Deep learning, past present and future: An odyssey

Anwaar Ulhaq, *Member, IEEE*,
Machine Vision and Digital
Health Research Group,
Charles Sturt University, NSW,
Australia
aulhaq@csu.edu.au

***Abstract*—Machine learning has grown in popularity and effectiveness over the last decade. It has become possible to solve complex problems, especially in artificial intelligence, due to the effectiveness of deep neural networks. While numerous books and countless papers have been written on deep learning, new researchers want to understand the field's history, current trends and envision future possibilities. This review paper will summarise the recorded work that resulted in such success and address patterns and prospects.**

***Index Terms*—Deep Learning, Multitask Learning, Self-supervised learning, Contrastive Learning, Meta-Learning, Aderserial Learning.**

I. Introduction

Deep learning is an umbrella term for a wide variety of machine learning techniques that learn from large amounts of data. The deep learning algorithms are essentially artificial neural networks that learn from data over and over again, each time fine-tuning the task a little more. We refer to "deep learning" because neural networks have multiple layers that offer complex learning. In "deep learning," the word "deep" refers to the number of layers through which the data is processed. Deep learning allows machines to make the best decision possible for a wide variety of complex, diverse, unstructured and interconnected data sets. We present some of the formal definitions of deep learning:

Definition 1: According to [1], "Deep learning is about learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features." Deep learning (also known as deep structured learning or hierarchical learning) is learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.

Definition 2: It's deep if it has more than one stage of non-linear feature transformation [2].

Definition 3: Deep Learning builds a system by assembling parameterised modules into a (possibly dynamic) computation graph and training it to perform a task by optimising the parameters using a gradient-based method [3].

Deep learning is considered a subset of machine learning and artificial intelligence. The Venn diagram in Figure 1 is showing the relationships between deep learning and other approaches in AI.

Due to its superior performance and accuracy, deep learning has become increasingly popular. Figure 2 shows a Google Trend graph that displays an increasing interest over time in the term "Deep Learning". The Turing Award, presented for excellence in artificial intelligence research, was awarded to three of deep learning's most influential architects, Yann LeCun, Geoffrey Hinton, and Yoshua Bengio. The members of this trio and many of their colleagues have developed the algorithms, systems, and techniques behind the recent AI-fueled wave of products and services, which they expect will dominate the future.

This review paper is organised as follows: In section 2, we provide a development panorama and timeline of deep learning in two episodes of its first emergence as artificial neural networks and reemergence in the revolution of deep learning. Section 3 presents a discussion and future research directions followed by a conclusion and references.

Due to its superior performance and accuracy, deep learning has become increasingly popular. Figure 2 shows a Google Trend graph that displays an increasing interest over time in the term "Deep Learning". The Turing Award, presented for excellence in artificial intelligence research, was awarded to three of deep learning's most influential architects, Yann LeCun, Geoffrey Hinton, and Yoshua Bengio. The members of this trio and many of their colleagues have developed the algorithms, systems, and techniques behind the recent AI-fueled wave of products and services, which they expect will dominate the future.

II. On deep learning odyssey

The story of deep learning is inspiring. It started with the evolution of neuroscience and computing when neuroscientists and, subsequently, computer scientists began thinking of ways to let the computer behave like the human brain and include this kind of intelligence in our everyday life.

We are born to function in a world of complexity and decisions. Our ability to think makes us intelligent creature. However still, the full functionality of our human brain has yet to be fully understood. Neuroscientists want to know how the human brain works, logicians and mathematicians like to investigate the decision-making process, and computer scientists develop mechanisms for computers that learn.

