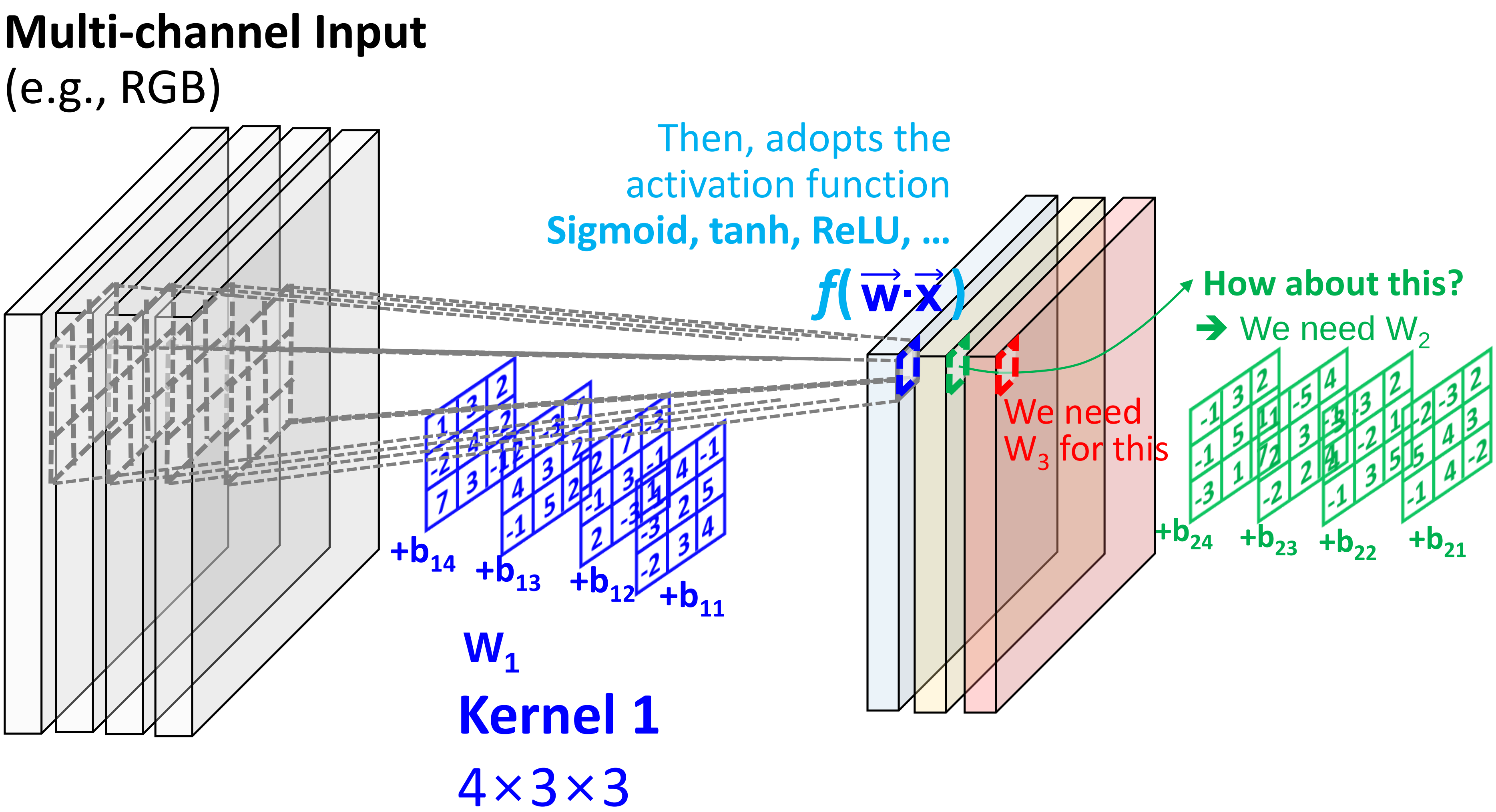
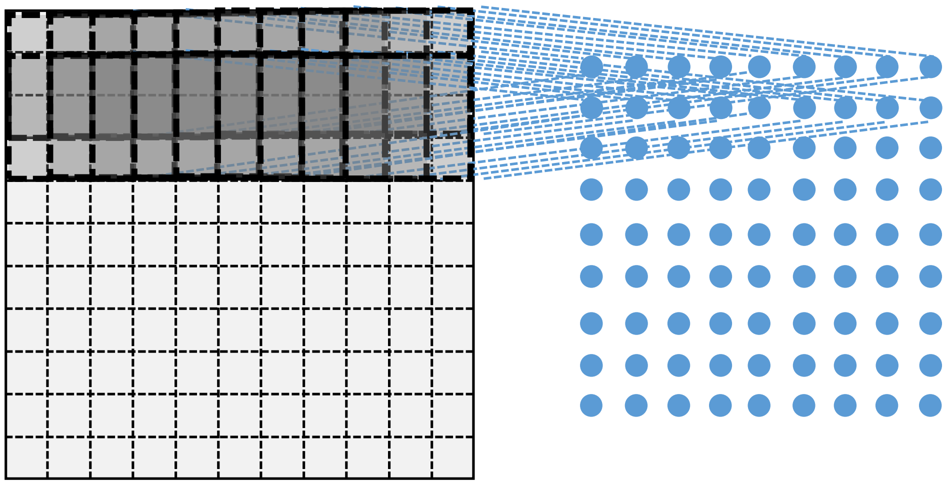
Revisit the Convolution Layer
Multi-Channel Input



Key Hyper-parameters in Convolution Layer

- Kenel or filter size? If input layer has *k* channel,
  - W can be k×h×h, where h = 2, 3, 4, 5, …
- Stride; degree of overlap

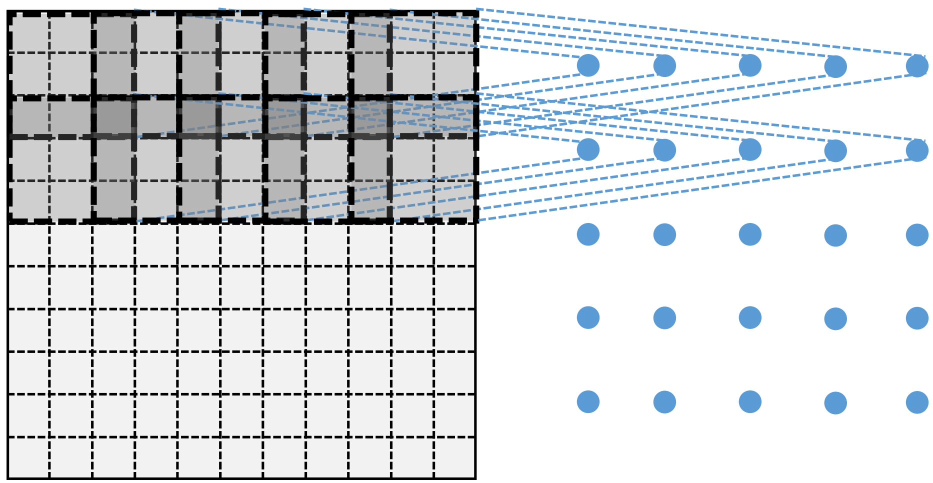
Stride = 1



11 x 11

9 x 9

Stride = 2

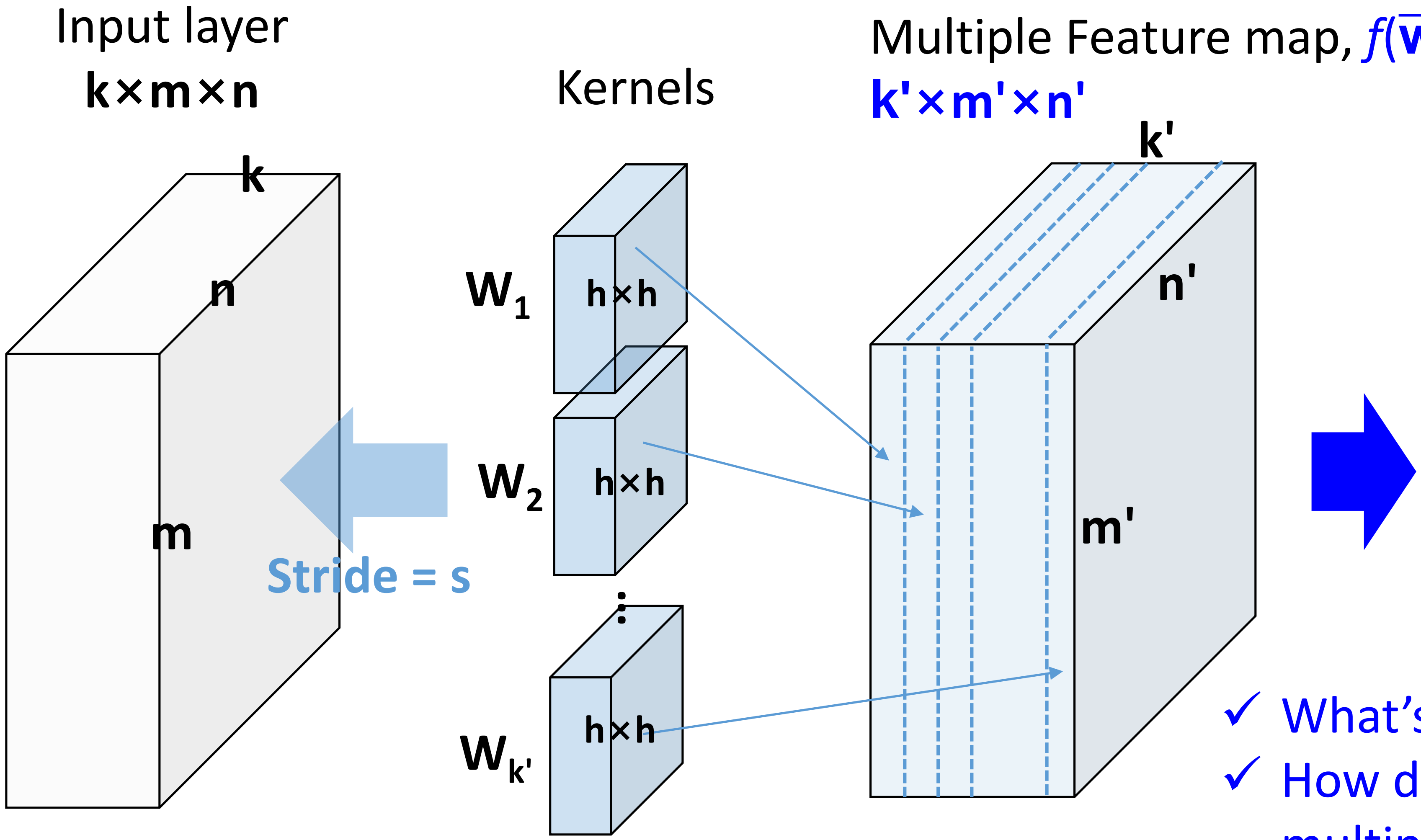


11 x 11

5 x 5

You see down-sampling effect?

Generalized Convolution Layer for CNN



- ✓ What's next?
- ✓ How do we construct multiple layers?

