Chapter 1

**Introduction**

At the time when programmable computers were first built, they quickly eclipsed most of the common approaches to problem solving in fields such as mathematics, physics, business, communication, education etc.

Abstract and formal tasks that are among the most problematic mental endeavors for a human being are amid the easiest for a computer.

Since then, people started wondering if such machines would have the potential to become intelligent and take calculated decisions with little to no human assistance.

One example for such a task is the ability to predict the movements and take the decision that will most likely lead to a victory in a chess game. IBM’s Deep Blue chess-playing system defeated the world champion, Garry Kasparov, in 1997. Chess is of course a very simple world due to its limitations (64 locations and 32 pieces that can move in a very restricted manner) and, even if developing a successful chess strategy is a considerable achievement, the challenge is not due to the difficulty of describing the set of chess pieces and legal moves to the computer, but rather to the implementation of efficient algorithms that conform to hardware limitations. Chess can be described by a short list of completely formal rules, which can be analyzed and provided ahead of time by the programmer.

Recently, in 2015, Google’s AlphaGo became the first computer program to beat a human professional Go player on a 19x19 board. This event was followed up by considerable victories against world champions and in 2017, Ke Jie, the world champion at the time, was beaten by AlphaGo Master (the successor of AlphaGo) in a three-game match. Go is considered more laborious for computers to win that other games such as chess due to its much larger branching factor (in games, it is the number of legal moves that a player has each turn). For comparison, chess has an average branching factor of 35, while Go’s average branching factor is 250. This greatly affects the ability of a computer to predict future moves. For example, if we consider the estimated number of possible board configurations in games, chess has a total of 10120 configurations, while Go has around 10174. Both numbers are astronomical (as comparison, it has been estimated that there are around 1080 atoms in the observable universe).

Due to the exponential increase in the complexity of Go compared to chess,

a new approach had to be taken to reduce the number of moves taken into consideration as potentially good moves. This requirement was similar to how humans think in such situations, generally sticking to strategies that proved to be efficient in the past and disregarding out of the blue moves that will probably lead to ineffective board configurations. The latter moves, although having the opportunity of being the better short-term action, will most likely lead to an unfamiliar board configuration that will require more thinking and will be prone to fail.

Such scenario required a system that would be bootstrapped from human gameplay expertise, trained to mimic human behavior by matching the moves of expert players from recorded historical matches, using a database of around 30 million moves. Once it had reached a certain degree of expertise, it was further trained by playing a substantial number of games against itself, and the better version was kept and trained against similar clones of itself, thus using a combination of machine learning and tree search techniques implemented using deep neural network technology.

* 1. Machine learning

Machine learning is defined as the set of automatic computing procedures based on binary or logic operations that learn a task from a series of examples. One of its aims is to generate expressions simple enough to be understood by the human. Those expressions must mimic human reasoning in order to provide some form of insight into the decision process.

Nowadays, the artificial intelligence field (commonly known as AI) is a rapidly growing field with many practical applications and active research topics, making intelligent software responsible for automating routine labor, speech and image recognition, supporting scientific research and making diagnoses in medicine.

Some artificial intelligence projects aimed to hard-code knowledge about the world in formal languages. A computer is able to reason automatically about statements in these formal languages using logical inference rules. This approach to artificial intelligence is known as the *knowledge base* approach. None of these projects turned out to be successful. One of the most known such projects was Cyc (developed by the Cycorp company, initially released in 1984). Cyc is the world’s longest-lived artificial intelligence project, having the goal of enabling artificial intelligence applications to perform human-like reasoning. It is an inference engine and a database of statements written in a language called CycL, and its knowledge base is divided into *microtheories*, defined as concepts referring to one particular realm of knowledge. The statements are entered by human supervisors, and the process of devising formal rules with enough complexity to accurately describe a world has been widely criticized and considered a struggle. One example for its absurd complexity was that Cyc failed to understand a story about a person named Fred shaving in the morning: it was aware that electric razors contained electrical parts and that people do not, but because Fred was holding an electric razor, it believed that the entity named “FredWhileShaving’ contained electrical parts. Therefore, it asked the question of whether Fred was still a person while he was shaving.

The difficulties that systems relying on hard-coded knowledge unveiled suggested that artificial intelligence systems must have the ability to acquire their own knowledge, by extracting patterns from raw data. This ability is now known as *machine learning* and it enabled computers to solve problems involving knowledge of the real world and take decisions that appear subjective.

