AI and Deep Learning

4-3 Convolution Neural Network II

Zhonglei Wang

WISE, SOE, CHOW

XMU

(Version v1)

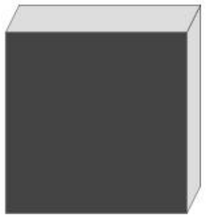
Contents

LeNet5

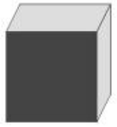
1. LeNet5 was proposed by LeCun in 1998
   - It is one of the earliest CNN
   - It is a landmark, showing success of CNN in CV
   - It is used to classify a 32×32 gray-scale handwritten digits (10 classes)
   - "5" means there exist 5 (or 7) layers
   - Tanh is used as the activation function for the hidden layers
2. LeNet5 broken away from the traditional approach that heavily relies on manually designed features

LeNet5

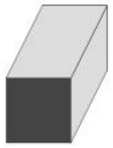
$f^{[1]} = 5$
$s^{[1]} = 1$
$p^{[1]} = 0$
$d_C^{[1]} = 6$



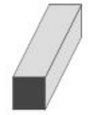
$f^{[2]} = 2$
$s^{[2]} = 2$
Average Pooling



$f^{[3]} = 5$
$s^{[3]} = 1$
$p^{[3]} = 0$
$d_C^{[3]} = 16$



$f^{[4]} = 2$
$s^{[4]} = 2$
Average Pooling



$f^{[5]} = 5$
$s^{[5]} = 1$
$p^{[5]} = 0$
$d_C^{[5]} = 120$

$\mathbb{R}^{1\times32\times32}$ $\mathbb{R}^{6\times28\times28}$ $\mathbb{R}^{6\times14\times14}$ $\mathbb{R}^{16\times10\times10}$ $\mathbb{R}^{16\times5\times5}$ $\mathbb{R}^{120\times1\times1}$

$\boldsymbol{W}^{[6]} \in \mathbb{R}^{84\times120}$
$\boldsymbol{b}^{[6]} \in \mathbb{R}^{84\times1}$
Fully Connected

$\boldsymbol{W}^{[7]} \in \mathbb{R}^{10\times84}$
$\boldsymbol{b}^{[7]} \in \mathbb{R}^{10\times1}$
Fully Connected

Softmax

$\mathcal{J}$

$\mathbb{R}^{10\times1}$ $\mathbb{R}^{10\times1}$ $\mathbb{R}^{84\times1}$ $\mathbb{R}^{120\times1}$

Remarks

1. Some features for early neural networks:
   - It used average pooling, not max pooling
   - Tanh, not ReLU, was used as the activation function
2. Key contribution:
   - Clearly demonstrates the core workflow of a CNN
   - Proved the good performance of CNN in computer vision tasks

AlexNet

1. It is also a landmark in CV
   - It won the ImageNet Large Scale Visual Recognition Challenge 2012
   - Its performance surpassed others significantly
   - It promoted the development of deep neural networks in CV greatly
   - It is used to classify a $224 \times 224 \times 3$ colorful images among 1000 classes
2. Contribution:
   - Use ReLU as activation function
   - Use dropout to increase generality
   - Use data augmentation to prevent overfitting