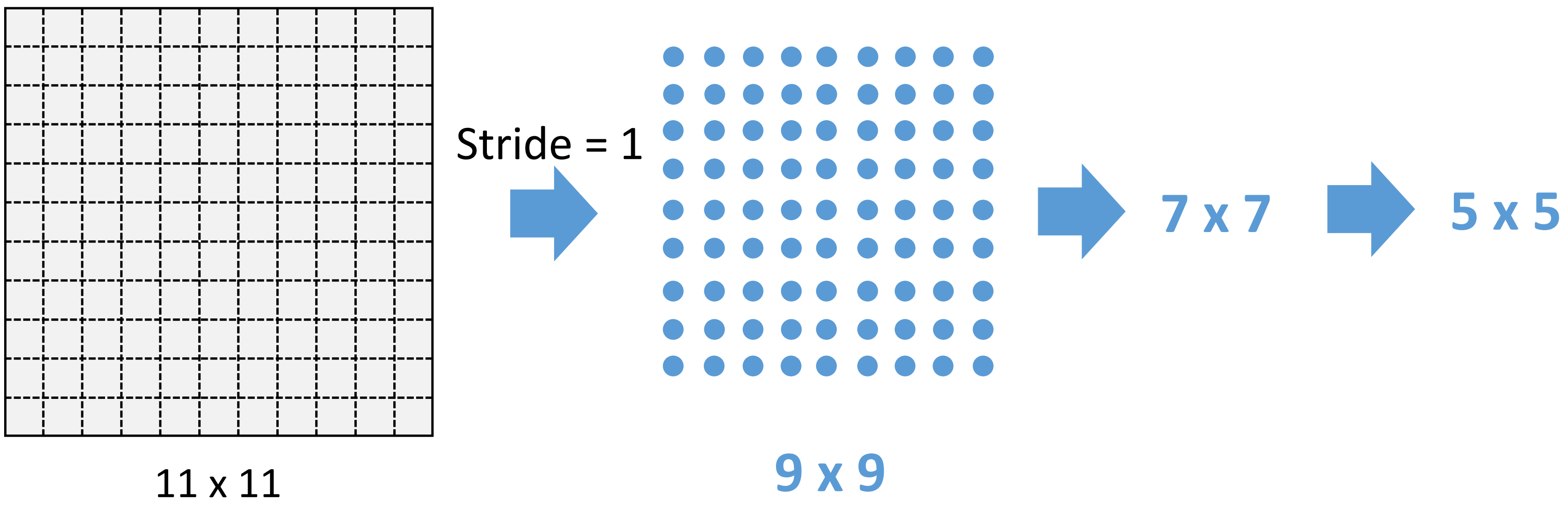
Convolutional Neural Network (2)

Hanwool Jeong

hwjeong@kw.ac.kr

Data Loss in CNN

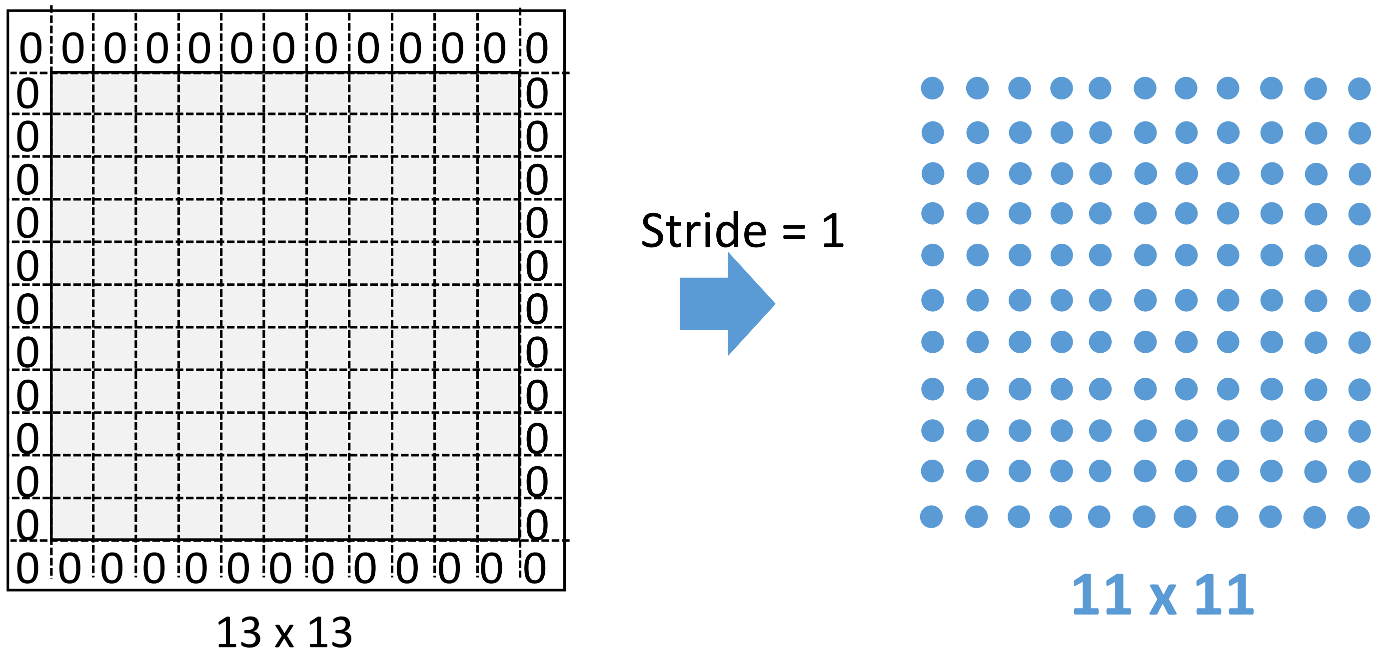
- Edge data component may have less portion



- To prevent this data size reduction, the padding can be performed.

Padding in CNN

- Besides the zero padding, copy padding exists.
- Can you imagine different p value?



p=2

We can retain the data size

Pooling or Subsampling Layer

- To extract summary statistics, pooling can be used.

➔ Accelerate the speed, reduce the noise effect, overfitting..

- Besides the max pooling, average, weighted average pooling or L_2 norm pooling.

0	5	4	3	0	0
8	2	0	0	6	9
9	0	0	6	4	0
8	6	0	8	9	6
4	0	6	9	0	2
0	7	6	4	10	7

s=2

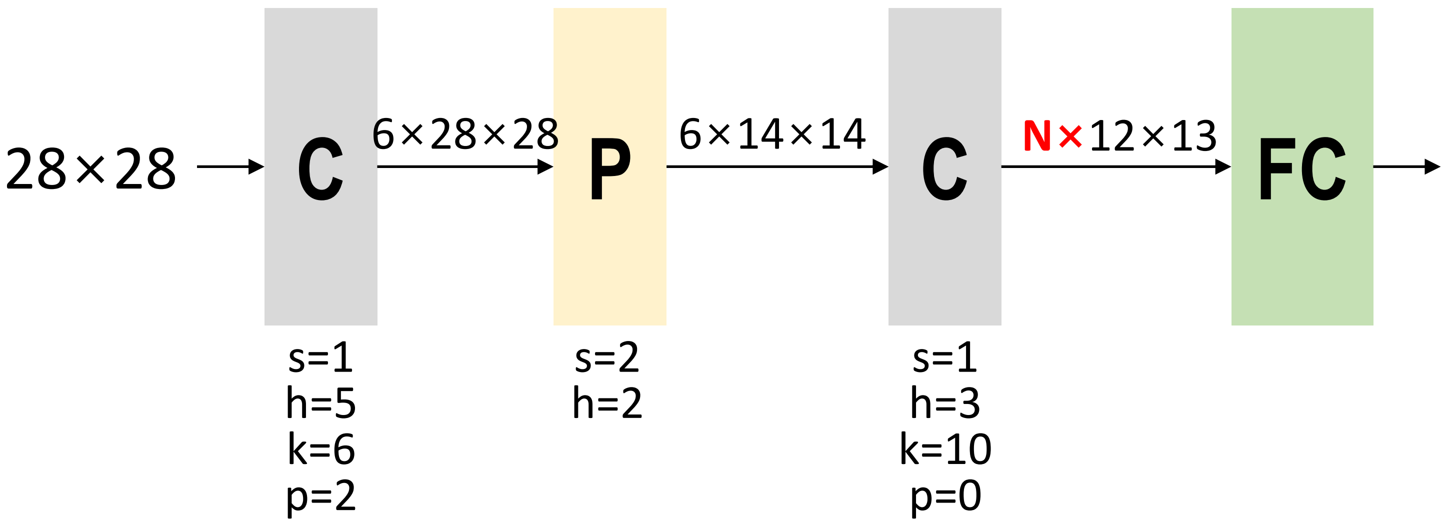
Max pooling

8	4	9
9	8	9
7	9	10

Now, We have Various Layers

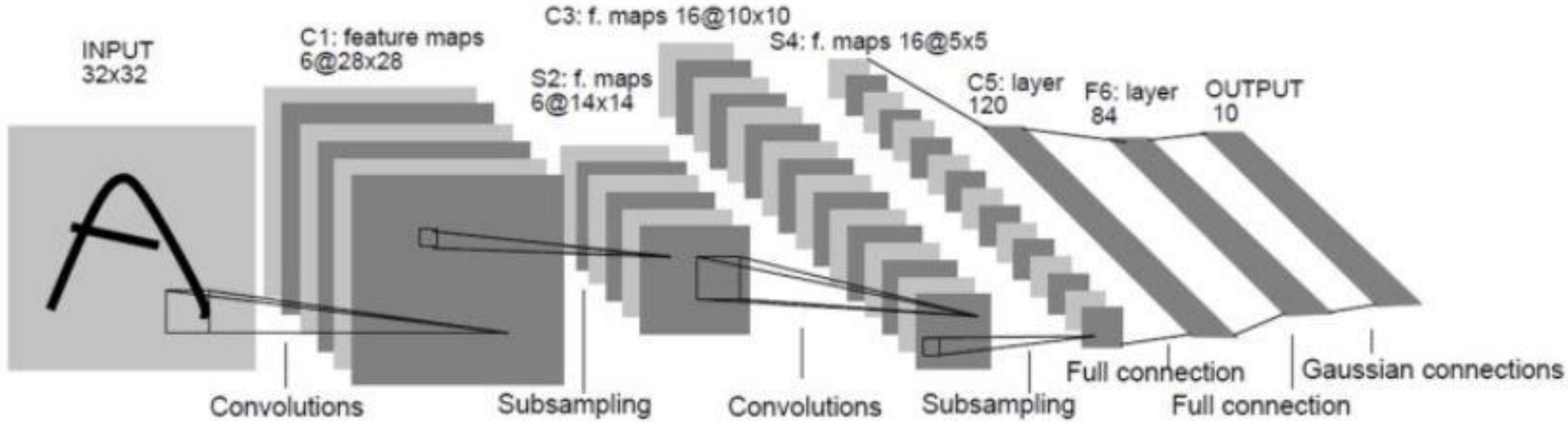
- Fully connected layer (dense layer)
- Convolution layer
- Pooling layer

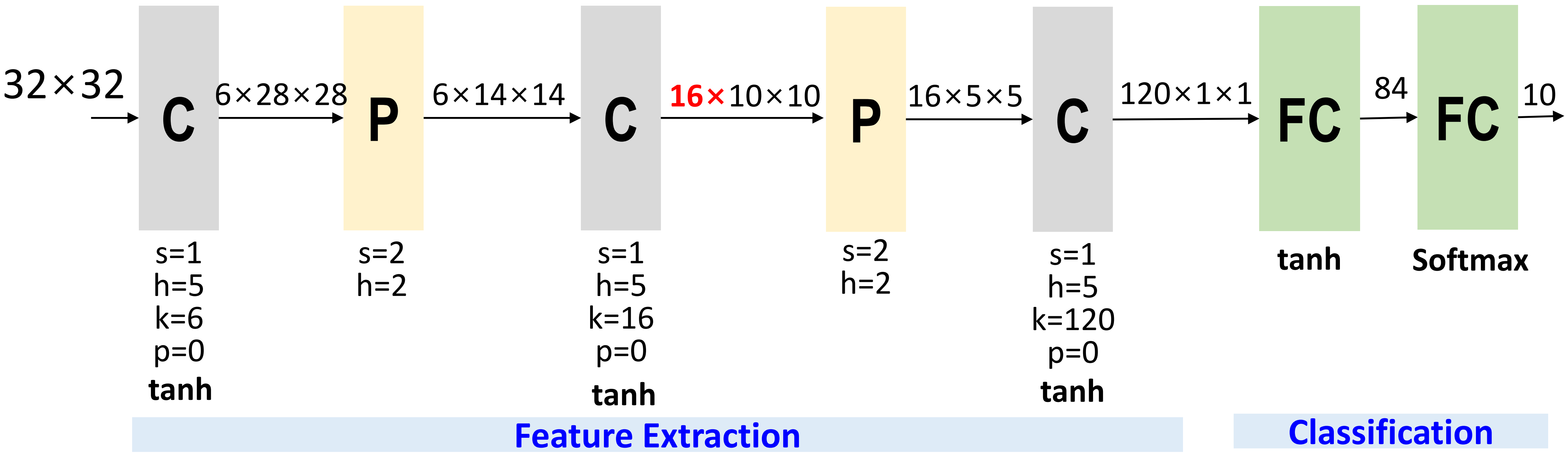
By combining above layers, we can construct multiple layers for certain goal