1.2 Deep learning

*Deep learning* was introduced as a subset of machine learning. Although commonly misused, those two terms imply different aspects of how models learn and a clear distinction must be made before continuing. For future references, the usage of the term deep learning denotes deep artificial neural networks, not deep reinforcement learning as it may be the case in other research papers.

Firstly, when referring to the structure, a machine learning model uses types of automated algorithms which learn to predict future decisions and model functions using the data fed to it, while a deep learning model interprets data features and its relationships using neural networks which pass the relevant information through several stages of data processing.

Secondly, machine learning algorithms are directed by the analysts to examine the different variables in the datasets, while deep learning algorithms, once implemented, are usually self-directed for the relevant data analysis.

Thirdly, machine learning models usually have a few thousand data points used for the analysis, while deep learning models usually have a few million.

Moreover, machine learning models usually have a numerical value (a score or classification) as the output, while the deep learning models can have anything from a score, an element, text, sound, etc.

Figure 1 .1 - Relationship between machine learning models and deep learning models

1.3 Artificial neural networks

A deep learning model is designed to continually analyze data with a logic structure similar to how a human would draw conclusions. To achieve this, deep learning uses a layered structure of algorithms called and artificial neural network (ANN).

Artificial neural network algorithms were inspired by the architecture and flow of neurons and their synapses in the human brain. Although simple in design compared to the real neuron networks, their ability to perform a wide area of information-processing tasks is uncontested. Some of those tasks seemed almost natural for people to perform, but difficult to describe formally – problems which we intuitively solve, such as recognizing patterns of speech, faces, words and digits written by hand. For a human, such tasks become second nature once he is accustomed to patterns that are homogeneous to the implied pattern, but once you try to come up with an algorithm to classify that pattern within a reasonable margin of error, it seems almost irrational for a human to give an exact answer whilst not being able to explain the decision making behind it.

