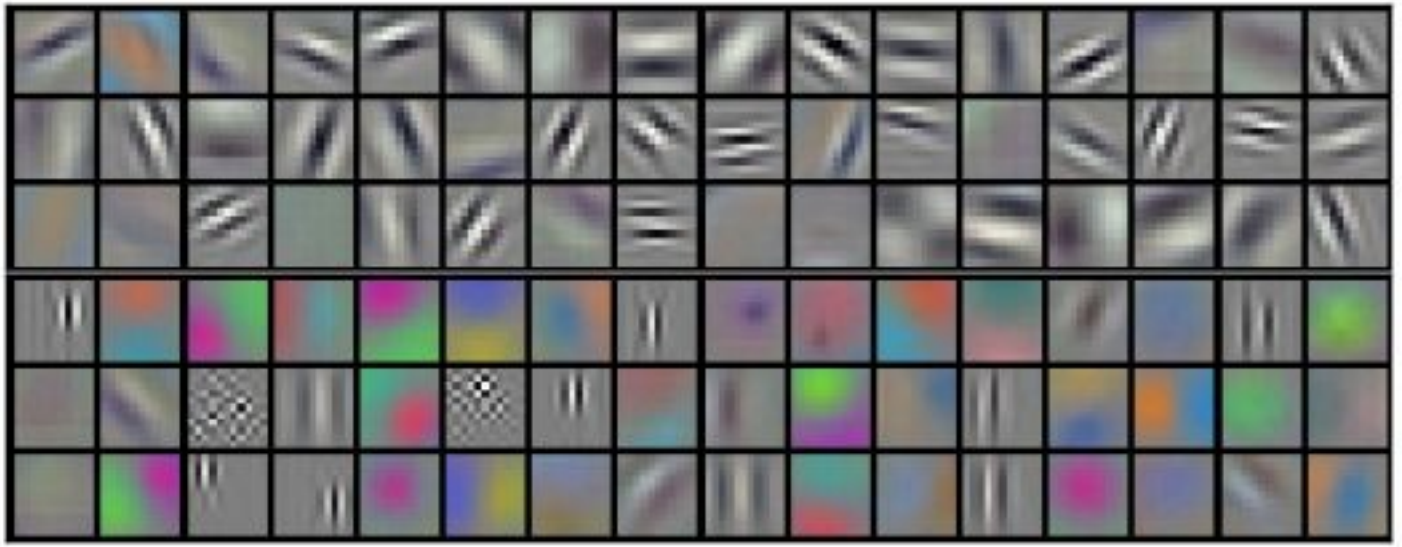
- 8 parameter layers (5 convolution, 3 fully connected)
- Softmax output
- 650,000 units
- 60 million free parameters
- Trained on two GPUs (two streams) for a week
- Ensemble of 7 nets used in ILSVRC challenge



Krizhevsky et al. ImageNet classification with deep convolutional neural networks. NIPS, 2012.

Filters learned by Alexnet

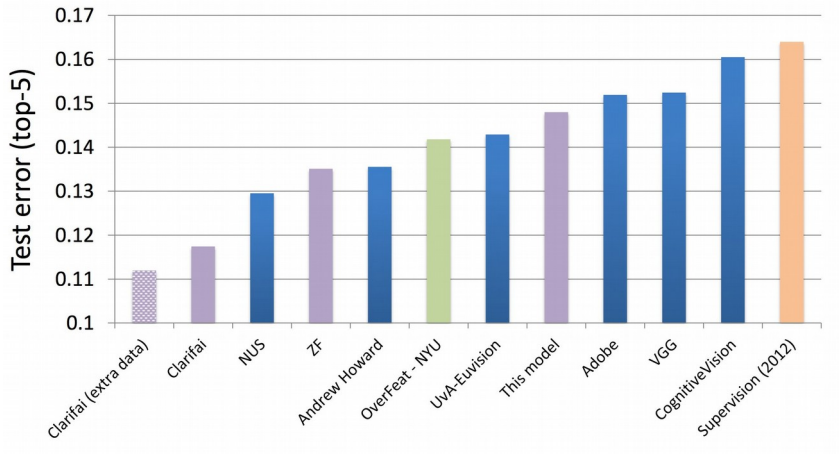
Visualization of the 96 11 x 11 filters learned by bottom layer



Krizhevsky et al. ImageNet classification with deep convolutional neural networks. NIPS, 2012.

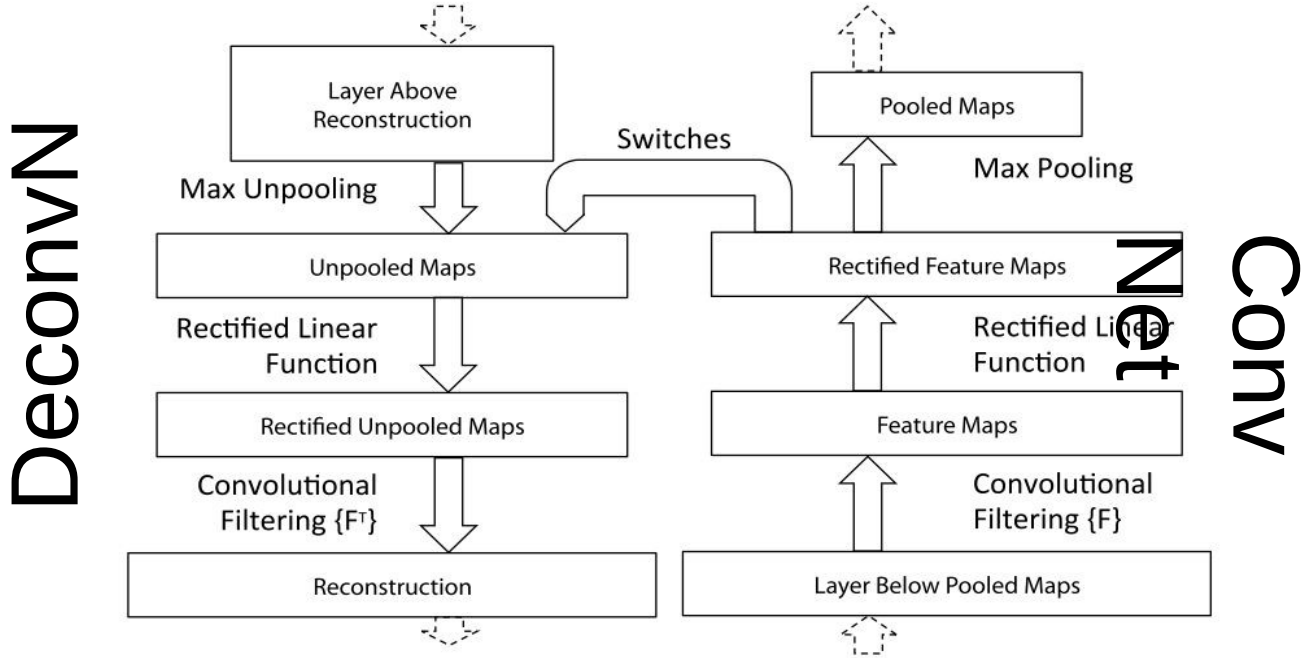
ILSVRC 2013: Zeiler and Fergus (ZFNet)

