Deep Learning (Convolutional Networks)

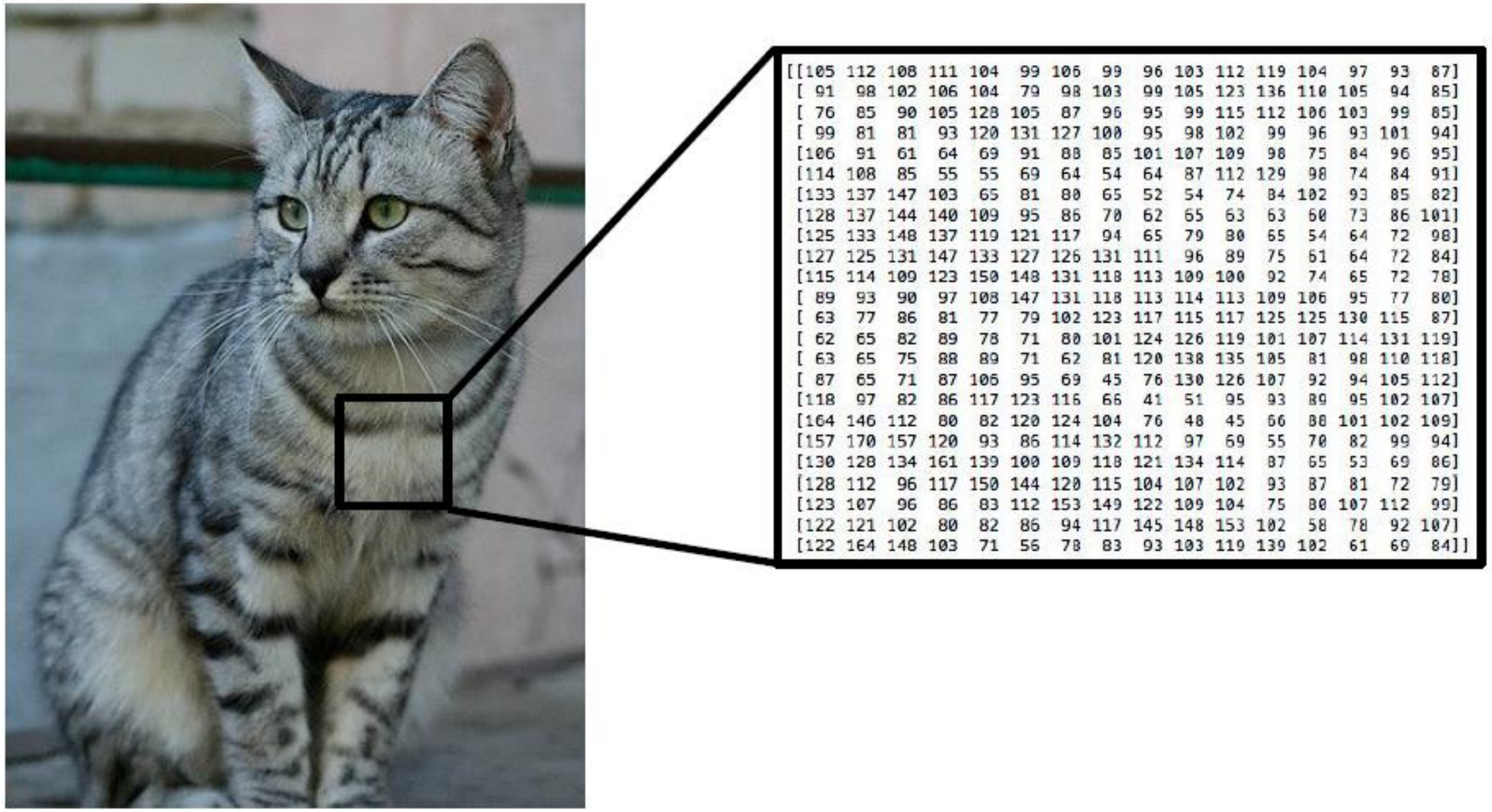
Sadegh Eskandari
Department of Computer Science, University of Guilan
eskandari@guilan.ac.ir

Today …

- Introduction
- Convolution Operation
- Pooling

Introduction

- **Convolutional neural networks (CNNs),** are a specialized kind of neural network for processing data that has a grid-like topology.



Gray image data can be thought of as a 2-D grid of pixels

Color image data can be thought of as a 3-D grid of pixels

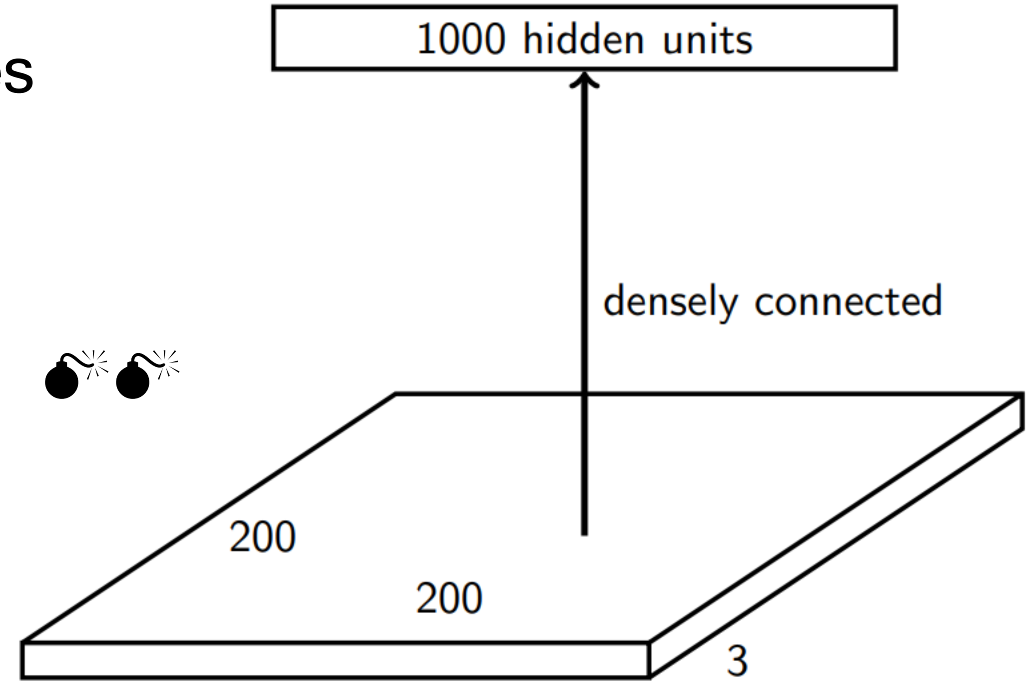
Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Introduction

- **Why we don't use fully connected networks?**

- Suppose we want to train a network that takes a 200×200 RGB (color) image as input.
    - Input size = $200 \times 200 \times 3 = 120\text{K}$
    - First layer parameters = $120\text{K} \times 1000 = 120 \text{ milion}$ 💣💣

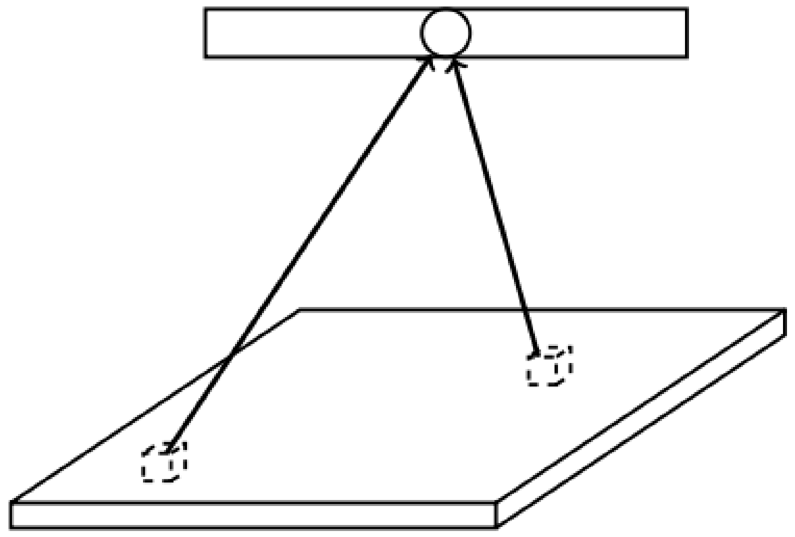
- What happens if the object in the image shifts a little?



Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

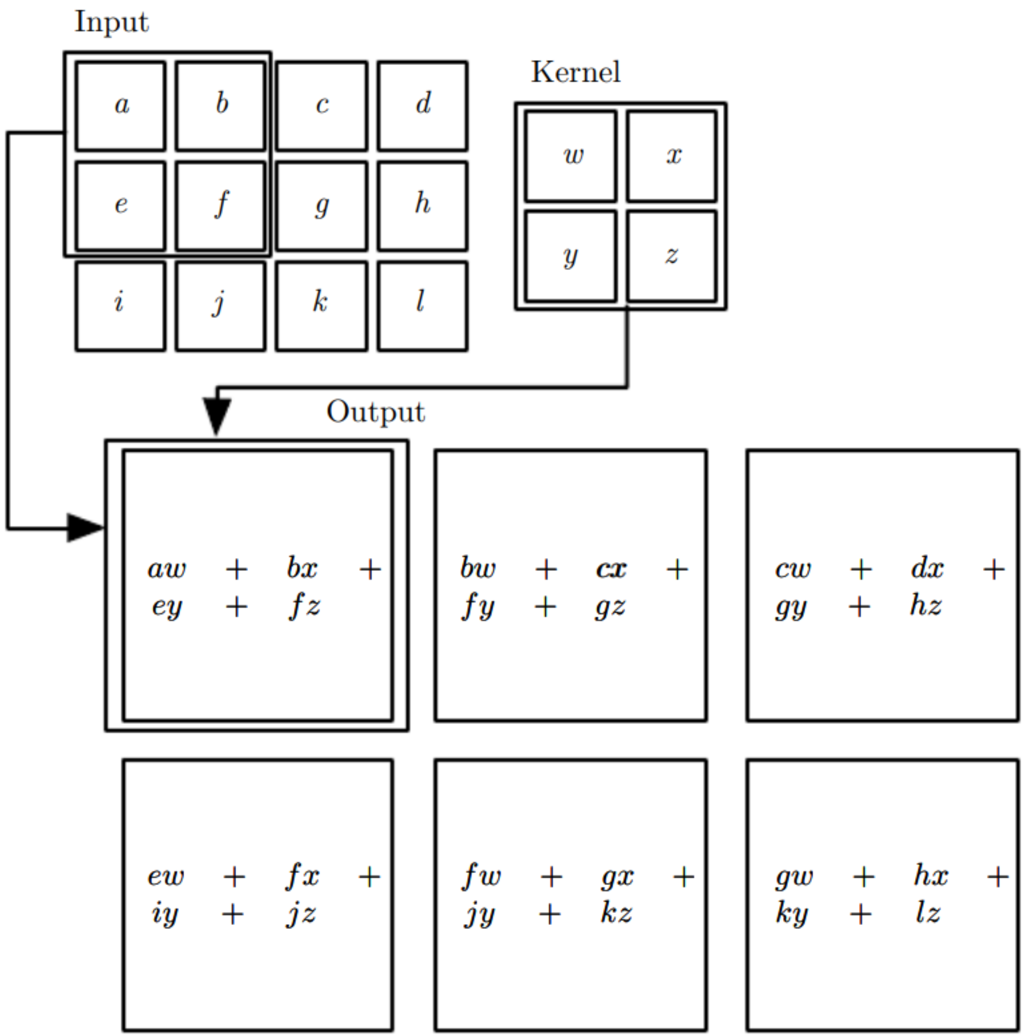
Introduction

- **Why we don’t use fully connected networks?**
- In the fully connected layer, each feature (hidden unit) looks at the entire image.
- The far away pixels will probably belong to completely different objects (or object sub-parts). Very little correlation.
- We want the incoming weights to focus on local patterns of the input image.



Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

Convolution (Cross-Correlation Operation)



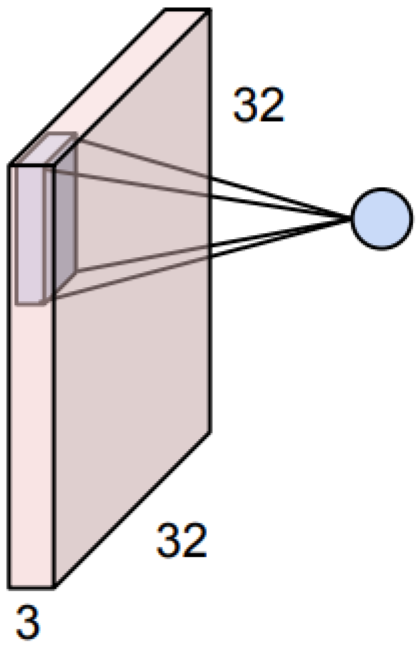
An example of 2-D convolution

$$S(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n)$$

Source: Goodfellow et al. (2016), Deep Learning

Convolution (Cross-Correlation Operation)

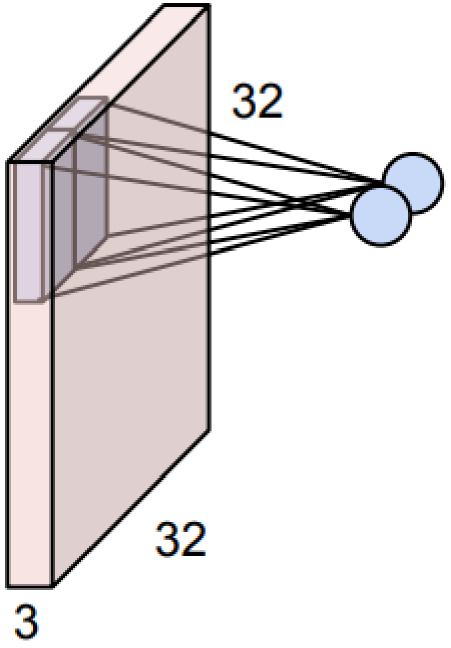
- **An example of a 3-D convolutional layer with a $5 \times 5 \times 3$ kernel (filter)**



Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

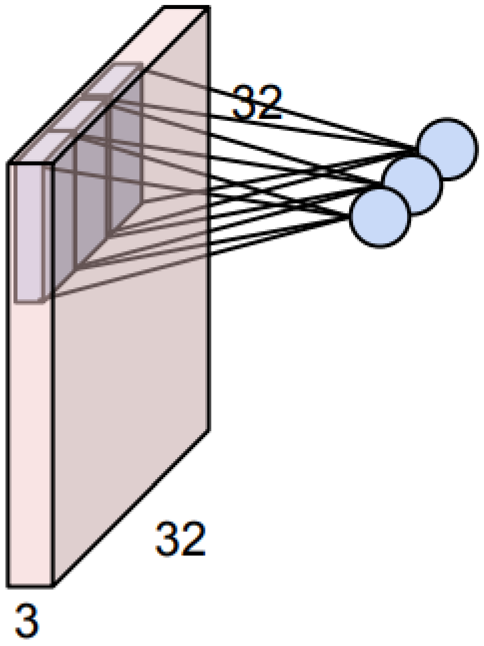
- **An example of a 3-D convolutional layer with a $5 \times 5 \times 3$ kernel (filter)**



Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

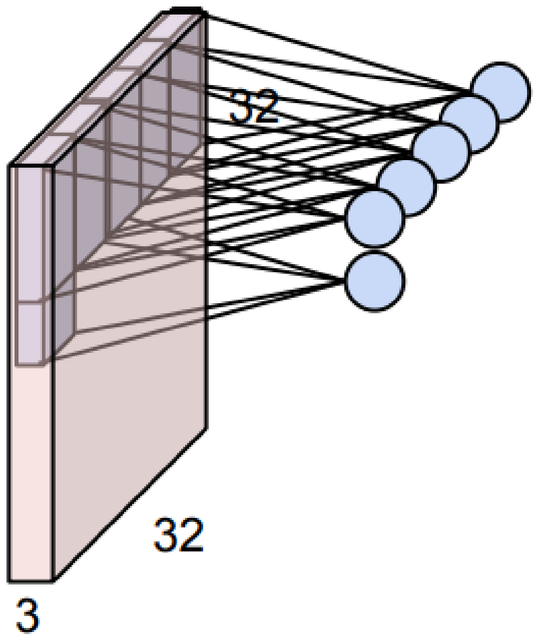
- **An example of a 3-D convolutional layer with a $5 \times 5 \times 3$ kernel (filter)**



Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

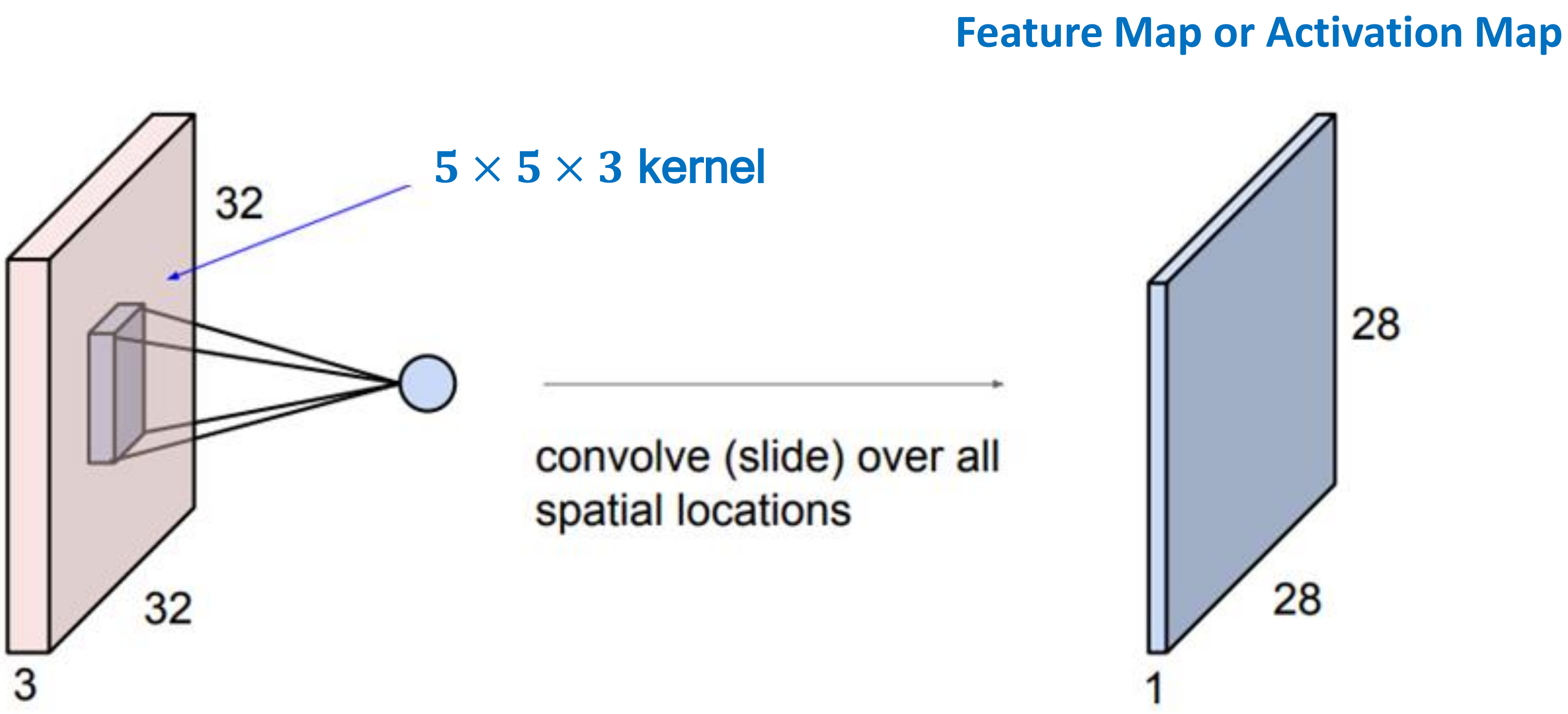
- **An example of a 3-D convolutional layer with a $5 \times 5 \times 3$ kernel (filter)**



Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

- **An example of a 3-D convolutional layer with a $5 \times 5 \times 3$ kernel (filter)**



Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

- **Example**

$$* \begin{bmatrix} 0 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

Convolution (Cross-Correlation Operation)

- **Example**

$*$

0	-1	0
-1	8	-1
0	-1	0

Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

Convolution (Cross-Correlation Operation)

- **Example**

$*$

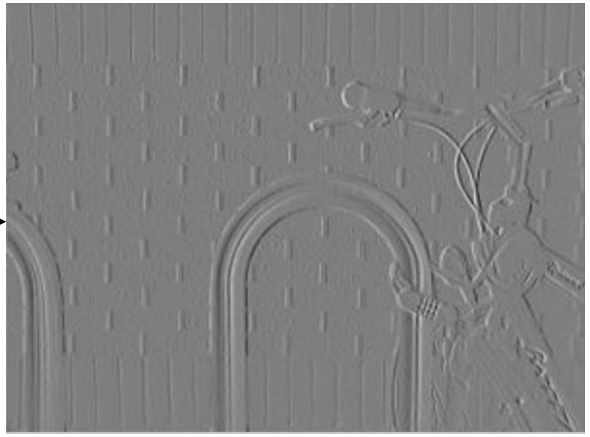
0	-1	0
-1	4	-1
0	-1	0

Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

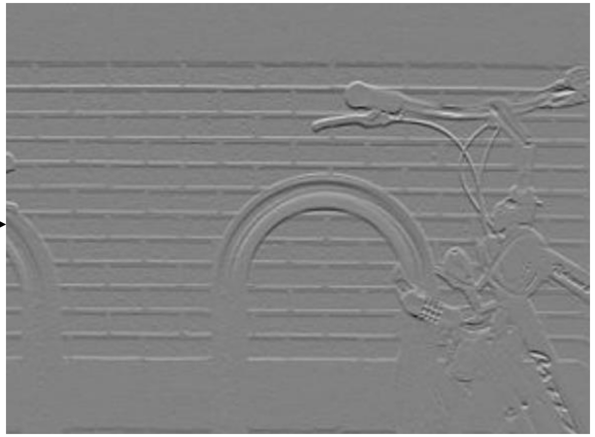
Convolution (Cross-Correlation Operation)

- **Example** (Sobel Edge Detection)

1	0	-1
2	0	-2
1	0	-1



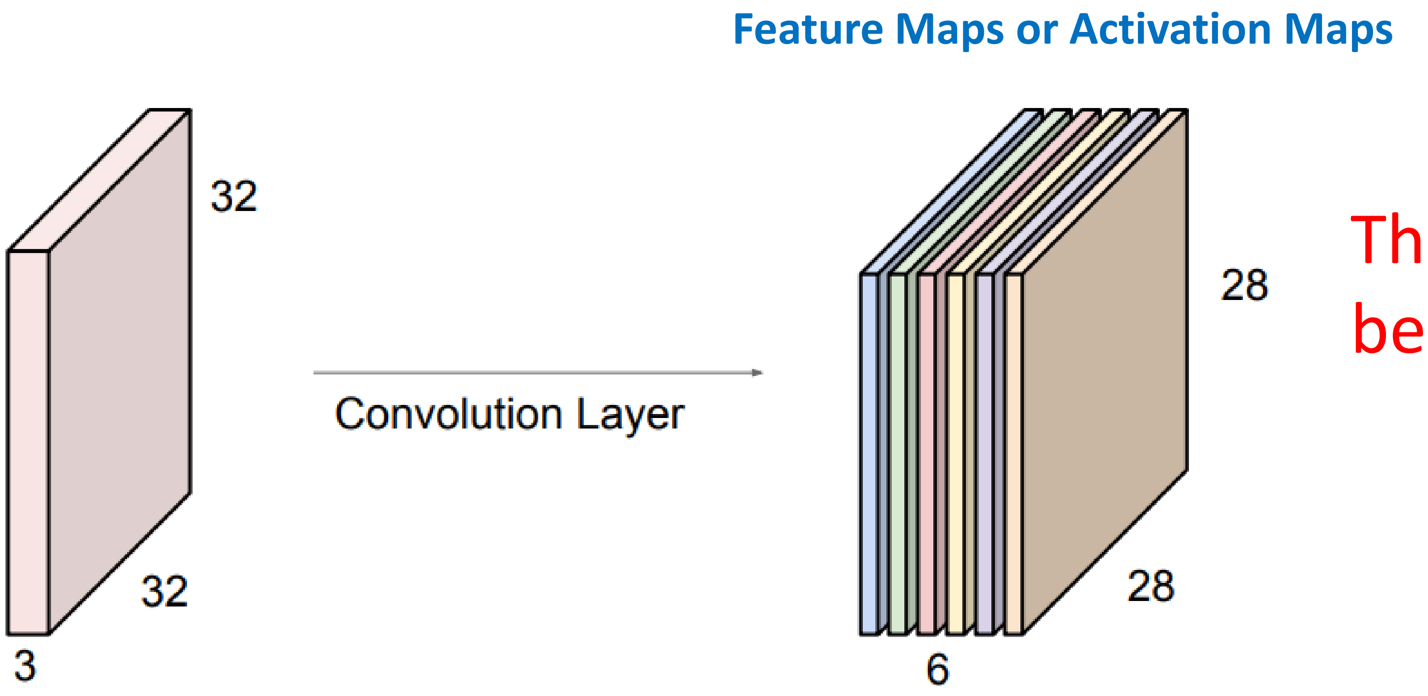
1	0	-1
2	0	-2
1	0	-1



Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

Convolution (Cross-Correlation Operation)

- If we use 6 different kernels, we'll get 6 separate feature maps (a new "image" of size $28 \times 28 \times 6$



The kernels should be learned!!

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

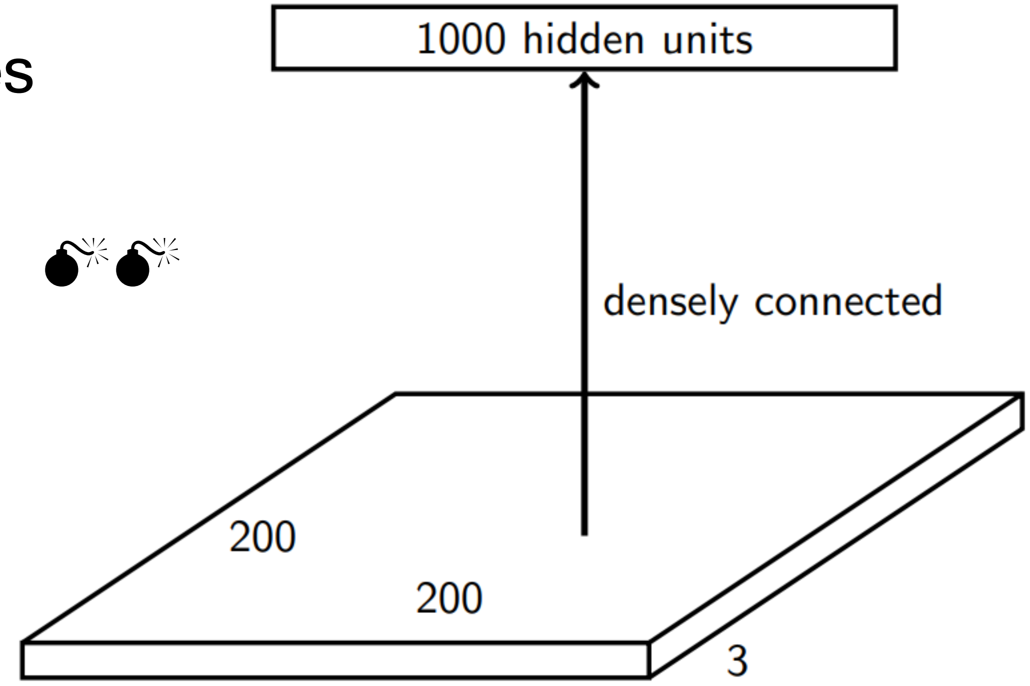
Convolution (Cross-Correlation Operation)

- **Remember …**

- Suppose we want to train a network that takes a 200×200 RGB (color) image as input.
    - First layer parameters = $120\text{K} \times 1000 = 120 \text{ milion}$ 💣💣