Human anatomists have been curious to know the functionality and structure of the brain. Initially, they dissected the brain to know what the brain is made of. However, even the microscope view was not that useful due to the complex structure of the brain. Camillo Golgi, an Italian

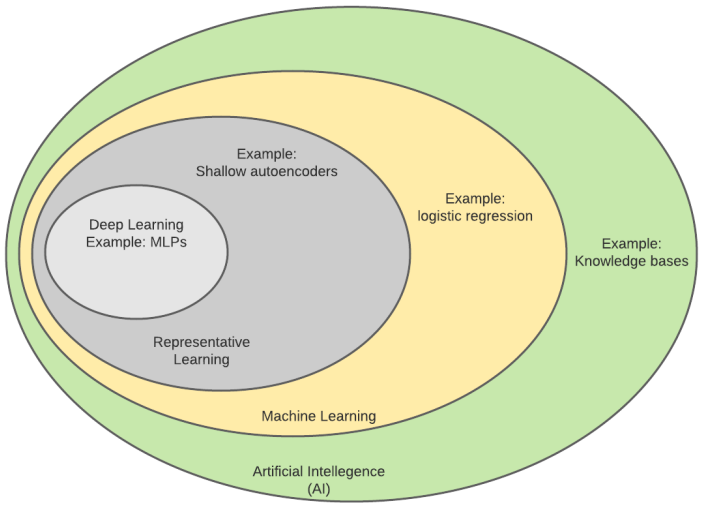

Fig. 1. A Venn diagram showing the relationships between deep learning, which is a kind of representation learning, which in turn is a form of machine learning, which is used for some, not all, of the approaches to AI. Each alternative of the Venn diagram includes explanations of the applications of AI. [4].

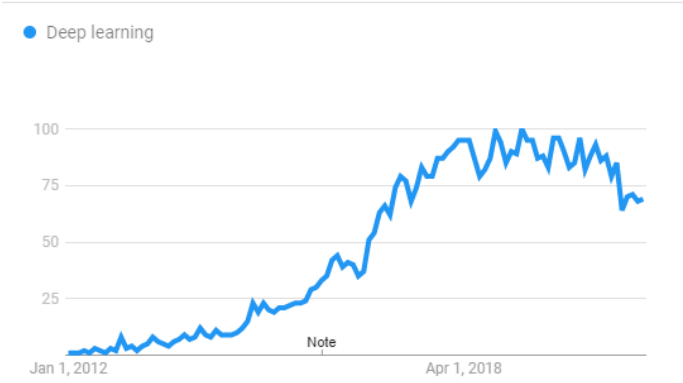

Fig. 2. Interest over time in term "Deep Learning" by Google Trends [5]: Numerals represent the degree of search interest above the chart's highest point for the given time and geographic region. A value of 100 represents the zenith of popularity for the term. The term "half as popular" has a value of 50, which means that it is half as popular. If there was not enough data for this term, then it scored 0.

biologist and pathologist, discovered a method of staining nervous tissues in 1873 [6]. He first used potassium dichromate to harden the tissue and then used silver nitrate to extract desired structure. The silver chromate particles created a black deposit on the soma (nerve cell body) and the axon and all dendrites, providing an exceedingly clear and well-contrasted picture of the neuron against a yellow background [7][293]. It exposed the structure of the nerve cells in the brain for the first time. Later on, Santiago Ramón y Cajal, a Spanish neuroscientist, used Golgi's black strain method to illustrate the arborizations ("tree growing") of brain cells in hundreds of his drawings [8], [9]. He and Camillo Golgi received the Nobel Prize in Physiology or Medicine in 1906 for their original investigations of the brain's microscopic structure made him a pioneer of modern neuroscience [10].

We can divide it into two epochs as described below:

A. The emergence of neural nets and learning rules

Inspired by structural designs of the brain by neuroscientists, Warren McCulloch and Walter Pitts, a neuroscientist and a logician, developed a biological neuron's first mathematical model (M-P Neuron). The paper, "A Logical Calculus of the ideas Imminent in Nervous Activity," [11] was published in 1943. M-P Neuron was designed out of the simple idea of thresholding. The main idea was to present a McCulloch-Pitts model to limit a noisy sigmoid function so that it worked for all data. The threshold control for neuron firing behaviour was named network learning. This network provides the same importance to each of the inputs, so there is no weight for each input in this first artificial neural network.

