Title: History of artificial neural networks

URL: https://en.wikipedia.org/wiki/History_of_artificial_neural_networks

PageID: 61541925

Categories: Category:Artificial neural networks, Category:Computational neuroscience, Category:Computational statistics

-----

https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (refs: 5, 148)

https://www.degruyter.com/view/books/9781400882618/9781400882618-002/9781400882618-002.xml (refs: 11, 12)

https://arxiv.org/abs/2212.11279 (refs: 26, 70)

https://arxiv.org/abs/1404.7828 (refs: 52, 96)

https://arxiv.org/abs/1411.4555 (refs: 65, 103)

https://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf (refs: 67, 76)

https://arxiv.org/abs/1409.1556 (refs: 101, 106)

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

Artificial neural networks (ANNs) are models created using machine learning to perform a number of tasks . Their creation was inspired by biological neural circuitry . [1] [a] While some of the computational implementations ANNs relate to earlier discoveries in mathematics, the first implementation of ANNs was by psychologist Frank Rosenblatt , who developed the perceptron . [1] Little research was conducted on ANNs in the 1970s and 1980s, with the AAAI calling this period an " AI winter ". [2]

Later, advances in hardware and the development of the backpropagation algorithm, as well as recurrent neural networks and convolutional neural networks , renewed interest in ANNs. The 2010s saw the development of a deep neural network (i.e., one with many layers) called AlexNet . [3] It greatly outperformed other image recognition models, and is thought to have launched the ongoing AI spring , and further increasing interest in deep learning . [4] The transformer architecture was first described in 2017 as a method to teach ANNs grammatical dependencies in language, [5] and is the predominant architecture used by large language models such as GPT-4 . Diffusion models were first described in 2015, and became the basis of image generation models such as DALL-E in the 2020s. [citation needed]

Perceptrons and other early neural networks

The simplest feedforward network consists of a single weight layer without activation functions. It would be just a linear map, and training it would be linear regression. Linear regression by least squares method was used by Adrien-Marie Legendre (1805) and Carl Friedrich Gauss (1795) for the prediction of planetary movement. [6] [7] [8] [9]

A Logical Calculus of the Ideas Immanent in Nervous Activity (Warren McCulloch and Walter Pitts , 1943) studied several abstract models for neural networks using symbolic logic of Rudolf Carnap and Principia Mathematica . The paper argued that several abstract models of neural networks (some learning, some not learning) have the same computational power as Turing machines. [10] This model paved the way for research to split into two approaches. One approach focused on biological processes while the other focused on the application of neural networks to artificial intelligence . This work led to work on nerve networks and their link to finite automata . [11]

In the early 1940s, D. O. Hebb [12] created a learning hypothesis based on the mechanism of neural plasticity that became known as Hebbian learning . Hebbian learning is unsupervised learning . This evolved into models for long-term potentiation . Researchers started applying these ideas to computational models in 1948 with Turing's B-type machines . B. Farley and Wesley A. Clark [13] (1954) first used computational machines, then called "calculators", to simulate a Hebbian network. Other neural network computational machines were created by Rochester , Holland , Habit and Duda (1956). [14]

Frank Rosenblatt [1] (1958) created the perceptron , an algorithm for pattern recognition. A multilayer perceptron (MLP) comprised 3 layers: an input layer, a hidden layer with randomized weights that did not learn, and an output layer. With mathematical notation, Rosenblatt described circuitry not in the basic perceptron, such as the exclusive-or circuit that could not be processed by neural networks at the time. In 1959, a biological model proposed by Nobel laureates Hubel and Wiesel was based on their discovery of two types of cells in the primary visual cortex : simple cells and complex cells . [15] He later published a 1962 book also introduced variants and computer experiments, including a version with four-layer perceptrons where the last two layers have learned weights (and thus a proper multilayer perceptron). [16] : section 16 Some consider that the 1962 book developed and explored all of the basic ingredients of the deep learning systems of today. [17]

Some say that research stagnated following Marvin Minsky and Seymour Papert's Perceptrons (1969). [18]

Group method of data handling , a method to train arbitrarily deep neural networks was published by Alexey Ivakhnenko and Lapa in 1967, which they regarded as a form of polynomial regression, [19] or a generalization of Rosenblatt's perceptron. [20] A 1971 paper described a deep network with eight layers trained by this method. [21]

The first deep learning multilayer perceptron trained by stochastic gradient descent [22] was published in 1967 by Shun'ichi Amari . [23] In computer experiments conducted by Amari's student Saito, a five layer MLP with two modifiable layers learned internal representations to classify non-linearily separable pattern classes. [24] Subsequent developments in hardware and hyperparameter tunings have made end-to-end stochastic gradient descent the currently dominant training technique.