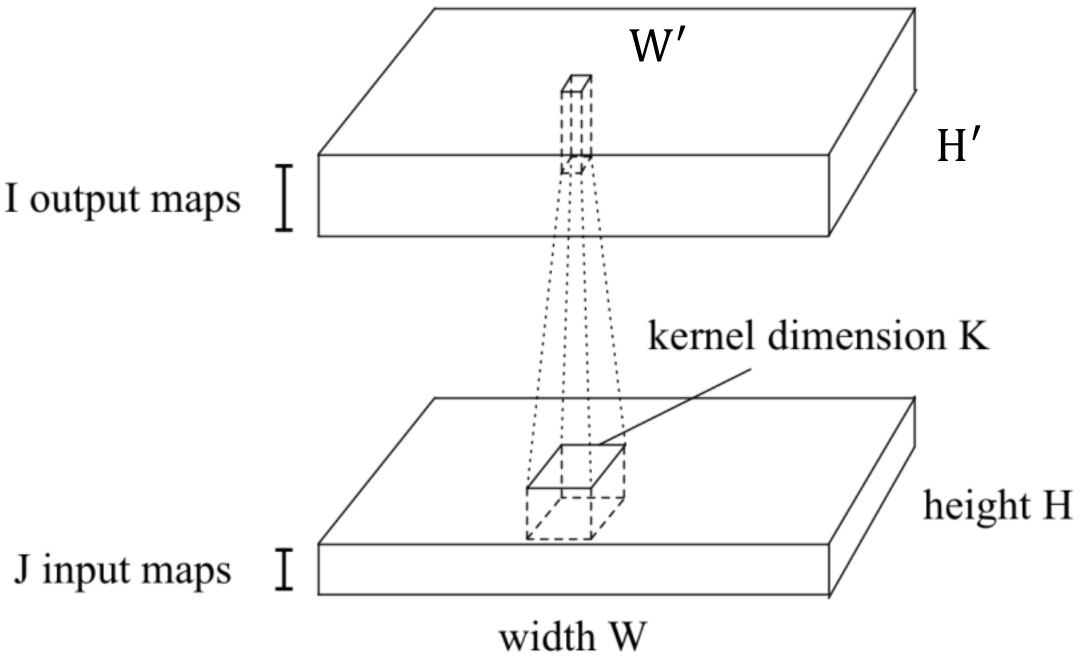
- If we use a convolutional layer with 64 filters of size $5 \times 5 \times 3$, how many parameters will the first layer have?

$$64 \times (5 \times 5 \times 3) = 4800$$



Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

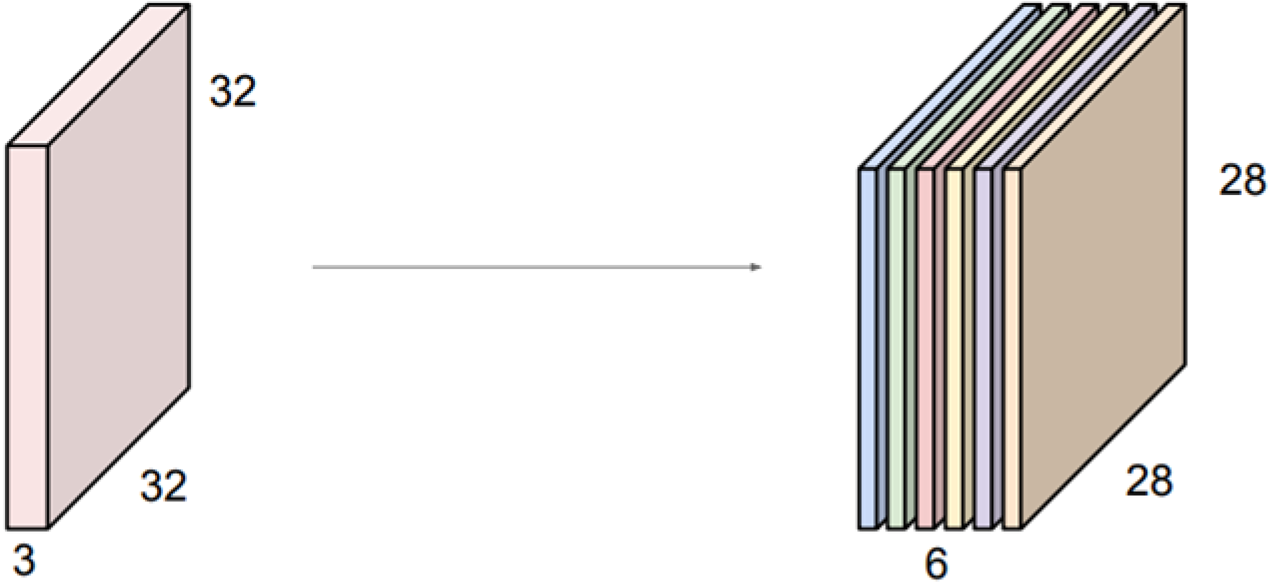
Convolution (Cross-Correlation Operation)



Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

	Fully connected layer	**Convolution layer**
#output units	$W' \times H' \times I$	$W' \times H' \times I$
#weights (params)	$W \times W' \times H \times H' \times I \times J$	$K^2 \times I \times J$

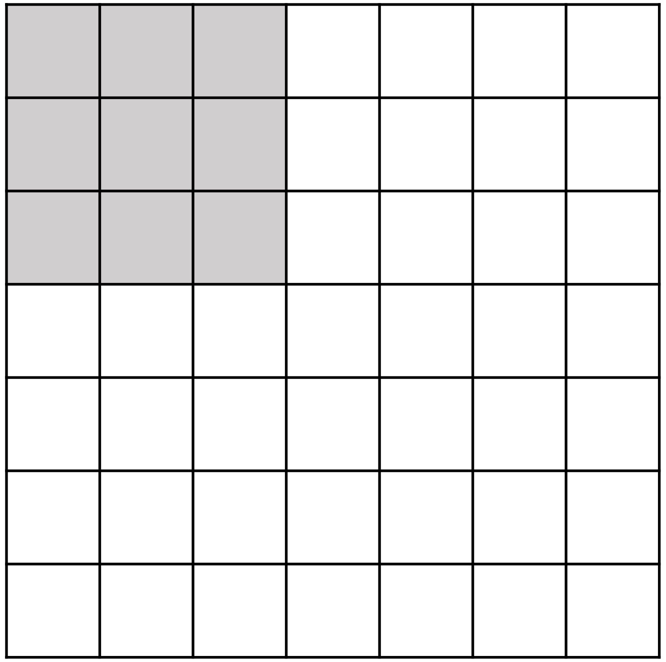
Convolution (Cross-Correlation Operation)



	Fully connected layer	Convolution layer
#output units	4368	4368
#weights (params)	14,450,688	450

Convolution (Cross-Correlation Operation)

Stride

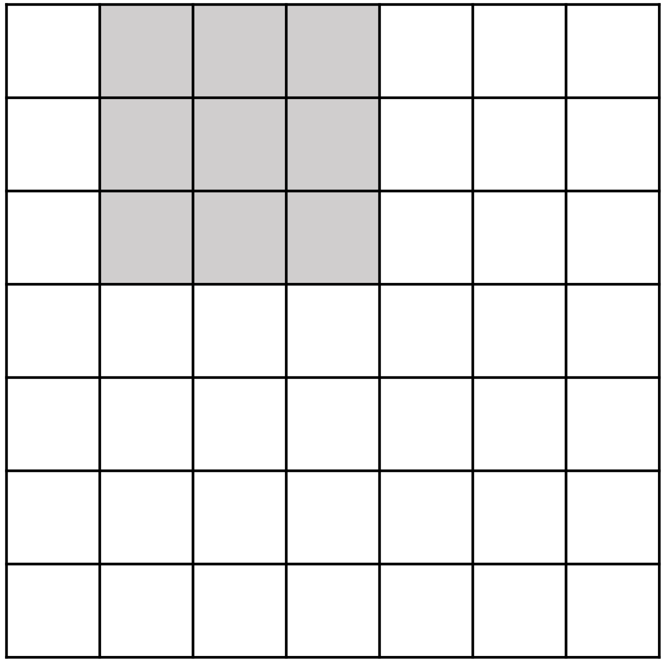


7x7 input (spatially)
assume 3x3 filter

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University:
http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride

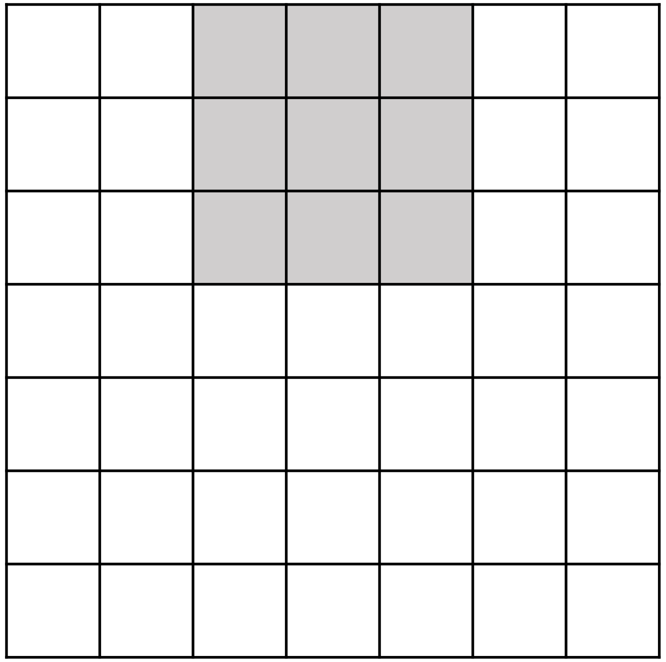


7x7 input (spatially)
assume 3x3 filter
Stride = 1

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride

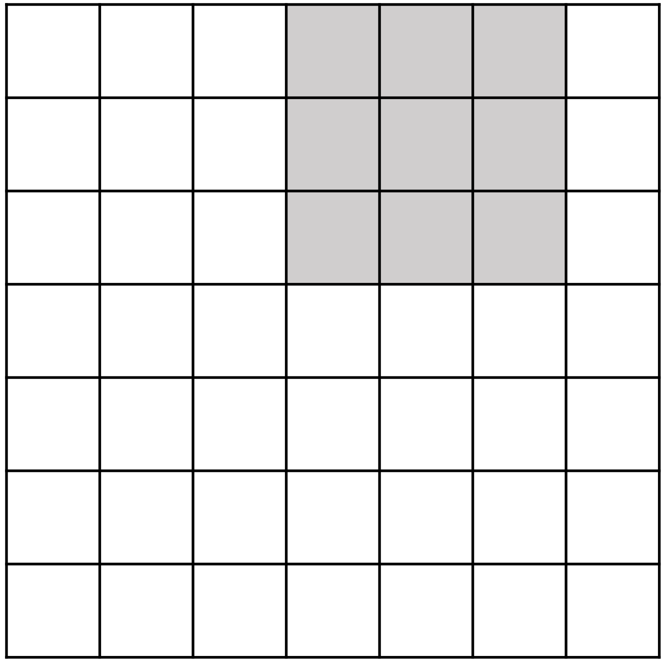


7x7 input (spatially)
assume 3x3 filter
Stride = 1

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride

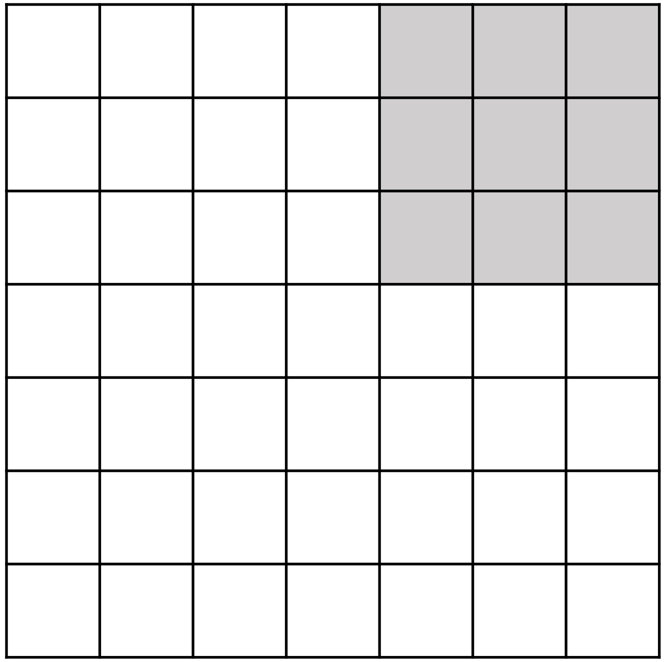


7x7 input (spatially)
assume 3x3 filter
Stride = 1

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride



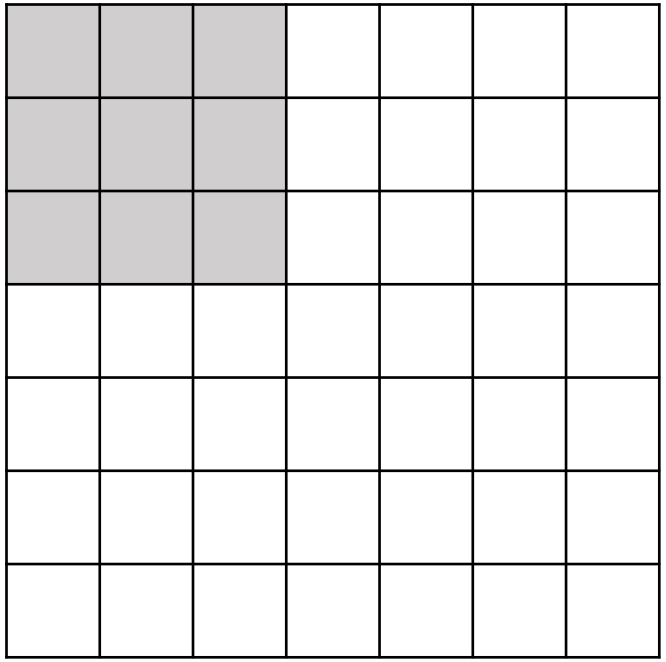
7x7 input (spatially)
assume 3x3 filter
Stride = 1

=> 5x5 output

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride

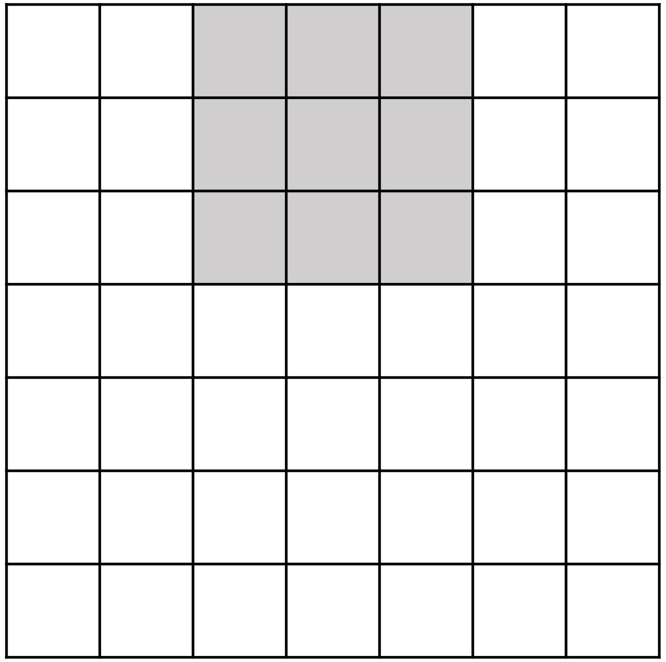


7x7 input (spatially)
assume 3x3 filter
stride = 2

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride

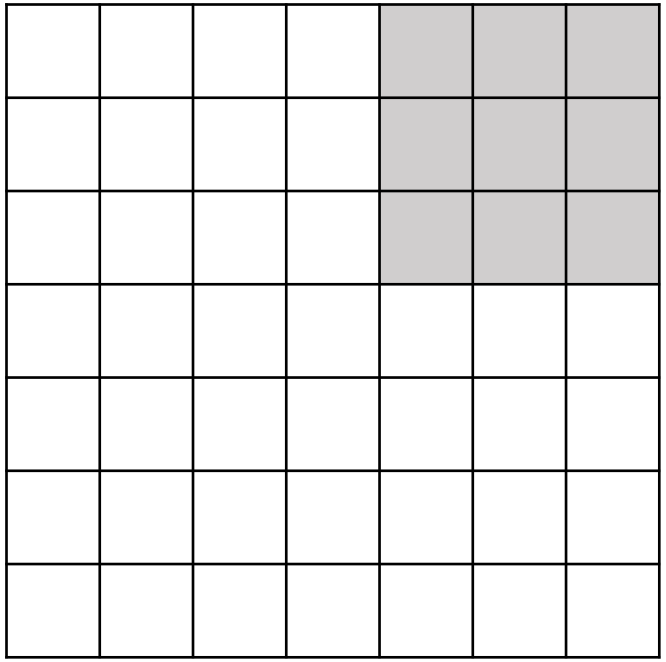


7x7 input (spatially)
assume 3x3 filter
stride = 2

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride



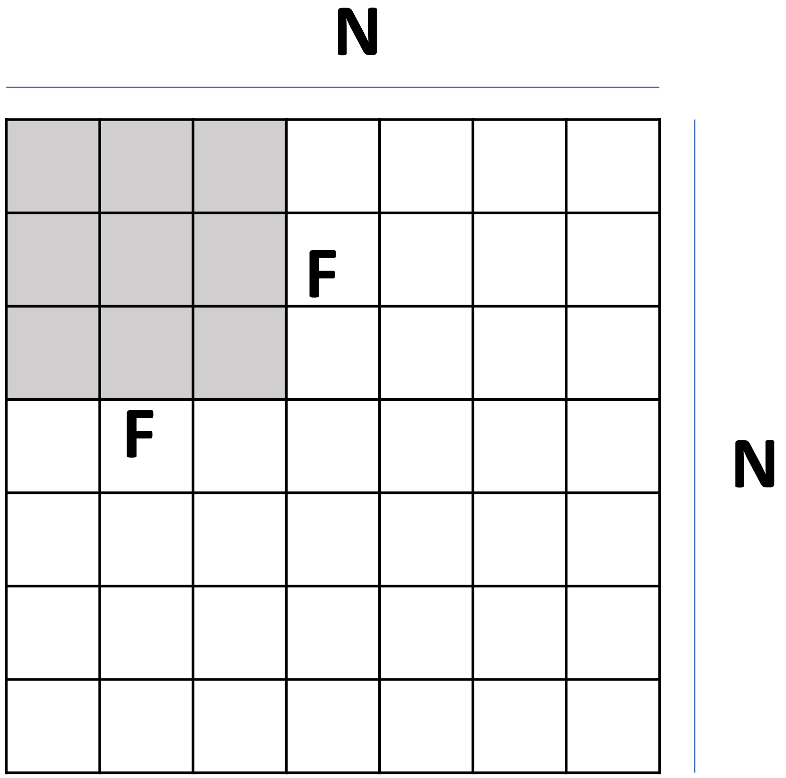
7x7 input (spatially)
assume 3x3 filter
stride = 2

=> 3x3 output

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Stride



Output size:

(N - F) / stride + 1

e.g. N = 7, F = 3:

stride 1 => (7 - 3)/1 + 1 = 5

stride 2 => (7 - 3)/2 + 1 = 3

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution (Cross-Correlation Operation)

Zero Padding

In practice: Common to zero pad the border

0	0	0	0	0	0	0	0	0
0								0
0								0
0								0
0								0
0								0
0								0
0								0
0	0	0	0	0	0	0	0	0

Input 7x7, **3x3** filter applied **with stride 1 pad with 1** pixel border => what is the output?