The first formal learning rule for neural nets was suggested in 1949 by Hebb in his book "The Organization of Behaviour." [12] The concept is based on neural plasticity and is summarised as "Cells that fire together, wire together.". The paper discussed how the strength of a connection between neurons relates to their scalar value. If the weight from neuron 1 to neuron 2 is greater, it indicates that neuron 1 has greater influence over neuron 2. The learning of these weights is more important to reduce the error (difference in desired and actual outcome. There is a rule of thumb to adjust these weights, called Learning Hebb's Rule, that states- "If two neurones in a network are triggered and deactivated simultaneously. This then forces the synaptic efficacy of these interneurons to increase". A crude example to understand is of two persons standing at a distance; if they both say hi to each other, their connections should be stronger than if one or none of them says hi to each other. However, Hebb's rule had issues. Specifically, it lacks a means of weakening connections, and there is no upper limit to the strength of connections.

Meanwhile, Frank Rosenblatt, a Cornell University psychologist, started studying the neural pathways present in the eye of a fly, which led to the idea of a Perceptron in 1958 [13]. Perceptron begins with a random set of weights. It represents

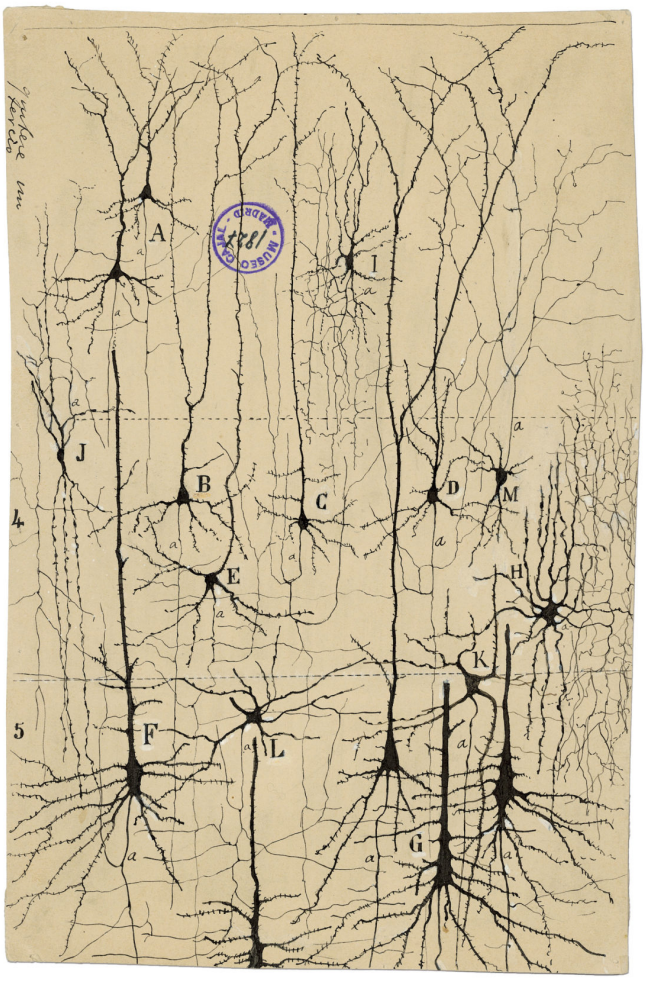

Fig. 3. A drawing done by Cajal of neurons that shows them as separate, individual cells

the error rate as the sum of the square of errors occurring for each individual in the learning sample, then its learning function adjusts the weights according to changes in the error function. Perceptron approached linear problems but failed in distinguishing problems. The intuition involves a dot product of two vectors and shifting the vectors to vary the degree to which they are oriented. This weight function implementation was done on a hardware package known as Perceptron Mark-1.

The development of Perceptron inspired a newfound interest in artificial networks for solving real-world problems. In 1959, Bernard Widrow and Marcian Hoff proposed a major development called the Delta Rule [14]. This update modifies the connection weights to account for the difference between the target and the output. In other words, the node weight change is equal to the product of error and input. The Delta Rule attempts to avoid weight changes if very small or zero error and increases as weight changes increase. Therefore, the idea of model calibration is to minimise the error between the machine learning output and the target vector that would later change the overall machine learning goal. It established the ADALINE and MADALINE systems to eliminate static in telephones and subsequent development of other networks [15].

