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 (AI).

Already, 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 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 severed twitching 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): can machines **can machines think and create?**.

On this point she 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.

"Only" foreseeing computation and turing completeness and yet missing the implication of non-human creativity arising from emergent complexity as we see it today in applications such as Generative Adversarial Networks (GANs) (to be touched upon in the final part) can arguably be forgiven.

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.

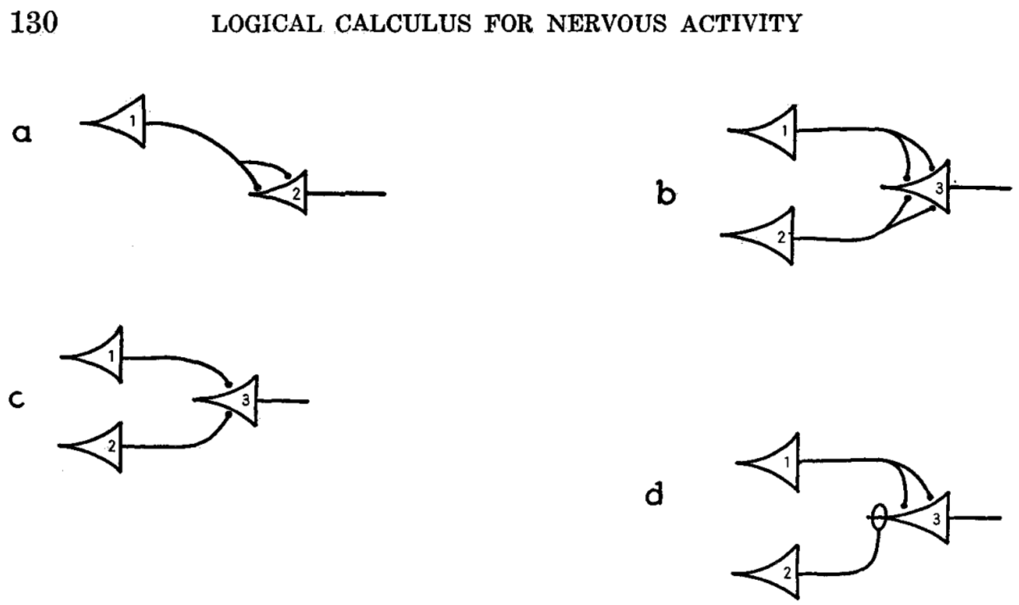

Figure 1: from McCulloch and Pitts (1943)

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.

Importantly however, Turing remarked that Babbage and Lovelace "had not claimed all that they could claim, nor had they reason to" - implying, I think, that they may well have foreseen that machines could create "surprising" (as he termed it) outputs, yet that these did not necessarily constitute true creativity, making their objection a more semantic difference.

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:

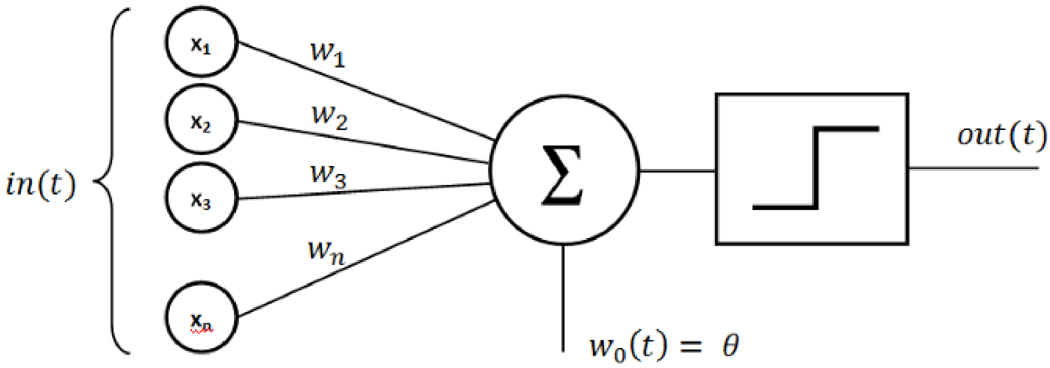

Figure 2: Perceptron Schematic, from wikipedia

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. Specifically, his early successes in training perceptrons to perform classification using his initial training algorithm:

Algorithm 1: Perceptron Learning Algorithm

Input: Training examples $\{\mathbf{x}_i, y_i\}_{i=1}^{m}$.

Initialize $\mathbf{w}$ and b randomly.

while *not converged* **do**

    # # # Loop through the examples.

    for $j = 1, m$ **do**

        # # # Compare the true label and the prediction.

        $error = y_j$ - $\sigma(\mathbf{w}^T\mathbf{x}_j + b)$

        ### If the model wrongly predicts the class, we update the weights and bias.

        if *error != 0* **then**

            ### Update the weights.

            $\mathbf{w} = \mathbf{w} + error \times x_j$

            ### Update the bias.

            $b = b + error$

    Test for convergence

Output: Set of weights $\mathbf{w}$ and bias b for the perceptron.

