2023-07-20

PAPER REVIEW:
VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCLAE IMAGE RECOGNITION

BOAZ 멘토&멘티 스터디
분석 20기 권정을(E)

Keyword: imagenet, convolution neural network, classification, object detection

❖ Information

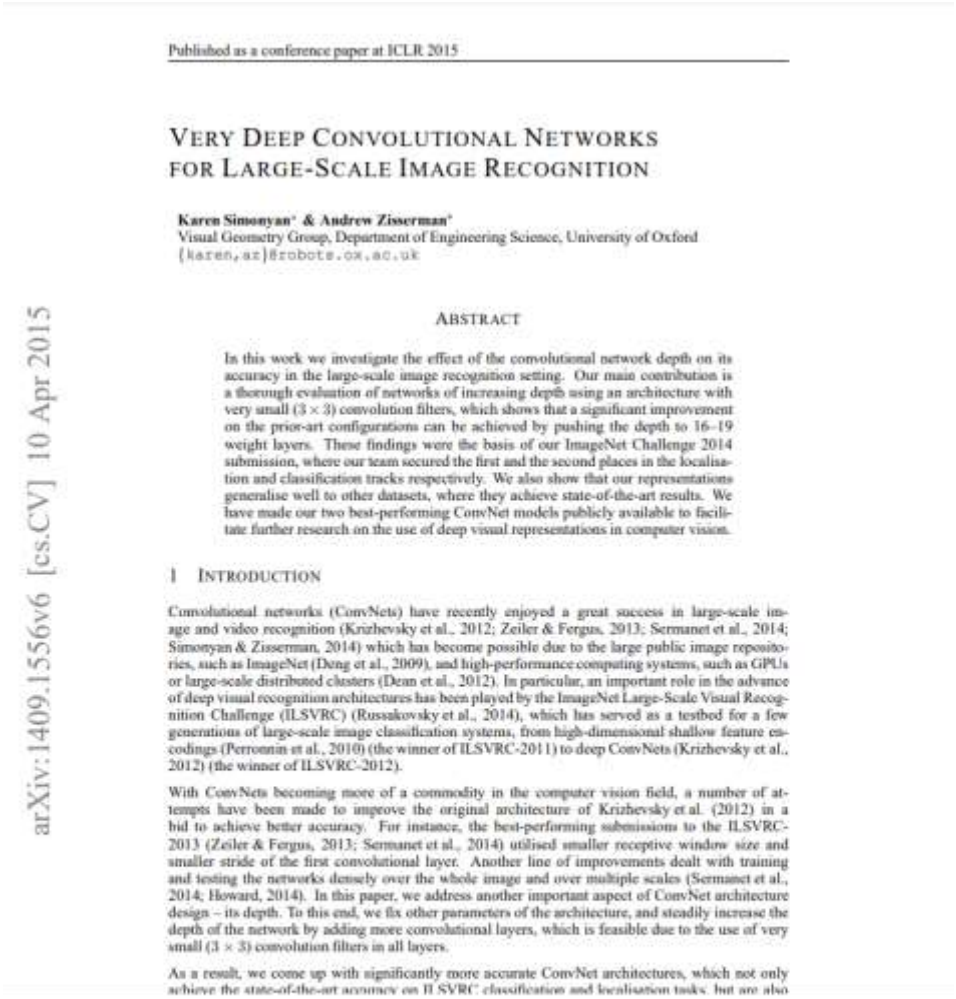
Title : Very deep convolutional networks for Large-Scale Image Recognition

Author : Karen Simonayan & Andrew Zisserman

Subjects : Computer Vision & Pattern Recognition

number of citations : 105,823 회 (23/07-17 기준)

Summary : Convolution network 의 depth 가 image 인식에 미치는 영향을 연구



Published as a conference paper at ICLR 2015

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.

1 INTRODUCTION

Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition (Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014; Simonyan & Zisserman, 2014) which has become possible due to the large public image repositories, such as ImageNet (Deng et al., 2009), and high-performance computing systems, such as GPUs or large-scale distributed clusters (Dean et al., 2012). In particular, an important role in the advance of deep visual recognition architectures has been played by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2014), which has served as a testbed for a few generations of large-scale image classification systems, from high-dimensional shallow feature encodings (Perronnin et al., 2010) (the winner of ILSVRC-2011) to deep ConvNets (Krizhevsky et al., 2012) (the winner of ILSVRC-2012).