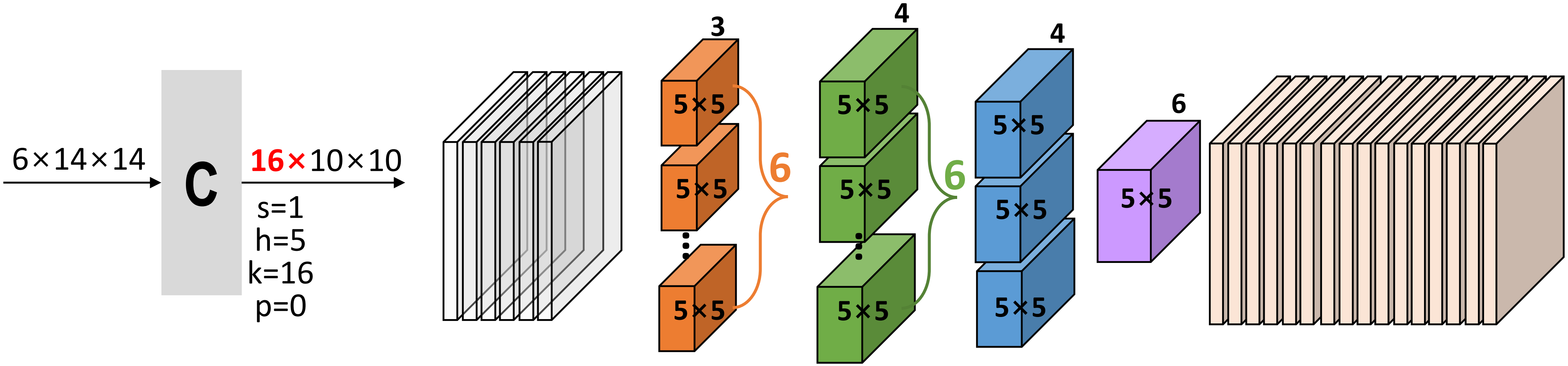
LeNet-5 (1998)

Gradient-based learning applied to document recognition
Y LeCun, L Bottou, Y Bengio… - Proceedings of the …, 1998 - ieeexplore.ieee.org
Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful **gradient based learning** technique. Given an appropriate network architecture, **gradient-based learning** algorithms can be used to synthesize a complex …
☆ 99 36854회 인용 관련 학술자료 전체 36개의 버전





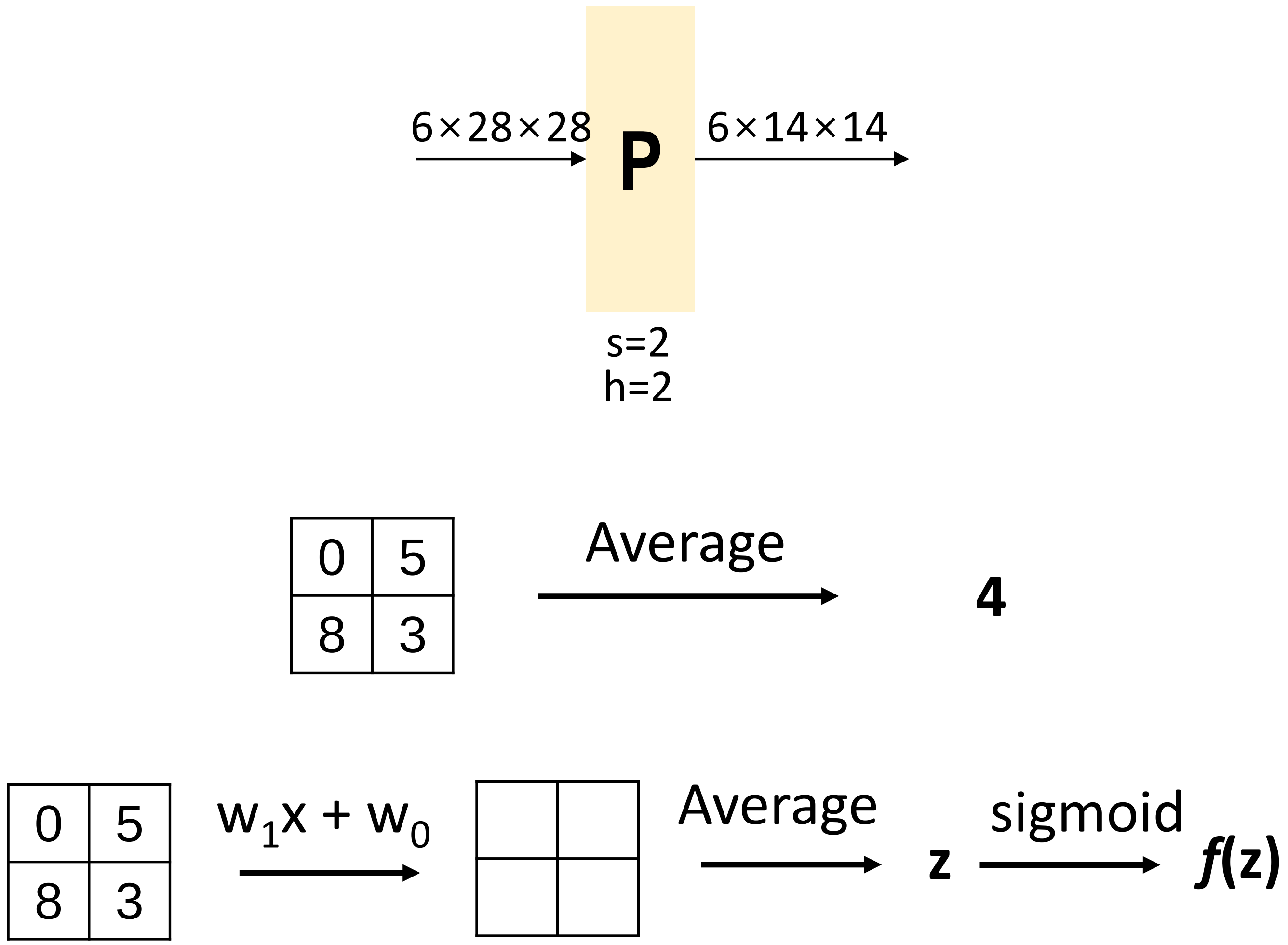
6 → 16?



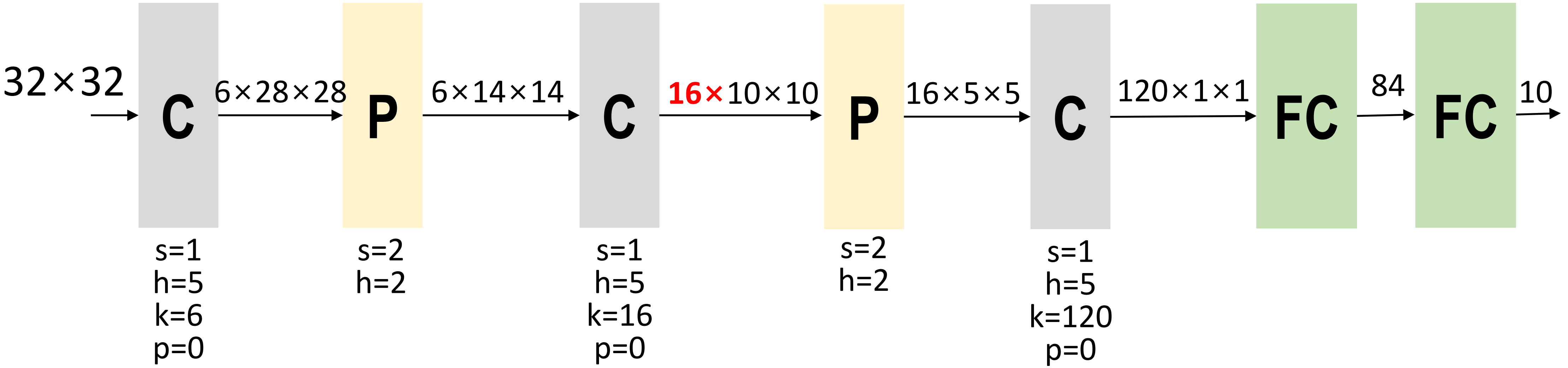
OUT

IN	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0																
1																
2																
3																
4																
5																

Enhancing Flexibility & Utility in Average Pooling Layer

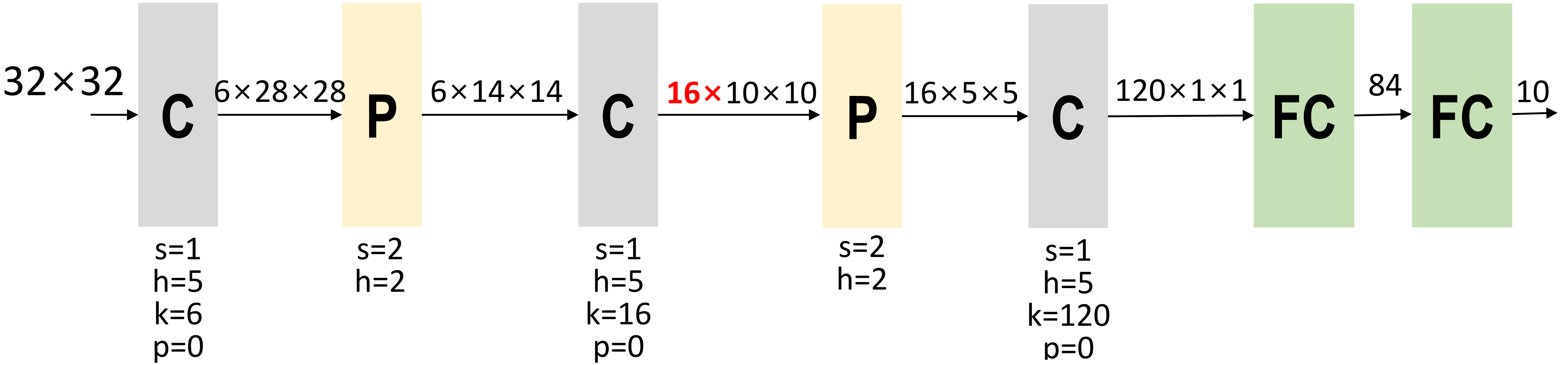


How Can You Train This?



- We already study this!
- Nothing but with the proper optimizer & cost function!
- Then we can perform back-propagation

How Many Parameters to be Trained?



$6\times(5\times5+1)$ **=156**

6×2 **=12**

$(5\times5\times3+1)\times6$
$+ (5\times5\times4+1)\times6$
$+ (5\times5\times4+1)\times3$
$+ (5\times5\times6+1)\times1$
=1516

16×2 **=32**

$(5\times5\times16+1)\times120$ **=48120**

$(120+1)\times84$ **=10164**

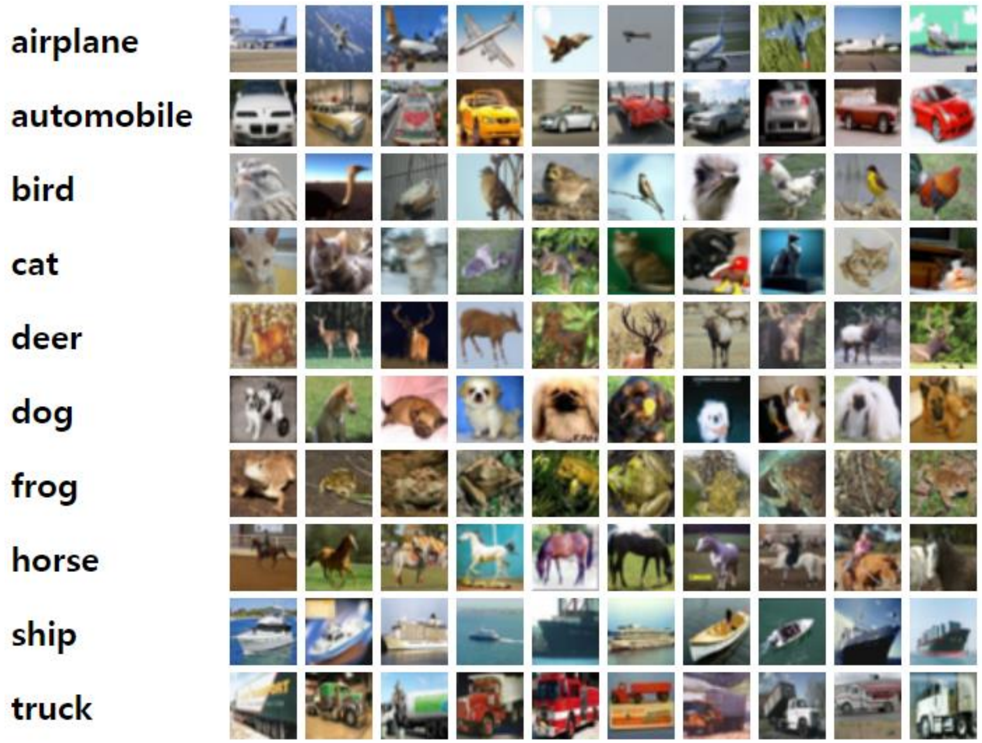
$(84+1)\times10$ **=850**

Application Cases of LeNet-5

- Recognizing the simple digit images
- Successfully applied to the recognition of handwritten zip code digits provided by the U.S. Postal Service.
- Contributes to the automating classifying the price of the bank check.

