Citation

For pioneering research and application of fundamental techniques of deep learning in artificial intelligence

Dr. Yoshua Bengio

Positions and Organizations :
Founder and Scientific Director, Quebec Artificial Intelligence Institute (Mila)
Scientific Director, The Institute for Data Valorisation (IVADO)
Full Professor, Dept. of Computer Science and Research (DIRO), University of Montreal
CIFAR Fellow and Program Director

Date of Birth : March 5, 1964

Degree :
Ph.D, Computer Science, McGill University (1991)

Brief Biography :

1988	MSc, Computer Science, McGill U
1991	PhD, Computer Science, McGill U
1992	Postdoc, Massachusetts Institute of Technology (MIT)
1993	Postdoc Fellow, AT&T Bell Labs, New Jersey
1993	Assistant Professor, DIRO, U. Montreal
1997	Associate Professor, DIRO, U. Montreal
2002 - Present	Full Professor, DIRO, U. Montreal
2016 - Present	Scientific Director, IVADO
2017 - Present	Founder and Scientific Director, Mila

Main Awards and Honors

2009	ACFAS Urgel Archambault Prize
2017	Marie-Victorin Prize
2017	Fellow of the Royal Society of Canada
2018	Lifetime Achievement Award from the Canadian AI Association
2018	A. M. Turing Award
2019	IEEE CIS Neural Networks Pioneer Award
2019	Killam Prize
2020	Fellow of the Royal Society of London
2022	Knight of the Legion of Honor (France)
2022	Princess of Asturias Award for Technical and Scientific Research (Spain)
2023	Member of the UN’s Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology

Main Achievements:

In recent years, practical applications of AI technologies such as generative AI, automatic translation, image recognition, and autonomous driving have garnered public attention. Dr. Yoshua Bengio is recognized as one of the leading researchers in the fundamental technologies of AI, including deep learning and artificial neural nets.

Deep Learning is a subset of AI technologies that learns and extracts features automatically based on a large amount of data. This is made possible by a multi-layered structure where the output of one layer becomes an input for downstream layers. Neural nets are inspired by computations in the brain, with learnable connection strengths between artificial neurons. Adapting these parameters slightly after each example is shown to the neural network makes it possible through numerous iterations to learn generic representations and computations from large quantities of data to capture underlying structure that can generalize well to new data.

Dr. Yoshua Bengio received his Ph.D. from McGill University. He was a post-doctoral fellow and researcher at MIT and AT&T Bell Labs, and later served as an assistant professor and associate professor at the University of Montreal. In 2002, he was appointed a full professor at the Department of Computer Science and Operations Research. He founded Quebec Artificial Intelligence Institute (Mila) and serves as the Scientific Director.

In 2018, Dr. Yoshua Bengio received the A.M. Turing Award along with Dr. Geoffrey Hinton and Dr. Yann LeCun. They are now known as the Godfathers of AI and Deep Learning. Dr. Bengio has also been awarded many honors and prizes from various countries and academic societies around the world.

Dr. Bengio has notably authored and published numerous research papers and they are frequently cited, now making him the most cited computer scientist in the world. In January 2024, over 750,000 citations to scientific publications were found by Google Scholar, with an H-index of 228, with over 121,000 citations in 2022 alone. His main contributions co-created the field of deep learning, in areas such as recurrent nets, methods enabling deep learning to succeed, theoretical understanding of deep learning, the development of novel architectures based on attention and making it possible for neural nets to process sets rather than just vectors and sequences, and the development of deep generative models such as the generative adversarial networks.

More recently, Dr. Yoshua Bengio turned his attention to the agent perspective for representation learning (and thus to deep reinforcement learning) and deep learning architectures for reasoning, causality, epistemic uncertainty quantification, systematic generalization and AI safety, and has taken part in national and global discussions (and documents) about the socially responsible development of AI, as well as contributing to the research on AI for social good applications, e.g. in healthcare, the environment and education.

For more detailed achievements by Dr. Yoshua Bengio, please refer to next page.