Backpropagation

Backpropagation is an efficient application of the chain rule derived by Gottfried Wilhelm Leibniz in 1673 [25] to networks of differentiable nodes. The terminology "back-propagating errors" was actually introduced in 1962 by Rosenblatt, [16] but he did not know how to implement this, although Henry J. Kelley had a continuous precursor of backpropagation in 1960 in the context of control theory . [26] The modern form of backpropagation was developed multiple times in early 1970s. The earliest published instance was Seppo Linnainmaa 's master thesis (1970). [27] [28] Paul Werbos developed it independently in 1971, [29] but had difficulty publishing it until 1982. [30] In 1986, David E. Rumelhart et al. popularized backpropagation. [31]

Recurrent network architectures

One origin of the recurrent neural network (RNN) was statistical mechanics . The Ising model was developed by Wilhelm Lenz [32] and Ernst Ising [33] in the 1920s [34] as a simple statistical mechanical model of magnets at equilibrium. Glauber in 1963 studied the Ising model evolving in time, as a process towards equilibrium (Glauber dynamics), adding in the component of time. [35] Shun'ichi Amari in 1972 proposed to modify the weights of an Ising model by Hebbian learning rule as a model of associative memory, adding in the component of learning. [36] This was popularized as the Hopfield network (1982). [37]

Another origin of RNN was neuroscience. The word "recurrent" is used to describe loop-like structures in anatomy. In 1901, Cajal observed "recurrent semicircles" in the cerebellar cortex . [38] In 1933, Lorente de Nó discovered "recurrent, reciprocal connections" by Golgi's method , and proposed that excitatory loops explain certain aspects of the vestibulo-ocular reflex . [39] [40] Hebb considered "reverberating circuit" as an explanation for short-term memory. [41] (McCulloch & Pitts 1943) considered neural networks that contains cycles, and noted that the current activity of such networks can be affected by activity indefinitely far in the past.

Two early influential works were the Jordan network (1986) and the Elman network (1990), which applied RNN to study cognitive psychology . In 1993, a neural history compressor system solved a "Very Deep Learning" task that required more than 1000 subsequent layers in an RNN unfolded in time. [42]

LSTM

Sepp Hochreiter 's diploma thesis (1991) [43] proposed the neural history compressor, and identified and analyzed the vanishing gradient problem . [43] [44] In 1993, a neural history compressor system solved a "Very Deep Learning" task that required more than 1000 subsequent layers in an RNN unfolded in time. [45] [42] Hochreiter proposed recurrent residual connections to solve the vanishing gradient problem. This led to the long short-term memory (LSTM), published in 1995. [46] LSTM can learn "very deep learning" tasks [47] with long credit assignment paths that require memories of events that happened thousands of discrete time steps before. That LSTM was not yet the modern architecture, which required a "forget gate", introduced in 1999, [48] which became the standard RNN architecture.

Long short-term memory (LSTM) networks were invented by Hochreiter and Schmidhuber in 1995 and set accuracy records in multiple applications domains. [46] [49] It became the default choice for RNN architecture.

Around 2006, LSTM started to revolutionize speech recognition , outperforming traditional models in certain speech applications. [50] [51] LSTM also improved large-vocabulary speech recognition [52] [53] and text-to-speech synthesis [54] and was used in Google voice search , and dictation on Android devices . [55]

LSTM broke records for improved machine translation , [56] language modeling [57] and Multilingual Language Processing. [58] LSTM combined with convolutional neural networks (CNNs) improved automatic image captioning . [59]

Convolutional neural networks (CNNs)

The origin of the CNN architecture is the " neocognitron " [60] introduced by Kunihiko Fukushima in 1980. [61] [62] It was inspired by work of Hubel and Wiesel in the 1950s and 1960s which showed that cat visual cortices contain neurons that individually respond to small regions of the visual field .