- Simple modifications of Alexnet designed to retain more information about features in early layers and reduce the number of dead filters
  - Reduce filter size on first layer to 7x7
  - Reduce stride on first layer to 2
  - Additional dropout on input layer
- Modifications motivated by visualizing activations of Alexnet
- Won ILSVRC 2013



Zeiler and Fergus. Visualizing and understanding convolutional networks *ECCV 2014*

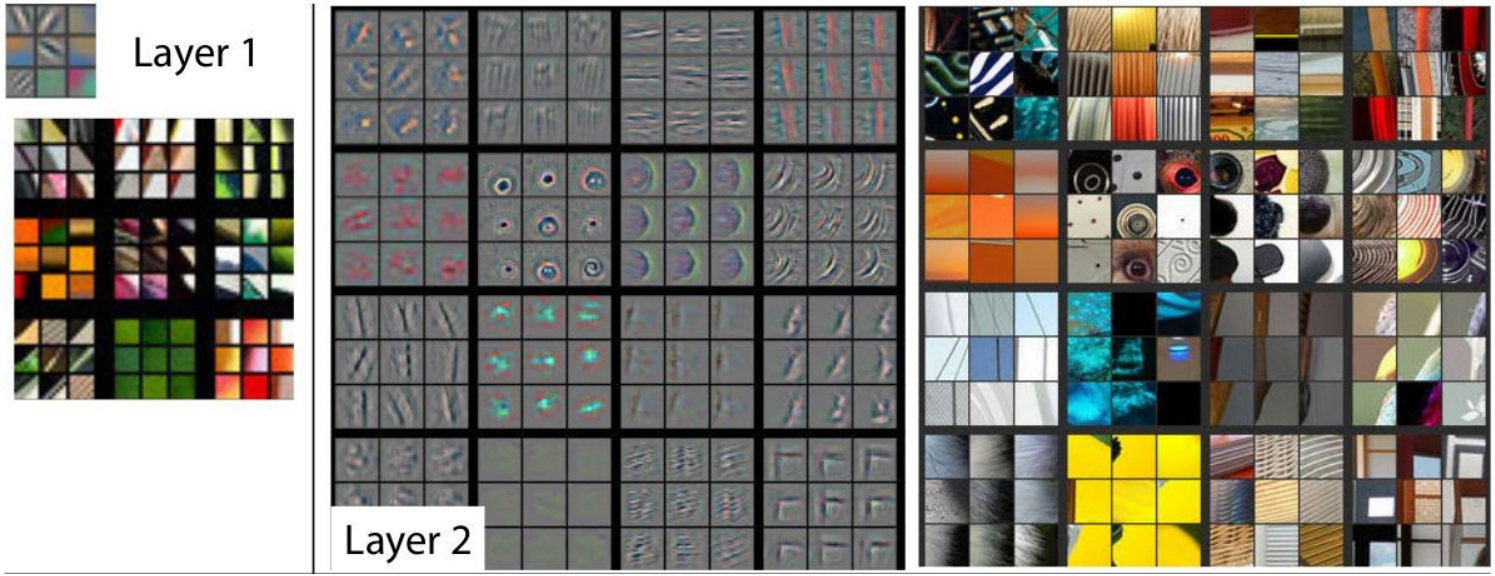
Using deconvolutions to visualize layer responses



Zeiler and Fergus. Visualizing and understanding convolutional networks *ECCV 2014*

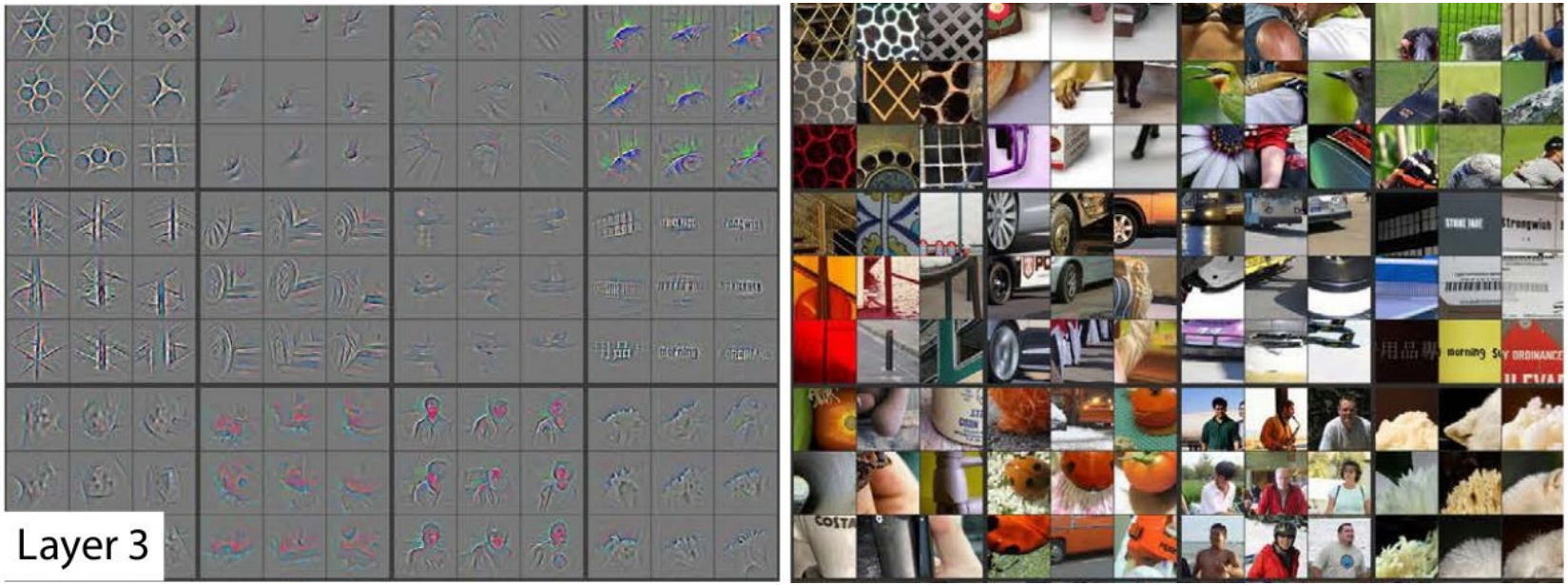
What do convnets learn?

