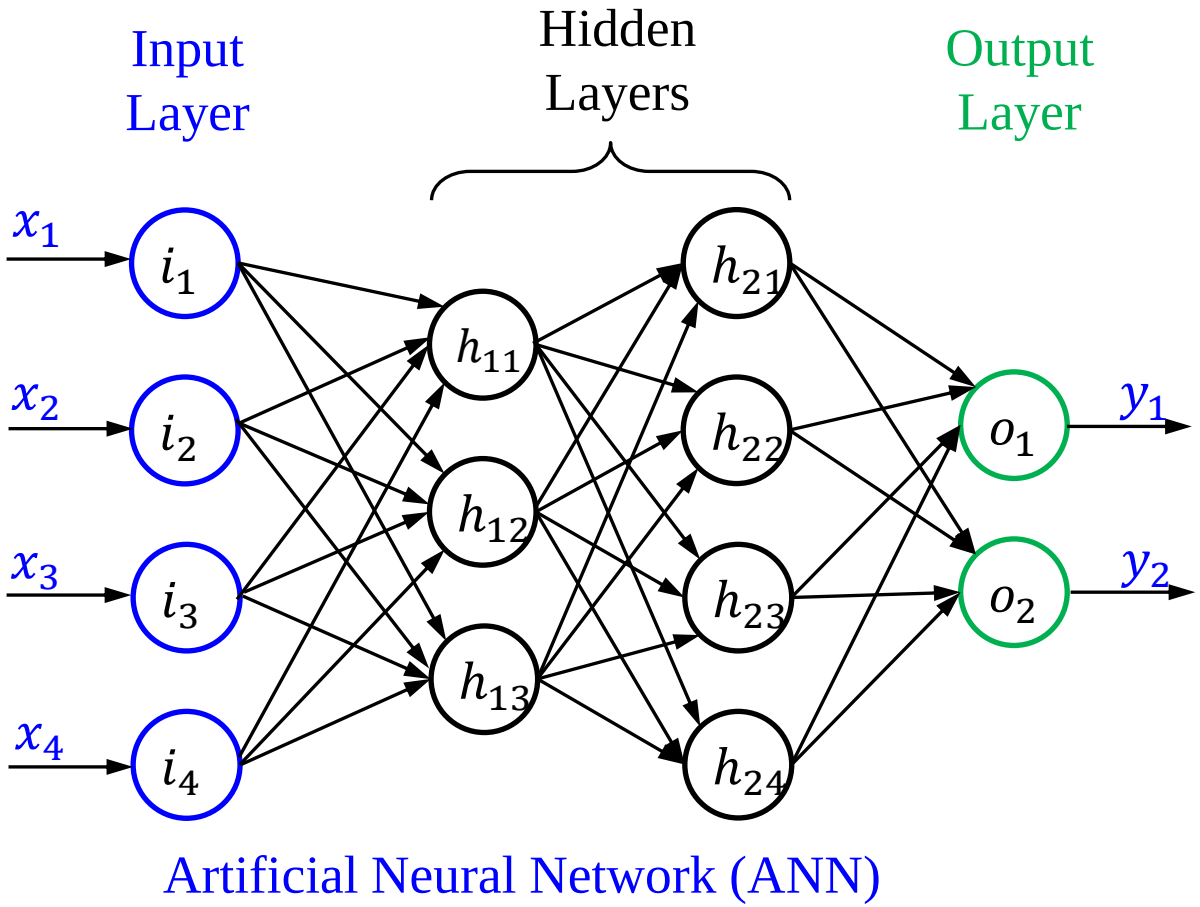
Module 3
Datasets
Convolution Neural Networks

Datasets



Artificial Neural Network (ANN)

- Image classification is one of the common benchmarks for neuromorphic tasks.
- Reaching high recognition performance of a neural network depends on training it on large-scale datasets of inputs and labels from large quantities of high-quality samples.
- Lacking objective and typical performance on standard benchmarks as a baseline to foster competition among proposed methods.
- Inherent bias and imbalance in datasets.

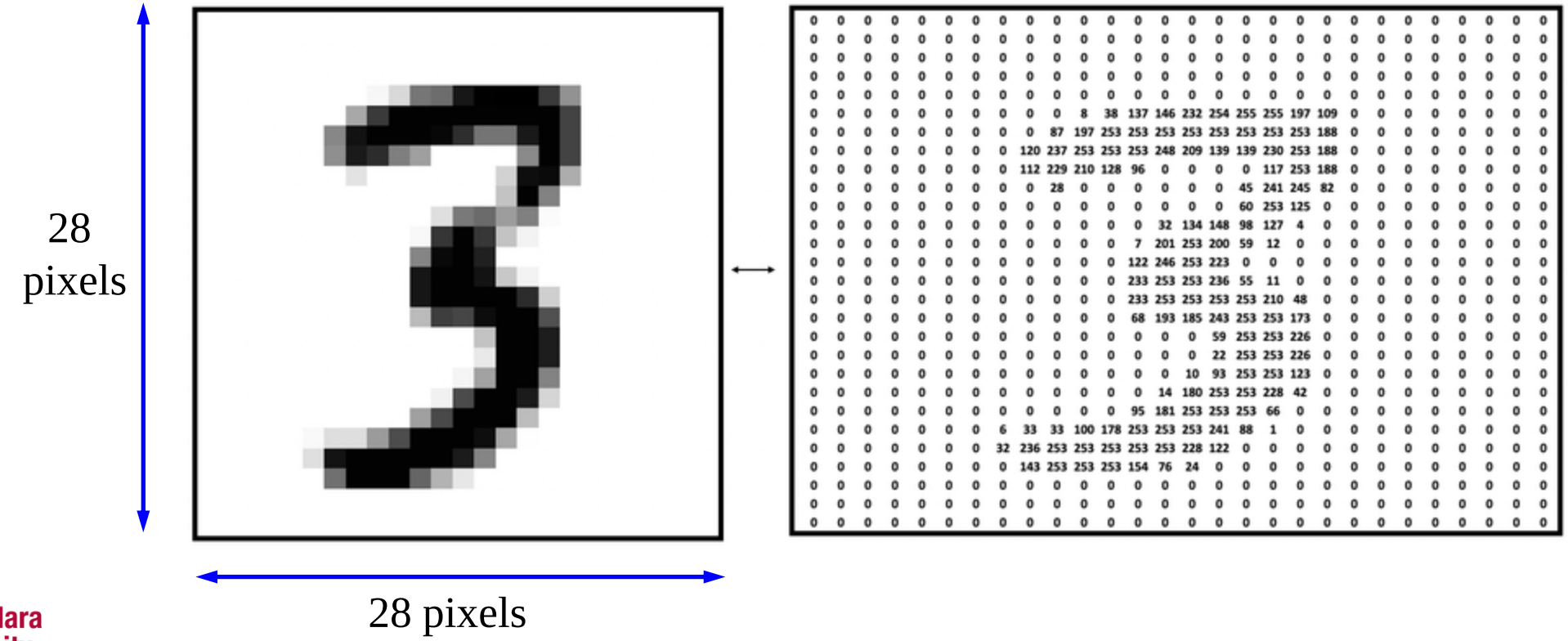
MNIST Dataset (1)

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

- The MNIST dataset (Modified National Institute of Standards and Technology) is an extensive collection of black and white images (28x28 pixels) of handwritten digits.
- The dataset contains 60,000 training and 10,000 testing images.
- The MNIST dataset is the most common benchmark for establishing a performance reference of a neural network on image recognition.

