Introduction to Neural Networks

Yasaman Amannejad,
Department of Mathematics and Computing,
Mount Royal University

Outline

- Introduction to neural networks
- Components of a neural network
  - Activation function
  - Loss function
  - Backpropagation
  - Stochastic gradient descent
- What is a deep learning network?

Activities

- In this session we will
  - Implementing a neural network classifier with Scikit-learn.
  - Implementing a neural network classifier with Keras.

ML vs. Programming

> Field of study that gives computers the ability to learn **without being explicitly programmed**.
> - Arthur Samuel, 1959

Traditional Programming

Program

If (eyes==2) & (legs == 4) & (tail == 1) ... then print “Cat”

Data

Computer

“Cat”

Output

Machine Learning

Output

“Cat”

Data

Computer

Program

Cat Recognition

Supervised vs. Unsupervised

Machine Learning

Supervised	Reinforcement Learning	Unsupervised

Output

Data

"Cat"

Computer

Program

Cat Recognition!

Data

Computer

Program

Grouping Cats!

Supervised Learning

Machine Learning

Supervised	Reinforcement Learning	Unsupervised

- Prediction
- Online Advertising

NN

- Photo Tagging

CNN

- Speech Recognition
- Language Translation

RNN

- Autonomous Driving

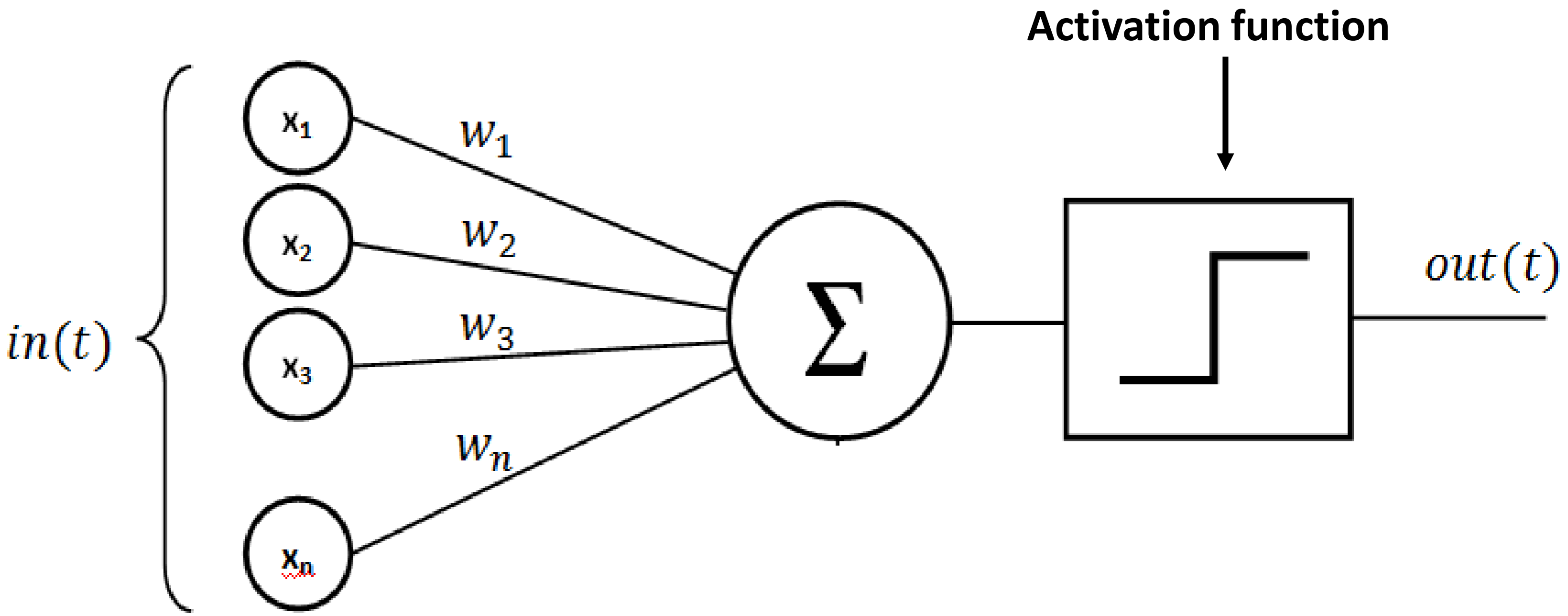
Hybrid NNs

Brief History of Neural Networks

- **1943:** McCulloch & Pitts show that neurons can be combined to construct a Turing machine (using ANDs, ORs, & NOTs)
- **1958:** Rosenblatt shows that perceptrons will converge if what they are trying to learn can be represented
- **1969:** Minsky & Papert showed the limitations of perceptrons, killing research for a decade
- **1985:** The backpropagation algorithm revitalizes the field
  - Geoff Hinton et al
- **2006:** The Hinton lab solves the training problem for DNNs