Visualization of layer responses using a deconvolutional neural network on alexnet



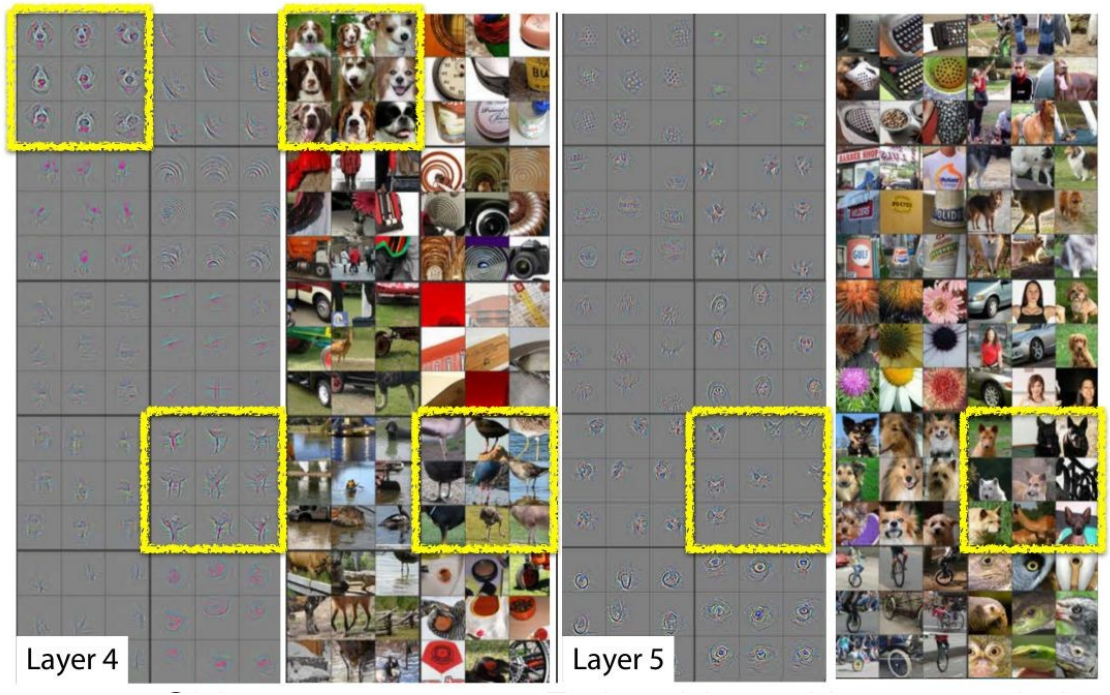
Zeiler and Fergus. Visualizing and understanding convolutional networks *ECCV 2014*

What do convnets learn?



Zeiler and Fergus. Visualizing and understanding convolutional networks *ECCV 2014*

What do convnets learn?

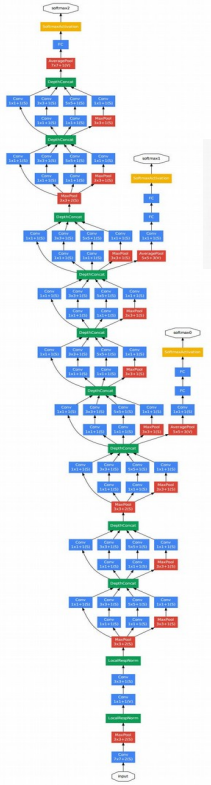


Zeiler and Fergus. Visualizing and understanding convolutional networks *ECCV 2014*

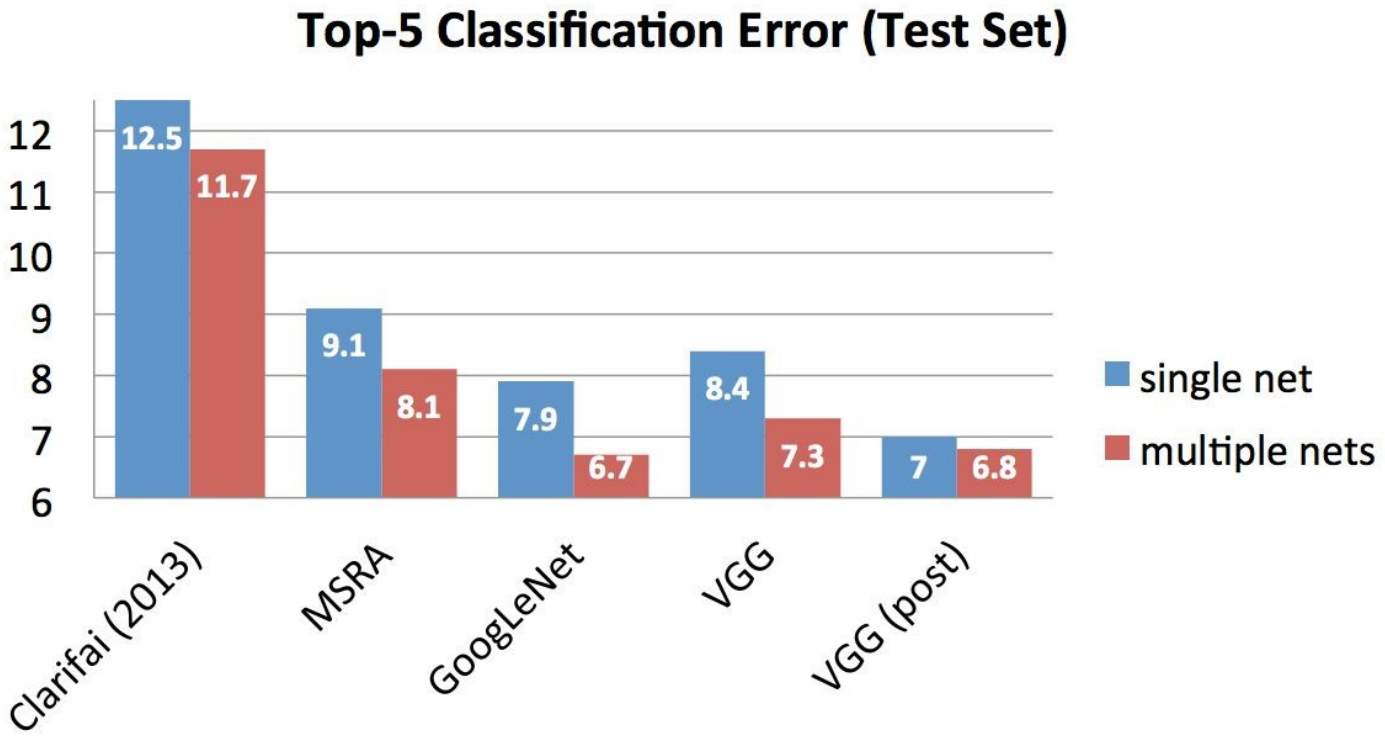
ILSVRC 2014: VGG and Inception

AlexNet	19-layer
image	image
conv-64	conv-64
conv-192	conv-64
conv-384	conv-128
conv-256	conv-128
conv-256	conv-256
FC-4096	conv-256
FC-4096	conv-256
FC-1000	conv-256
	conv-512
	conv-512
	conv-512
	conv-512
	conv-512
	conv-512
	conv-512
	conv-512
	FC-4096

image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
conv-256
conv-256
maxpool
conv-512
conv-512
conv-512
conv-512
maxpool
conv-512
conv-512
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax



ImageNet 2014: VGG and Inception



Simonyan, Karen, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR 2015*. [video] [slides] [project]

Modern convnets: VGG

VGG networks (Simonyan & Zisserman)

- Add more layers
- Stacked convolutions with smaller apertures (3x3) work better than a large (7x7) convolution

2 versions

- VGG-16 (16 parameter layers)
- VGG-19 (19 parameter layers)

ILSVRC 2014

- Top-5 error (16 layer): 7.5%
- Top-5 error (19 layer): 7.4%
- Top-5 error (ensemble): 7.0%

http://www.robots.ox.ac.uk/~vgg/research/very_deep/

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 × 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**
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**
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**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

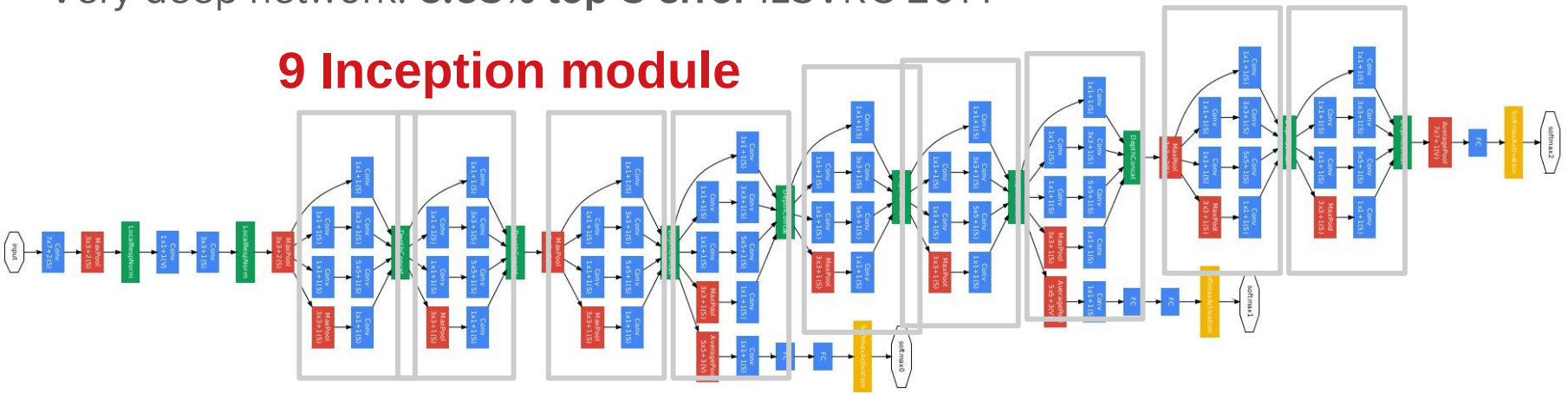
Simonyan, Karen, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR 2015*. [video] [slides] [project]

Modern convnets: GoogLeNet

Introduced **inception layer** for multiscale analysis

Extensive use of **1x1 convolutions** for dimensionality reduction

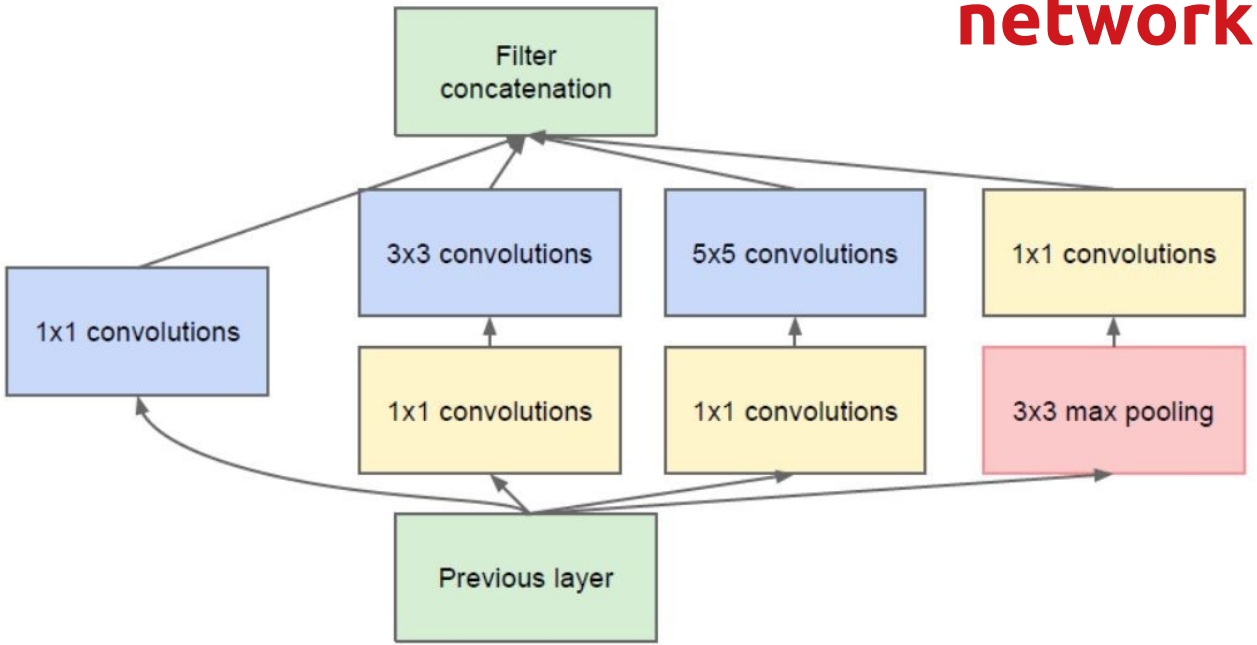
Very deep network. **6.65% top-5 error** ILSVRC 2014