MNIST Dataset (2)

- Each MNIST image is a black-and-white picture of 28x28 pixels.
- Pixel values of a black-and-white picture vary in a range of [white, black] or [0, 255]



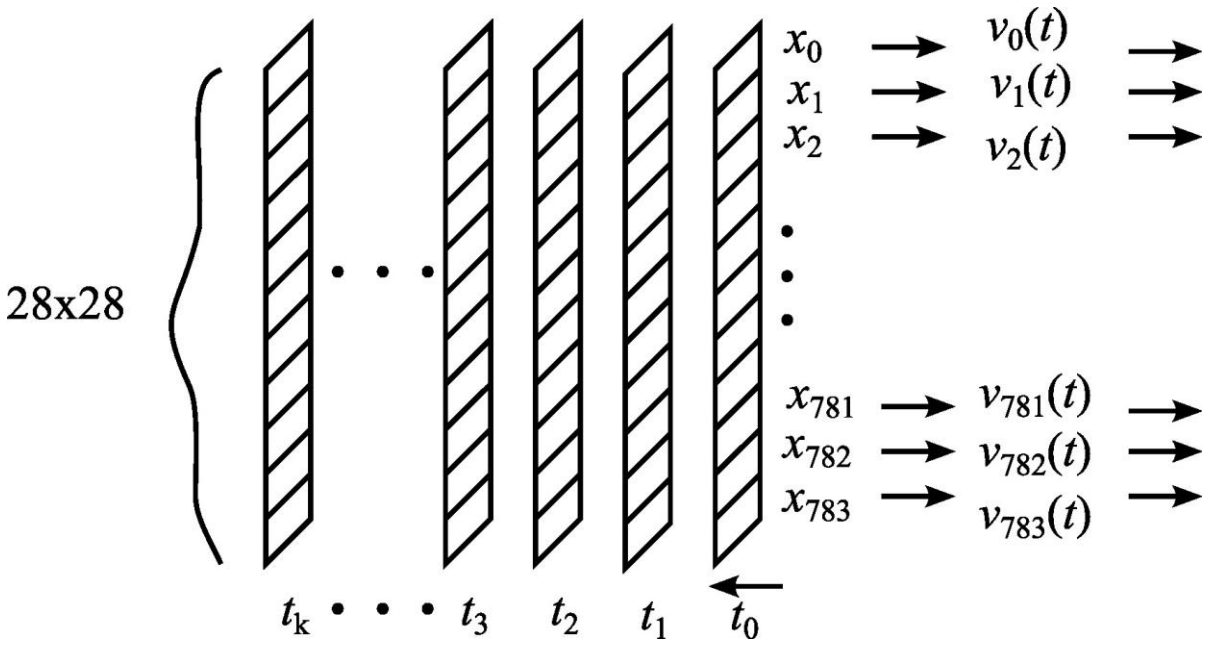
MNIST Dataset (3)

Digit	0	1	2	3	4	5	6	7	8	9	Total
Training	2,923	6,742	5,958	6,131	5,842	5,421	5,918	6,265	5,851	5,949	60,000
Testing	980	1,135	1,032	1,010	982	892	958	1,028	974	1,009	10,000

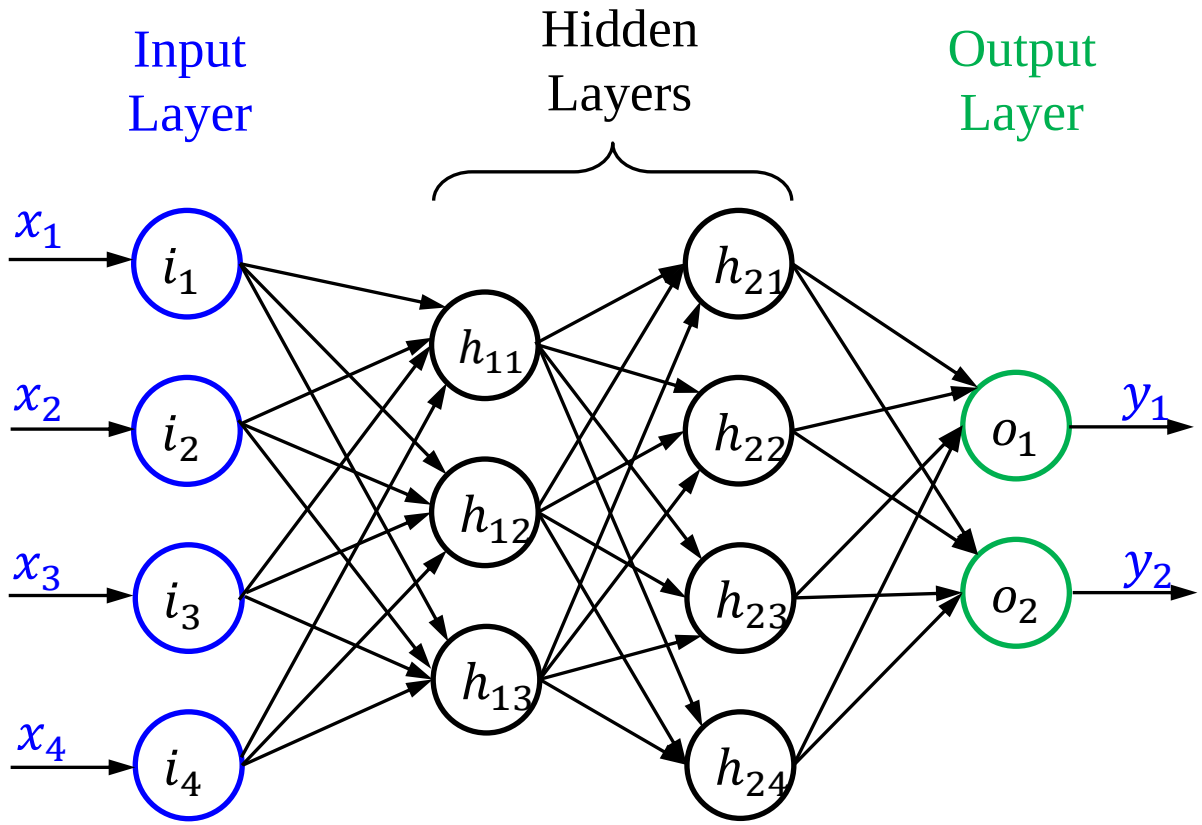
https://git-disl.github.io/GTDLBench/datasets/mnist_datasets/

MNIST Dataset (4)

- MNIST is a spatial dataset.



- Pixels are flattened into 784-length vectors and scaled from 0 to 1 (÷ 255.0).