Dr. Yoshua Bengio has made significant contributions to pioneering research in the fundamental techniques of deep learning in the field of artificial intelligence and its applications. In recognition of his considerable accomplishments, he is hereby awarded the Okawa Prize.

Dr. Yoshua Bengio

• 1989–1998 Convolutional and recurrent networks combined with probabilistic alignment (HMMs) to model sequences, as the main contribution of his PhD thesis (1991), NIPS'1988, NIPS'1989, Eurospeech'1991, PAMI'1991, IEEE Trans. Neural Nets 1992. These architectures were first applied to speech recognition in PhD (and rediscovered after 2010) and then with Yann LeCun et al. to handwriting recognition and document analysis (most cited paper is 'Gradient-based learning applied to document recognition', 1998, with over 19,000 citations).

• 1991–1995 Learning to learn papers with Samy Bengio, starting with IJCNN 1991, "Learning a synaptic learning rule." The idea of learning to learn (in particular by back-propagating through the whole process) has now become very popular (now called meta-learning) but they lacked the necessary computing power in the early '90s.

• 1993–1995 Uncovering the fundamental difficulty of learning in recurrent nets and other machine learning models of temporal dependencies, associated with vanishing and exploding gradients: ICNN'1993, NIPS'1993, NIPS'1994, IEEE Transactions on Neural Nets 1994, NIPS'1995. These papers (in particular the negative result) have had a major impact (turning the field of recurrent nets upside down) and motivated later papers on architectures to help learn long-term dependencies and deal with vanishing or exploding gradients. An important but subtle contribution of the IEEE Transactions 1994 paper is to show that the condition required to store bits of information reliably over time also gives rise to vanishing gradients, using dynamical systems theory. The NIPS'1995 paper introduced the use of a hierarchy of time scales to combat the vanishing gradients issue.

• 1999–2014 Understanding how distributed representations can bypass the curse of dimensionality by providing generalization to an exponentially large set of regions from those comparatively few occupied by training examples. This series of papers also highlights how methods based on local generalization, like nearest-neighbor and Gaussian kernel SVMs lack this kind of generalization ability. The NIPS'1999 introduced for the first time autoregressive neural networks for density estimation (the ancestor of the NADE and PixelRNN/PixelCNN models). The NIPS'2004, NIPS'2005 and NIPS'2011 papers on this subject show how neural nets can learn a local metric which can bring the power of generalization of distributed representations to kernel methods and manifold learning methods. Another NIPS'2005 paper shows the fundamental limitations of kernel methods due to a generalization of the curse of dimensionality (the curse of highly variable functions, which have many ups and downs). Finally, the ICLR'2014 paper shows in the case of piecewise-linear networks (like those with ReLUs) that the regions (linear pieces) distinguished by a one-hidden layer network is exponential in the number of neurons (whereas the number of parameters is quadratic in the number of neurons, and a local kernel method would require an exponential number of examples to capture the same kind of function).

• 2000–2008 Word embeddings from neural networks and neural language models. The NIPS'2000 paper introduces for the first time the learning of word embeddings as part of a neural network which models language data. The JMLR'2003 journal version expands this (these two papers together get around 3000 citations) and also introduces the idea of asynchronous SGD for distributed training of neural nets. Word embeddings have become one of the most common fixtures of deep learning when it comes to language data and this has basically created a new sub-field in the area of computational linguistics. He also introduced the use of importance sampling (AISTATS'2003, IEEE Trans. on Neural Nets, 2008) as well as of a probabilistic hierarchy (AISTATS 2005) to speedup computations and face larger vocabularies.

• 2006–2014 Showing the theoretical advantage of depth for generalization. The NIPS'2006 oral shows experimentally the advantage of depth and is one of the most cited papers in the field (over 2600 citations). The NIPS'2011 paper shows how deeper sum-product networks can represent functions which would otherwise require an exponentially larger model if the network is shallow. Finally, the NIPS'2014 paper on the number of linear regions of deep neural networks generalizes the ICLR'2014 paper mentioned above, showing that the number of linear pieces produced by a piecewise linear network grows exponentially in both width of layers and number of layers, i.e. depth, making the functions represented by such networks generally impossible to capture efficiently with kernel methods (short of using a trained neural net as the kernel).