(recall:) (N - F) / stride + 1

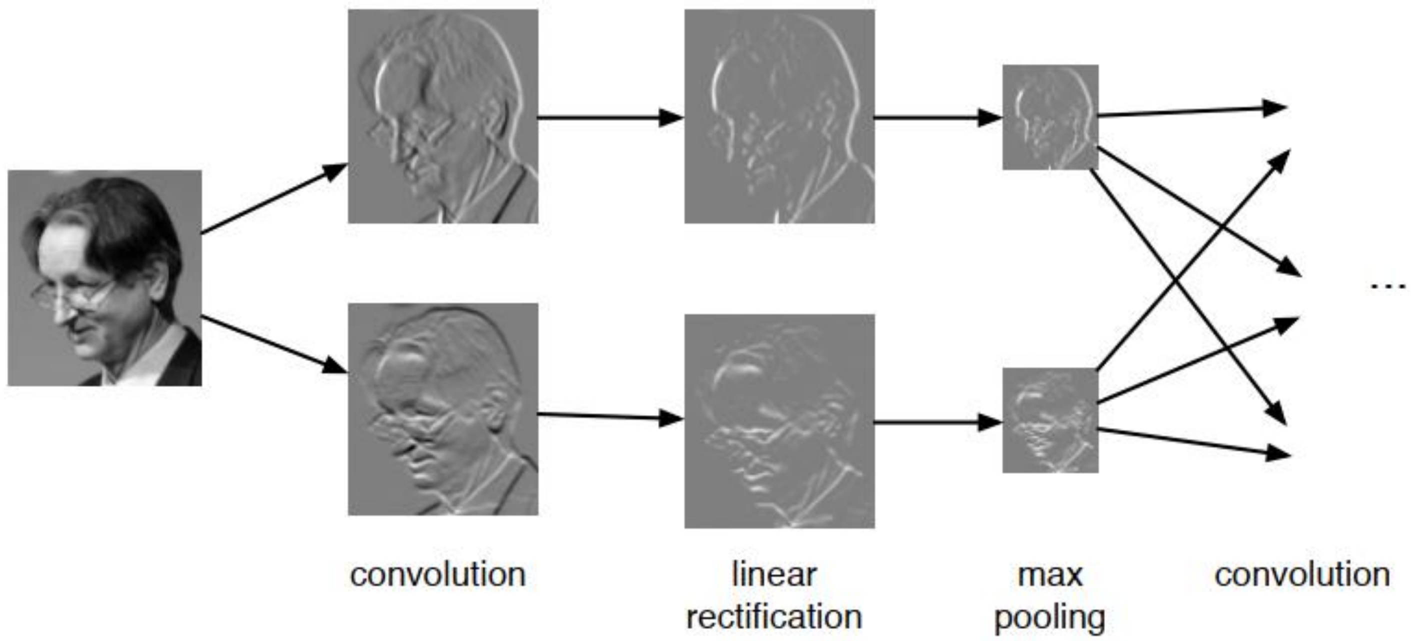
(N + 2P - F) / stride + 1

=> 7x7

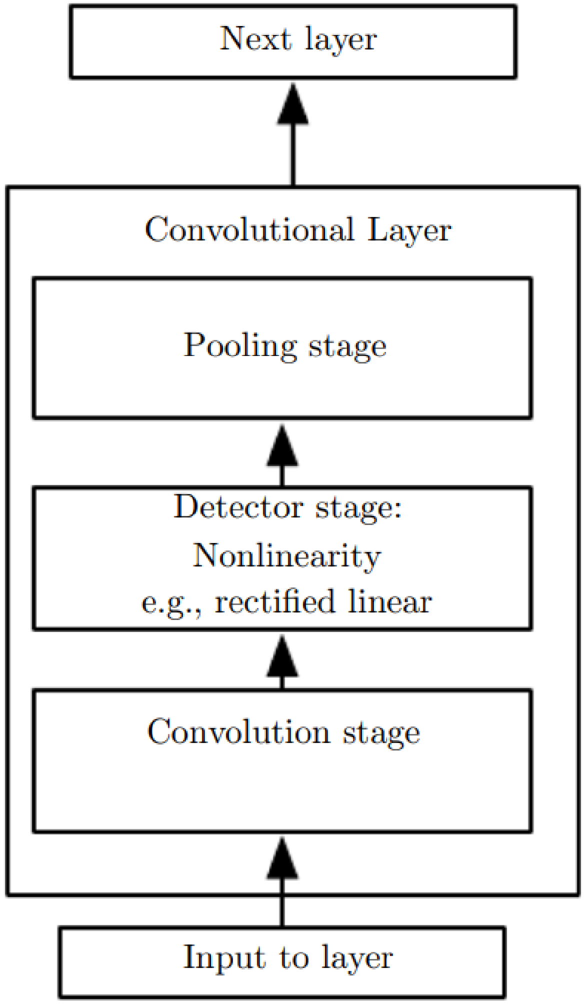
Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Convolution Layer

- A typical layer of a convolutional network consists of three stages



Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

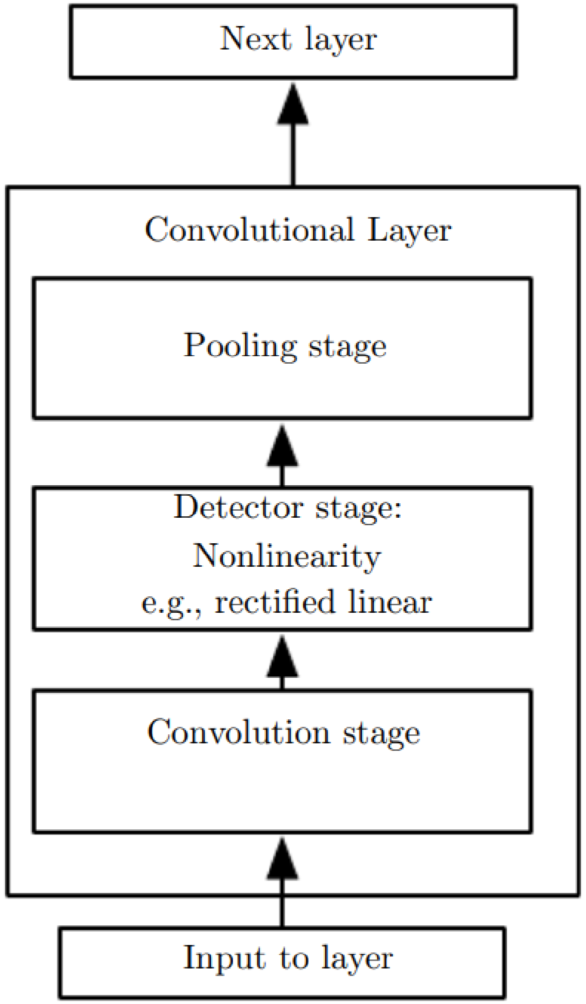


Source: Goodfellow et al. (2016), Deep Learning

Convolution Layer

- A typical layer of a convolutional network consists of three stages

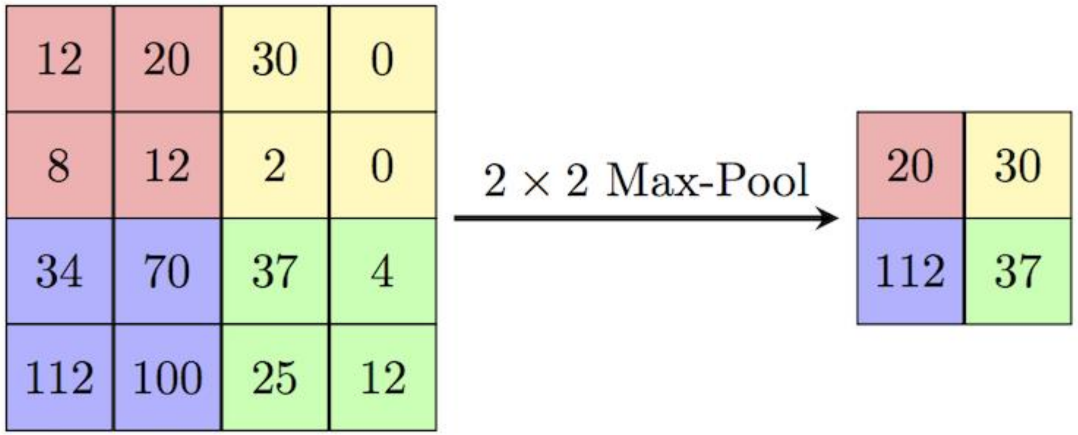
- Why do we use a detector stage?
  - Convolution is a linear operation. Therefore, we need a nonlinearity, otherwise 2 convolution layers would be no more powerful than 1.



