

NLP for me

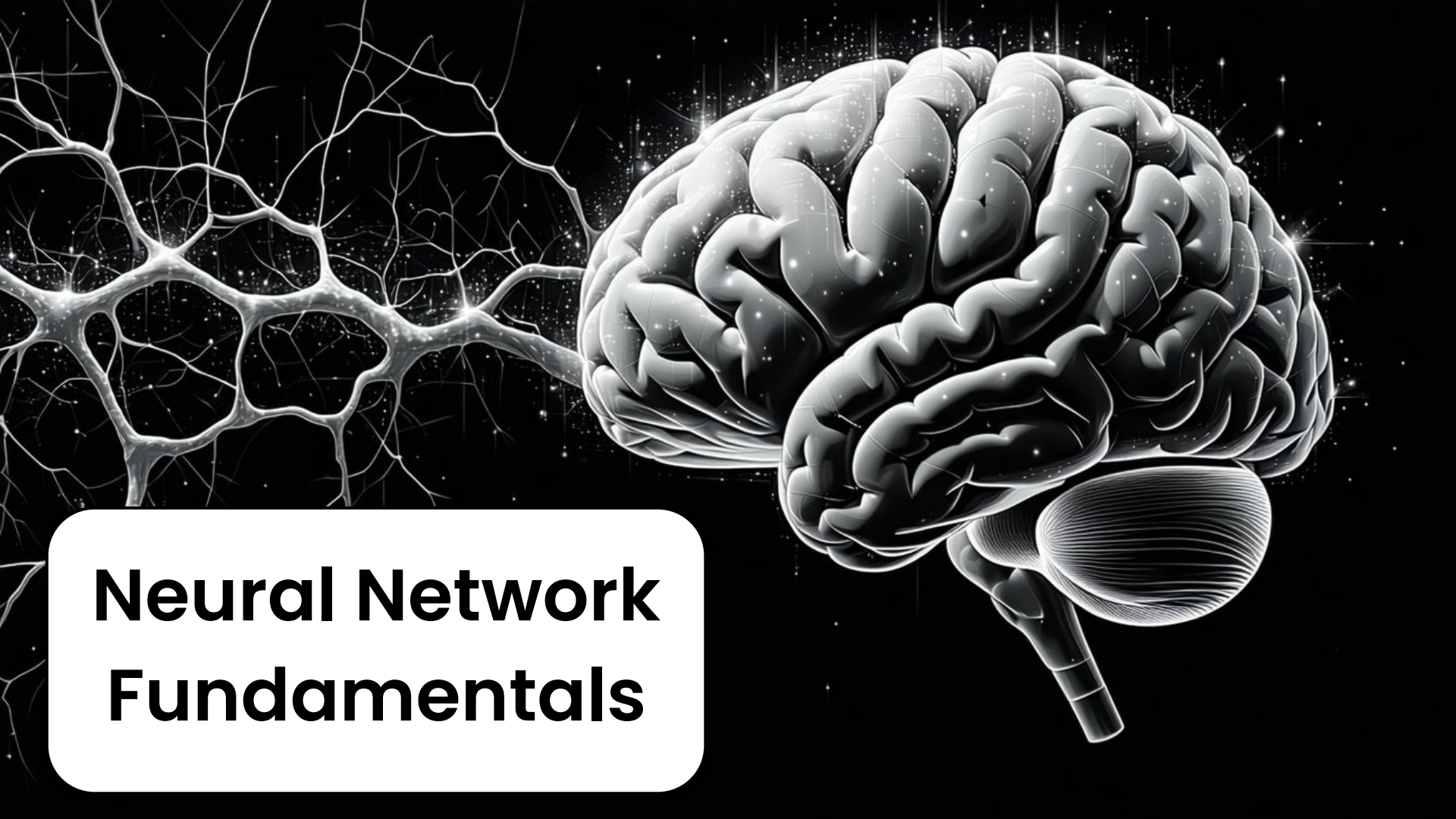
PWYC Microcourse in Natural Language Processing
October 2024

Part 5 – Deep Learning for Natural Language
Monday, November 4th, 2024



nlpfor.me

NLP from scratch 



Neural Network Fundamentals

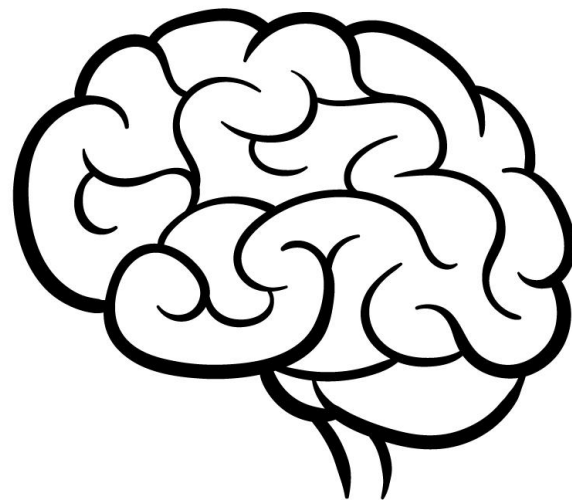
What is Deep Learning?

Deep Learning is a specialized type of machine learning that takes motivations from the structure of the human brain.

Unlike other machine learning models, deep learning models - or artificial neural networks - are composed of many nodes which can be viewed as individual "sub-models"

The theoretical foundations for deep learning have existed since the 1960s (or even earlier), but it only recently been realized with the rise of inexpensive and powerful computing available at scale.

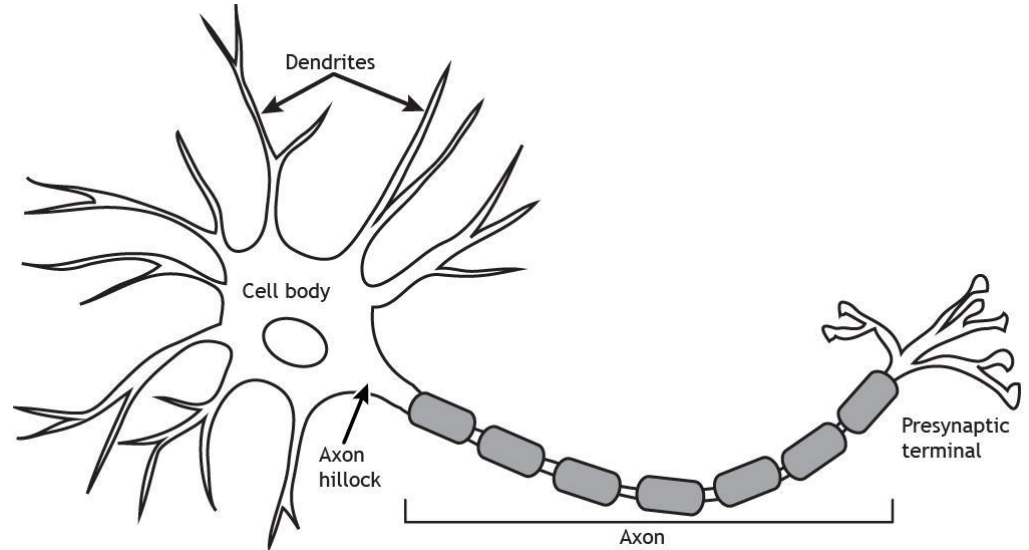
In NLP, deep learning models form the basis for state-of-the-art large language models (LLMs).