With ConvNets becoming more of a commodity in the computer vision field, a number of attempts have been made to improve the original architecture of Krizhevsky et al. (2012) in a bid to achieve better accuracy. For instance, the best-performing submissions to the ILSVRC-2013 (Zeiler & Fergus, 2013; Sermanet et al., 2014) utilised smaller receptive window size and smaller stride of the first convolutional layer. Another line of improvements dealt with training and testing the networks densely over the whole image and over multiple scales (Sermanet et al., 2014; Howard, 2014). In this paper, 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.

As a result, we come up with significantly more accurate ConvNet architectures, which not only achieve the state-of-the-art accuracy on ILSVRC classification and localisation tasks, but are also

❖ ILSVRC history

- ImageNet Large Scale Visual Recognition Challenge(ILSVRC)는 Image Classification, Object detection, Segmentation 등 과 같은 **Computer Vision** 국제 대회임
- ImageNet 이라는 데이터셋을 기반으로 대회가 진행되며, Top-5 error, MAP와 같은 평가 지표를 사용하여 모델의 성능을 평가

  연구자들은 이 대회를 이용하여 이미지 모델을 개선하면서 많은 논문을 작성하고, 실적을 달성하고… 대학원생은 죽어가고…

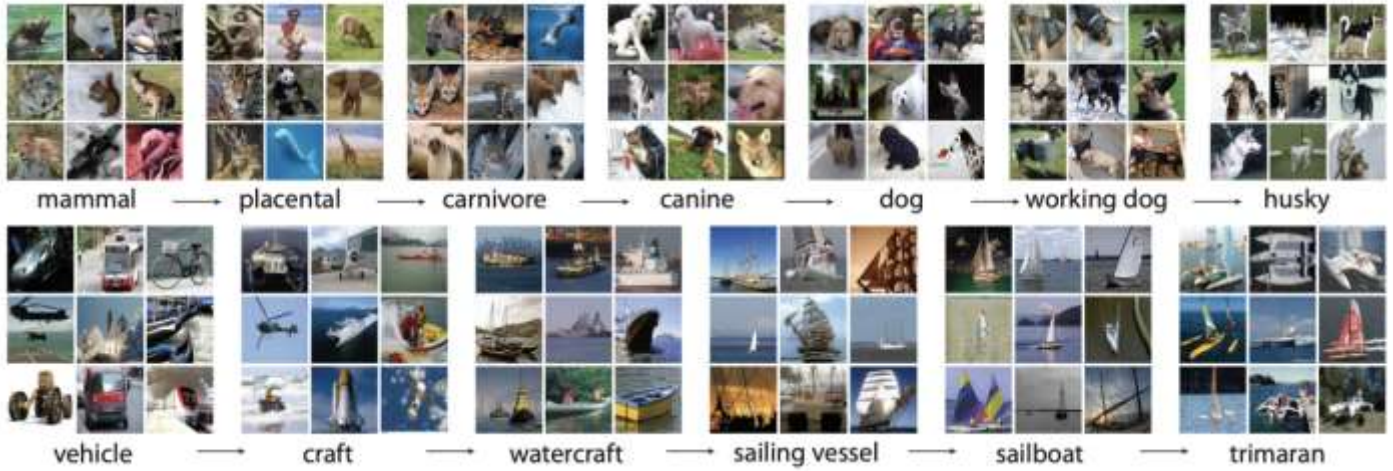
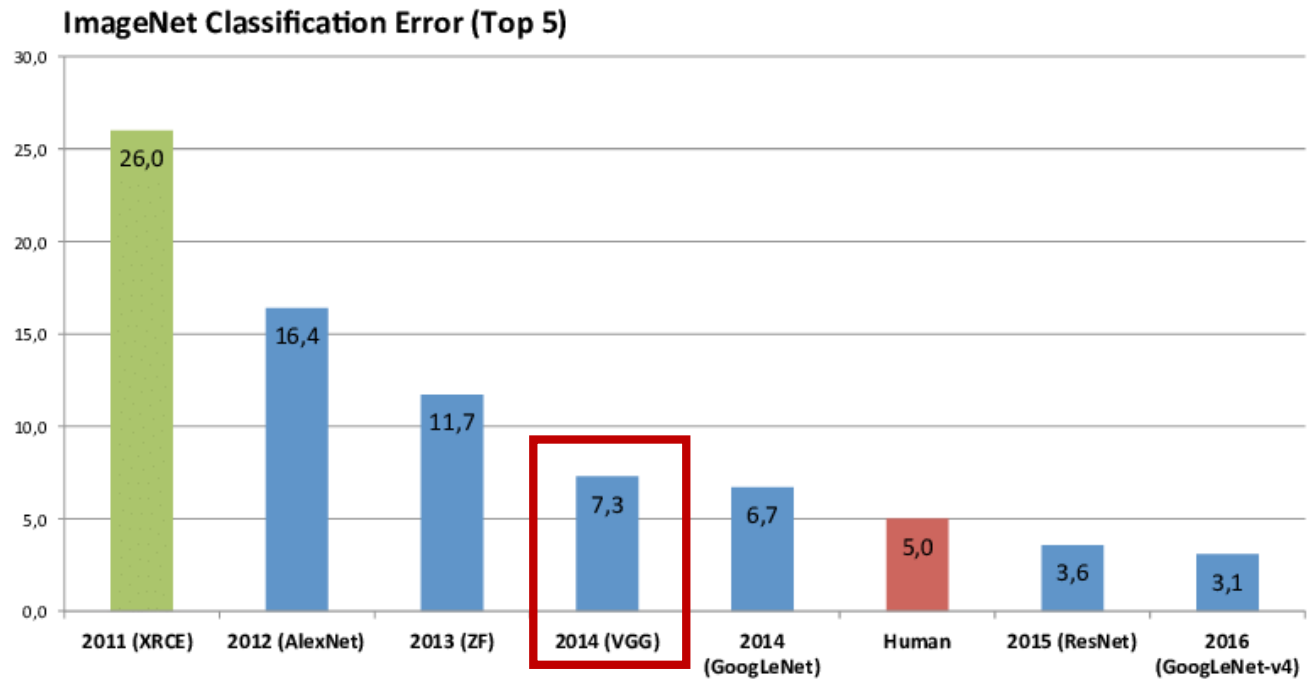


