A history of Artificial Neural Networks

Verfasser:

Paul Leitner, BA

1910837299

Abgabedatum:

31. October 2021

Contents

1. Beginnings and Foundational Theory

1.1 Introduction

Today, artificial neural networks are a central building block of the machine learning landscape and are assumed to hold the biggest promise for the budding advent of true artificial intelligence. Neural Networks are integral to a wide array of applications as of the writing of this paper, such as image recognition, voice assistants, natural language processing and more. While debates over the safety and the implications of the looming Artificial Intelligence (AI) singularity are entering the popular dialogue, most potential applications have not yet reached application in business as of yet.

Throughout the history of science, the inner workings of the human mind had been modeled along the most current understanding of outwardly applied science and engineering principles.

1.2 predating 1900

While in 335 BC Greek philosopher Aristotle assumed the brain to be a cooling mechanism for the blood with the seat of intelligence being the heart, the physician Galen posed in his *balloonist theory* that nerves carried fluid as a signal and inflated muscles like balloons.

These from today's standpoint rather cartoonish conceptions of the nervous system gave way to the first indication that electricity played a part when in 1791 Luigi Galvani showed with his famous demonstrations of amphibian limbs that nerves carried electrical impulses from the brain.

In 1906 the Nobel Prize in Physiology or Medicine was awarded to Camillo Golgi and Santiago Ramon y Cajal, Golgi and Cajal (1906), for their description of neurons as the building blocks of the brain. Thus the groundwork from the biological side had been laid.

Concerning the mathematical and engineering side, Charles Babbage's' *Analytical Engine*, Bromley (1982), a steam-powered mechanical general purpose computation machine (which was sadly never completed) was recognized by the 19th century Mathematician Ada Lovelace as revolutionary. She writes:

> *"The Analytical Machine does not occupy common ground with mere 'calculating machines.' It holds a position wholly it's own, and the considerations it suggests are more interesting in their nature [. . .] we mean any process which alters the mutual relation of two or more things, be this relation of what kind it may. This is the most general definition, and would include all subjects in the universe."*

Alluding to what we would today call *turing complete*. However, and probably most importantly, Lady Lovelace already foresaw the question that anticipated the question of artificial intelligence (by almost a century) she further notes:

> *""The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform."*

i.e. a system may only do what it has been explicitly programmed to do, and not *create*. Due to the fact that the first general purpose computer would not be built for several decades, her ideas would remain in the realm of (albeit stunningly brilliant and prescient) theory.

1.3 mid 20th century, Mathematical modeling of neurons and the Perceptron

1.3.1 McCulloch, Pitts and Turing

The first intersection between biological understanding of the brain and logical/mathematical theory arrived when Warren McCulloch and Walter Pitts proposed that the connections between neurons actually constituted a "logical calculus" in 1943 McCulloch and Pitts (1943), which could approximate functions.

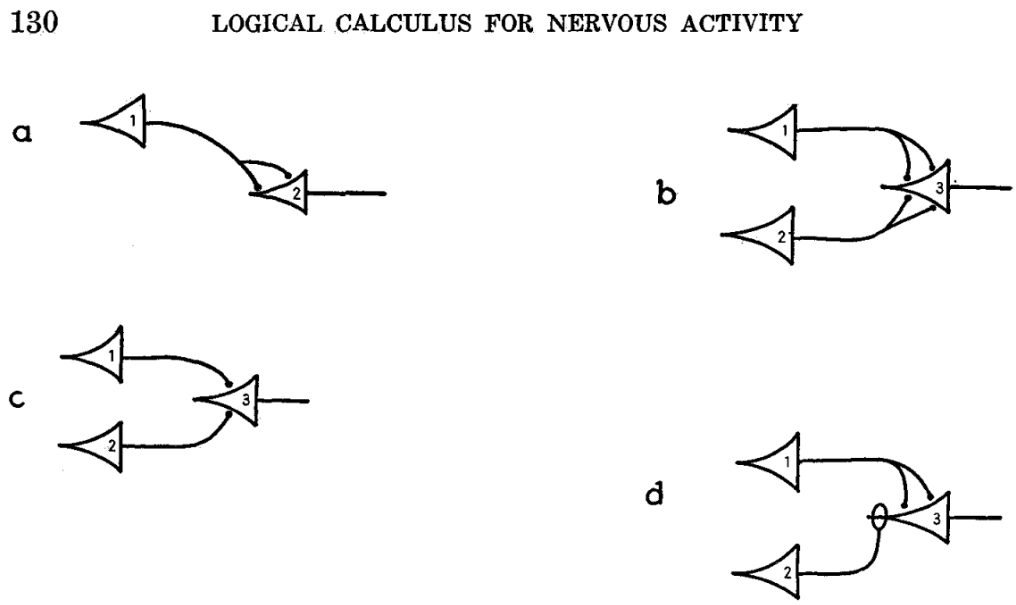

Figure 1: from McCulloch and Pitts (1943)

In their conclusion, McCulloch and Pitts (1943) even allude to the future possibility of understanding of the mind from nervous structure and claim that therefore *"in such systems, 'Mind' no longer 'goes more ghostly than a ghost'"*.