The neocognitron introduced the two basic types of layers in CNNs: convolutional layers, and downsampling layers. A convolutional layer contains units whose receptive fields cover a patch of the previous layer. The weight vector (the set of adaptive parameters) of such a unit is often called a filter. Units can share filters. Downsampling layers contain units whose receptive fields cover patches of previous convolutional layers. Such a unit typically computes the average of the activations of the units in its patch. This downsampling helps to correctly classify objects in visual scenes even when the objects are shifted.

In 1969, Kunihiko Fukushima also introduced the ReLU (rectified linear unit) activation function . [63] [64] The rectifier has become the most popular activation function for CNNs and deep neural networks in general. [65]

The time delay neural network (TDNN) was introduced in 1987 by Alex Waibel and was one of the first CNNs, as it achieved shift invariance. [66] It did so by utilizing weight sharing in combination with backpropagation training. [67] Thus, while also using a pyramidal structure as in the neocognitron, it performed a global optimization of the weights instead of a local one. [66]

In 1988, Wei Zhang et al. applied backpropagation to a CNN (a simplified Neocognitron with convolutional interconnections between the image feature layers and the last fully connected layer) for alphabet recognition. They also proposed an implementation of the CNN with an optical computing system. [68] [69]

Kunihiko Fukushima published the neocognitron in 1980. [70] Max pooling appears in a 1982 publication on the neocognitron. [71] In 1989, Yann LeCun et al. trained a CNN with the purpose of recognizing handwritten ZIP codes on mail. While the algorithm worked, training required 3 days. [72] [73] It used max pooling. Learning was fully automatic, performed better than manual coefficient design, and was suited to a broader range of image recognition problems and image types.

Subsequently, Wei Zhang, et al. modified their model by removing the last fully connected layer and applied it for medical image object segmentation in 1991 [74] and breast cancer detection in mammograms in 1994. [75]

In a variant of the neocognitron called the cresceptron, instead of using Fukushima's spatial averaging, J. Weng et al. also used max-pooling where a downsampling unit computes the maximum of the activations of the units in its patch. [76] [77] [78] [79]

LeNet-5, a 7-level CNN by Yann LeCun et al. in 1998, [80] that classifies digits, was applied by several banks to recognize hand-written numbers on checks (British English : cheques) digitized in 32x32 pixel images. The ability to process higher-resolution images requires larger and more layers of CNNs, so this technique is constrained by the availability of computing resources.

In 2010, Backpropagation training through max-pooling was accelerated by GPUs and shown to perform better than other pooling variants. [81] Behnke (2003) relied only on the sign of the gradient (Rprop) [82] on problems such as image reconstruction and face localization. Rprop is a first-order optimization algorithm created by Martin Riedmiller and Heinrich Braun in 1992. [83]

Deep learning

The deep learning revolution started around CNN- and GPU-based computer vision.

Although CNNs trained by backpropagation had been around for decades and GPU implementations of NNs for years, [84] including CNNs, [85] faster implementations of CNNs on GPUs were needed to progress on computer vision. Later, as deep learning becomes widespread, specialized hardware and algorithm optimizations were developed specifically for deep learning. [86]

A key advance for the deep learning revolution was hardware advances, especially GPU. Some early work dated back to 2004. [84] [85] In 2009, Raina, Madhavan, and Andrew Ng reported a 100M deep belief network trained on 30 Nvidia GeForce GTX 280 GPUs, an early demonstration of GPU-based deep learning. They reported up to 70 times faster training. [87]

In 2011, a CNN named DanNet [88] [89] by Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella , and Jürgen Schmidhuber achieved for the first time superhuman performance in a visual pattern recognition contest, outperforming traditional methods by a factor of 3. [47] It then won more contests. [90] [91] They also showed how max-pooling CNNs on GPU improved performance significantly. [92]

Many discoveries were empirical and focused on engineering. For example, in 2011, Xavier Glorot, Antoine Bordes and Yoshua Bengio found that the ReLU [63] worked better than widely used activation functions prior to 2011.

