CS540 Introduction to Artificial Intelligence

Convolutional Neural Networks (II)

University of Wisconsin-Madison

Spring 2023

Announcements

- **Homeworks:**
  - HW 7 due in two weeks
- Midterms are being graded; solutions on Canvas.
- Final exam is May 12, 5:05 - 7:05 pm.
- **Class roadmap**:

Thursday, Mar 30	Deep Learning II
Tuesday, April 4	Neural Network Review
Thursday, April 6	Uninformed Search
Tuesday, April 11	Informed Search

Today's goals

Today’s goals

- Review (some of) convolutional computations.

Today's goals

- Review (some of) convolutional computations.
  - 2D convolutions, multiple input channels, pooling.

Today's goals

- Review (some of) convolutional computations.
  - 2D convolutions, multiple input channels, pooling.
- Understand how convolutions are used as layers in a (deep) neural network.

Today's goals

- Review (some of) convolutional computations.
  - 2D convolutions, multiple input channels, pooling.
- Understand how convolutions are used as layers in a (deep) neural network.
- Build intuition for output of convolutional layers.

Today's goals

- Review (some of) convolutional computations.
  - 2D convolutions, multiple input channels, pooling.
- Understand how convolutions are used as layers in a (deep) neural network.
- Build intuition for output of convolutional layers.
- Overview the evolution of deeper convolutional networks

How to classify

Cats vs. dogs?

How to classify

Cats vs. dogs?

How to classify

Cats vs. dogs?

Dual

12MP

wide-angle and
telephoto cameras

How to classify
Cats vs. dogs?

Dual

12MP

wide-angle and
telephoto cameras

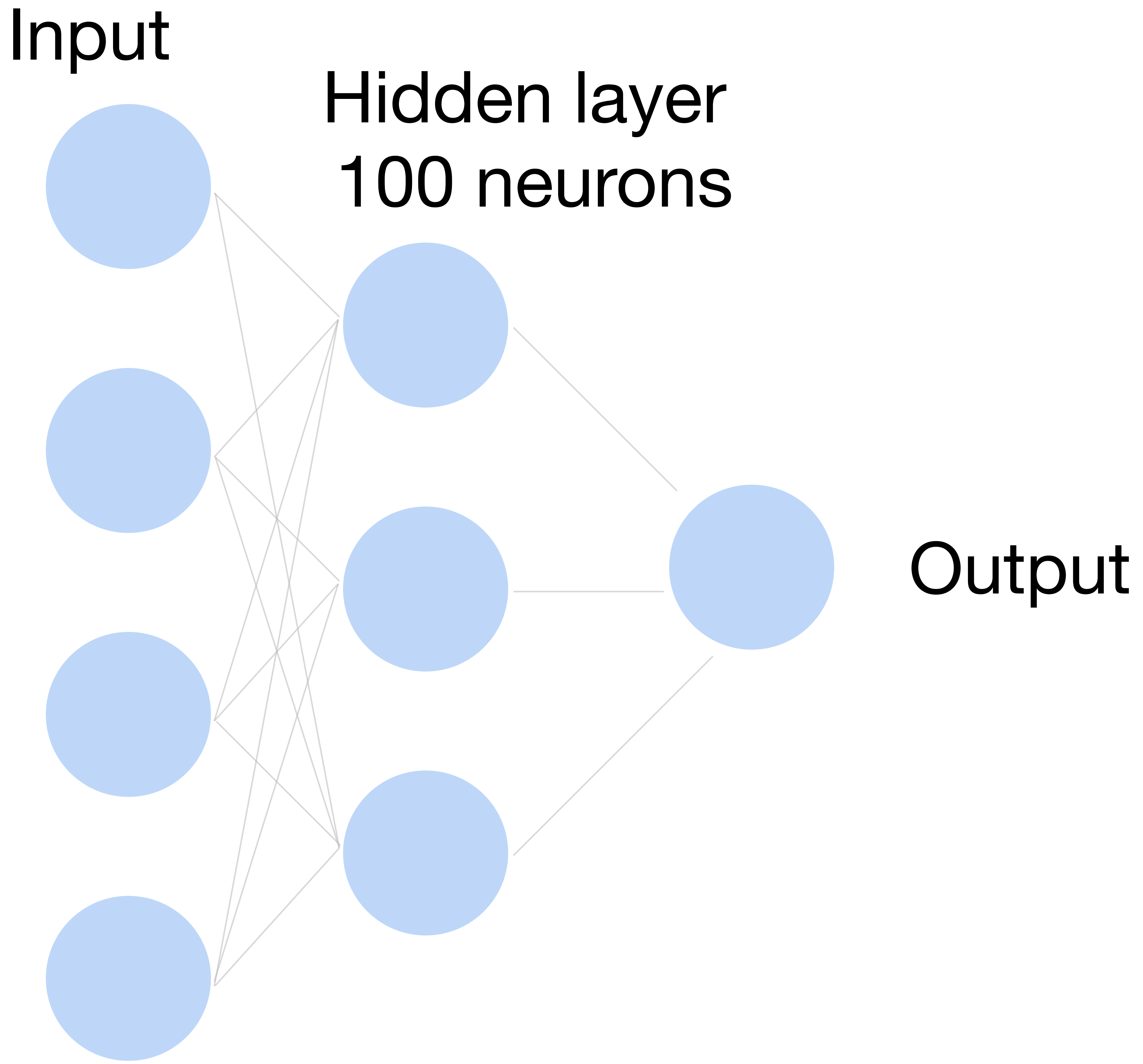
36M floats in a RGB image!

Fully Connected Networks

Cats vs. dogs?

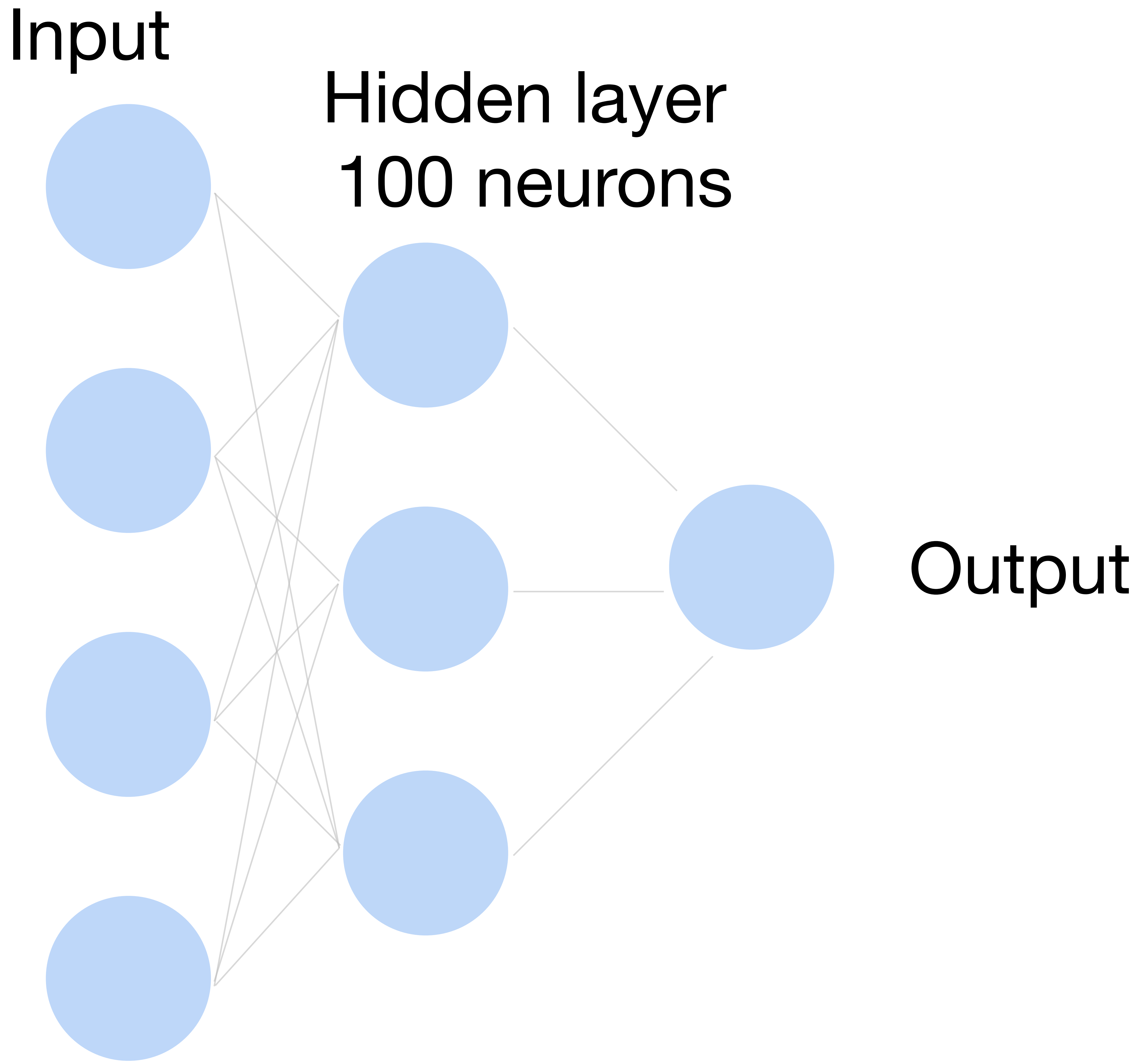
Fully Connected Networks

Cats vs. dogs?



Fully Connected Networks

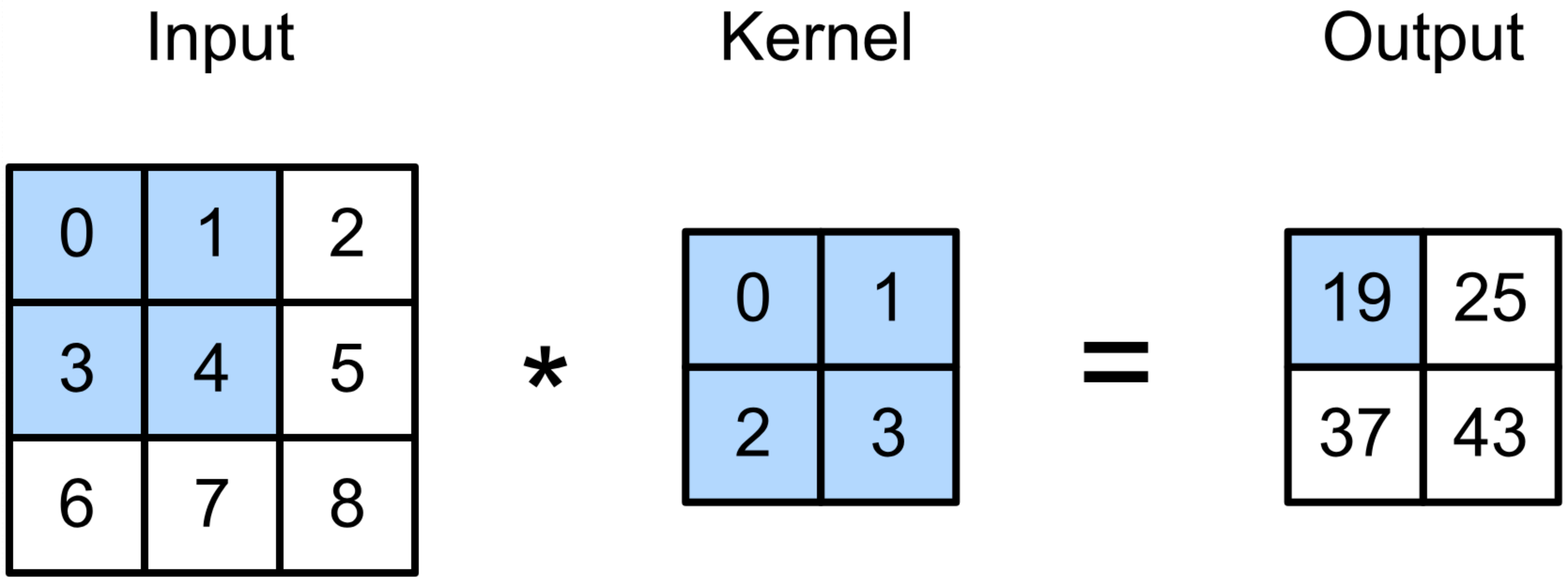
Cats vs. dogs?



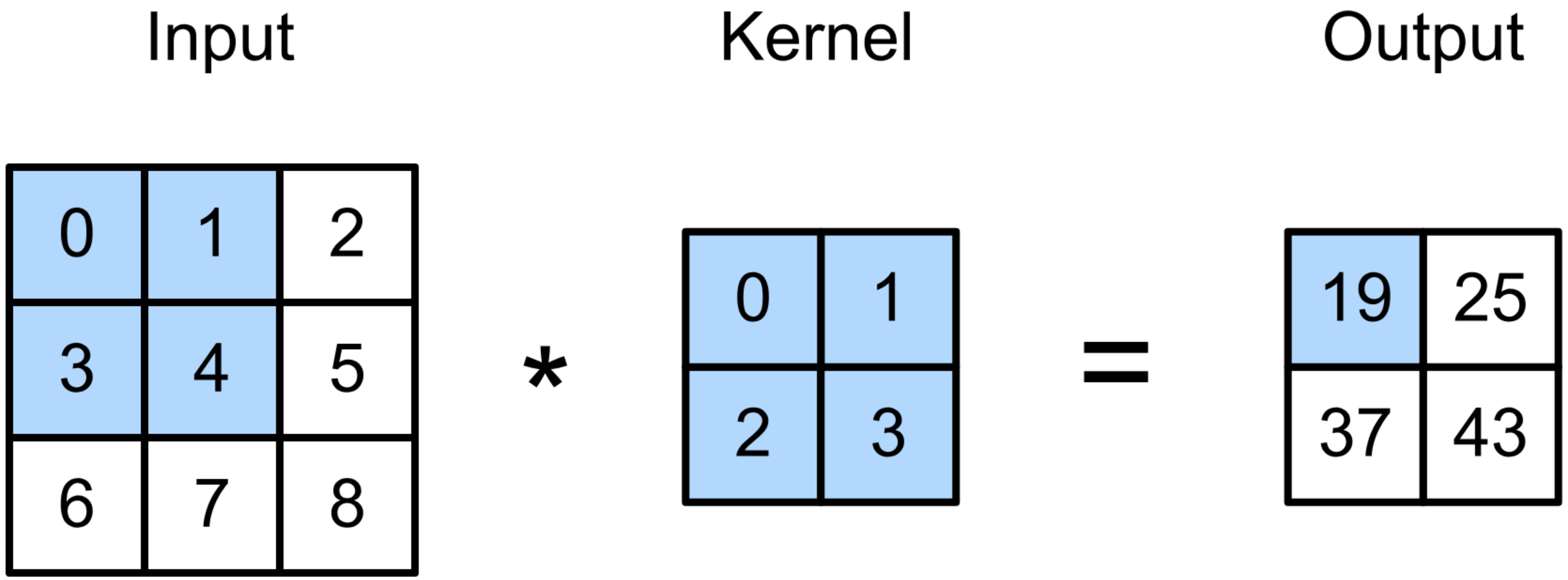
36M elements x 100 = **3.6B** parameters!

Review: 2-D Convolution

Review: 2-D Convolution



Review: 2-D Convolution

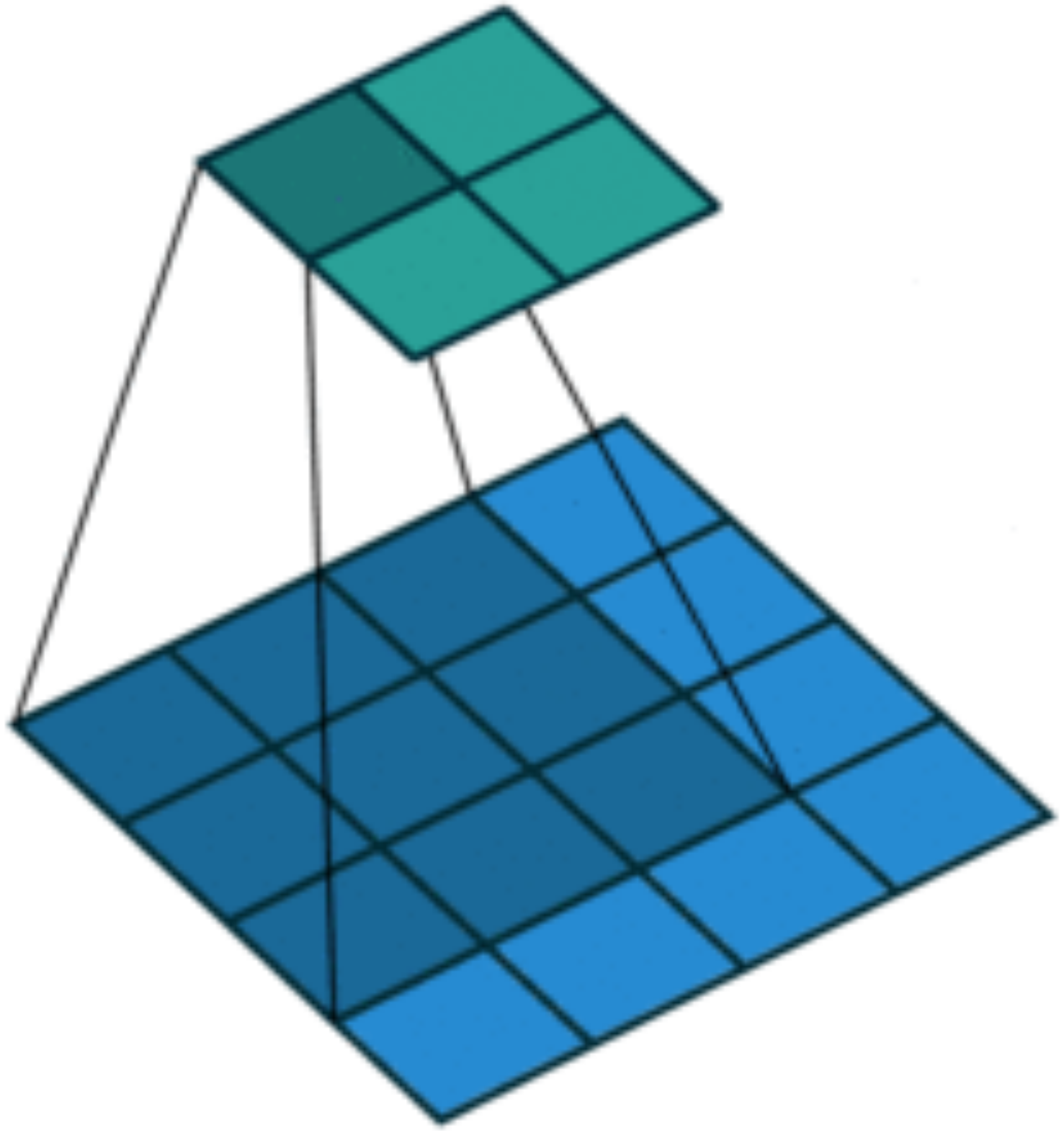


$$0 \times 0 + 1 \times 1 + 3 \times 2 + 4 \times 3 = 19,$$
$$1 \times 0 + 2 \times 1 + 4 \times 2 + 5 \times 3 = 25,$$
$$3 \times 0 + 4 \times 1 + 6 \times 2 + 7 \times 3 = 37,$$
$$4 \times 0 + 5 \times 1 + 7 \times 2 + 8 \times 3 = 43.$$

Review: 2-D Convolution

$$\underset{\text{Input}}{\begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{bmatrix}} * \underset{\text{Kernel}}{\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}} = \underset{\text{Output}}{\begin{bmatrix} 19 & 25 \\ 37 & 43 \end{bmatrix}}$$

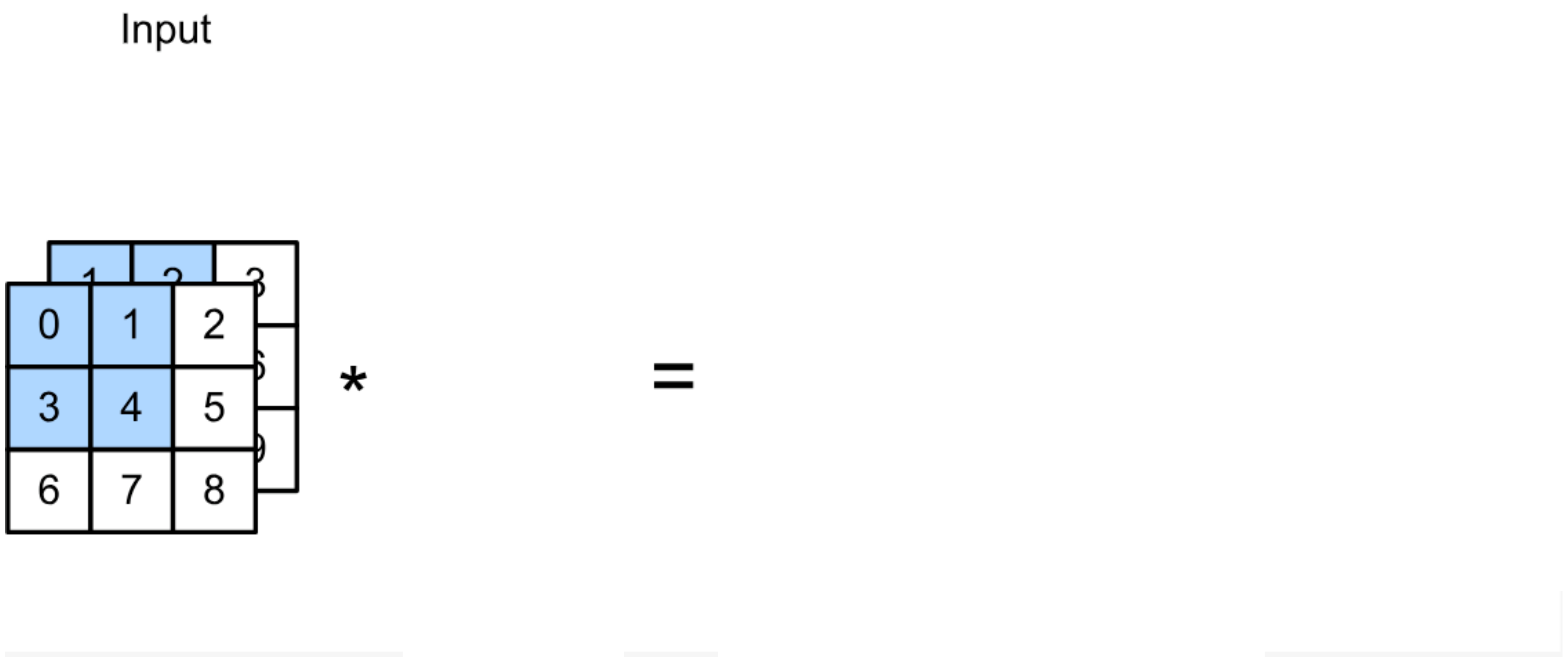
$$\begin{aligned}
0 \times 0 + 1 \times 1 + 3 \times 2 + 4 \times 3 &= 19, \\
1 \times 0 + 2 \times 1 + 4 \times 2 + 5 \times 3 &= 25, \\
3 \times 0 + 4 \times 1 + 6 \times 2 + 7 \times 3 &= 37, \\
4 \times 0 + 5 \times 1 + 7 \times 2 + 8 \times 3 &= 43.
\end{aligned}$$



(vdumoulin@ Github)

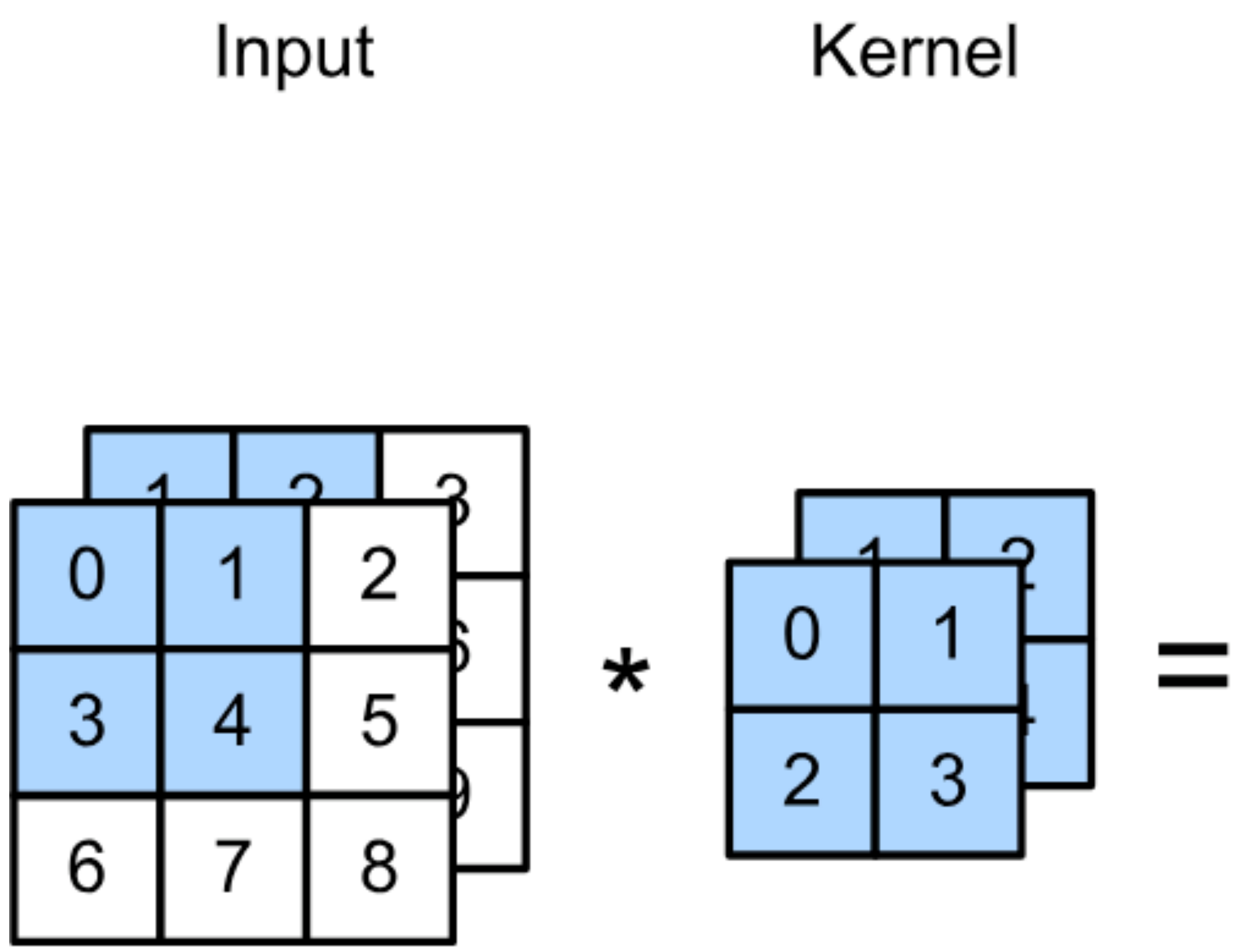
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