Source: Goodfellow et al. (2016), Deep Learning

Pooling

- A pooling function takes the output of the previous layer at a certain location L and computes a summary statistic of the neighborhood around L.
- **Example**: the max pooling [1]

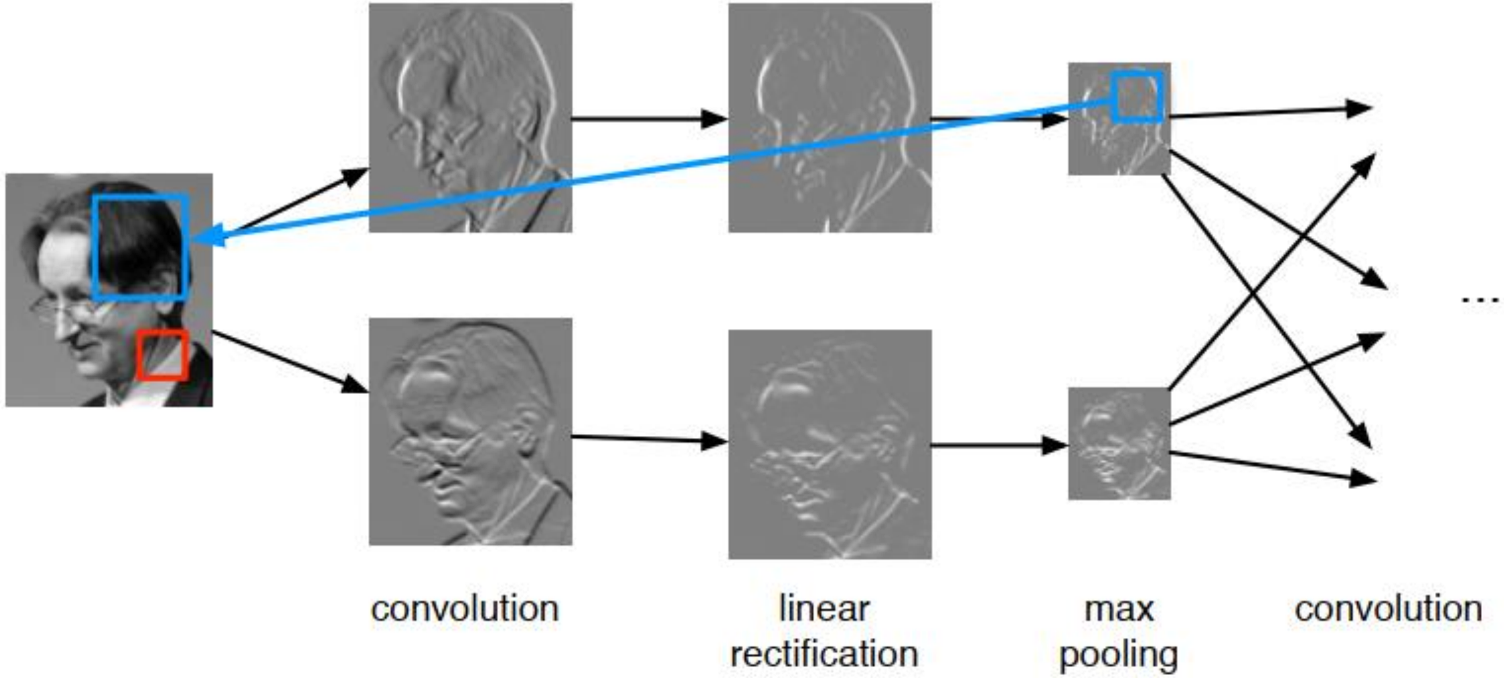


https://paperswithcode.com/method/max-pooling

[1] Zhou, Y.T. and Chellappa, R., 1988, July. Computation of optical flow using a neural network. In ICNN (pp. 71-78)

Pooling

- Pooling layers reduce the size of the representation.
  - Higher-layer filters can cover a larger region of the input than equal-sized filters in the lower layers.
  - Reduces the computational burden on the next layer.

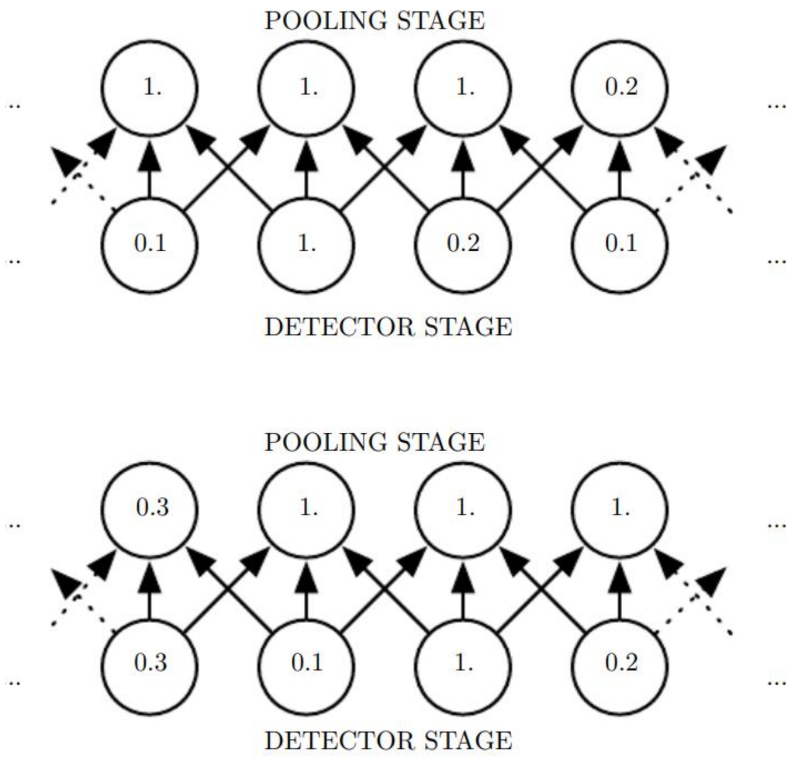


Source: Neural Networks and Deep Learning course by Jimmy Ba, 2020, University of Toronto: https://csc413-2020.github.io/

Pooling

- Pooling helps to make the representation approximately invariant to small translations of the input.
- i.e. if we translate the input by a small amount, the values of most of the pooled outputs do not change.

Figure: After the input has been shifted to the right by one pixel, every value in the bottom row has changed, but only half of the values in the top row have changed, because the max pooling units are sensitive only to the maximum value in the neighborhood, not its exact location.



Source: Goodfellow et al. (2016), Deep Learning

CNN Applications from the Literature
(Image Classification)

Object Recognition

- Object recognition is the task of identifying which object category is present in an image.

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Object Recognition

Challenges: Background Clutter

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Object Recognition

Challenges: Illumination

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Object Recognition

Challenges: Deformation

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Object Recognition

Challenges: Occlusion

Source: CS231n: Convolutional Neural Networks for Visual Recognition by Fei-Fei Li, 2017, Stanford University: http://cs231n.stanford.edu/2017/index.html

Object Recognition

- Why object recognition is important?
  - It is an integral part of all the image search techniques
  - It is closely related to object detection
  - **Object detection**: locating all instances of an object in an image

Source: Yang, S., Luo, P., Loy, C.C. and Tang, X., 2016. Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5525-5533).

Datasets

- MNIST

  Categories: 10 digit classes
  Source: Scans of handwritten zip codes from envelopes
  Size: 60,000 training images and 10,000 test images, grayscale, of size 28 × 28

Source: https://en.wikipedia.org/wiki/MNIST_database

Datasets

- ImageNet [1]
  - ImageNet is an image dataset organized according to the WordNet [2] hierarchy.
  - Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset".
  - There are more than 100,000 synsets in WordNet; the majority of them are nouns (80,000+).
  - In ImageNet, the aim is to provide on average 1000 images to illustrate each synset.

[1]: https://www.image-net.org/about.php
[2]: Princeton University "About WordNet." WordNet. Princeton University. 2010.

Datasets

- ImageNet
  - The most highly-used subset of ImageNet is the **ImageNet Large Scale Visual Recognition Challenge (ILSVRC)** 2012-2017 image classification and localization dataset [1].

    Categories: 1000 object classes

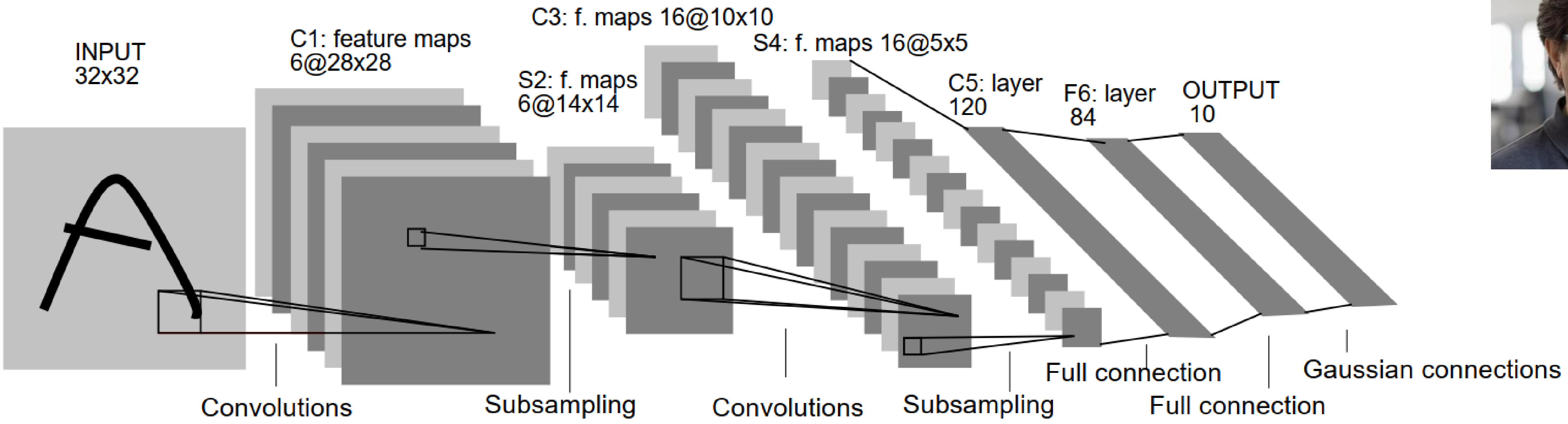
    Size: 1,281,167 training images and 50,000 validation images and 100,000 test images.