In 1950, meanwhile, Alan Turing argued in his essay "Computing Machinery and Intelligence" that machines in fact could think, and after having stumbled on Babbage and Lovelace's notes, he famously addressed the latter's point in his section on contrary views as **Lady Lovelace's Objection**. In this section, Turing argues that original work produced by humans may merely be due to "a seed planted" in them by learning, and furthermore creative acts may be more sensibly construed as "machines taking us by surprise" which they often do Turing (1950) - the whole essay is really quite well-written, and broader awareness of Turing's arguments would go a long way towards improving our present debate about artificial intelligence and its' challenges.

1.3.2 Rosenblatt, Minsky and Papert

Several years later, in 1958, the psychologist Frank Rosenblatt, inspired by Hebb (1949) and the famous adage derived from his book, "the Organization of Behavior", *neurons that fire together, wire together* proposed the **Perceptron** in his paper Rosenblatt (1958), the basic building block as it is used in neural networks today:

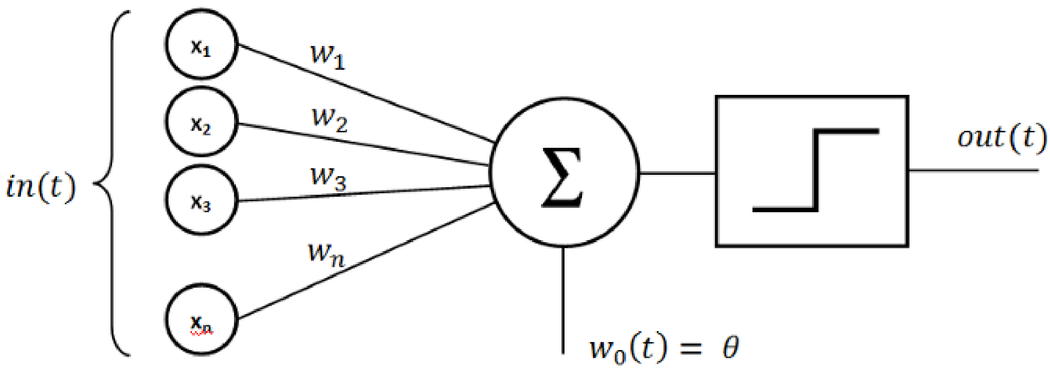

Figure 2: Perceptron Schematic, from wikipedia

as shown here, Rosenblatt's perceptron already exhibited all the hallmarks of modern units of neural networks, multiple inputs coming in with attendant weights, a bias term, summing inputs with the bias and an **activation threshold** which decides whether or not the perceptron "fires" and an output.

Rosenblatt showed that he was able to train perceptrons and perform classification, which garnered a large amount of interest. However, the perceptron proved fairly "brittle" in practice and when Mycielski (1972) showed in their book on perceptrons "Perceptrons: an introduction to computational geometry" that they were unable to emulate a logical XOR gate. A XOR gate, or exclusive or, is a logic gate which returns TRUE if either of the inputs evaluates to TRUE, but not both - it had been an important litmus test in nonlinear classification.

While multiple layers of Perceptrons *were* able to emulate a XOR logic gate, there was no algorithm to train multiple layers of perceptrons.

Arguably Minsky and Papert triggered what is now commonly known as the **first AI winter**.

2. AI Winters and Resurgence

2.1 First AI Winter

Arguably, the initial enthusiasm for his early results had led Rosenblatt to overpromise.

Together with Minsky and Paperts book, Mycielski (1972), "Artificial Intelligence: a general Survey" published in 1973 by James Lighthill, Lighthill (1973), colloquially known as the Lighthill Report, caused public opinion on the field of Artificial Intelligence and its' outlook to drastically swing into the opposite direction.

Ultimately the Lighthill Report bemoaned the practically nonexistent results in the field in contrast with the sums invested. At the heart of the issue seemed to be, at the time, the "combinatorial explosion".

Framing multilayer perceptrons as a **search space** of all possible ideal weights and biases - each perceptron having a number of weights equal to their number of inputs, as seen in 2 - combinatorial explosion refers to the problem that search time in such spaces increases exponentially with the number of parameters.