AlexNet

$\mathbb{R}^{3\times224\times224}$

$f^{[1]} = 11$
$s^{[1]} = 4$
$p^{[1]} = 0$
$d_C^{[1]} = 96$

$\mathbb{R}^{96\times55\times55}$

$f^{[2]} = 3$
$s^{[2]} = 2$
Max
Pooling

$\mathbb{R}^{96\times27\times27}$

$f^{[3]} = 5$
$s^{[3]} = 1$
$p^{[3]} = 2$
$d_C^{[3]} = 256$

$\mathbb{R}^{256\times27\times27}$

$f^{[4]} = 3$
$s^{[4]} = 2$
Max
Pooling

$\mathbb{R}^{256\times13\times13}$

$f^{[5]} = 3$
$s^{[5]} = 1$
$p^{[5]} = 1$
$d_C^{[5]} = 384$

$\mathbb{R}^{384\times13\times13}$

$f^{[6]} = 3$
$s^{[6]} = 1$
$p^{[6]} = 1$
$d_C^{[6]} = 384$

$\mathbb{R}^{384\times13\times13}$

$f^{[7]} = 3$
$s^{[7]} = 1$
$p^{[7]} = 1$
$d_C^{[6]} = 256$

$\mathbb{R}^{256\times13\times13}$

$f^{[8]} = 3$
$s^{[8]} = 2$
Max
Pooling

$\mathbb{R}^{256\times6\times6}$

$\mathbf{W}^{[9]} \in \mathbb{R}^{4096\times9216}$
$\mathbf{b}^{[9]} \in \mathbb{R}^{4096\times1}$
Fully
Connected

$\mathbb{R}^{9216\times1}$

$\mathbf{W}^{[10]} \in \mathbb{R}^{4096\times4096}$
$\mathbf{b}^{[10]} \in \mathbb{R}^{4096\times1}$
Fully
Connected

$\mathbb{R}^{4096\times1}$

$\mathbf{W}^{[11]} \in \mathbb{R}^{1000\times4096}$
$\mathbf{b}^{[11]} \in \mathbb{R}^{1000\times1}$
Fully
Connected