In 1960, Henry J. Kelley demonstrated the first continuous backpropagation model in his paper "Gradient Theory of Optimal Flight Paths." [16] His model is consistent with Control Theory, but it will be refined and used for artificial neural networks in the future. In 1962, in his paper, "The Numerical Solution of Variational Problems," [17] The backpropagation algorithm used by Stuart Dreyfus illustrates how to use a derivative chain rule, instead of dynamic programming, to train an artificial neural network [18].

In 1965, working in tandem with Valentin Grigorievich Lapa, Alexey Grigoryevich Ivakhnenko created a hierarchical neural network that uses polynomial activation functions and is trained using the Group Method of Data Handling (GMDH). [19] It was the first multi-layer perceptron, and Ivakhnenko is credited with laying the foundation for deep learning.

In 1969, In their book "Perceptrons," Marvin Minsky and Seymour Papert demonstrated that Rosenblatt's perceptron could not solve difficult mathematical problems like XOR. This setback initiates the long and bitter winter of neural network research [20]. In 1970, the general method for automatic differentiation by Seppo Linnainmaa, including backpropagation, is published in addition to which he also incorporates backpropagation in computer code. [21] Backpropagation has advanced greatly, but it could not be implemented in neural networks until the next decade.

In 1971, Alexey Grigoryevich Ivakhnenko created an 8-layer Deep neural network using the Group Method of Data Handling (GMDH). In 1980, Kunihiko Fukushima came up with Neocognitron, the first convolutional neural network architecture which could recognise visual patterns such as handwritten characters.

Based on his 1974 PhD thesis, Paul Werbos publicly advocates the use of Backpropagation for propagating errors during Neural Network training [22]. His PhD thesis findings will potentially contribute to the neural network community's functional implementation of backpropagation in the future. Hopfield Network [23] is nothing more than a recurrent neural network developed by John Hopfield. It's a content-addressable memory device that'll be useful for future RNN models in the modern deep learning age.

The Boltzmann machine [24], a stochastic recurrent neural network, was developed by David H. Ackley, Geoffrey Hinton, and Terrence Sejnowski in 1985. There is no output layer in this neural network, just an input layer and a hidden layer.

In 1986, In their paper Learning Representations through Back-Propagating Mistakes," Geoffrey Hinton, Rumelhart, and Williams demonstrate the effective implementation of back-propagation in the neural network [25]. It made it possible to easily train complex deep neural networks, which was previously a big roadblock in this study field. In the same year, Paul Smolensky proposes a Boltzmann Machine variant in which there is no intralayer relation between the input and hidden layers. Restricted Boltzmann Machine is the name given to this machine (RBM) [26]. It would become popular in the coming years, especially for creating recommender systems. Terry Sejnowski develops NeTalk, a neural network

that learns to pronounce written English text by being fed text and comparing it to phonetic transcriptions [27].

Sepp Hochreiter discovered the vanishing gradient problem in 1991 [28], which makes deep neural network learning incredibly slow and nearly impossible. For many years to come, this topic would annoy the deep learning community.

Yann LeCun et al. applied the standard backpropagation algorithm to a deep neural network in 1989 to recognise handwritten ZIP codes on mail, which had been around as the reverse mode of automatic differentiation since 1970 [29]. Although the algorithm worked, it took three days to practise. In the same year, George Cybenko publishes "Approximation by superpositions of a sigmoidal function," which includes the first version of the Universal Approximation Theorem [30]. He argues that continuous functions can be approximated by feed-forward neural networks with a single hidden layer, which is built on a neural network containing a finite number of neurones. Deep learning is accepted as being even more credible as a result.

A landmark study with the concept of Long Short-Term Memory was published in 1997 by Sepp Hochreiter and Jürgen Schmidhuber (LSTM) [31]. This form of recurrent neural network architecture will profoundly affect future deep learning technologies.