A Neuron in the Brain

The human brain is composed of billions of neurons, electrically excitable cells composed of a cell body, dendrites, an axon, and terminal.

Neurons receive input through their dendrites, and when firing, an electrical impulse travels down the axon to the terminal and release neurotransmitters to the next cell.



An Artificial Neuron (Perceptron)

An artificial neuron, referred to as a *perceptron*, is structured similarly: inputs to the model are the data plus a constant which are then multiplied by a set of *weights* (corresponding to dendrites in a physical neuron).

These together make a weighted sum of the inputs, which are processed through an *activation function* producing the perceptron output or *activation*, analogous to the axon and terminal in a physical neuron firing.

Many perceptrons combined together make up what was historically referred to as a Multilayer Perceptron (MLP) that we now refer to as a feed-forward, or fully connected, neural network.

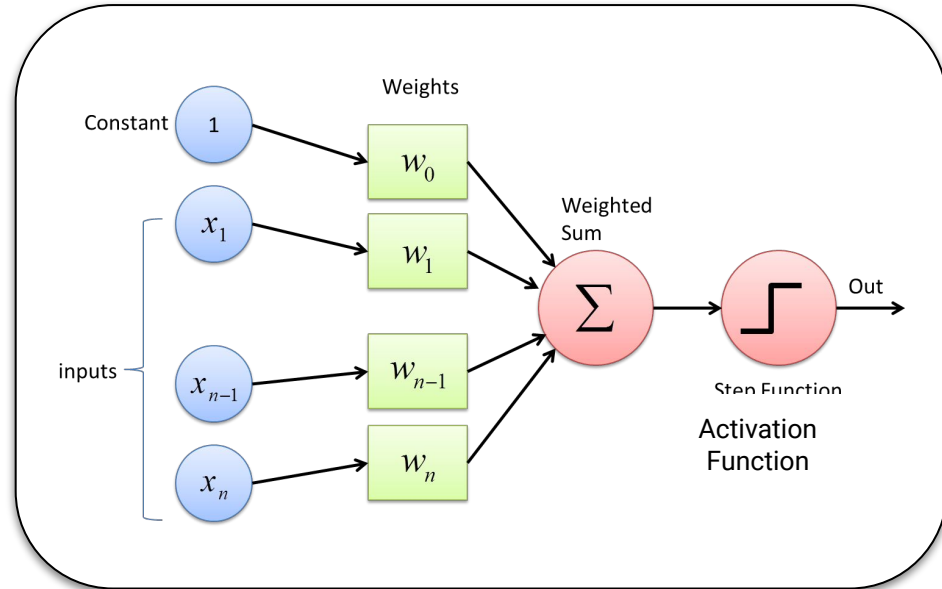


Image from:

<https://deepai.org/machine-learning-glossary-and-terms/perceptron>

Structure of A Neural Network

Multiple perceptrons are put together into *layers* composed of *nodes* (each perceptron) to create a *neural network*.

The outputs, or *activations*, which come out of previous layers become the inputs of the following layer.

The number of layers and number of nodes in the network - known as its *architecture* - is arbitrary and up to the modeler. There are also specific architectures that are well suited to particular types of problems.

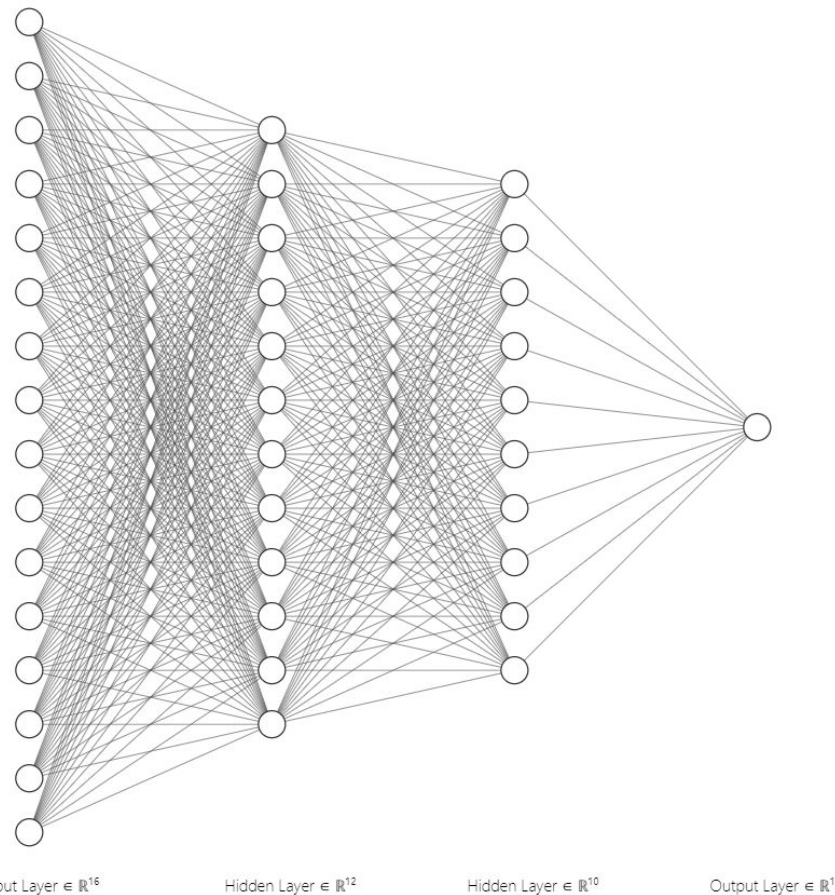



Image from: <https://alexlenail.me/NN-SVG/>

NLP from scratch 

Structure of A Neural Network

Multiple perceptrons are put together into *layers* composed of *nodes* (each perceptron) to create a *neural network*.