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

Inception module

A network in a network



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

ILSVRC 2015: Resnet

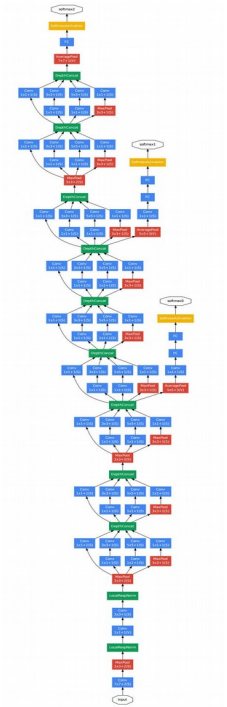
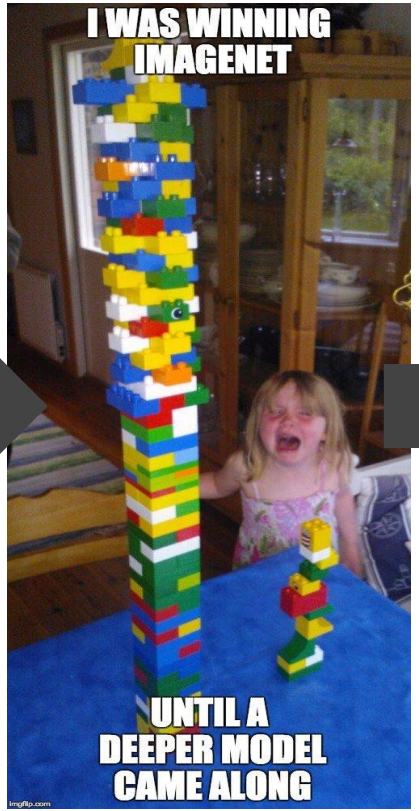
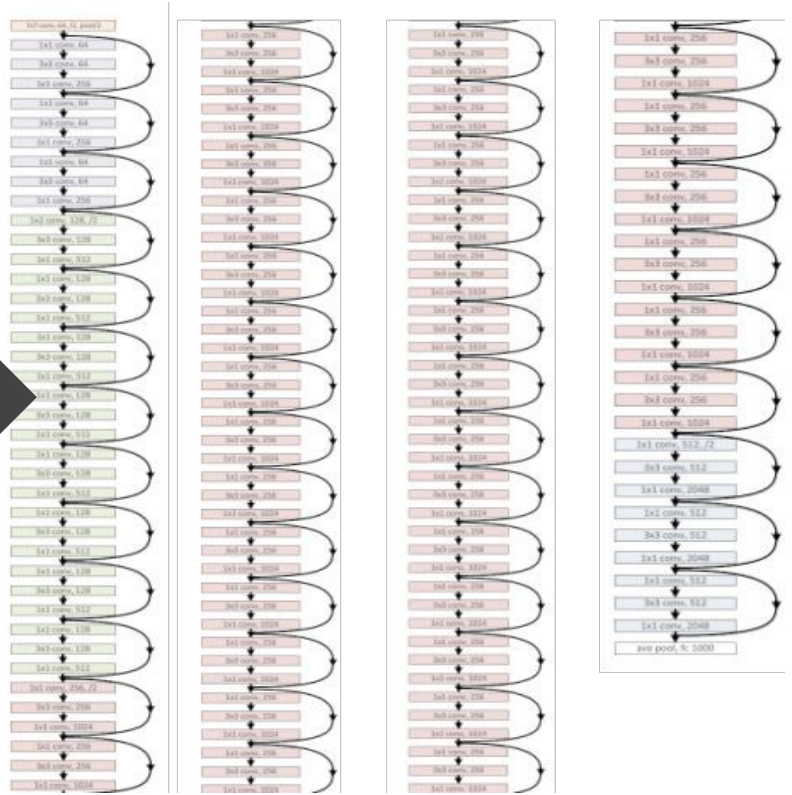
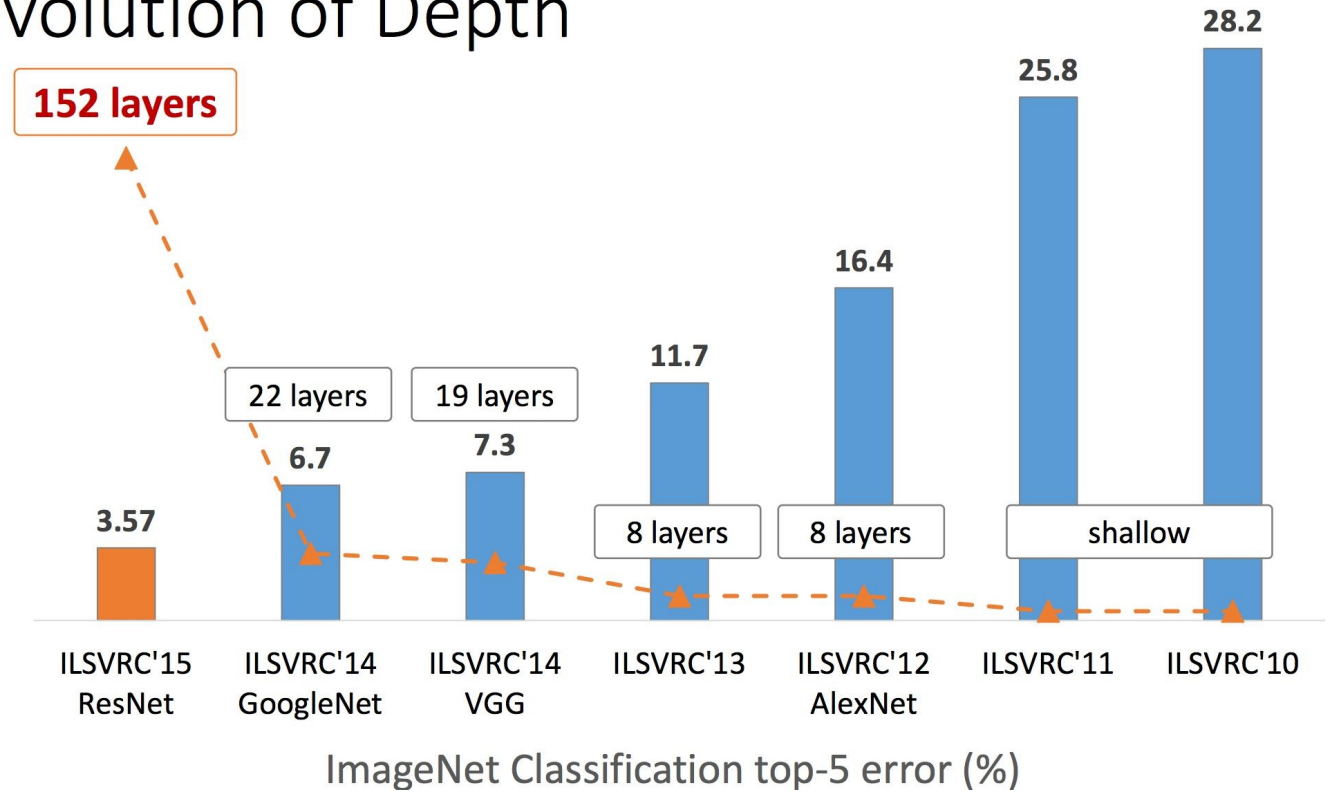


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
conv-256
conv-256
maxpool
conv-512
conv-512
conv-512
conv-512
maxpool
conv-512
conv-512
conv-512
conv-512
maxpool
FC-4096
FC-4096
FC-1000
softmax