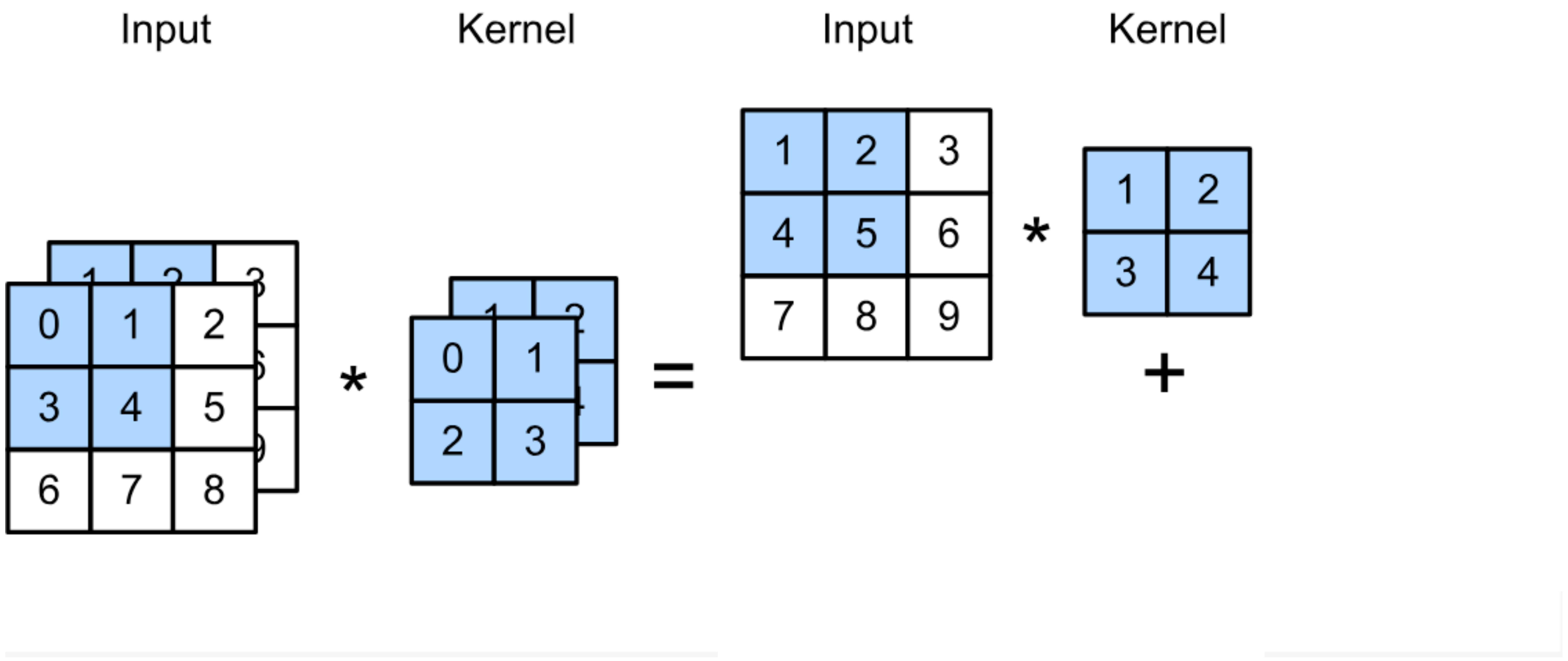
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



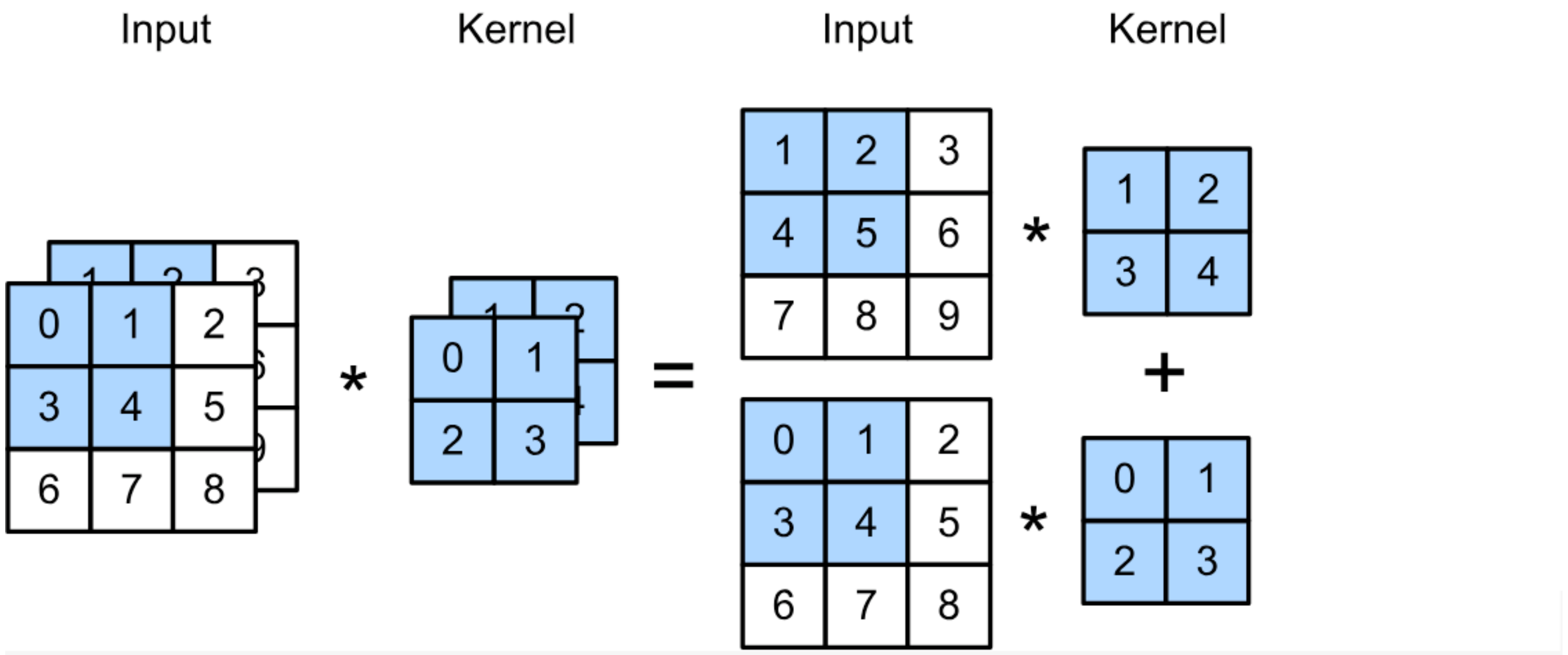
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



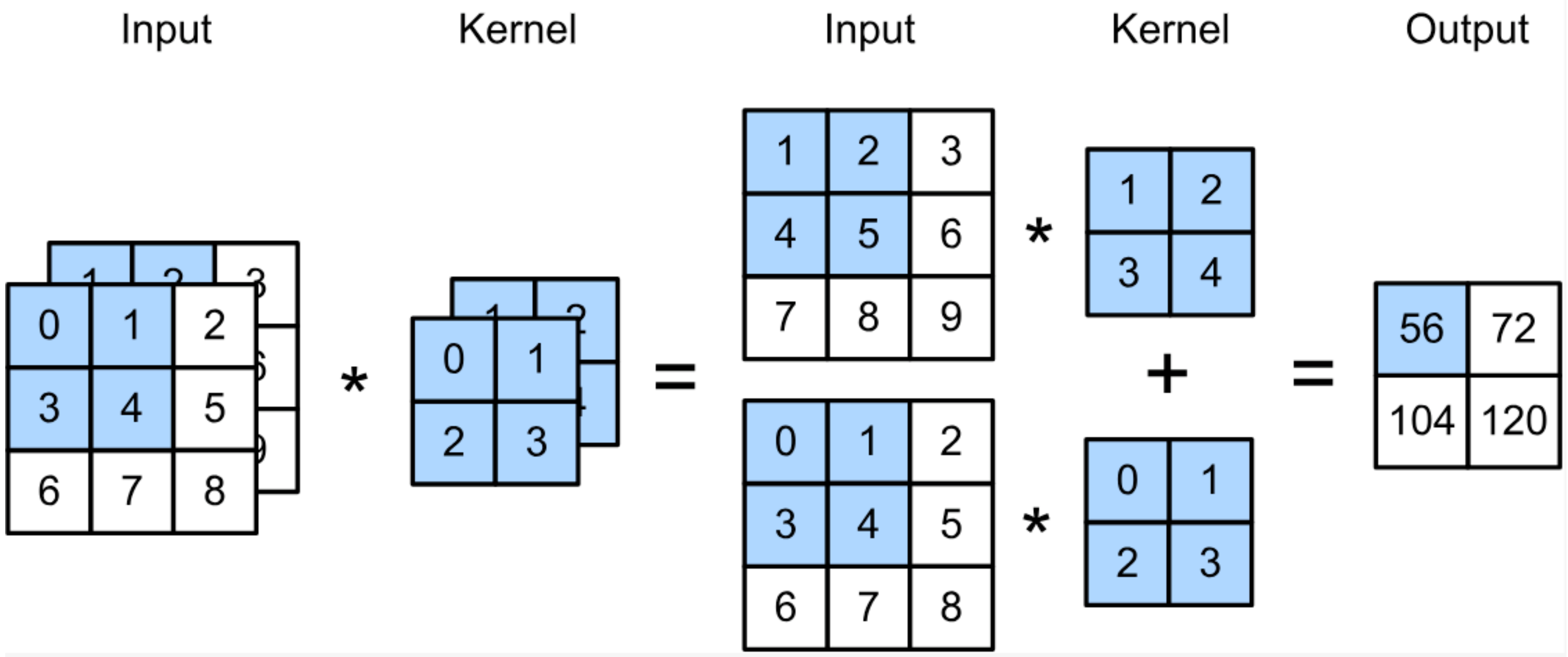
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



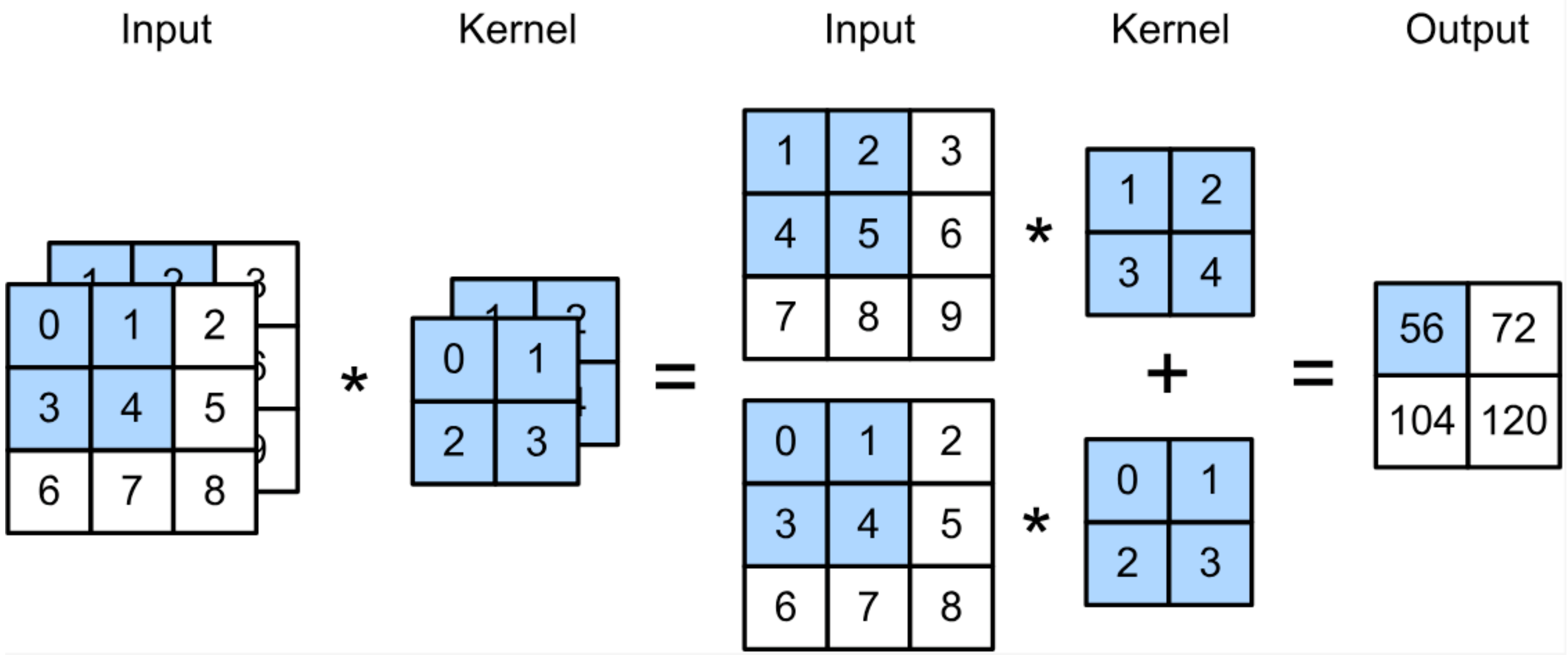
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



Review: Multiple Input Channels

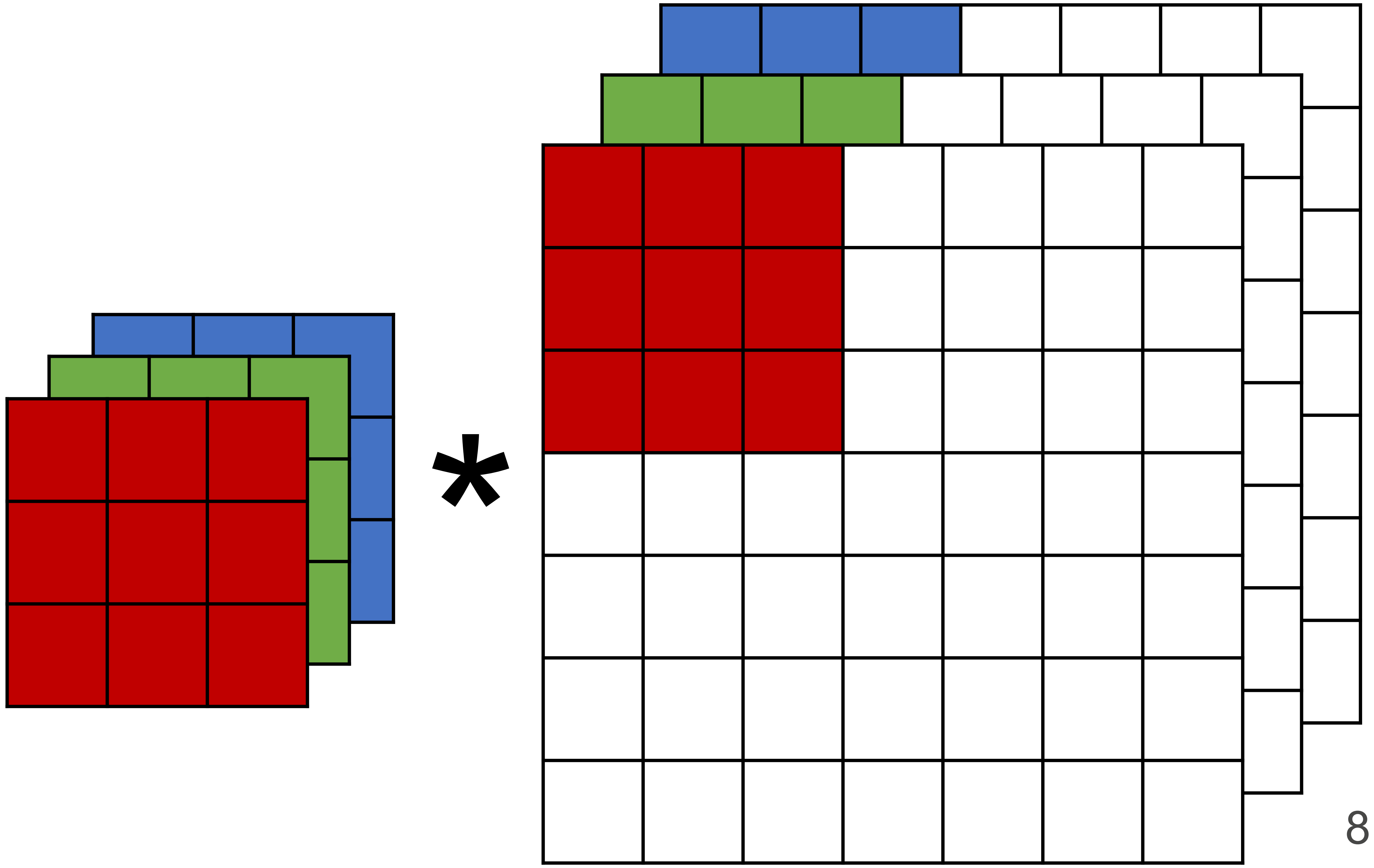
- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



$$(1\times 1+2\times 2+4\times 3+5\times 4)$$
$$+(0\times 0+1\times 1+3\times 2+4\times 3)$$
$$=56$$

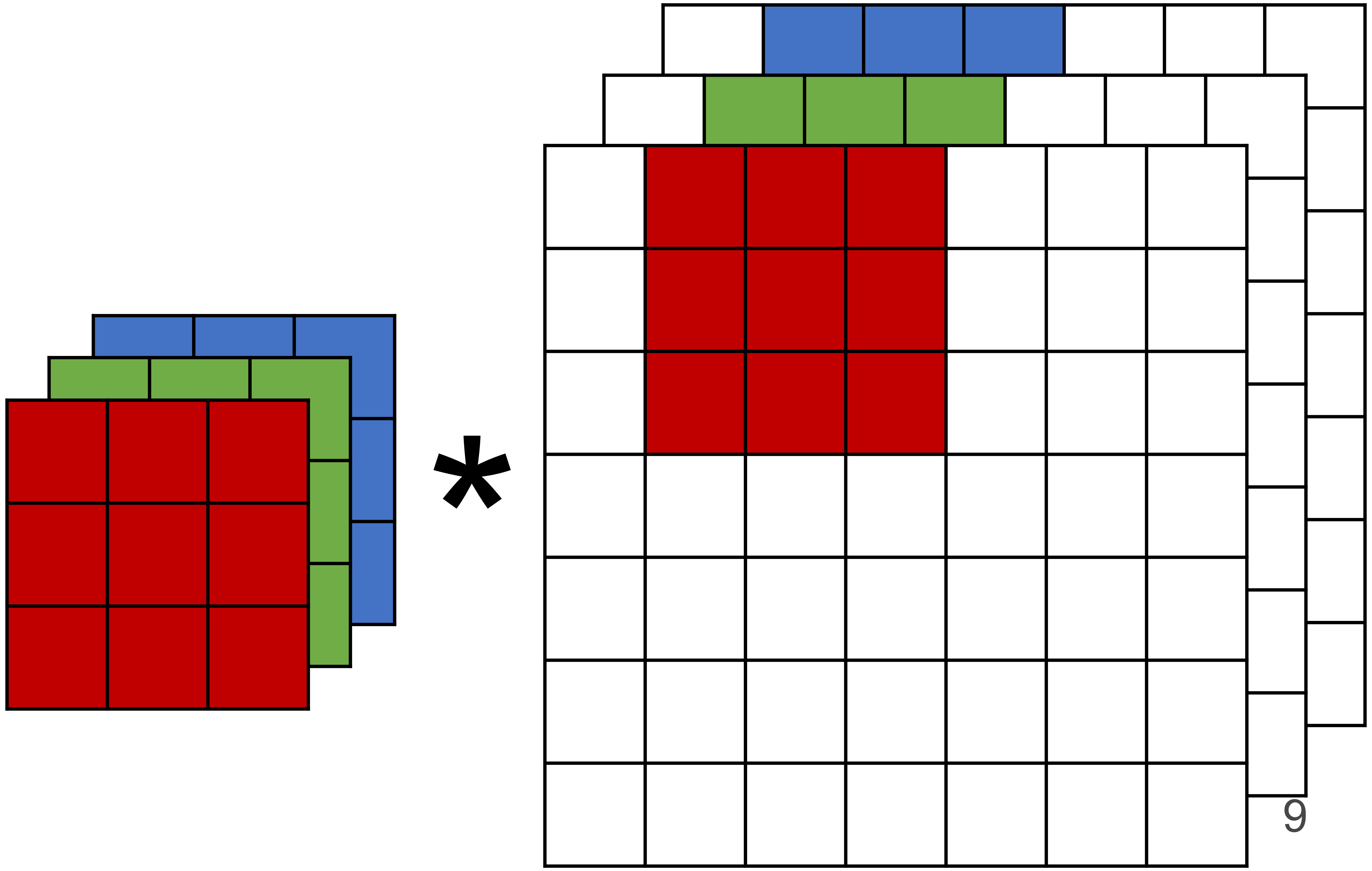
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



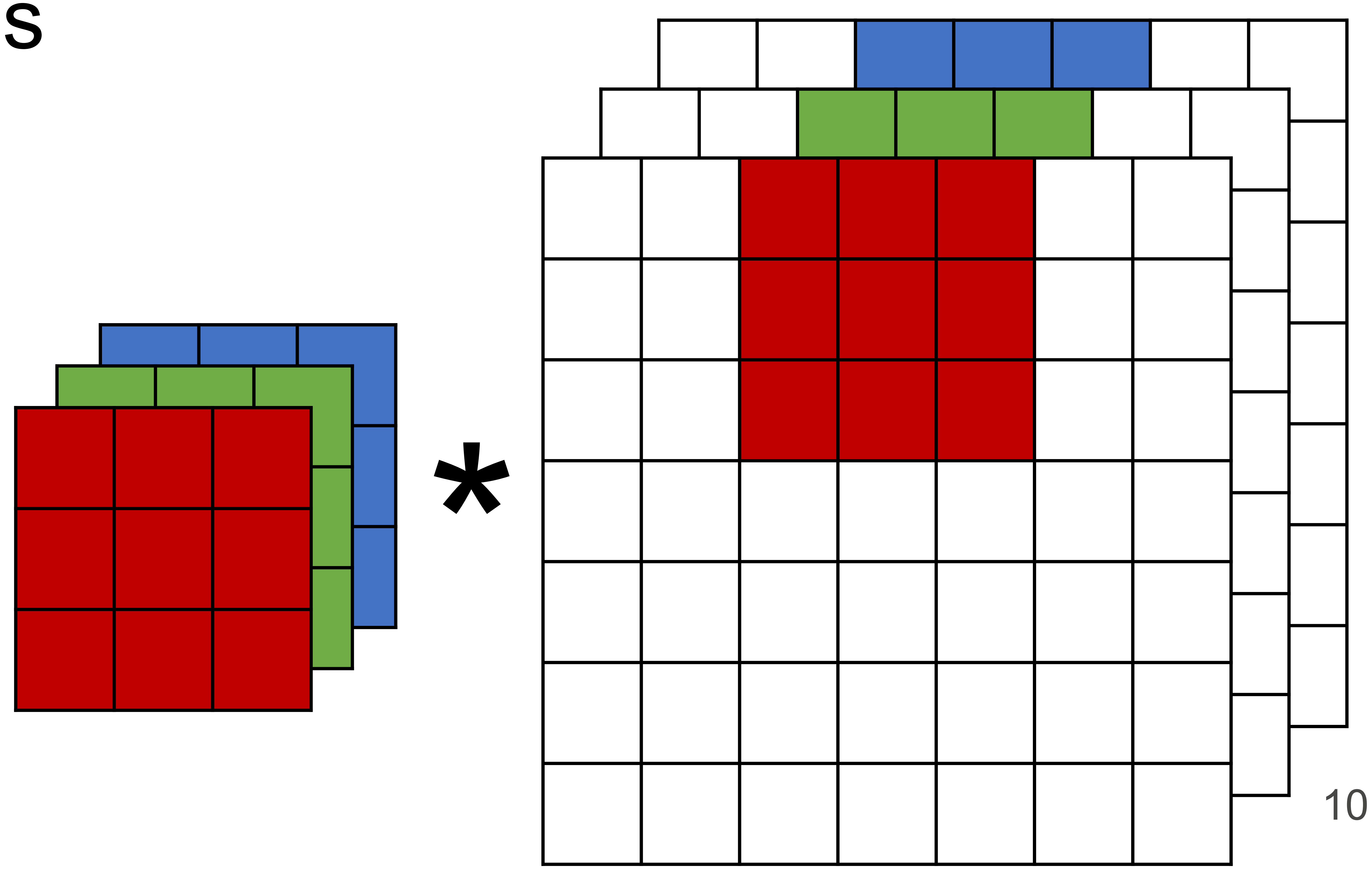
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



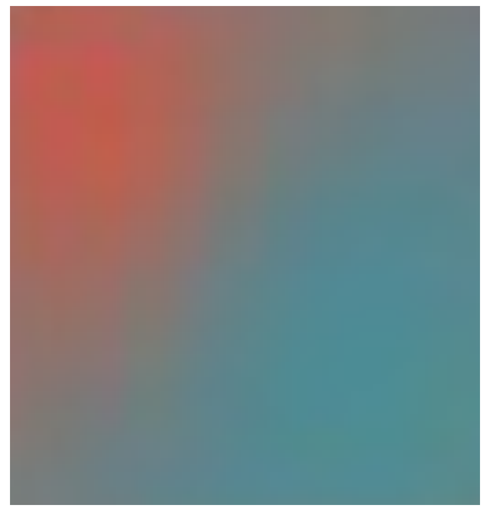
Review: Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image have 3 channels
- Have a kernel for each channel, and then sum results over channels



Multiple Input Channels

- Input and kernel can be 3D, e.g., an RGB image has 3 channels
- Also call each 3D kernel a "**filter**", which produce only **one** output channel (due to summation over channels)



*

One filter
(3 channels)

RGB (3 input channels)

Multiple filters (in one layer)

- Apply multiple filters on the input
- Each filter may learn different features about the input
- Each filter (3D kernel) produces one output channel

*

A different filter

RGB (3 input channels)

Output shape

Kernel/filter size

$$\lfloor (n_h - k_h + p_h + s_h)/s_h \rfloor \times \lfloor (n_w - k_w + p_w + s_w)/s_w \rfloor$$

Input size

Pad

Stride

Consider a convolution layer with 16 filters. Each filter has a size of 11x11x3, a stride of 2x2. Given an input image of size 22x22x3, if we don't allow a filter to fall outside of the input, what is the output size?

- 11x11x16
- 6x6x16
- 7x7x16
- 5x5x16

Consider a convolution layer with 16 filters. Each filter has a size of 11x11x3, a stride of 2x2. Given an input image of size 22x22x3, if we don't allow a filter to fall outside of the input, what is the output size?