Lator on, it was discussed that multi-layered neural networks do not extend their function set (which determines a neural network's ability to learn) and that neural networks can not be trained in a feasible time (resulted from both training algorithm and computing capacity). Imagine you have trained a deep neural net for one or two weeks, but it didn't get the right answers. On the other hand, other Machine Learning methods like Random Forest [32] and SVM [33] performed better than ANN, which implies that the latter is not as promising.

B. Deep Learning: The re-emergence of neural nets

Researchers believed that ANN would work eventually and kept trying. Meanwhile, thanks to the availability of big data, the improvement of both computing (GPU) and network architecture testing, researchers became able to examine deep network architecture in a relatively short time The term 'Deep Learning' was first introduced to the machine learning community by Rina Dechter in 1986 [34] [29] and to artificial neural networks by Igor Aizenberg and colleagues in 2000 as a method for deriving decision thresholds in Boolean systems [35].

In 2006, Geoffrey Hinton et al. published a paper in which they described building RBMs by stacking multiple layers on top of each other and referring to them as Deep Belief Networks [36]. With a large amount of data, the training process is far more efficient.

In 2008, GPUs usage was largely advocated to quickly train Deep Neural Networks by many magnitude orders [37], [38]. It led to GPU practicality in Deep Learning about training on large volumes of data. The modern deep learning era began in 2009 when Stanford University created ImageNet [39]. This large training dataset provided researchers with the ability to build computer vision algorithms and thus assisted in developing similar approaches for natural language processing and other critical AI technologies. This competition created a new world of opportunities where friendly efforts paved the way for the new AI revolution. Finding labelled data has always been a difficult problem for the Deep Learning community. Since then, it has served as a reference point for the many deep learning researchers who participate in the annual ImageNet competitions (ILSVRC) [40].

In 2011, GPU speed had increased significantly, allowing convolutional neural networks to be trained "without" layer-by-layer pre-training [41]. As computing speeds increased, it became obvious that Deep Learning offered considerable advantages in terms of efficiency and speed. An illustrative example is AlexNet [42], a convolutional neural network that won several international competitions between 2011 and 2012, employing its architecture. Refined linear units were employed to improve the top speed and dropout. The paper "Deep Sparse Rectifier Neural Networks" [43] by Yoshua Bengio, Antoine Bordes, and Xavier Glorot shows that the ReLU activation function can avoid the vanishing gradient problem. It means that apart from GPUs, the deep learning community now has another tool to circumvent the challenge of training deep neural networks for long periods.

In 2012, AlexNet [42], a GPU-based CNN model created by Alex Krizhevsky, takes first place in Imagenet's image classification competition with a winning accuracy of 84%. It's a huge leap from the 75% accuracy models had previously obtained. This victory signifies the beginning of a worldwide expansion of deep learning. Google Brain also published the results of an unusual project known as the "Cat Experiment" in 2012 [44]. The free-spirited project studied "unsupervised learning" and how challenging it can be. The cat experiment utilised a neural network distributed across 1,000 computers. A billion "unlabeled" images were randomly picked from YouTube, shown to the system, and then the training software was given the green light to proceed. At the end of the training, the strongest response to the images of cats was observed in the highest layer neurone. It turns out that the Cat Experiment does about 70% better than previous similar experiments at processing unlabeled images. Despite these shortcomings, however, it recognised only 16% of the objects used for training and performed even worse with rotated or moved objects.

GAN, also known as Generative Adversarial Neural Network [45], is created in 2014 by Ian Goodfellow. Using GANs, new doors in the application of deep learning in fashion, art, and science are opened. This is due to its ability to synthesise real-world data; Facebook introduced DeepFace [46] [41], which uses deep learning to automatically identify and tag Facebook users in photographs, in 2015. Computational algorithms provide superior face recognition capabilities using deep neural networks with 120 million parameters to take into account [47]. DeepID2 achieved 99.15% on the Labeled Faces in the Wild (LFW) dataset, [48] , which is the better-than-human performance of 97.53%. In 2016: AlphaGo [49], an algorithm developed by Google DeepMind, masterfully plays the complex board game Go and bests professional go player

Lee Sedol in a highly publicised tournament in Seoul.