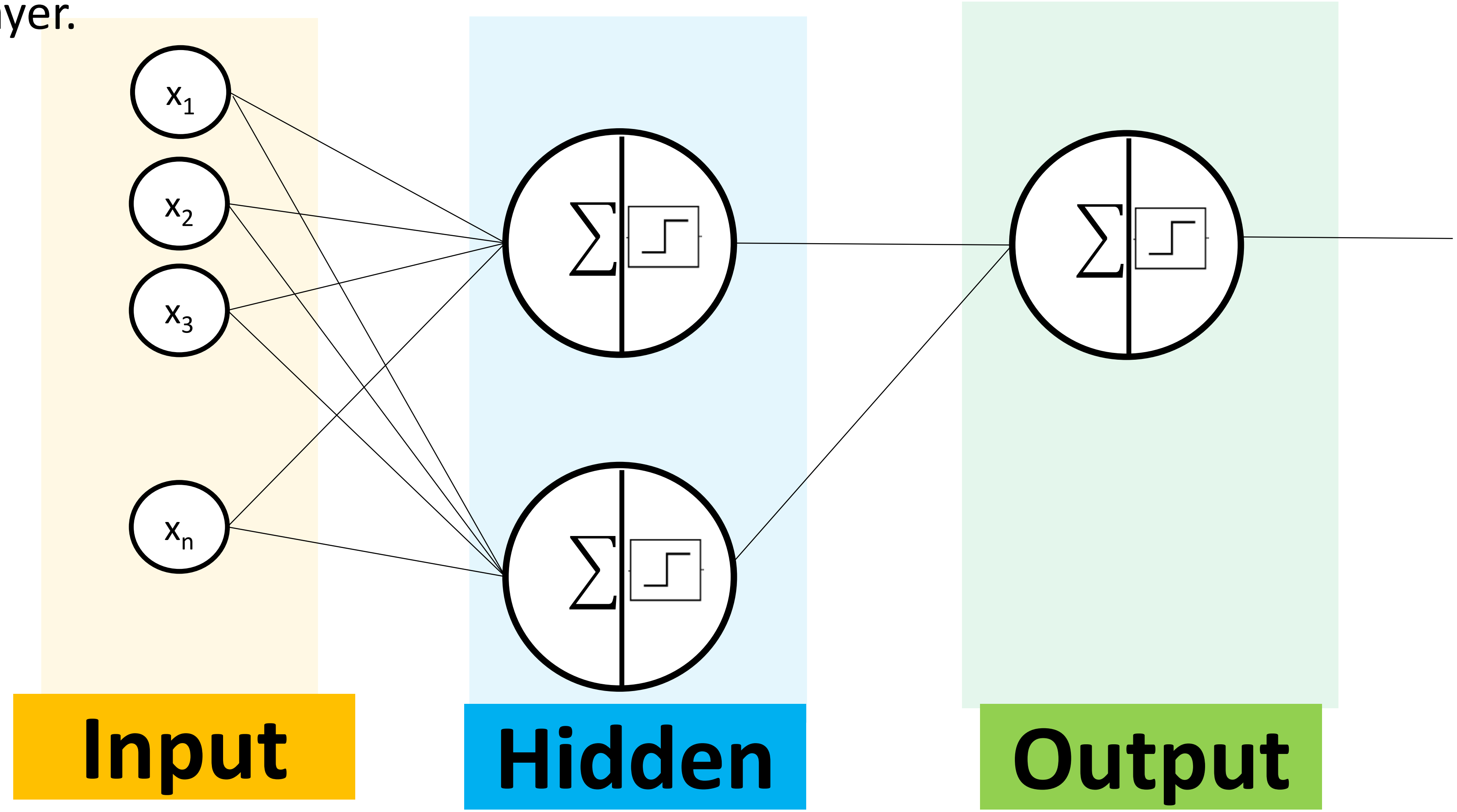
Fundamental of Neural Network

- Perceptron



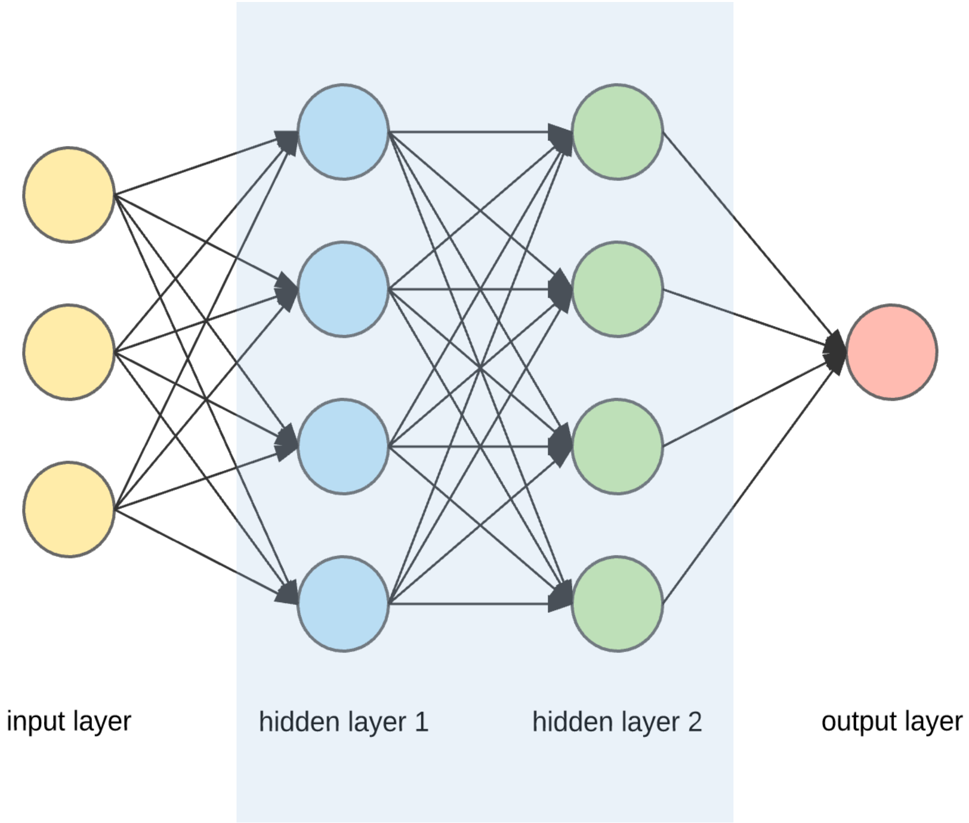
Multi Layer Perceptron (MLP)

- Consists of at least three layers of nodes: an input layer, a hidden layer and an output layer.



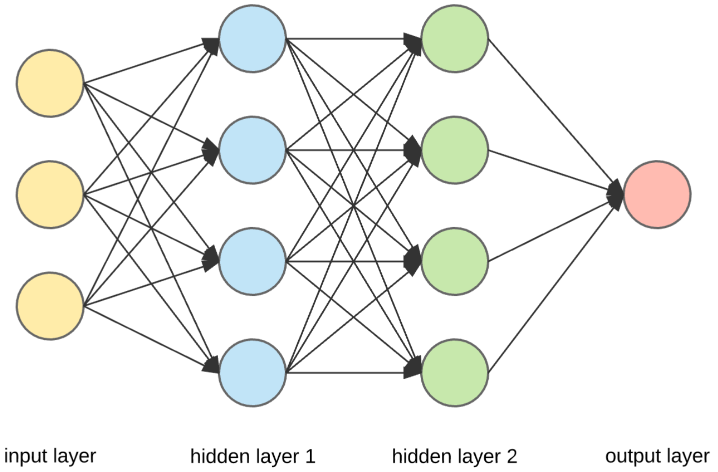
Neural Network (NN)

- Input Layer
- Hidden Layers
- Output Layer

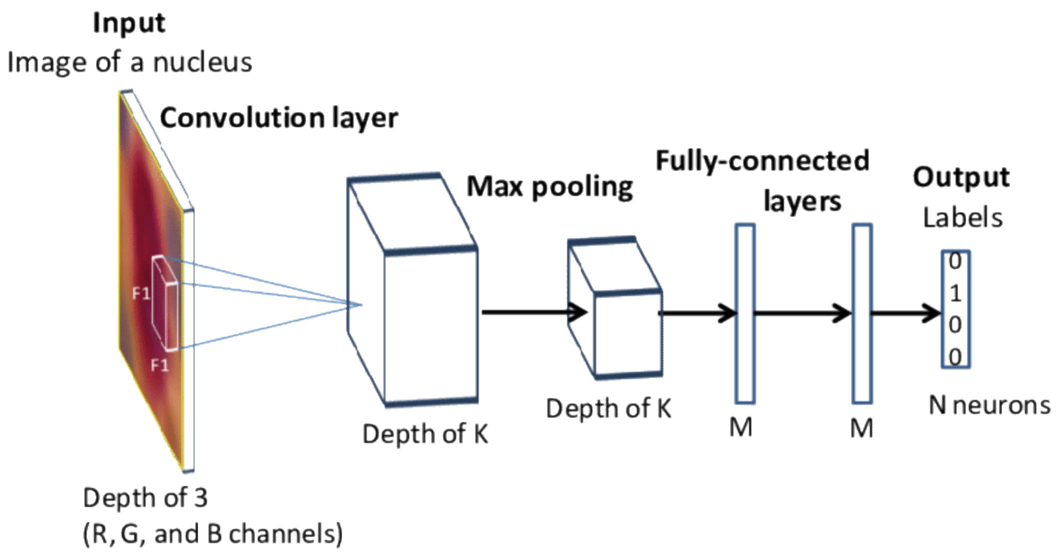


- Given enough data, NNs are good at finding functions that map input X to output Y.