The outputs, or *activations*, which come out of previous layers become the inputs of the following layer.

The number of layers and number of nodes in the network - known as its *architecture* - is arbitrary and up to the modeler. There are also specific architectures that are well suited to particular types of problems.

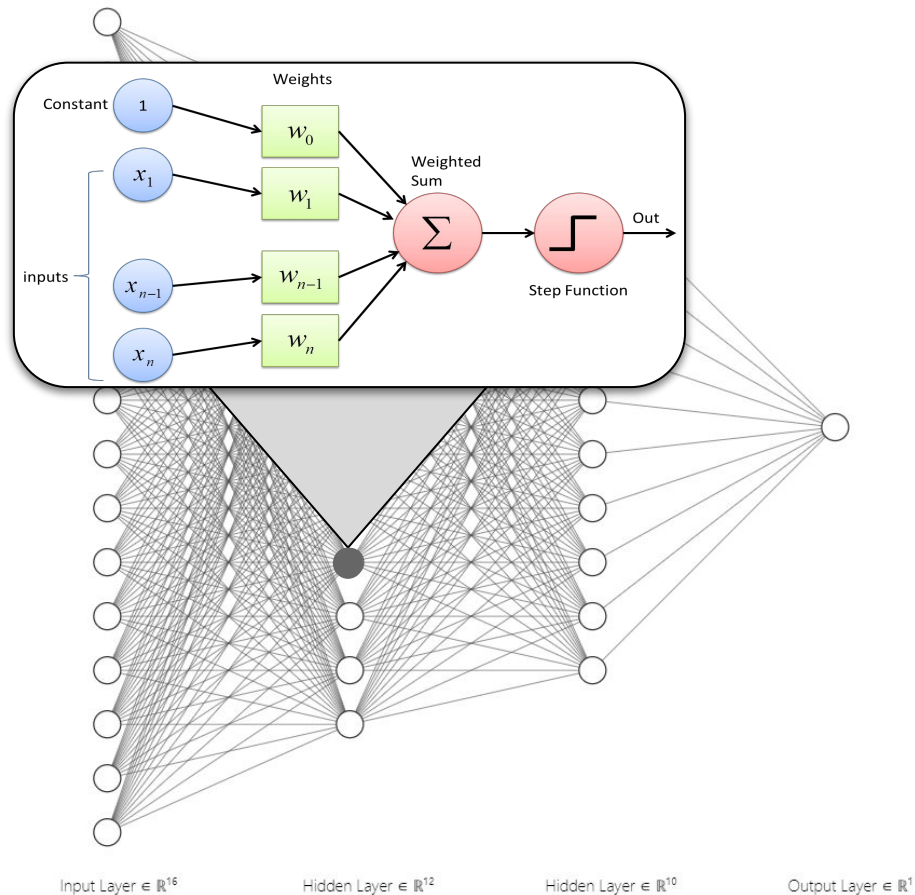



Image from: <https://alexlenail.me/NN-SVG/>

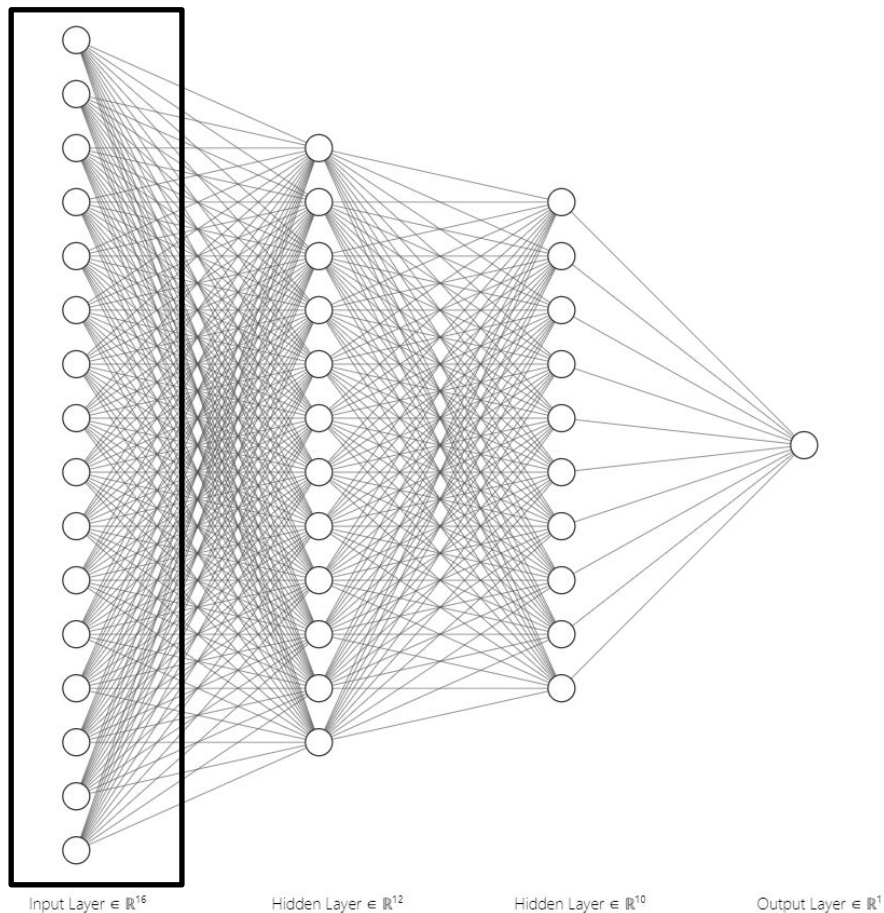
NLP from scratch 

Input Layer

The *input layer* is not a “true” layer but just passes the data through to the following layers - we say either that there is no activation or that the activation function here is a linear passthrough (the function $y = x$).

Each neuron in the input layer represents a feature of the data, so there will be an equal number of nodes in the input layer as there are features in the data.