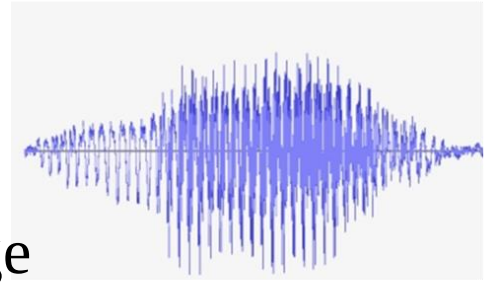
Artificial Neural Network (ANN)

Isolated Spoken Digits (1)

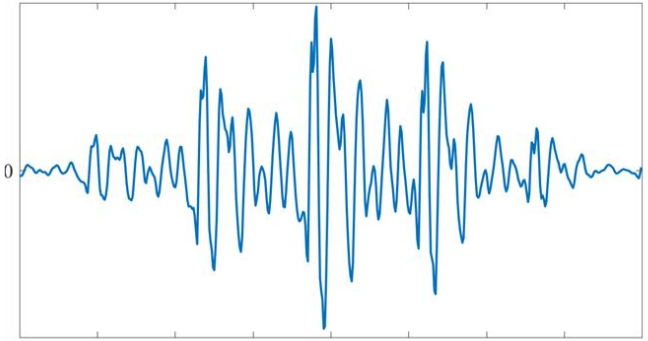
- With the widespread growth in the use of digital electronic objects, the need to communicate with these devices has increased, especially through human-friendly instructions such as uttering keywords like spoken digits.
- TI-46 digit corpus is often used for isolated-word automatic speech recognition (licensed dataset).
- Use a non-license digit dataset created by Jackson (https://github.com/Jakobovski/free-spoken-digit-dataset).
- The dataset contains the sound recordings of spoken digits at 8kHz
- The dataset has 1,500 recordings of digits 0 to 9 from various English speakers.
- The dataset is divided into two non-overlapping sets: 1,000 digits for training and 500 digits for testing.

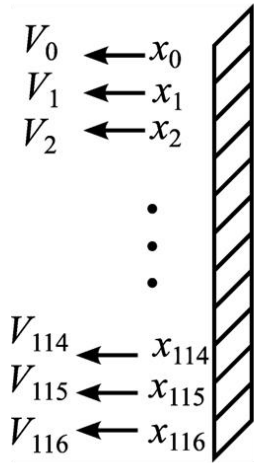
Isolated Spoken Digits (2)

- Preprocessing sound recordings using Mel-frequency Cepstral Coefficients.
- Use the Python package (https://pypi.org/project/python_speech_features/)
- Isolated spoken digit dataset is both temporal and spatial input data.



Mel-frequency cepstral Coefficients (MFCC)



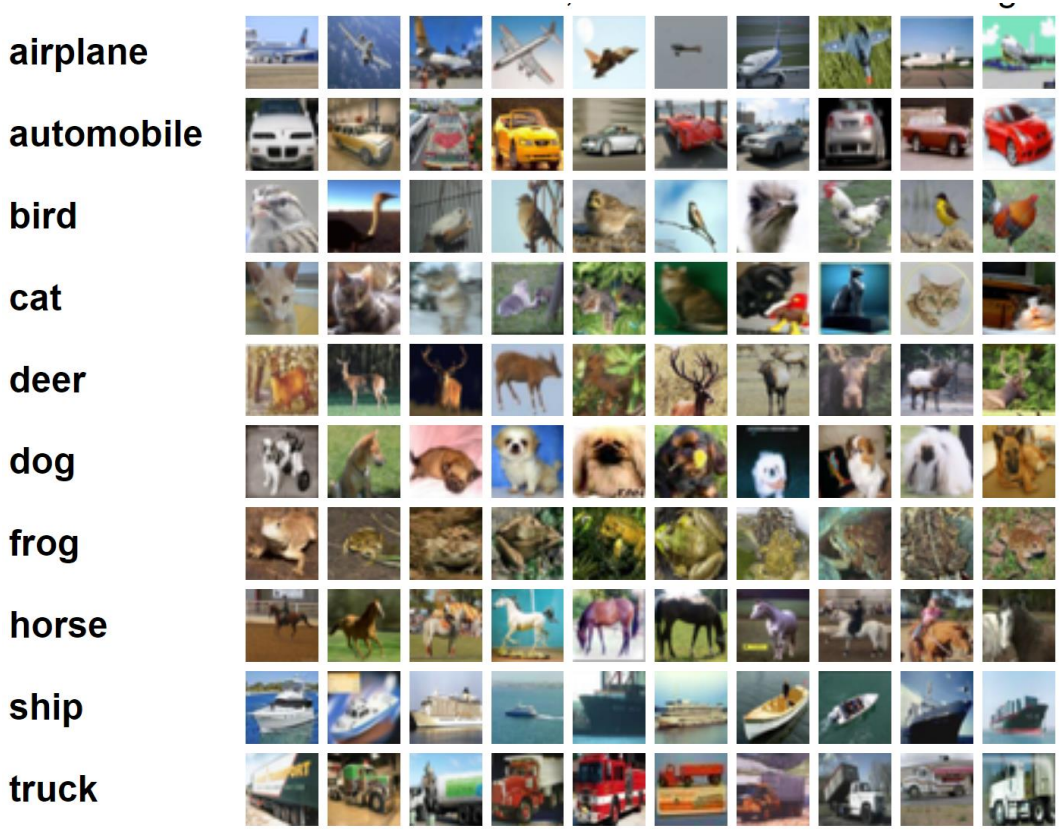


117-coefficient vectors

delta and delta-delta coefficients

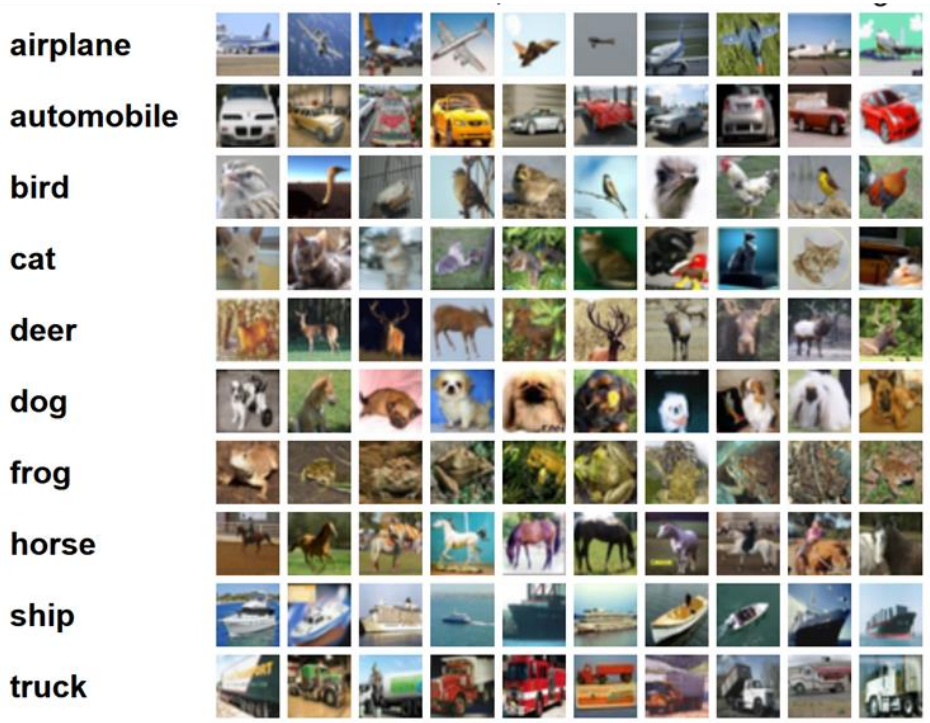
CIFAR-10 (1)

- The CIFAR-10 dataset (Canadian Institute For Advanced Research) is a collection of color images with a resolution of 32x32 pixels (https://www.cs.toronto.edu/kriz/cifar.html).
- The dataset has 50000 training images and 10000 test images of 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.
- CIFAR-100 is an extension of CIFAR-10 with 50000 training images and 10000 test images for 100 classes.