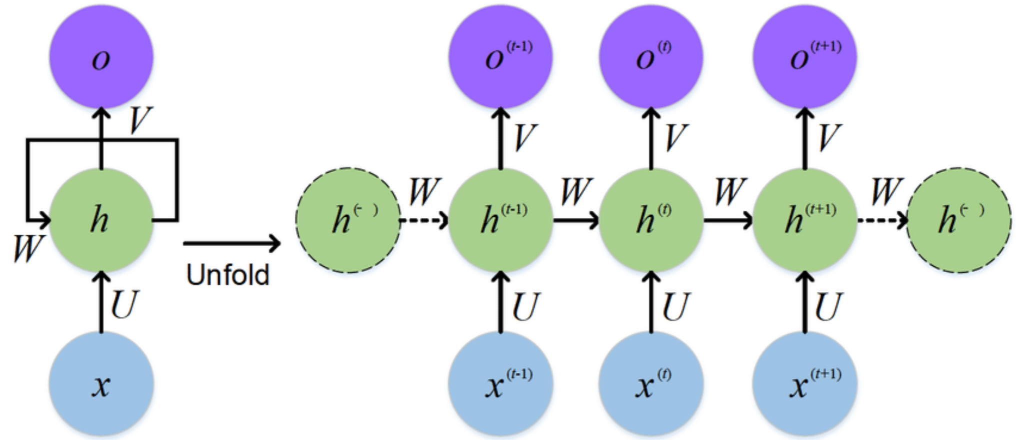
Different Type of Networks



NN



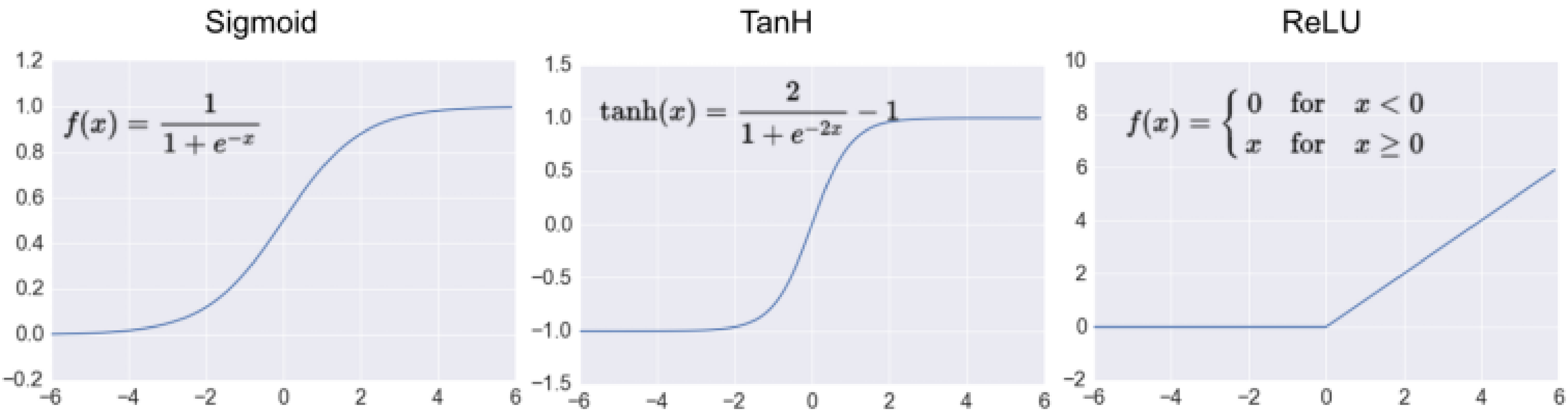
Convolutional NN (CNN)



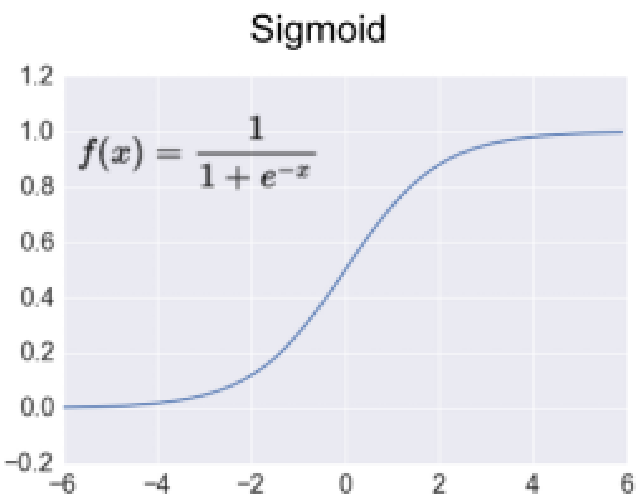
Recurrent NN (RNN)

Activation Functions

- Some well-known activation functions are:
  - Sigmoid or Logistic
  - Tanh — Hyperbolic tangent
  - ReLu -Rectified linear unit



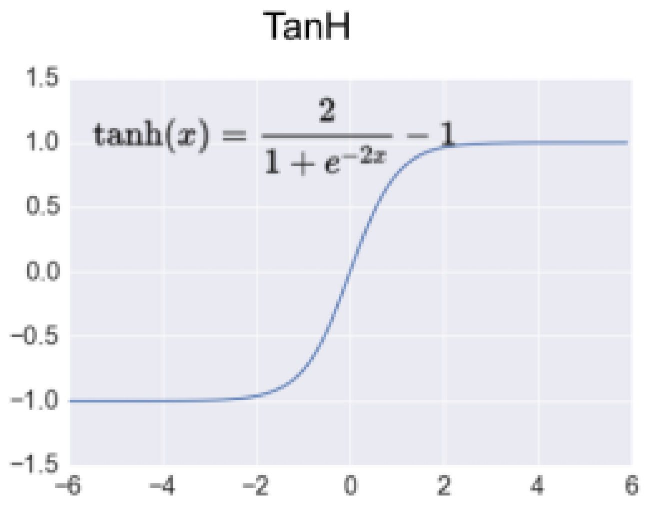
Sigmoid



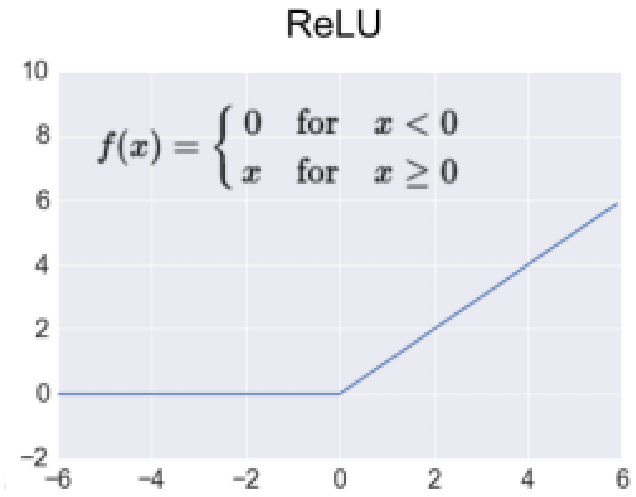
- Its Range is between 0 and 1.
- It is a S — shaped curve. It is easy to understand and apply

- **Issues:**
  - Vanishing gradient problem.
  - Output isn't zero centered. It makes the gradient updates go too far in different directions. **It makes optimization harder.**
  - Sigmoids saturate and kill gradients.
  - Sigmoids have slow convergence