Famous Datasets

- MNIST: 28×28 handwritten 0-9 digits (60000 training + 10000 test)
- CIFAR-10: 32×32 color images in 10 classes (50000 training + 10000 test)

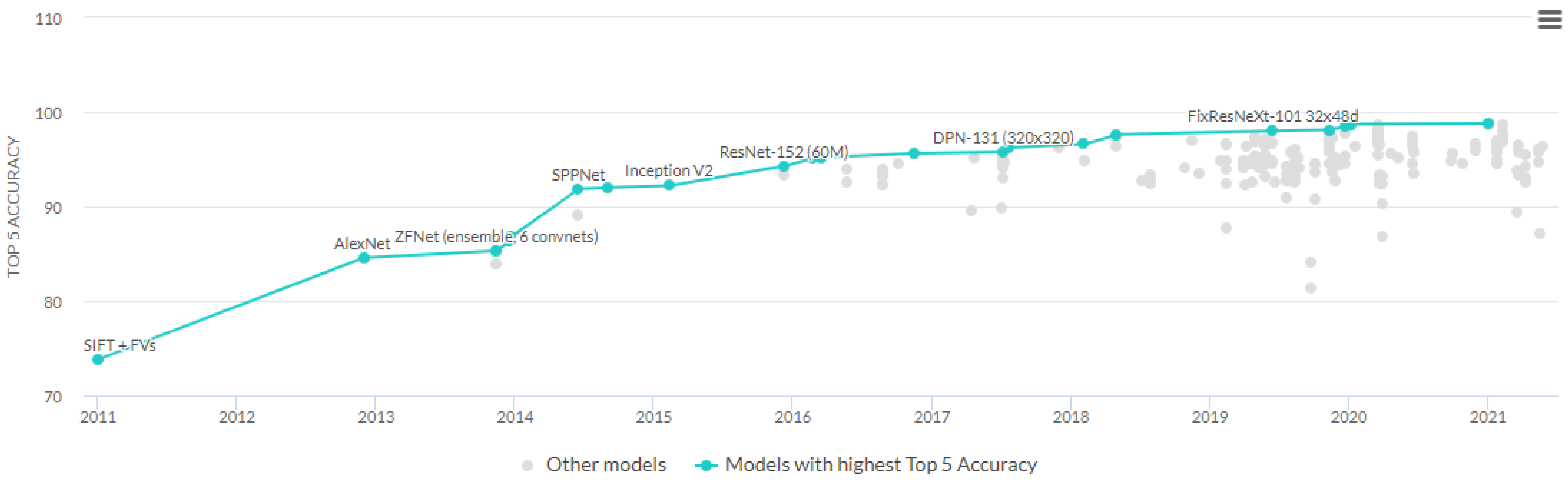


ImageNet & ILSVRC

- ImageNet: More than 14 million images have been hand-annotated by MTurk(the Amazon Mechanical Turk) into more than 20,000 categories
- ILSVRC (ImageNet Large Scale Visual Recognition Challenge)

ILSVRC

- https://paperswithcode.com/sota/image-classification-on-imagenet



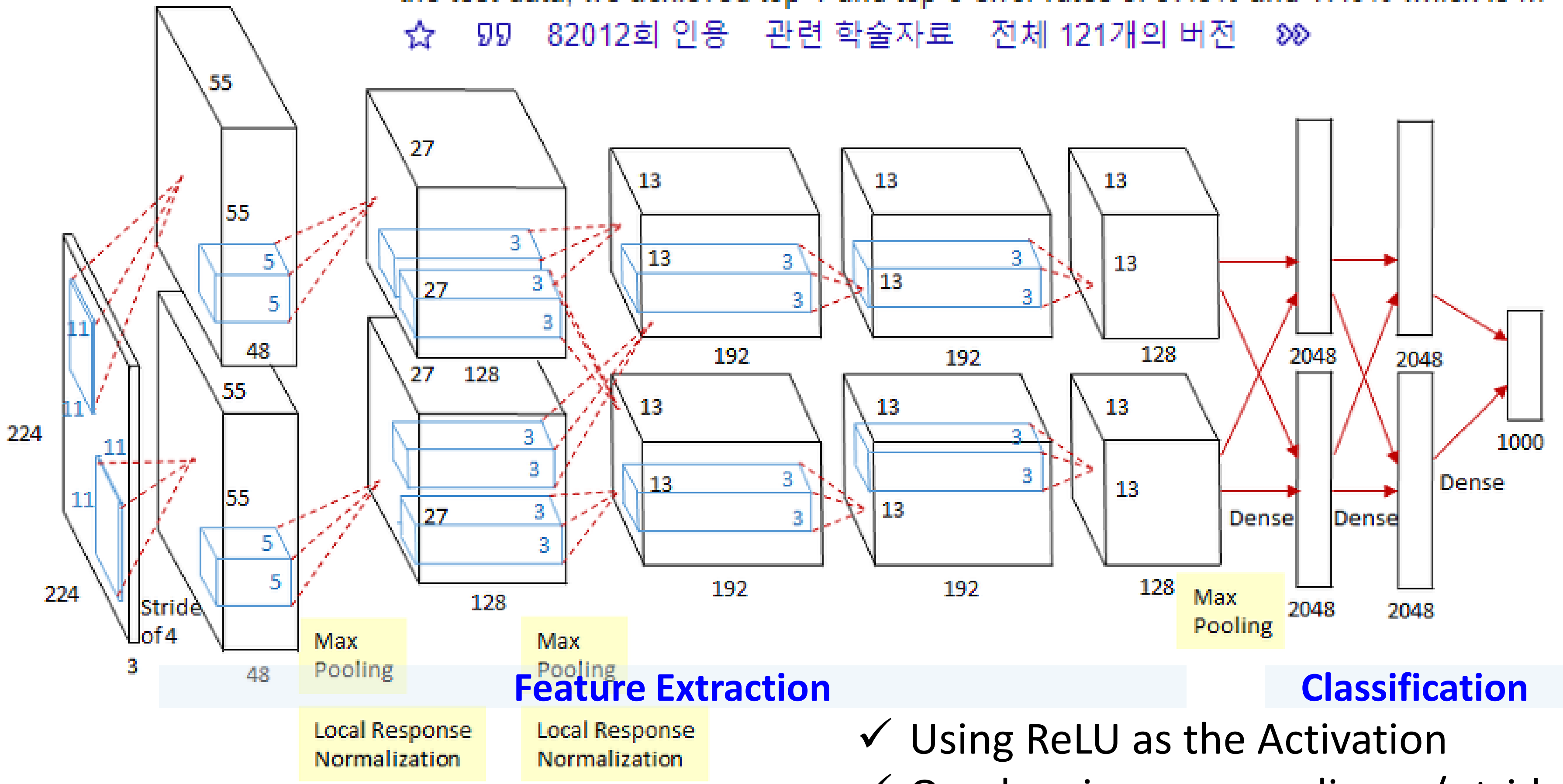
AlexNet

[PDF] **Imagenet classification** with **deep convolutional neural networks**

A Krizhevsky, I Sutskever, GE Hinton - Advances in **neural** information ..., 2012 - kr.nvidia.com

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is ...

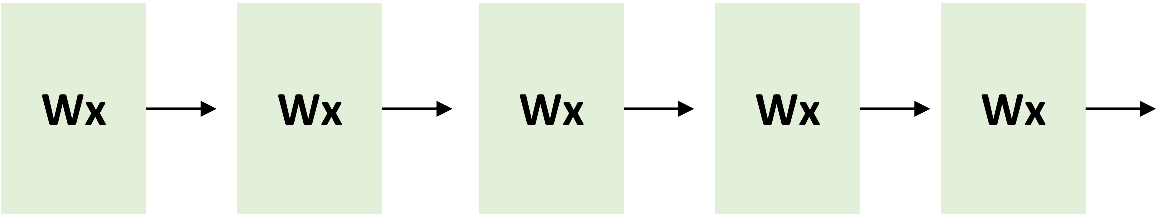
☆ 99 82012회 인용 관련 학술자료 전체 121개의 버전 »



- ✓ Using ReLU as the Activation
- ✓ Overlapping max-pooling w/ stride =2
- ✓ Using local response normalization
- ✓ Using drop-out

Why is Deep Layer Good?

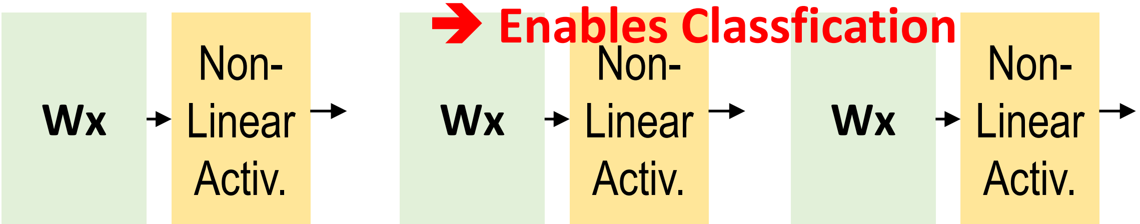
Multiple layers of linear combination:



- ✓ No vanishing gradient
- ✓ No use for deep layer

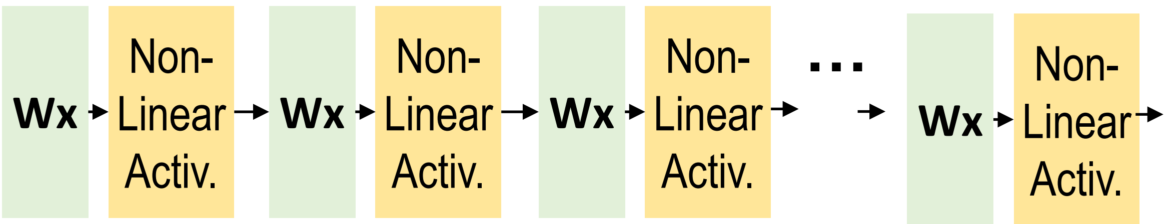
Deep MLP:

Non-linear function + Many layer
➔ Enables Classfication



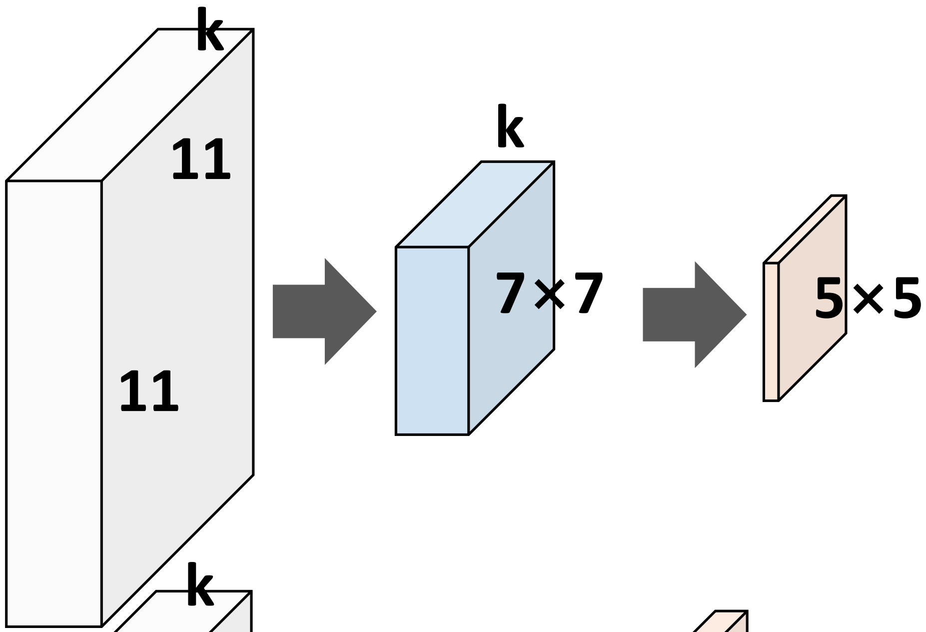
- ✓ Limited number of layers
- ✓ Vanishing gradient
- ✓ Overfitting