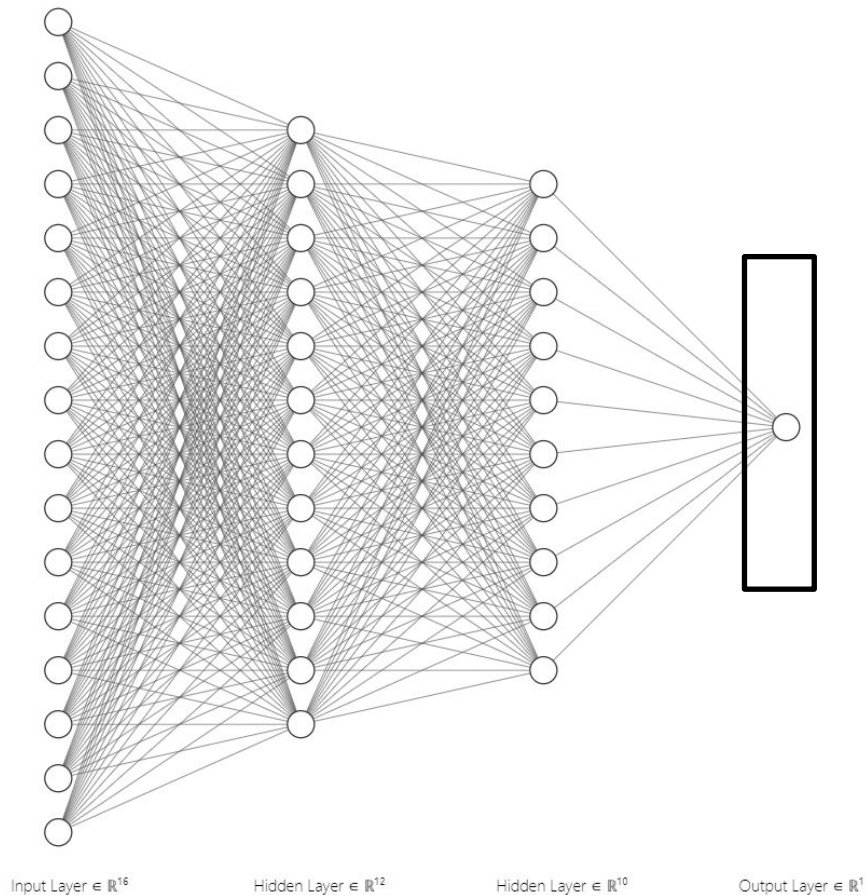
For example, the neural network depicted on the right could be used to make predictions if we tabular data with 16 features (columns) describing each observations (row) in the dataset.



Output Layer

On the other hand, the *output layer* is the final layer of a neural network that produces the network's predictions (output). The number of nodes in the output layer is dependent on the problem type.

A single node can be used for binary classification or regression, since for each observation of input there will only be one output: a probability between 0 and 1 of lying in the positive class (class 1) in the case of binary classification, and a single numeric value in the case of regression - predicting a continuous value associated with each observation of the input data.

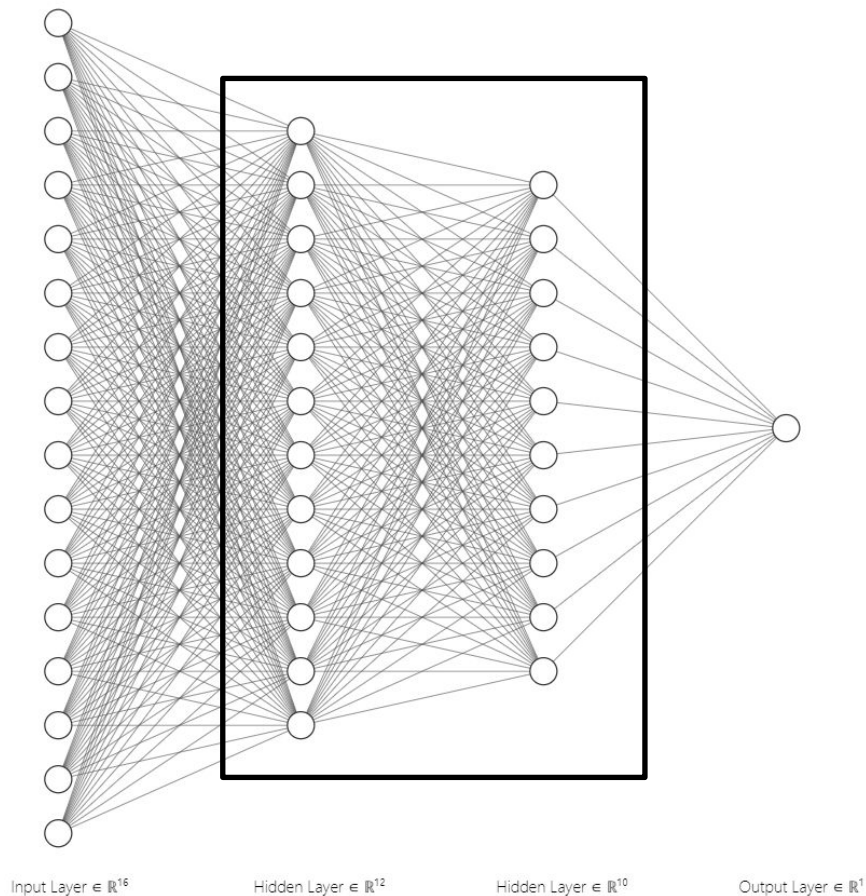


Hidden Layers

The intermediate layers between the input and output of the network are the so-called *hidden layers*.

Each hidden layers perform computations on outputs of previous layer, a linear combination of these multiplied by coefficients - or *weights*. The activation function is then applied to this result. The weights are what is learned by the model in training.

Here we are speaking only of *fully-connected (feed-forward)* networks, the simplest type of neural network. There are other layer types for different types of models specifically suited to certain types of tasks (e.g. computer vision)



Activation Functions

The ability for neural networks to learn highly complex, non-linear relationships is greatly due to activation functions. As these are applied at each layer, as the information flows through the network from left to right, the functions are applied atop one another in previous outputs.

Each layer may have a different type of activation function, though it is not uncommon to use the same for most, if not all of the layers. Again, these choices are up to the modeler.

There exist families of well-known mathematical functions that perform well and have desirable mathematical properties and serve as standard choices to use when training deep learning models.