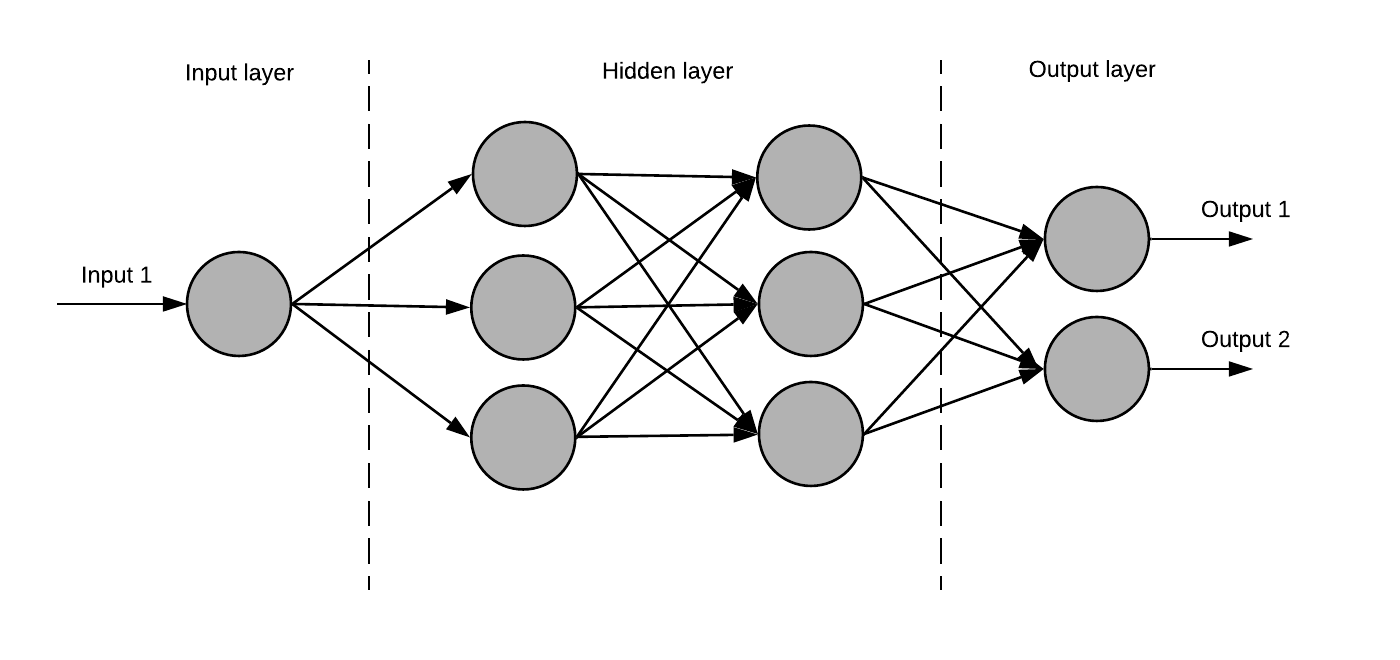


Figure 1.2 - Artificial neural network architecture

A first example for a problem that is nowadays solvable by a machine, while outperforming the human counterpart is some cases, is the object recognition in images, for instance in the sequence of camera images taken by a self-driving car. Recently, usage of neural networks in object recognition has registered an upturn, due to a steady increase in demand for face recognition and self-driving car software in the industry. Other arguments for this demand are the availability of better image databases, thus offering a wide support for the training the networks, and the improvement of hardware that reduces the time needed to have a production-ready network.

Artificial neural networks excel in analyzing large sets of data where it may be difficult to determine beforehand which properties of that data are of interest. In such cases, unsupervised learning is often taken into account for developing the network. Unsupervised learning algorithms allow the network to learn without a training set by determining in terms of which categories the data can be analyzed. In this way, networks can detect familiarity (which input patterns had occurred most often) and other structures in the input data. Those algorithms excel when there is redundancy in the input data which is not obvious at first because the data is high dimensional.

The other artificial neural network training method is by means of supervised learning. Supervised learning must be closely monitored by a specialized individual. During the training of the network under supervised learning, the input data is vectorized and transformed into an output vector. The output will be compared to the desired result and, if there is a difference between the actual output and the desired output, an error signal will be generated. This error signal will determine the network to adjust itself until the actual output is matched with the desired one. This research paper will focus more on the supervised learning algorithms and how networks may be adapted for overperforming in such scenarios.

1.4 History and Personalities in Artificial Intelligence

Although comprehensive, the historical context of deep learning can be better understood by identifying some key trends:

1. Deep learning has proven to be more useful and versatile as the amount of available training data increased.
2. Deep learning models have grown in size over time as computer infrastructure, represented by both the hardware and the software counterpart, for deep learning has improved.
3. Deep learning has solved progressively complicated applications with increasing accuracy over time.
4. Deep learning has had a long history during which it has taken many names and has risen and fell in popularity.

Deep learning dates back to the 1940s, the model seems because it was rather unpopular for several years preceding its current popularity and because it has gone through many distinct names, only recently being called “deep learning”. The field has been rebranded many times, reflecting the influence of different researchers and different perspectives.

Predominantly, there have been three waves of development:

1. Deep learning known as cybernetics in the 1940s-1960s, started with the development of theories of biological learning and the implementation of the first models, such as the perceptron (Frank Rosenblatt, 1958).
2. Deep learning known as connectionism in the 1980s-1990s when the concept of backpropagation was experimentally proven to generate useful internal representations in the hidden layers of neural networks (David Rumelhart, Geoffrey Everest Hinton, Ronald J. Williams, 1986).
3. The current reemergence that has taken the name deep learning, which began in 2006 when the previous viewpoint that training deep architectures was too difficult to optimize had been proven wrong.

From a chronological point of view, the major milestones that have been achieved in machine learning are:

1763 - Thomas Bayes’ work “An Essay towards solving a Problem in the Doctrine of Chances” is published two years after his death, setting the foundation for *Bayes’ theorem* that describes the probability of an event based on prior knowledge of conditions that might be related to the event.

*Bayer’s theorem*: Given a hypothesis H and evidence E, Bayer’s theorem states that the relationship between the probability of hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(E|H) is:

1805 – Adrien-Marie Legendre introduces the *least squares method*, widely used nowadays in data fitting.

1812 – Pierre-Simon Laplace expands upon the work of Bayes and defines what is now known as Bayes’ Theorem.

1913 – Andrey Markov studied the distribution of vowels in a poem and described the techniques used, which are now known as *Markov chains*.

1950 – Alan Turing proposes a ‘learning machine’ with the potential of becoming artificially intelligent, foreshadowing genetic algorithms.