*Image Net: 수백만 개의 이미지와 수천 개의 카테고리로 구성된 대규모 이미지 데이터셋

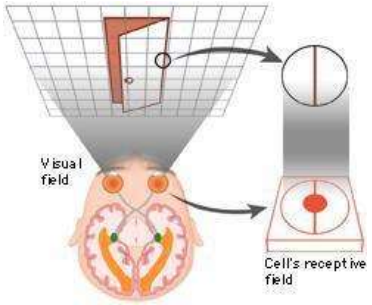
❖ ILSVRC history

- VGGNet 은 초기 등장했던 AlexNet(2012), ZFNet(2013) 보다 더 우수한 성능을 보였음
- 동시에 CNN 의 layer depth 가 모델 성능을 향상시키는 데 중요한 역할을 한다는 것을 증명하였음

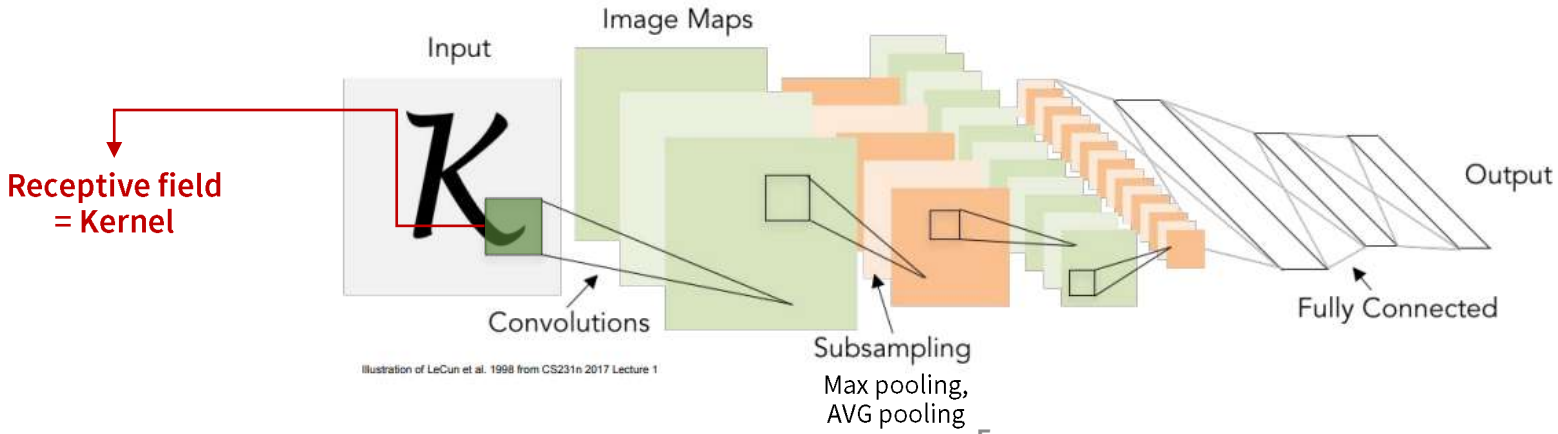


❖ Receptive Field & Convolution

- 사람의 눈으로 사물을 관찰할 때 일단은 local receptive field(국부 수용영역)를 보게 되며, 이러한 receptive field가 서로 겹치게 되면 사물 전체를 관찰하게 됨(저 수준의 패턴의 조합으로 복잡한 패턴을 인식)