Tanh — Hyperbolic Tangent

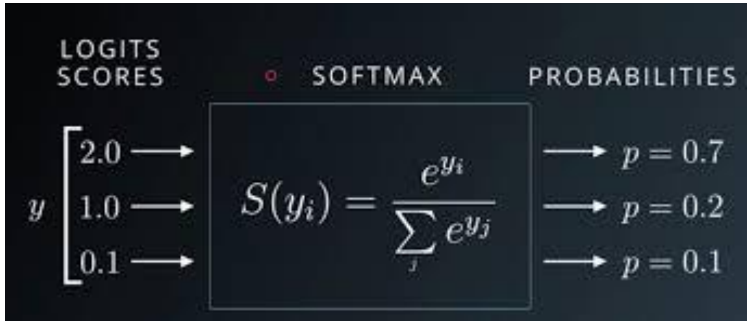


- it’s output is zero centered because its range in between -1 to 1 .
- Hence optimization is easier in this method hence in practice it is always preferred over Sigmoid function.
- But still it suffers from Vanishing gradient problem.



ReLu- Rectified Linear Units

- It was recently proved that it had 6 times improvement in convergence from Tanh function.
- Use it almost always except for the last layer of non-binary.
  - **Softmax** can be used for non-binary output layer.



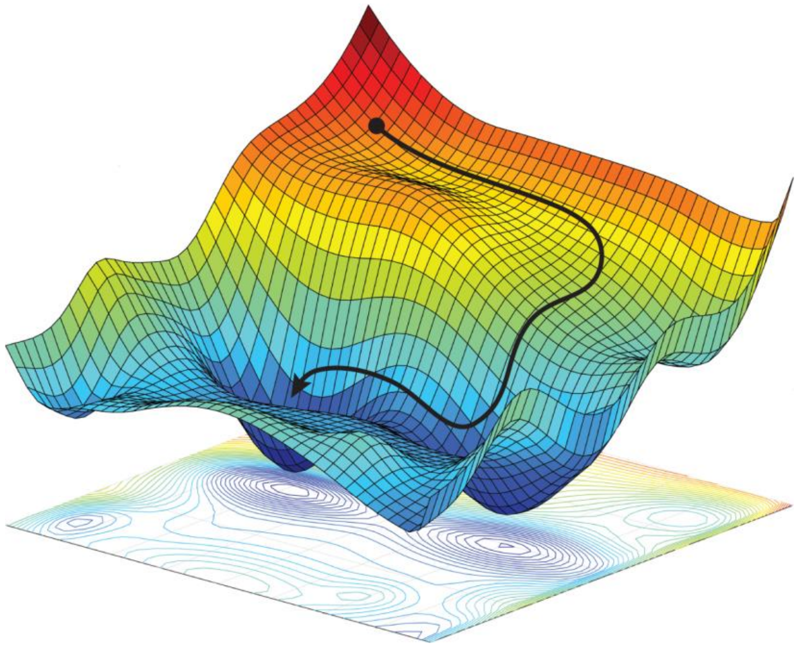
Loss Function

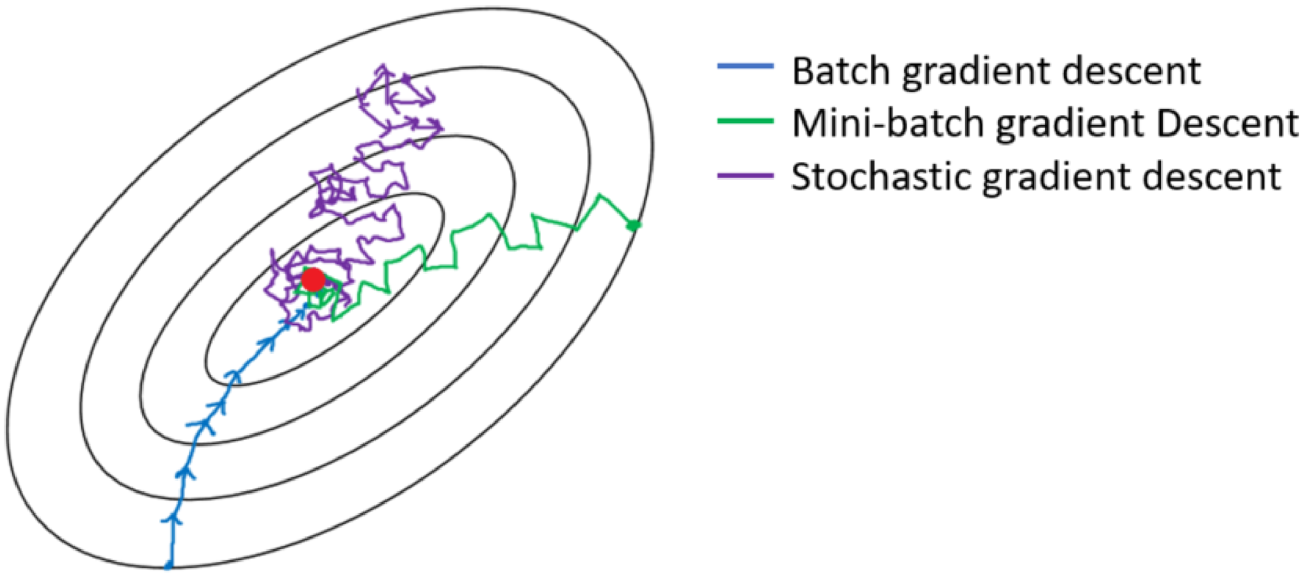
- Neural networks are trained using an optimization process that requires a **loss function** to calculate **the model error**.

- Cross-entropy and mean squared error are the two main types of loss functions to use.

- ".. Reduces all the various good and bad aspects of a possibly complex system down to a single number, a scalar value. .."
  - Page 155, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, 1999.

Stochastic Gradient Descent





If cost function is non-convex, **trying different initial random weights** can help to find the **global minimum** (e.g., by applying optimization techniques like PSO to find the best weights).

