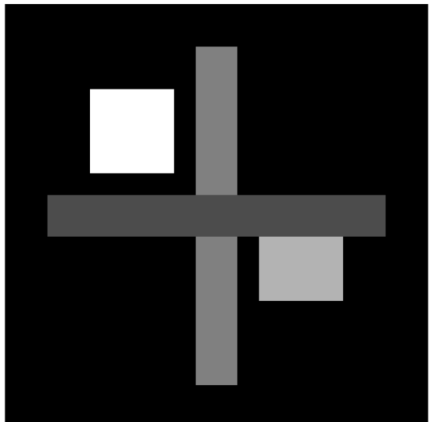
Introduction to Machine Learning

Convolutional Neural Network (CNN)

Prof. Subrahmanya Swamy Peruru
Department of Electrical Engineering
IIT Kanpur

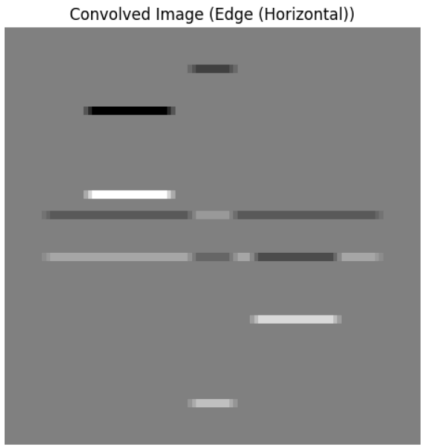
Convolution Filters for Feature Extraction

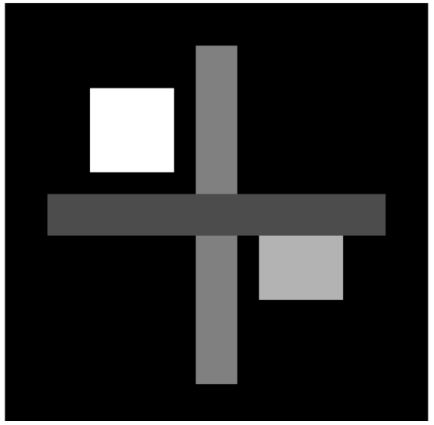
Original Image



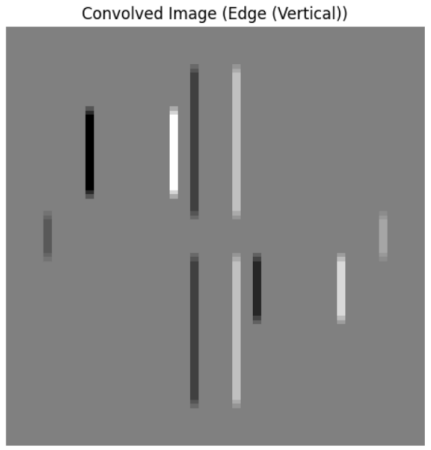
$$\begin{matrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{matrix}$$