Geoffrey Hinton, Yann LeCun, and Yoshua Bengio receive the Turing Award in 2019 for their enormous contributions to advancements in the field of deep learning and artificial intelligence [50]. This is a defining moment for those who were involved in neural networks during the 1970s, when the machine learning community shifted away from neural networks.

III. Discussion: DL capabilities and Future Directions

Overall, we have seen a remarkable success profile of deep learning. The emergence of big data and powerful computing has revolutionalised the field. The interest and attention of the deep learning research community, academia and industry sponsorship has contributed towards the developments of applied deep learning in business and industry. The last decade has seen the development of AlexNet [42], ResNet [51], GAN [45], Deep Q-learning [52], and Transformer networks [53]. There is excitement about new learning strategies like self-supervised learning [54], contrastive learning [55], and meta-learning [56].

We have seen that fully adequate methods for deep learning turn out to be data-intensive. Data efficient deep understanding has become a very common debatable issue: Few-Shot learning [57], where the training dataset contains limited information, has come to the rescue. A shot is simply one example to learn from. Zero-shot learning [58] aims to identify unseen groups, even though we've never seen them before. One-shot learning [59] means that we have only one instance of each class. Now the job is to identify the correct test image type using that restriction. Siamese Neural Networks [60] brought about significant change, which helped propel us to even better results.

An important aspect of intelligence is to deal with versatility – the capability of doing many different things. Intelligent agents need to learn how to learn new tasks faster by leveraging previous experience rather than considering each new task in isolation. This learning approach, or meta-learning [56], is a key step towards developing versatile agents that can continually learn a wide variety of tasks throughout their lifetimes. Recently meta-learning has become a hot topic, with a flurry of recent papers claiming improved performance on existing tasks [61], [62].

However, still in the age of big data, unlabeled data is being generated all the time. The development of a dataset with cleanly assigned labels is costly. To make use of this broader data collection appropriate learning goals to supervise the data, Facebook AI Scientist Yann LeCun presented his "cake'analogy'" during NIPS 2016 [63]."If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning (RL)." [63]

Three organisations are leading AGI research; OpenAI aims to achieve or promote artificial intelligence (and ensure that it is responsibly used). Similar targets are being pursued by Google's DeepMind and the Human Brain Project [64].

OpenAI demonstrated a robotic hand that was completely trained in simulation to manipulate objects into different orientations in mid-2018 [65]. Even though the tasks seem to be straightforward, the most important thing it accomplished was the ability to do well in unfamiliar situations despite never having been explicitly trained to behave in such cases. This was accomplished using Domain Randomisation [66], a training technique that enabled the system to recognise the environment's key features and generalise to new situations.

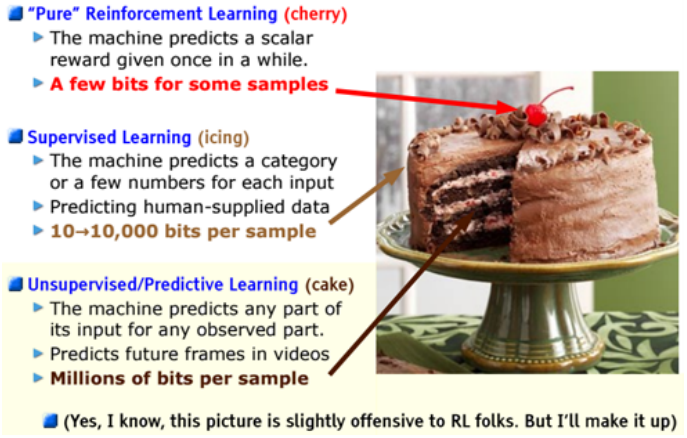

Fig. 4. Original LeCun cake analogy slide presented at NIPS 2016, the highlighted area has now been updated. Source [63]

The self-supervised task is also known as the pretext task. The pretexts task helps us to a supervised loss function. Generally, however, we do not focus on the long-term outcome of this manufactured mission. In addition, we're interested in the learned intermediate representation, which carries semantic or structural meaning and can be useful to a number of downstream tasks. For example, we can do this with images by rotating and training a model to predict which input image is next. Contrastive learning [55] has also emerged as a promising technique for self-supervised learning.