Gradient Descent

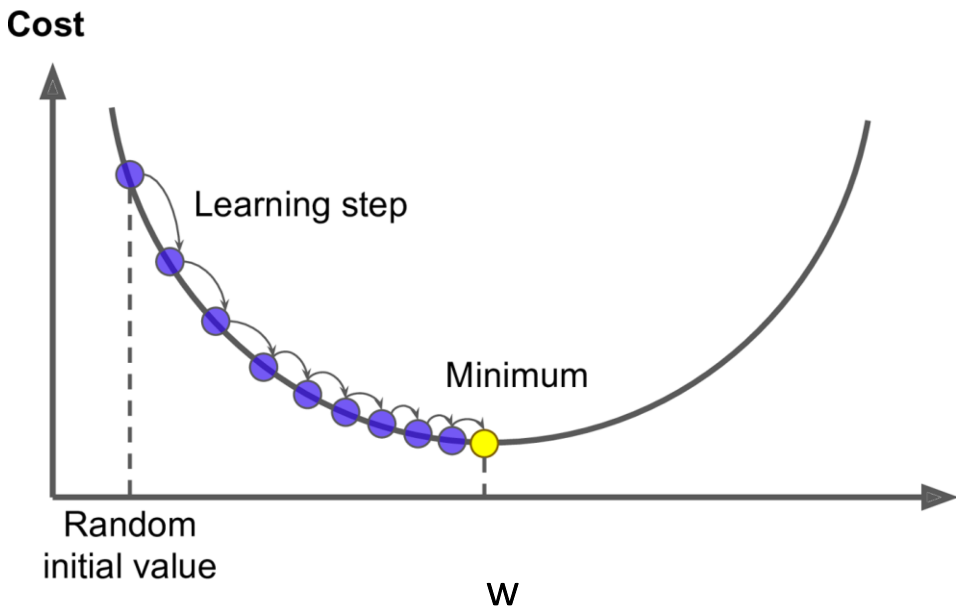
Repeat{

$$w := w - \alpha * \frac{d(J)}{d(w)}$$

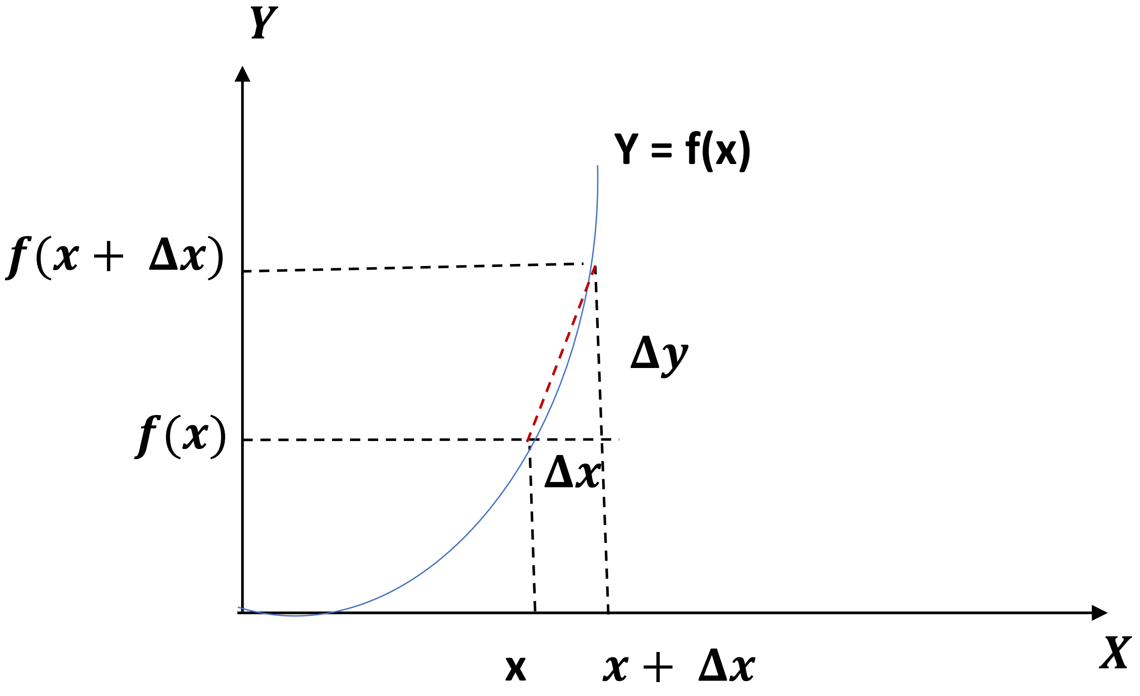
$$b := b - \alpha * \frac{d(J)}{d(b)}$$

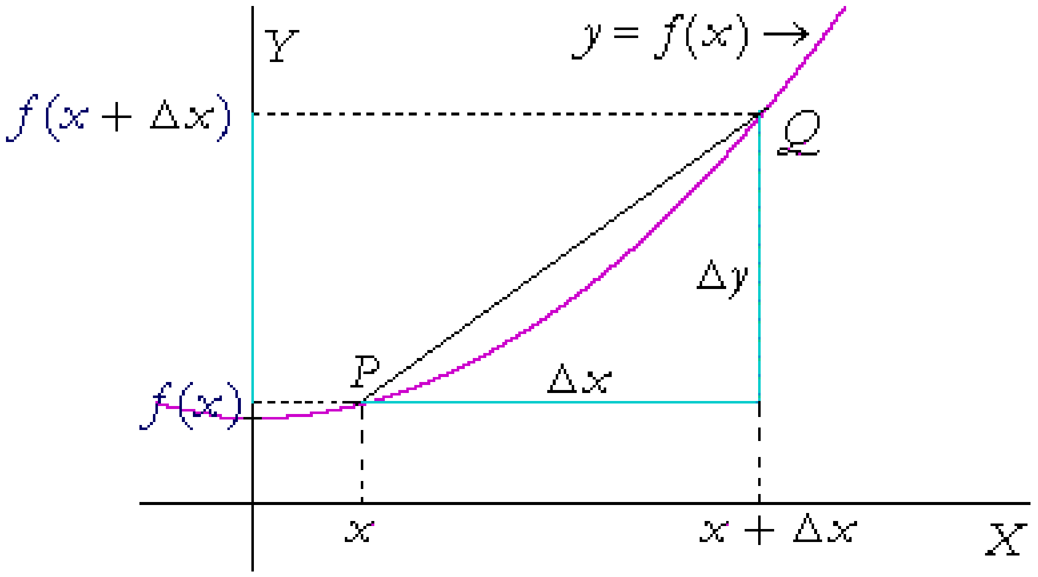
}

α : learning rate



Derivatives



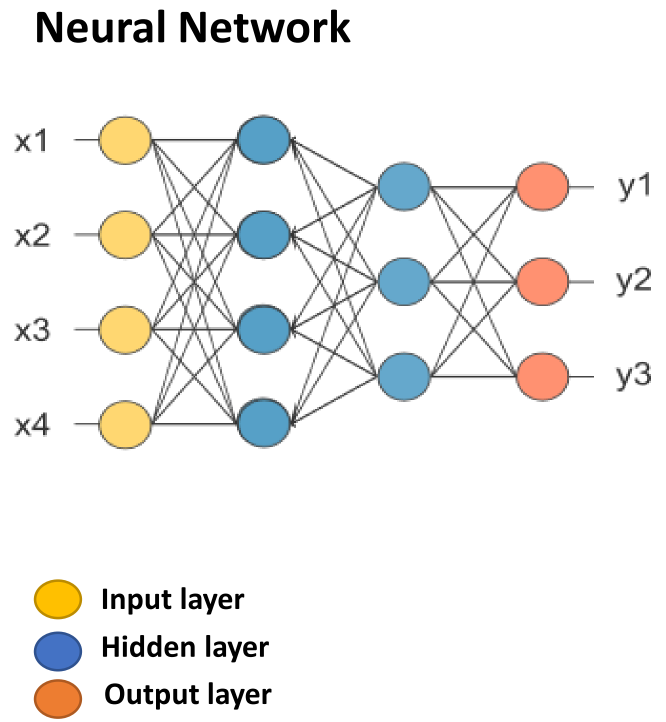


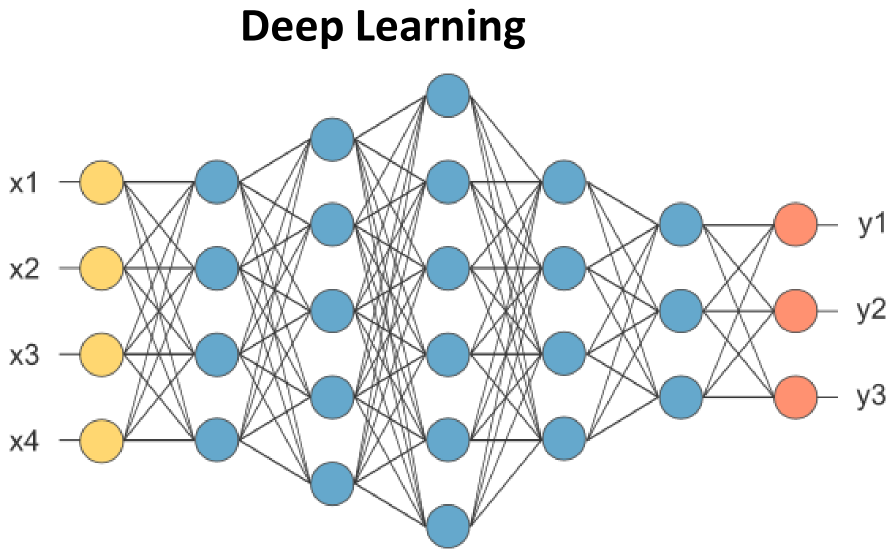
Some Terminology

- **Epoch:** one forward pass and one backward pass of *all* the training examples.