In October 2012, AlexNet by Alex Krizhevsky , Ilya Sutskever , and Geoffrey Hinton [93] won the large-scale ImageNet competition by a significant margin over shallow machine learning methods. Further incremental improvements included the VGG-16 network by Karen Simonyan and Andrew Zisserman [94] and Google's Inceptionv3 . [95]

The success in image classification was then extended to the more challenging task of generating descriptions (captions) for images, often as a combination of CNNs and LSTMs. [96] [97] [98]

In 2014, the state of the art was training “very deep neural network” with 20 to 30 layers. [99] Stacking too many layers led to a steep reduction in training accuracy, [100] known as the "degradation" problem. [101] In 2015, two techniques were developed concurrently to train very deep networks: highway network [102] and residual neural network (ResNet). [103] The ResNet research team attempted to train deeper ones by empirically testing various tricks for training deeper networks until they discovered the deep residual network architecture. [104]

Generative adversarial networks

In 1991, Juergen Schmidhuber published "artificial curiosity", neural networks in a zero-sum game . [105] The first network is a generative model that models a probability distribution over output patterns. The second network learns by gradient descent to predict the reactions of the environment to these patterns. GANs can be regarded as a case where the environmental reaction is 1 or 0 depending on whether the first network's output is in a given set. [106] It was extended to "predictability minimization" to create disentangled representations of input patterns . [107] [108]

Other people had similar ideas but did not develop them similarly. An idea involving adversarial networks was published in a 2010 blog post by Olli Niemitalo. [109] This idea was never implemented and did not involve stochasticity in the generator and thus was not a generative model. It is now known as a conditional GAN or cGAN. [citation needed] An idea similar to GANs was used to model animal behavior by Li, Gauci and Gross in 2013. [110]

Another inspiration for GANs was noise-contrastive estimation, [111] which uses the same loss function as GANs and which Goodfellow studied during his PhD in 2010–2014.

Generative adversarial network (GAN) by (Ian Goodfellow et al., 2014) [112] became state of the art in generative modeling during 2014-2018 period. Excellent image quality is achieved by Nvidia 's StyleGAN (2018) [113] based on the Progressive GAN by Tero Karras et al. [114] Here the GAN generator is grown from small to large scale in a pyramidal fashion. Image generation by GAN reached popular success, and provoked discussions concerning deepfakes . [115] Diffusion models (2015) [116] eclipsed GANs in generative modeling since then, with systems such as DALL·E 2 (2022) and Stable Diffusion (2022).

Attention mechanism and Transformer

The human selective attention had been studied in neuroscience and cognitive psychology. [117] Selective attention of audition was studied in the cocktail party effect (Colin Cherry , 1953). [118] (Donald Broadbent , 1958) proposed the filter model of attention . [119] Selective attention of vision was studied in the 1960s by George Sperling 's partial report paradigm . It was also noticed that saccade control is modulated by cognitive processes, in that the eye moves preferentially towards areas of high salience . As the fovea of the eye is small, the eye cannot sharply resolve all of the visual field at once. The use of saccade control allows the eye to quickly scan important features of a scene. [120]

These researches inspired algorithms, such as a variant of the Neocognitron . [121] [122] Conversely, developments in neural networks had inspired circuit models of biological visual attention. [123] [124]

A key aspect of attention mechanism is the use of multiplicative operations, which had been studied under the names of higher-order neural networks , [125] multiplication units , [126] sigma-pi units , [127] fast weight controllers , [128] and hyper-networks . [129]

Recurrent attention

During the deep learning era, attention mechanism was developed solve similar problems in encoding-decoding. [130]

The idea of encoder-decoder sequence transduction had been developed in the early 2010s. The papers most commonly cited as the originators that produced seq2seq are two papers from 2014. [131] [132] A seq2seq architecture employs two RNN, typically LSTM, an "encoder" and a "decoder", for sequence transduction, such as machine translation. They became state of the art in machine translation, and was instrumental in the development of attention mechanism and Transformer .

An image captioning model was proposed in 2015, citing inspiration from the seq2seq model. [133] that would encode an input image into a fixed-length vector. (Xu et al. 2015), [134] citing (Bahdanau et al. 2014), [135] applied the attention mechanism as used in the seq2seq model to image captioning.

Transformer

One problem with seq2seq models was their use of recurrent neural networks, which are not parallelizable as both the encoder and the decoder processes the sequence token-by-token. The decomposable attention attempted to solve this problem by processing the input sequence in parallel, before computing a "soft alignment matrix" ("alignment" is the terminology used by (Bahdanau et al. 2014) [135]). This allowed parallel processing.

The idea of using attention mechanism for self-attention, instead of in an encoder-decoder (cross-attention), was also proposed during this period, such as in differentiable neural computers and neural Turing machines . [136] It was termed intra-attention [137] where an LSTM is augmented with a memory network as it encodes an input sequence.

