Very Deep Convolutional Networks For Large-Scale Image recognition

Karen Simonyan∗ & Andrew Zisserman
Visual Geometry Group, Department of Engineering Science, University of Oxford

rayjang111@gmail.com
HyunSukJang

CONTENTS

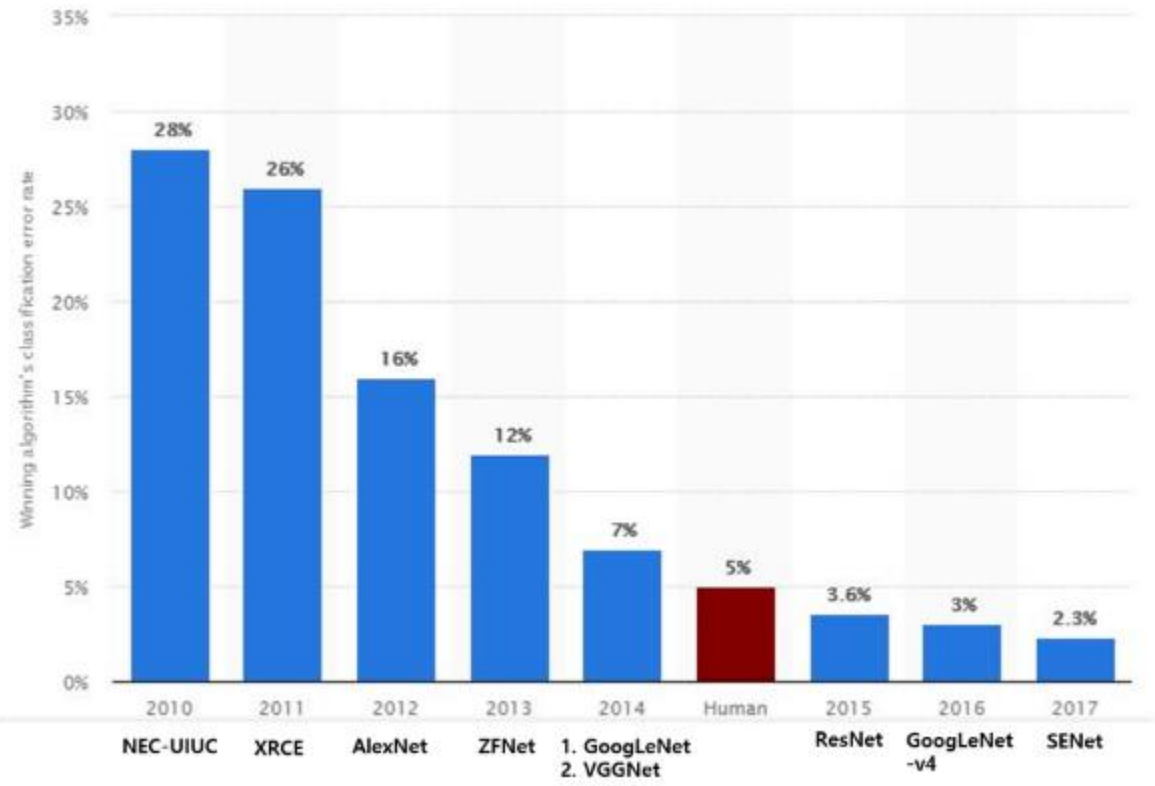
01 Introduction

What is ILSVRC???

- Imagenet Large Scale Visual Recognition Challenge
- More than 14 million images have been hand-annotated by the project to indicate what objects are pictured
- Testbed for a few generations of large-scale image classification systems, from high-dimensional shallow feature encodings to deep ConvNets

01. Introduction

Winners of ILSVRC challenge



01.Introduction

Before 2012 ILSVRC=> feature extraction by hand

Frame of classification

Data Acquisition → Preprocessing → Dimension reduction& optimal feature selection → Classfication