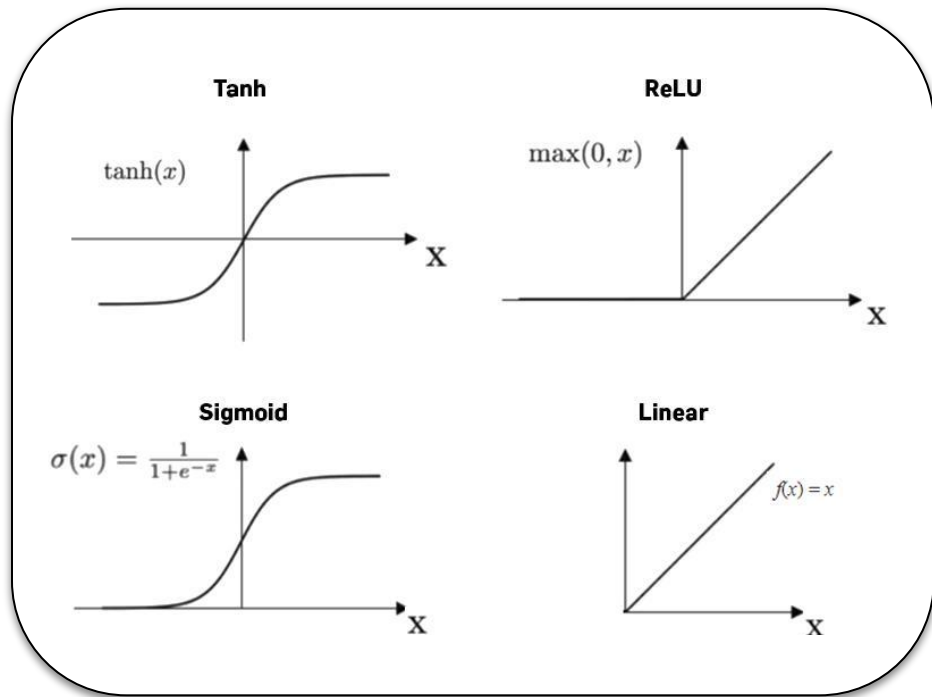


Image source: <https://machine-learning.paperspace.com/wiki/activation-function>

Loss Functions

Every neural network also has a *loss function*. This is just a technical term for a function which measures the model's error. They compares the values output from the network to those expected or known to compute the error.

Depending on the type of problem the network is being applied to, there are different families of loss functions which are used.

Any machine learning model is never "right" - it is only less wrong, and the goal of training a neural network is to find the optimal values for the weights such that the loss (error) is minimized.

But how do we find these values for the weights? That brings us to the details of how neural networks are trained.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$L(\hat{y}, y) = - \sum_k y^{(k)} \log \hat{y}^{(k)}$$

How do Neural Networks Learn?

The details of how neural networks are trained and the optimal values for the model weights found are quite mathematically complex.

At a high level, it can be viewed as an optimization problem where we want to minimize the error (loss) as a function of how much we should change the weights.

Imagine you wish to find the lowest point in a mountain range - which way should you walk to make the steepest descent and reach the bottom?

Analogously, there are numerical methods which can perform computations over many millions of parameters in a neural network to determine how to adjust the weights to those which minimize the error.

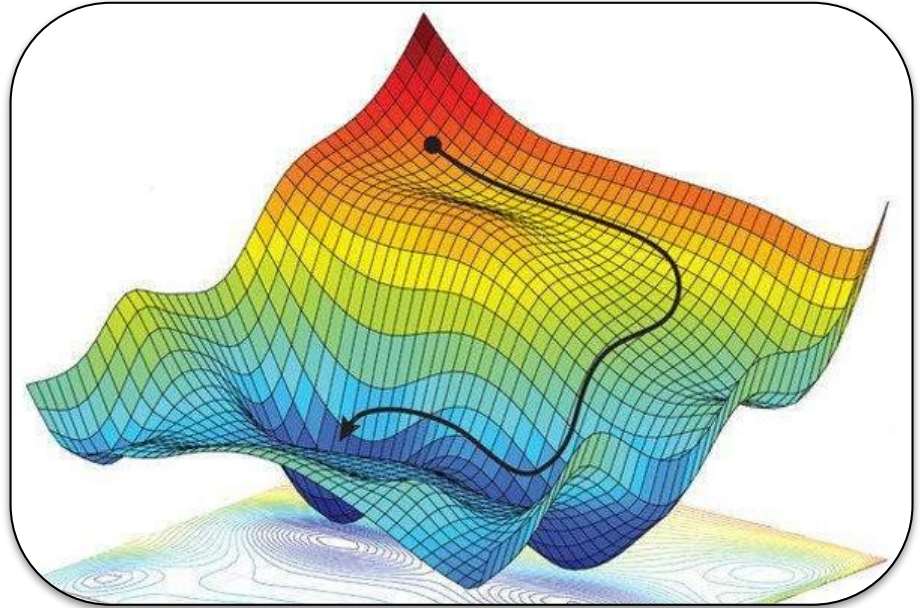


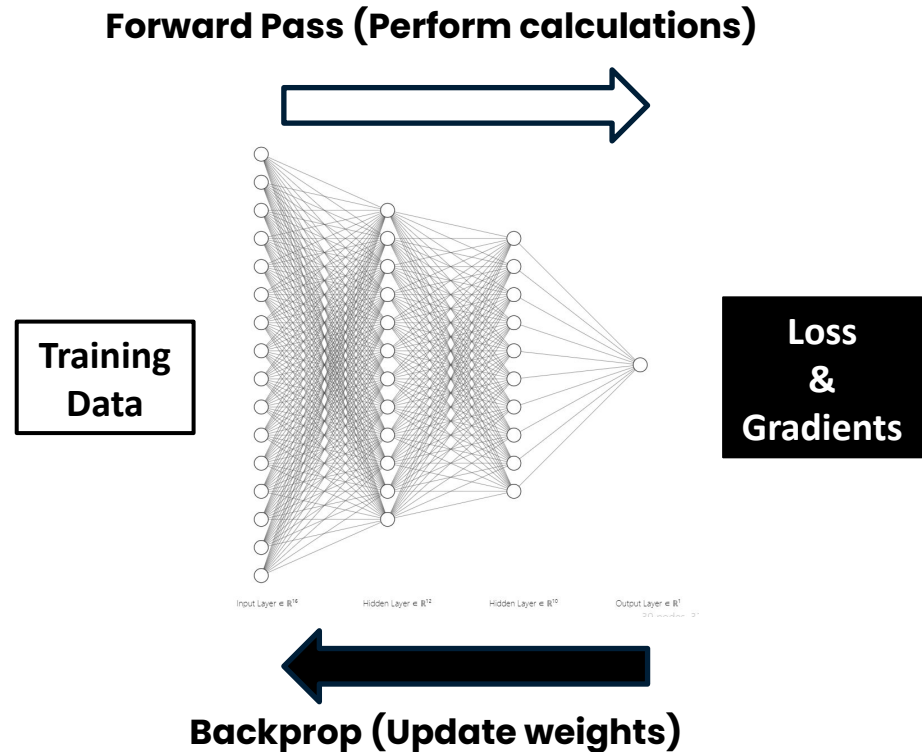
Image from:
<https://www.mygreatlearning.com/academy/learn-for-free/courses/stochastic-gradient-descent>

Forward Pass and Backpropagation

This logic is applied during neural network training. Unlike other types of machine learning models, neural networks are trained in two steps which are repeated many times.