These strands of development were combined in the Transformer architecture, published in Attention Is All You Need (2017). Subsequently, attention mechanisms were extended within the framework of Transformer architecture.

Seq2seq models with attention still suffered from the same issue with recurrent networks, which is that they are hard to parallelize, which prevented them to be accelerated on GPUs. In 2016, decomposable attention applied attention mechanism to the feedforward network , which are easy to parallelize. [138] One of its authors, Jakob Uszkoreit, suspected that attention without recurrence is sufficient for language translation, thus the title "attention is all you need". [139]

In 2017, the original (100M-sized) encoder-decoder transformer model was proposed in the " Attention is all you need " paper. At the time, the focus of the research was on improving seq2seq for machine translation , by removing its recurrence to processes all tokens in parallel, but preserving its dot-product attention mechanism to keep its text processing performance. [140] Its parallelizability was an important factor to its widespread use in large neural networks. [141]

Unsupervised and self-supervised learning

Self-organizing maps

Self-organizing maps (SOMs) were described by Teuvo Kohonen in 1982. [142] [143] SOMs are neurophysiologically inspired [144] artificial neural networks that learn low-dimensional representations of high-dimensional data while preserving the topological structure of the data. They are trained using competitive learning .

SOMs create internal representations reminiscent of the cortical homunculus , a distorted representation of the human body , based on a neurological "map" of the areas and proportions of the human brain dedicated to processing sensory functions , for different parts of the body.

Boltzmann machines

During 1985–1995, inspired by statistical mechanics, several architectures and methods were developed by Terry Sejnowski , Peter Dayan , Geoffrey Hinton , etc., including the Boltzmann machine , [145] restricted Boltzmann machine , [146] Helmholtz machine , [147] and the wake-sleep algorithm . [148] These were designed for unsupervised learning of deep generative models. However, those were more computationally expensive compared to backpropagation. Boltzmann machine learning algorithm, published in 1985, was briefly popular before being eclipsed by the backpropagation algorithm in 1986. (p. 112 [149]).

Geoffrey Hinton et al. (2006) proposed learning a high-level internal representation using successive layers of binary or real-valued latent variables with a restricted Boltzmann machine [150] to model each layer. This RBM is a generative stochastic feedforward neural network that can learn a probability distribution over its set of inputs. Once sufficiently many layers have been learned, the deep architecture may be used as a generative model by reproducing the data when sampling down the model (an "ancestral pass") from the top level feature activations. [151] [152]

Deep learning

In 2012, Andrew Ng and Jeff Dean created an FNN that learned to recognize higher-level concepts, such as cats, only from watching unlabeled images taken from YouTube videos. [153]

Other aspects

Knowledge distillation

Knowledge distillation or model distillation is the process of transferring knowledge from a large model to a smaller one. The idea of using the output of one neural network to train another neural network was studied as the teacher-student network configuration. [154] In 1992, several papers studied the statistical mechanics of teacher-student network configuration, where both networks are committee machines [155] [156] or both are parity machines. [157]

Another early example of network distillation was also published in 1992, in the field of recurrent neural networks (RNNs). The problem was sequence prediction. It was solved by two RNNs. One of them ("atomizer") predicted the sequence, and another ("chunker") predicted the errors of the atomizer. Simultaneously, the atomizer predicted the internal states of the chunker. After the atomizer manages to predict the chunker's internal states well, it would start fixing the errors, and soon the chunker is obsoleted, leaving just one RNN in the end. [158]

A related methodology was model compression or pruning , where a trained network is reduced in size. It was inspired by neurobiological studies showing that the human brain is resistant to damage, and was studied in the 1980s, via methods such as Biased Weight Decay [159] and Optimal Brain Damage. [160]

Hardware-based designs

The development of metal–oxide–semiconductor (MOS) very-large-scale integration (VLSI), combining millions or billions of MOS transistors onto a single chip in the form of complementary MOS (CMOS) technology, enabled the development of practical artificial neural networks in the 1980s. [161]

Computational devices were created in CMOS , for both biophysical simulation and neuromorphic computing inspired by the structure and function of the human brain. Nanodevices [162] for very large scale principal components analyses and convolution may create a new class of neural computing because they are fundamentally analog rather than digital (even though the first implementations may use digital devices). [163]

Notes

References

External links

"Lecun 2019-7-11 ACM Tech Talk" . Google Docs . Retrieved 2020-02-13 .