

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Karen Simonyan* & Andrew Zisserman*

Visual Geometry Group, Department of Engineering Science, University of Oxford
{karen,az}@robots.ox.ac.uk

ABSTRACT

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve state-of-the-art results. We have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

Large size image recognition 문제에서 conv. net의 깊이가 정확도에 끼치는 영향

ARCHITECTURE

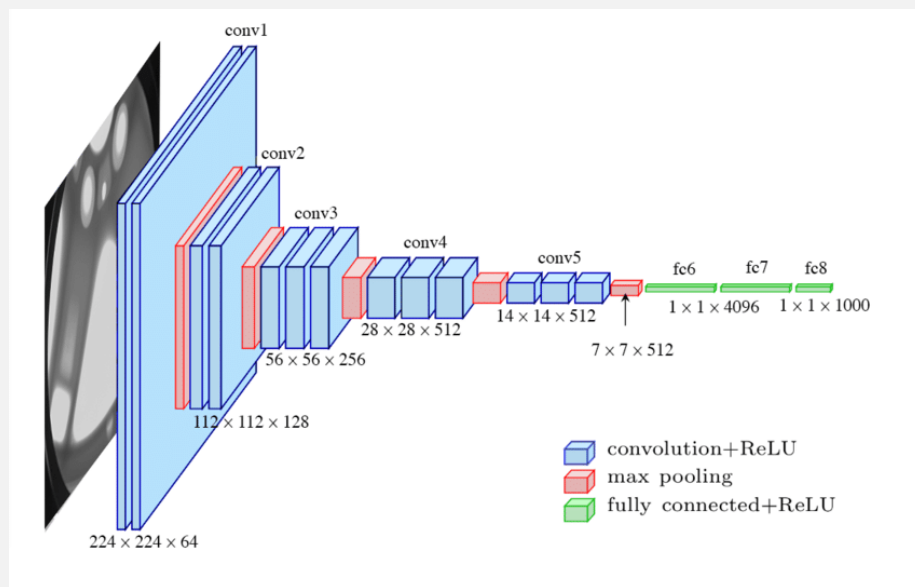


Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as “conv(receptive field size)-(number of channels)”. The ReLU activation function is not shown for brevity.

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
maxpool					
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	conv3-256	conv1-256	conv3-256	conv3-256
maxpool					
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv1-512	conv3-512	conv3-512
maxpool					
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv1-512	conv3-512	conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: Number of parameters (in millions).

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

- 224X224 RGB image with subtracting mean RGB value
- Small conv filter with relu : 3X3 conv layer와 1X1 conv layer
- 5 max pooling layer : 2X2 maxpool with stride 2
- 3 FC layer with relu: 4096, 4096, 1000 channels
- Softmax output

DISCUSSION

기존의 receptive field(7X7, 11X11)보다 매우 작은 3X3 receptive field

3개의 3X3 receptive field의 stack은 7X7 receptive field와 동일하면서도
다음의 차별점을 가진다.

- A. Layer마다 relu가 있기에 좀 더 비선형 근사를 세밀하게 할 수 있을 것으로 추정
- B. 동일한 receptive field 대비 적은 parameter 숫자
C채널 3X3 conv layer의 경우 $27c^2$ 개의 weight, 7X7 layer는 $49c^2$ 개의 weight 필요

1X1 conv. layer는 Network in network에서 사용되었고, receptive field의 변화없이 더
비선형화 시키는 역할을 한다.

Small conv. layer는 Ciresan도 사용했으나 깊이가 얕았다.

Goodfellow의 글에서도 depth의 긍정적인 영향에 대해 서술되었고,
GoogLeNet은 유사하지만 구조가 복잡하고, 연산 효율을 위해 해상도를 공격적으로 낮춤

CLASSIFICATION FRAMEWORK

Training

Minibatch GD with momentum

Batch size=256

Momentum=0.9

Learning rate= $1e-2$ accuracy에 진전이 없을 때마다 1/10

Regularization

Parameter norm penalty= $5e-4$

0.5 Dropout to 1,2 FC layer

Data augmentation : random crop, horizontal flip, RGB shift

결과적으로 학습률은 세 번 떨어졌고 370K번 74 epochs만에 학습 종료.
많은 parameter와 깊이에도 적은 학습으로 가능했던 데는 다음이 영향을 끼쳤다.

- a) Depth와 small conv filter에 따른 implicit regularization
- b) Pre-initialization

CLASSIFICATION FRAMEWORK

Training

Pre-initialization

Depth가 낮은 A에 random initialize 후 pre-train한 후 A의 첫 4개 conv. Layer와 FC layer 전체를 다른 모델에 initialization으로 사용했다. (학습률은 고정)

중간 layer들은 평균 0, 분산 0.01인 정규분포에서 sampling하여 random initialize, bias=0

Single scale image training

Training image size는 256과 384를 사용했고, input은 224로 crop.

학습속도 증진을 위해 384는 256의 pre-trained initialize와 학습률 $1e-3$ 을 사용했다.