- **Batch size:** the number of training examples in one forward/backward pass. The higher the batch size, the more memory space you'll need.
- **Number of iterations** = number of passes, each pass using [batch size] number of examples.

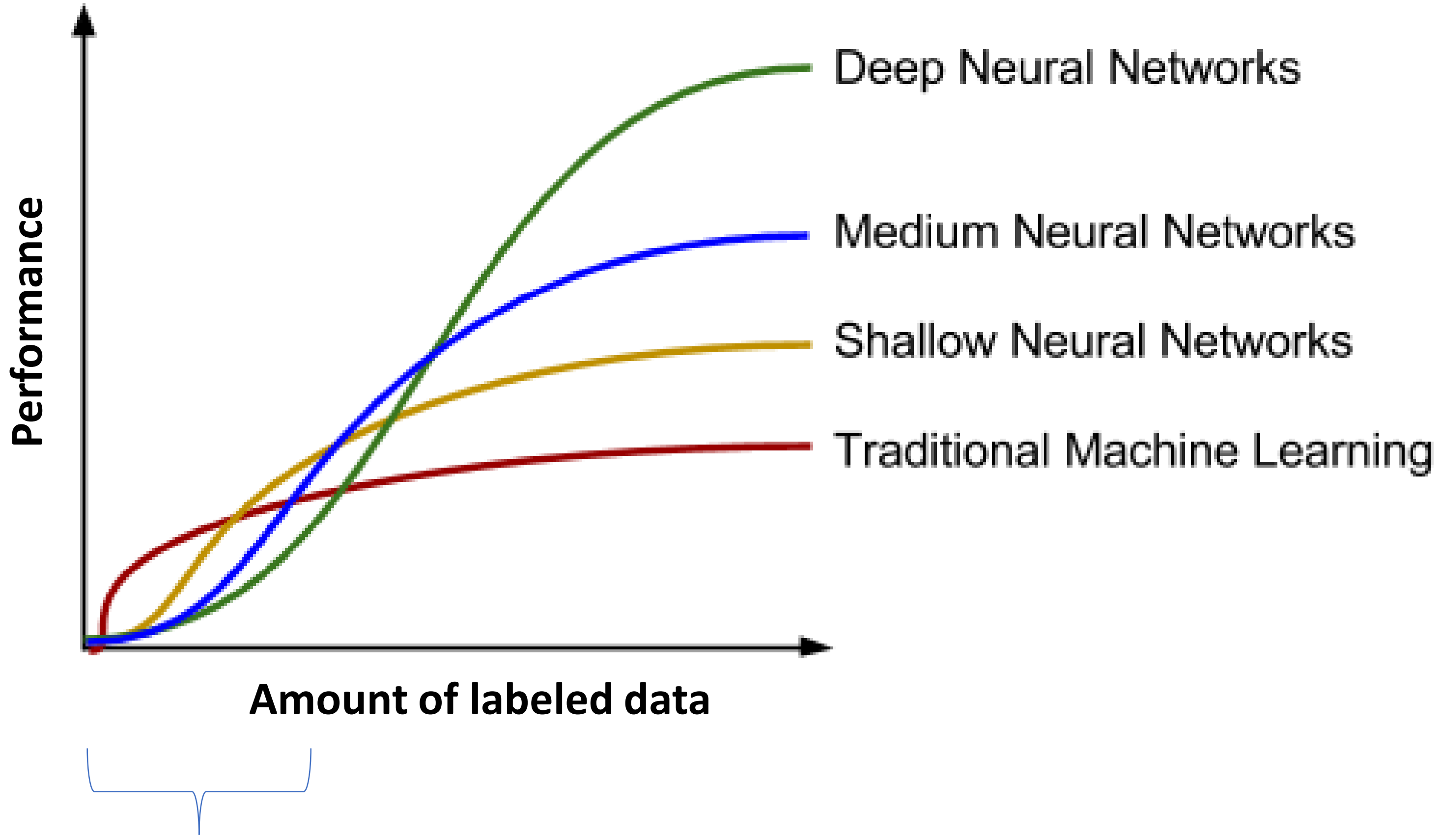
Deep Learning

The term "deep" refers to the number of hidden layers and the size of the layers in the network.