Data Acquisition
- Image
- Video
- Audio

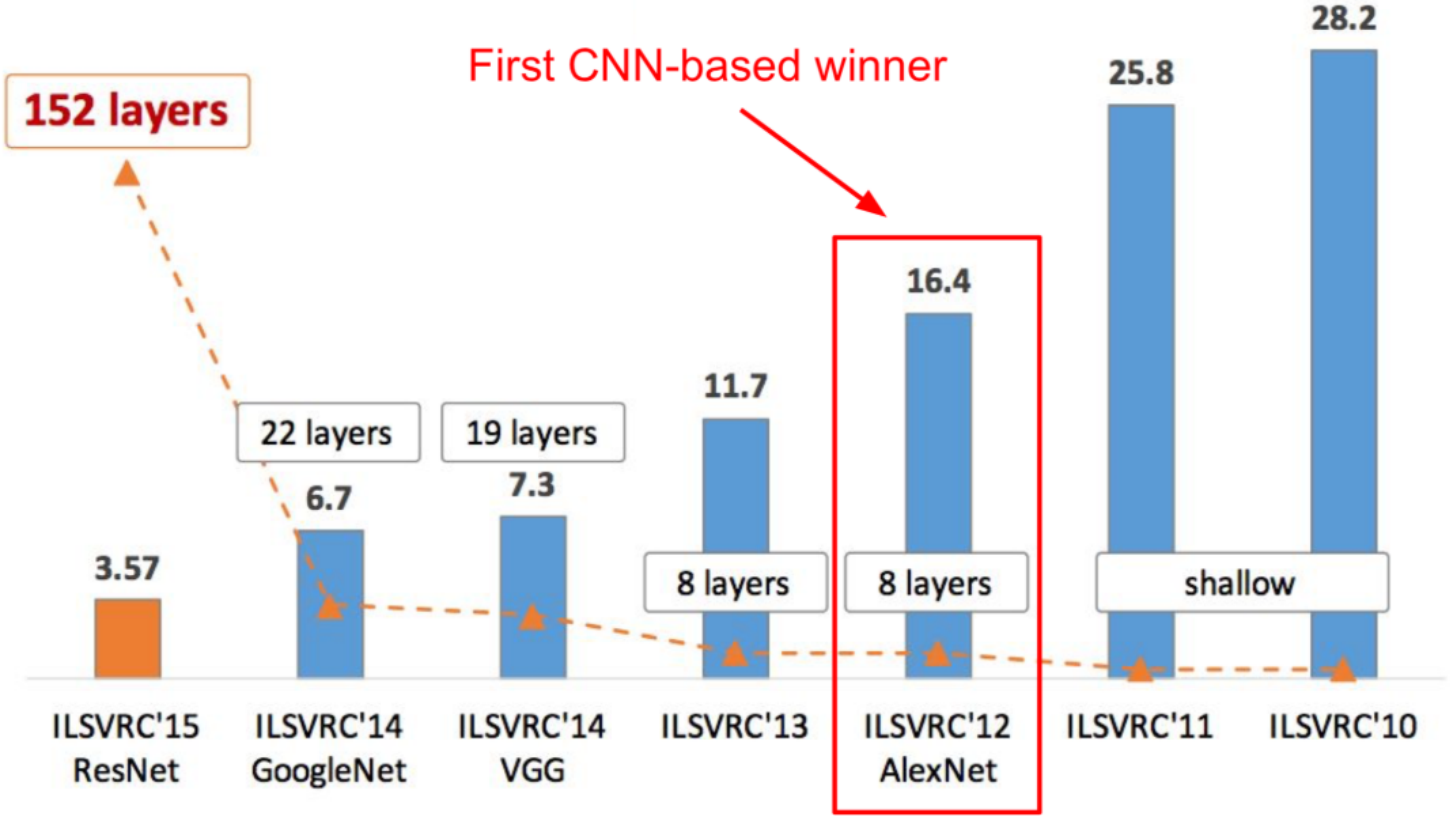
Preprocessing
- Noise reduction
- Image enhancement

Dimension reduction& optimal feature selection
- PCA(Principal component Analysis)
- LDA(Linear Discriminal Analysis)
- LBP(Local Binary Pattern)

Classfication
- Logistic Regression
- SVM

01.Introduction

After ILSVRC 2012



To train deep layers and many parameters=> data augmentation, regulation

02 Vggnet Model architecture

02. Model Architecture

Main Disussion

What if we simply increase the depth of Nets?

We address another important aspect of ConvNet architecture design – its depth. To this end, we fix other parameters of the architecture, and steadily increase the depth of the network by adding more convolutional layers, which is feasible due to the use of very small (3 × 3) convolution filters in all layers.

02. Model Architecture

what have we gained by using, for instance, a stack of three 3×3 conv. layers instead of a single 7×7 layer?

First, we incorporate three non-linear rectification layers instead of a single one, which makes the decision function more discriminative.

Second, we decrease the number of parameters

02. Model Architecture

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 **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					

A vs A-LRN:
Local Response Normalization

B vs C:
Use of 1*1 conv layer
To add Nonlinearity

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Model B: Number of parameters

(3*3*3)*64+(3*3*64)*64+(3*3*64)*128+(3*3*128)*128+(3*3*128)*256+(3*3*256)*256+(3*3*256)*512+(3*3*256)*512+(3*3*512)*512+(3*3*512)*512+(7*7*512*4096)+4096*4096+4096*1000=133m

03training&test method

03. training&test method

ILSVRC-2012 has 1000images for each 1000classes
=>not enough image (overfitting)

Need Data Augmentation

03. training&test method

Training Scale: S

Single Scale=> S=256, 384 fixed

Cropped 224*224 image from each fixed size image

Multi Scale=>Scale jittering

After training with S=384 fixed image,

Select S from Smin=256 Smax=512 => do fine tuning

PCA Color Augmentation

is designed to shift those values based on which values are the most present in the image.

Test Scale: Q

Convert image size to Q

Single test scale=> use only one Q

Multi test scale=> use multiple Qs
If test Scale S is fixed to single scale
Q is {S-32,S,S+32} since too much scale difference from trainset can lead worse performance

04 evaluation

Single test scale

Table 3: **ConvNet performance at a single test scale.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

04. evaluation

Multiple test scale

Table 4: **ConvNet performance at multiple test scales.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	$[256; 512]$	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	$[256; 512]$	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	$[256; 512]$	256,384,512	**24.8**	**7.5**

04. evaluation

Comparison to other models

Table 7: **Comparison with the state of the art in ILSVRC classification**. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

Thank you