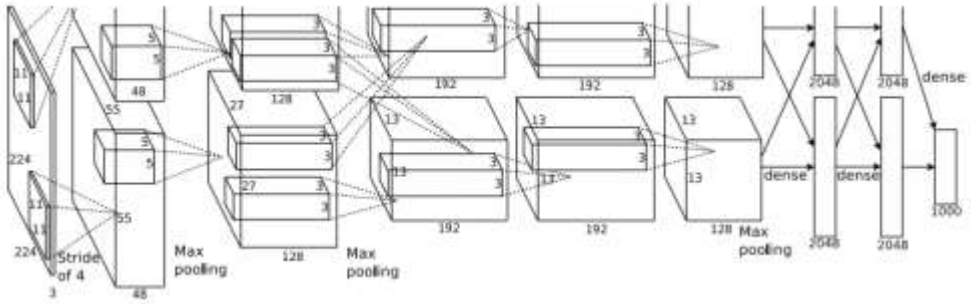


- 이는 CNN의 기원이 되었고, 이미지에서 Window를 통해 전체 이미지의 특징을 추출하는 Convolution이 등장



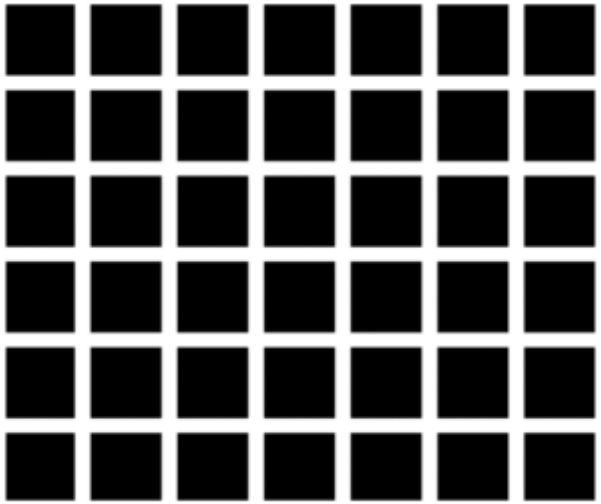
❖ AlexNet

- AlexNet 은 두 개의 GPU를 사용하여 연산을 분할하였음(90개의 kernel 이 있을 때 45개, 45개로 나누어서 병렬적 연산 진행)
- Local Response Normalization(LRN) 도입: 한 픽셀의 값이 너무 커서 주변 픽셀을 모델이 보지 못하는 현상을 예방



Receptive field: (11x11, stride = 4)

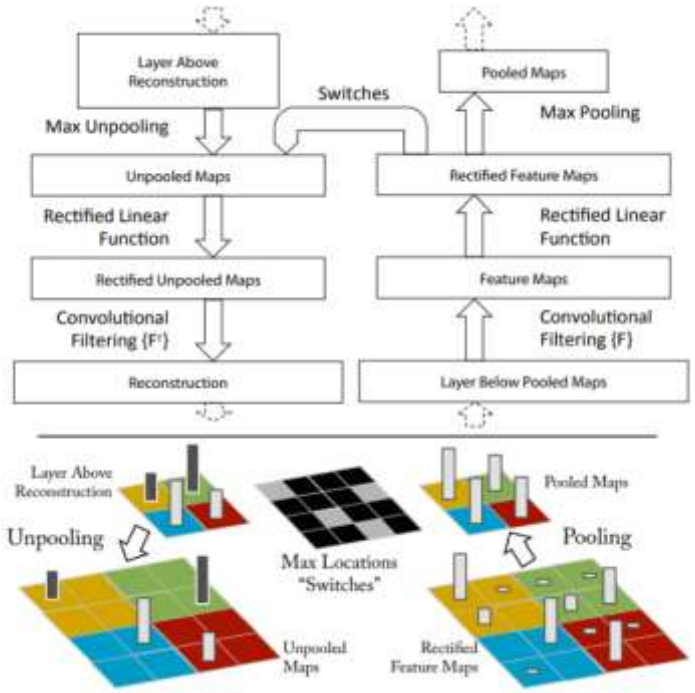
$$b^i_{x,y} = \frac{a^i_{x,y}}{\left(k + \alpha \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (a^j_{x,y})^2\right)^\beta}$$