The principles of quantum computing naturally lend themselves to machine learning [67]. As a result, there is ongoing research to utilise these principles in the area of ML representation power and computational efficiency. Quantum computing, which utilises qubits, relies on the qubit as the basic unit of information. Quantum computing's superiority over classical computing is due to superposition and entanglement, which qubits exhibit. Quantum neural network (QNN) as a quantum analogue to a neural networks is emerging [68]. Based on the theories of convolutional and pooling layers used in classical CNNs, quantum CNNs are proposed [52] [69], A quantum CNN uses a quantum circuit model that takes inspiration from both classical CNNs and convolutional and pooling layers. Cloud and edge computing are providing the availability and scalability of broader quantum computing [70]. New platforms will contribute to the feasibility of quantum computing. The next decade is expected to see the emergence of a new type of algorithm that takes advantage of quantum computing's advantages and utilises its strengths. However, a dream of attaining AGI is still far away.

IV. Conclusion

This overview paper takes readers on a journey through the history, current, and future of deep learning. While significant progress has been made in solving various AI problems, the field faces an enormous challenge in achieving artificial general intelligence. Researchers must address several challenges related to data-efficient learning and improved performance of deep learning systems to accomplish this feat.

References

[1] WJ Zhang, Guosheng Yang, Yingzi Lin, Chunli Ji, and Madan M Gupta. On definition of deep learning. In *2018 World automation congress (WAC)*, pages 1–5. IEEE, 2018.

[2] Guillaume Chassagnon, Maria Vakalopolou, Nikos Paragios, and Marie-Pierre Revel. Deep learning: definition and perspectives for thoracic imaging. *European radiology*, pages 1–10, 2019.

[3] Aaai 2020 — a turning point for deep learning? hinton, lecun, and bengio might have different approaches — synced. https://syncedreview.com/2020/02/10/aaai-2020-whats-next-for-deep-learning-hinton-lecun-and-bengio-share-their-visions/. (Accessed on 04/15/2021).

[4] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[5] deep learning - explore - google trends. https://trends.google.com/trends/explore?date=all&geo=US&q=deep%20learning. (Accessed on 04/17/2021).

[6] Paolo Mazzarello et al. *The hidden structure: A scientific biography of Camillo Golgi*. Oxford University Press on Demand, 1999.

[7] Henry R Viets. Camillo golgi, 1843-1926. *Archives of Neurology & Psychiatry*, 15(5):NP–627, 1926.

[8] Larry W Swanson, Eric Newman, Alfonso Araque, and Janet M Dubinsky. *The beautiful brain: the drawings of Santiago Ramón y Cajal*. Abrams, 2017.

[9] Constantino Sotelo. Viewing the brain through the master hand of ramón y cajal. *Nature Reviews Neuroscience*, 4(1):71–77, 2003.

[10] Francisco López-Muñoz, Jesús Boya, and Cecilio Alamo. Neuron theory, the cornerstone of neuroscience, on the centenary of the nobel prize award to santiago ramón y cajal. *Brain research bulletin*, 70(4-6):391–405, 2006.

[11] Cornelius Borck. Toys are us. *Critical Neuroscience*, pages 111–133, 2012.

[12] Donald Olding Hebb. The organization of behavior; a neuropsychological theory. *A Wiley Book in Clinical Psychology*, 62:78, 1949.

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

[14] F. Lara. Artificial neural networks: an introduction. *Instrumentation and development*, 3:5–10, 1998.

[15] Capt Rodney Winter and B Widrow. Madaline rule ii: A training algorithm for neural networks. In *Secondf Annual International Conference on Neural Networks*, pages 1–401, 1988.

[16] Henry J Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

[17] Stuart Dreyfus. The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications*, 5(1):30–45, 1962.