$\mathbb{R}^{4096\times1}$

Softmax

$\mathbb{R}^{1000\times1}$

$\mathcal{J}$

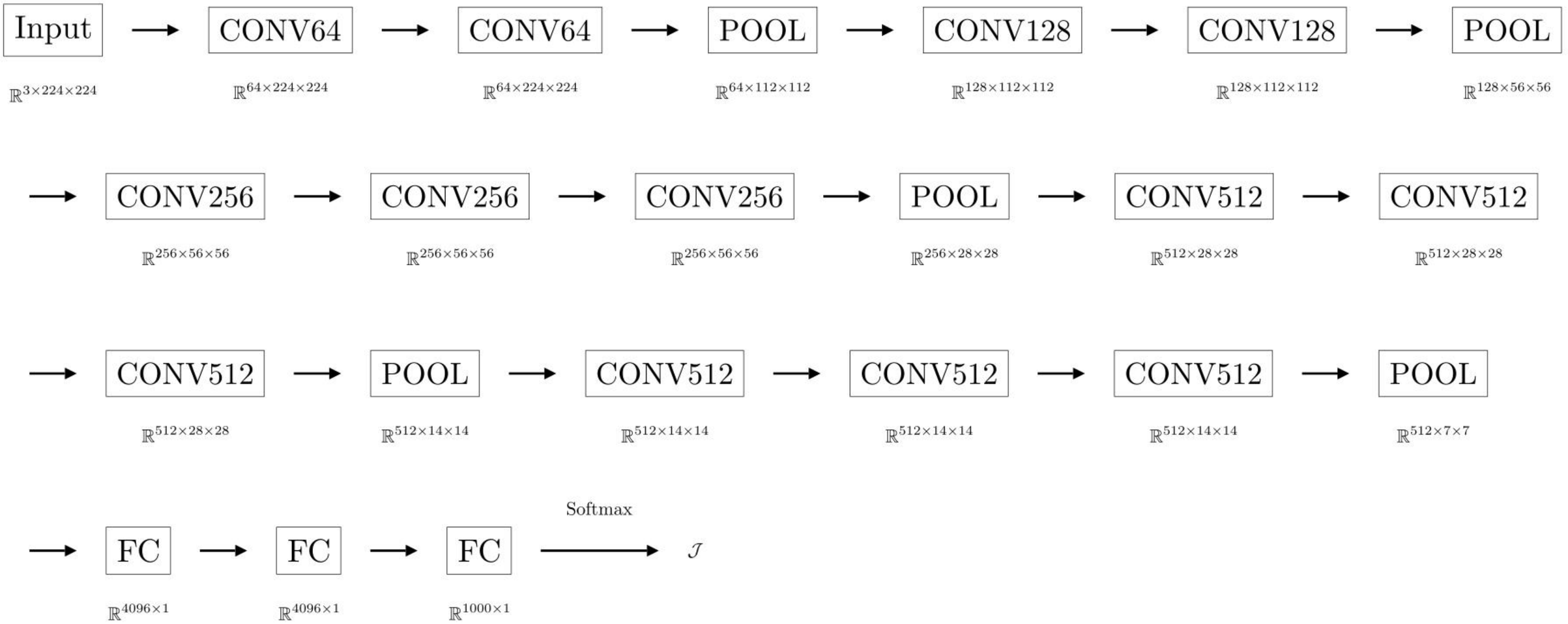
Remarks

1. Compared with LeNet5, AlexNet has the following features
   - Input is much larger: from small gray-scale images to large colorful ones
   - Network is much deeper
   - Channel is much wider
   - Pooling and intermediate layers are more suitable for complex natural images
2. LeNet5 shows the feasibility of CNN
3. AlexNet proved that CNN could become mainstream for CV

VGG16

1. It is an important CNN in CV
   - It won the ImageNet Large Scale Visual Recognition Challenge 2014
   - It is used to classify a $224 \times 224 \times 3$ colorful images among 1000 classes
2. Some notations
   - CONV[num]: Conduct "num" convolution operations with size 3×3 stride 1 and padding 1
   - POOL: Conduct Max pooling with size 2×2 stride 2
3. Remark
   - Convolution operations do not change the image size
   - Max pooling operations decrease half of the image size

VGG16


Input
CONV64
CONV64
POOL
CONV128
CONV128
POOL
$\mathbb{R}^{3\times224\times224}$
$\mathbb{R}^{64\times224\times224}$
$\mathbb{R}^{64\times224\times224}$
$\mathbb{R}^{64\times112\times112}$
$\mathbb{R}^{128\times112\times112}$
$\mathbb{R}^{128\times112\times112}$
$\mathbb{R}^{128\times56\times56}$
CONV256
CONV256
CONV256
POOL
CONV512
CONV512
$\mathbb{R}^{256\times56\times56}$
$\mathbb{R}^{256\times56\times56}$
$\mathbb{R}^{256\times56\times56}$
$\mathbb{R}^{256\times28\times28}$
$\mathbb{R}^{512\times28\times28}$
$\mathbb{R}^{512\times28\times28}$
CONV512
POOL
CONV512
CONV512
CONV512
POOL
$\mathbb{R}^{512\times28\times28}$
$\mathbb{R}^{512\times14\times14}$
$\mathbb{R}^{512\times14\times14}$
$\mathbb{R}^{512\times14\times14}$
$\mathbb{R}^{512\times14\times14}$
$\mathbb{R}^{512\times7\times7}$
Softmax
FC
FC
FC
$\mathcal{J}$
$\mathbb{R}^{4096\times1}$
$\mathbb{R}^{4096\times1}$
$\mathbb{R}^{1000\times1}$

Remarks

1. The structure of VGG16 is unified
   - CONV keeps the resolution: Kernel of size 3×3, and padding is 1
   - POOL halves image size: Kernel of size 2×2, and stride is 2
   - Separate two tasks: CONV is for feature extraction, and POOL for resolution reduction
2. LeNet5 shows the feasibility of CNN
3. AlexNet proved that CNN could become mainstream for CV
4. VGG16 established the importance of depth and unified designs

Summary

1. LeNet5:
   - It laid the foundational architecture of CNN
   - It provides the idea of hierarchical feature stacking
2. AlexNet:
   - It breaks the "bottleneck" of deep networks
   - It provides several regularization techniques, including the use of ReLU and dropout
3. VGG16:
   - A "deeper and simpler" CNN is the key path to achieving high accuracy
   - A unified design: Small convolutional kernel + max pooling