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

❖ ZFNet

- Alexnet 에서 1개의 GPU를 사용하고, conv layer의 kernel size, stride를 일부 조정
- ZFNet 은 CNN의 동작 과정을 가시화하여 CNN을 이해하는 데 중요한 역할을 함

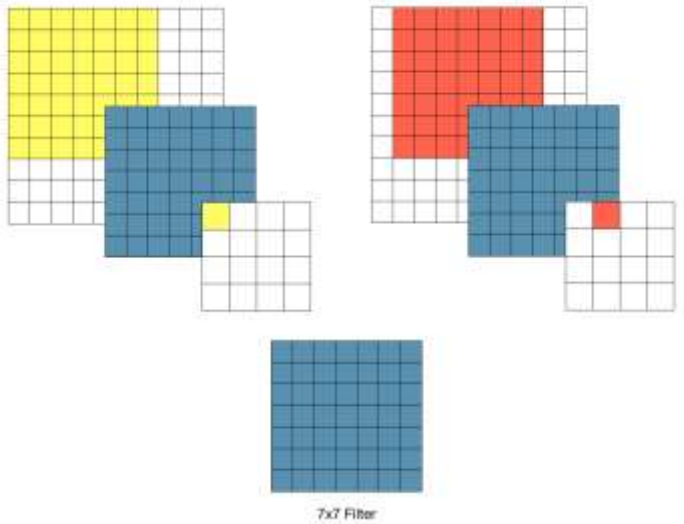


Receptive field: (7x7, stride = 2)

Layer 4

Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing, 2014.

❖ Convolution Factorization



Receptive Field의 사이즈가 크다라는 것은 한 픽셀과 주변 픽셀 사이의 관계를 모델이 쉽게 파악할 수 있음을 의미

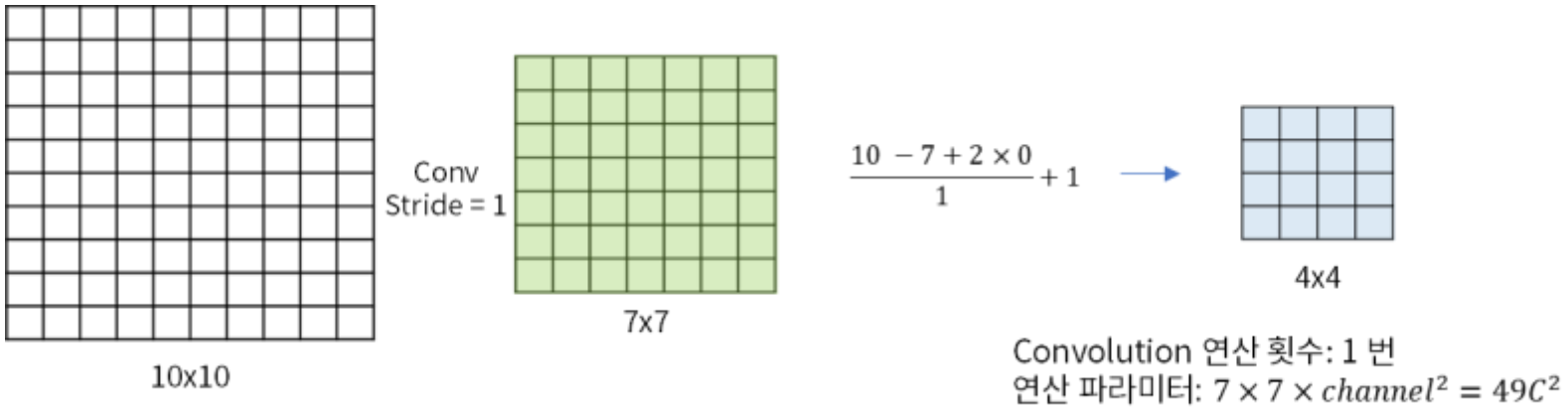
또한 Receptive Field 의 사이즈가 크다면 Stride 를 작게 잡아도 Convolution 연산을 빠르게 끝낼 수 있음