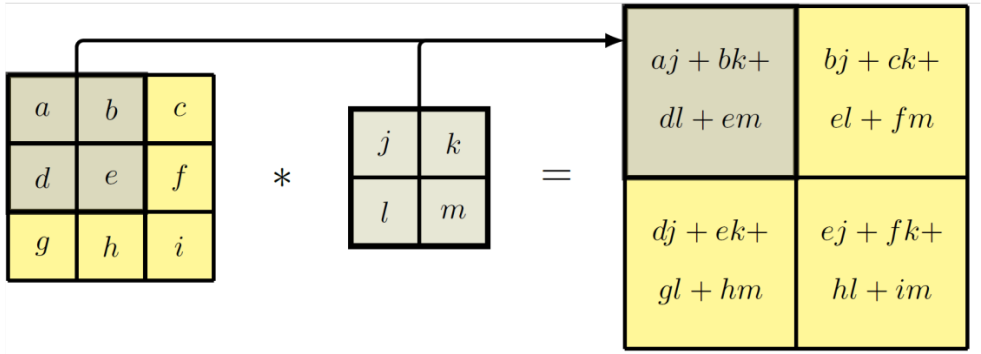
Extracted Features





$$\begin{matrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{matrix}$$





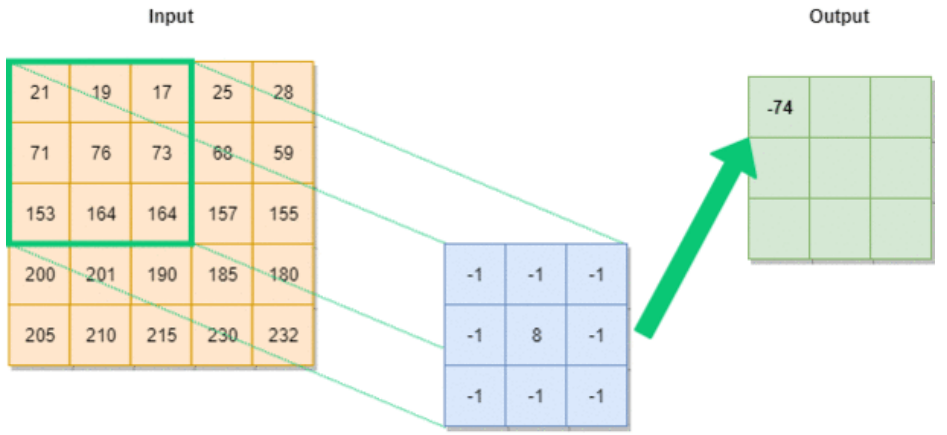
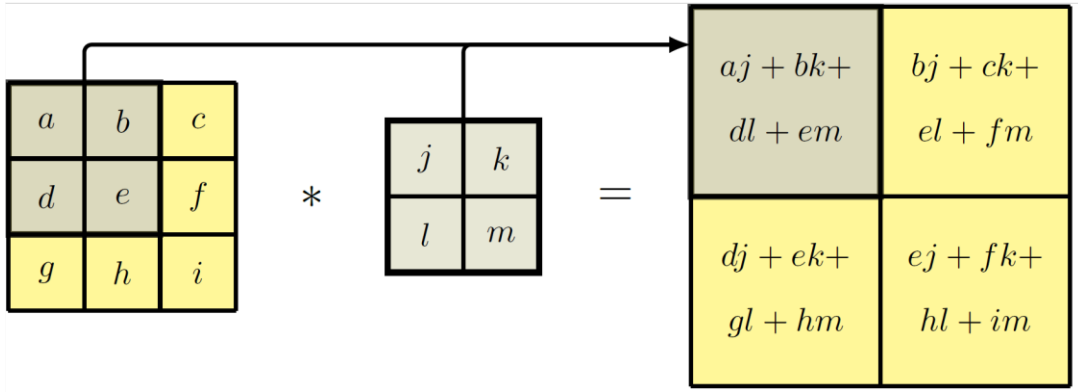


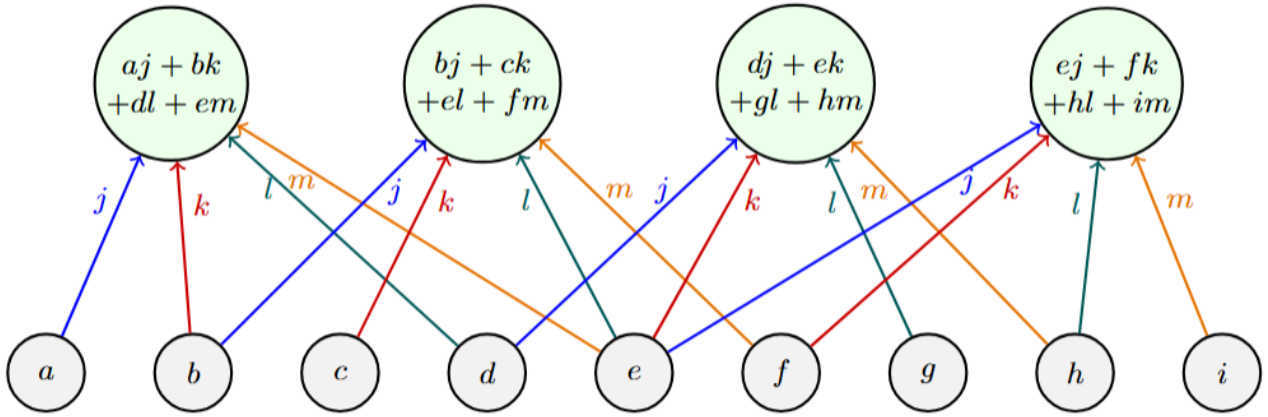
Fig Src: https://commons.wikimedia.org/wiki/File:CNN-filter-animation-1.gif

Filters Demo

Convolutional Neural Networks (CNN)

- CNNs use connections between layers that are different from MLPs in two key ways