In the *forward pass*, data is run through the network to compute the output, and error calculated from the loss function comparing this output against the known true value associated with the input.

Backpropagation (“backprop”) follows, and applies changes to weights' values in the network as determined from *gradients*, or slopes, the direction of greatest decrease of error.



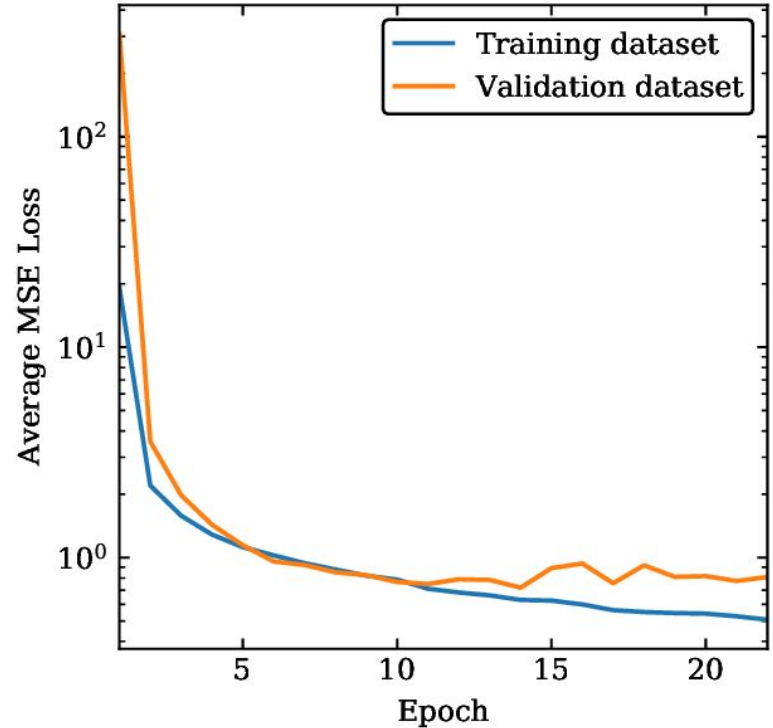
Epochs and Batches

In addition to differing from other types of machine learning based on these two steps, deep learning also differs in that data is not passed through the network once in its entirety, but in smaller subsets known as *batches*.

When all the data has gone through the network once, this is referred to as a single *epoch* of training

One epoch is composed of many batches, and networks are trained for many epochs and see the whole training dataset multiple, even hundreds or thousands of times, depending on the problem.

With each epoch of training, the model weights are adjusted and the network better learns the relationships between the features and target variable.



(Python) Neural Network Frameworks



TensorFlow

- Google product
- Graph-based computation, GPU training
- Other deployment options (Tensorflow Lite, TF.js)
- Easy with integration of Keras into TF 2.x



PyTorch

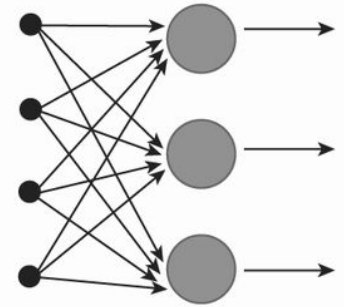
- Facebook product
- Graph-based computation, GPU training
- Pytorch Mobile for embedded, no web (ONNX?)
- OOP dev focus (ML eng), Lightning equivalent to Keras



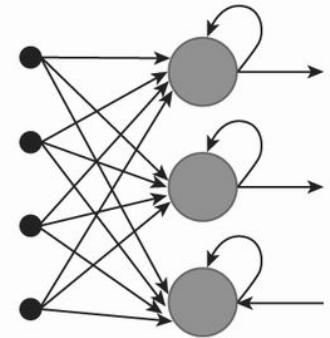
Deep Learning for Language

Recurrent Neural Networks

- Designed for **sequence data**, where each output depends on prior inputs (*recurrence*), making them useful for tasks like text, speech, and time-series analysis.
- RNNs address sequential dependencies that traditional neural networks do not
- Developed in 1980s, RNNs evolved with improvements like LSTM (1997) and GRU (2014) cells to overcome memory and computational limitations and enable longer dependency learning
- RNNs capture linguistic patterns over time, generating coherent sentences or paragraphs based on learned language structure - e.g. character-level and word-level RNN text generation



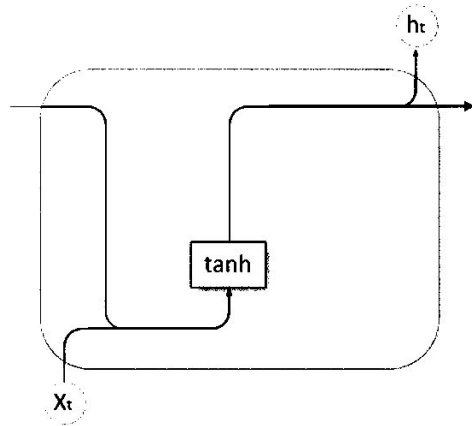
(b) Feed-Forward Neural Network



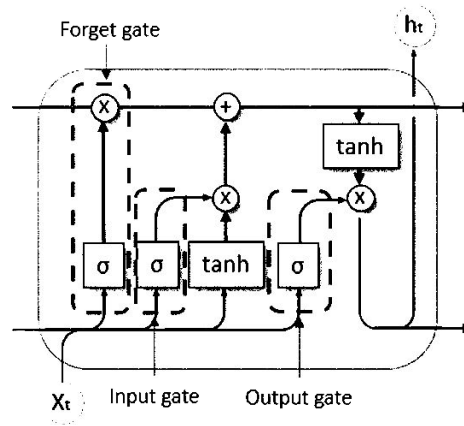
(a) Recurrent Neural Network

Image credit: [geeksforgeeks.org/introduction-to-recurrent-neural-network](https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/)