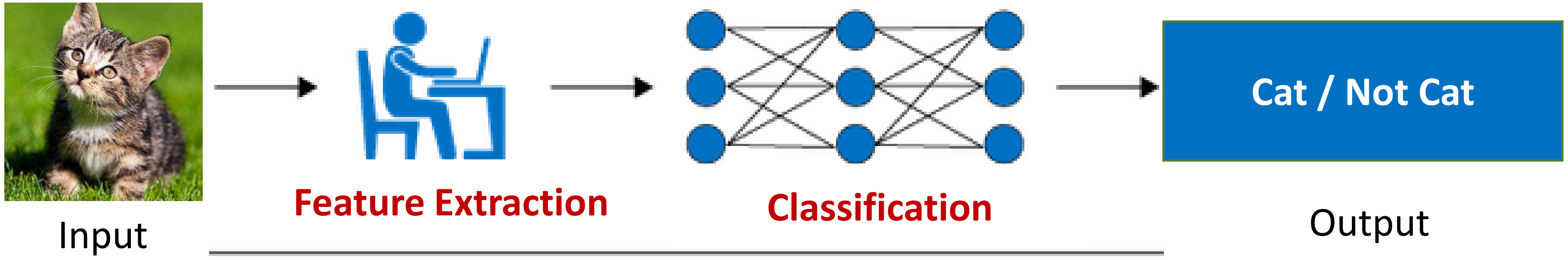
Deep Learning Progress



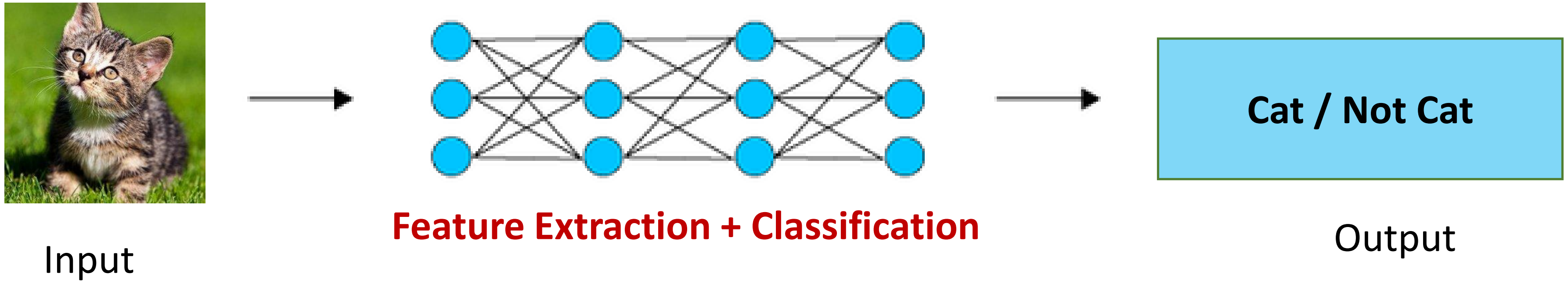
Small Data → Feature selection plays an important role!

DL vs. Classical ML

Classical Machine Learning

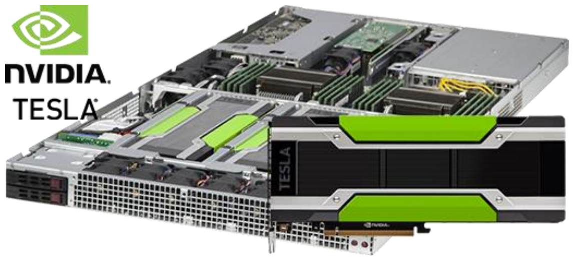


Deep Learning



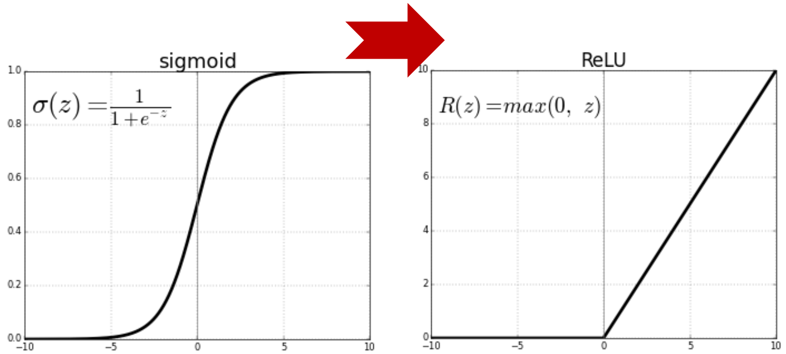
Drivers of Deep learning

- Big Data



- Faster Computation

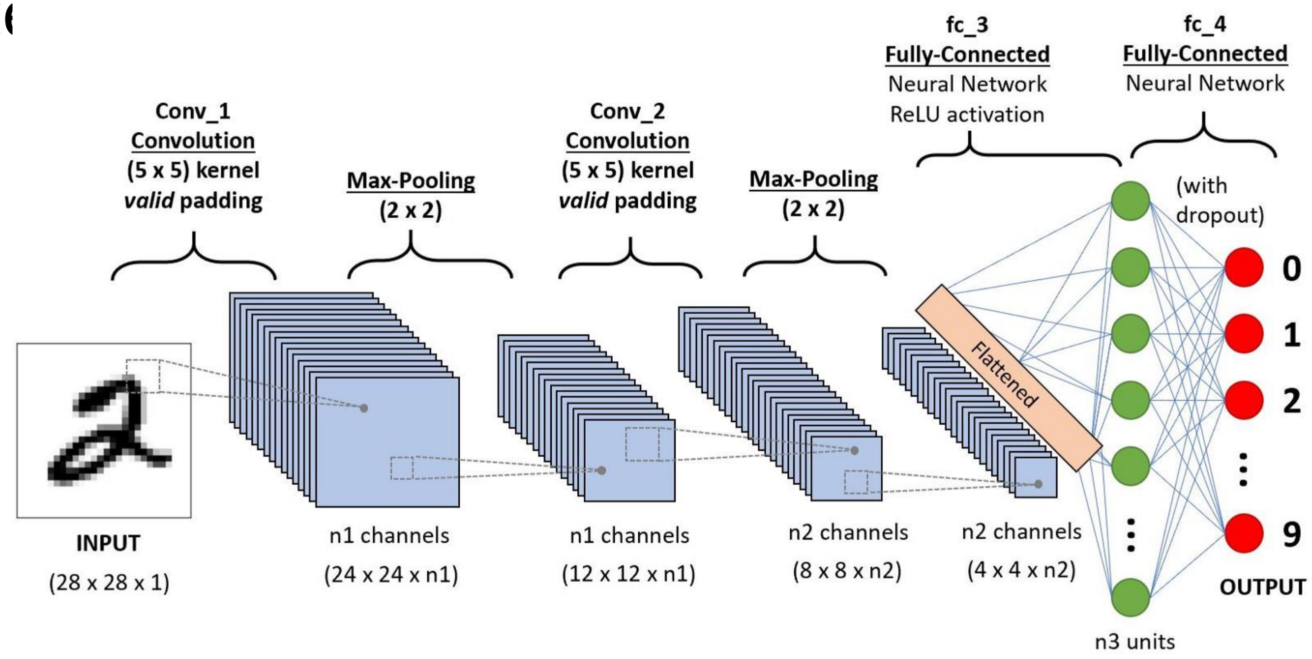
(GPU, TPU processing power)



- Better algorithms

Convolutional Neural Networks

- Specialized networks for image processing
- Pooling layers → Reduce the input size
- Convolution layers → Det



Yasaman Amannejad, Assistant Professor, Mount Royal University
Email: yamannejad@mtroyal.ca

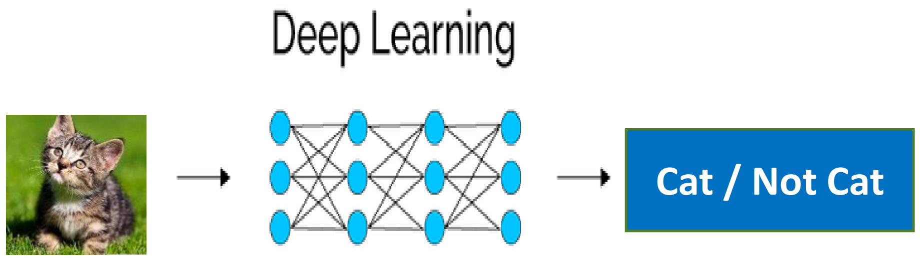
Image Patterns

- **Convolution layers** are responsible for detecting **patterns**.

- Patterns:
  - Edges
  - Shapes
  - Objects
  - Texture
  - Corners

- **Filters** detect patterns.