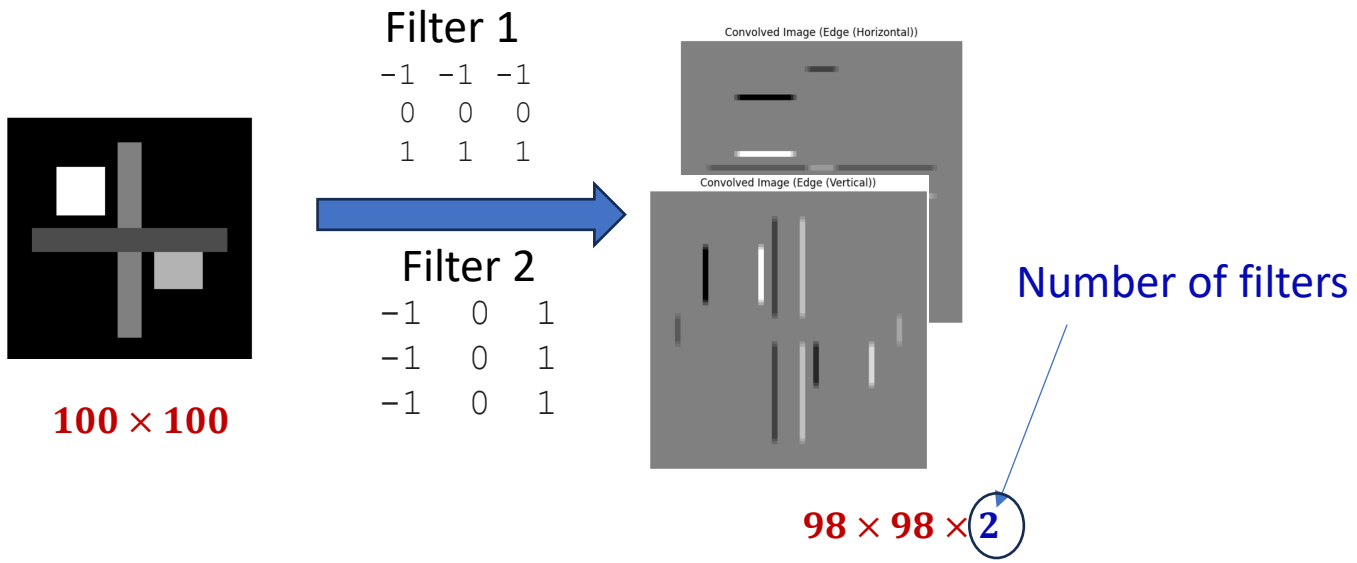




- Change 1: Each hidden layer node is connected only to a local patch in previous layer
- Change 2: Same set of weights used for each local patch (blue, red, green, orange is one set of weights, and this same set of used for all patches)
- These changes help in
  - Substantial reduction on the number of weights to be learned (9×4 Vs 4)
  - Learning the local structures within the inputs

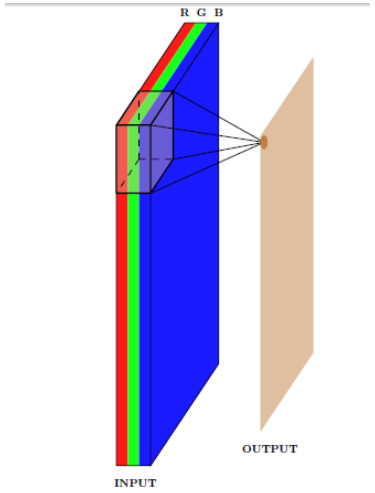
Convolution: Multiple Filters

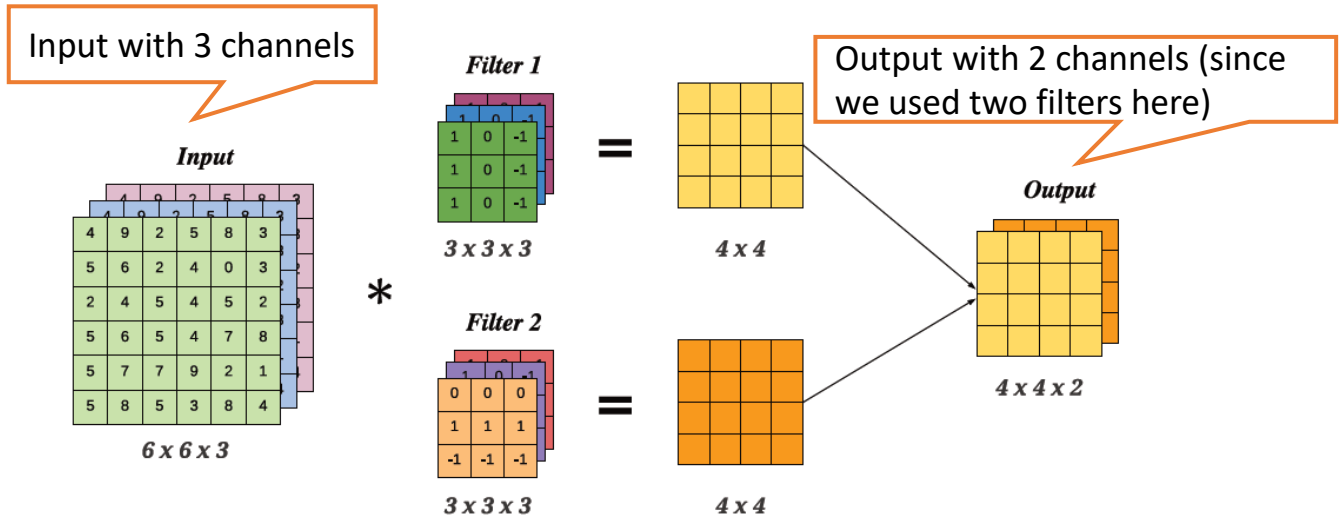
- High “match” of a filter/kernel with a patch gives high values in the feature map
- In CNN, these weights/filters are learnable. Also, usually multiple filters are used
  - Each filter gives us a different feature map (K filters will give K feature maps)
  - Each map can be seen as representing a different type of feature in the inputs



Convolution: Multiple Input Channels

- If the input has multiple channels (e.g., images with R,G,B channels), then each filter/kernel also needs to have multiple channels, as shown below (left figure)
- We perform per-channel convolution followed by an aggregation (sum across channels)