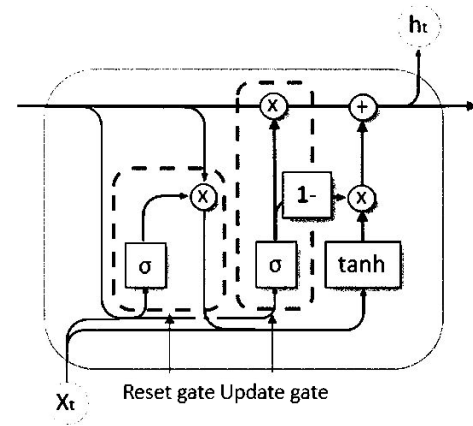
Recurrent Neural Network Types (Cells)



RNN

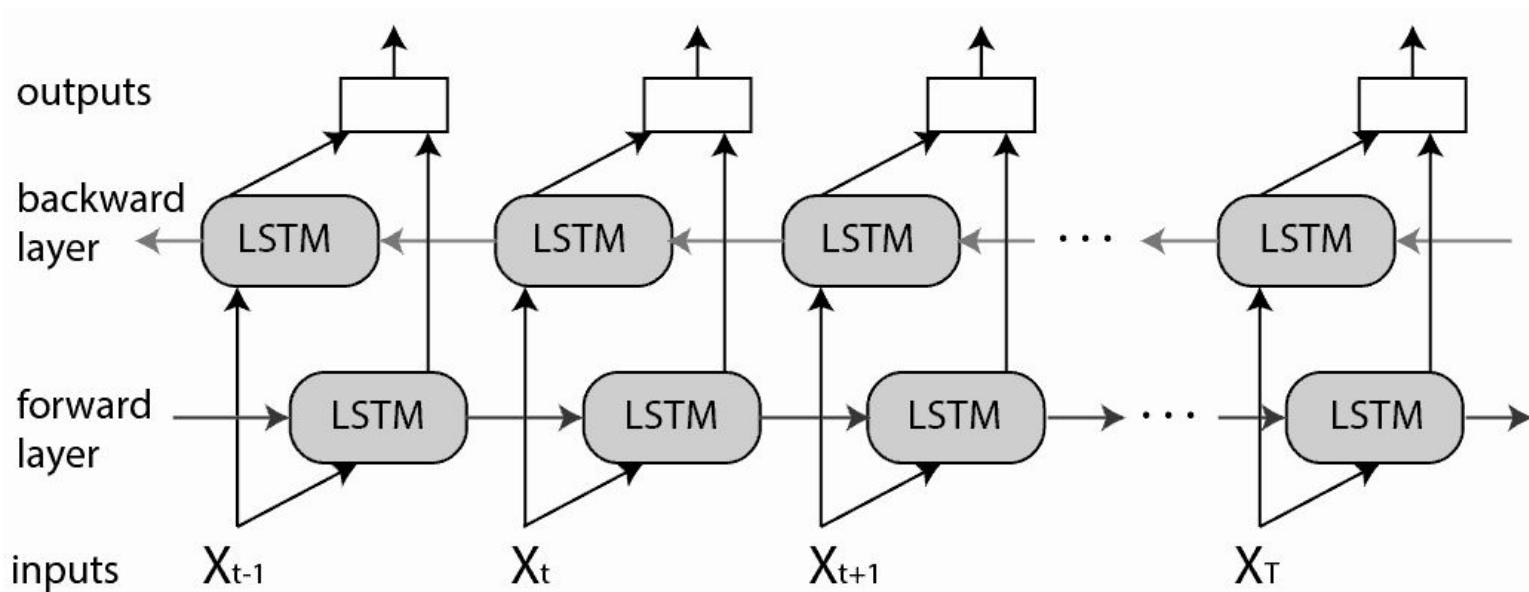


LSTM



GRU

Bidirectional LSTM



Applications of RNNs

AB
CD

**Text
Generation**



**Time Series
Forecasting**



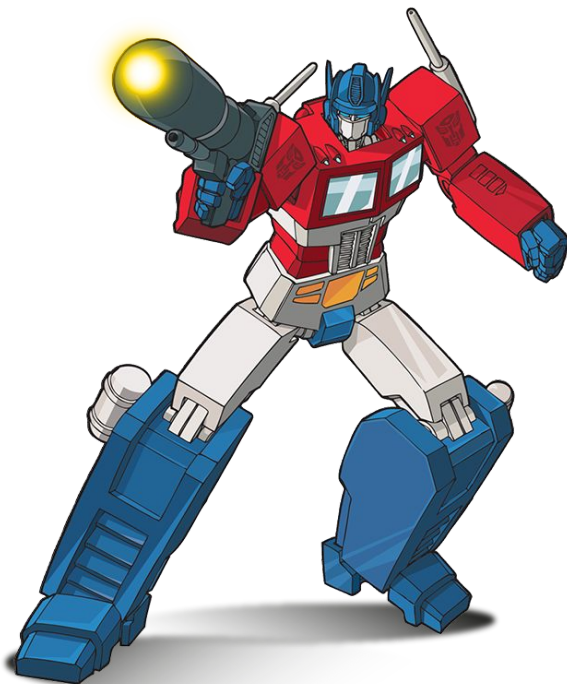
**Audio & Video
Processing**

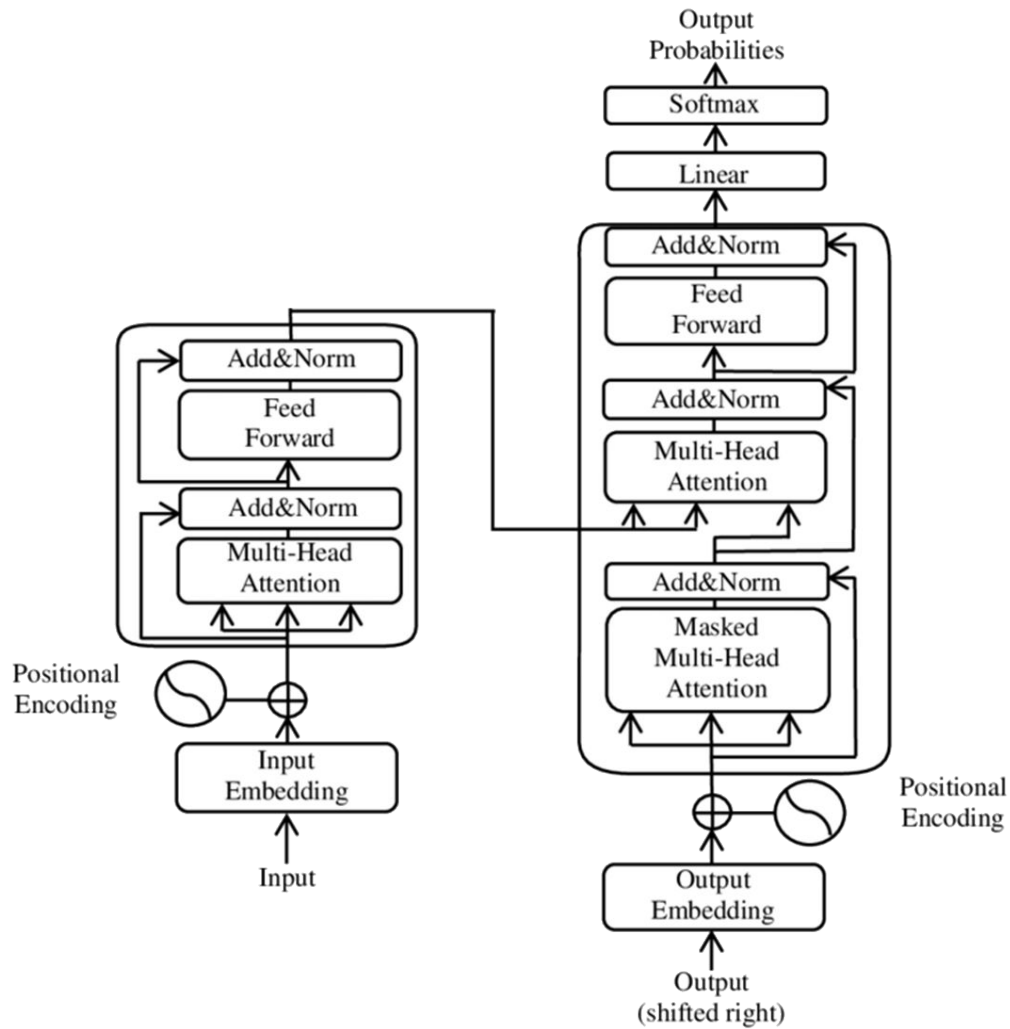


**Translation &
Summarization**

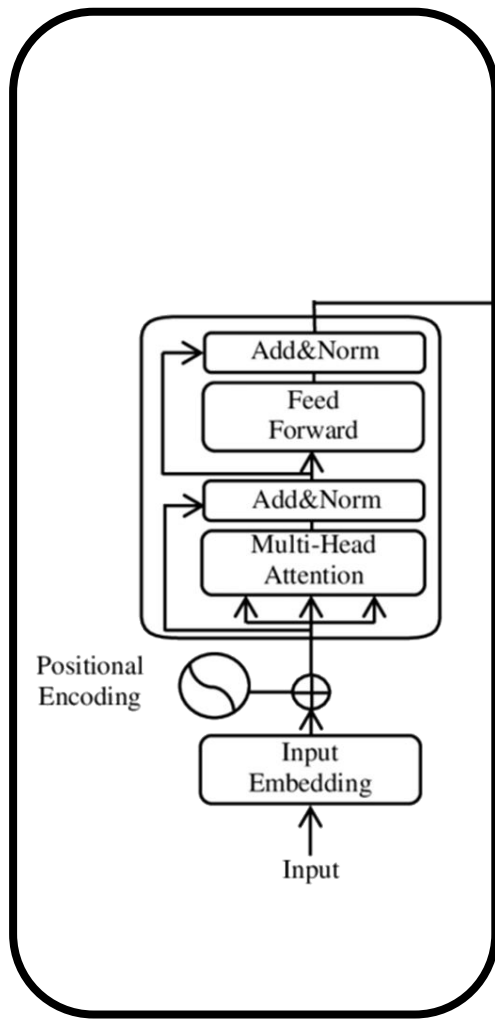
The Transformer Architecture

- Groundbreaking paper "Attention is All You Need" from Google researchers (Vaswani et al, 2017) introduced Transformer architecture