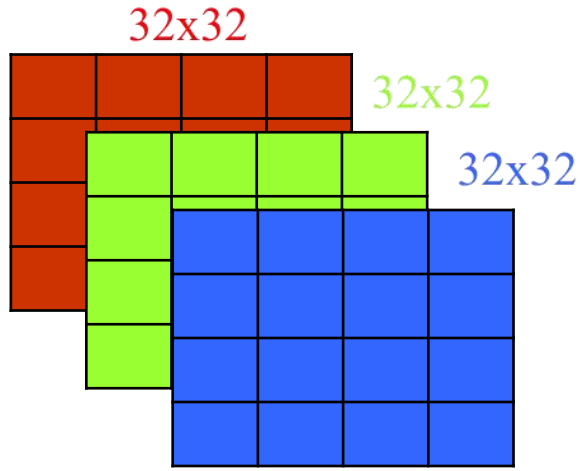


https://www.cs.toronto.edu/~kriz/cifar.html

CIFAR-10 (2)



https://www.cs.toronto.edu/~kriz/cifar.html



Convert color to grayscale

32x32 = 1024
High-dimensional network

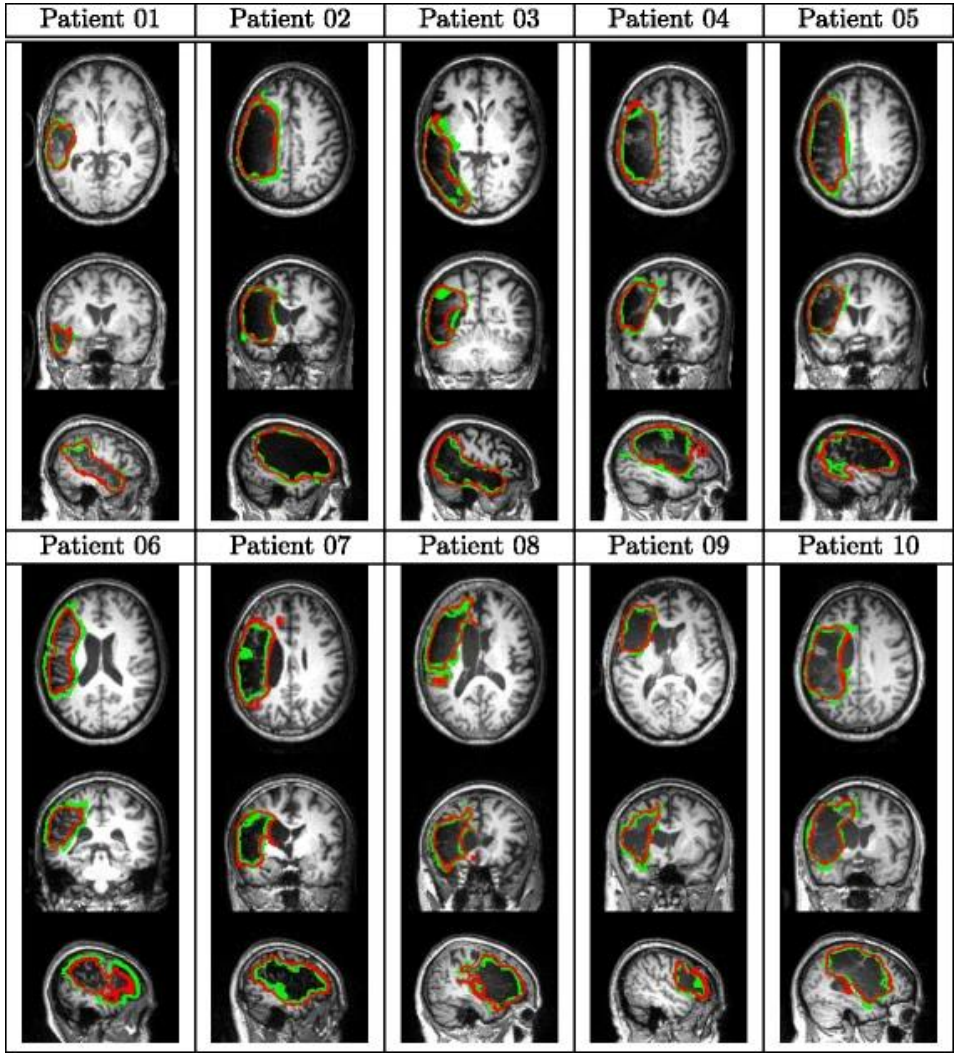
32x32x3 = 3072
Very high-dimensional network !!!

ImageNet

- ImageNet is a publicly available large-scale database with annotated images, designed for use in multiple computer vision tasks (https://www.image-net.org/).
- It contains over 14 million images, with each image annotated using WordNet synonym sets. It is one of the largest resources available for training deep learning models in object recognition tasks.
    - Over 14 million images in high resolution (~ 469x387 pixels).
    - Around 22000 WordNet synonym sets (also known as synsets).
    - A synset is a phrase that describes a meaningful concept in WordNet and ImageNet.
    - Over one million annotated images with bounding boxes.
    - 10,000+ synsets with scale-invariant feature transform (SIFT) features.
    - Over 1.2 million images with SIFT features.

Convolution Neural Network (1)

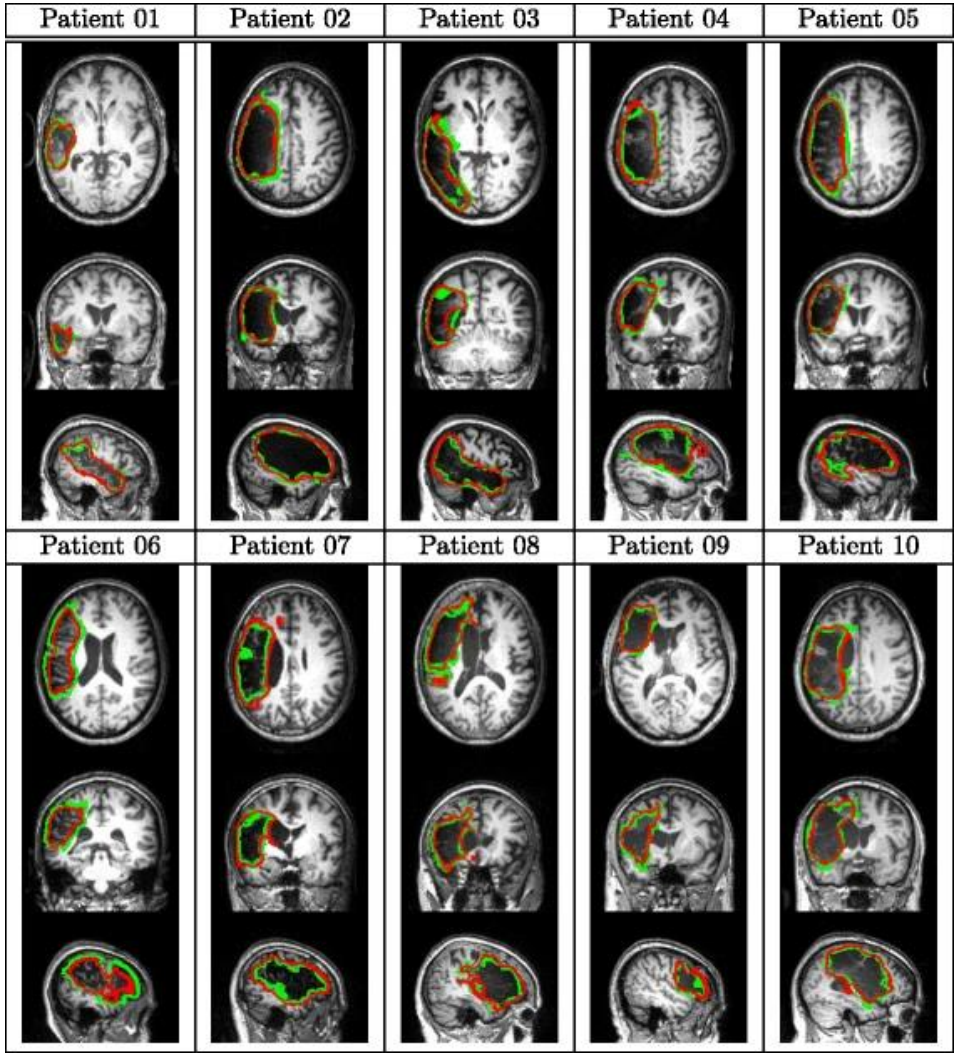
- Convolutional neural networks (CNNs) have been a dominant method in computer vision tasks on the object recognition such as the ImageNet Large Scale Visual Recognition Competition (ILSVRC).
- CNNs has been extremely helpful for medical applications:
  - Lesion detection: detecting an abnormal radiologic sign on MRI or CT scans obtained using radiocontrast.