→ 하지만 VGGNet은 AlexNet, ZFNet 에 비해 비교적 적은 3x3 필터만 사용했음에도 이미지 분류 정확도를 **비약적으로 개선**하였음

❖ Convolution Factorization

- Convolution Factorization 은 컨볼루션 연산을 더 작은 크기의 커널로 분해하여 계산 효율성을 높이는 트릭임

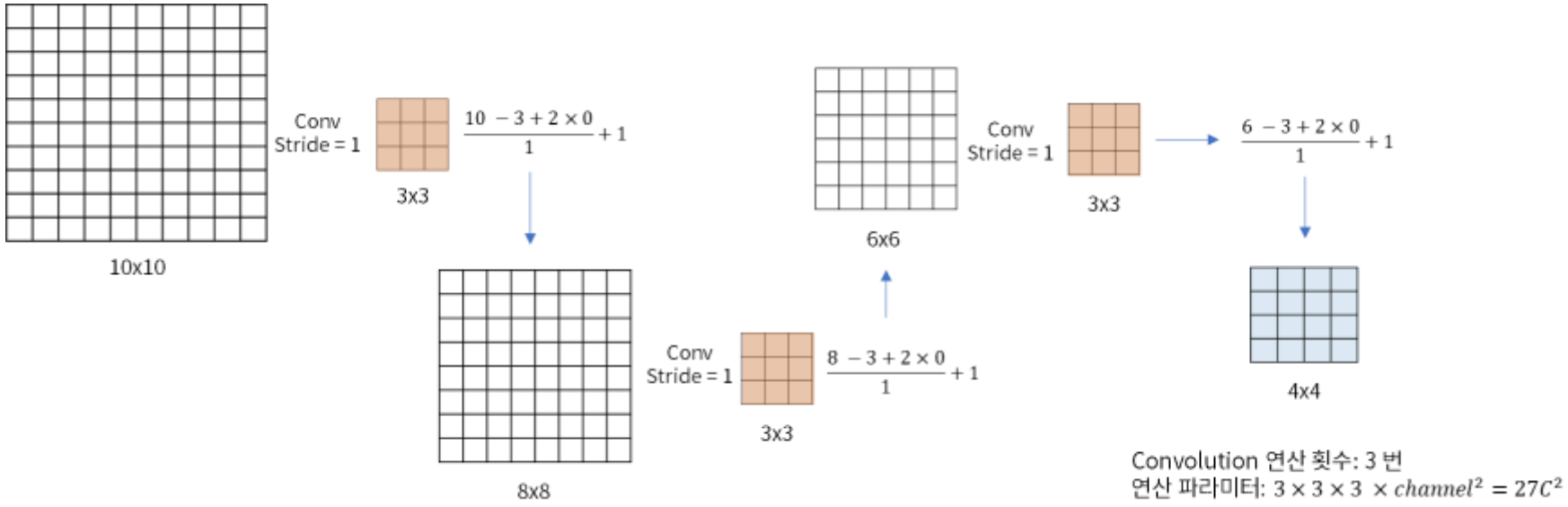
  ex) 10 x 10 이미지를 7x7 receptive field로 conv 연산하여 4x4 feature map을 얻는 경우



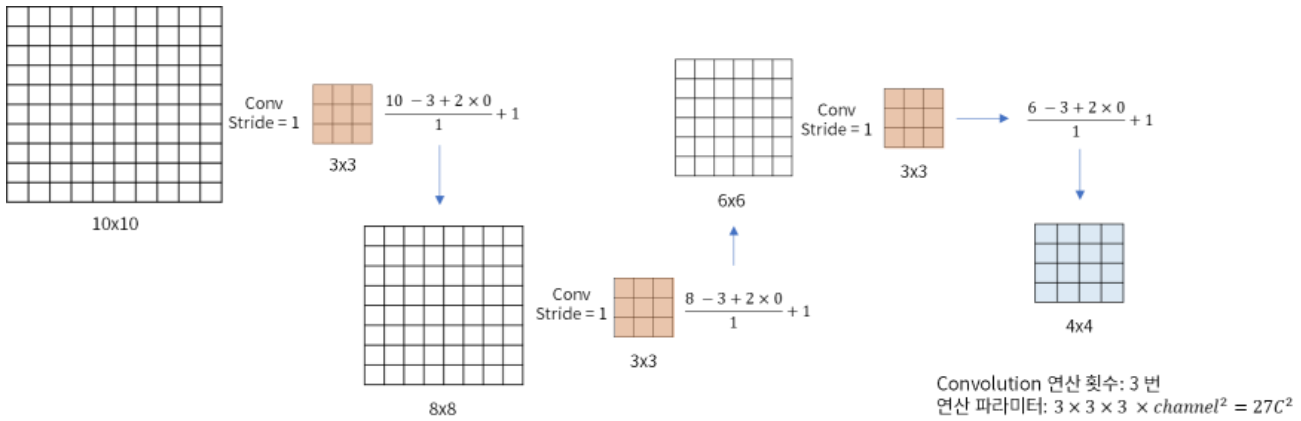
Convolution 연산 횟수: 1 번
연산 파라미터: $7 \times 7 \times channel^2 = 49C^2$