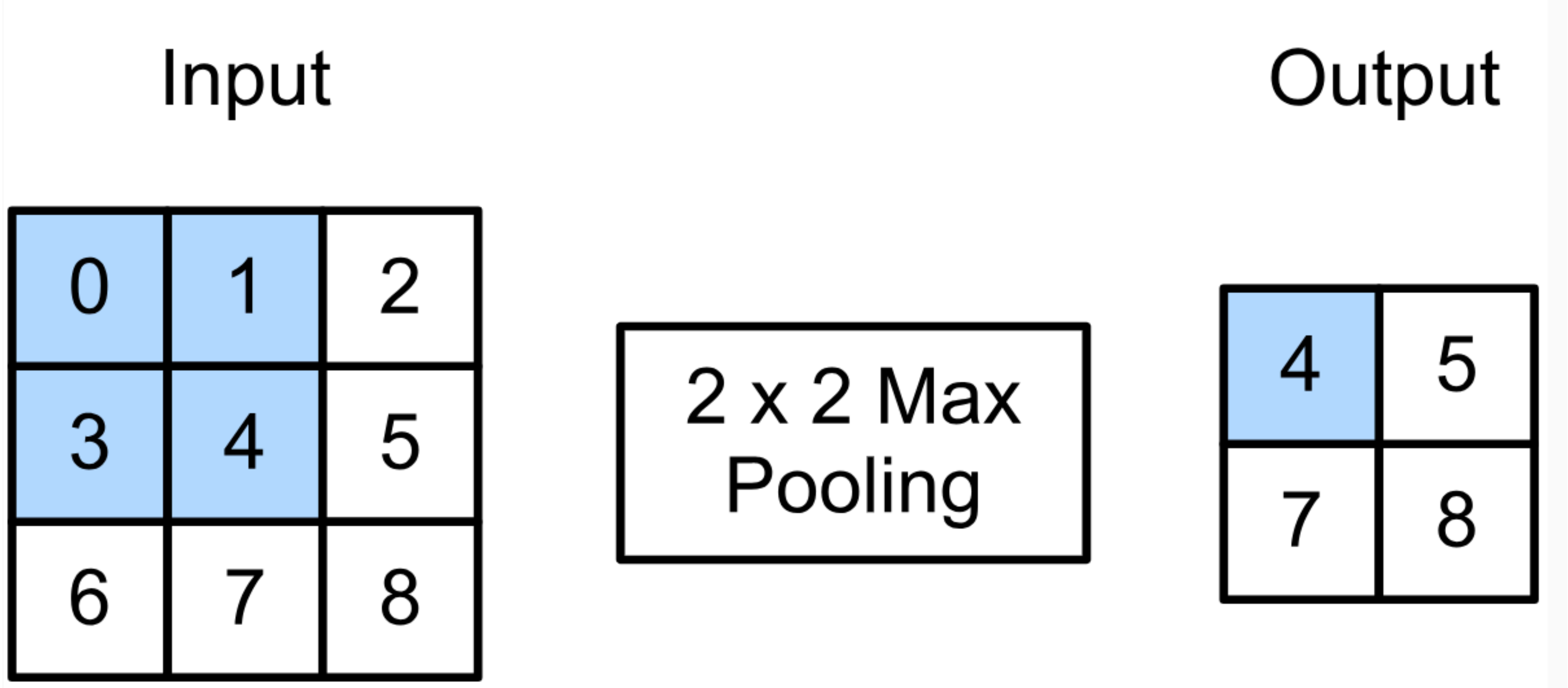
- 11x11x16
- 6x6x16
- 7x7x16
- 5x5x16

$$\lfloor (n_h - k_h + p_h + s_h)/s_h \rfloor \times \lfloor (n_w - k_w + p_w + s_w)/s_w \rfloor$$

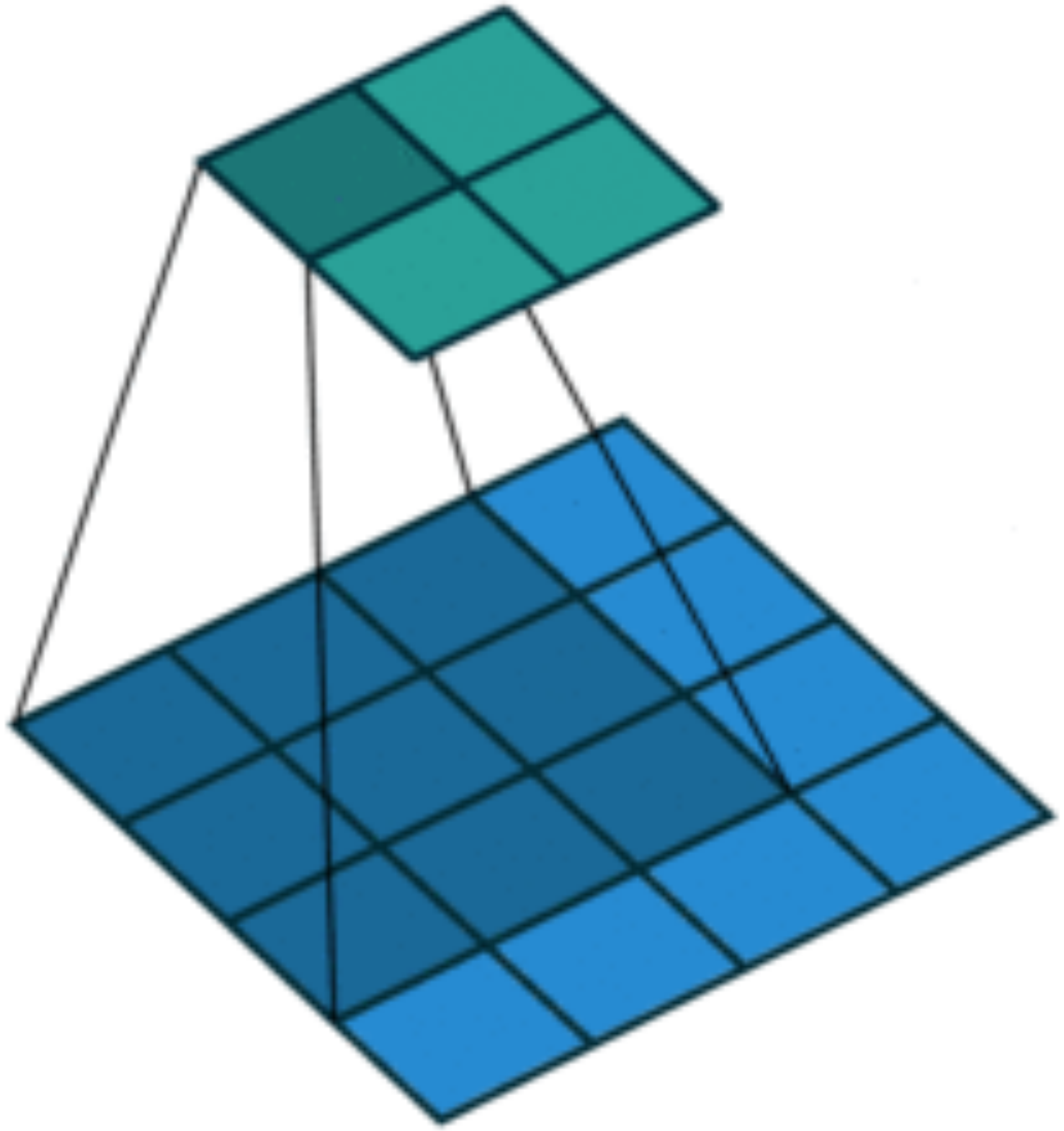
Pooling Layer

2-D Max Pooling

- Returns the maximal value in the sliding window

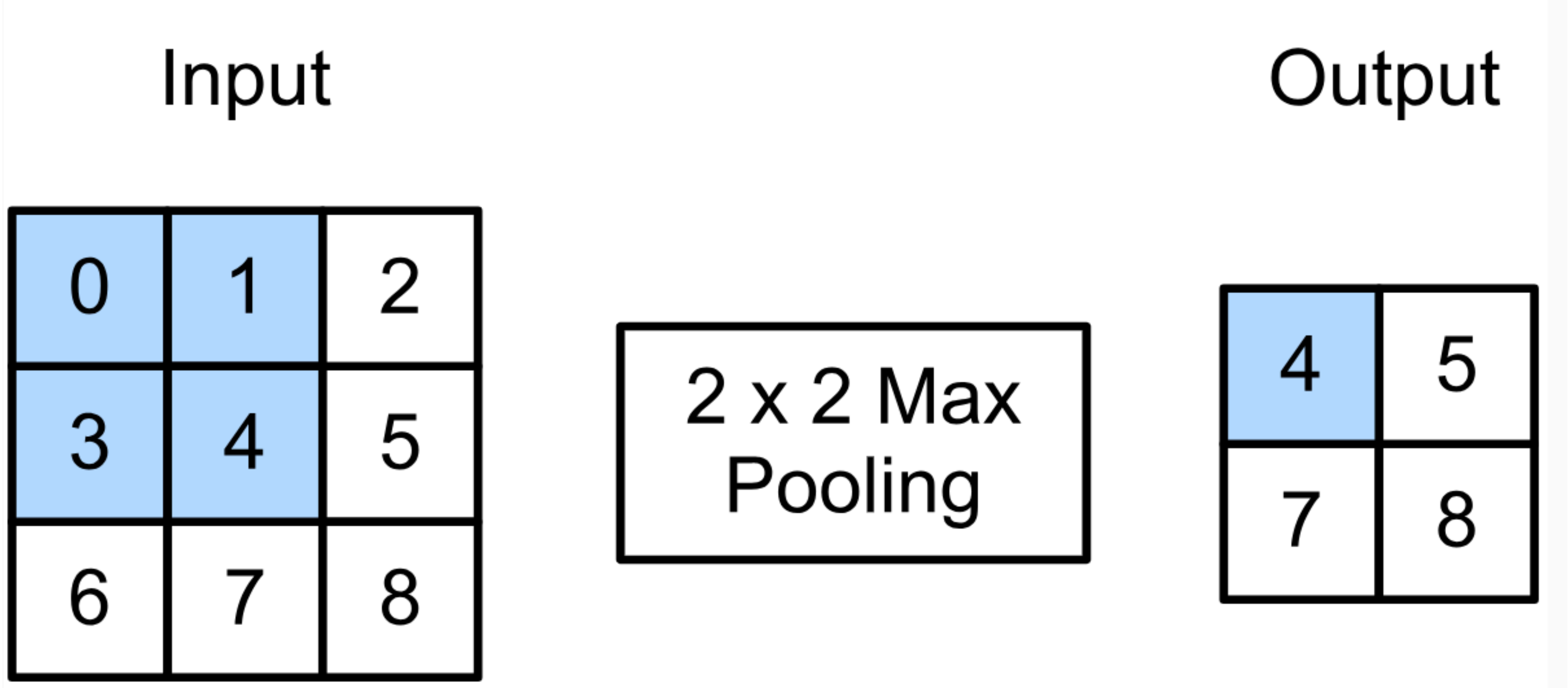


$$\max(0,1,3,4) = 4$$

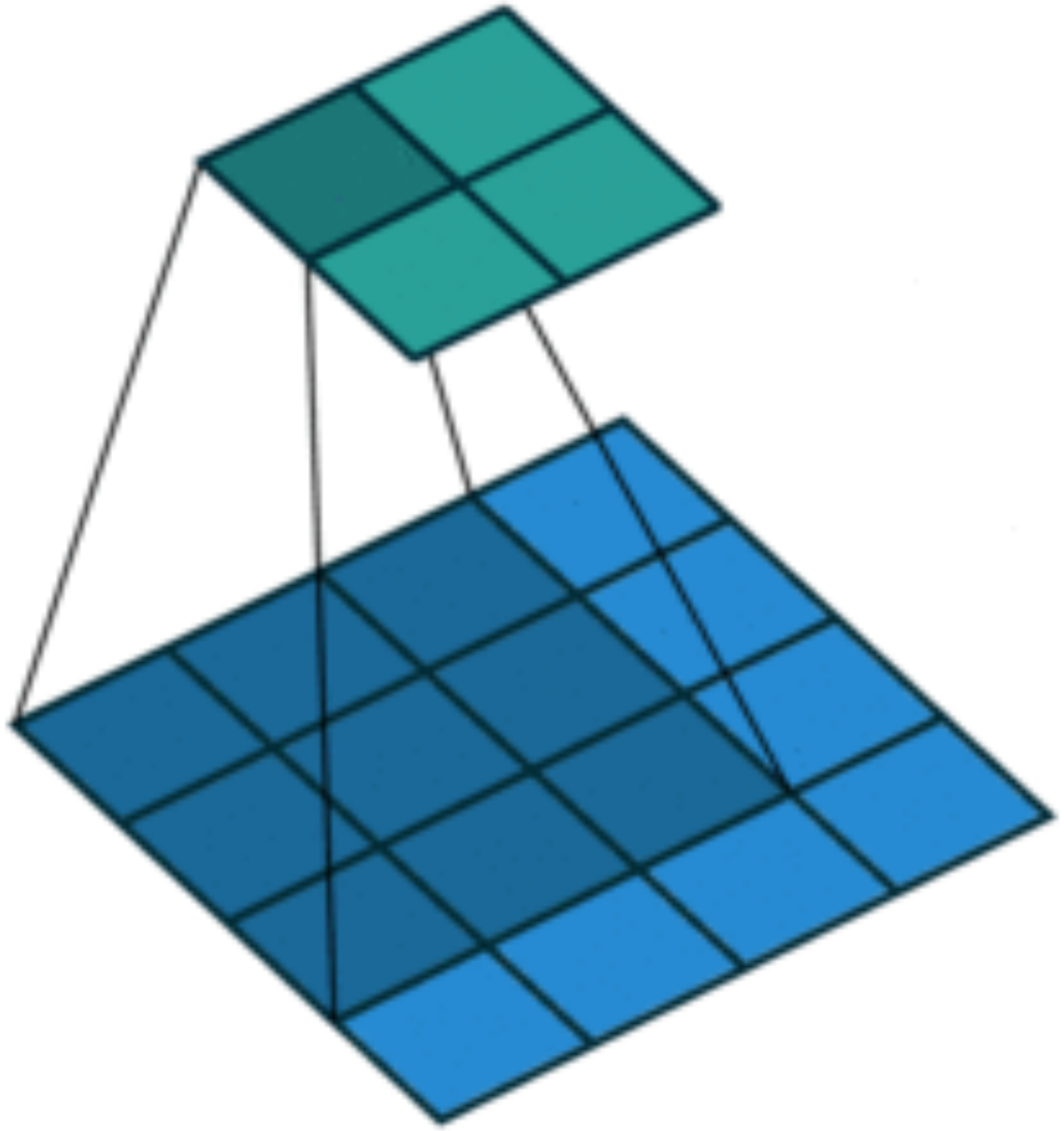


2-D Max Pooling

- Returns the maximal value in the sliding window



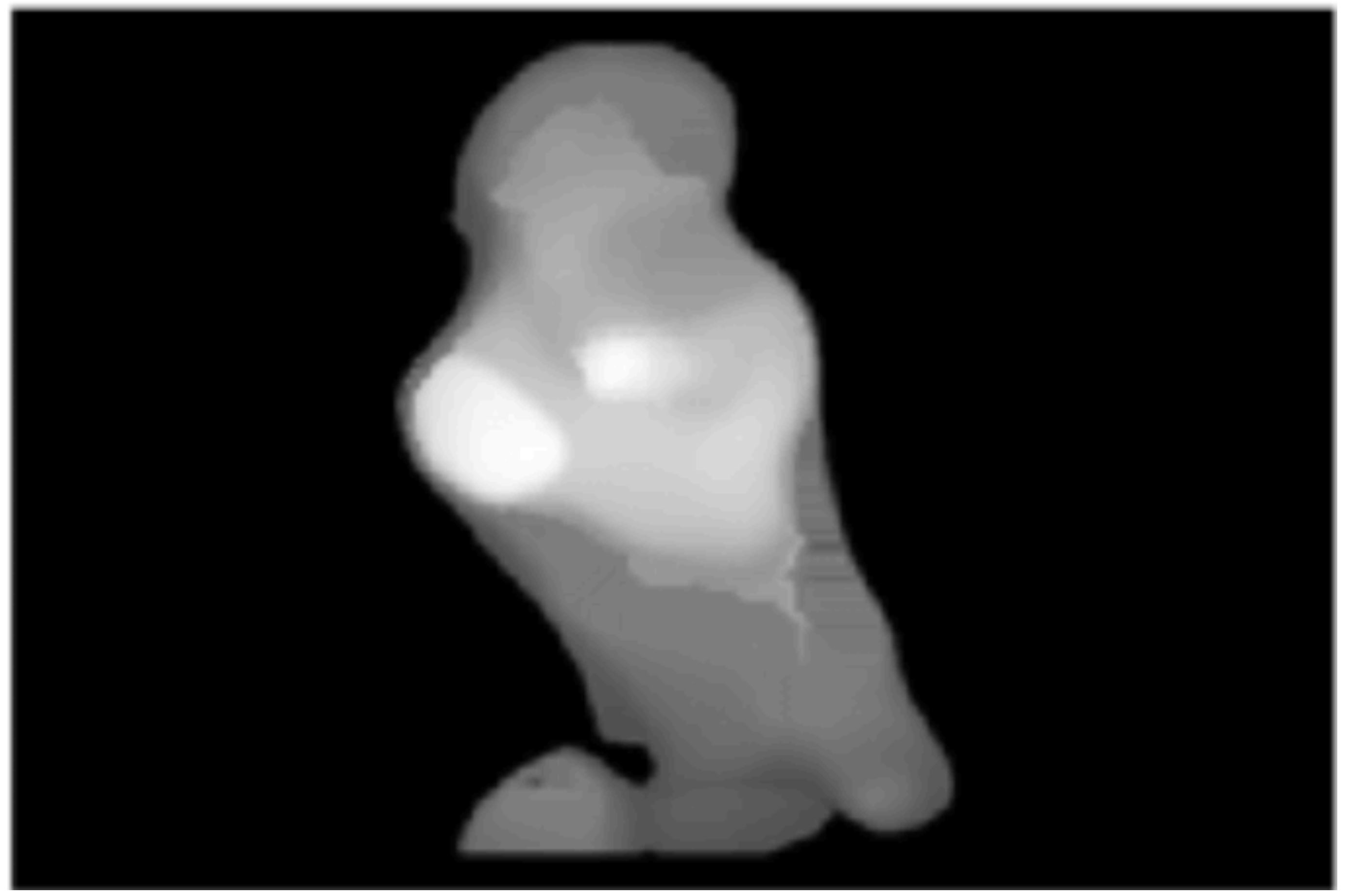
$$\max(0,1,3,4) = 4$$



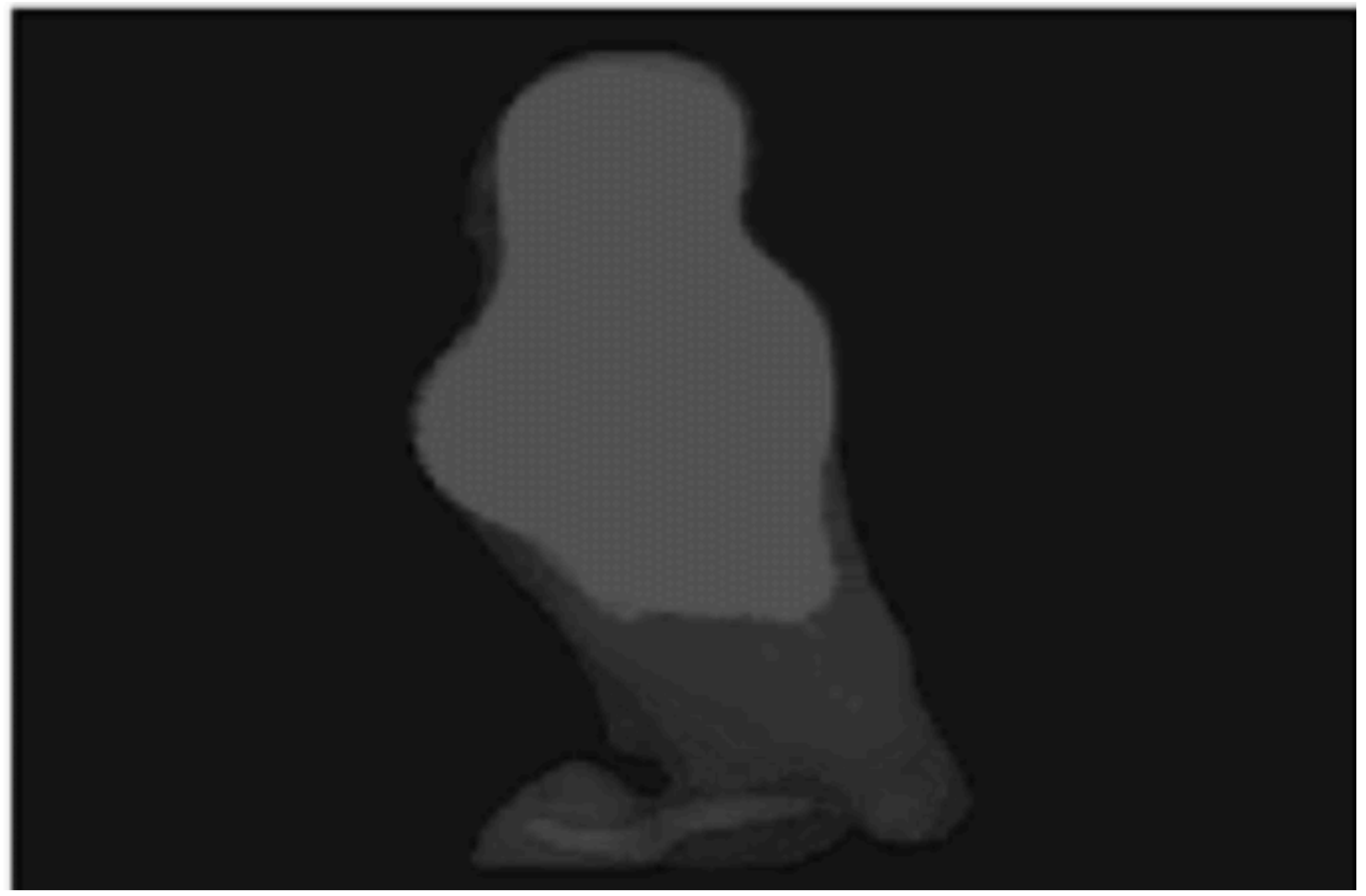
Average Pooling

- Max pooling: the strongest pattern signal in a window
- Average pooling: replace max with mean in max pooling
  - The average signal strength in a window

Max pooling



Average pooling



Convolutional Neural Network Architecture

Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner

Convolutional Neural Network Intuition

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Look for features of cats or dogs in the image and use for decision.

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Look for features of cats or dogs in the image and use for decision.

- Example: cats have cat-like faces, dogs have dog-like faces.

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Look for features of cats or dogs in the image and use for decision.

- Example: cats have cat-like faces, dogs have dog-like faces.
- How do you determine what is a “cat-like” face vs a “dog-like” face?

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Look for features of cats or dogs in the image and use for decision.

- Example: cats have cat-like faces, dogs have dog-like faces.
- How do you determine what is a “cat-like” face vs a “dog-like” face?

Look for features of “cat-like” faces and “dog-like” faces.

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Look for features of cats or dogs in the image and use for decision.

- Example: cats have cat-like faces, dogs have dog-like faces.
- How do you determine what is a “cat-like” face vs a “dog-like” face?

Look for features of “cat-like” faces and “dog-like” faces.

- Example: Dogs have longer snouts.

Convolutional Neural Network Intuition

Early layers recognize simple visual features, later layers recognize more complex visual features.

Suppose we want to classify pictures of either cats or dogs. How would you do this?

Look for features of cats or dogs in the image and use for decision.

- Example: cats have cat-like faces, dogs have dog-like faces.
- How do you determine what is a “cat-like” face vs a “dog-like” face?

Look for features of “cat-like” faces and “dog-like” faces.

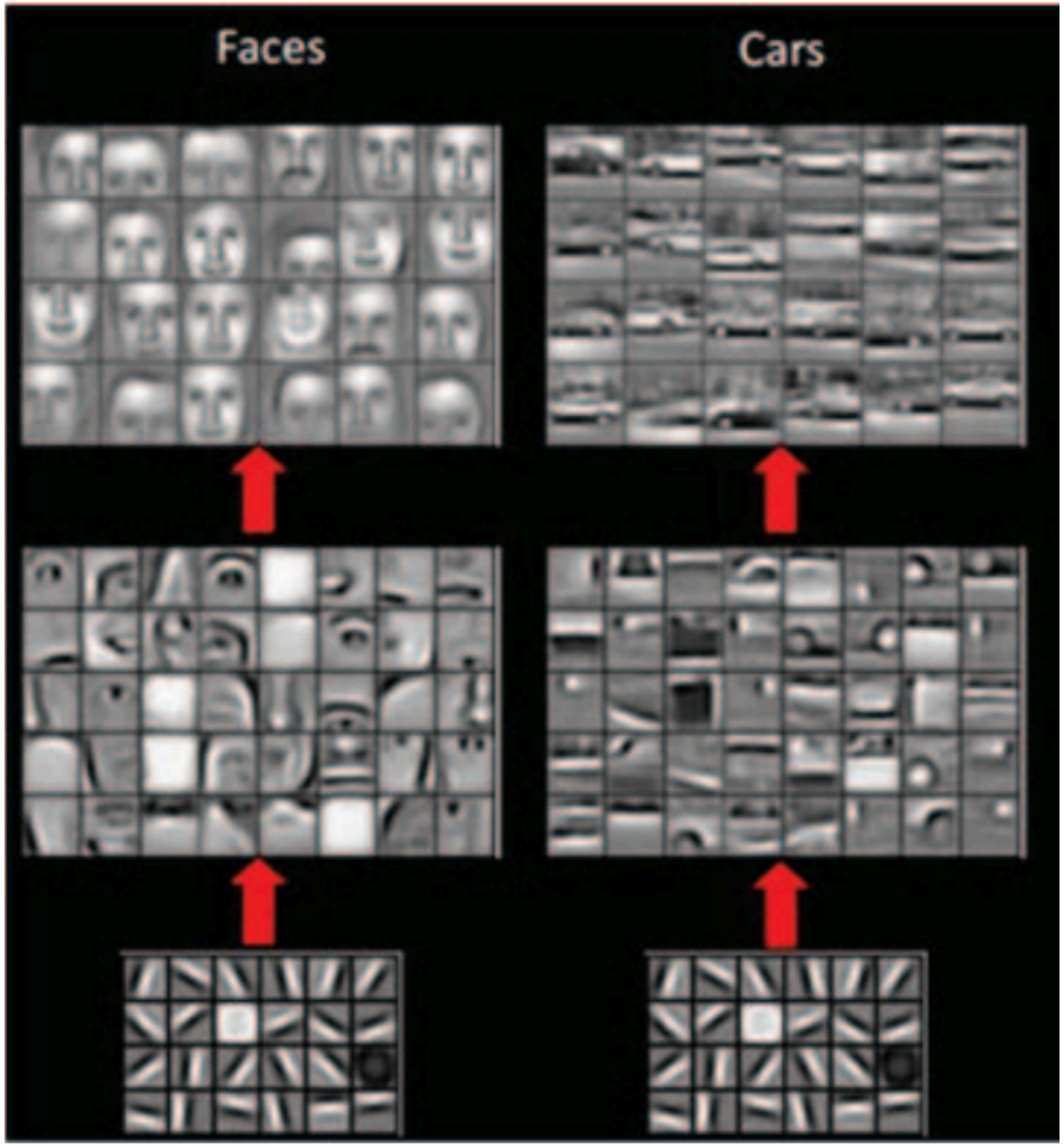
- Example: Dogs have longer snouts.
- How do you determine what is a long snout?

Feature Learning

Later layers recognize complete objects

Middle layers recognize parts of objects

Early layers recognize simple patterns



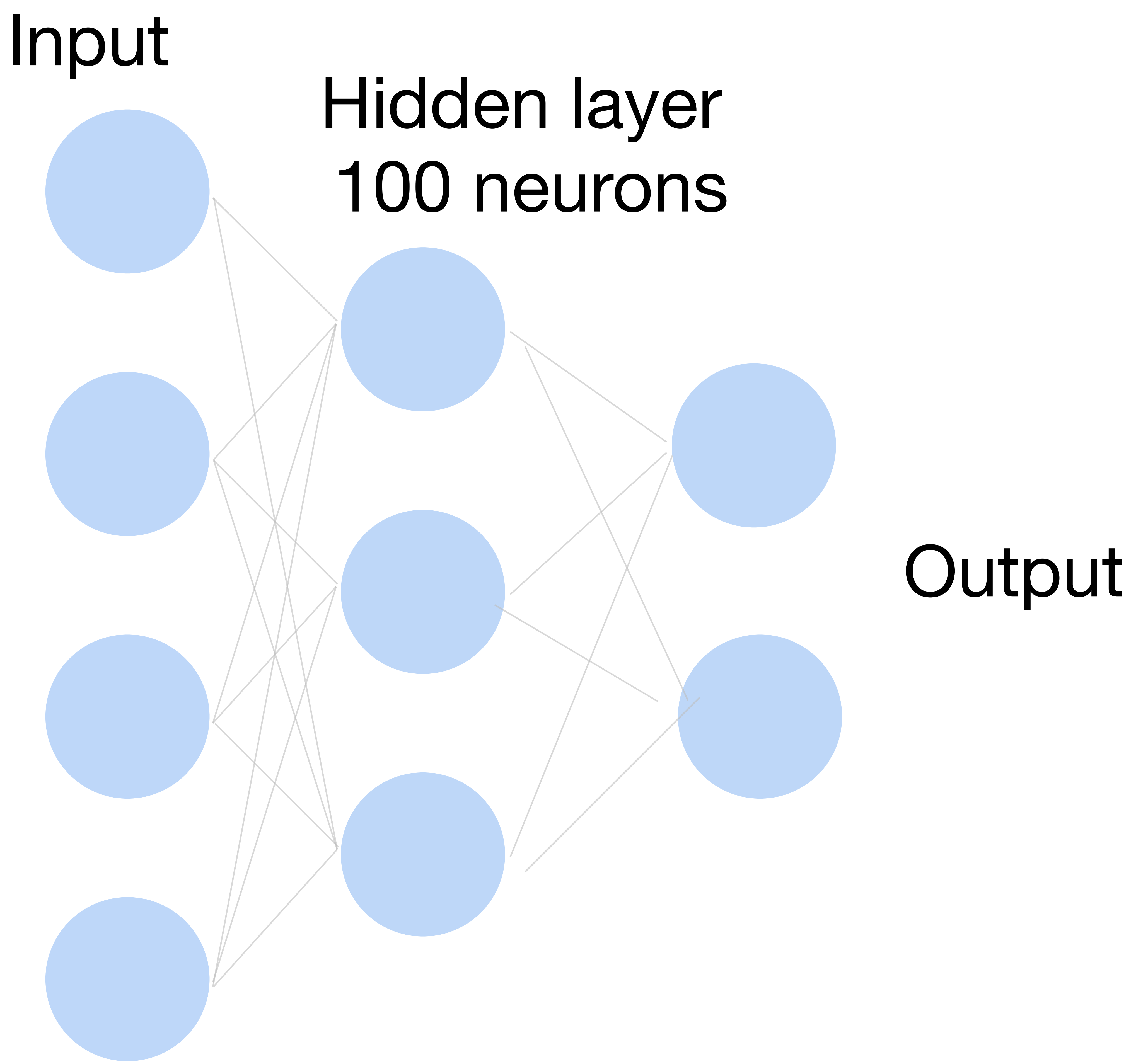
Adaptive Neuron Apoptosis for Accelerating Deep Learning on Large Scale Systems. Seigel et al. 2016.

Convolutional Neural Network Architecture

Gradient-based learning applied to document recognition,
by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner

How to train a neural network?