https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0092-x

Convolution Neural Network (2)

- Convolutional neural networks (CNNs) have been a dominant method in computer vision tasks on the object recognition such as the ImageNet Large Scale Visual Recognition Competition (ILSVRC).
- CNNs has been extremely helpful for medical applications:
  - Lesion detection: detecting an abnormal radiologic sign on MRI or CT scans obtained using radiocontrast.

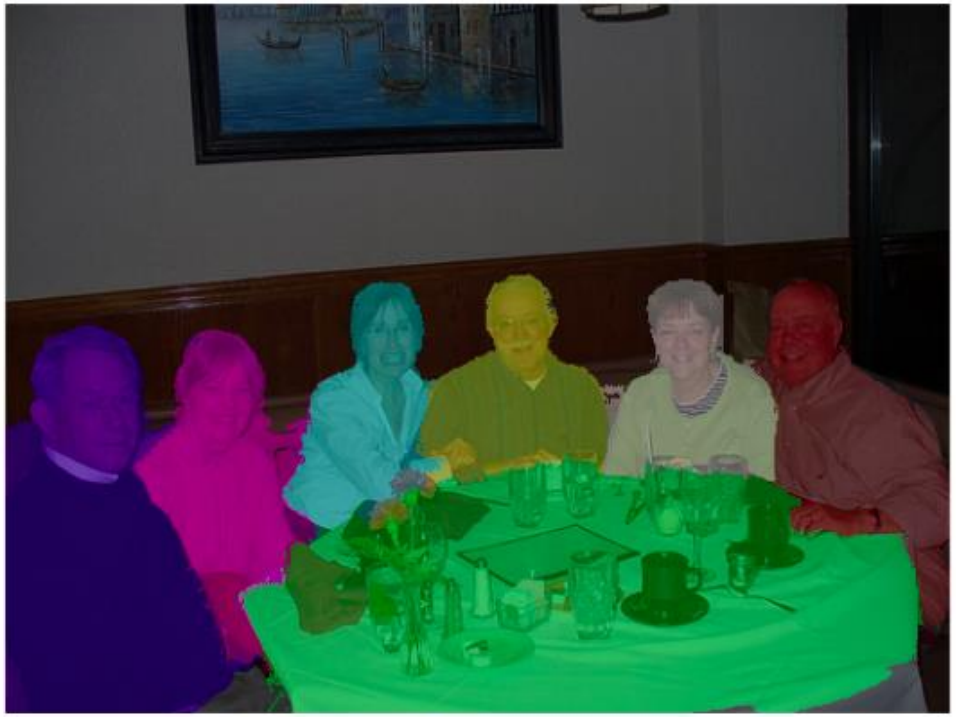


https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0092-x

Convolution Neural Network (3)

- Image Segmentation: the process of dividing an image into multiple segments, where pixels are associated with an object type

Semantic Segmentation

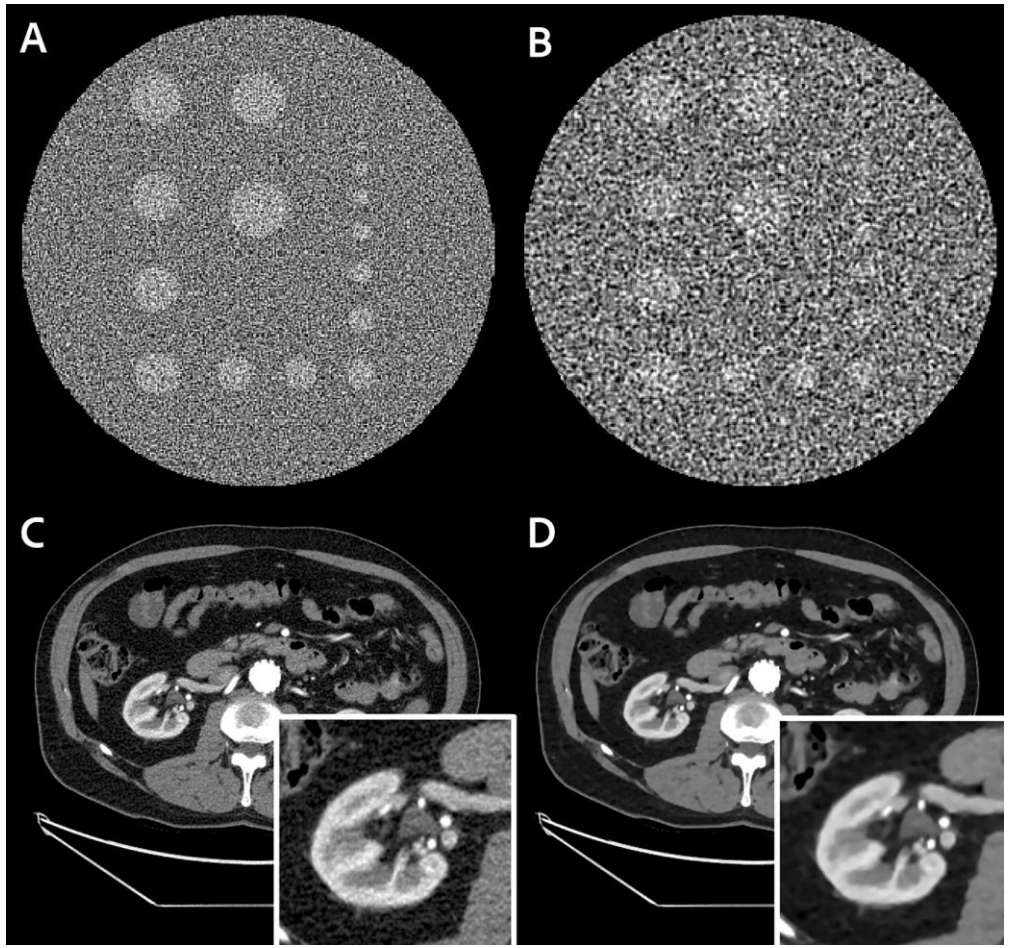


Instance Segmentation

https://neptune.ai/blog/image-segmentation

Convolution Neural Network (4)