CNN: Partial connection + Down sampling + Drop-out + ReLU

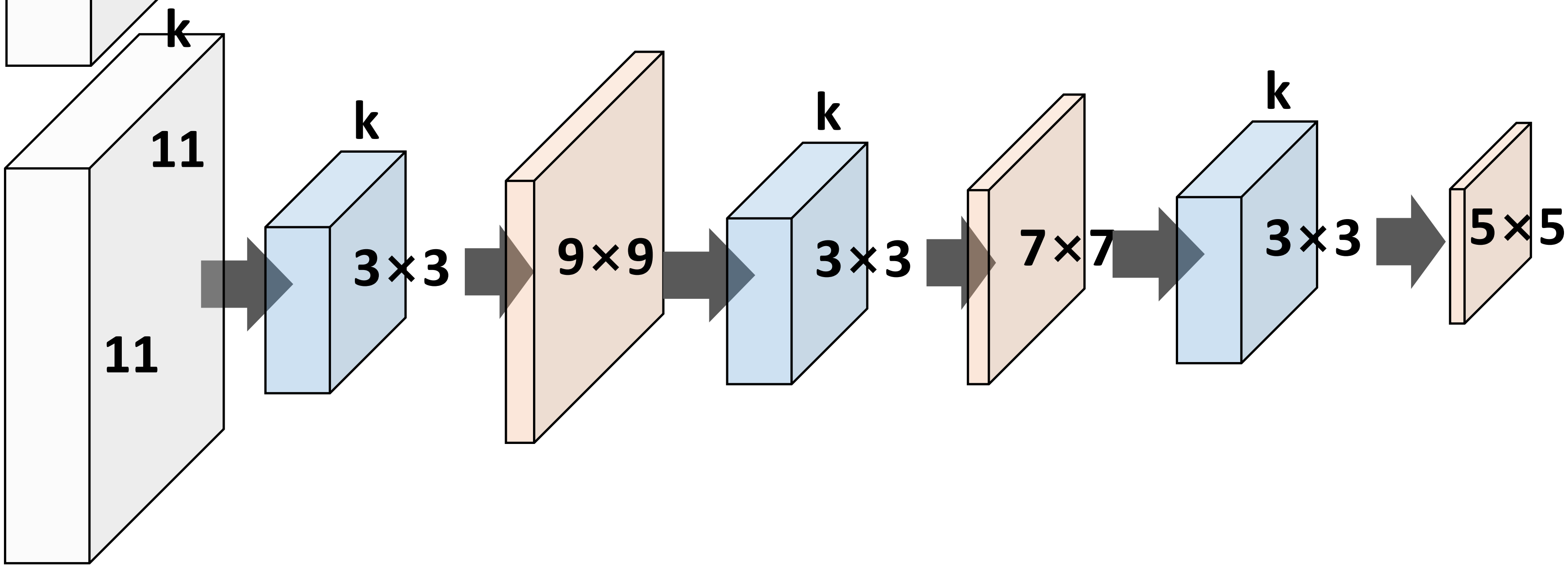


- ✓ Can use more layer
- ➔ better classification capability
- ✓ Less calculation and parameters
- ✓ Less vanishing gradient
- ✓ Less overfitting

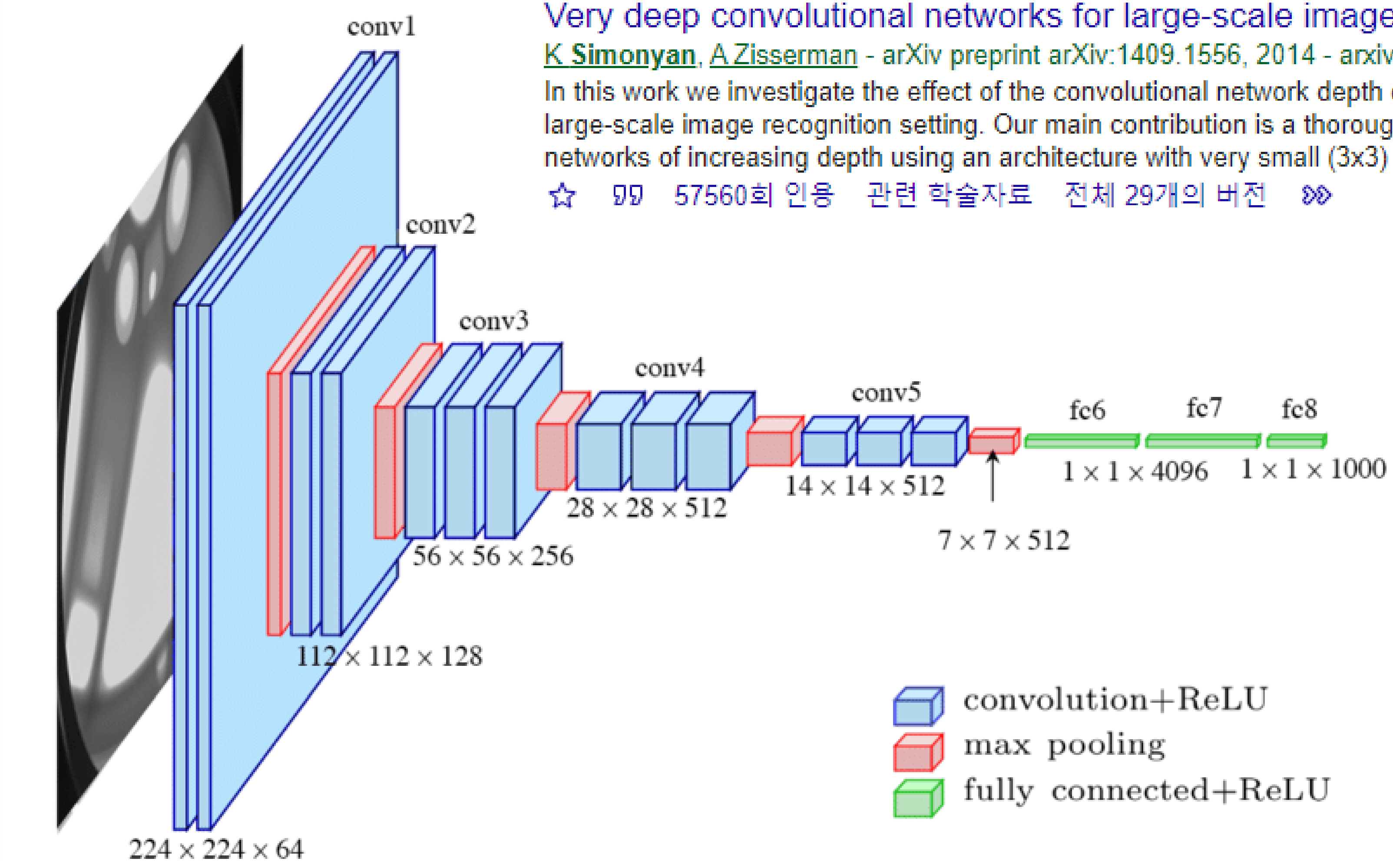
7×7 Kernel vs. 3×3 Kernel



Same effect on
7×7 one time vs. 3×3 three times
Except more non-linear activations
can be applied in the latter
➔ Better accuracy in 3×3 three times



VGGNet; VGG-16



Very deep convolutional networks for large-scale image recognition

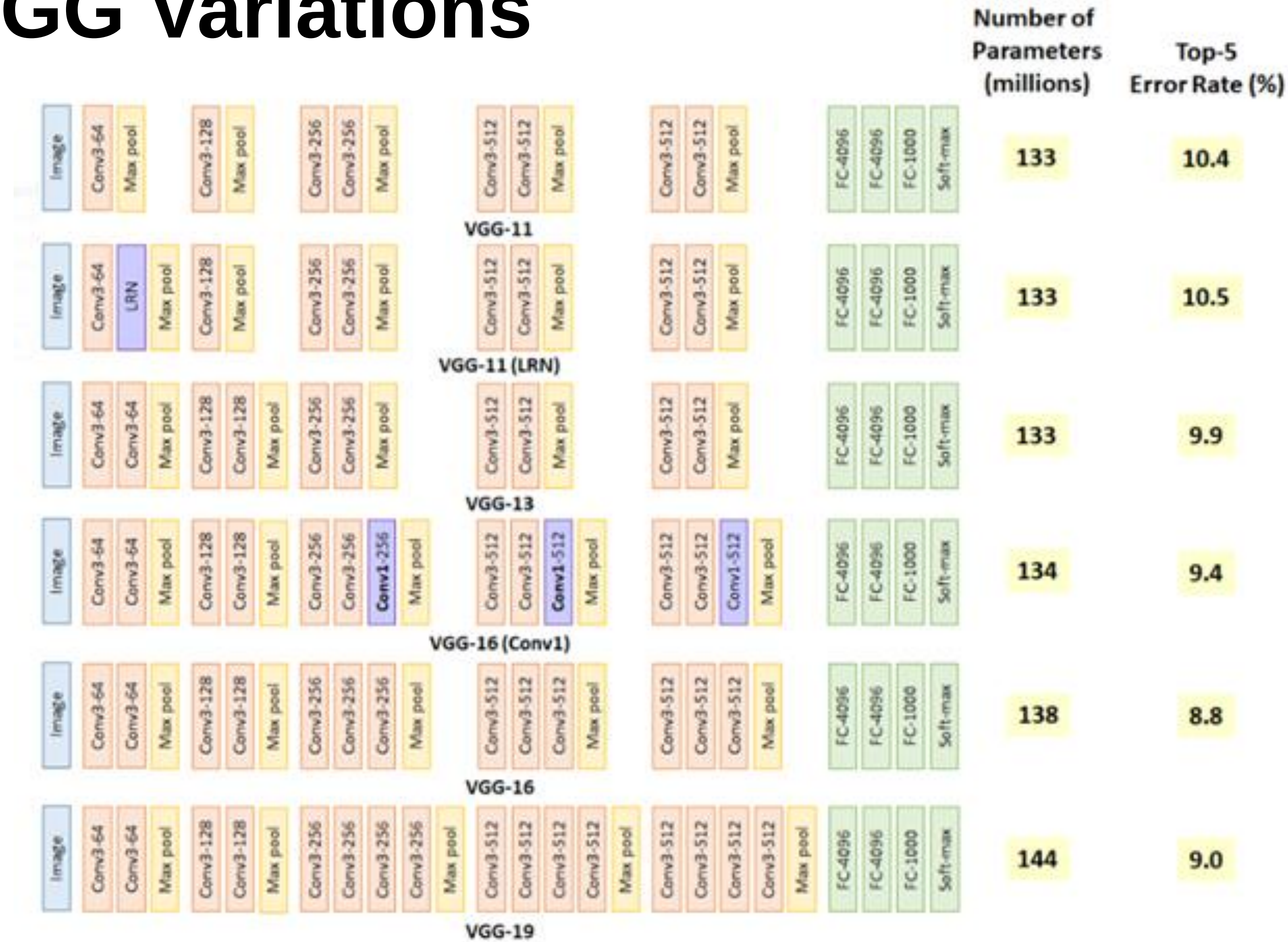
K Simonyan, A Zisserman - arXiv preprint arXiv:1409.1556, 2014 - arxiv.org

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 (3x3) convolution filters ...

☆ 99 57560회 인용 관련 학술자료 전체 29개의 버전 »