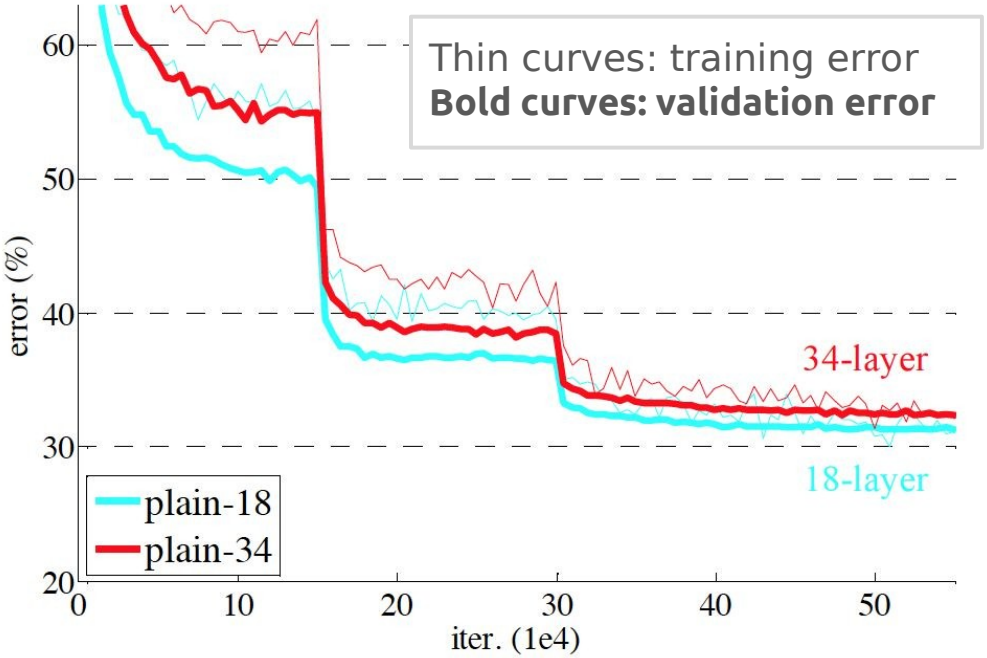


Revolution of Depth



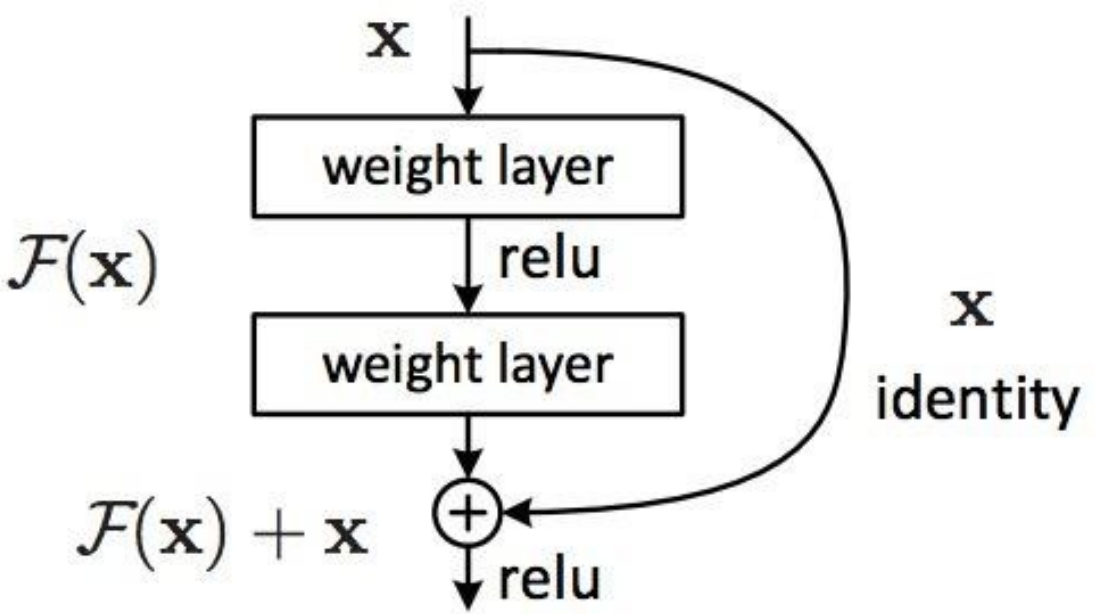
He et al. Deep Residual Learning for Image Recognition CVPR 2016

The problem with training very deep networks



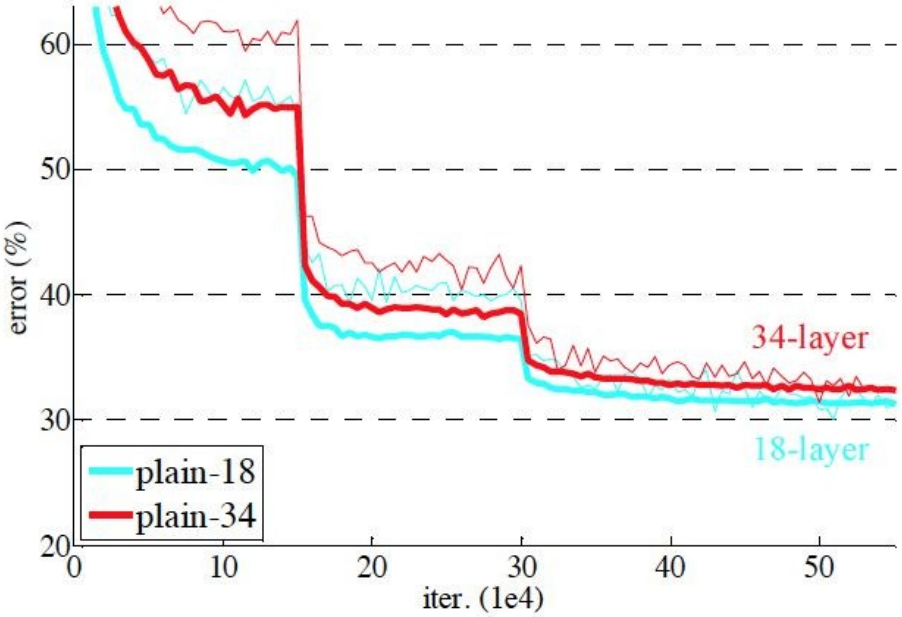
He et al. Deep Residual Learning for Image Recognition CVPR 2016