Multiple scale image training

256~512 random rescaled image

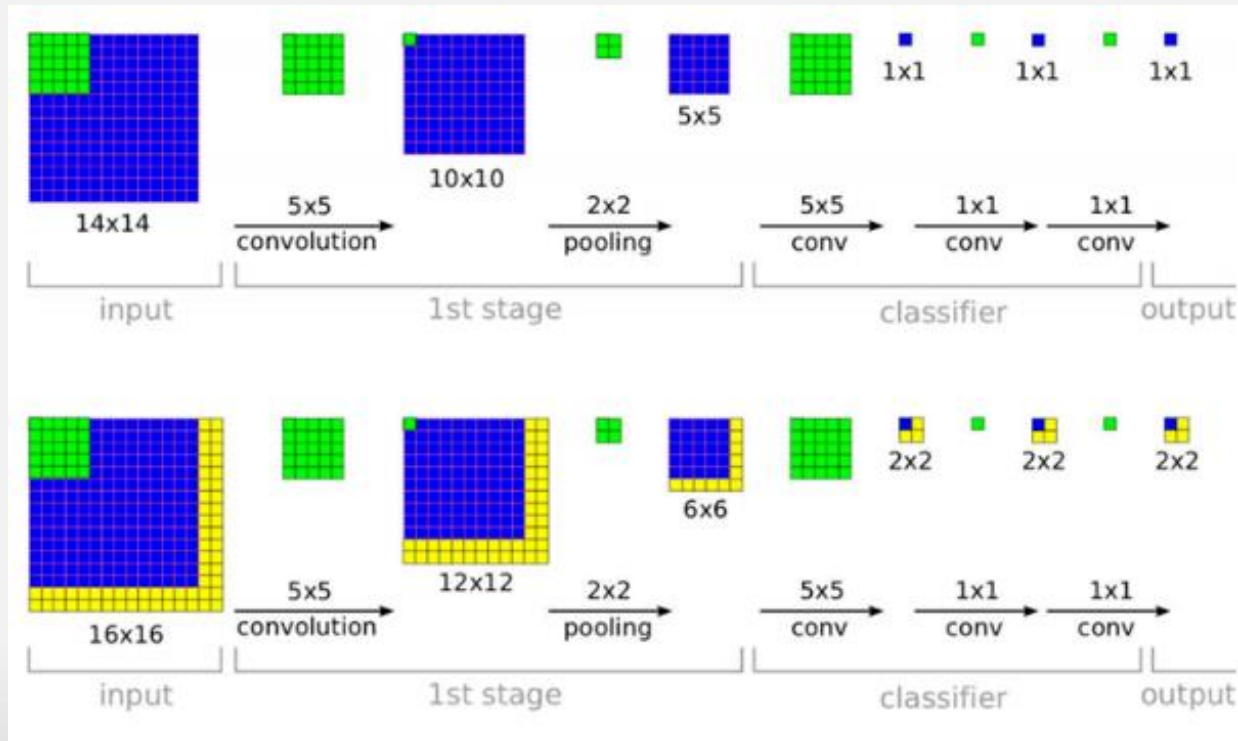
학습속도 증진을 위해 384 모델을 사용했다.

CLASSIFICATION FRAMEWORK

Testing

Image

Use non-fixed size whole image(Overfeat)



CLASSIFICATION EXPERIMENT

Single scale evaluation

ILSVRC-2012 dataset

Test image size(Q)=256, 384 or 0.5(min S + max S)

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

Table 1: ConvNet configurations (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv(receptive field size)-(number of channels)". The ReLU activation function is not shown for brevity.

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
maxpool					
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	conv3-256	conv1-256	conv3-256	conv3-256 conv3-256
maxpool					
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv1-512	conv3-512	conv3-512 conv3-512
maxpool					
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv1-512	conv3-512	conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: Number of parameters (in millions).

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

CLASSIFICATION EXPERIMENT

Multi scale evaluation

ILSVRC-2012 dataset

Test image size(Q)=S+-32 or min S, 0.5(min S + max S), max S

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
	256	224,256,288	27.7	9.2
C	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

Table 1: ConvNet configurations (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv(receptive field size)-(number of channels)". The ReLU activation function is not shown for brevity.

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	conv3-256 conv3-256 conv3-256
maxpool					
conv3-512	conv3-512	conv3-512 conv3-512	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	conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: Number of parameters (in millions).

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

CLASSIFICATION EXPERIMENT

Comparison between multi crop and dense eval.

Table 5: **ConvNet evaluation techniques comparison.** In all experiments the training scale S was sampled from $[256; 512]$, and three test scales Q were considered: $\{256, 384, 512\}$.

ConvNet config. (Table 1)	Evaluation method	top-1 val. error (%)	top-5 val. error (%)
D	dense	24.8	7.5
	multi-crop	24.6	7.5
	multi-crop & dense	24.4	7.2
E	dense	24.8	7.5
	multi-crop	24.6	7.4
	multi-crop & dense	24.4	7.1

CLASSIFICATION EXPERIMENT

ConvNet fusion result

Table 6: **Multiple ConvNet fusion results.**

Combined ConvNet models	Error		
	top-1 val	top-5 val	top-5 test
ILSVRC submission			
(D/256/224,256,288), (D/384/352,384,416), (D/[256;512]/256,384,512) (C/256/224,256,288), (C/384/352,384,416) (E/256/224,256,288), (E/384/352,384,416)	24.7	7.5	7.3
post-submission			
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), dense eval.	24.0	7.1	7.0
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), multi-crop	23.9	7.2	-
(D/[256;512]/256,384,512), (E/[256;512]/256,384,512), multi-crop & dense eval.	23.7	6.8	6.8

CLASSIFICATION EXPERIMENT

Comparison with the state of the art

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	-

CONCLUSION

1. 최대 19개의 weight layer를 가지는 deep conv. net을 평가해보았다.
2. Depth가 이미지 분류의 accurac에 긍정적인 영향을 준다.
3. 기존 Conv. Net을 통해서도 ImageNet dataset을 충분히 분류할 수 있다.
4. 이미지 분류 외에도 다양한 분야에 적용할 수 있다.