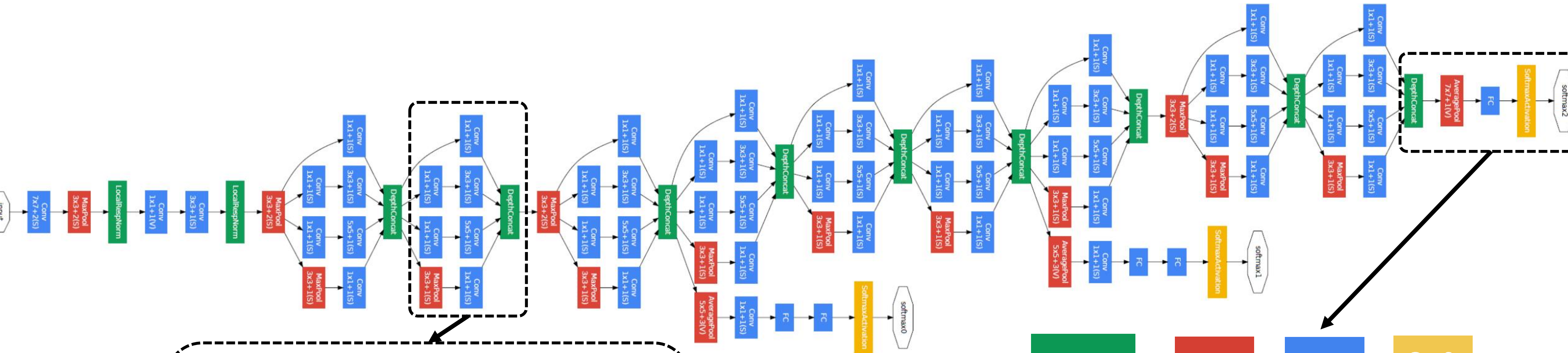
Feature Extraction **Classification**

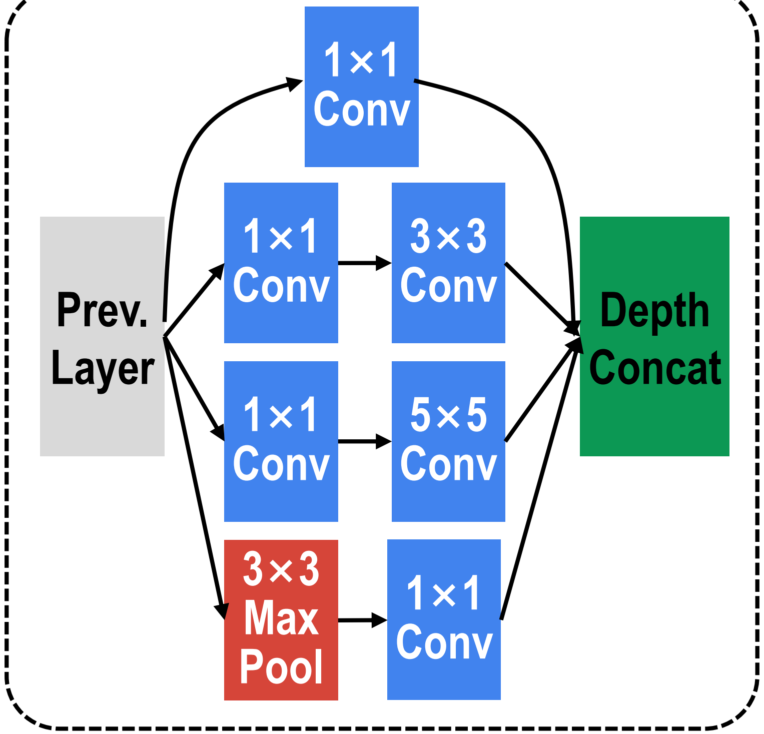
VGG Variations



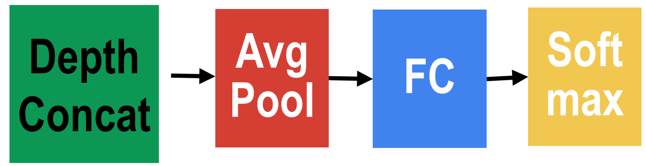
GoogLeNet

Convolution
Pooling
Softmax
Concatenation Reshape





Inception module

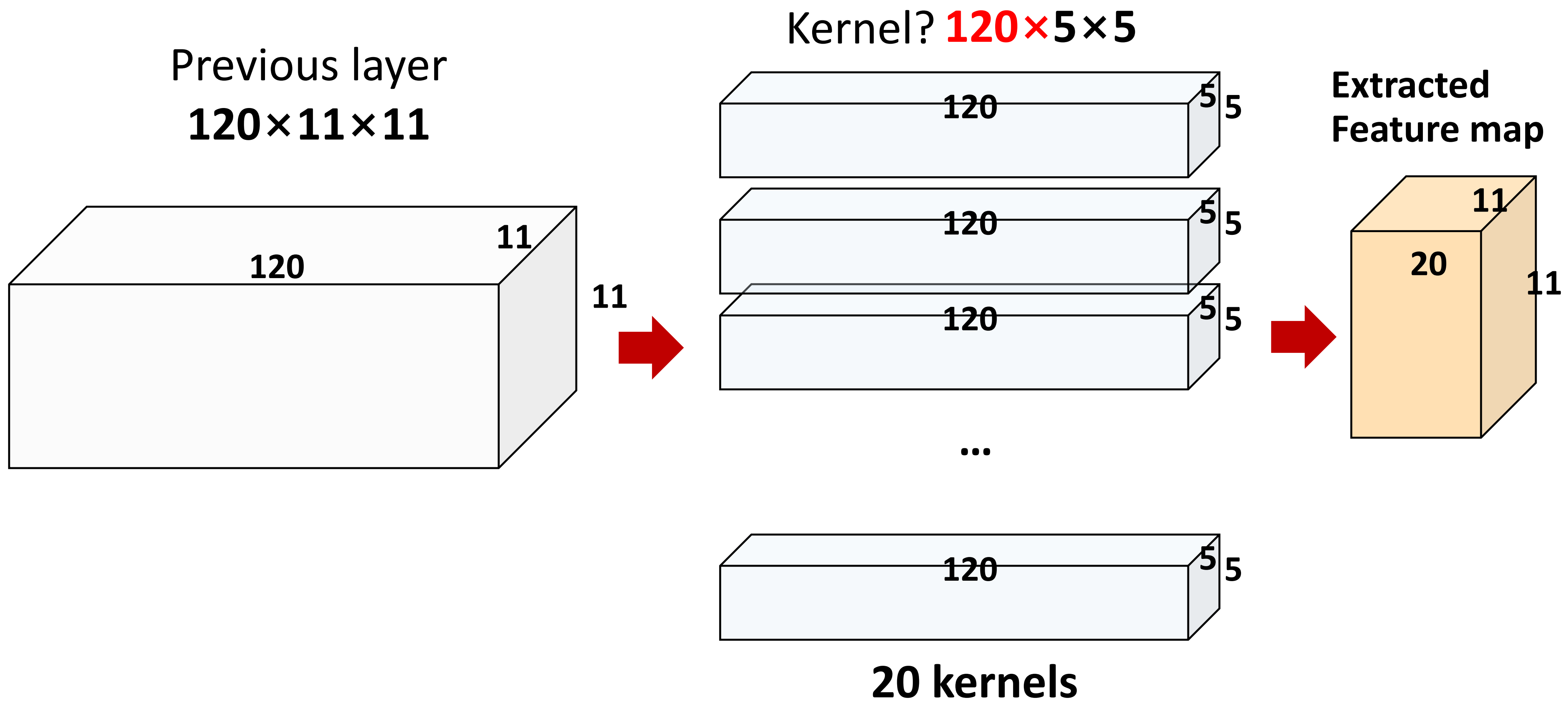


Global average pooling

Three key noticeable techniques:

1) 1×1 Convolution kernel
2) Global average pooling
3) Inception module

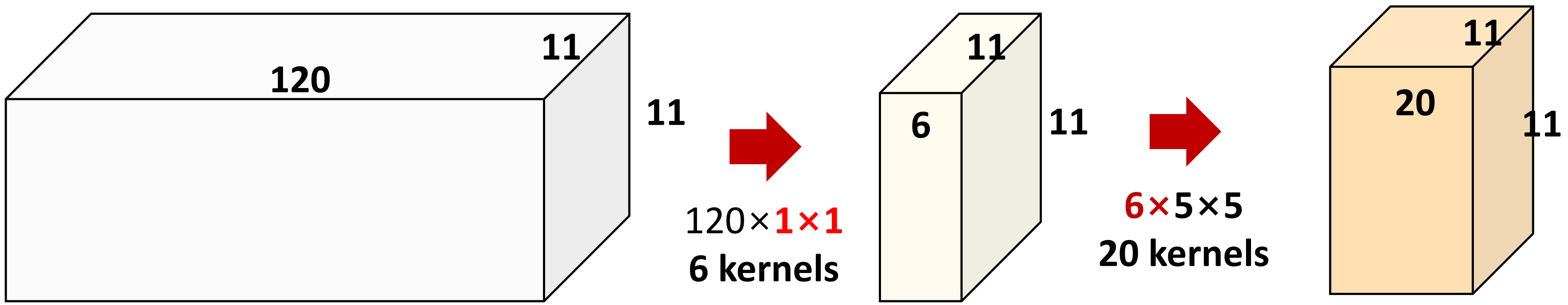
1) 1×1 Convolution kernel;
Required Number of Calculations in Conv



Number of calculations : 20×11×11×(120×5×5) = 7260k

1) 1×1 Convolution kernel; How About This?
Advantage of 1×1 Kernel Convolution

By using 1×1 kernel, we can reduce the required calculation numbers

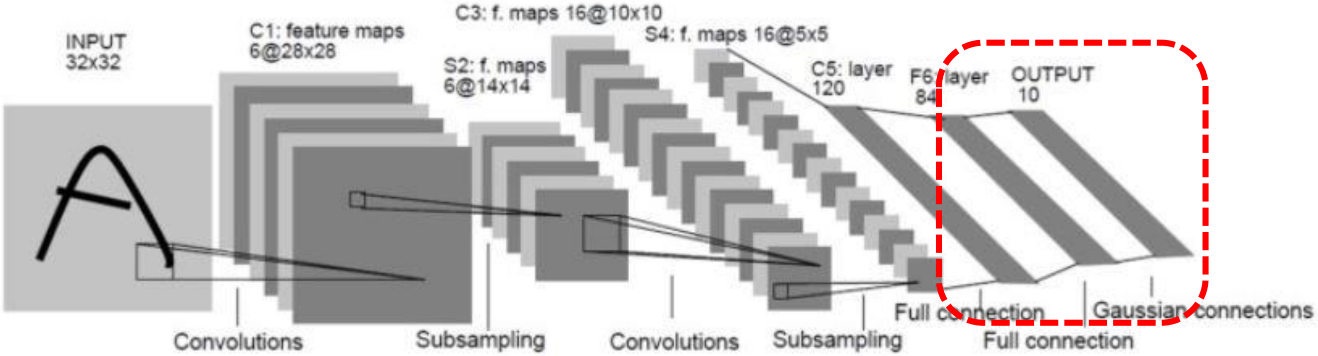


Same results,

Number of calculations :

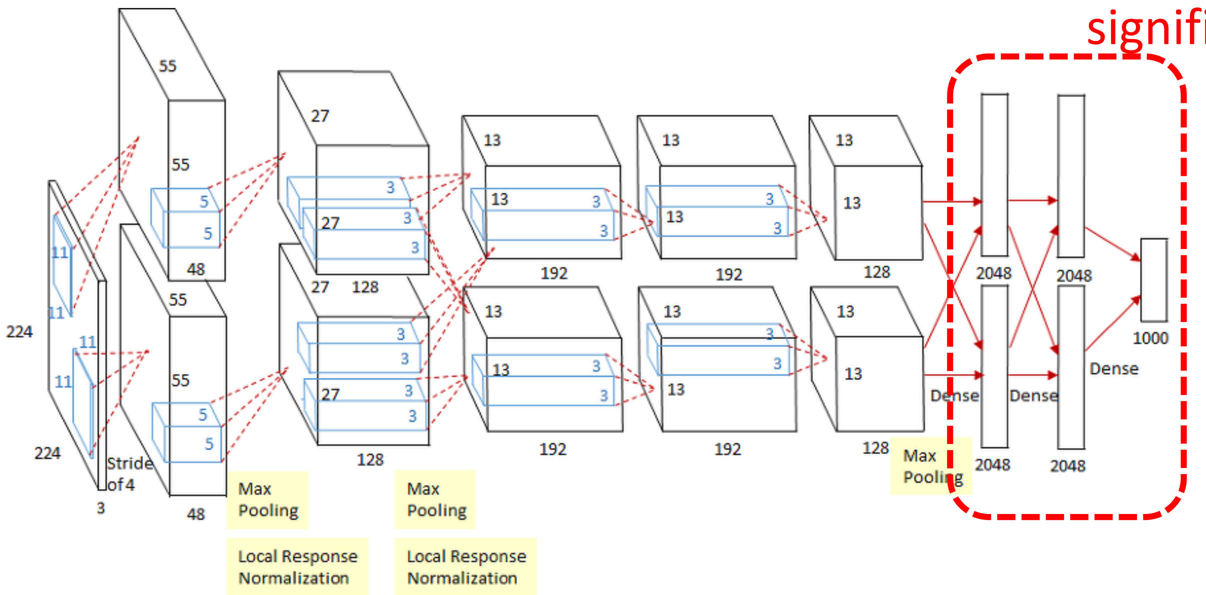
6×11×11×(120×1×1) + 20×11×11×(6×5×5) = 450k

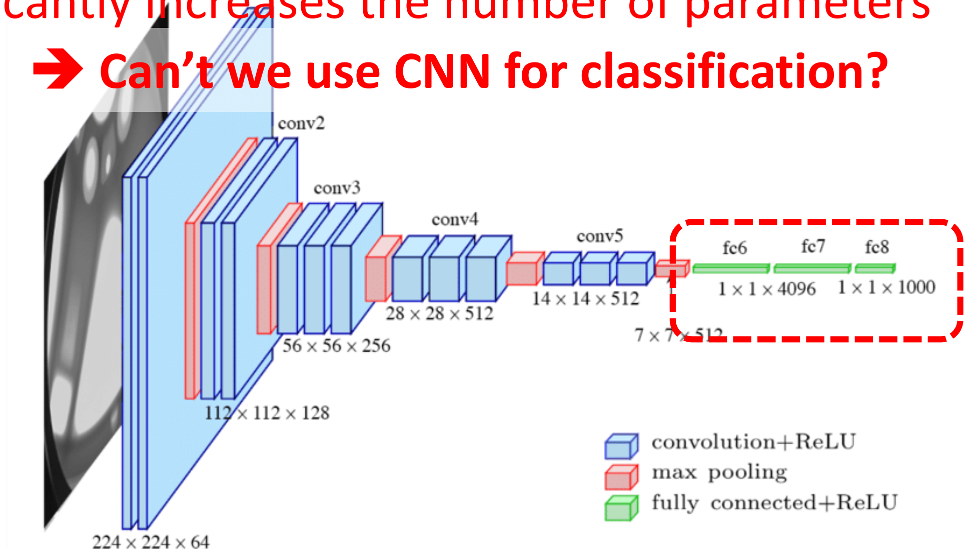
2) Global Average Pooling;
Cause of Large Parameters in Prev. CNN