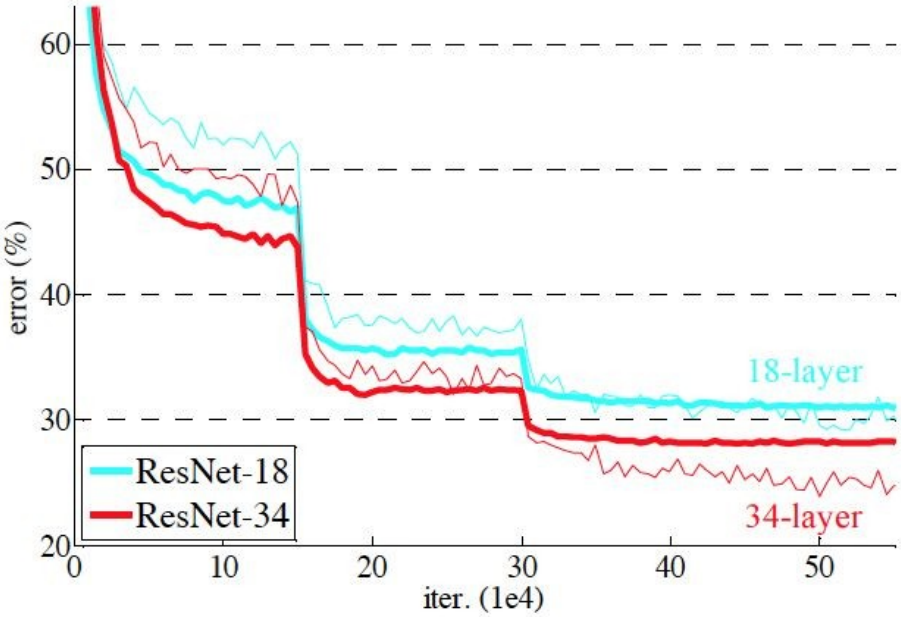
Residual blocks



He et al. Deep Residual Learning for Image Recognition CVPR 2016

34 layer plain network vs 34 layer residual network





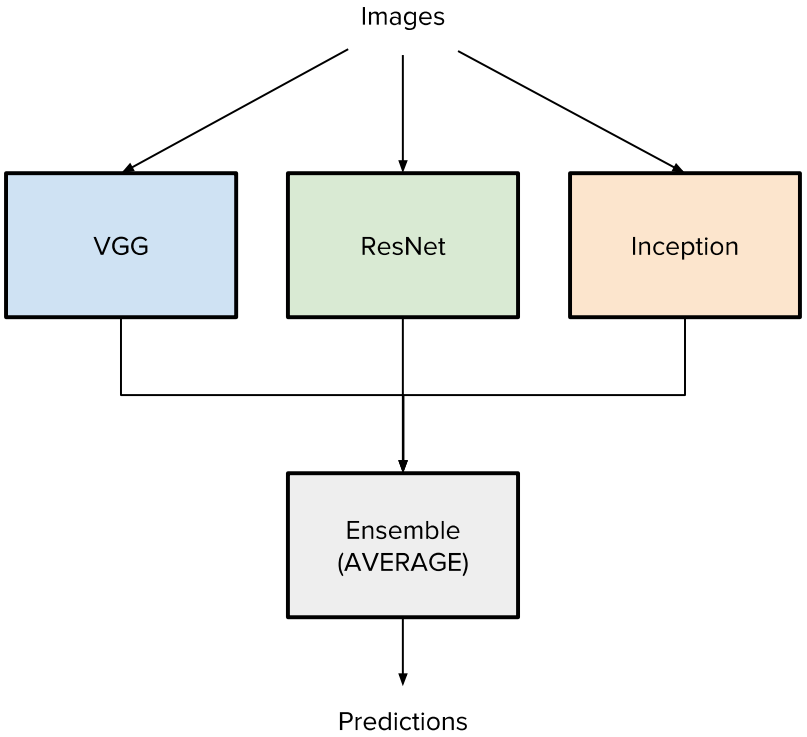
He et al. Deep Residual Learning for Image Recognition CVPR 2016

ILSVRC 2016

Source: http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf

Ensembles (Hikivision)

- More than 20 models, including VGG, Inception, ResNet and variations of it.
- Novel data augmentation.
- Novel learning rate policy.
- ...and “some small tricks”



ResNext = ResNet + Inception

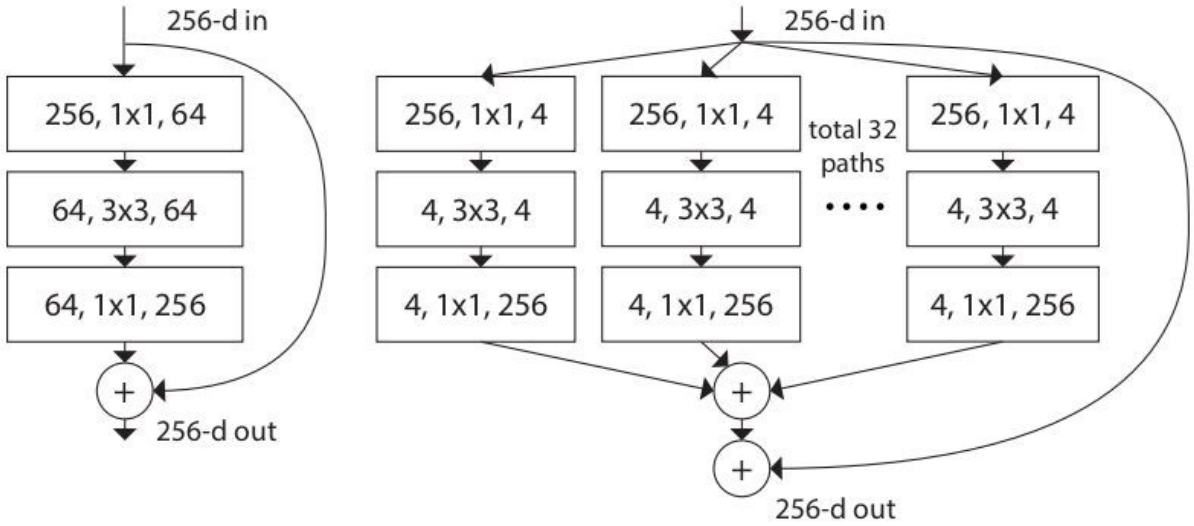
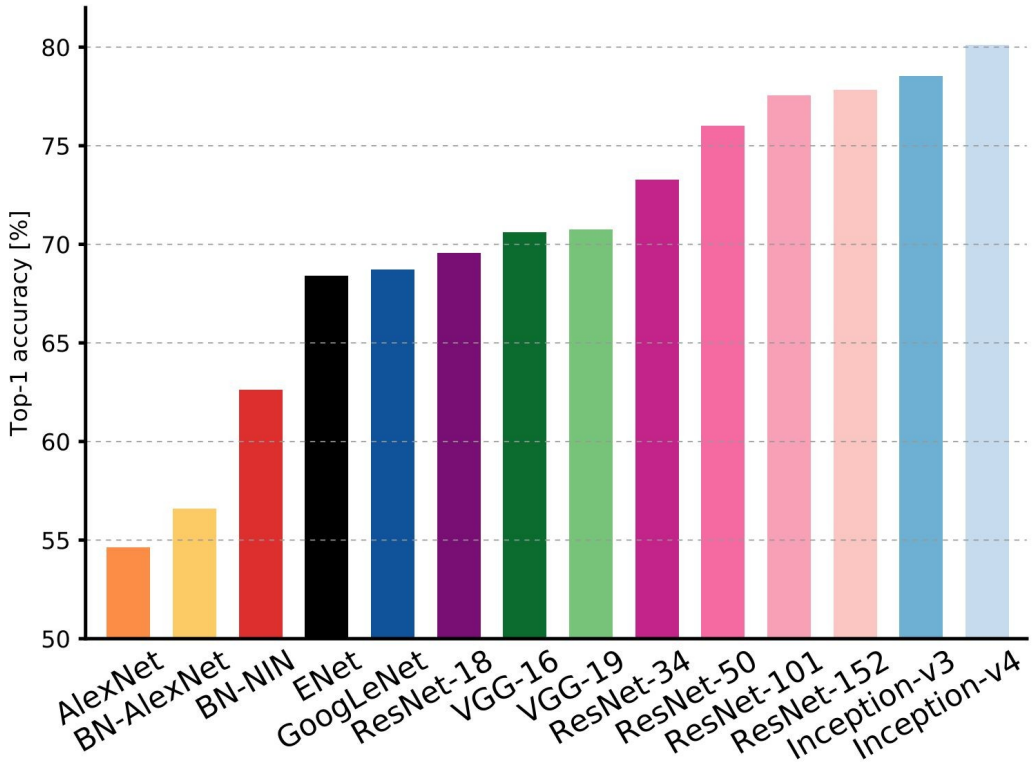