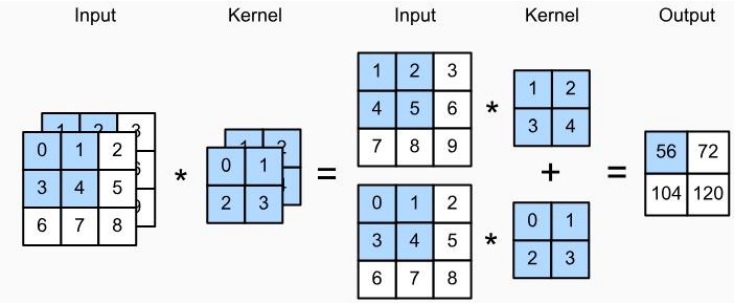
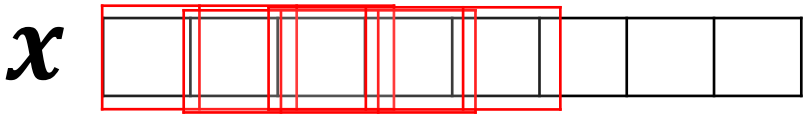


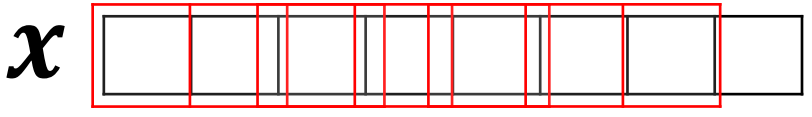
Figure credit: PML-1 (Murphy, 2022),

Convolution: Stride size

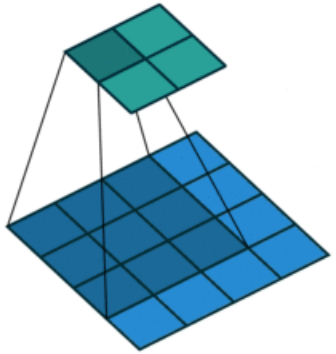
- When “moving” the filter across the input, the stride size can be one or more than one
  - Stride means how much the filter moves between successive convolutions



Stride of 1



Stride of 2



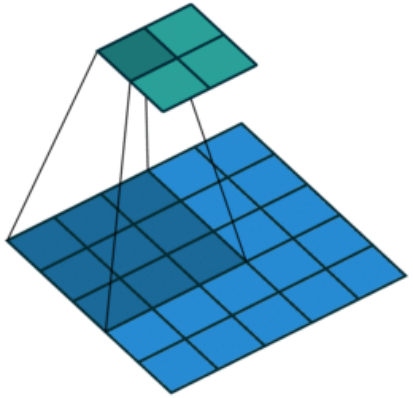
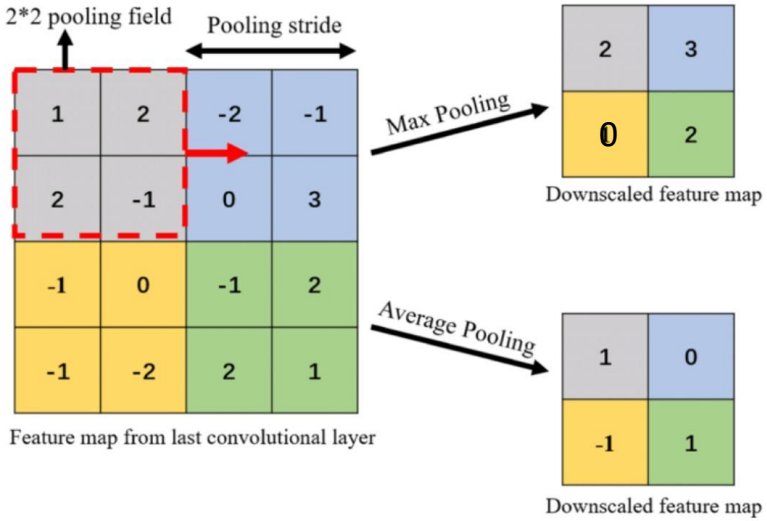


Fig Src: https://hannibunny.github.io/mlbook/neuralnetworks/convolutionDemos.html

Pooling

- CNNs also consist of a pooling operation after each conv layer
- Pooling plays two important roles
  - Reducing the size of the feature maps
  - Combining local features to make global features
- Need to specify the size of group to pool, and pooling stride
- Max pooling and average pooling are popular pooling methods



Convolutional Neural Networks (CNN)

- CNN consists of a sequence of operations to transform an input to output
  - Convolution (a linear transformation but more "local" than the one in MLP)
  - Nonlinearity (e.g., sigmoid, ReLU, etc) after the convolution operation
  - Pooling (aggregates local features into global features and reduce representation size)