[1]: https://www.image-net.org/challenges/LSVRC/index.php

Models

LeNet

- Proposed by Yann LeCun and colleagues in 1998 [1] which was able to classify digits with 98.9% test accuracy.
- It was good enough to be used in a system for automatically reading numbers on checks.
- Parameters: 60000



[1]: LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

Models

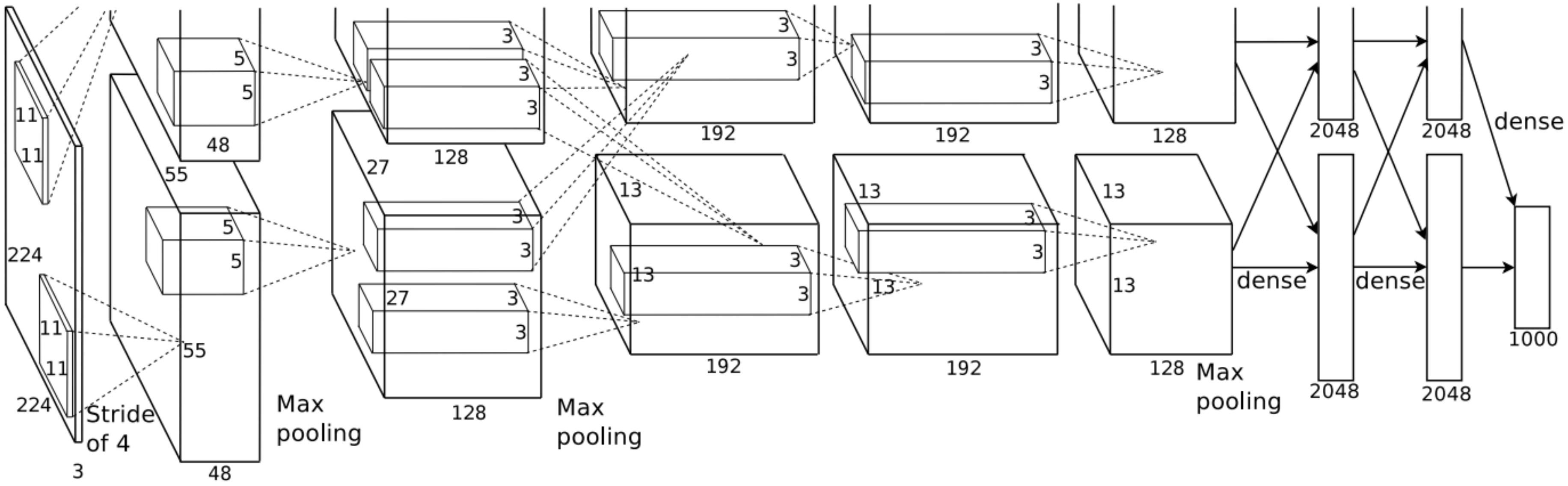
Top-5 error rate over time on ILSVRC

• 2012: AlexNet	16.5%	✓
• 2013: ZF	11.7%	✗
• 2014: VGG	7.3%	✗
• 2014: GoogLeNet	6.7%	✓
• 2015: ResNet	3.6%	Homework
• 2016: GoogLeNet-v4	3.1%	Homework

Human error rate: 5.1% [Russakovsky et al. 2015]

Models

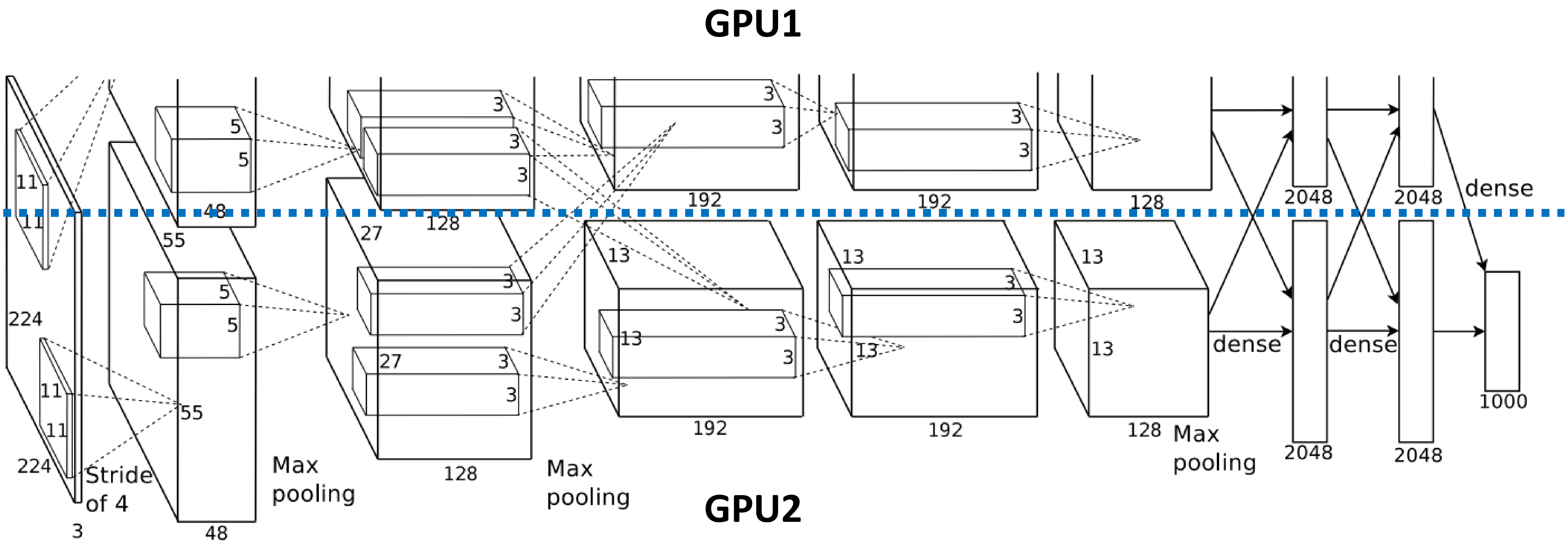
- AlexNet [1]
  - 16.4% top-5 error on ILSVRC
  - Lots of tricks (ReLU units, weight decay, SGD with momentum, dropout, data augmentation)
  - Parameters: 60 million



[1]: Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. NIPS 2012.

Models

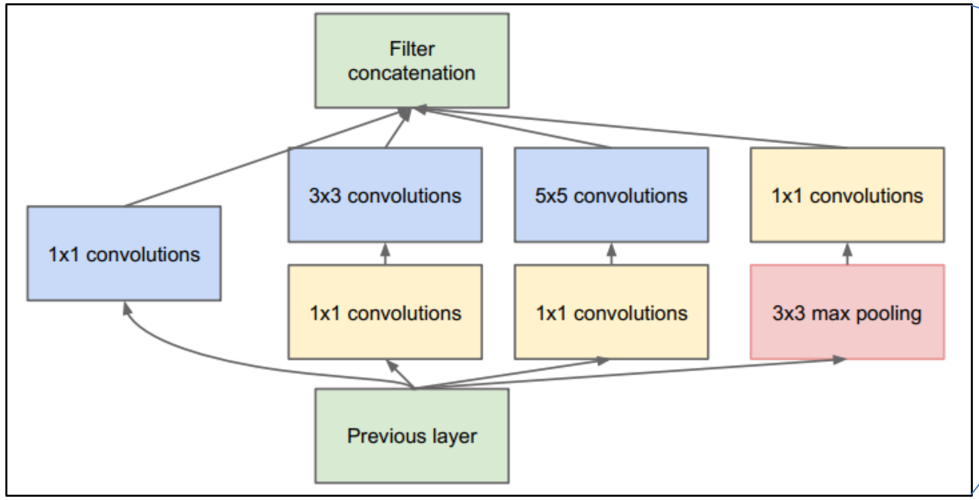
- AlexNet [1]

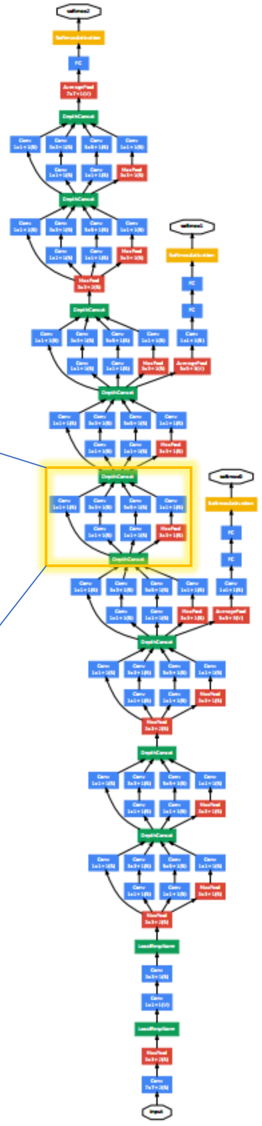


[1]: Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. NIPS 2012.

Models

- GoogLeNet [1]
  - 6.6% top-5 error on ILSVRC
  - Fully convolutional (no fully connected layers)
  - Convolutions are broken down into a bunch of smaller convolutions (Inception modules)





[1]: Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).