Fully connected layers for classification significantly increases the number of parameters

➔ **Can't we use CNN for classification?**

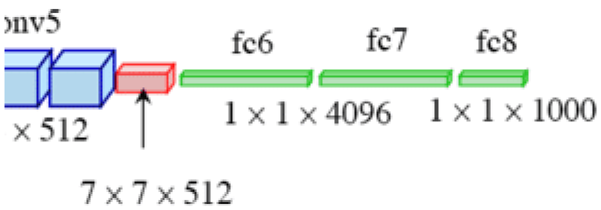




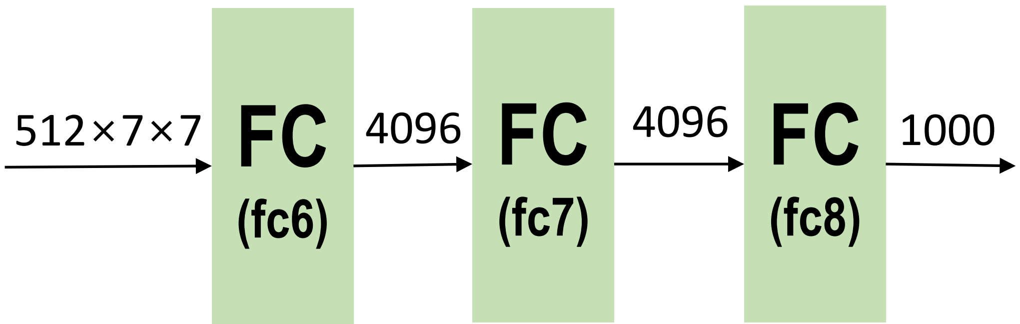
2) Global Average Pooling; Parameters of FC in VGGNet

- Classification with 1000 categories for ImageNet

VGGNet



✓ Large computation
✓ Sensitive to overfitting



Parameter Number	512×7×7×4096 **=102,760,448**	4096×4096 **=16,777,216**	4096×1000 **=4,096,000**

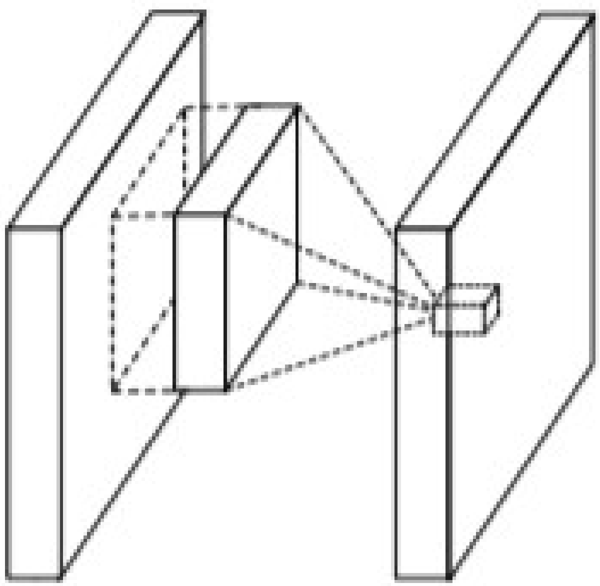
2) Global Average Pooling; MLP Convolution in NIN

Network in network

M Lin, Q Chen, S Yan - arXiv preprint arXiv:1312.4400, 2013 - arxiv.org

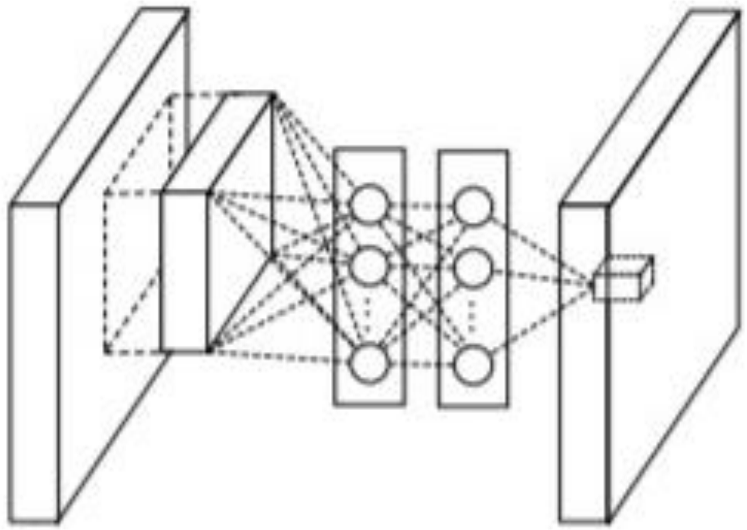
We propose a novel deep network structure called" Network In Network"(NIN) to enhance model discriminability for local patches within the receptive field. The conventional convolutional layer uses linear filters followed by a nonlinear activation function to scan the ...

☆ 99 5151회 인용 관련 학술자료 전체 11개의 버전 »



Convolution layer in CNN

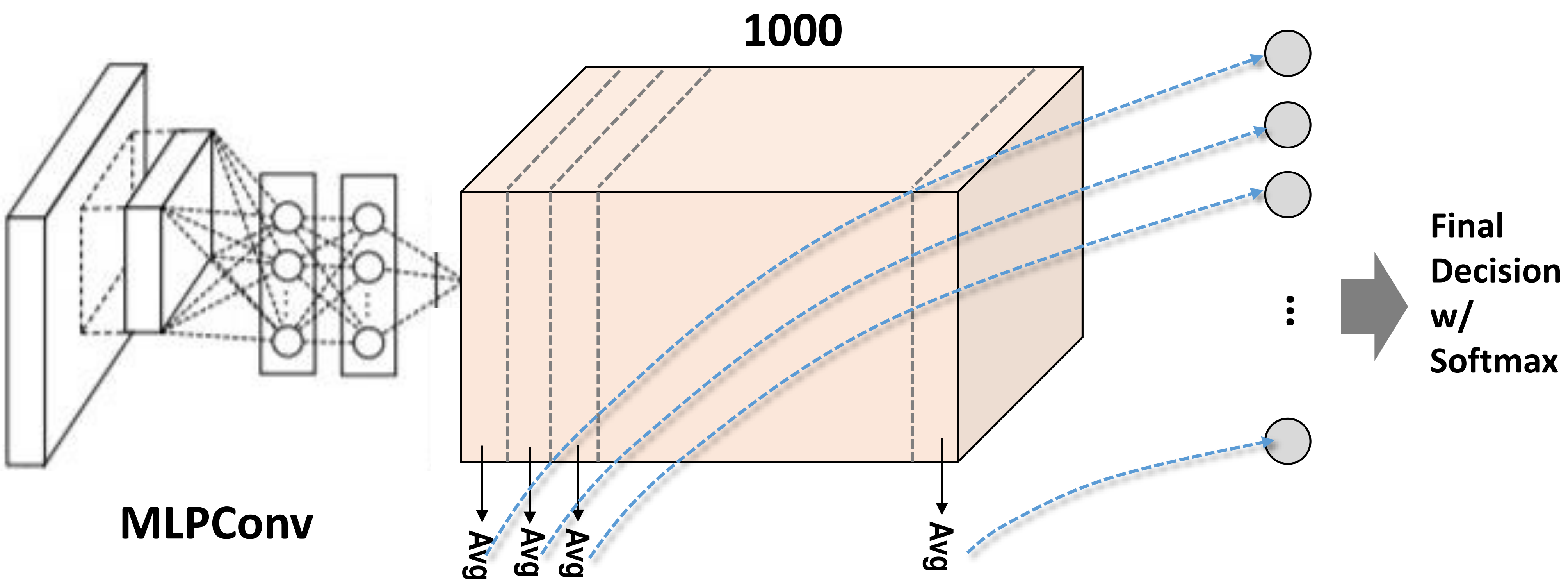
vs.