Loss function: $$\frac{1}{|D|}\sum_i \ell(\mathbf{x}_i, y_i)$$

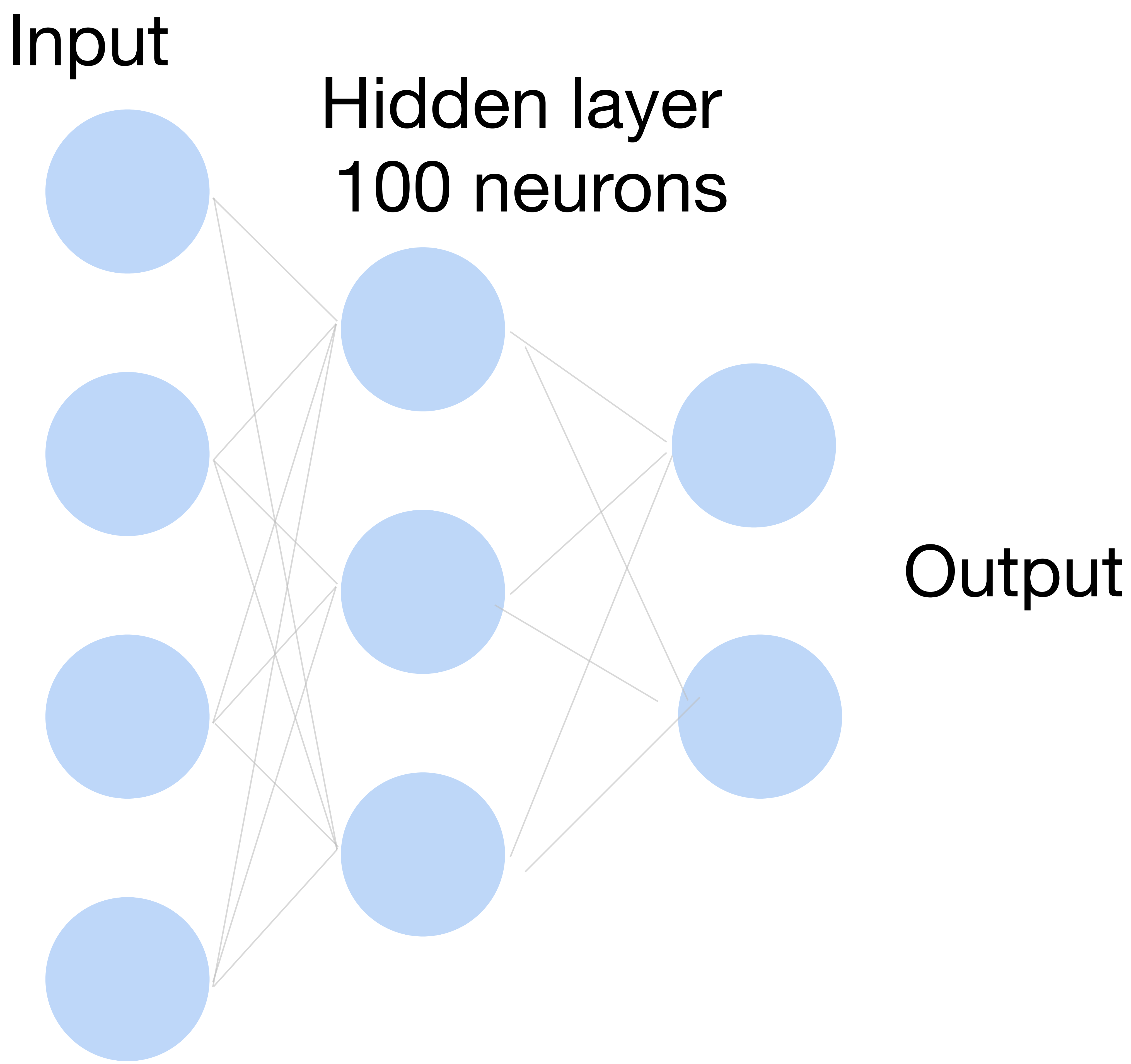


How to train a neural network?

Loss function: $$\frac{1}{|D|}\sum_i \ell(\mathbf{x}_i, y_i)$$

Per-sample loss:

$$\ell(\mathbf{x}, y) = \sum_{j=1}^{K} - y_j \log p_j$$



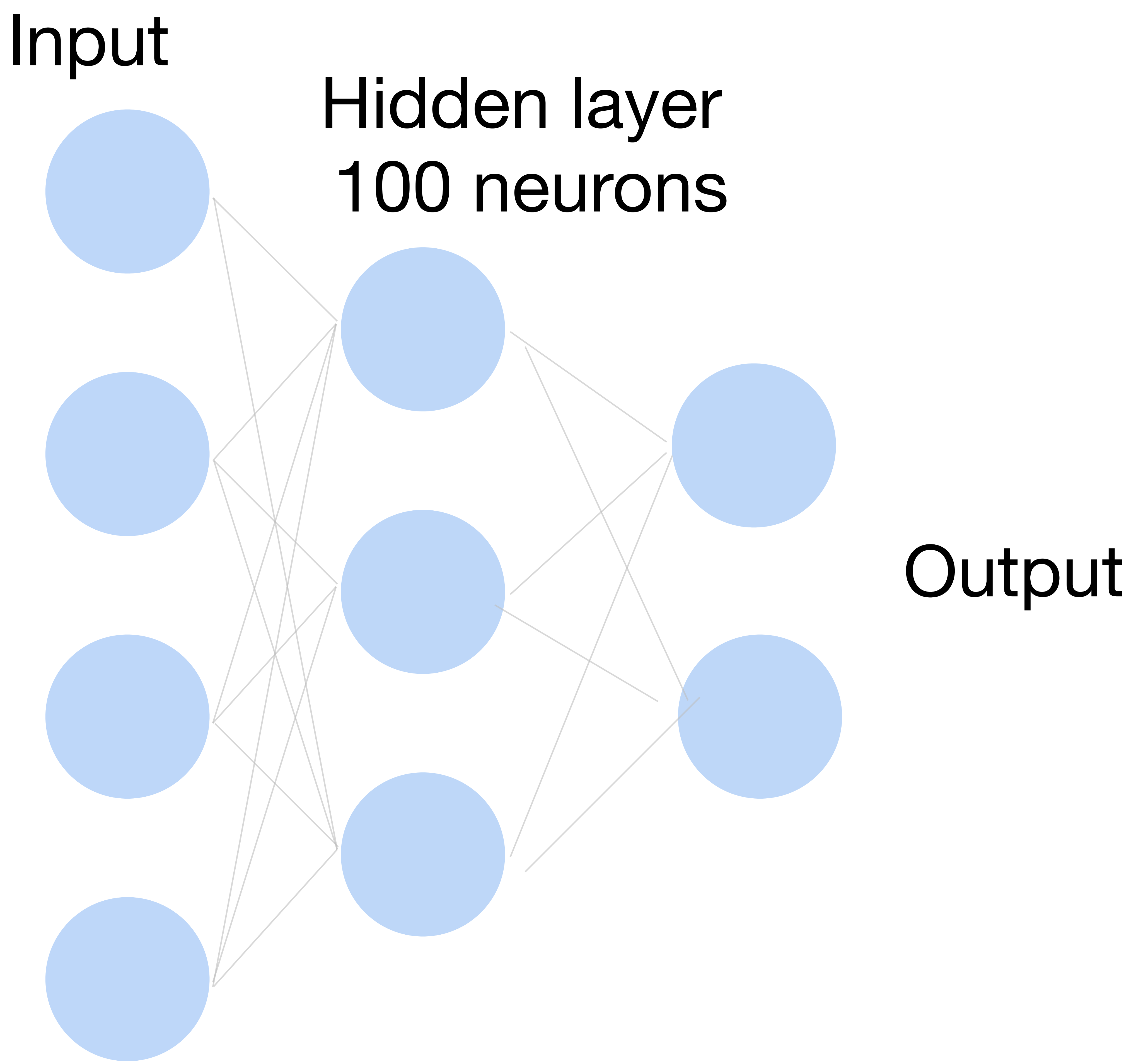
How to train a neural network?

Loss function: $$\frac{1}{|D|}\sum_i \ell(\mathbf{x}_i, y_i)$$

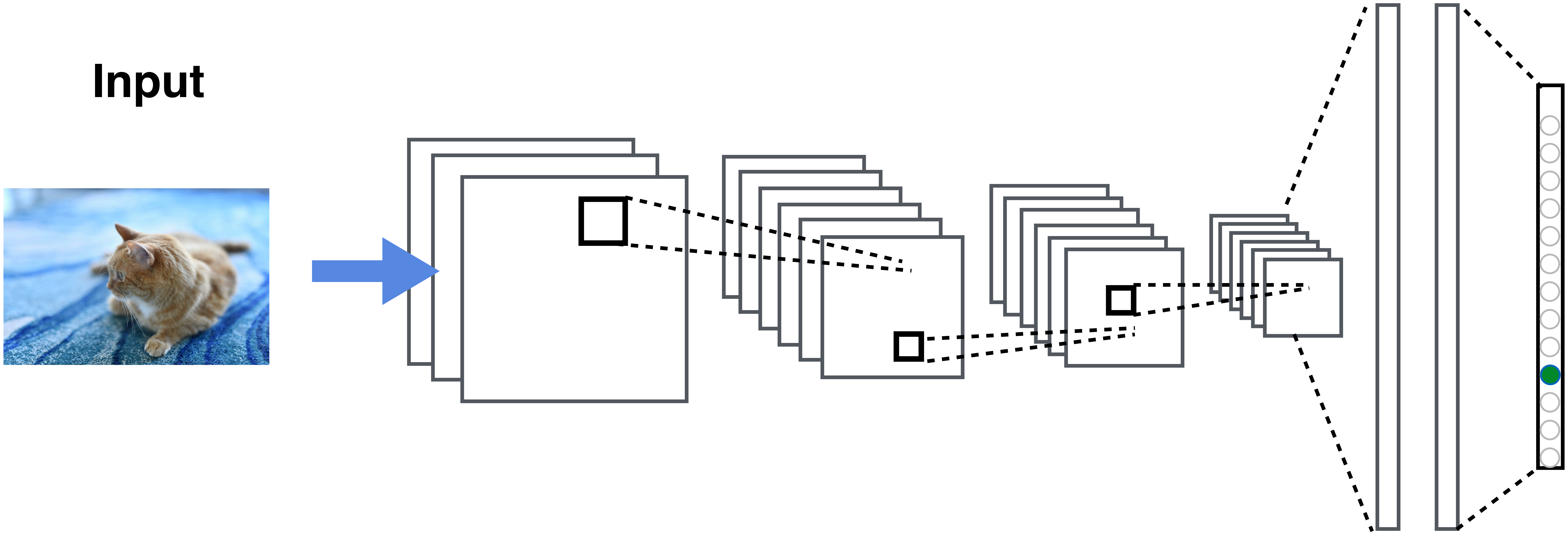
Per-sample loss:

$$\ell(\mathbf{x}, y) = \sum_{j=1}^{K} - y_j \log p_j$$

Also known as cross-entropy loss or softmax loss

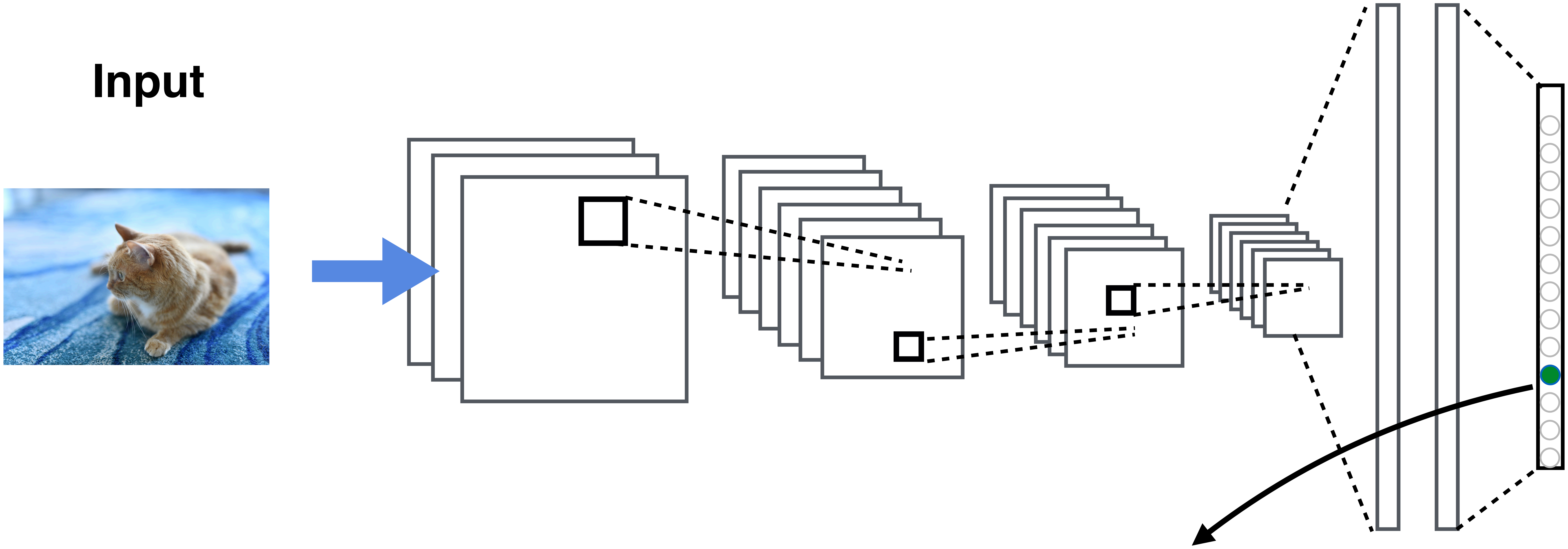


How to train a convolutional neural network?



How to train a convolutional neural network?

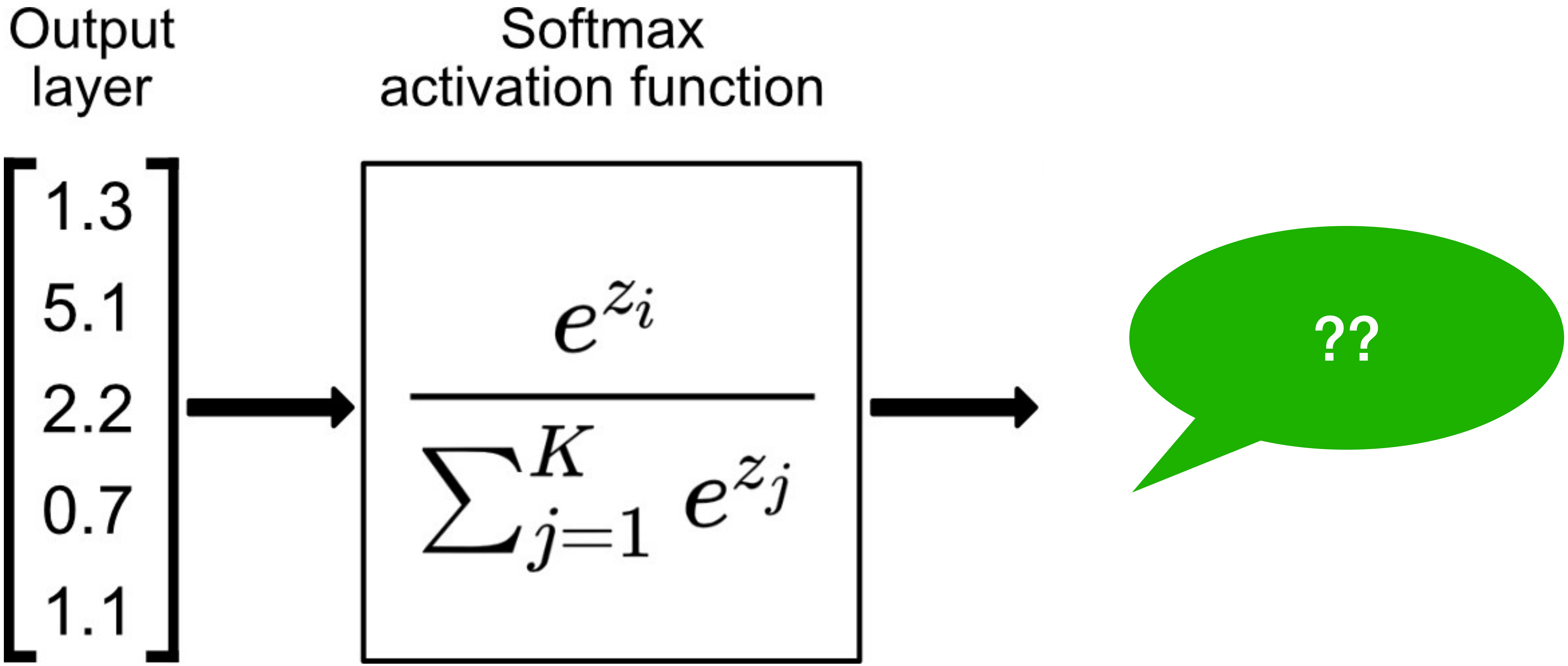
Input



$$p_i(\boldsymbol{x}) = \frac{\exp\left(f_i(\boldsymbol{x})\right)}{\sum_{j=1}^{N} \exp\left(f_j(\boldsymbol{x})\right)}, \quad \text{softmax}$$

Recall Softmax

Turns outputs f into probabilities (sum up to 1 across k classes)



Recall Softmax

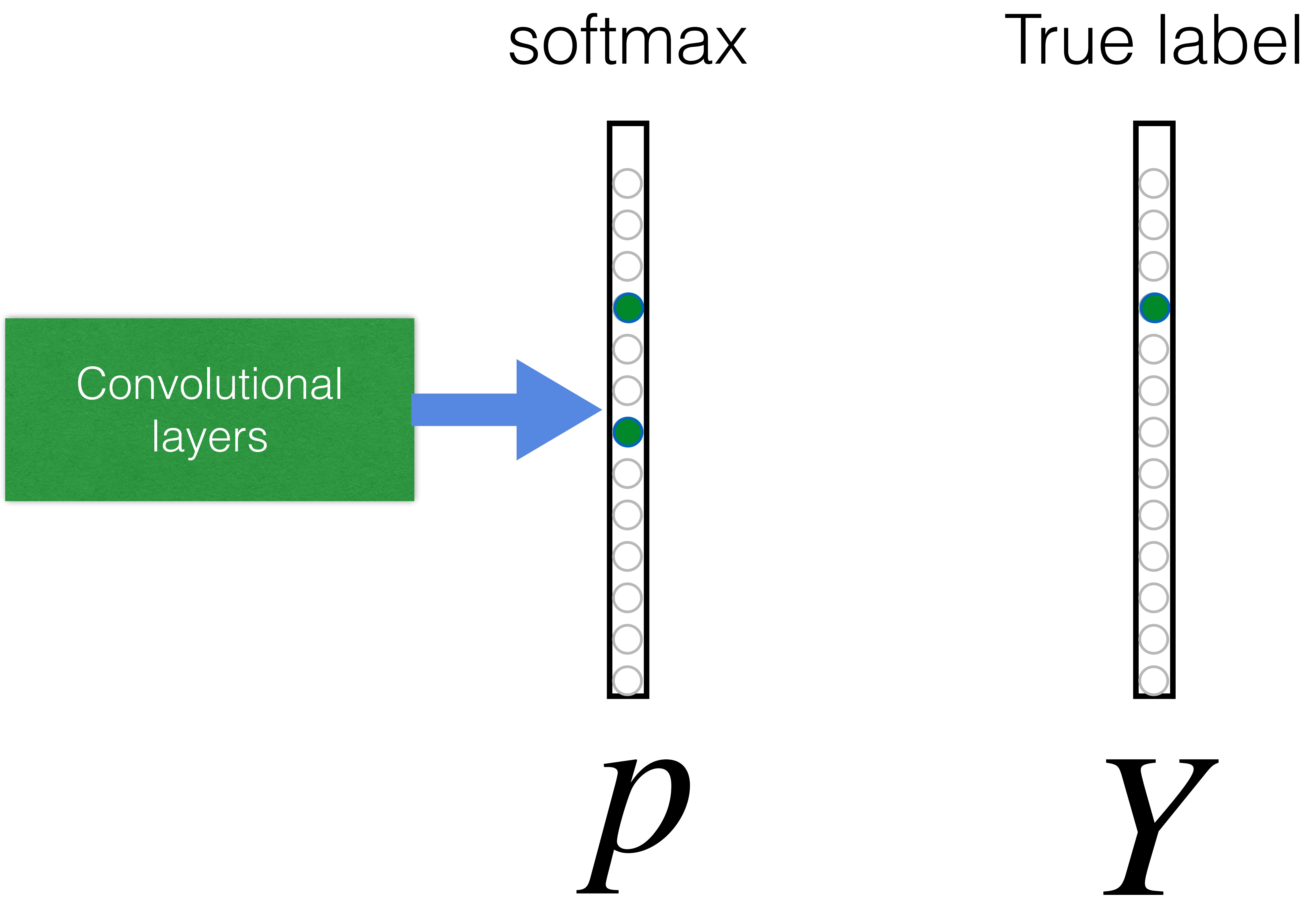
Turns outputs f into probabilities (sum up to 1 across k classes)

Output layer | Softmax activation function | Probabilities

$$\begin{bmatrix} 1.3 \\ 5.1 \\ 2.2 \\ 0.7 \\ 1.1 \end{bmatrix} \rightarrow \boxed{\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}} \rightarrow \begin{bmatrix} 0.02 \\ 0.90 \\ 0.05 \\ 0.01 \\ 0.02 \end{bmatrix}$$

