Artificial neural networks are computing systems inspired by the biological neural network of the brain. Such systems can progressively improve their ability to do tasks and recognize patterns by learning from examples. Artificial neural networks are in their essence computational networks that can perform certain specific tasks like clustering, classification, pattern recognition. They do this by representing patterns in data as networks of connections between nodes on the network. They then learn by altering the strength of the connections between the nodes to create new network structures that can represent new patterns. For example, neural nets are now widely used for image recognition, where they learn to identify images that contain say a house by analyzing example images that have been manually labeled as such and using the results to identify houses in other images.

As the name implies they are directly inspired and modeled on the working of the brain. Thus to understand neural networks it is of value to understand a little how the brain works to represent and process information. The brain is composed of neurons and connections between them called axons, which have synapses where the different neurons meet. Neurons generate electrical signals that travel along their axons. Any given neuron has a number of inputs from other neurons if those inputs are above a given threshold then it is activated and fires. If a neuron is activated it then sends out signals to other neurons that it is connected to. The synapses change in their chemical composition as one learns in order to create stronger connections between networks of neurons. In such a way the cognitive system can adapt and changes over time to form new patterns of neural networks. If two neurons are turned on when a pattern is stimulated then the synaptic connection between them becomes stronger. The brain is physically built as a neural network and cognition happens in patterns. Networks of interconnected neurons form a pattern which corresponds to an idea or memory.

An artificial neural network is based on this very same architecture as a collection of connected nodes that form a network. Each connection between nodes can transmit a signal to another node. The receiving node can process the signal and then signal downstream nodes connected to it. They're all connected to each other going down through the layers. The nodes mimic neurons in that they're little triggers. Each neuron has something of a threshold where it makes a decision. Nodes take inputs from connected nodes and have some internal function to determine when or if they will fire to send a signal downstream. The nodes and connections may also have a weight which can increase or decrease the strength of the signal that is send downstream which varies as learning proceeds. Typically weight represents the strength of the interconnection between neurons inside the neural network. As such artificial neural networks can be viewed as weighted directed graphs in which artificial neurons are nodes and directed edges with weights are connections between neuron outputs.
Layering

Typically, neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first (input), to the last (output) layer, possibly after traversing the layers multiple times. The input layer contains those units (artificial neurons) which receive input from the outside world on which the network will learn or process. The output layer contains units that respond to information that the network has learned to represent or process. These units between input and output layers are termed hidden layers. A neural network with many layers in between is called a deep learning neural network. When there are no layers in between it is simply a neural network. Because there are very few practical applications for a two-layer neural network virtually all are going to be of the deep learning form. As of 2017, neural networks typically have a few thousand to a few million units and millions of connections. These different layers of the network can be used to represent the different levels of abstraction during classification and identification as we will discuss further in the coming module on deep learning.

Back Propagation

The key innovation required to actually get a functioning neural network is what is called backpropagation. This involves the network learning through an iterative process where you send all the information back through the network and the network adjusts its weights so as to learn to match the output better. The basic idea in backprop is quite intuitive, it is simply looking for the difference between the network's actual output and the desired output. The error in the network's output is used to determine how much the network adjusts and changes on the next iteration. The error between the desired output and the real output we are trying to minimize. So if we put in a graphic of a circle and asked the system to identify it and it outputted an estimate that it was .8 likely to be a circle we know that there is an error of .2 We can then adjust the nodes up and down by a very small amount to see how the error changes. The amount that they're adjusted is determined by how big the error is, with a large error they are adjusted a lot, with a small error just a bit, with no error they're not adjusted at all.

We are doing this alteration to try to go down the gradient to where the error is minimal. We keep making these adjustments to the weight of the nodes and connections all the way back to the input and this is why it's called back propagation, because what we are doing is propagating back the errors and updating the weights to make the error as small as possible. This backpropagation algorithm is what is at the heart of deep learning and these days is used for just about everything you hear about in machine learning. Very early on people use it to do things like predicting the stock market, these days it is used for search, for automatic adding color to Black and White images, for video recognition,
for automatic handwriting generation, for generating captions on images, for speech recognition, for simultaneous translation, social network filtering, playing board and video games, medical diagnosis, and many other applications.\(^5\)

**Conclusion**

The idea of neural networks has been around since the sixties but at that time computers didn't have enough processing power to effectively handle the work required by large neural networks. Neural network research slowed until computers achieved far greater processing power. Support vector machines and other, much simpler methods such as linear classifiers gradually overtook neural networks in machine learning popularity. It is only very recently that this has changed as available computing power increased through the use of GPUs and distributed computing; neural networks are now starting to be deployed on a large scale. They have found best usage in applications difficult to express with a traditional computer algorithm using rule-based programming. Neural networks are the basis of deep learning, that has become highly popular in the past years.

**Citation**