Figure 3: from Rosenblatt (1958)

which started the first AI boom.

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 (in practice, see below).

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.

Importantly, 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 over-promise.

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.

While some argued that in no branch of science results had been expected as quickly and it was therefore not entirely fair to expect practically usable results from the Perceptron in practice yet, public opinion had turned.

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", a phrase that is still used today.

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, funding and therefore 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 hard-coded by experts, constituting (very) large boolean logic systems encoding knowledge gained from experts (hence the name).

Many companies started inhouse AI departments, leading to enthusiastic coverage once again when in 1984 Business Week even posed in a headline: **AI - it's here!**, via Schuchmann (2019). However, it quickly became apparent that expert systems completely lacked common sense at were not practically viable.

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 backpropagation 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 and reduce their generalization error).

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 followed. Both new improvements to the approach were developed and applications of established research followed, some the most important of which are:

- BatchNormalization - Ioffe and Szegedy (2015)
- Wide & Deep learning for Recommender Systems - Cheng et al. (2016)
- Generative Adversarial Networks - Goodfellow et al. (2014)
- Monte-Carlo Dropout - Gal and Ghahramani (2016)
- Word Embeddings for natural language understanding - Mikolov et al. (2013)
- Recurrent Neural Networks with Long Short Term Memory (LSTM) for long Sequence and time-series processing - Hochreiter and Schmidhuber (1997)
- Convolution and Max Pooling used in Neural Networks for computer vision - Weng et al. (1993)

The advances since 2012 have kicked off a veritable tsunami of research and development of neural networks and the field itself is still very much in flux. Arguably, only a very small percentage of these advances have, in turn, as of yet found their way into application in business and public sectors at all - undoubtedly, many of these improvements will find their business models and will therefore be applied more broadly in various industries.

As for the ever-elusive promise of **true** artificial intelligence, or artificial general intelligence as it is now referred to - some researchers and public intellectuals such as Sam Harris have been very vocal about the risks of the race towards AI which, as they argue, have fallen by the wayside. A good summary of this is his interview with Stuart Russel, co-author of "Artifitial Intelligence, a modern approach" here: https://samharris.org/podcasts/the-dawn-of-artificial-intelligence1/.

Others like Hawkins (2021) in his book "a thousand brains" argue that while AI is coming, it is not cause for alarm, simply because the analogy towards the brain does not completely hold.

While Geoffrey Hinton himself apparently stated that deep networks alone may not be enough for true AI and we need to look towards other ways than backpropagation such as reinforcement learning, transfer learning, genetic algorithms, one-shot learning and other approaches.

In his recent interview with Lex Fridman (https://lexfridman.com/peter-norvig/), Peter Norvig (himself co-author of "Artificial Intelligence: A modern Approach" together with Steward Russel and others, Russell and Norvig (2002)) contrasted the more "probabilistic" approach modeled in deep learning to expert systems based on booleans and binary logic employed in expert systems and noted that for true AI a synthesis of the two would probably be necessary.

Similarly, Ian Goodfellow (creator of GANs) pointed out that while especially LSTM provided something like an internal representation of "knowledge" in a

Neural Network (NN), effectively propagating forward the signal as it is done in the LSTM approach does not actually model short-term memory as it is defined in the brain - he therefore also noted that NNs would have to interact with or completely incorporate a knowledge base in some way (interview found here: https://lexfridman.com/ian-goodfellow/).

While some very promising results have been achieved in steps relating to perception and multi-step processing of multidimensional data in NNs, the search for true AI seems to be missing some fundamental ingredients. Maybe even in the practical application of a theoretically explored approach from decades ago that is simply waiting for its' revival.

It may therefore be taken for granted that the field itself will remain dynamic and vibrant for years to come, even though another AI winter is not out of the realm of possibility - see Schuchmann (2019).

Bibliography

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

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., and Shah, H. (2016). Wide and deep learning for recommender systems.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

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.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Hawkins, J. (2021). *A Thousand Brains: A new Theory of Intelligence*.

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.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality.

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.

Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach.

Schuchmann, S. (2019). *Analyzing the Prospect of an Approaching AI Winter*. PhD thesis, Master's thesis. https://doi. org/10. 13140/RG. 2.2. 10932.91524.

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

Weng, J., Ahuja, N., and Huang, T. (1993). Learning recognition and segmentation of 3-d objects from 2-d images.

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

List of Acronyms

AI Artificial Intelligence

LSTM Long Short Term Memory

NN Neural Network

GAN Generative Adversarial Network

List of Figures