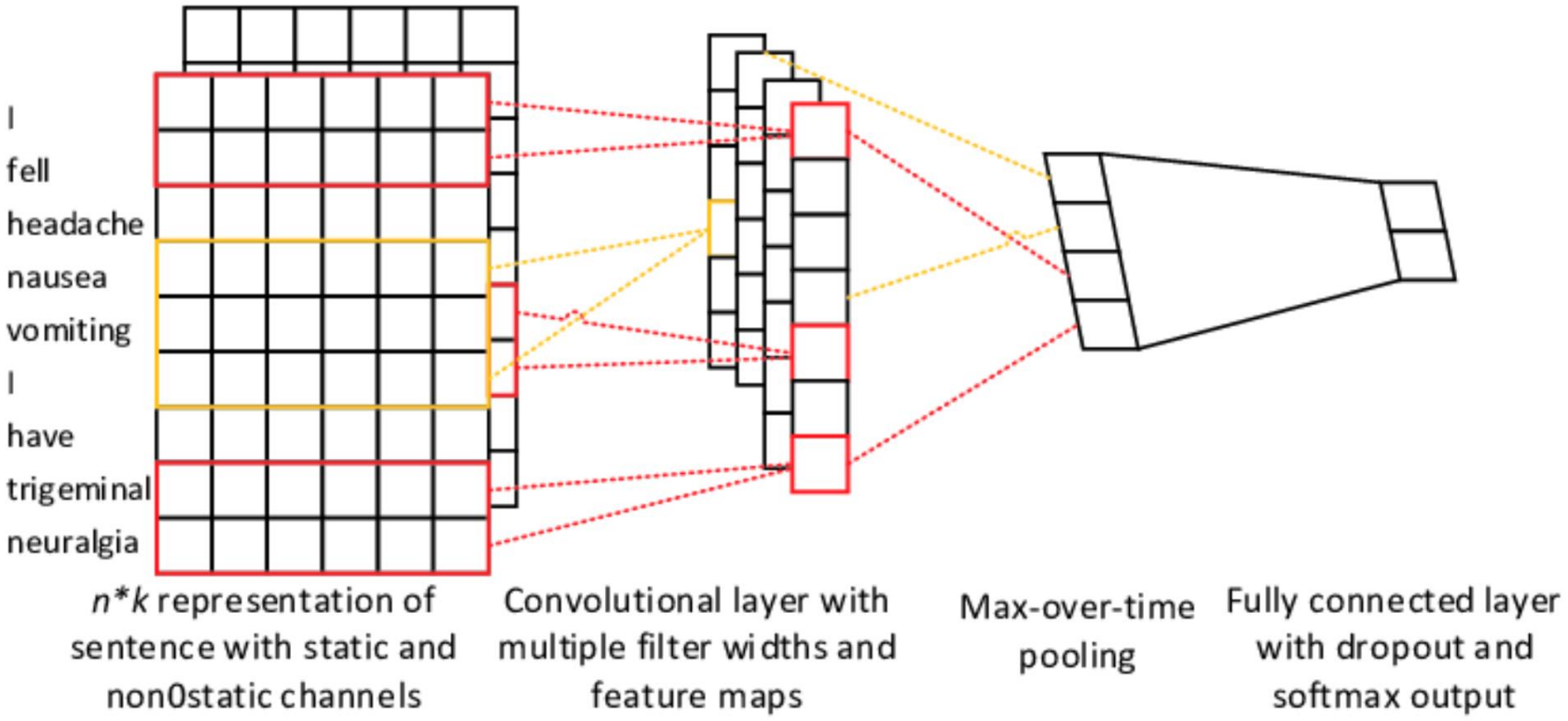
- Image reconstruction is a mathematical process that enhances images from X-ray projection data acquired at many different angles around patients.
- In computed tomography (CT), reconstructing and improving image quality can be translated into a reduction in radiation dose because images of the same quality can be reconstructed at a lower dose.



https://pubs.rsna.org/doi/10.1148/radiol.221257

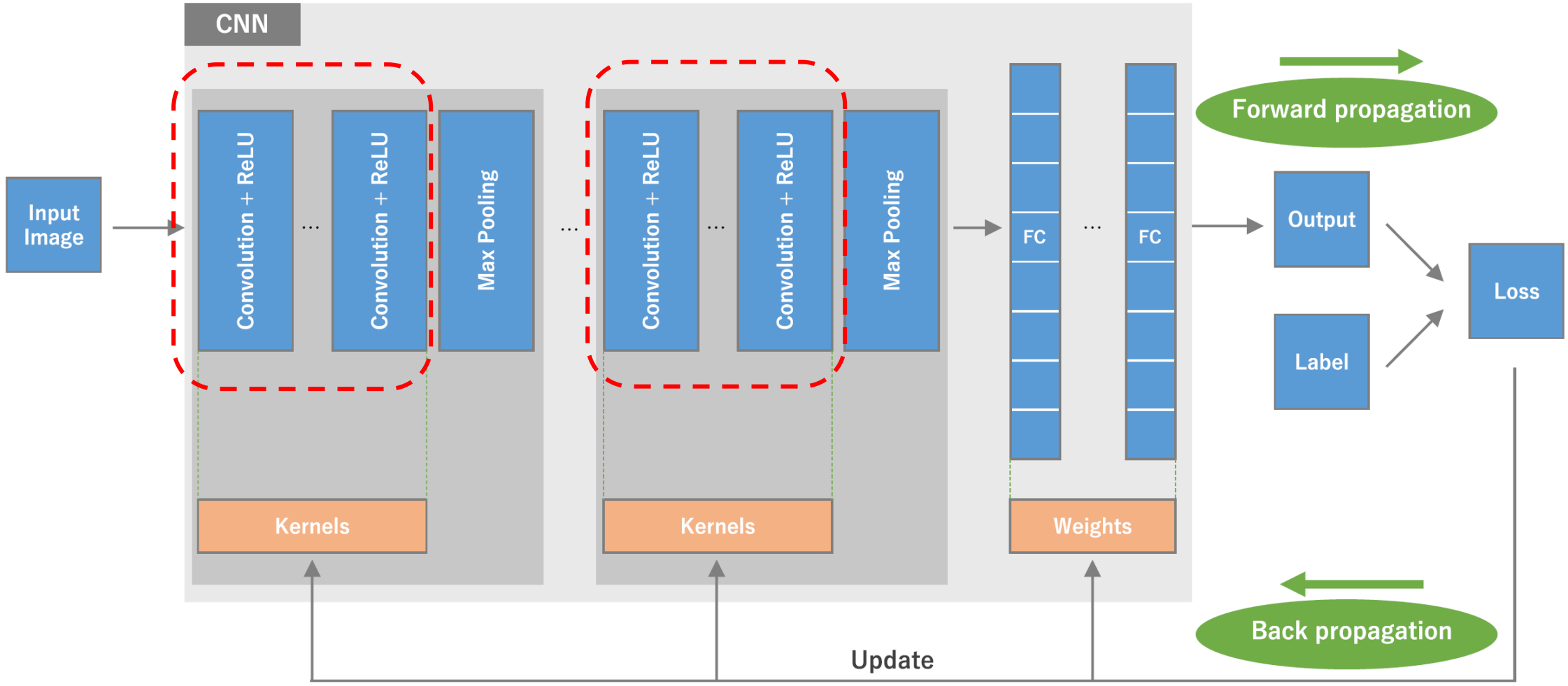
Convolution Neural Network (5)