Normalized

Cross-Entropy Loss

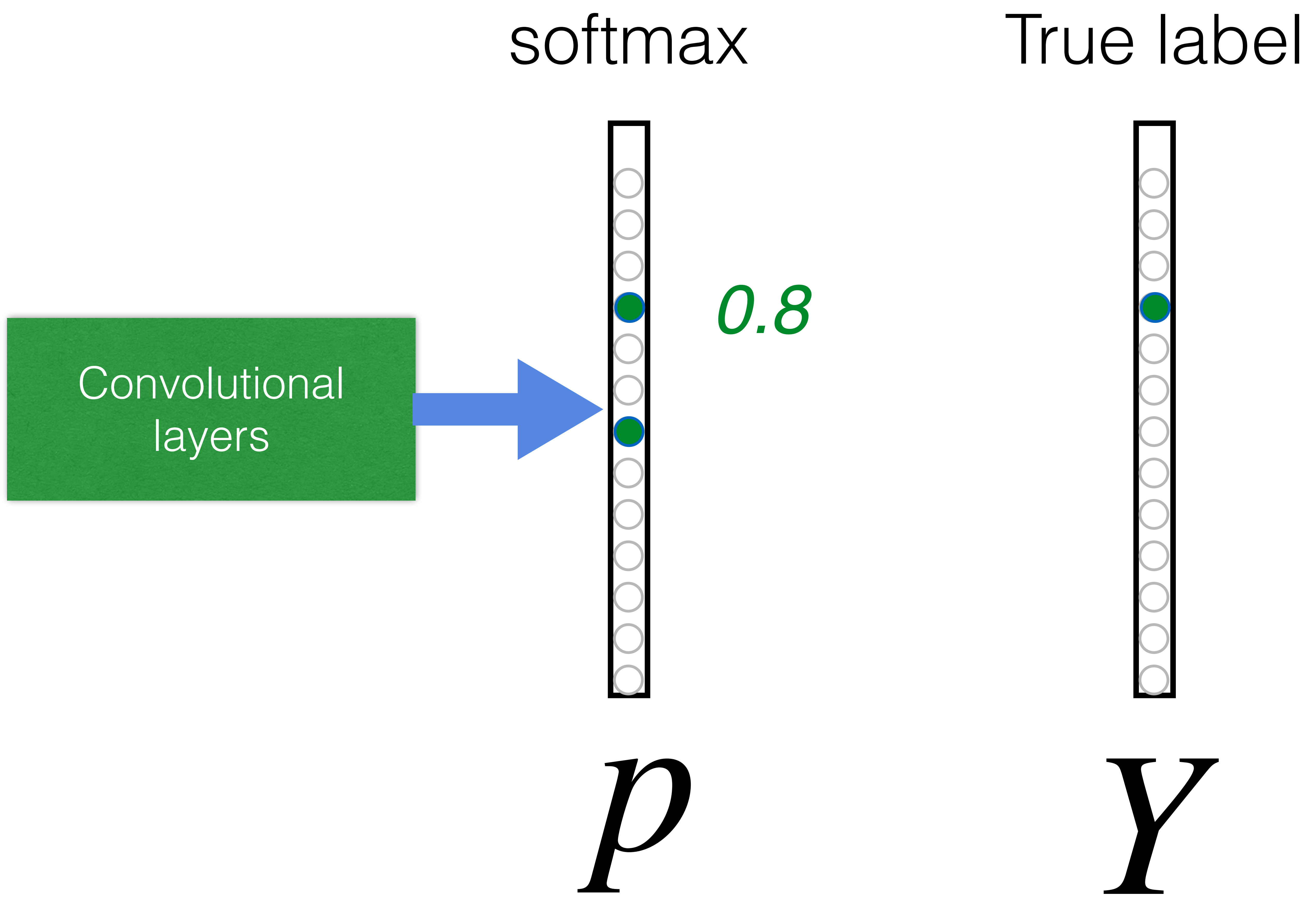


$$L_{CE} = \sum_{i} -Y_i \log(p_i)$$

$$= -\log(0.8)$$

Goal: push **p** and **Y** to be identical

Cross-Entropy Loss

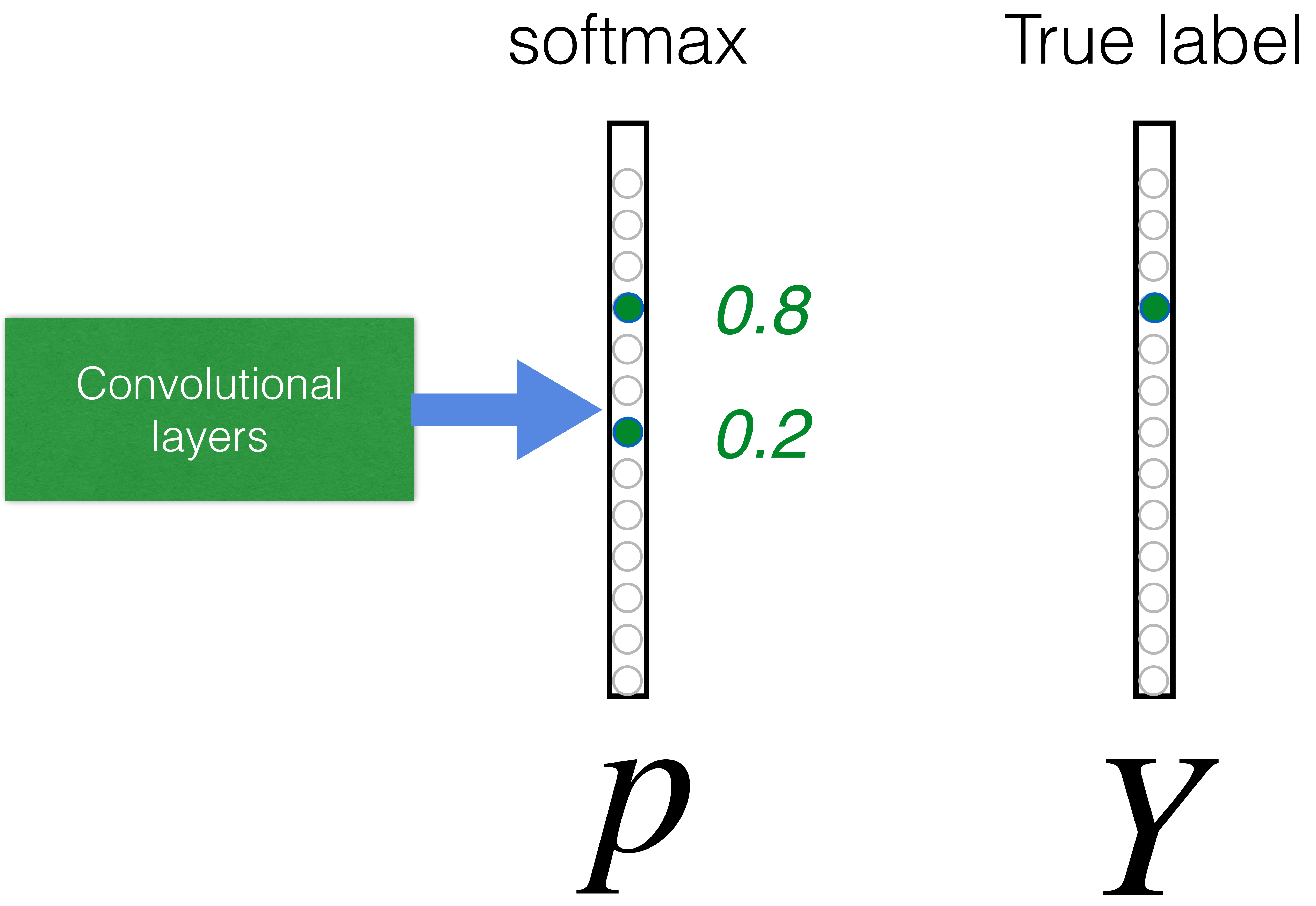


$$L_{CE} = \sum_i -Y_i \log(p_i)$$

$$= -\log(0.8)$$

Goal: push **p** and **Y** to be identical

Cross-Entropy Loss

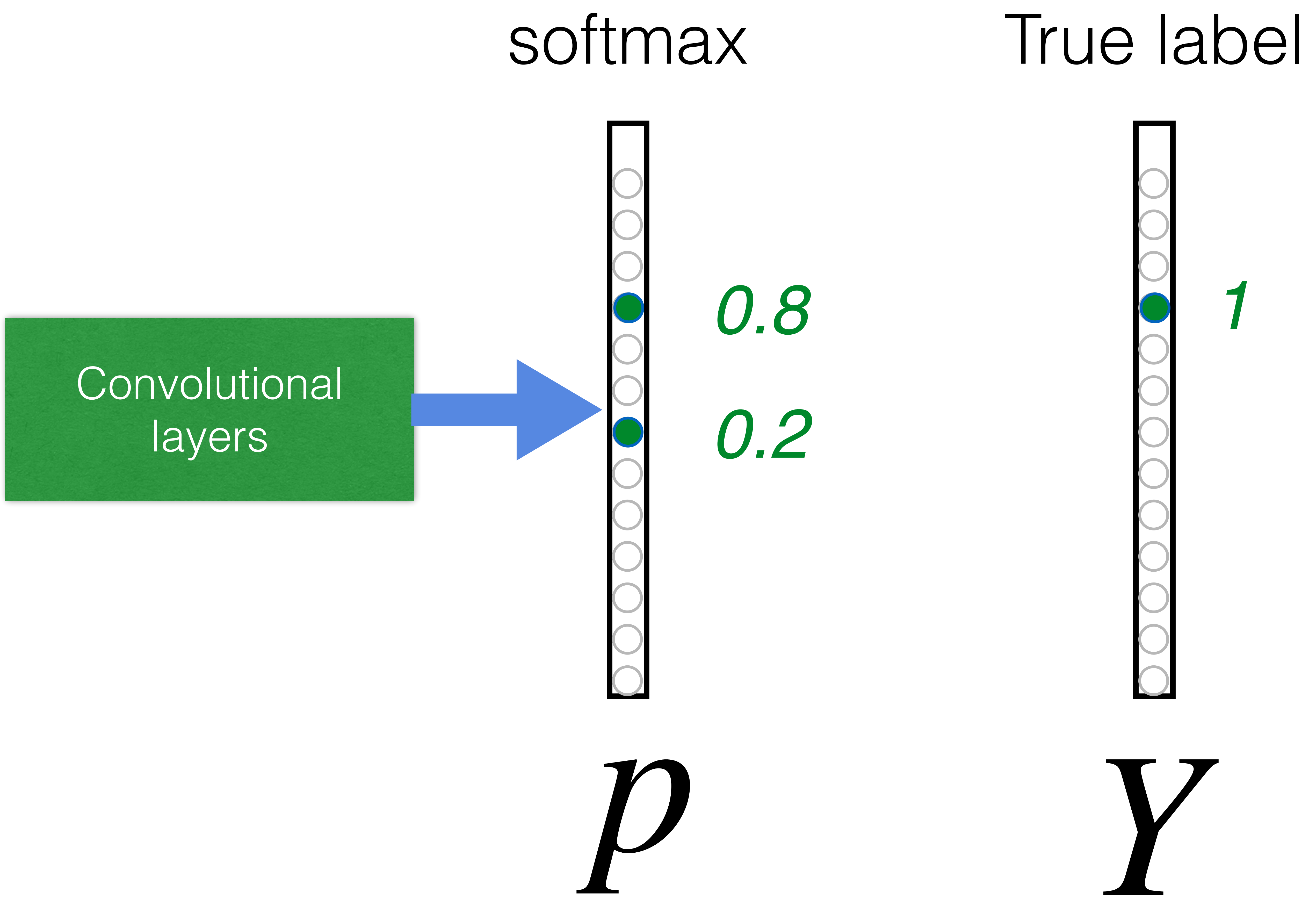


$$L_{CE} = \sum_i - Y_i \log(p_i)$$

$$= -\log(0.8)$$

Goal: push **p** and **Y** to be identical

Cross-Entropy Loss



$$L_{CE} = \sum_i - Y_i \log(p_i)$$

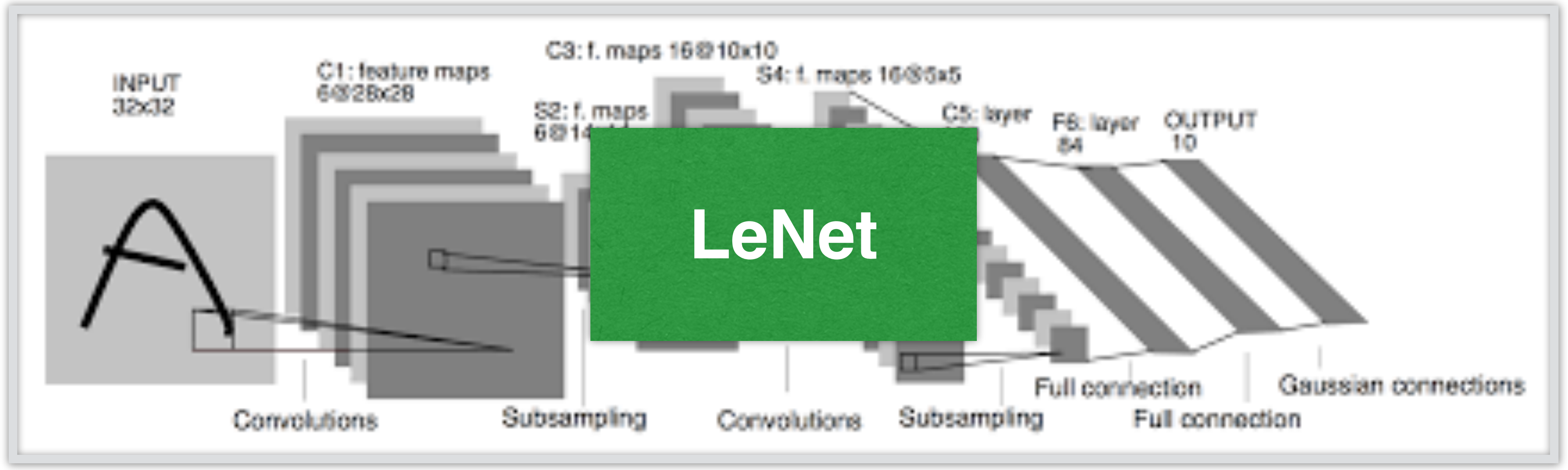
$$= -\log(0.8)$$

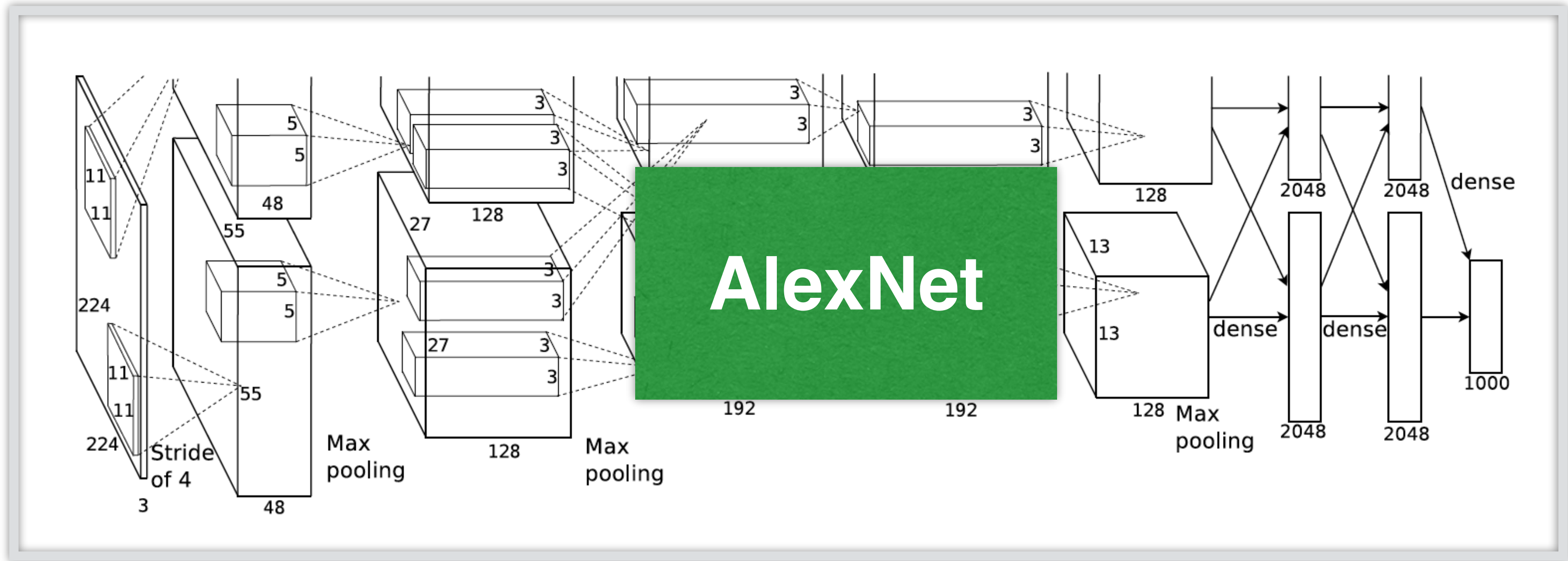
Goal: push **p** and **Y** to be identical

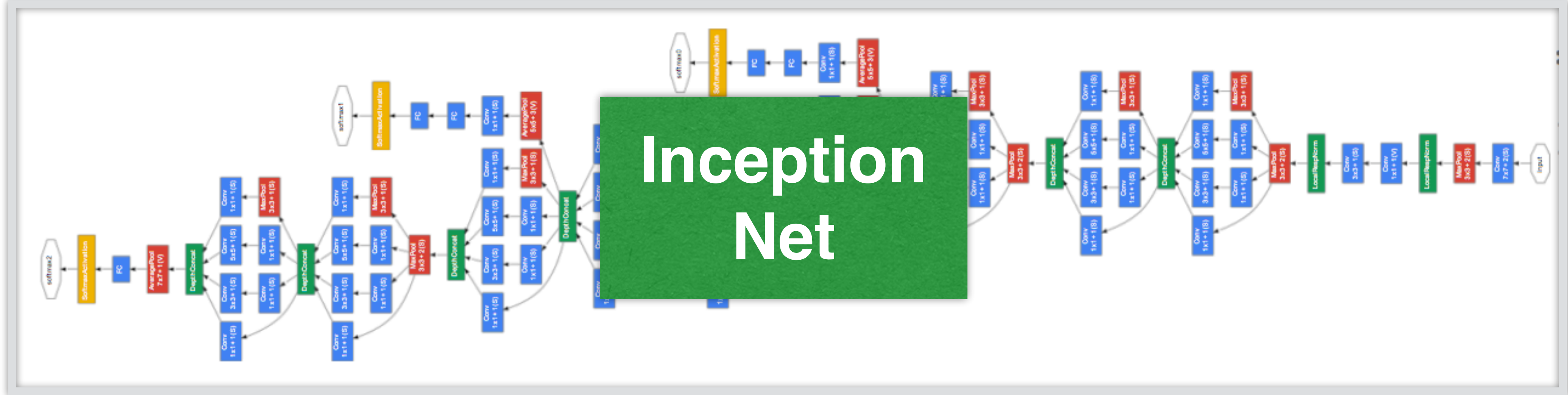
Convolutional Neural Networks

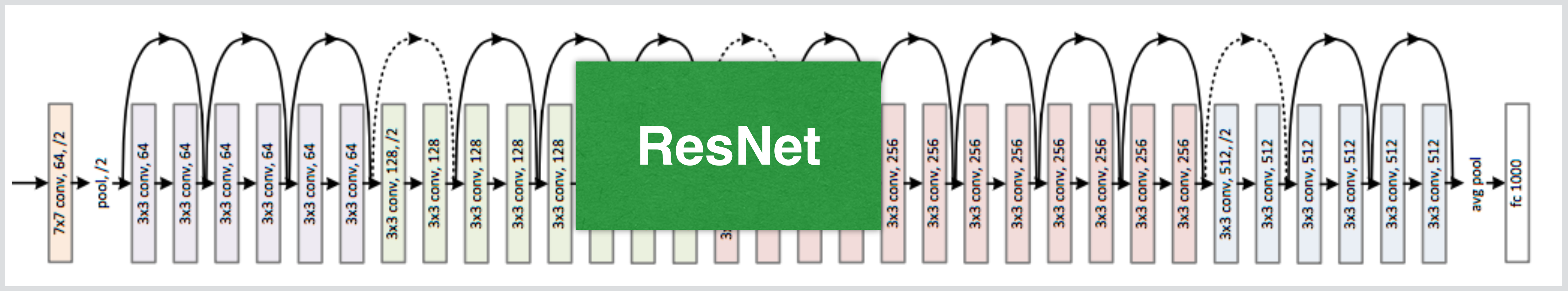
Evolution of neural net architectures

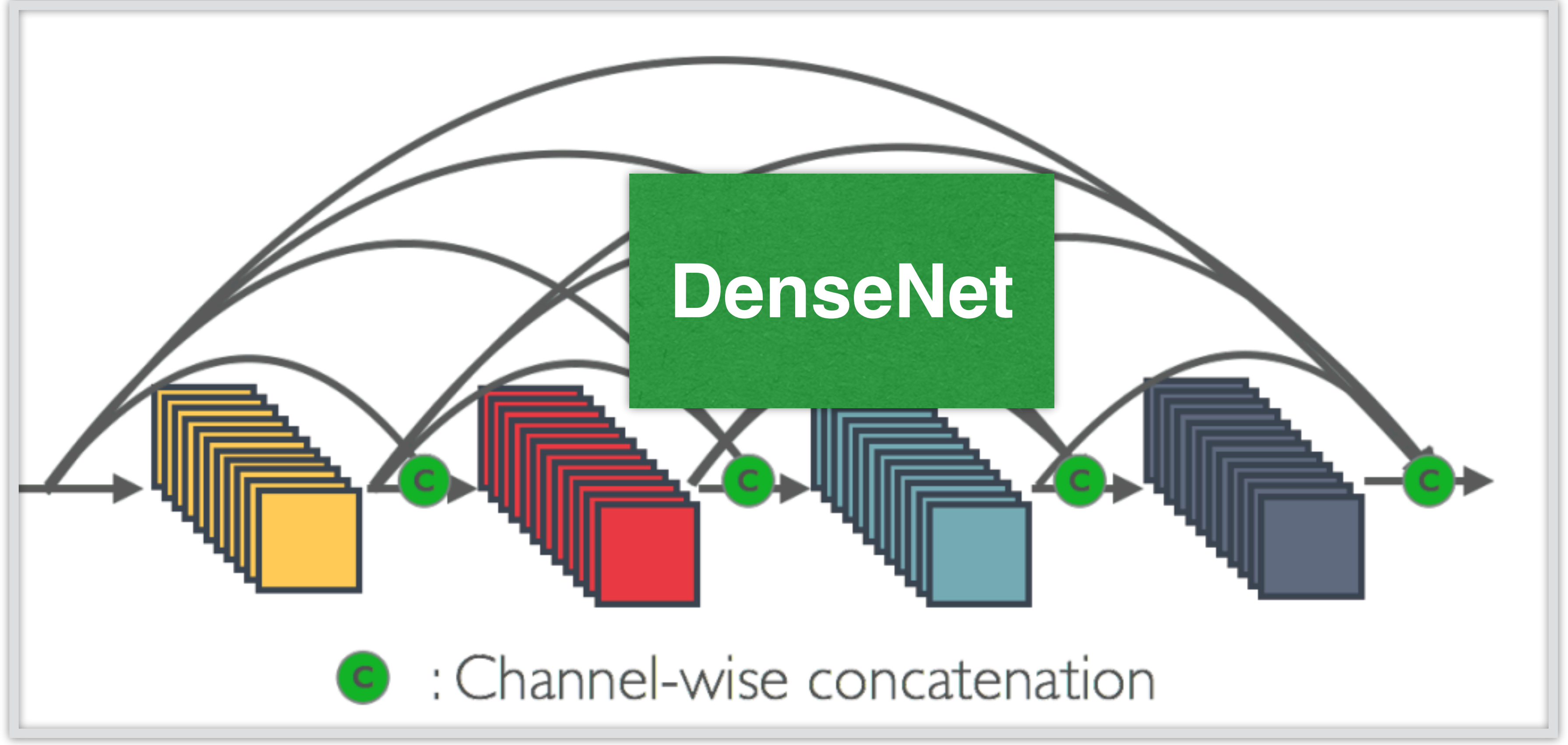
Evolution of neural net architectures







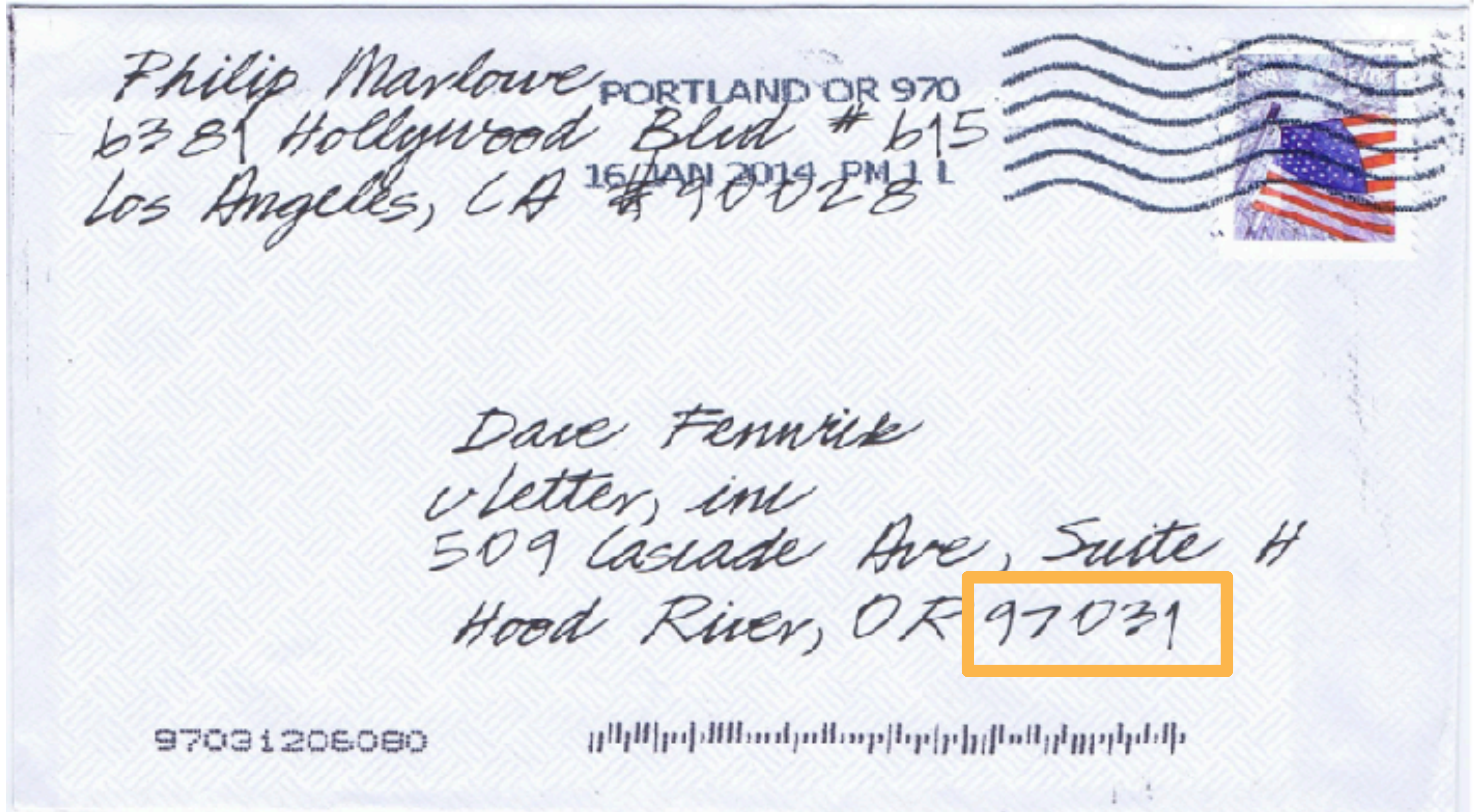


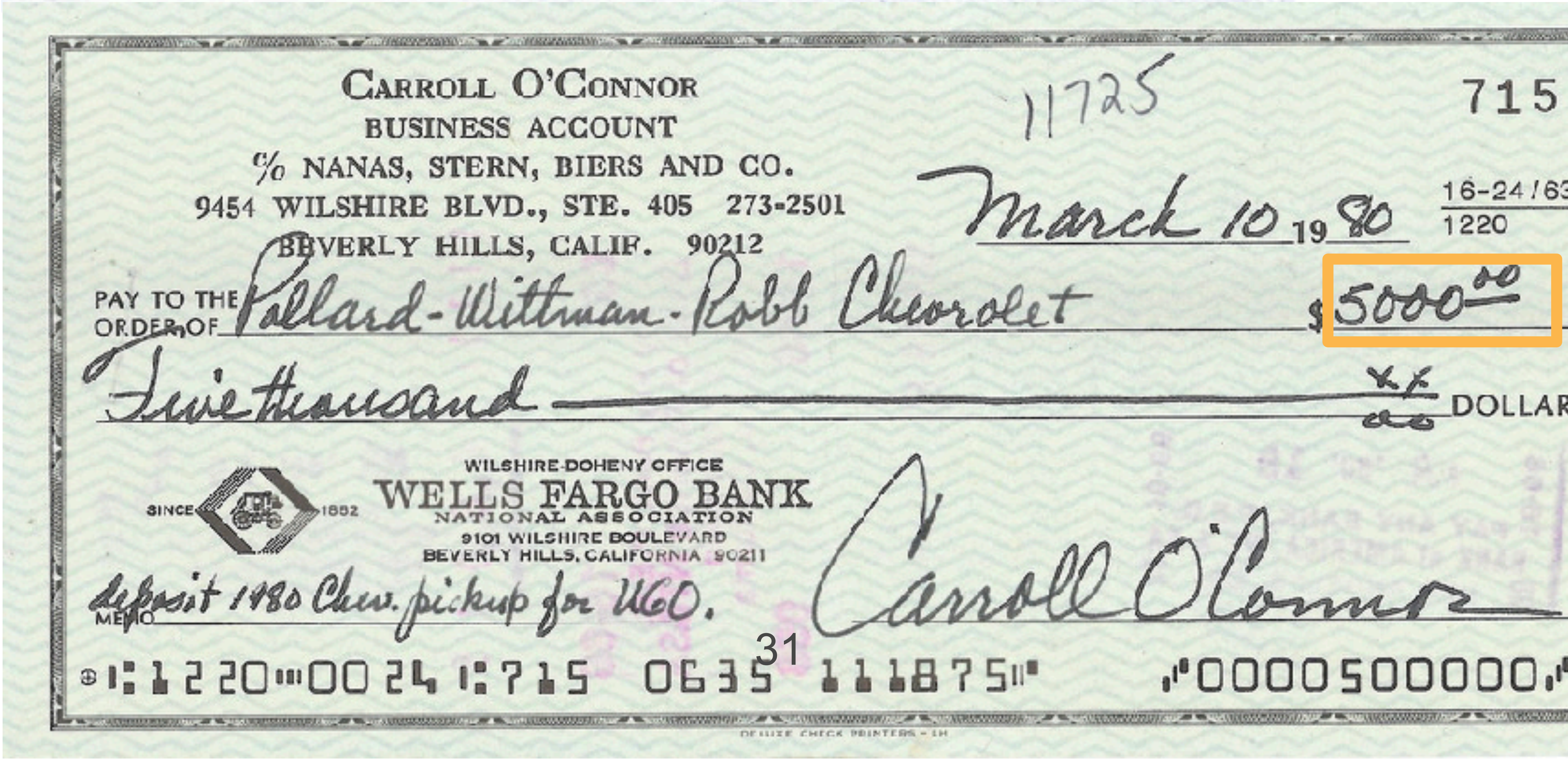


LeNet Architecture (first conv nets)

Gradient-based learning applied to document recognition,
by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner

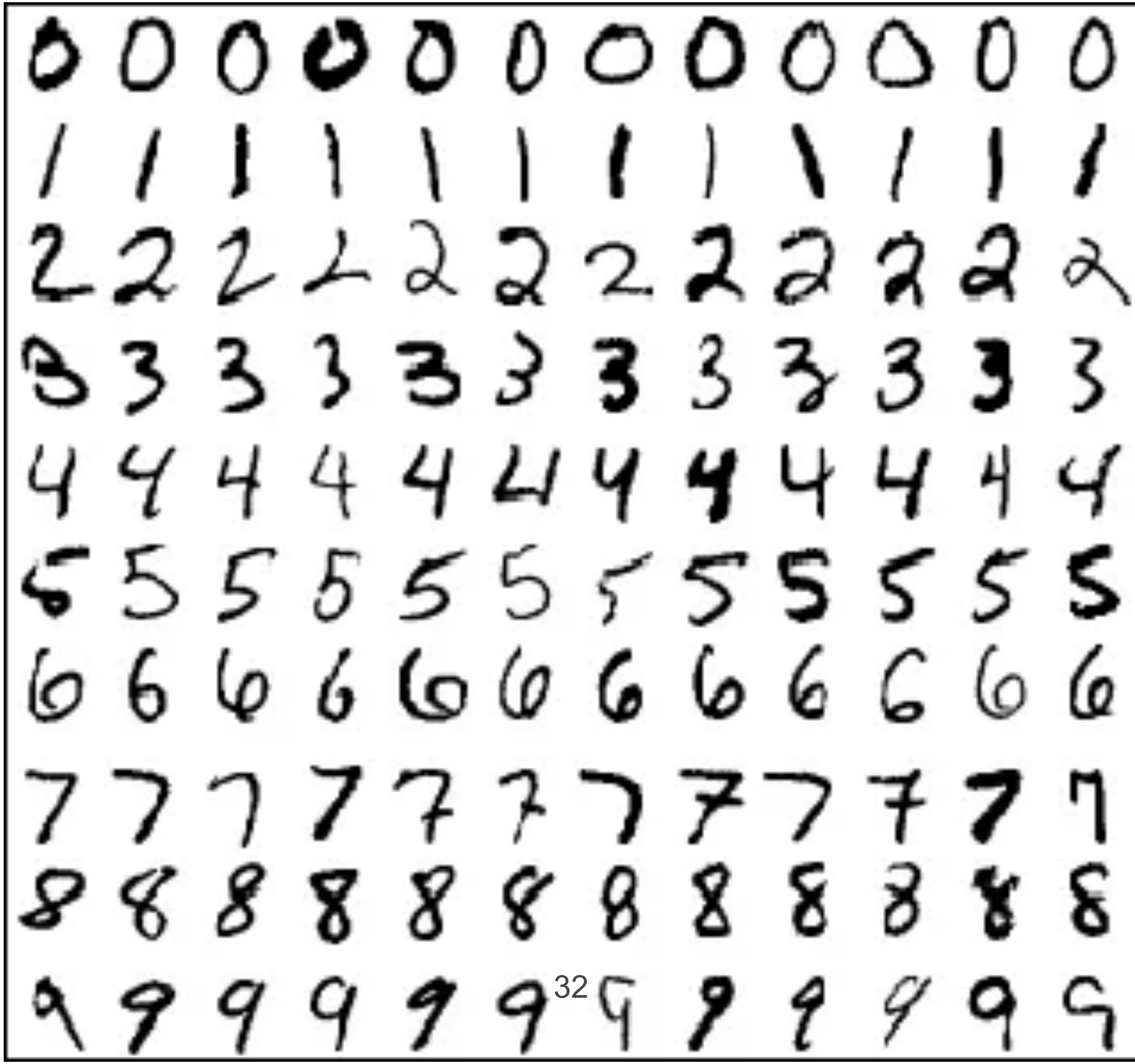
Handwritten Digit Recognition

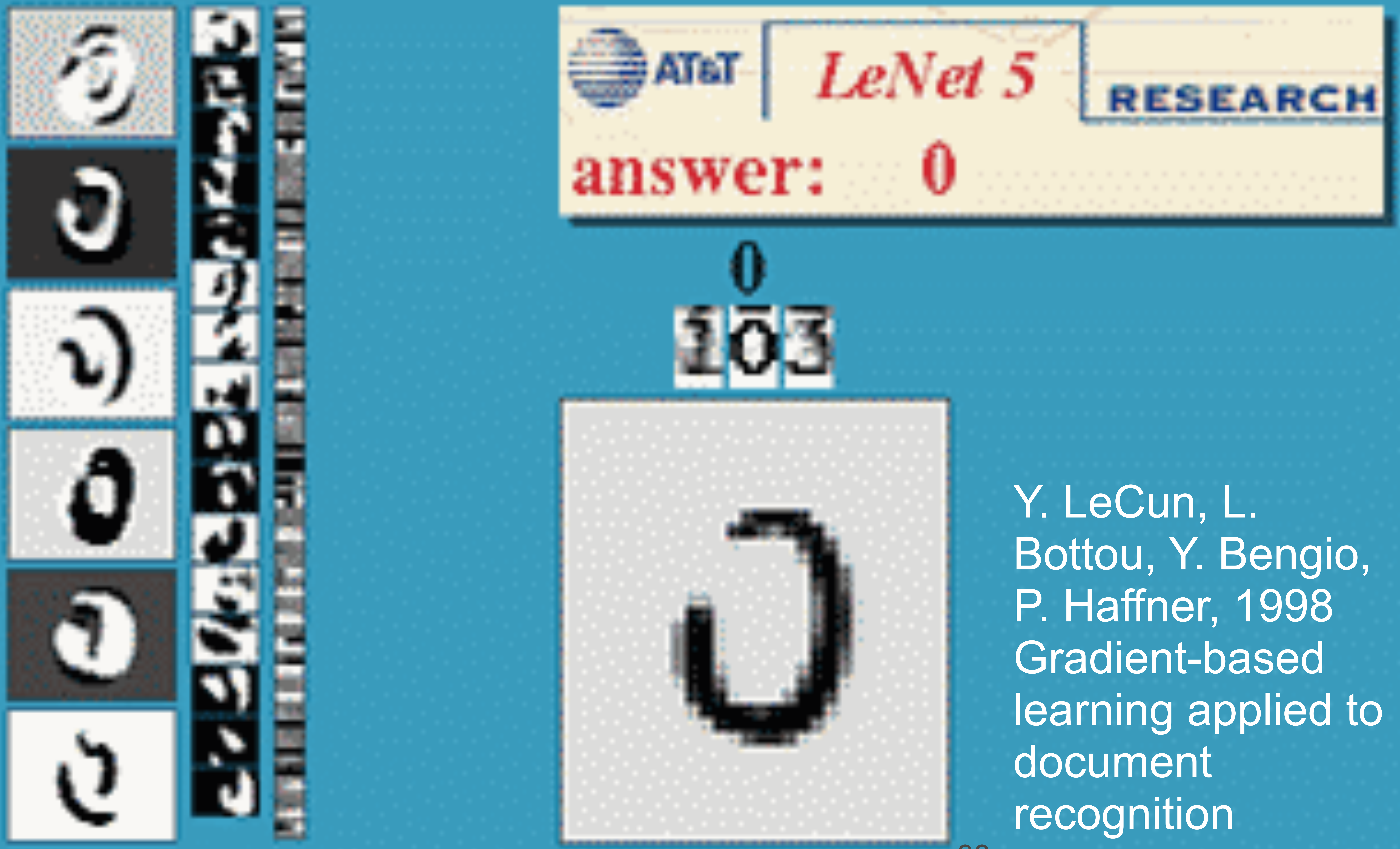




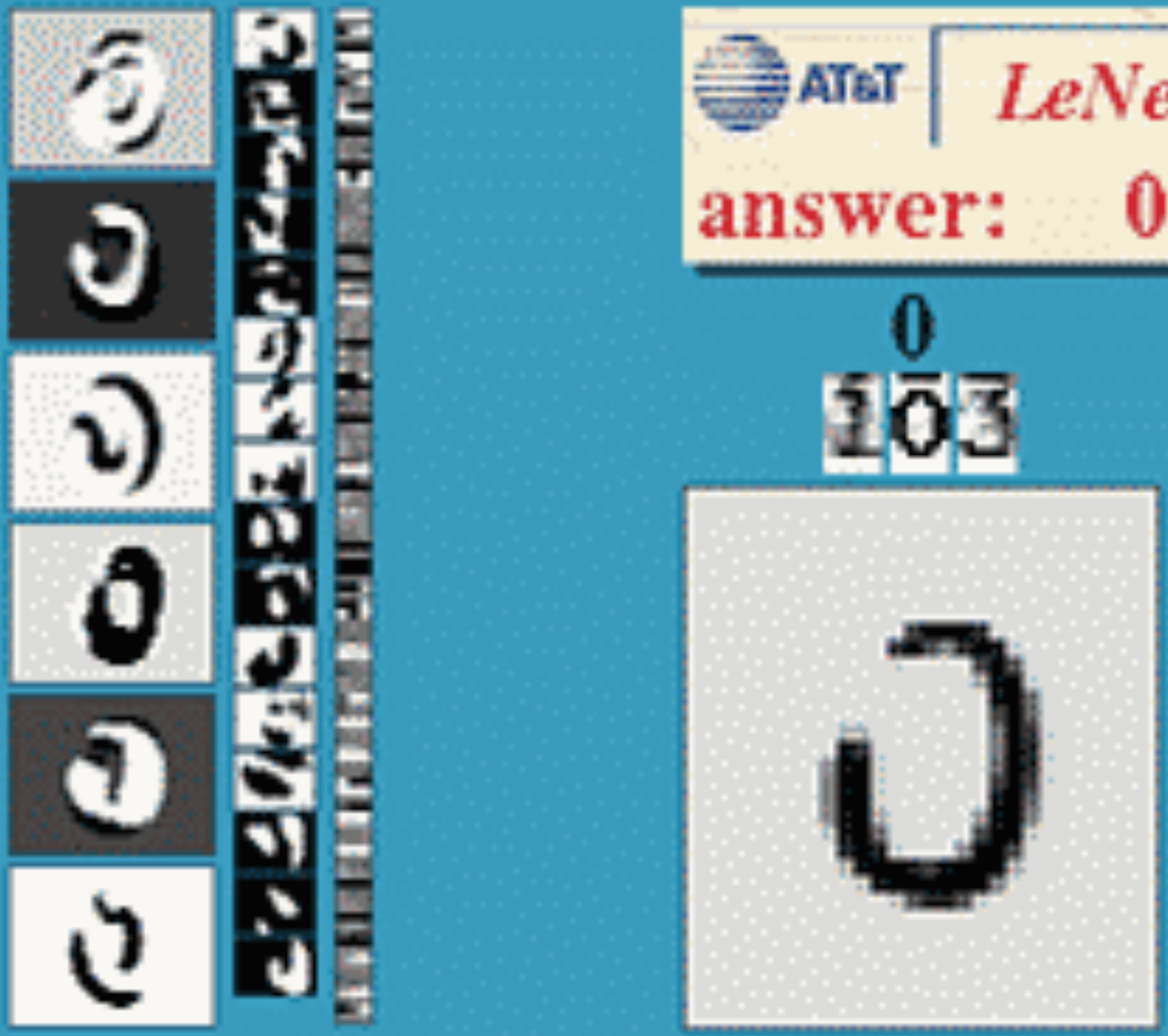
MNIST

- Centered and scaled
- 50,000 training data
- 10,000 test data
- 28 x 28 images
- 10 classes





Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, 1998 Gradient-based learning applied to document recognition



Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, 1998 Gradient-based learning applied to document recognition

LeNet Architecture

Gradient-based learning applied to document recognition, by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner

LeNet in Pytorch

```
def __init__(self):
    super(LeNet5, self).__init__()
    # Convolution (In LeNet-5, 32x32 images are given as input. Hence padding of 2 is done below)
    self.conv1 = torch.nn.Conv2d(in_channels=1, out_channels=6, kernel_size=5, stride=1, padding=2, bias=True)
    # Max-pooling
    self.max_pool_1 = torch.nn.MaxPool2d(kernel_size=2)
    # Convolution
    self.conv2 = torch.nn.Conv2d(in_channels=6, out_channels=16, kernel_size=5, stride=1, padding=0, bias=True)
    # Max-pooling
    self.max_pool_2 = torch.nn.MaxPool2d(kernel_size=2)
    # Fully connected layer
    self.fc1 = torch.nn.Linear(16*5*5, 120)   # convert matrix with 16*5*5 (= 400) features to a matrix of 120 features (col
    self.fc2 = torch.nn.Linear(120, 84)       # convert matrix with 120 features to a matrix of 84 features (columns)
    self.fc3 = torch.nn.Linear(84, 10)        # convert matrix with 84 features to a matrix of 10 features (columns)
```

https://github.com/bollakarthikeya/LeNet-5-PyTorch/blob/master/lenet5_gpu.py

```python
def forward(self, x):
    # convolve, then perform ReLU non-linearity
    x = torch.nn.functional.relu(self.conv1(x))
    # max-pooling with 2x2 grid
    x = self.max_pool_1(x)
    # convolve, then perform ReLU non-linearity
    x = torch.nn.functional.relu(self.conv2(x))
    # max-pooling with 2x2 grid
    x = self.max_pool_2(x)
    # first flatten 'max_pool_2_out' to contain 16*5*5 columns
    # read through https://stackoverflow.com/a/42482819/7551231
    x = x.view(-1, 16*5*5)
    # FC-1, then perform ReLU non-linearity
    x = torch.nn.functional.relu(self.fc1(x))
    # FC-2, then perform ReLU non-linearity
    x = torch.nn.functional.relu(self.fc2(x))
    # FC-3
    x = self.fc3(x)

    return x
```

LeNet in Pytorch

Let’s walk through an example using PyTorch

https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

AlexNet

11
11
224
11
11
224
Stride of 4
3
5
5
48
55
5
5
55
48
Max pooling
3
3
3
3
27
128
3
3
27
3
3
128
Max pooling
3
3
192
13
13
3
3
192
3
3
192
13
13
3
3
192
128
13
13
128
Max pooling
2048
2048
dense
dense
dense
2048
2048
1000

Deng et al. 2009

AlexNet

AlexNet

- AlexNet won ImageNet competition in 2012

AlexNet

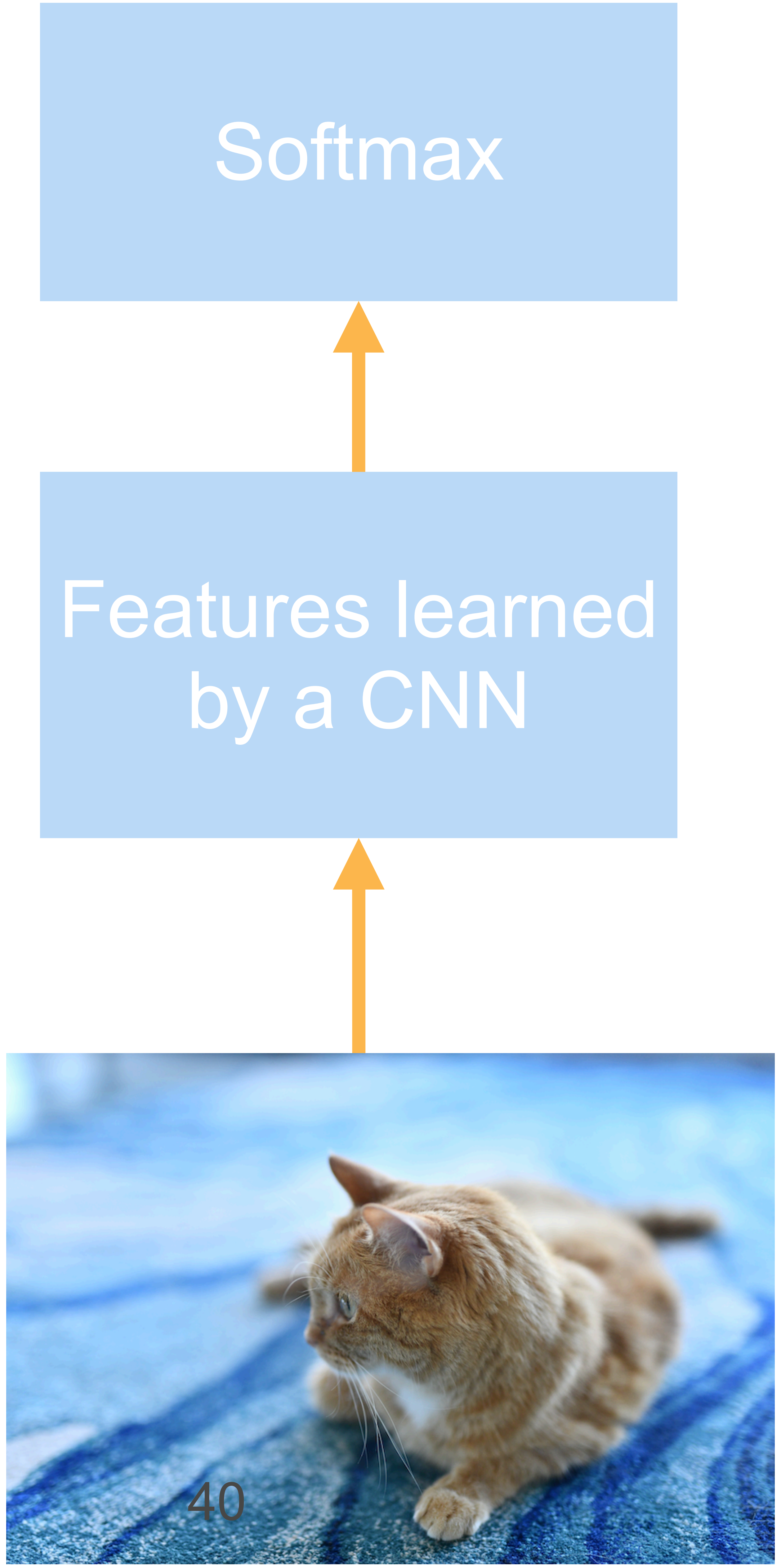
- AlexNet won ImageNet competition in 2012
- Deeper and bigger LeNet

AlexNet

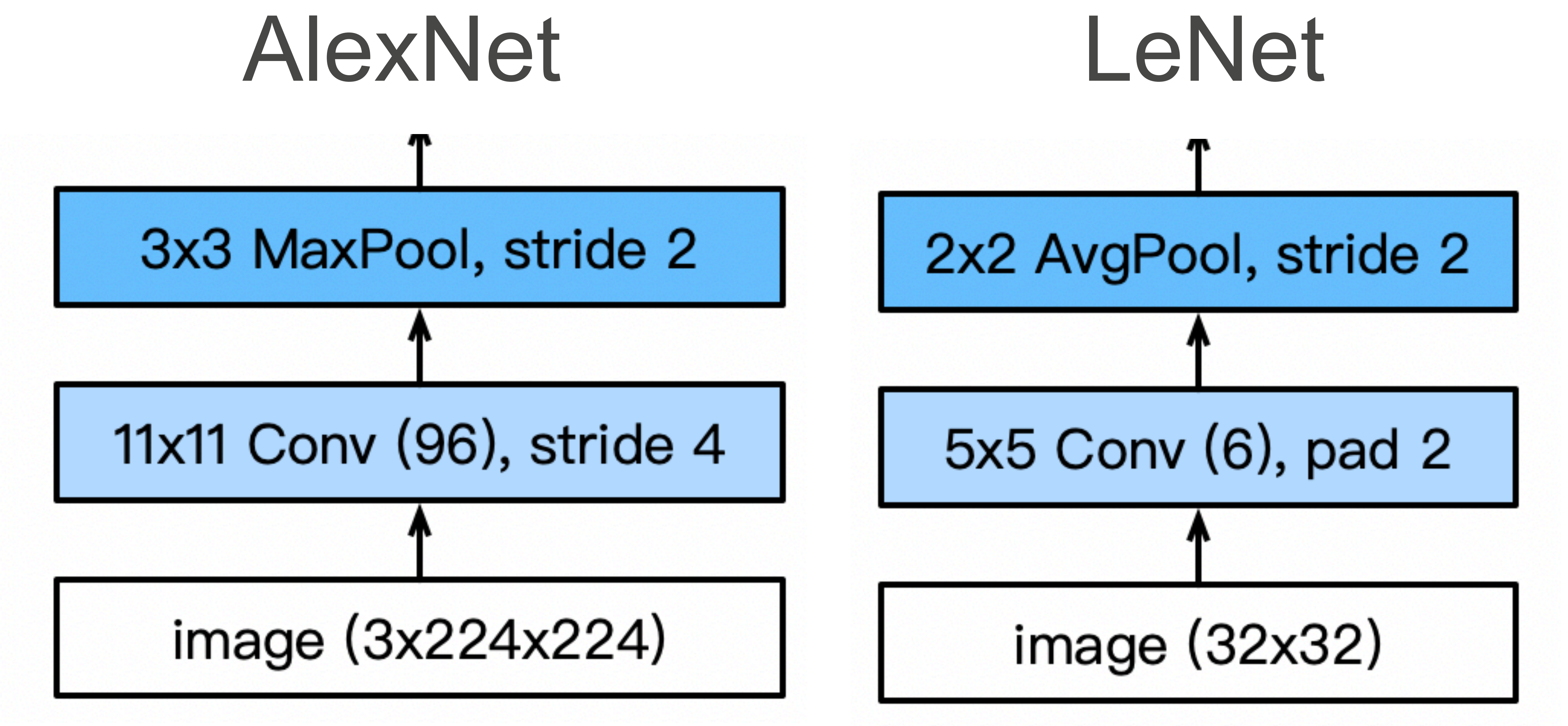
- AlexNet won ImageNet competition in 2012
- Deeper and bigger LeNet
- Paradigm shift for computer vision

AlexNet

- AlexNet won ImageNet competition in 2012
- Deeper and bigger LeNet
- Paradigm shift for computer vision

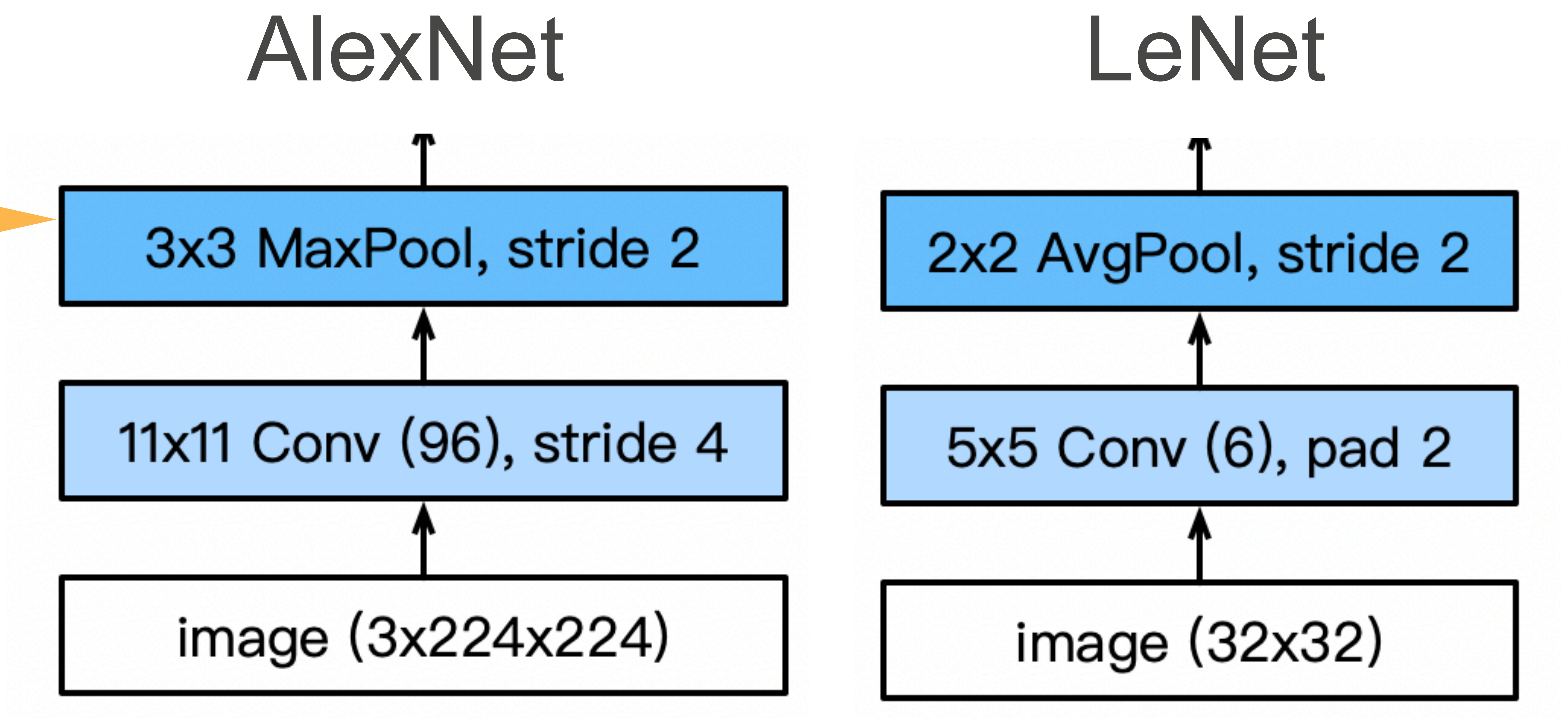


AlexNet Architecture



AlexNet Architecture

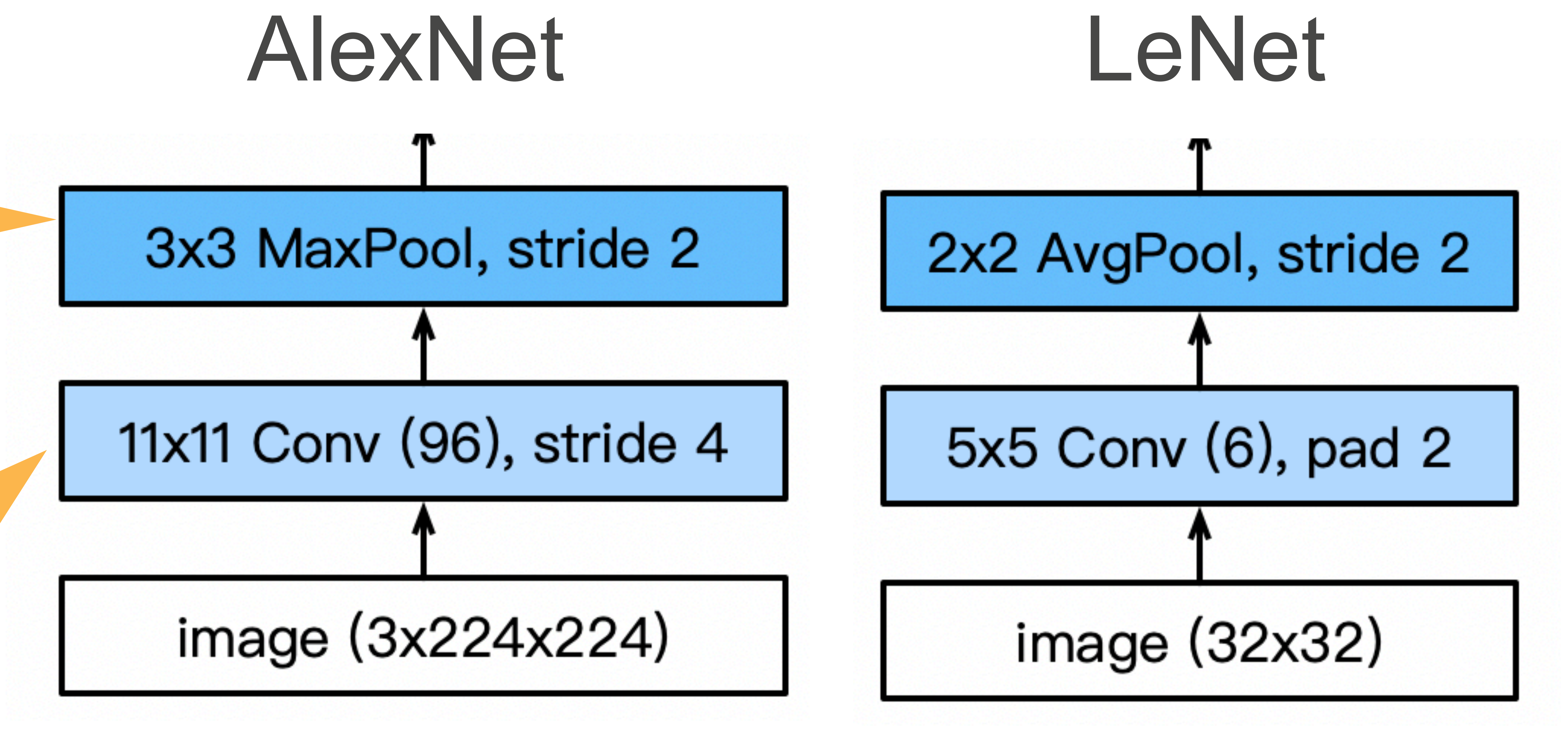
Larger pool size



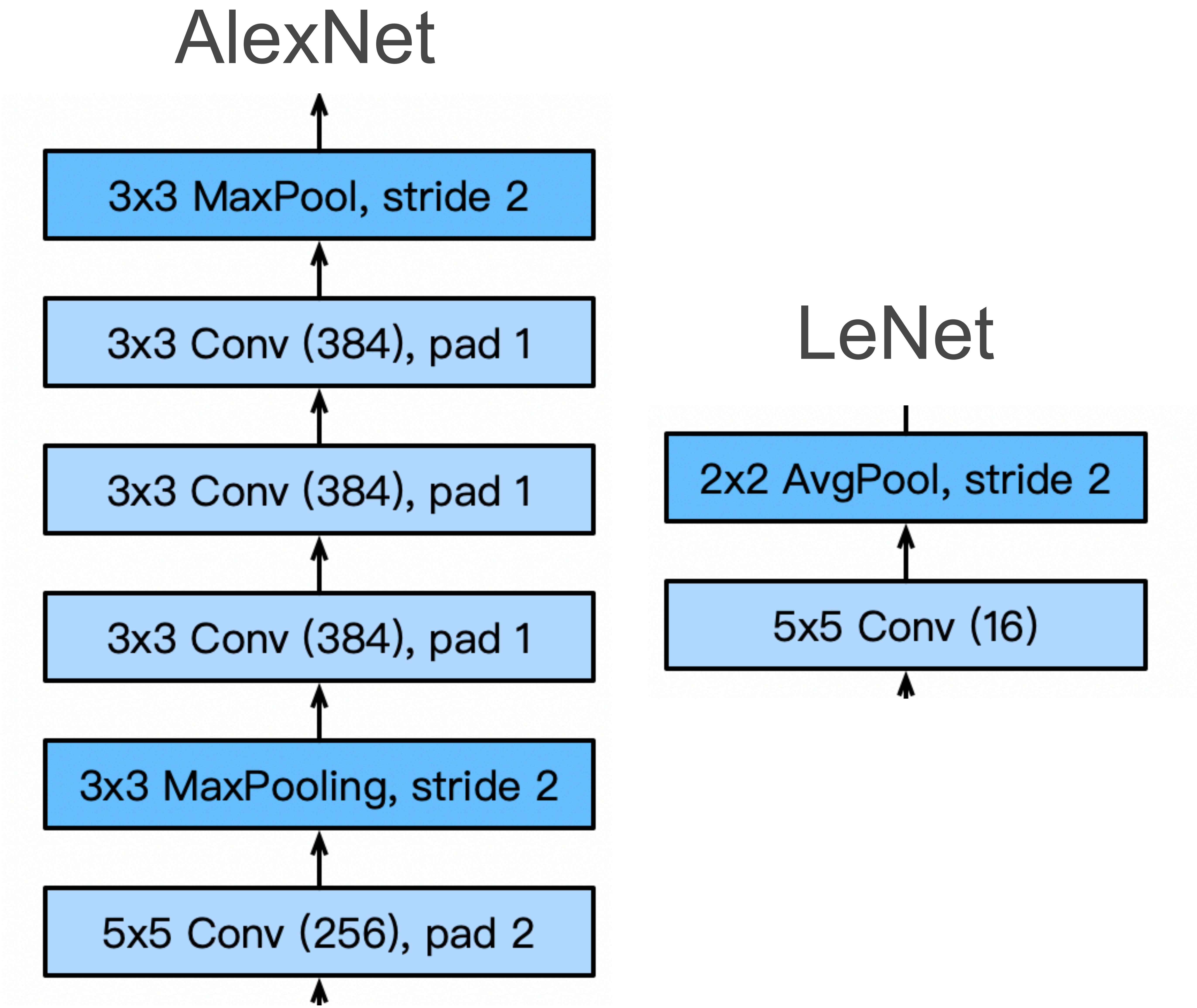
AlexNet Architecture

Larger pool size

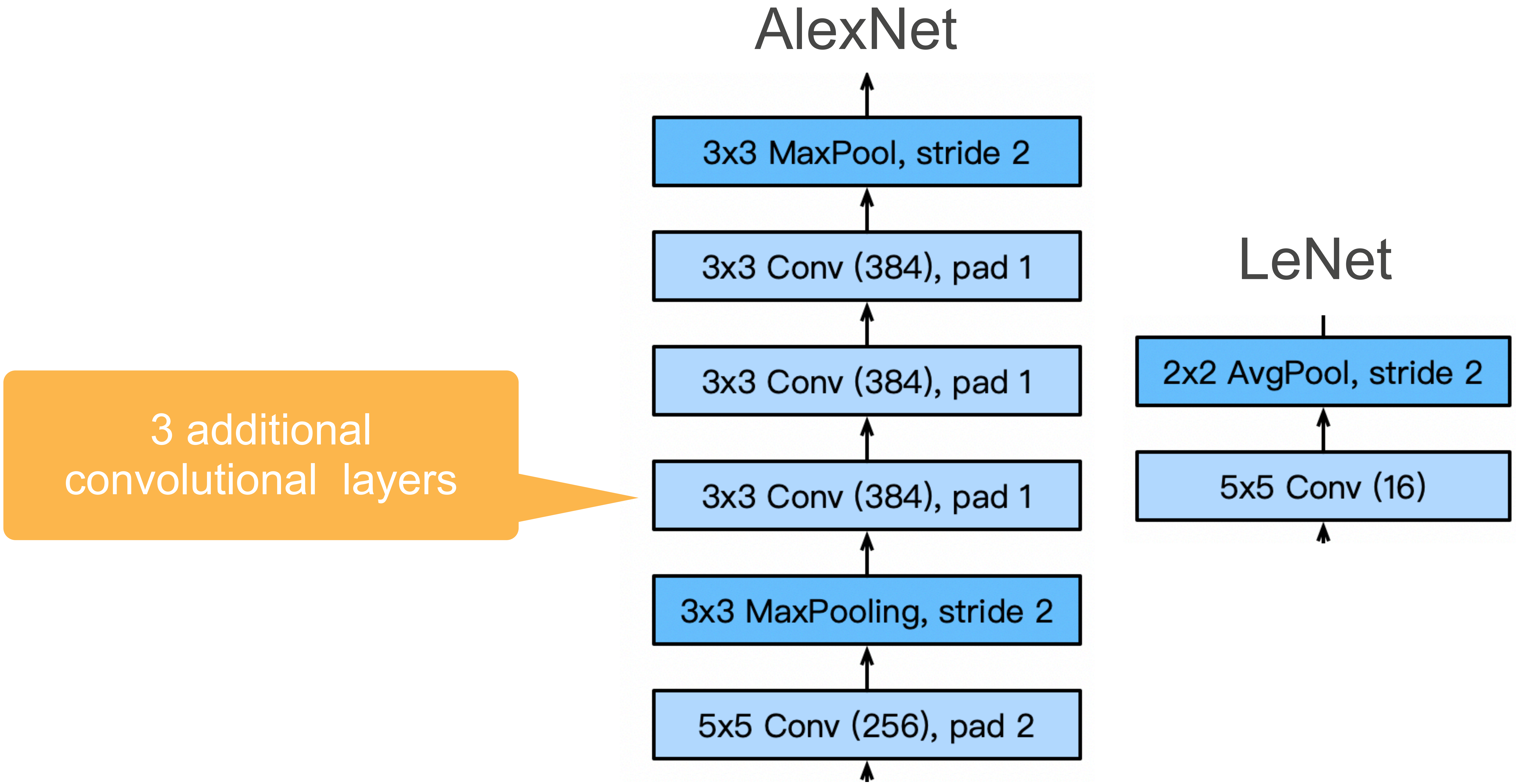
Larger kernel size, stride because of the increased image size, and more output channels.