In the following years, research into the topic quickly dried up.

2.2 Brief revival and Second AI Winter

For a few years there was renewed interest in AI due to the advent of *expert systems*, which were not neural networks in any sense, but knowledge hardcoded by experts, constituting (very) large if-then loops (hence the name).

Therefore funding, especially by DARPA dried up quickly again by the end of the 1980s.

Meanwhile, without much fanfare, after a few years of almost completely nonexistent funding of AI projects, a previously missing piece of the puzzle was discovered - twice.

Both Rumelhart et al. (1985) and Werbos and John (1974) independently described a method for propagating errors back trough a network of layered perceptrons. Thereby, the algorithm to train deep networks was in place.

Ultimately, even though the theoretical groundwork had been mostly laid, AI lay dormant for almost 3 decades.

3. Revival and Recent Explosion

In 2007 Hinton (2009) published his paper on "Deep Belief Networks"; a layered composite model that was used in image recognition. However, the approach still suffered from several limitations. Firstly, using Hinton's method of back-propagation was promising, but the deep belief nets at the time still suffered from what is known as "vanishing gradients". The error that was backpropagated through the network to adjust all parameters grew vanishingly small after propagation through multiple layers - due to the fact that backpropagation involved multiple multiplications by numbers between 0 and 1.

This was due to the fact that backpropagation builds a partial derivative of the loss function with regard to each weight, which was in relation to the slope of the activation function. The most popular activation function at the time, the *sigmoid* activation function, has a gradient approaching 0 at large and small values. This leads to vanishing gradients, or saturated activation functions.

At the time it was standard procedure, therefore, to pre-train lower layers of the network separately, in order adjust their weights.

In 2010 Glorot and Bengio (2010) proposed a new initialization function for weights, a normalized random activation, henceforth the de-facto standard known as glorot-initialization (which is still used by default in packages like keras). This initialization, with the advent of new activation functions such as the hyperbolic tangent activation function, significantly ameliorated the problem of saturation.

Finally in 2012, two more important pieces emerged to kick off the recent explosion off deep learning, which is still ongoing.

Dropout which randomly deactivates a number of neurons of a neural network layer during training in order to make sure training signals "saturate" each region and that neurons do not overfit each other's signals too closely as described by Hinton et al. (2012) (thereby making neural networks much more robust).

Secondly, *RMSPropagation* as described by Hinton (2012), in which batches are split into smaller mini batches and adjust the change in weights based on the root mean square (hence RMS) of the batch, thereby enabling smaller batches of very large datasets being used.

As soon as these pieces were in place, a veritable cambrian explosion of deep learning improvements and technologies followed soon after, the most important of which are:

- BatchNormalization
- Wide & Deep learning Recommender Systems
- Generative Adversarial Networks
- Monte-Carlo Dropout
- Word Embeddings for natural language understanding

4. further reading

Bibliography

Bromley, A. G. (1982). Charles babbage's analytical engine, 1838. *Annals of the History of Computing*, 4(3):196–217.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Golgi, C. and Cajal, S. R. (1906). On the structure of the nervous system.

Hebb, D. O. (1949). *The organisation of behaviour: a neuropsychological theory*. Science Editions New York.

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5):5947.

Hinton, G. E. (2012). Batch gradient descent.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors.

Lighthill, I. (1973). Artificial intelligence: A general survey. In *Artificial Intelligence: A Paper Symposium. London: Science Research Council*.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mycielski, J. (1972). Marvin minsky and seymour papert, perceptrons, an introduction to computational geometry. *Bulletin of the American Mathematical Society*, 78(1):12–15.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Turing, A. M. (1950). Mind. *Mind*, 59(236):433–460.

Werbos, P. and John, P. (1974). Beyond regression : new tools for prediction and analysis in the behavioral sciences /.

List of Acronyms

AI Artificial Intelligence

CNN Convolutional Neural Network

GB Gradient Boosting

NN Neural Network

ML Machine Learning

SMOTE synthetic minority oversampling technique

GAN Generative Adversarial Network

DCGAN Deep Convolutional GAN

EM Earth Mover's Distance

SVC Support Vector Machine Classification

KNN K-Nearest-Neighbor Classification

bagging bootstrap aggregating

GDPR General Data Protection Regulation

HIPAA US Health Insurance Portability and Accountability Act

API Application Programming Interface

SDV Synthetic Data Vault

DS DataSynthesizer

SVM Support Vector Machine

ROC Receiver Operand Characteristic

AUC Area Under the Curve

pypi Python Packaging Index

List of Listings

List of Figures