Filters



Feature Extraction + Classification

"Convolution"

3	1	1	2	8	4
1	0	7	3	2	6
2	3	5	1	1	3
1	4	1	2	6	5
3	2	1	3	7	2
9	2	6	2	5	1

Original image 6x6

✖

1	0	-1
1	0	-1
1	0	-1

Filter 3x3

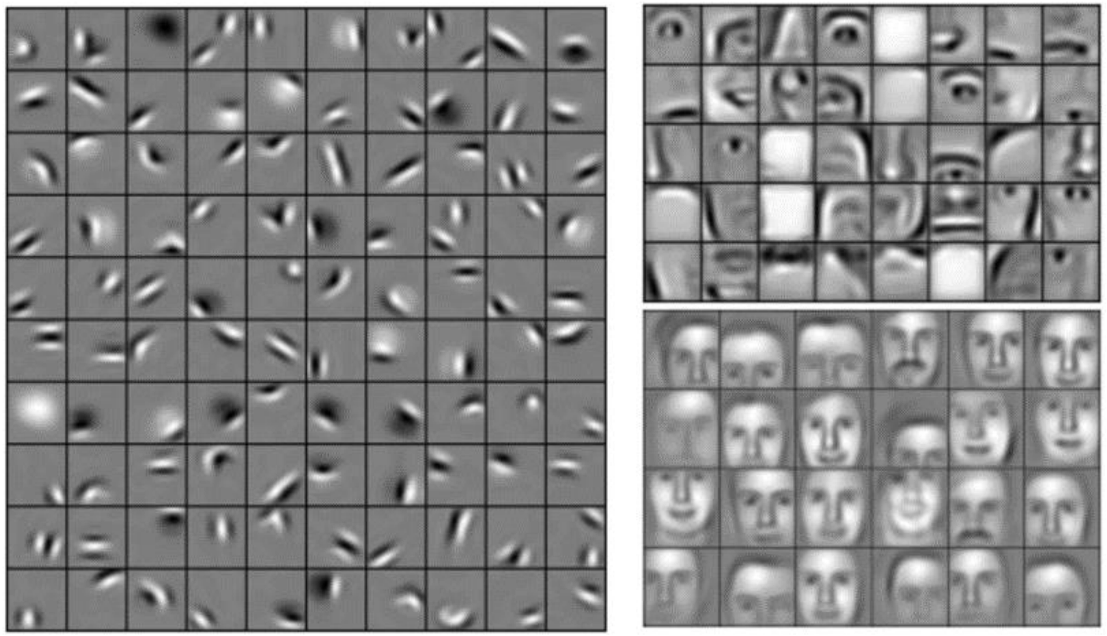
=

-7	...		
...	...		

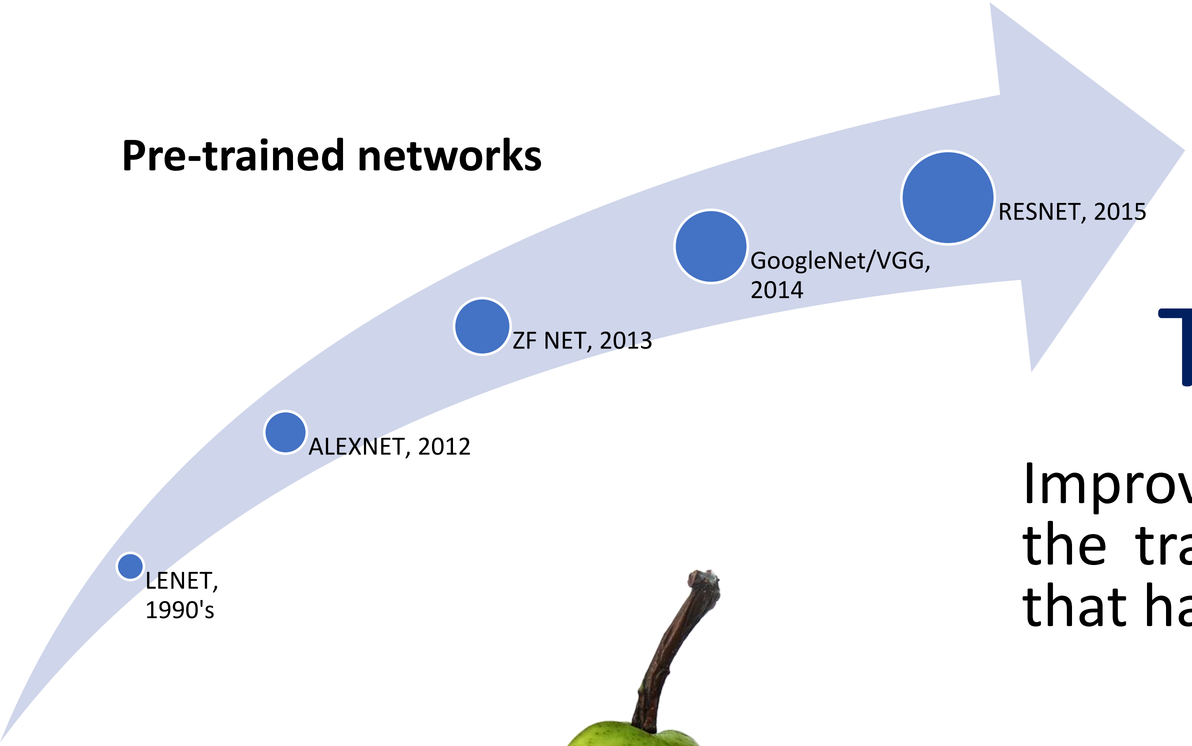
Output 4x4

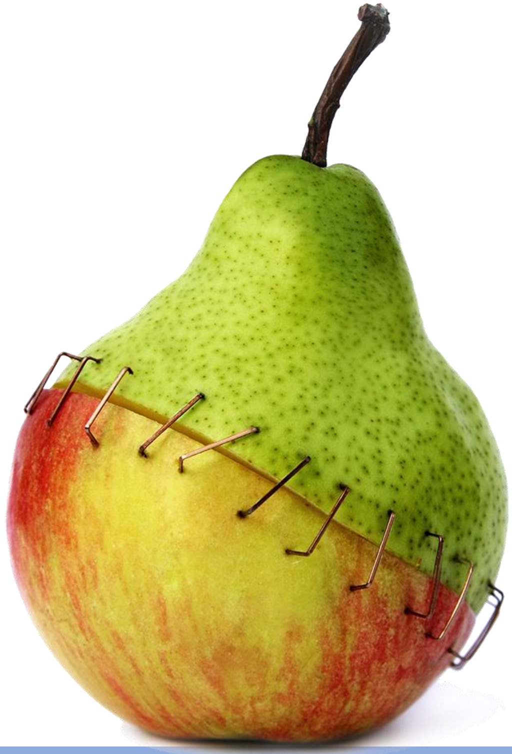
Result of the element-wise product and sum of the filter matrix on the original image.

Feature extraction is happening at the same time as the classification is happening



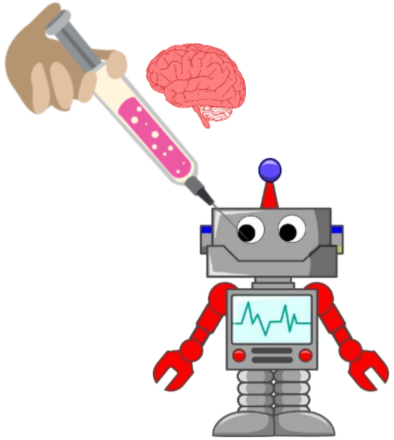
Pre-trained networks





Transfer Learning

Improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.



Yasaman Amannejad, PhD
Assistant Professor, Mount Royal University

Email: yamannejad@mtroyal.ca
Website: mru.ca/amannejad