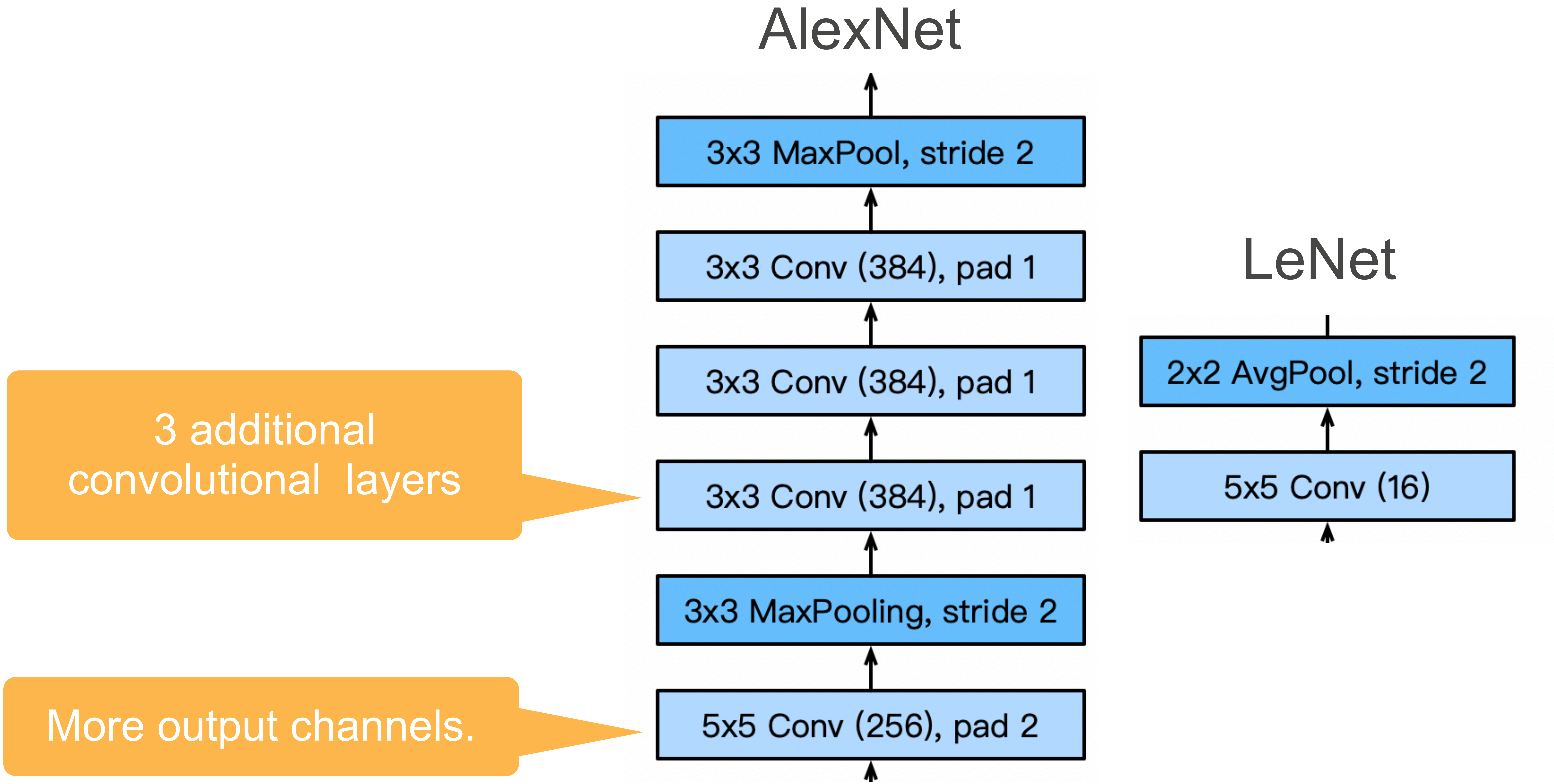
AlexNet Architecture



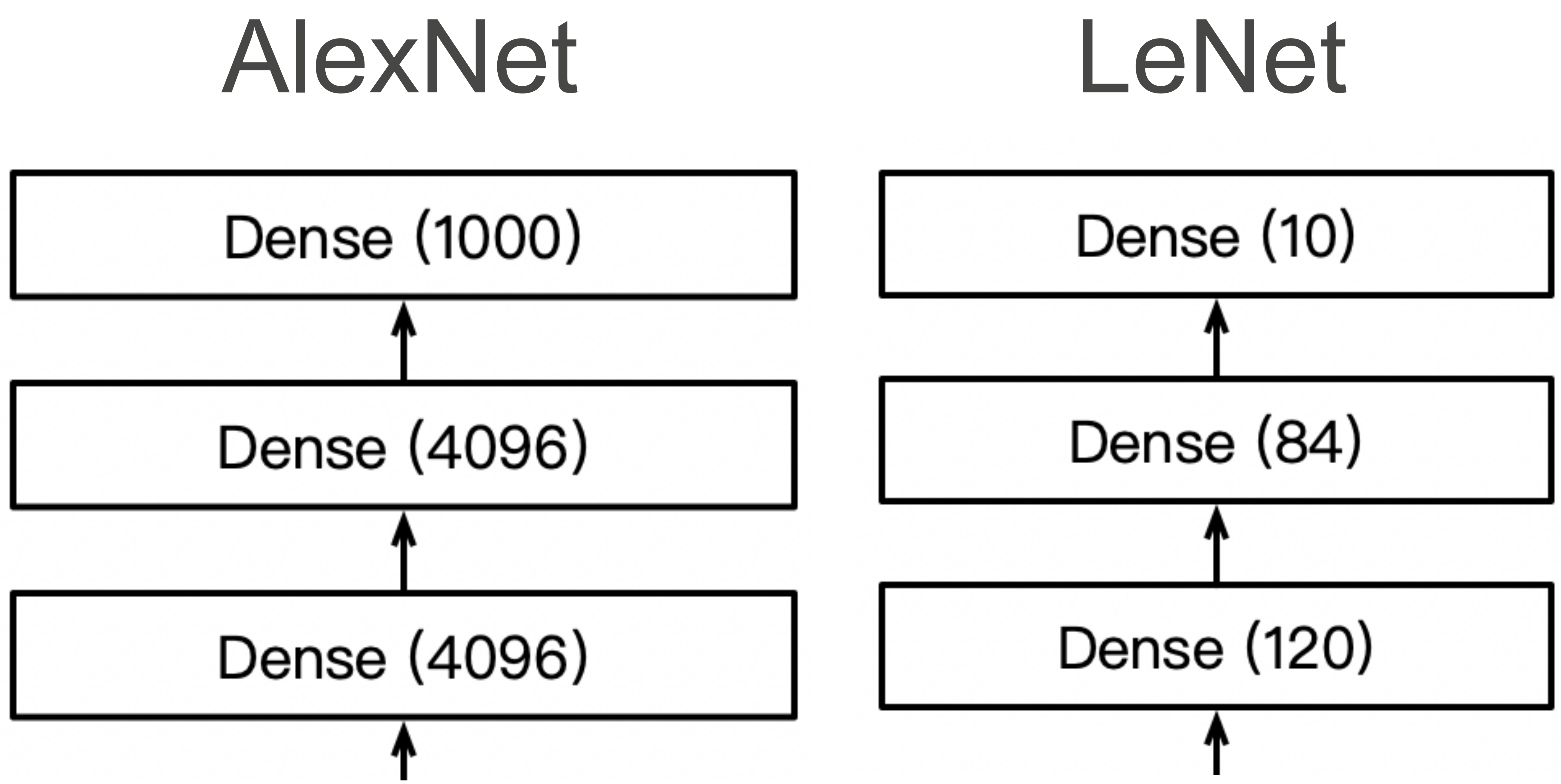
AlexNet Architecture



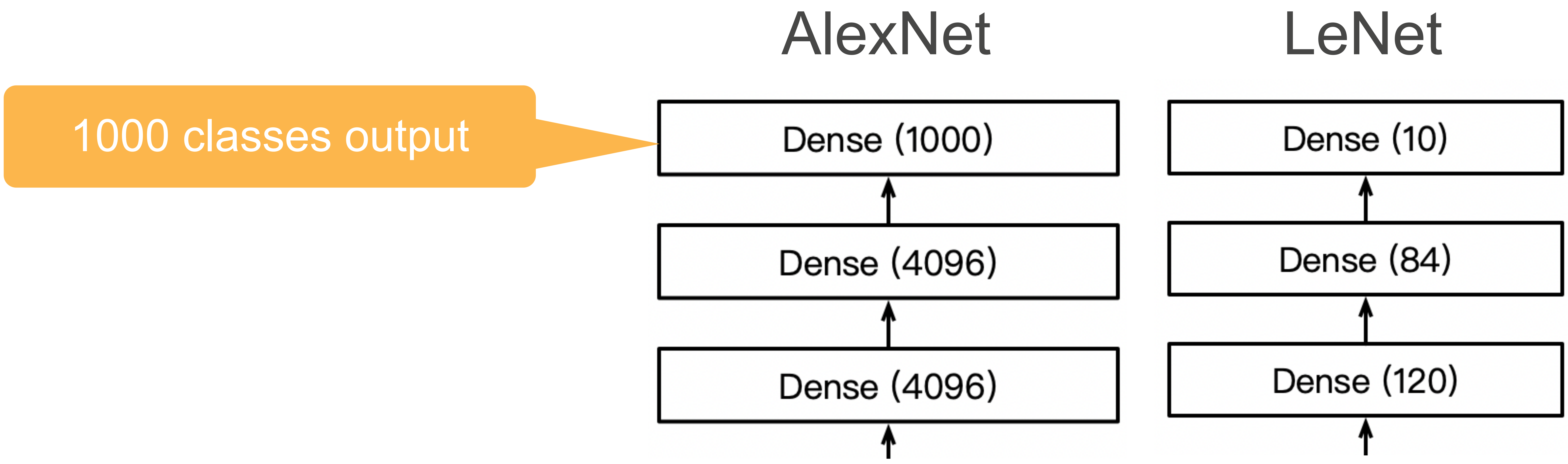
AlexNet Architecture



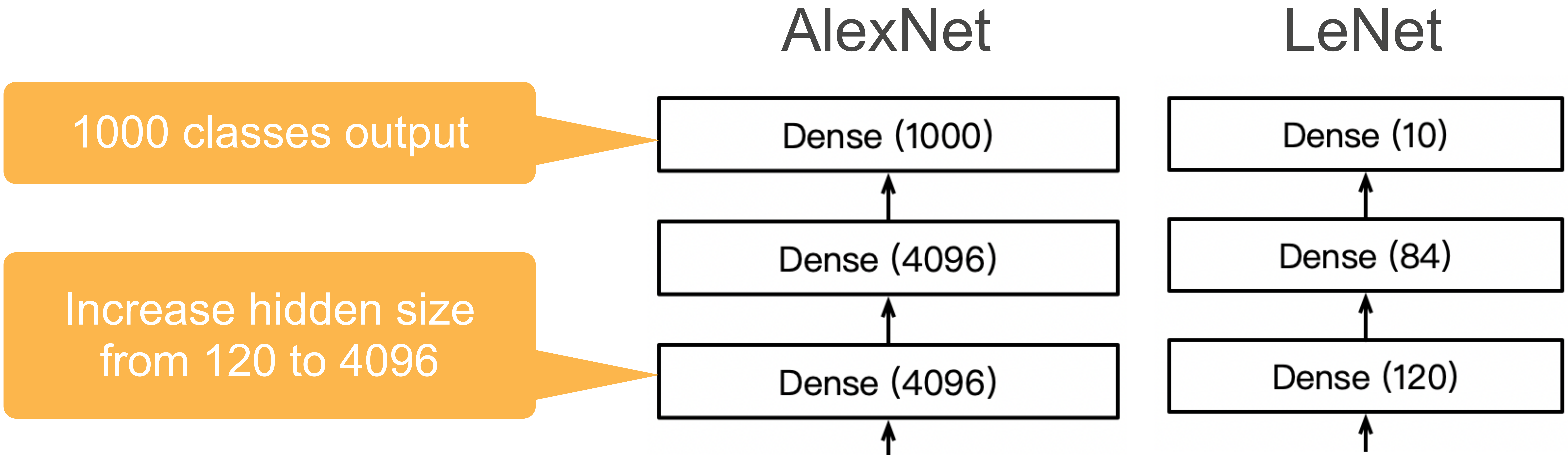
AlexNet Architecture



AlexNet Architecture

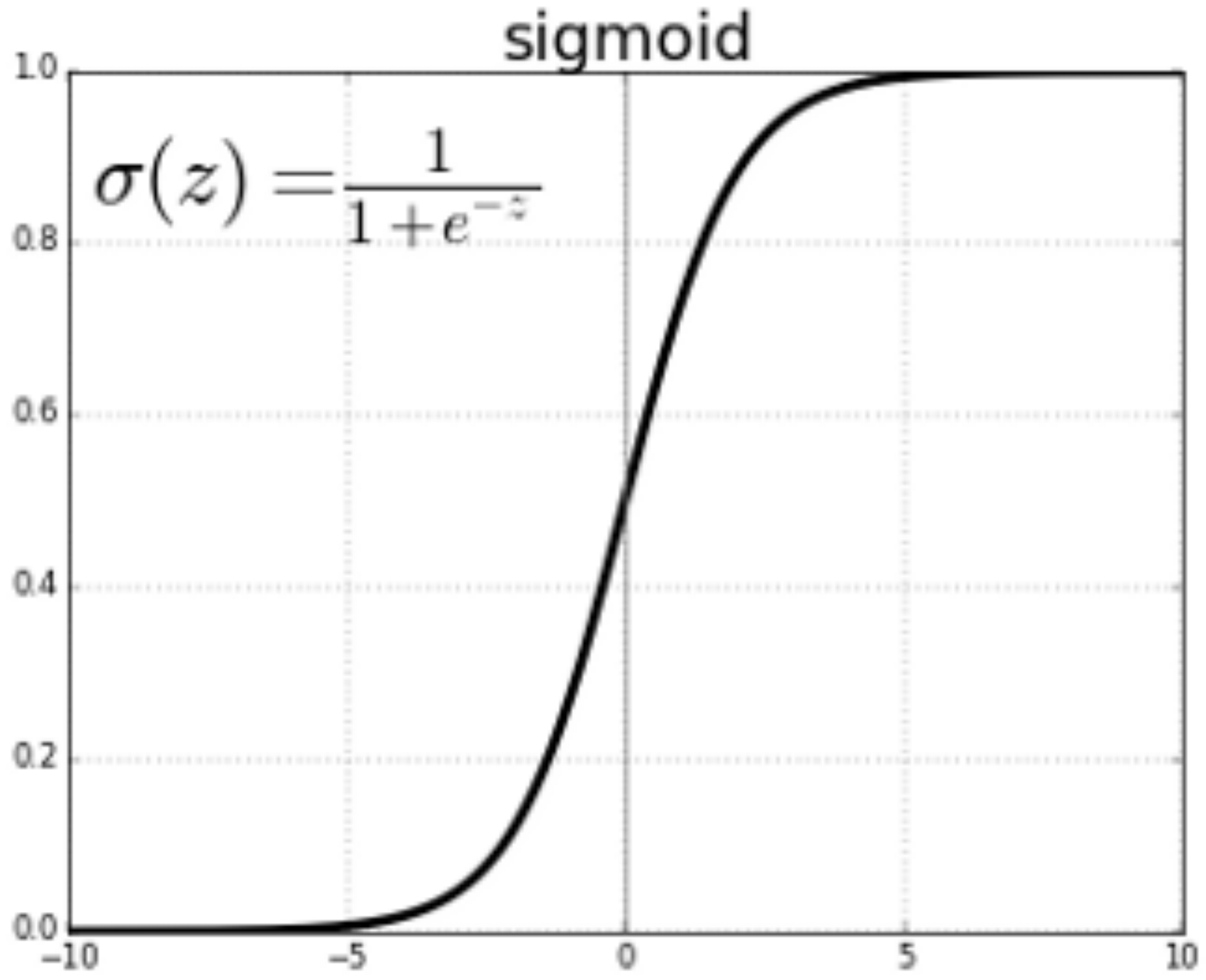


AlexNet Architecture



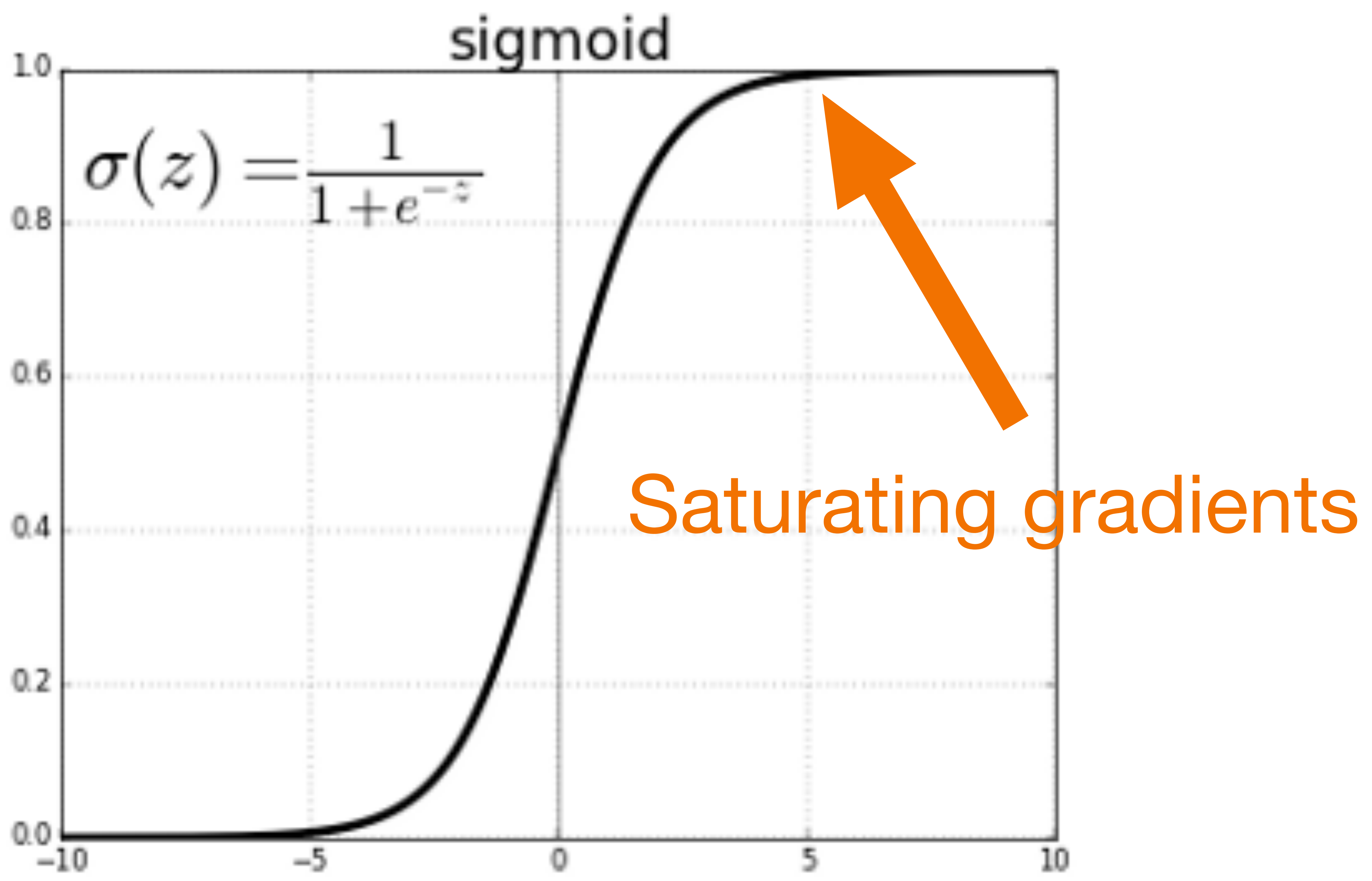
More Differences…

- Change activation function from sigmoid to ReLu (no more vanishing gradient)



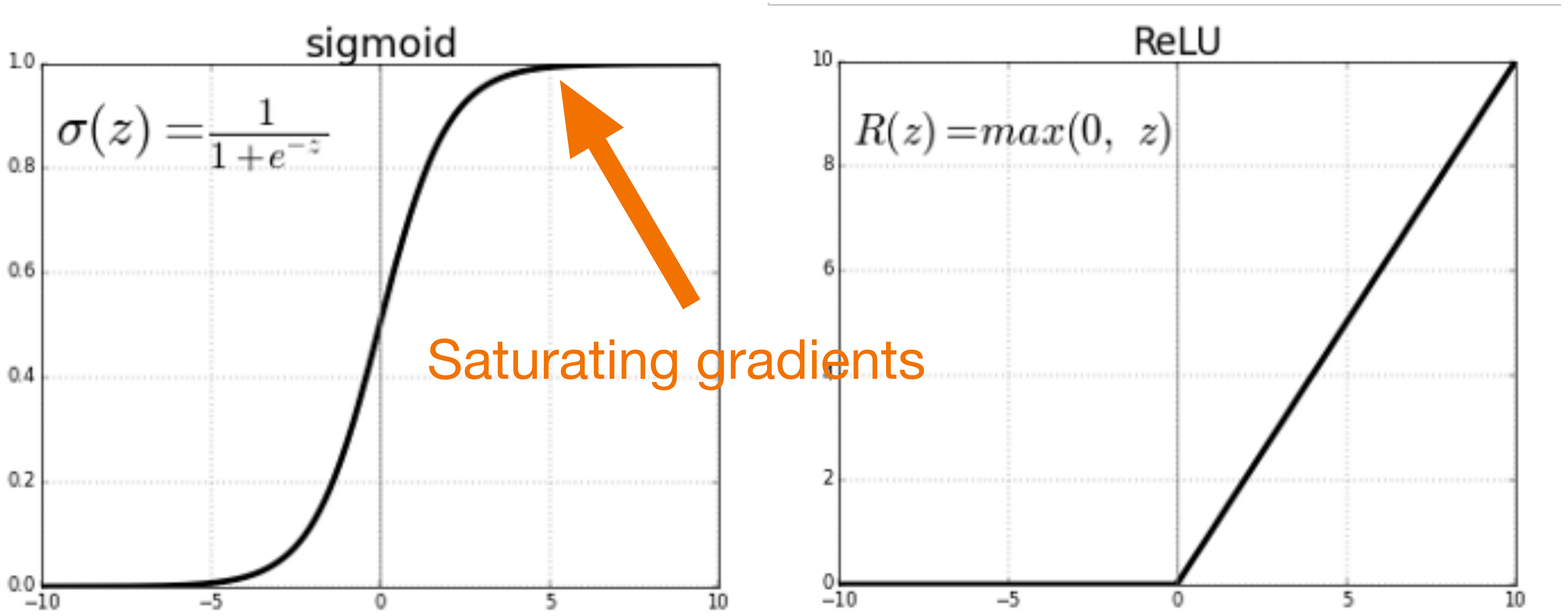
More Differences…

- Change activation function from sigmoid to ReLu (no more vanishing gradient)



More Differences…

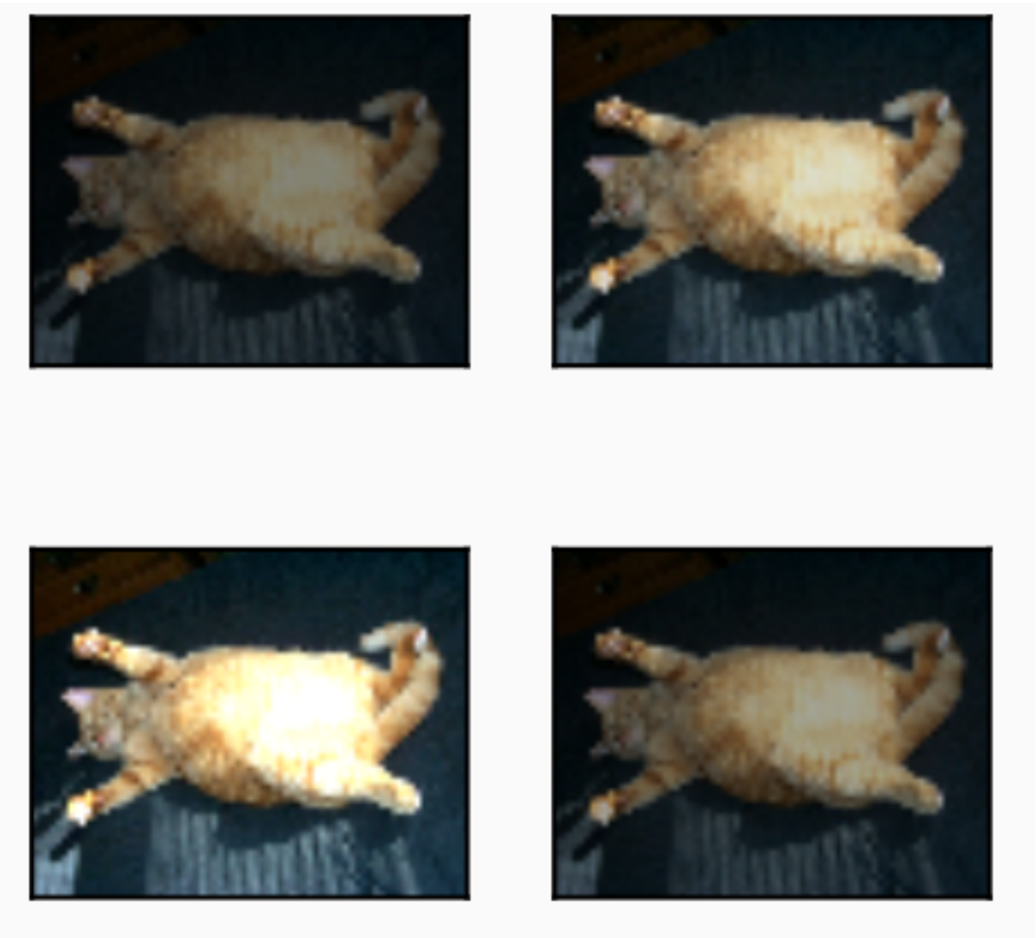
- Change activation function from sigmoid to ReLu (no more vanishing gradient)