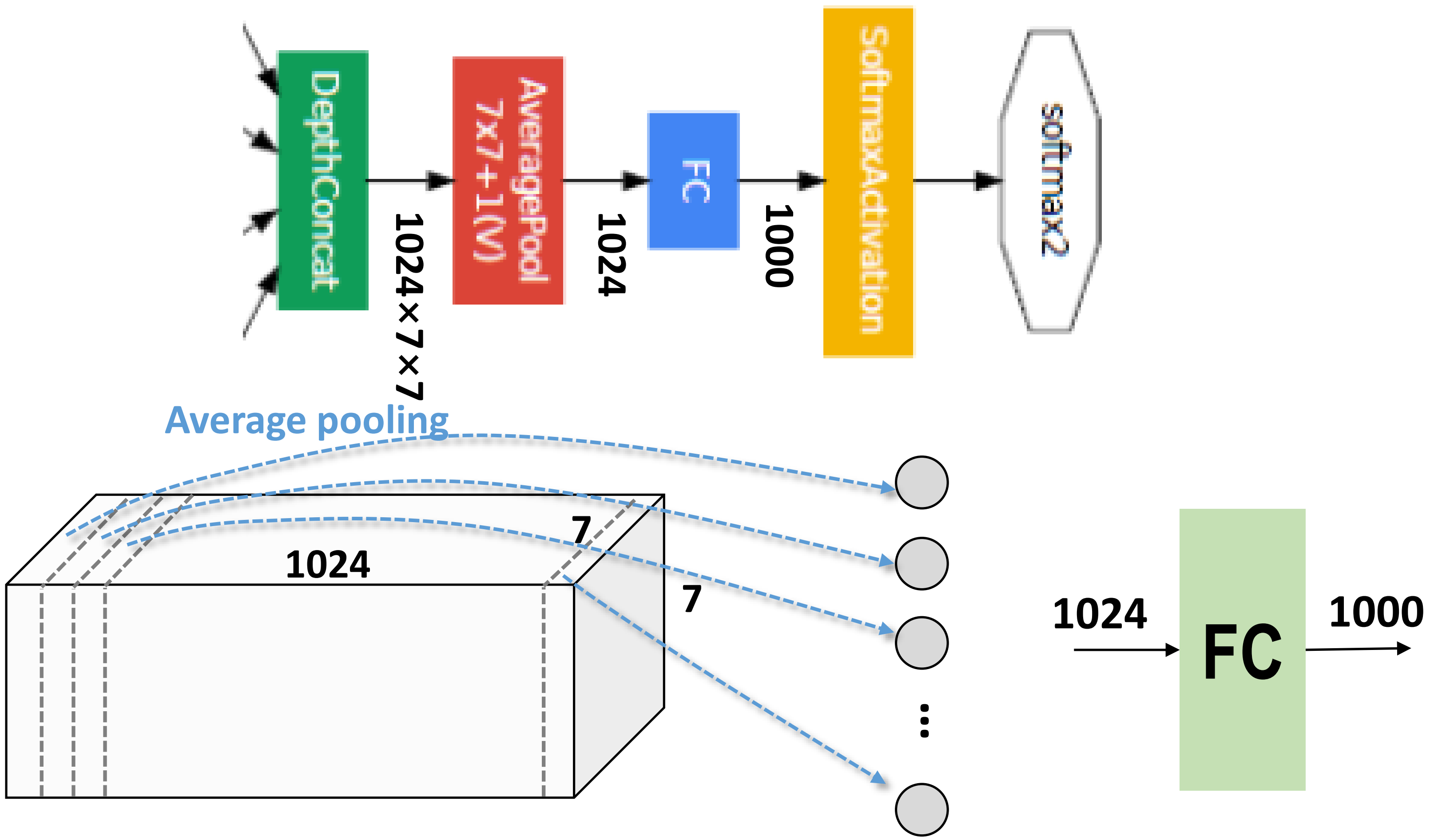
MLP-Convolution layer

2) Global Average Pooling; Global Average Pooling w/ MLPConv

- For 1000 categories,

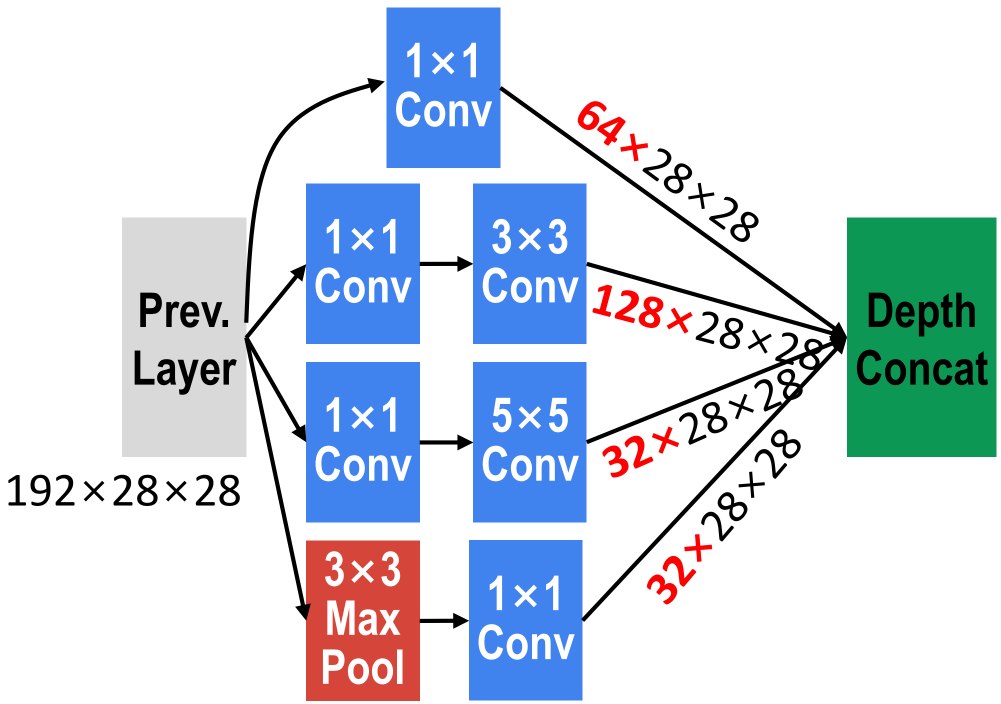


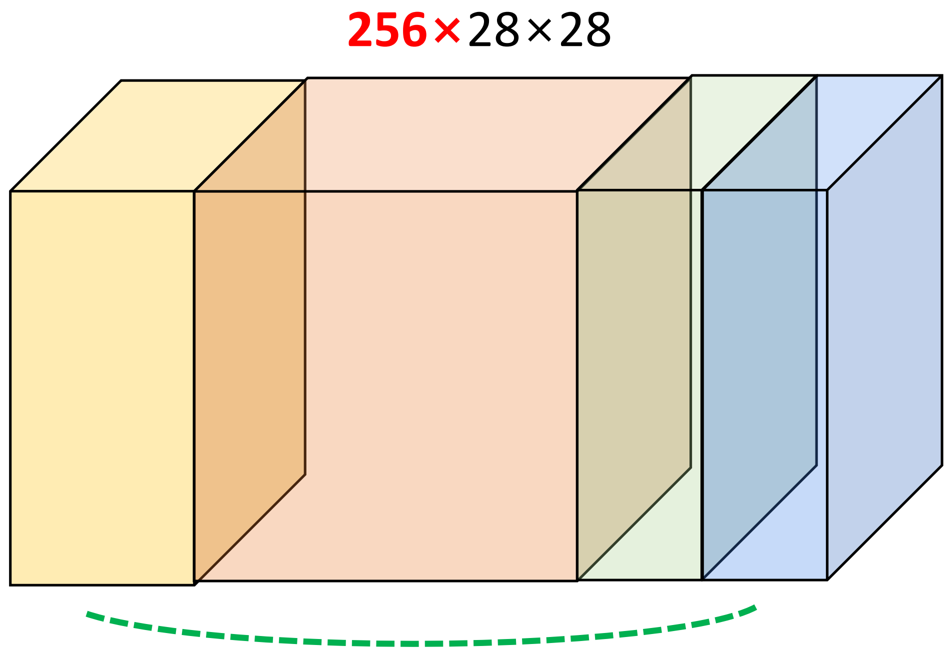
2) Global Average Pooling; Classification in GoogLeNet



Parameter Number = 1024×1000

3) Inception Module

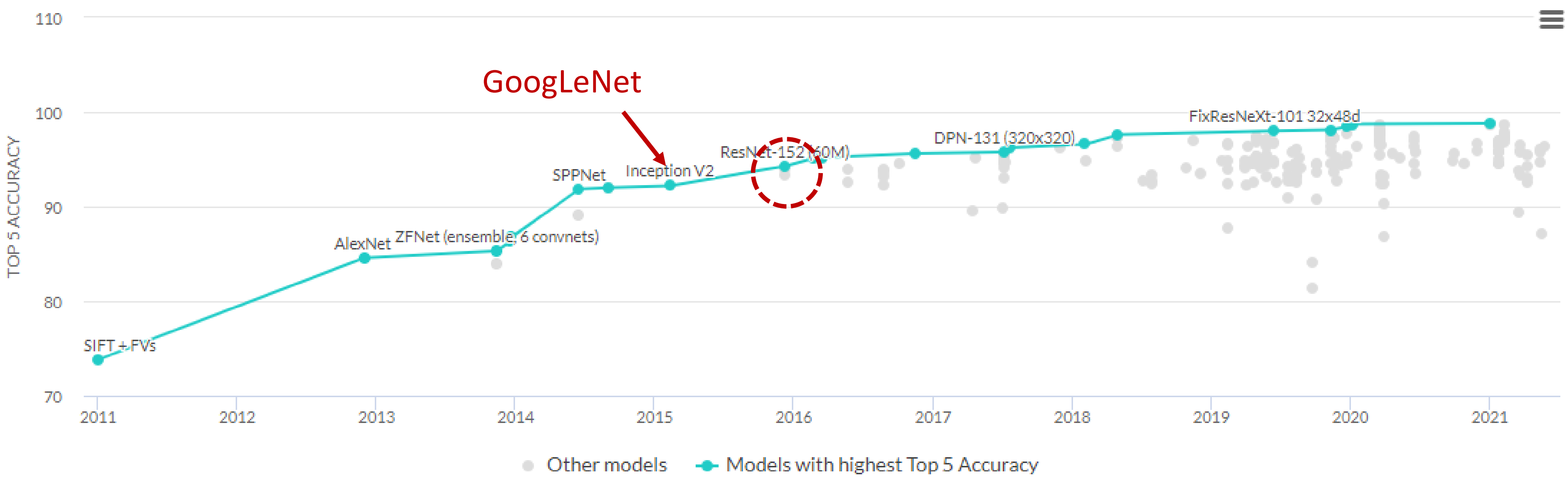




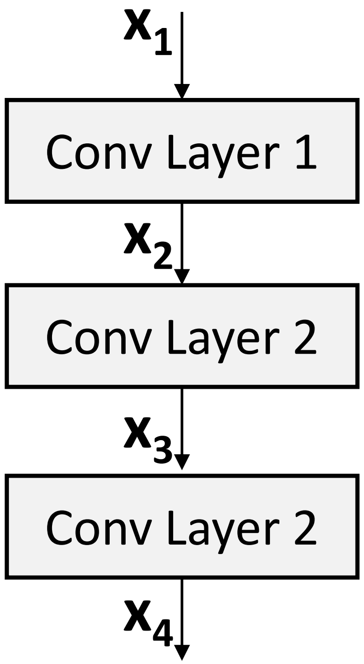
Depth Concatenation

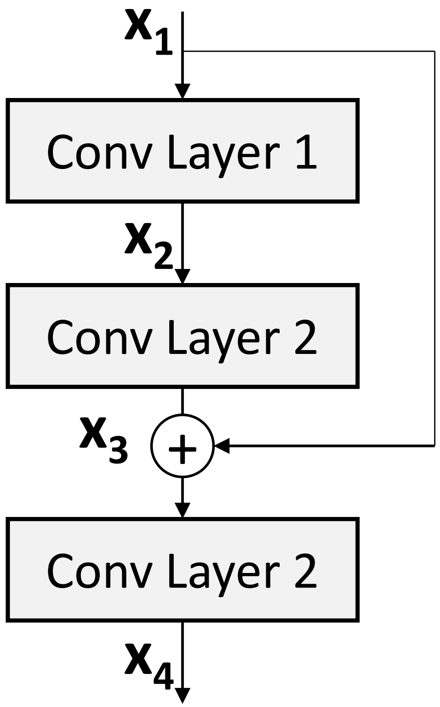
➔ Can extract various features

Revisit IRSVRC Results



Resolving Vanishing Gradient Problem





$$\nabla J(\mathbf{x}) = \frac{\partial J(\mathbf{x}_1)}{\partial \mathbf{x}_1} = \frac{\partial \mathbf{x}_4}{\partial \mathbf{x}_1} = \frac{\partial \mathbf{x}_4}{\partial \mathbf{x}_3} \times \frac{\partial \mathbf{x}_3}{\partial \mathbf{x}_2} \times \frac{\partial \mathbf{x}_2}{\partial \mathbf{x}_1}$$

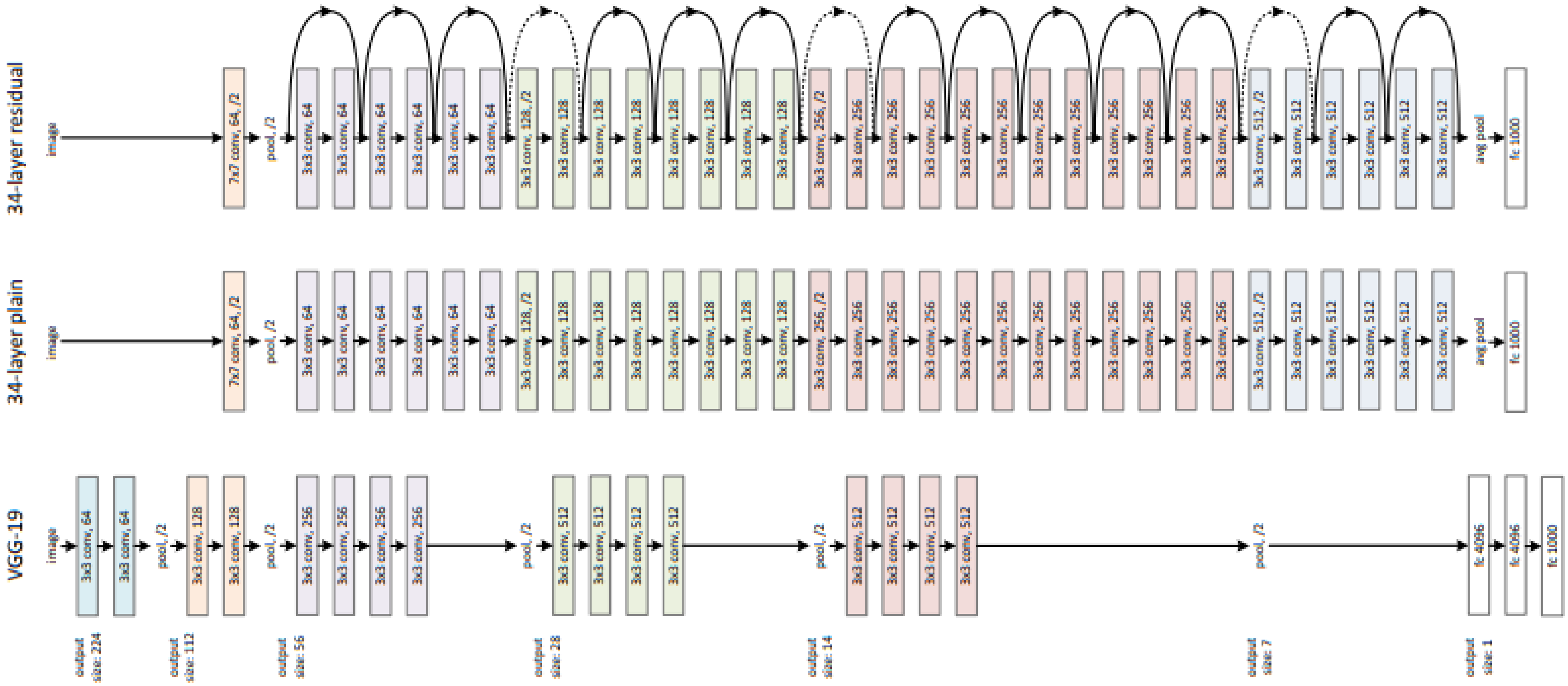
$$\frac{\partial \mathbf{x}_4}{\partial \mathbf{x}_1} = \frac{\partial \mathbf{x}_4}{\partial \mathbf{x}_3} \times \frac{\partial \mathbf{x}_3}{\partial \mathbf{x}_2} \times \frac{\partial \mathbf{x}_2}{\partial \mathbf{x}_1} + \frac{\partial \mathbf{x}_4}{\partial \mathbf{x}_1}$$

Is there any method to x_1 more directly affect to x_4?
So the gradient is not vanishing

Formation of short-cut path
Prevents the vanishing gradient problem

ResNet

Deep residual learning for image recognition
K He, X Zhang, S Ren, J Sun - Proceedings of the IEEE …, 2016 - openaccess.thecvf.com
Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer …
☆ 99 79599회 인용 관련 학술자료 전체 55개의 버전 »



Checkpoints

- Padding & Pooling
- Various CNNs
  - LeNet (
  - AlexNet
  - VGGNet
  - GoogLeNet
  - ResNet