❖ Convolution Factorization

ex) 10 x 10 이미지를 3x3 receptive field로 conv 연산하여 4x4 feature map을 얻는 경우



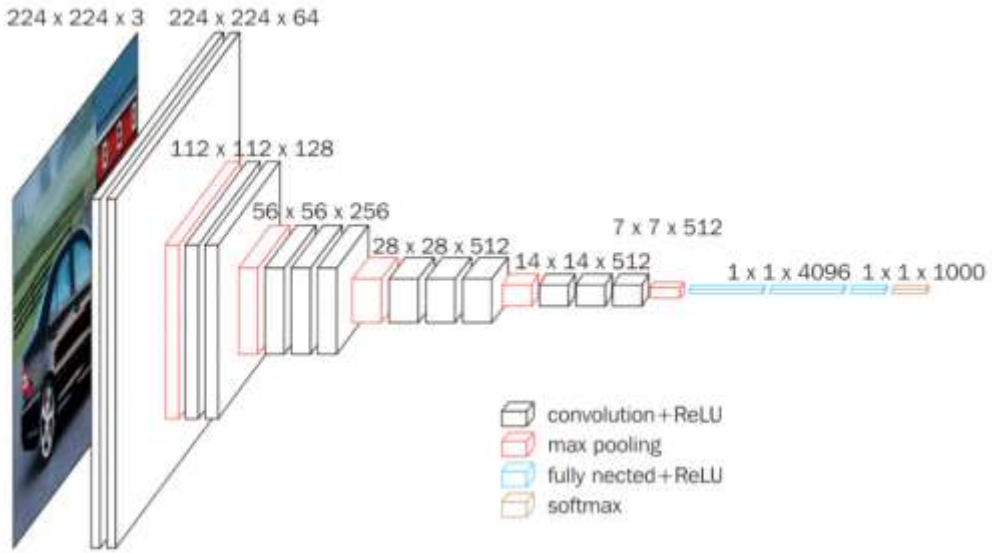
❖ Convolution Factorization



- Stride 가 1 일 때, 3차례 3x3 conv 필터링을 반복한 특징맵은 원본 이미지의 **7x7 Receptive field 와 같은 효과**를 볼 수 있음
- **결정 함수의 비선형성 증가 효과**를 얻을 수 있음
  - ✓ Convolution 연산은 ReLU 함수를 포함하기 때문
- **학습 파라미터 수 감소**
  - ✓ 3x3 conv를 3번 반복하는 것보다 7x7 conv 1번 사용하는 것이 81(49/27)%정도 더 연산

❖ VGGNet

- VGGNet 은 3x3 Receptive Field 를 사용하되 신경망을 깊게 쌓은 구조
  - ✓ 기존에 제안된 모델 보다 파라미터의 수를 효과적으로 줄일 수 있으며
  - ✓ 층이 깊어질 수록 활성함수가 많이 적용되어 Task를 discriminative 하게 만들 수 있음



❖ Classification Framework - Training

- **Batch Size : 256**
- **Momentum : 0.9**
- **L2 Regularization : 0.0005**
- **Learning Rate : 0.01**
- **Input Image : 224 X 224 X 3**

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					

❖ Classification Framework - Training

- **Single-scale training**: Resizing image Size(fix) S (256 or 384)
- **Multi-scale training**: The input image is individually rescaled by randomly sampling S from a certain range $[S_{min}, S_{max}]$(we used $S_{min} = 256, S_{max} = 512$).
  - ✓ This can also be seen as **training set augmentation** by scale jittering