- Natural Language Processing: A convolution is a window that slides over a larger input data set, emphasizing a subset of the input matrix (each row is vector that represents a word).



https://www.researchgate.net/figure/Architecture-of-CNN-for-Text-Classification_fig4_305695642

Convolution Neural Network (6)

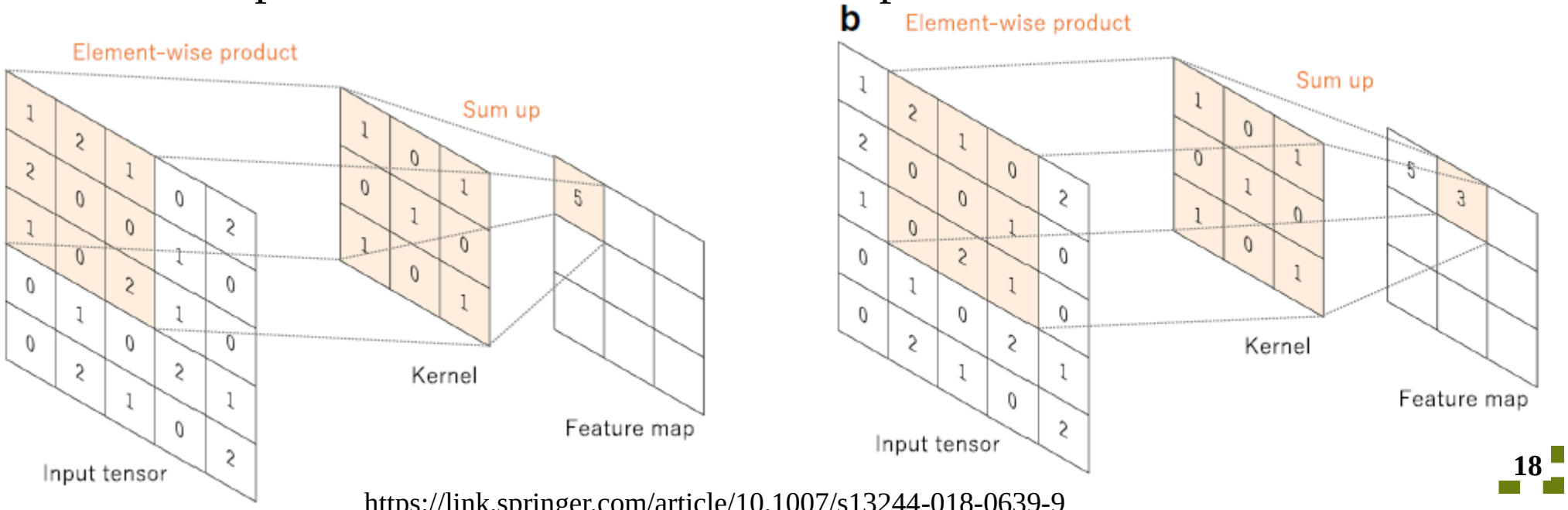


Yamashita, Rikiya, et al. "Convolutional neural networks: an overview and application in radiology." Insights into imaging 9 (2018): 611-629.

Convolution Layer (1)

Building blocks of CNNs:

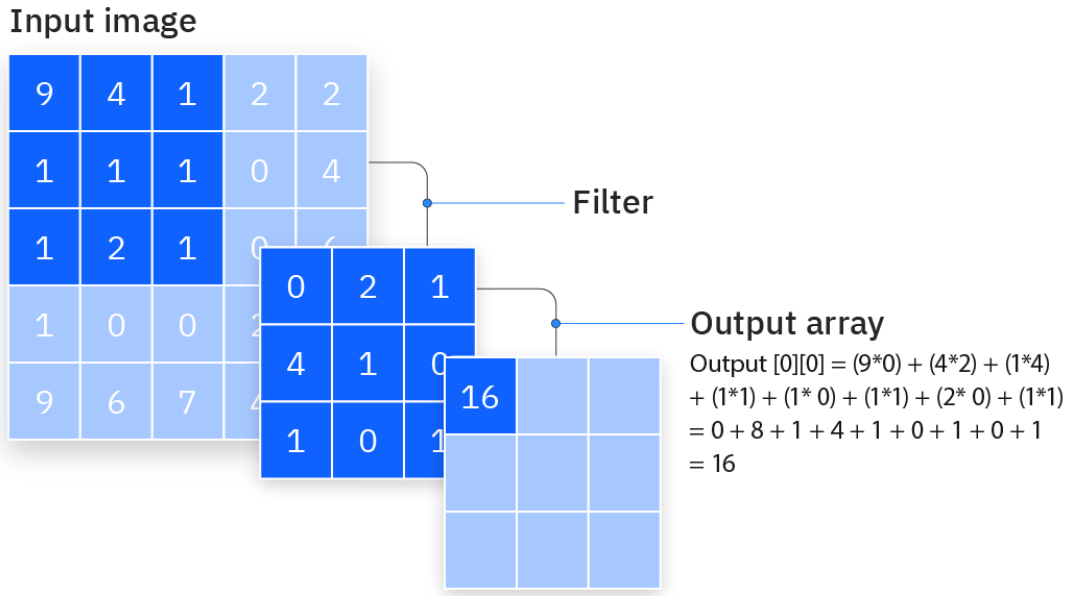
- Convolution Layer: This layer performs feature extractions, which involve a combination of linear and nonlinear convolution operations. In this layer, we look at the image through smaller sections and move a filter (kernel) (3x3) over the image, finding features in a particular section → feature maps.



https://link.springer.com/article/10.1007/s13244-018-0639-9

Convolution Layer (2)

- The number of filters affects the depth of the output: 3 distinct filters ⇒ three different feature maps ⇒ a depth of three.
- Stride is the distance, or number of pixels, where the filter moves over the input matrix.
- Zero-padding is usually used when the filters do not fit the input image. This sets all elements that fall outside of the input matrix to zero.
- Kernel or filter weights are updated along with full-connected network weights using the backpropagation technique.