Figure 1. **Left**: A block of ResNet [13]. **Right**: A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).

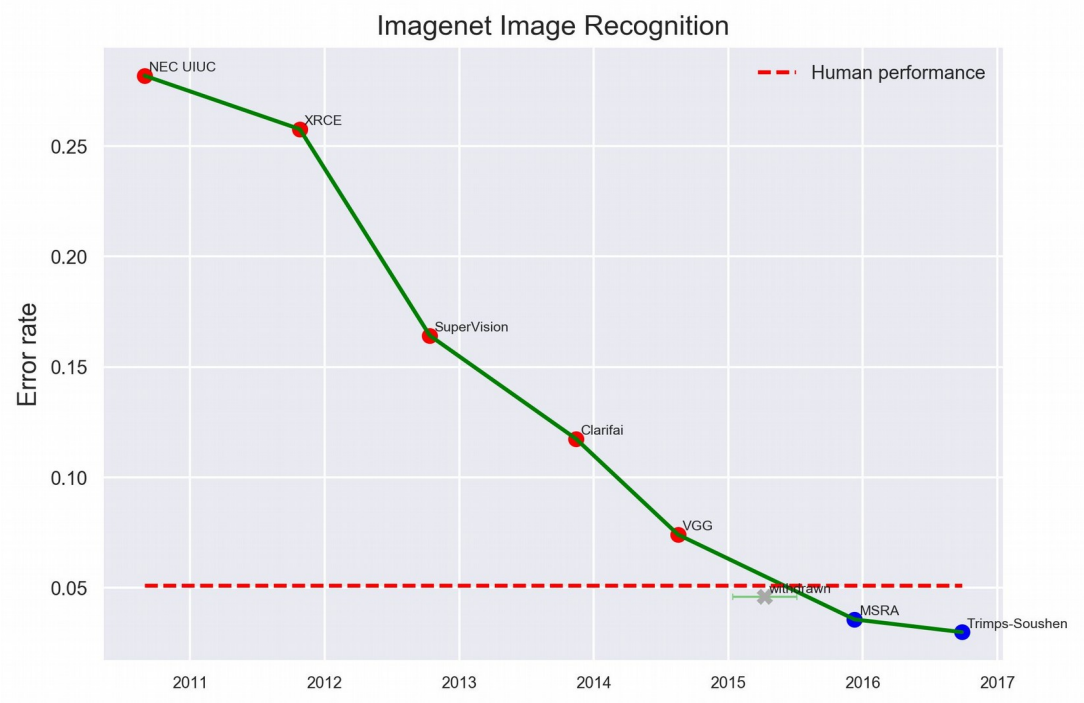
facebook research

rmations for deep neural networks. CVPR 2017 [code]



Canziani, Alfredo, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications." arXiv 2016.

ILSVRC 2017: the end of the challenge



Electronic Frontier Foundation: "Measuring the Progress of AI Research" (2017)

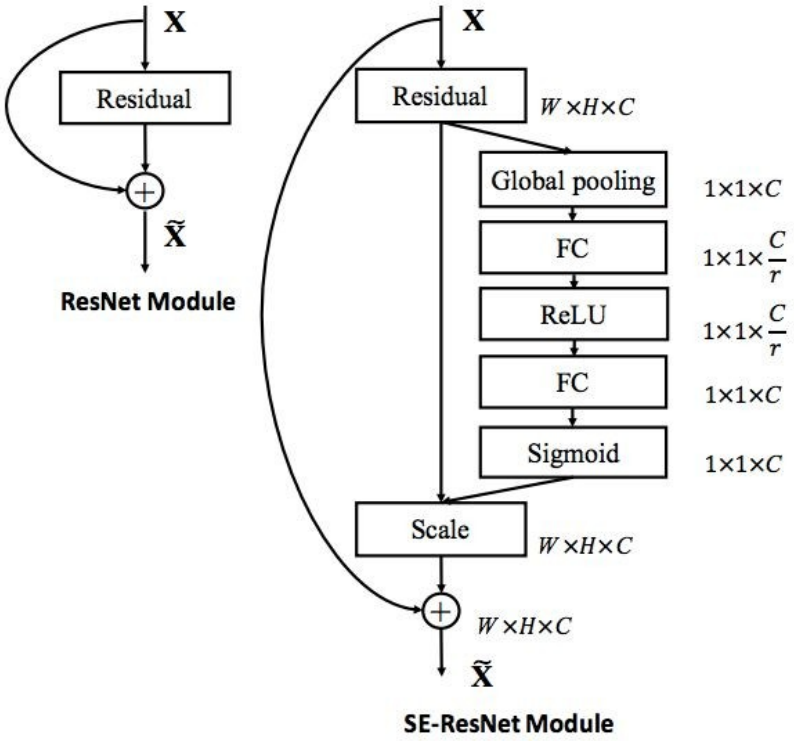
The end of the challenge

Beyond ImageNet Large Scale Visual Recognition Challenge

July 26th in conjunction with CVPR 2017

http://image-net.org/challenges/beyond_ilsvrc

Squeeze and excitation networks



Used in ILSVRC 2017

25% improvement over ResNet and ResNext

Jie Hu, Li Shen, Gang Sun, Squeeze-and-Excitation Networks, CVPR 2018

Other datasets and challenges

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Visual Genome is a dataset, a knowledge base, an ongoing effort to connect structured image concepts to language.

- 108,077 images
- 5.4M region descriptions
- 1.7M visual question answers
- 3.8M object instances
- 2.8M attributes
- 2.3M relationships
- Everything mapped to wordnet synsets