1951 – Marvin Minsky and Dean Edmonds build SNARC (Stochastic Neural Analog Reinforcement Calculator), the first neural network machine.

1952 – Arthur Samuel begins working on the first machine learning programs, creating programs able to play checkers.

1957 – Frank Rosenblatt invents the *perceptron*, an algorithm used for supervised learning of binary classifiers.

1965 – Alexey Invakhneko introduces *Deep Networks Based on Group Method of Data Handling (GMDH)*, considered the first deep learning systems of the feed-forward multilayer perceptron type.

1967 – Thomas Cover, Peter E. Hart create the nearest neighbor decision rule, which is the start of basic pattern recognition.

1970 – Seppo Linnainmaa formalizes *backpropagation*.

1979 – Students at Stanford University develop a cart able to navigate and avoid obstacles within a room.

1980 – Kunihiko Fukushima publishes his work on *Neocognitron*, a type of artificial neural network from which the convolutional neural networks are later inspired. Nowadays, it is believed that the Neocognitron was the first artificial neural network to deserve the attribute deep and the very first one to incorporate neurophysiological insights.

1981 – Gerald Dejong proposes *Explanation Based Learning*, which allows a computer algorithm to analyze data and create a general rule while discarding unimportant features.

1985 – Terry Sejnowski develops a program able to learn the pronunciations of words similar to a newborn baby.

1986 - David Rumelhart, Geoffrey Everest Hinton and Ronald J. Williams experiment with Seppo Linnainmaa’s reverse mode of automatic differentiation.

1995 – Corinna Cortes and Vladimir Vapnik publish their research on support vector machines (supervised learning models used for classification and regression analysis).

1997 – IBM’s Deep Blue beats the world champion at chess.

1998 – A team led by Yann LeCun releases the MNIST database (a dataset of handwritten digits from American Census), which would shortly become a benchmark for evaluating handwriting recognition.

2002 – Torch, an open-source machine learning library, is released.

2009 – ImageNet, considered the catalyst for the artificial intelligence boom of the 21st century, is released. ImageNet consists of a large visual database and aims to help in visual object recognition software research.

2014 – Facebook’s research department publishes its work on DeepFace, a revolutionary neural network that is able to identify human faces with an accuracy of 97.35%.

2016 – Google’s AlphaGo becomes the first Go program to bean a professional human player.

? Classification

*Classification* tasks come as second nature to humans, having a vast importance in our evolution. The term covers any context in which some form of decision or forecast is made on the basis of available information beforehand and a *classification procedure* is then some formal method for repeatedly making such judgements in new situations. A classification procedure is constructed from a set of data, for which the expected output is known in advance and it is commonly known as *pattern recognition* or *supervised learning*. When the expected output is unknown and the procedure’s aim is to establish the existence of classes or clusters in the data, the procedure is known as *clustering* or *unsupervised learning*.

Some contexts in which a classification task is required include, for example, mechanical procedures for sorting postal packages based on their destination, determining the eligibility of individuals for cred programs based on their financial status and preliminary diagnoses of a patient’s disease while waiting for the definitive test results. Most pressing problems in science, industry and commerce can be treated as a classification problem using complex and extensive data.

The reasoning behind those classification procedures varies for each context, for the examples presented above the arguments for those procedures are as follows:

1. Mechanical classification procedures can be much faster. Human assistance may still be needed in some cases but the work is greatly reduced.
2. Human-specific biases are removed. A bank must take a decision about the eligibility of a person to a credit program based on its financial status and other criterion, whilst a human may include other irrelevant information and turn away customers.
3. Some preliminary diagnoses in the medical field may exclude the need for a surgery that would be required for definitive results.

On the other hand, there are some pressing issues that affect those procedures, such as:

1. Time to learn. In some fast-changing domains, it is required to establish classification rules quickly or dynamically adjust to changes by changing the defined rules.
2. Accuracy. In some environments, accuracy errors may lead to disasters, while in other areas even a 50% accuracy may have benefits.
3. Speed. A link between speed and accuracy is often present, classifiers with higher accuracies falling behind those with lower values when referring to the time it takes in order to give a result.
4. Comprehensibility. Human operators must easily understand the classification procedure in order to intervene when necessary. Complex procedures may lead to interventions of the supervisors even when the result was correct.
   * 1. Class definitions

The nature of classes and their associated definitions ….

2.1 Machine learning