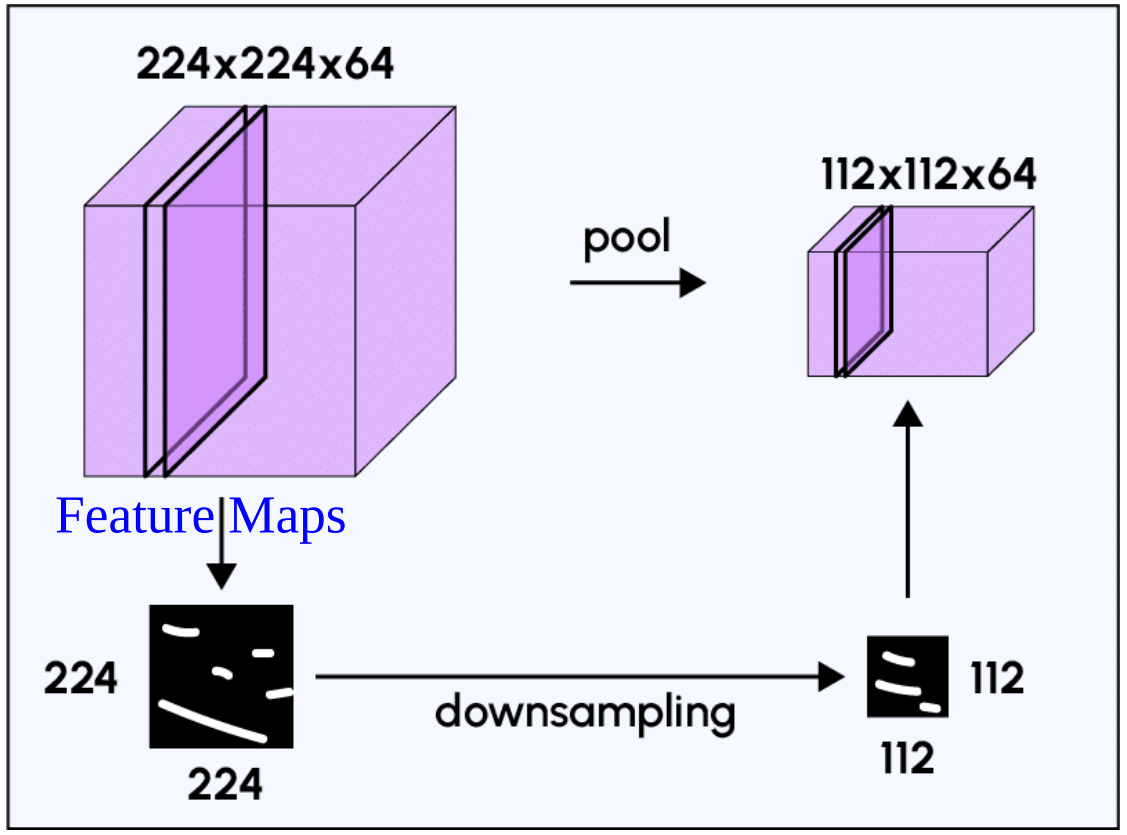


A kernel size of 3×3, no padding, and a stride of 1

Max-pooling Layer (1)

Max-pooling Layer:

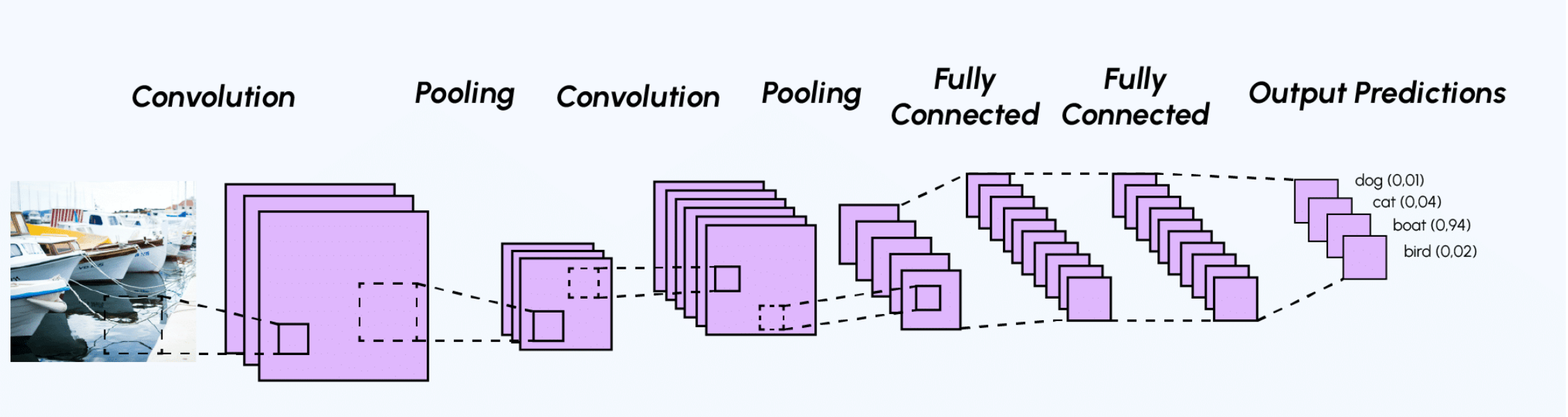
- The pooling layer is typically applied between two convolutional layers.
- It takes the feature maps generated by the convolutional layer and reduces the size of the images while preserving their essential characteristics.
- Pooling methods are max-pooling and average pooling, which calculates the average value of the filter window at each step.



https://datascientest.com/en/convolutional-neural-network-everything-you-need-to-know

Max-pooling Layer (2)

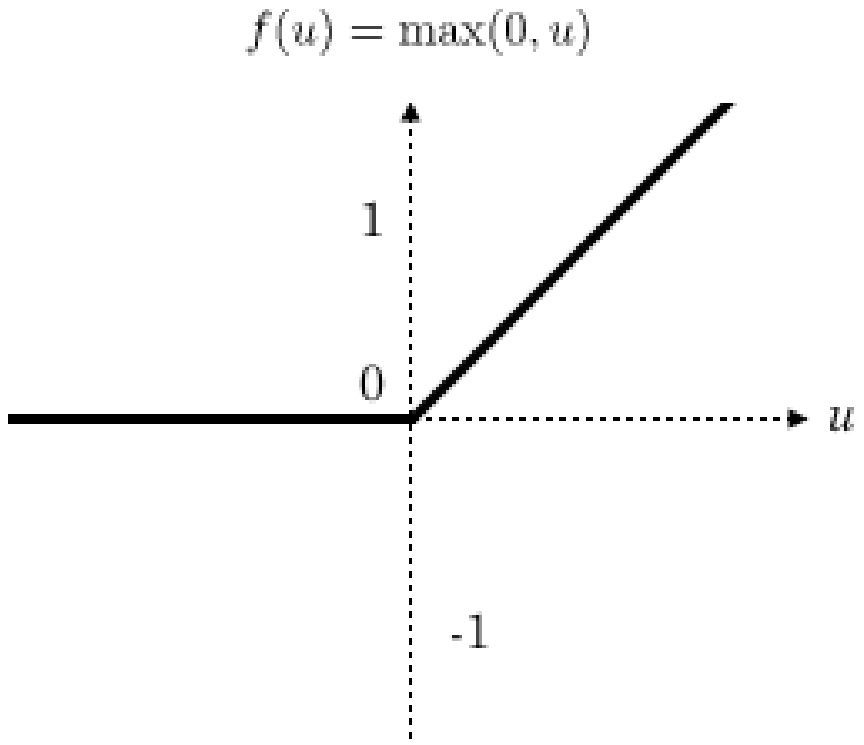
- Max-pooling Layers can be more than 1 depending on the architecture of a CNN.



https://datascientest.com/en/convolutional-neural-network-everything-you-need-to-know

Rectified Linear Units (ReLU) Activation

- The ReLU activation function replaces all negative input values with zeros.
- The purpose of these activation layers is to make the model nonlinear and, therefore, more complex.
- The final output of the pooling layer retains the same number of feature maps as the input but in a considerably compressed form.



Fully Connected (FC) Layers

- FC layers are placed at the end of the CNN architecture and are fully connected to all output neurons.
- After receiving an input vector, the FC layer applies a linear combination followed by an activation function, ultimately aiming to classify the input image (see the following diagram).
- In the end, it outputs a vector of size d, corresponding to the number of classes, where each component represents the probability of the input image belonging to a class.