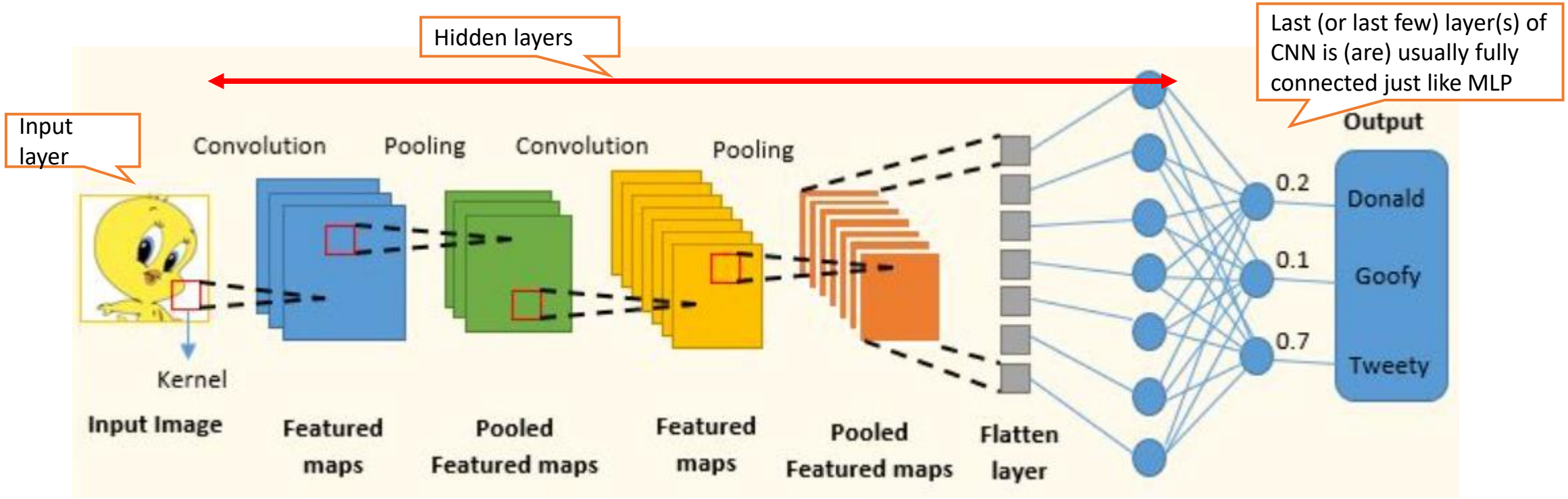
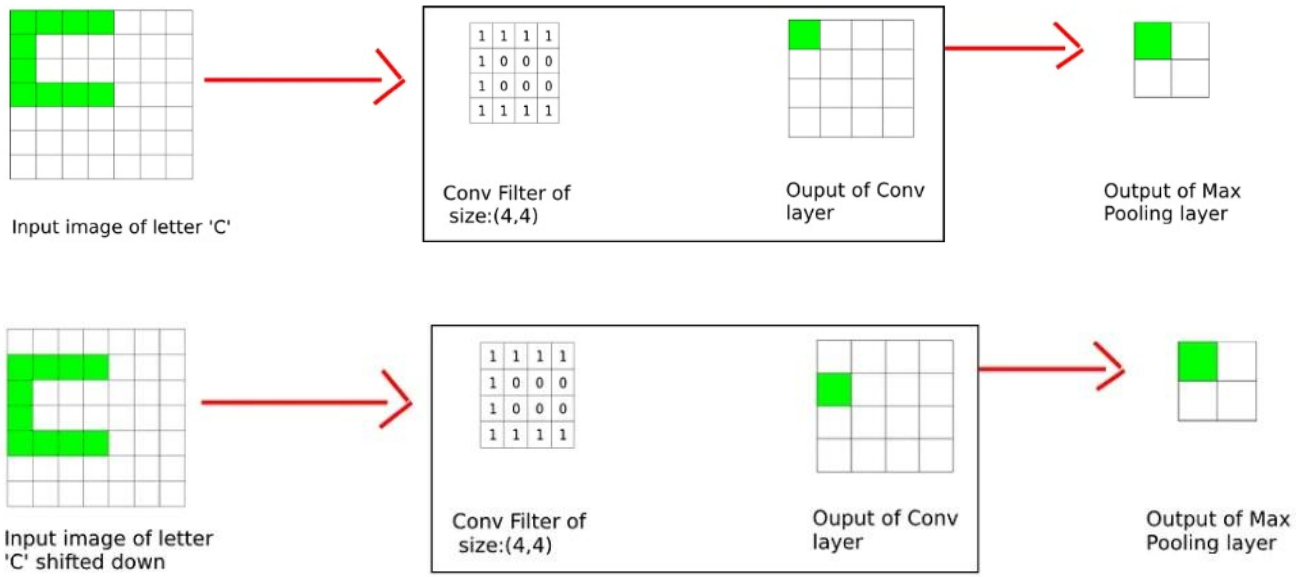


Figure credit: https://www.analyticsvidhya.com/blog/2022/01/convolutional-neural-network-an-overview/

CNNs have Translation Invariance!

- Even if the object of interest has shifted/translated, CNN don't face a problem (it will be detected regardless of its location in the image)
- The simple example below shows how (max) pooling helps with this



Pic credit: https://divsoni2012.medium.com/translation-invariance-in-convolutional-neural-networks-61d9b6fa03df