More Differences…

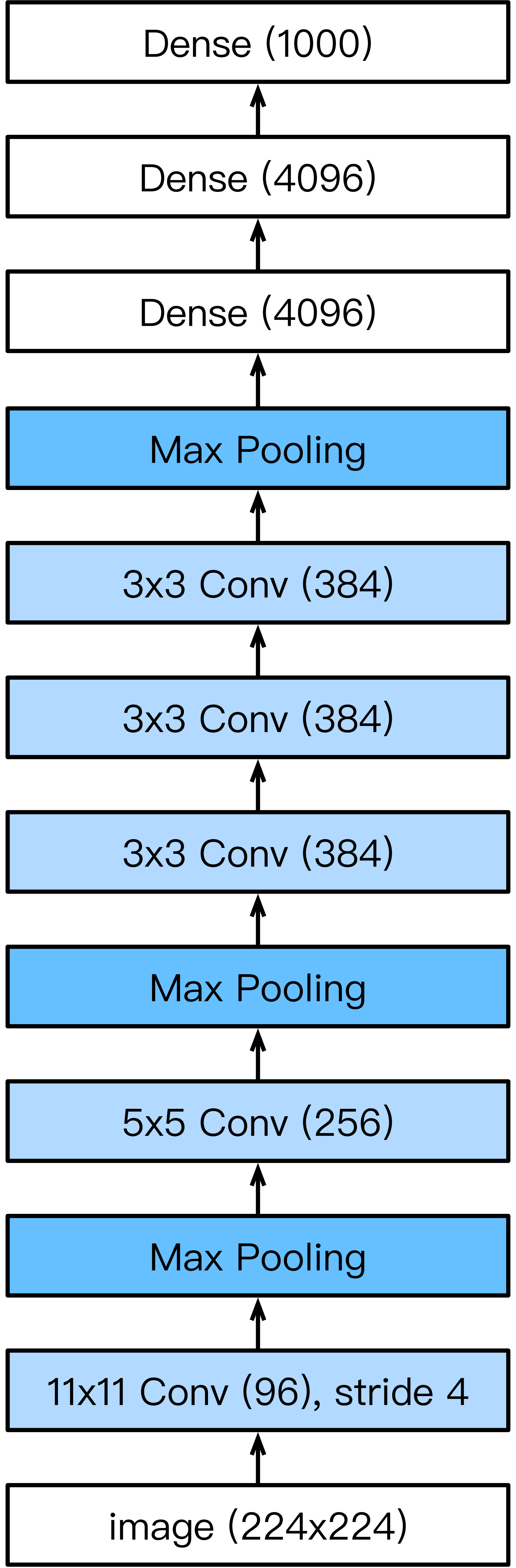
- Change activation function from sigmoid to ReLu (no more vanishing gradient)
- Data augmentation





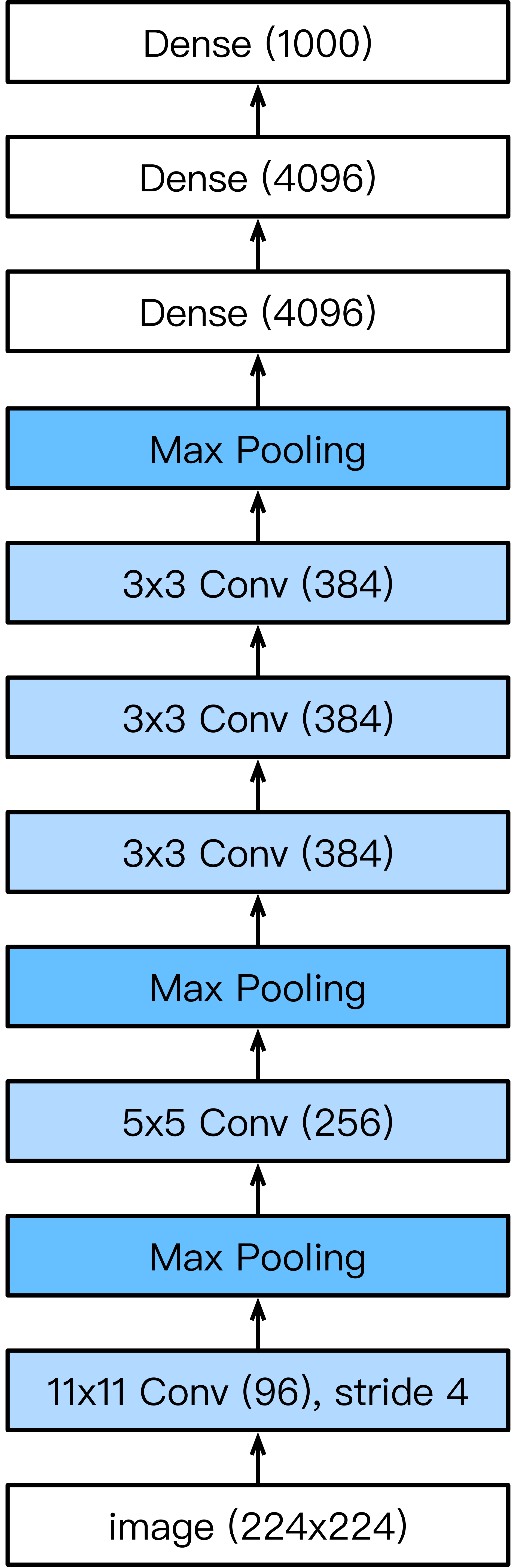

Complexity

	#parameters	
	AlexNet	LeNet
Conv1	35K	150
Conv2	614K	2.4K
Conv3-5	3M	
Dense1	26M	0.048M
Dense2	16M	0.01M
Total	46M	0.06M



Complexity

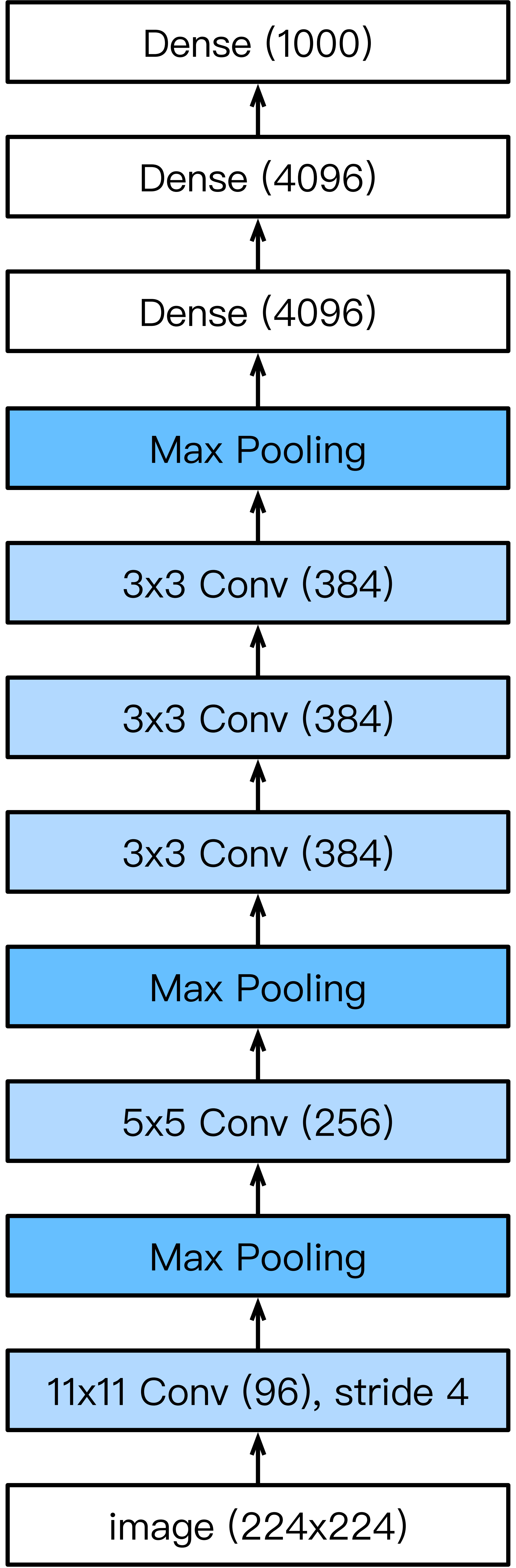
	#parameters	
	AlexNet	LeNet
Conv1	35K	150
Conv2	614K	2.4K
Conv3-5	3M	
Dense1	26M	0.048M
Dense2	16M	0.01M
Total	46M	0.06M

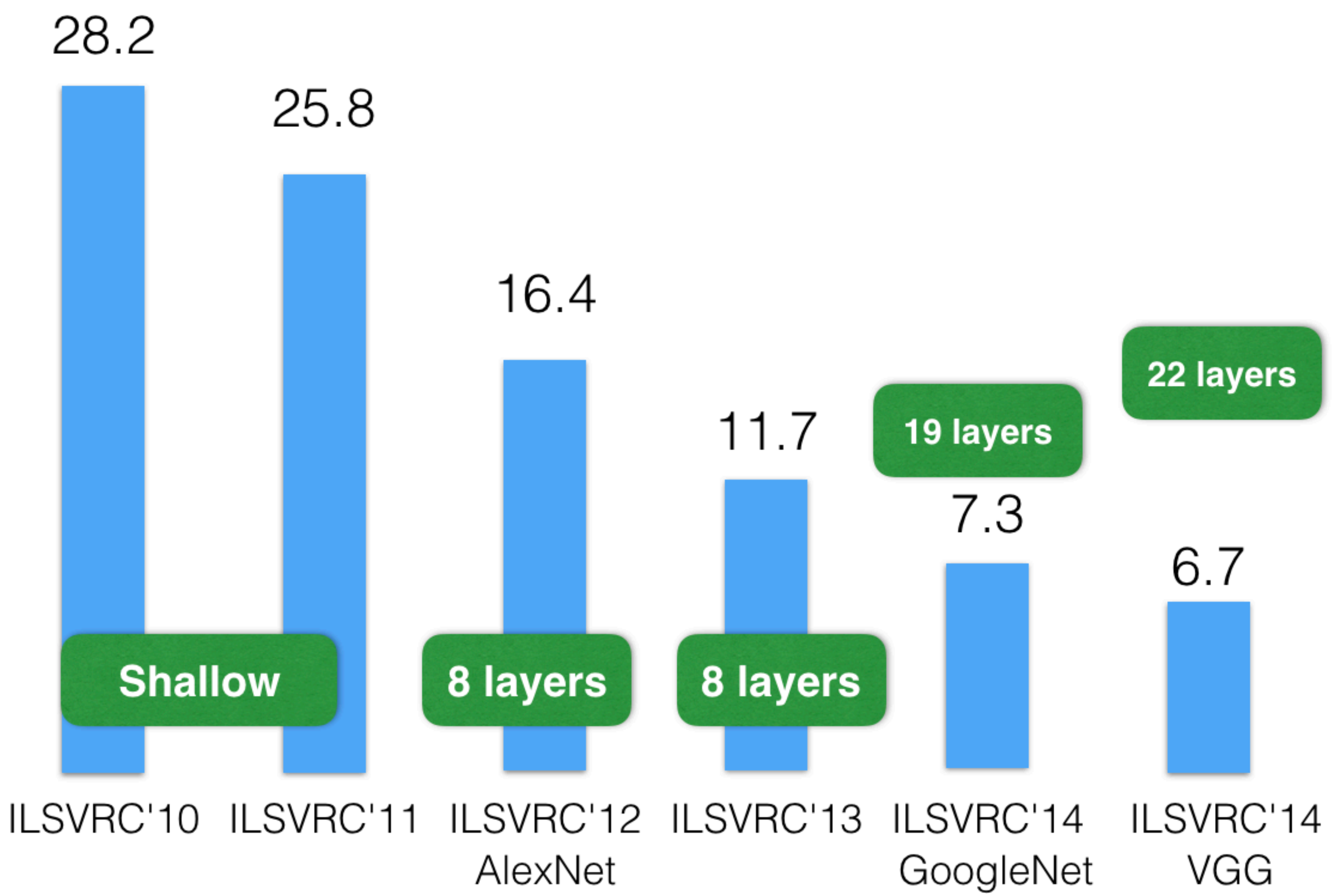


Complexity

	#parameters	
	AlexNet	LeNet
Conv1	35K	150
Conv2	614K	2.4K
Conv3-5	3M	
Dense1	26M	0.048M
Dense2	16M	0.01M
Total	46M	0.06M

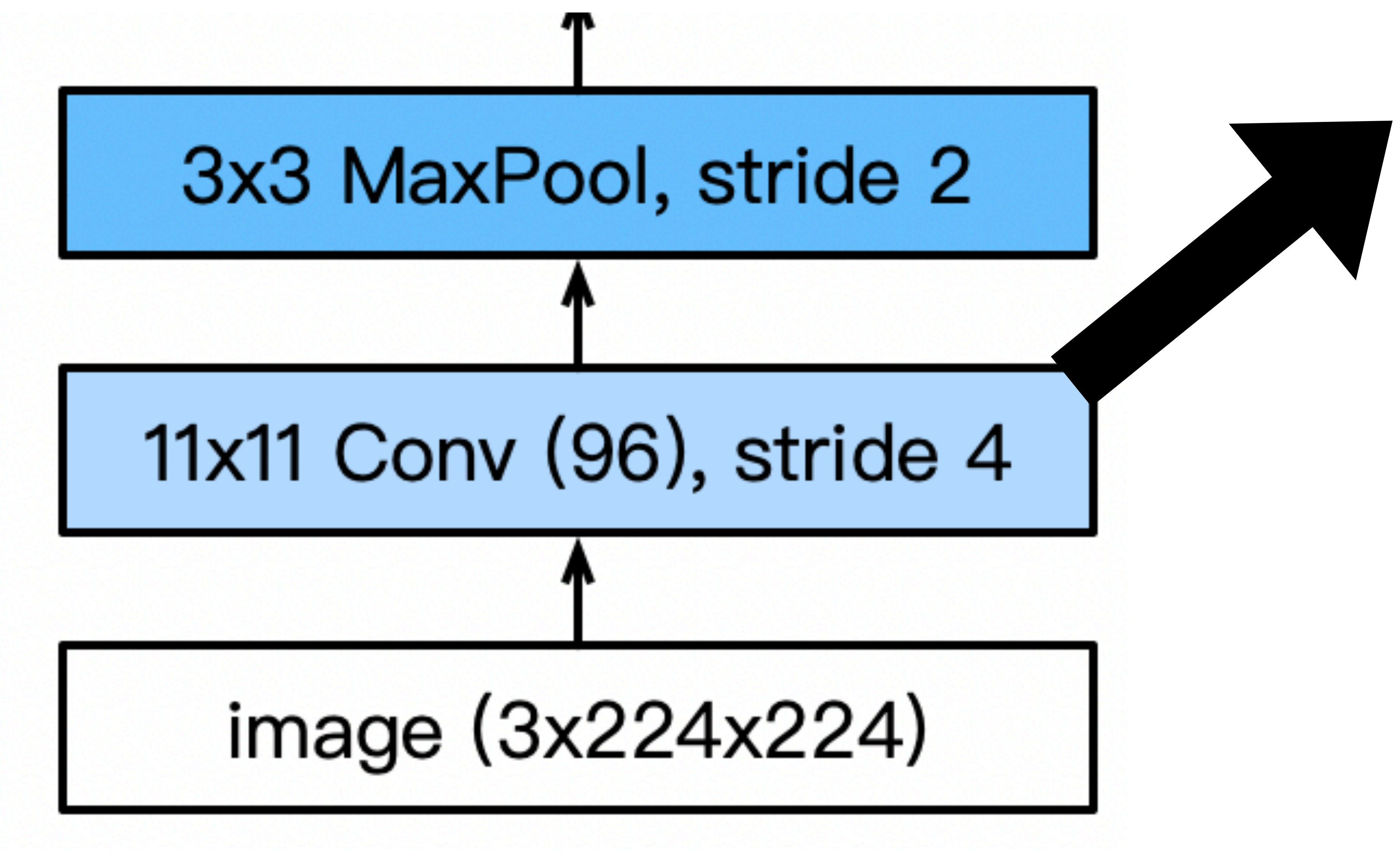
11x11x3x96=35k



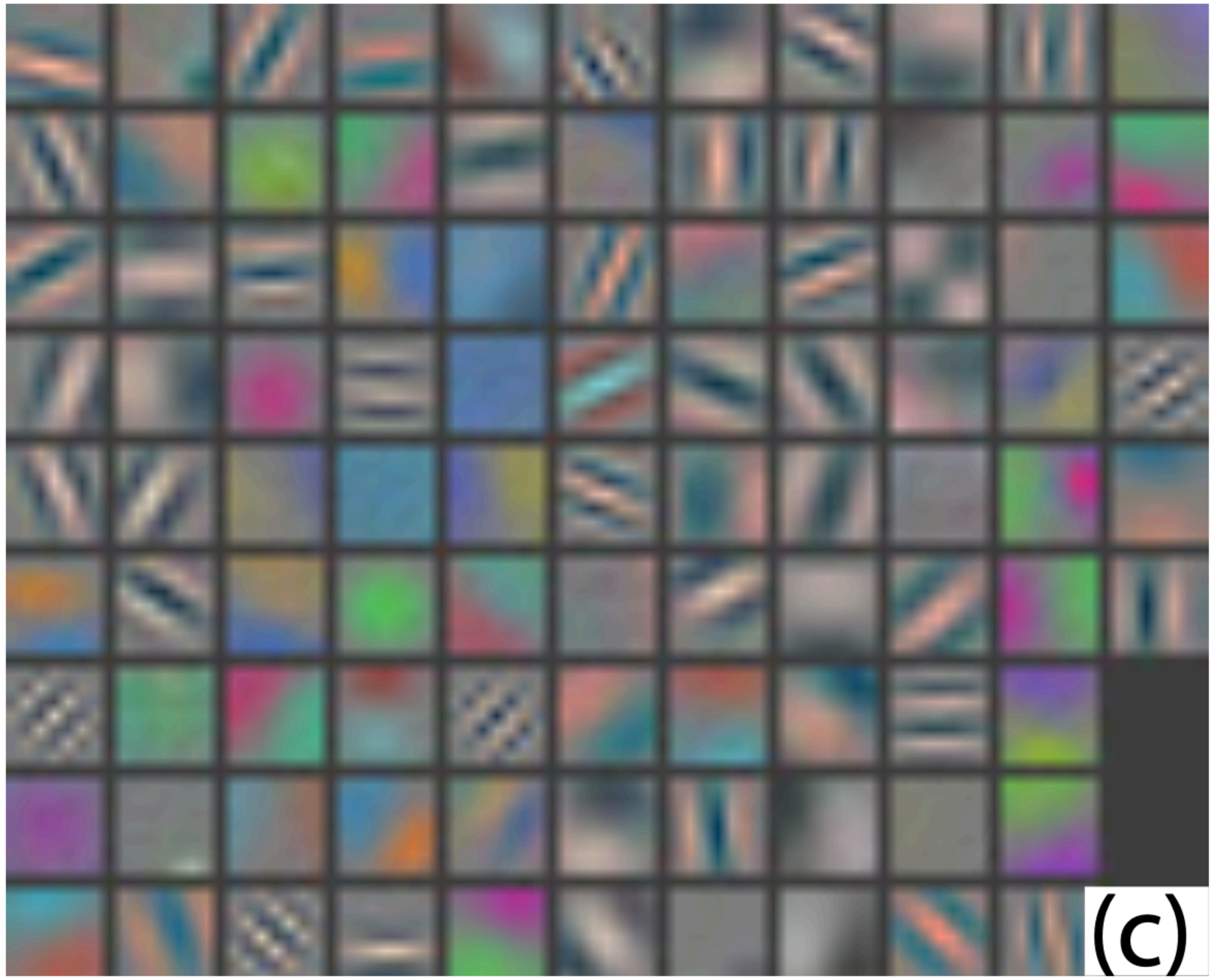


ImageNet Top-5 Classification Accuracy (%)

AlexNet



Each Conv1 kernel is 3x11x11, can be visualized as an RGB patch:



[Visualizing and Understanding Convolutional Networks. M Zeiler & R Fergus 2013]

Which of the following are true about AlexNet? Select all that apply.

A. AlexNet contains 8 conv/fc layers. The first five are convolutional layers.

B. The last three layers are fully connected layers.

C. some of the convolutional layers are followed by max-pooling (layers).

D. AlexNet achieved excellent performance in the 2012 ImageNet challenge.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097–1105).

Which of the following are true about AlexNet? Select all that apply.

A. AlexNet contains 8 conv/fc layers. The first five are convolutional layers.

B.The last three layers are fully connected layers.

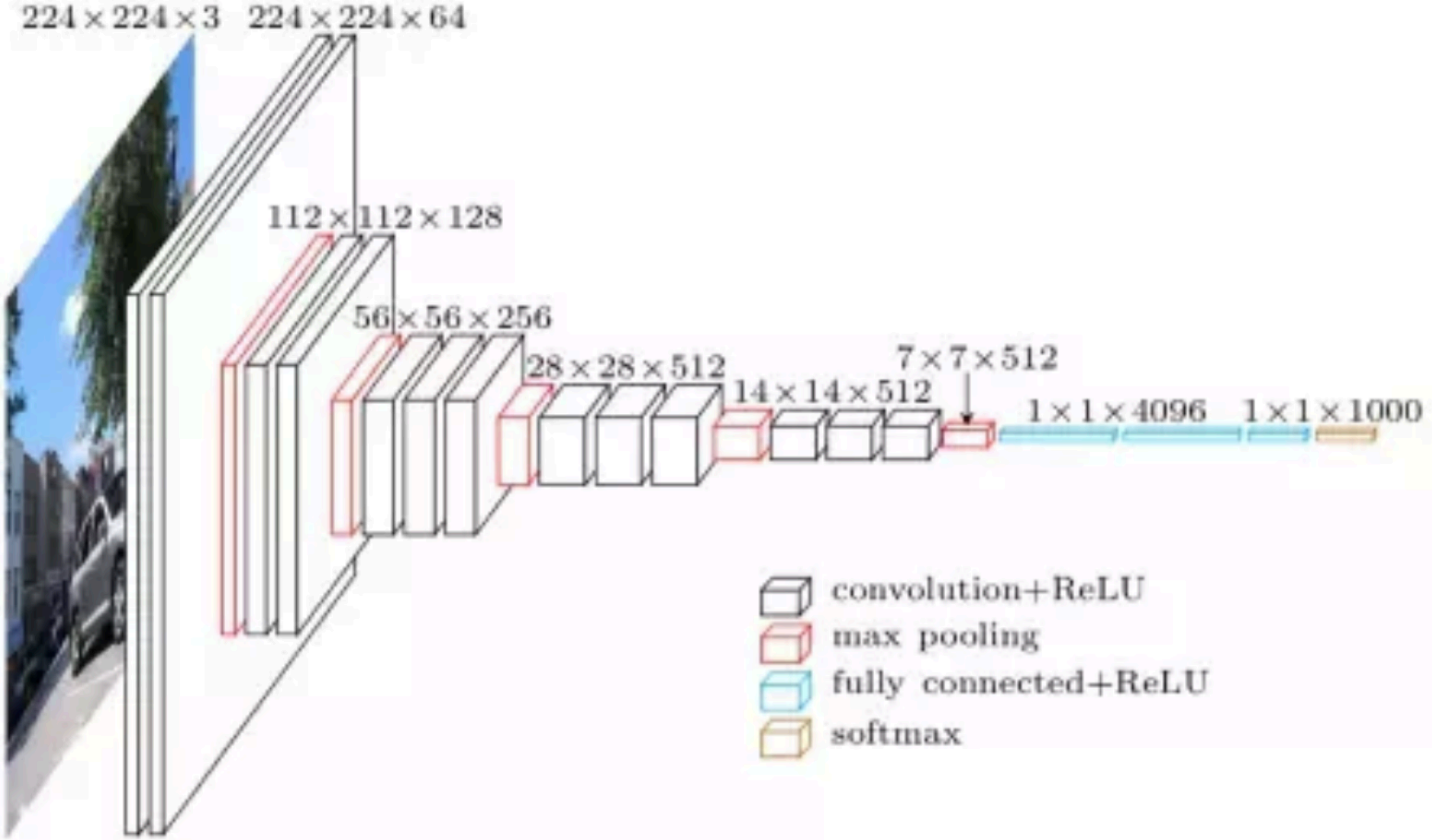
C.some of the convolutional layers are followed by max-pooling (layers).

D. AlexNet achieved excellent performance in the 2012 ImageNet challenge.

All options are true!

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097–1105).

VGG



VGG Block: Multiple convolution layers followed by pooling.

Progress

- LeNet (1995)
  - 2 convolution + pooling layers
  - 2 hidden dense layers
- AlexNet
  - Bigger and deeper LeNet
  - ReLu, preprocessing
- VGG
  - Bigger and deeper AlexNet (repeated VGG blocks)

Which of the following statement is True for the success of deep models?

- Better design of the neural networks
- Large scale training dataset
- Available computing power
- All of the above

Which of the following statement is True for the success of deep models?

- Better design of the neural networks
- Large scale training dataset
- Available computing power
- All of the above

Summary of today

Summary of today

- Reviewed (some of) convolutional computations.

Summary of today

- Reviewed (some of) convolutional computations.
    - 2D convolutions, multiple input channels, pooling.

Summary of today

- Reviewed (some of) convolutional computations.
    - 2D convolutions, multiple input channels, pooling.
- Shown how convolutions are used as layers in a (deep) neural network.

Summary of today

- Reviewed (some of) convolutional computations.
    - 2D convolutions, multiple input channels, pooling.
- Shown how convolutions are used as layers in a (deep) neural network.
- Built intuition for output of convolutional layers.

Summary of today

- Reviewed (some of) convolutional computations.
    - 2D convolutions, multiple input channels, pooling.
- Shown how convolutions are used as layers in a (deep) neural network.
- Built intuition for output of convolutional layers.
- Overviewed the evolution of deeper convolutional networks

Other Deep Architectures

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.
  - Main use is for processing images.

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.
  - Main use is for processing images.
  - Also can be useful for handling time series.

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.
  - Main use is for processing images.
  - Also can be useful for handling time series.
- Other common architectures:

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.
    - Main use is for processing images.
    - Also can be useful for handling time series.
- Other common architectures:
    - Recurrent neural networks: hidden activations are a function of input and activations from previous inputs. Designed for sequential data such as text.

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.
  - Main use is for processing images.
  - Also can be useful for handling time series.
- Other common architectures:
  - Recurrent neural networks: hidden activations are a function of input and activations from previous inputs. Designed for sequential data such as text.
  - Graph neural networks: take graph data as input.

Other Deep Architectures

- Convolutional neural networks are one of many special types of layers.
  - Main use is for processing images.
  - Also can be useful for handling time series.
- Other common architectures:
  - Recurrent neural networks: hidden activations are a function of input and activations from previous inputs. Designed for sequential data such as text.
  - Graph neural networks: take graph data as input.
  - Transformers: take sequences as input and learn what parts of input to pay attention to.

Acknowledgement:
Some of the slides in these lectures have been adapted/borrowed from materials developed by Yin Li (https://happyharrycn.github.io/CS540-Fall20/schedule/), Alex Smola and Mu Li:
https://courses.d2l.ai/berkeley-stat-157/index.html