256x256

512x512

변환된 이미지가 작을수록 개체의 전체적인 측면을 학습할 수 있고,
변환된 이미지가 클수록 개체의 특정한 부분을 학습에 반영할 수 있음

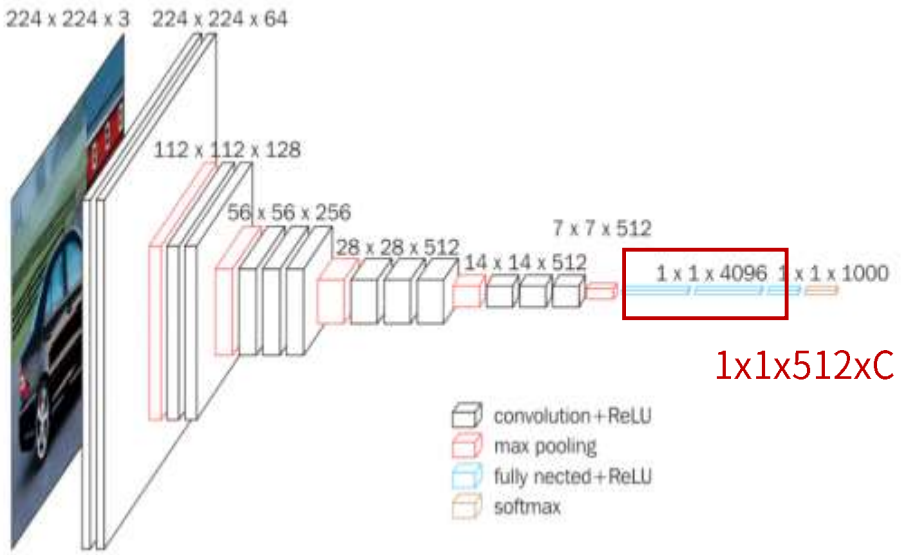
❖ Classification Framework - Testing

- 입력되는 이미지의 크기가 학습하는 과정에서 바뀌기 때문에 FC layer의 차원을 매번 바꿔줘야하는 문제

  → FC layer 를 1x1 convolution layer로 대체 (차원에 제약을 제거)

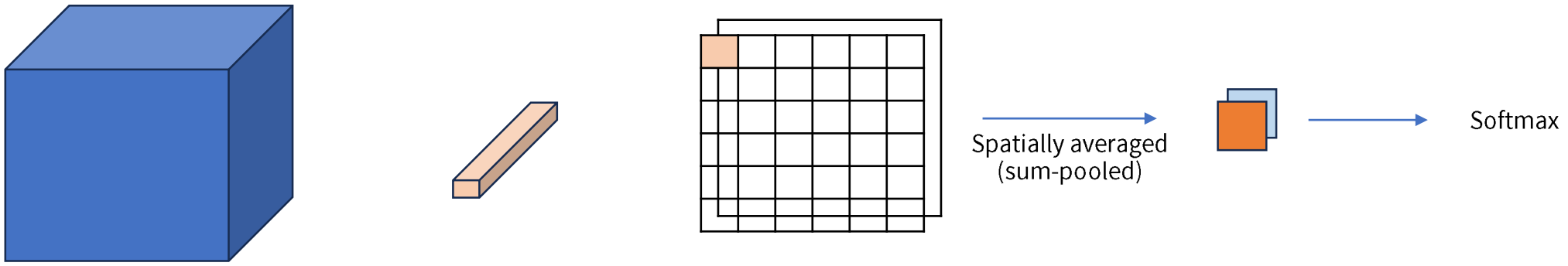
  이에 따라 하나의 입력 이미지를 다양한 스케일로 사용한 결과들을 앙상블하여 이미지 분류 정확도를 개선하는 것도 가능

- 이때 1x1 conv의 채널의 개수는 클래스의 개수로 설정
  - 1x1 conv 의 결과로 나온 Feature map 을 spatial average 후
  - 채널 방향으로 Softmax를 태운다면 바로 이미지 분류가 가능



❖ Classification Framework - Testing

- **1x1 Convolution & Testing**
  - ✓ 채널의 수를 줄이고 싶을 때 사용
  - ✓ 6 x 6 x 10 이미지를 6 x 6 x 2 로 줄이기 위해서
  - ✓ 1 x 1 x 10 필터를 2개 사용하면 됨



6 x 6 x 10 (1 x 1 x 10) x 2 6 x 6 x 2

❖ Classification Experiments

▪ 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**

top-1:오분류율
top-5: Top 5 예측 범주 내 정답 클래스가 속하지 않은 비율

❖ Classification Experiments

- **Multi-Scale Evaluation**

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**

❖ Discussion

- 본 논문이 가장 크게 기여한 부분은?
- 본 논문에서 아쉬운 부분은?
- 본 논문의 핵심은?