CNN: Summary of the overall architecture

- The overall structure of a CNN looks something like this

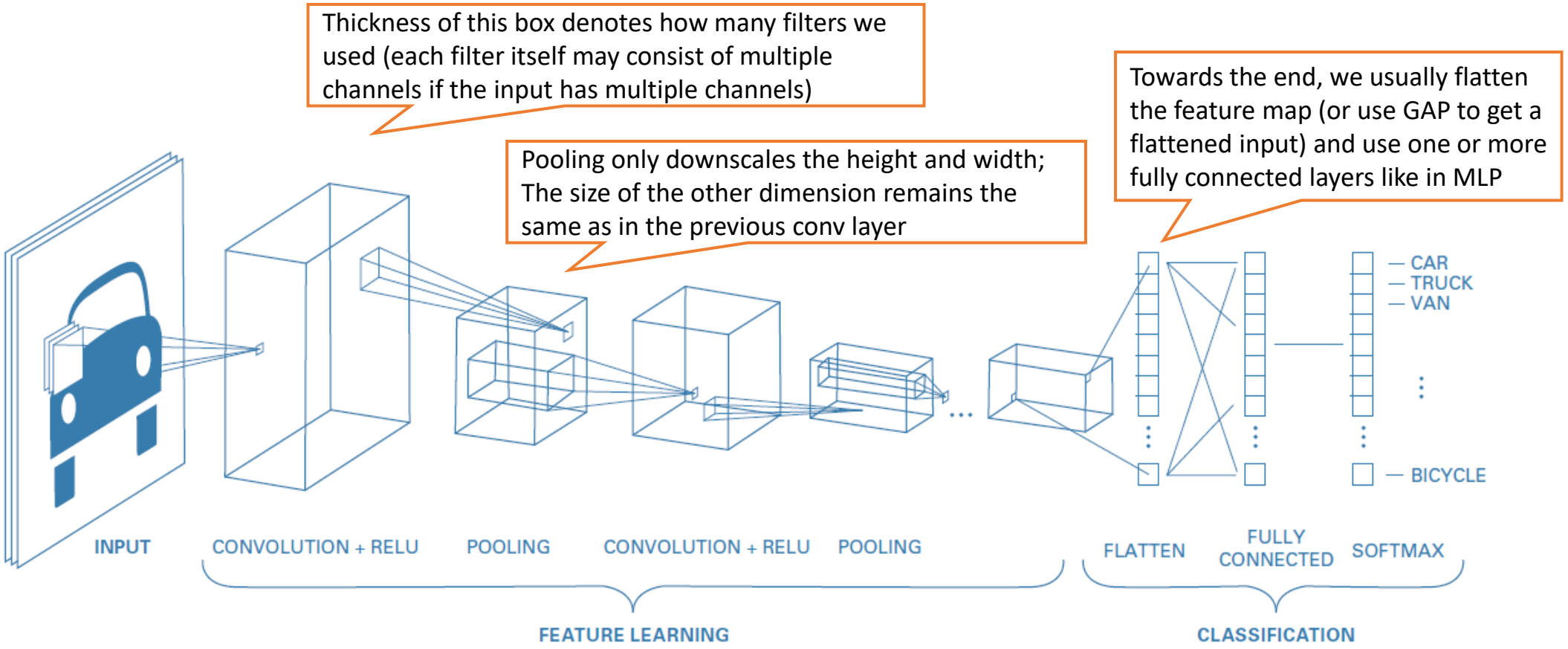


Figure credit: PML-1 (Murphy, 2022),

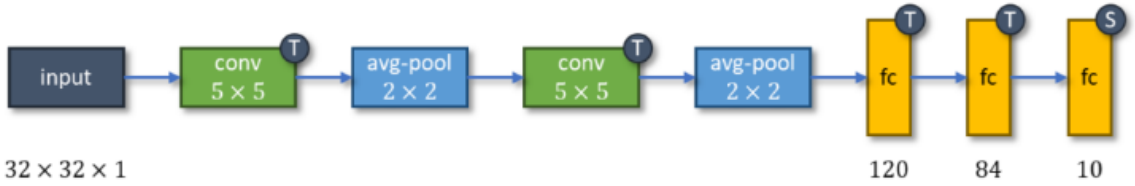
LENET-5

LeNet-5

This starts it all. Excluding pooling, LeNet-5 consists of 5 layers:

- 2 convolution layers with kernel size 5×5, followed by
- 3 fully connected layers.

Each convolution layer is followed by a 2×2 average-pooling, and every layer has tanh activation function except the last (which has softmax).

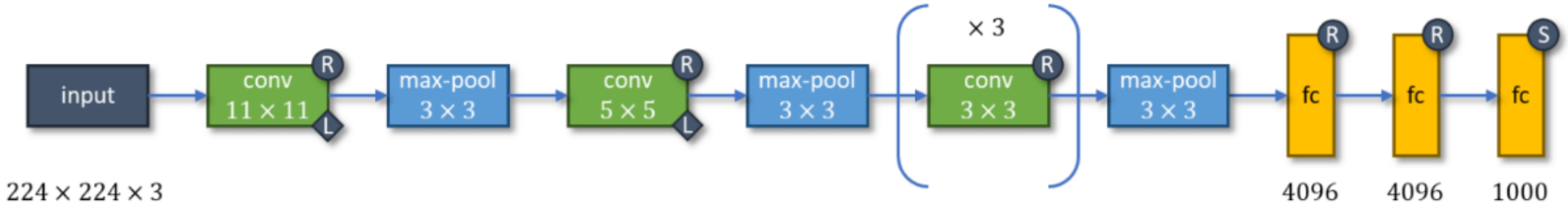


LeNet-5 architecture | Image by author

LeNet-5 has **60,000** parameters. The network is trained on greyscale 32×32 digit images and tries to recognize them as one of the ten digits (0 to 9).