- Original application in machine translation but now general purpose and applied to a myriad of other tasks
- Represents the state of the art for LLMs and also applied in domains outside of language (image generation) - virtually all new models based on this architecture
- Popularized by OpenAI and the Generative Pretrained Transformer (GPT) series of models

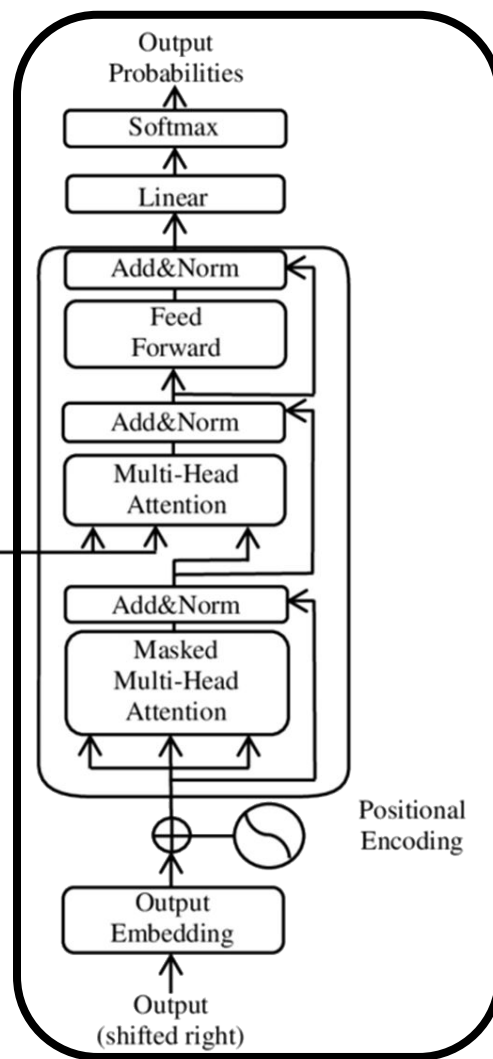




ENCODER (e.g. BERT)



DECODER (e.g. GPT)



End of Part 5

[NLPfor.me](https://nlpfor.me)

PWYC Microcourse in Natural Language Processing
October 2024

Part 5 - Deep Learning for Natural Language
Monday, November 4th, 2024



nlpfor.me

NLP from scratch 