Very Deep Convolutional Networks for Large-Scale Image Recognition

Abstract

increasing depth using an architecture with very small (3 x 3) convolution filters
16–19 weight layers
secured the first and the second places in the localisation and classification tracks respectively
generalise well to other datasets

Conclusion

Very Deep Convolution Networks (19 weight Layers)

Classification Experiments

ImageNet
1000 classes
training (1.3M images), validation (50K images), and testing (100K images with held-out class labels)
the top-1 and top-5 error

Single Scale Evaluation

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**

1 local response normalisation은 normalisation layers없이는 의미가 없다

2 ConvNet depth에 따라 성능이 좋아진다
근데 1x1 세 번보다 3x3 하나가 더 낫다
nonlinearity도 좋지만 spatial context by using conv도 중요
19 이상은 한계인 것 같지만 even deeper models might be beneficial for larger datasets
a deep net with small filters outperforms a shallow net with larger filters.

3 scale jittering at training time이 좋다

Multi Scale Evaluation

scale jittering at test time leads to better performance

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**

MULTI-CROP EVALUATION

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**

using multiple crops performs slightly better than dense evaluation

CONVNET FUSION

ensemble

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**

COMPARISON WITH THE STATE OF THE ART

Introduction

Krizhevsky 이후 발전
smaller receptive window size and smaller stride of the first convolutional layer
densely over the whole image and over multiple scales

very small filter -> depth

ConvNet Configurations

Architecture
a fixed-size 224 × 224 RGB image. The only preprocessing we do is subtracting the mean RGB value,
3x3 which is the smallest size to capture the notion of left/right, up/down, center
3 fc layers
ReLU
Local Response Normalisation

Configuration

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

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

Discussion
3x3
왜 7x7 안 쓰나?
1 three non-linearities
2 decrease the number of parameters

Classification Framework

Training
Krizhevsky랑 똑같음
the nets required less epochs to converge
(a) implicit regularisation imposed by greater depth and smaller conv. filter sizes; (b) pre-initialisation of certain layers.

The initialisation of the network weights is important

224 224 구하기
random crop
random horizontal flipping and random RGB colour shift

Smallest side
1 Fix S
2 random S in range

Testing
isotropically rescaled
Finally, to obtain a fixed-size vector of class scores for the image, the class score map is spatially averaged (sum-pooled).
We also augment the test set by horizontal flipping of the images;
the soft-max class posteriors of the original and flipped images are averaged to obtain the final scores for the image.

IMPLEMENTATION DETAILS