Slide credit: https://towardsdatascience.com/5-most-well-known-cnn-architectures-visualized-af76f1f0065e/

AlexNet



AlexNet architecture | Image by author

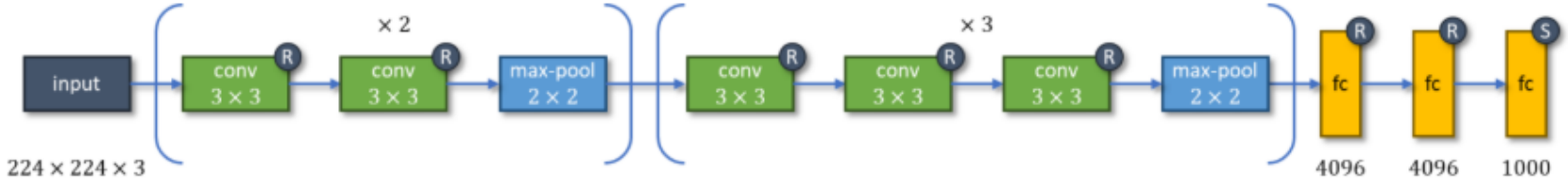
The network consists of 8 layers:

- 5 convolution layers with non-increasing kernel sizes, followed by
- 3 fully connected layers.

60 million parameters

Slide credit: https://towardsdatascience.com/5-most-well-known-cnn-architectures-visualized-af76f1f0065e/

VGG-16



VGG-16 architecture | Image by author

This story focuses on VGG-16, a deep CNN architecture with, well, 16 layers:

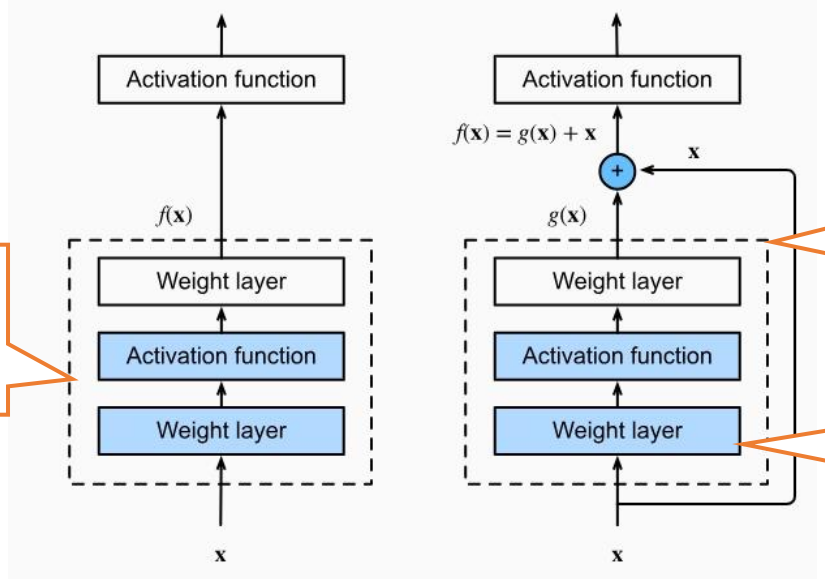
- 13 convolution layers with kernel size 3×3, followed by
- 3 fully connected layers.

138 million parameters

Slide credit: https://towardsdatascience.com/5-most-well-known-cnn-architectures-visualized-af76f1f0065e/

Residual/Skip Connections (ResNet-50)

- In general, just stacking lots of layer doesn't necessarily help a deep learning model
  - Vanishing/exploding gradient may make learning difficult
- Skip connections or "residual connections" help if we want very deep networks
  - This idea was popularized by "Residual Networks"* (ResNets) which can have hundreds of layers
- Basic idea: Don't force a layer to learn everything about a mapping