• 2006–2014 Unsupervised deep learning based on auto-encoders (with the special case of GANs as decoder-only models, see below). The NIPS'2006 paper introduced greedy layer-wise pre-training, both the in the supervised case and the unsupervised case with auto-encoders. The ICML'2008 paper introduced denoising auto-encoders and the NIPS'2013, ICML'2014 and JMLR'2014 papers cast their theory and generalize them as proper probabilistic models, at the same time introducing alternatives to maximum likelihood as training principles.

• 2014 Dispelling the local-minima myth regarding the optimization of neural networks, with the NIPS'2014 paper on saddle points, showing that it is the large number of parameters which makes it very unlikely that bad local minima exist.

• 2014 Introducing Generative Adversarial Networks (GANs) at NIPS'2014, which innovates in many ways to train deep generative models, outside of the maximum likelihood framework and even outside of the classical framework of having a single objective function (instead entering into the territory of multiple models trained in a game-theoretical way, each with their objective). One of the hottest research areas in deep learning, as of this writing, with almost 2000 citations mostly from papers which introduce variants of GANs, which have been producing impressively realistic synthetic images, one would not imagine computers being able to generate just a few years ago.

• 2014–2016 Introducing content-based soft attention and the breakthrough it brought to neural machine translation, mostly with Kyunghyun Cho and Dima Bahdanau. They first introduced the encoder-decoder (now called sequence-to-sequence) architecture (EMNLP'2014) and then achieved a big jump in BLEU scores with content-based soft attention (ICLR'2015). These ingredients are now the basis of most commercial machine translation systems. Another whole subfield has been created using these techniques.

Yoshua Bengio's Listing of the top 20 most significant publications

[1] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[2] Anirudh Goyal and Yoshua Bengio. "Inductive biases for deep learning of higher-level cognition." *Proceedings of the Royal Society A* 478.2266 (2022), pp. 20210068.

[3] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. "Flow network based generative models for non-iterative diverse candidate generation." *Advances in Neural Information Processing Systems* 34 (2021), pp. 27381–27394.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *ICLR'2015, arXiv*:1409.0473. 2015.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[6] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In: *NIPS'2014*.

[7] Guido F. Montufar, Razvan Pascanu, KyungHyun Cho, and Yoshua Bengio. "On the Number of Linear Regions of Deep Neural Networks". In: *NIPS'2014*. 2014.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." In: *NIPS'2014*. 2014.

[9] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. "On the number of inference regions of deep feed forward networks with piece-wise linear activations." In: *ICLR'2014*. 2014.

[10] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. "Generalized Denoising Auto-Encoders as Generative Models." In: *NIPS'2013*. 2013.

[11] Xavier Glorot, Antoine Bordes, and Yoshua. Bengio. "Deep Sparse Rectifier Neural Networks." In: *AISTATS'2011*.

[12] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In: *AISTATS'2010*. 2010.

[13] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. "Curriculum Learning". In: *ICML'2009*.

[14] Yoshua Bengio. "Learning deep architectures for AI." In: *Foundations and Trends in Machine Learning* 2.1. (2009), pp. 1–127.

[15] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and Composing Robust Features with Denoising Autoencoders." In: *ICML'2008*. 2008, pp. 1096–1103.

[16] Yoshua Bengio, Pascal Lamblin, D. Popovici, and H. Larochelle. "Greedy Layer-Wise Training of Deep Networks." In: *NIPS'2006*. 2007.

[17] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A Neural Probabilistic Language Model." In: *Journal of Machine Learning Research* 3 (2003), pp. 1137–1155.

[18] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-Based Learning Applied to Document Recognition." In: *Proceedings of the IEEE* 86.11 (Nov. 1998), pp. 2278–2324.

[19] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning Long-Term Dependencies with Gradient Descent is Difficult." In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166.

[20] Yoshua Bengio, Samy Bengio, Jocelyn Cloutier, and Jan Gescei. "Learning a Synaptic Learning Rule." In:*IJCNN'1991*. Seattle, WA, 1991, II—A969