[18] Stuart E Dreyfus. Artificial neural networks, back propagation, and the kelley-bryson gradient procedure. *Journal of guidance, control, and dynamics*, 13(5):926–928, 1990.

[19] Stanley J Farlow. *Self-organizing methods in modeling: GMDH type algorithms*. CrC Press, 2020.

[20] Minsky Marvin and A Papert Seymour. Perceptrons, 1969.

[21] Jürgen Schmidhuber. Who invented backpropagation? *More in [DL2]*, 2014.

[22] Paul John Werbos. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*, volume 1. John Wiley & Sons, 1994.

[23] John J Hopfield. Hopfield network. *Scholarpedia*, 2(5):1977, 2007.

[24] Geoffrey E Hinton. Boltzmann machine. *Scholarpedia*, 2(5):1668, 2007.

[25] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[26] Ilya Sutskever, Geoffrey E Hinton, and Graham W Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in neural information processing systems*, pages 1601–1608, 2009.

[27] Terrence J Sejnowski and Charles R Rosenberg. Parallel networks that learn to pronounce english text. *Complex systems*, 1(1):145–168, 1987.

[28] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[29] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[30] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[33] Vladimir Vapnik, Isabel Guyon, and Trevor Hastie. Support vector machines. *Mach. Learn*, 20(3):273–297, 1995.

[34] Rina Dechter. Learning while searching in constraint-satisfaction problems. 1986.

[35] IN Aizenberg, NN Aizenberg, and J Vandewalle. Multi-valued and universal binary neurons: Theory, learning, and applications. *IEEE Transactions on Neural Networks*, 12(3):647, 2001.

[36] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[37] Adam Coates, Paul Baumstarck, Quoc Le, and Andrew Y Ng. Scalable learning for object detection with gpu hardware. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4287–4293. IEEE, 2009.

[38] Rajat Raina, Anand Madhavan, and Andrew Y Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880, 2009.

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[41] Pierre Sermanet and Yann LeCun. Traffic sign recognition with multi-scale convolutional networks. In *The 2011 International Joint Conference on Neural Networks*, pages 2809–2813. IEEE, 2011.

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[43] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[44] Quoc V Le. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8595–8598. IEEE, 2013.

[45] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[46] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[47] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. *arXiv preprint arXiv:1406.4773*, 2014.

[48] Gary B Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, 14(003), 2014.

[49] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[50] Charles C Tappert. Who is the father of deep learning? In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 343–348. IEEE, 2019.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[52] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[54] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[55] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[56] Fengwei Zhou, Bin Wu, and Zhenguo Li. Deep meta-learning: Learning to learn in the concept space. *arXiv preprint arXiv:1802.03596*, 2018.

[57] Shruti Jadon. An overview of deep learning architectures in few-shot learning domain. *arXiv preprint arXiv:2008.06365*, 2020.

[58] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[59] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[60] Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, pages 73–94, 2021.

[61] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[62] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

[63] Yann LeCun. Nips 2016 schedule. https://nips.cc/Conferences/2016/Schedule?showEvent=6197, 2016. (Accessed on 04/15/2021).

[64] Henry Markram. The human brain project. *Scientific American*, 306(6):50–55, 2012.

[65] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[66] Tianhong Dai, Kai Arulkumaran, Tamara Gerbert, Samyakh Tukra, Feryal Behbahani, and Anil Anthony Bharath. Analysing deep reinforcement learning agents trained with domain randomisation. *arXiv preprint arXiv:1912.08324*, 2019.

[67] Jozef Gruska. Quantum computing. https://www.fi.muni.cz/usr/gruska/quantum/, 1999. (Accessed on 04/15/2021).

[68] A Ezhov and D Ventura. Quantum neural networks. future directions for intelligent systems and information sciences. *Studies in Fuzziness and Soft Computing*, 45, 2000.

[69] Seunghyeok Oh, Jaeho Choi, and Joongheon Kim. A tutorial on quantum convolutional neural networks (qcnn). In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 236–239. IEEE, 2020.

[70] Davide Castelvecchi. Ibm's quantum cloud computer goes commercial. *Nature News*, 543(7644):159, 2017.