These layers trying to learn some function $f(x)$

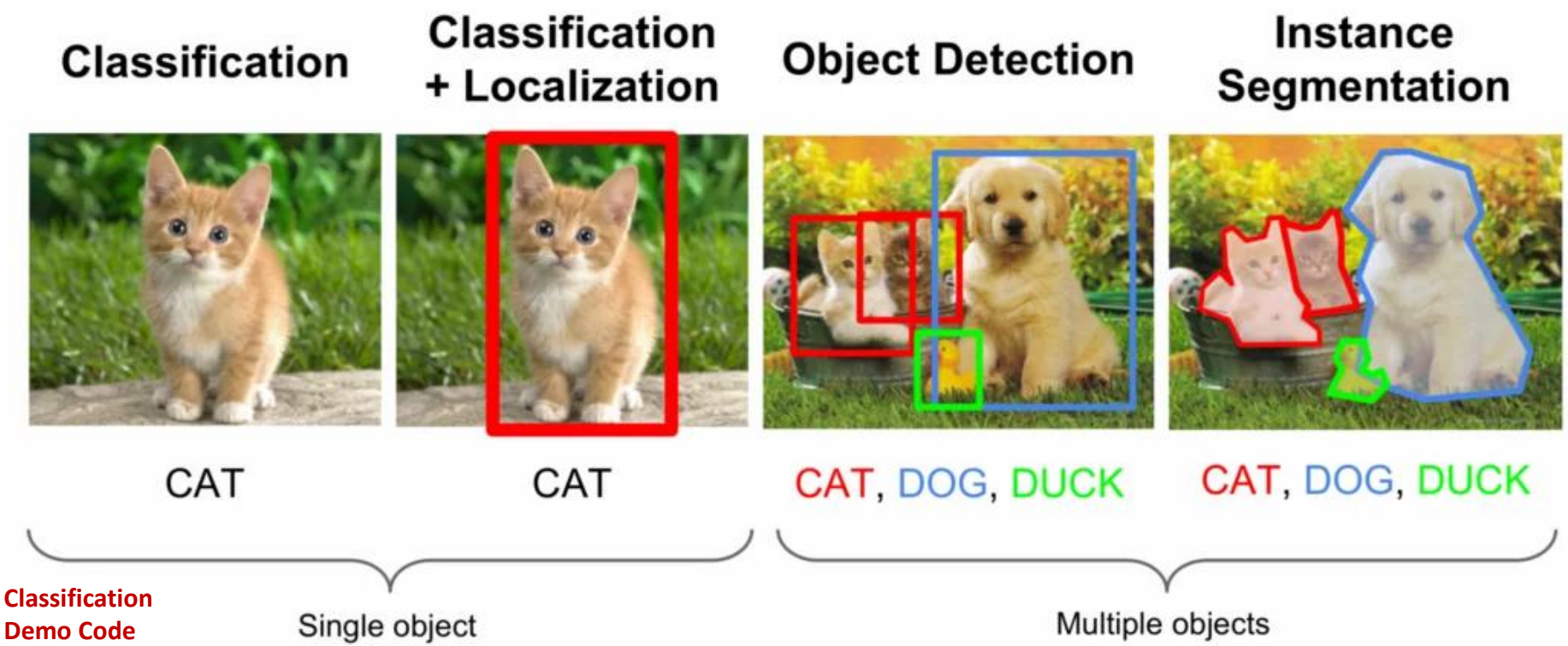
Added a "residual branch" or "short-cut" connection to connect x to the residual output $g(x)$ of these layers

Reducing their burden by just asking them to learn the "residual" $g(x) = f(x) - x$

*Deep Residual Learning for Image Recognition (He et al, 2015)

Pic source: https://www.d2l.ai/index.html

Different tasks with Images



Classification Demo Code

Task	Output	Typical models
Classification	one label	VGG, ResNet
Detection	boxes + labels	YOLO, Faster R-CNN
Semantic Segmentation	pixel classes	U-Net, DeepLab
Instance Segmentation	separate object masks	Mask R-CNN

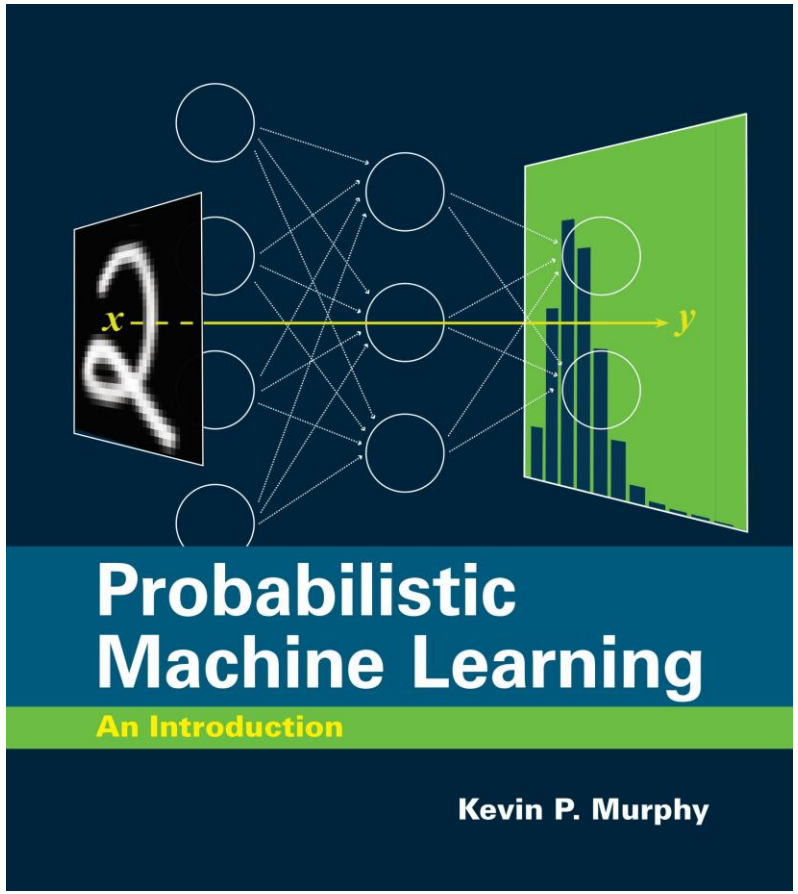
Figure credit: https://svitla.com/blog/cnn-for-image-processing/

Acknowledgments

- Slides are adapted from
  - Prof. Piyush Rai's course at IITK

References

Chapter 14



Thank you