2023年度 大川賞受賞者

受賞理由

人工知能における深層学習の基礎技術における先駆的研究とその応用

ヨシュア ベンジオ 博士

現　職　Mila−ケベック人工知能研究所　創設者・科学ディレクター
データ・バロリゼーション研究所（IVADO）　科学ディレクター
モントリオール大学　計算機科学・オペレーションズリサーチ学科（DIRO）　教授
カナダ先端研究機構　フェロー・プログラムディレクター

生年月日　1964年3月5日

学　位　Ph.D.（マギル大学、1991 年）

略　歴
1988年　マギル大学 計算機科学修士号
1991年　マギル大学 計算機科学博士号
1992年　マサチューセッツ工科大学（MIT） 博士研究員
1993年　AT&Tベル研究所 博士研究員
1993年　モントリオール大学 計算機科学・オペレーションズリサーチ学科 助教授
1997年　同 准教授
2002年−現在　同 教授
2016年−現在　データ・バロリゼーション研究所 科学ディレクター
2017年−現在　Mila−ケベック人工知能研究所　創設者・科学ディレクター

主な受賞歴
2009年　フランス系カナダ科学進歩協会（ACFAS） Urgel Archambault賞
2017年　マリー・ビクトリン賞
2017年　カナダ王立協会フェロー
2018年　カナダ人工知能協会 生涯功労賞
2018年　A・M・チューリング賞
2019年　米国電気電子学会（IEEE）CIS Neural Networks Pioneer Award
2019年　キラム賞
2020年　ロンドン王立協会フェロー
2022年　レジオンドヌール勲章シュヴァリエ（フランス）
2022年　アストゥリアス女公賞学術・技術研究部門（スペイン）
2023年　国連科学諮問委員会委員

主な業績

近年、生成AIや自動翻訳、画像認識、自動運転などの人工知能（AI）技術の実際的応用が一般の関心を集めている。ヨシュアベンジオ博士は、深層学習や人工ニューラルネットワークを含むAIの基本技術分野の主要な研究者の一人と認められている。

深層学習（ディープラーニング）は大量のデータをもとに自動で特徴量を抽出し、学習していくAI関連技術である。この技術は、ある層の出力が下流層の入力となる多層構造により実現される。ニューラルネットワークは人間の脳内の演算をヒントにしており、人工ニューロン間の学習可能なつながりの強さを利用する。各事例がニューラルネットワークに示された直後にこれらのパラメータを適合させることにより、多数の繰り返しを通じて、大量のデータから一般的な表現や計算を学習し、新しいデータに適切に一般化できる基礎構造を把握することが可能になる。

ベンジオ博士はカナダのマギル大学で博士号を取得後、MITやベル研究所で博士研究員をつとめたのち、モントリオール大学において助教授、准教授を経て2002年に計算機科学・オペレーションズリサーチ学科の教授に任じられた。さらに、Mila−ケベック人工知能研究所を創立し、科学ディレクターを務めている。

博士は2018年にはヒントン博士、ルカン博士とともにACMチューリング賞を受賞されており、この3名は、現在のAIのゴッドファーザー、あるいは、深層学習のゴッドファーザーと称されている。その他にも、数多くの諸外国や学会からの栄誉や賞を受けている。

博士は非常に多数の研究論文を執筆・発表しており、いずれも頻繁に引用され、現在、計算機科学分野では世界で最も被引用数の多い研究者である。2024年1月の時点で、学術文献の検索サービスGoogle Scholarで科学論文への引用が75万件以上確認され、科学研究への貢献度を示すh指数は228、2022年に限った被引用数でも12万1,000件を超える。ベンジオ博士の主要な研究は、再帰型ネットワーク、深層学習の成功を可能にする手法、深層学習の理論的理解、「Attention（注意）」という概念に基づく新たなアーキテクチャの開発による、ベクトルとシーケンスだけでなくセットを処理できるニューラルネットワークの実現、敵対的生成ネットワークなどの深層生成モデルの開発などの領域において、深層学習分野の創始に寄与した。

最近では、ベンジオ博士は表現学習におけるエージェントの観点（および深層強化学習）や、推論、因果関係、認識論的不確実性の定量化、系統的生成、AIの安全性のための深層学習アーキテクチャに関心を向け、社会的に責任あるAI開発に関する国内外の議論（および文書策定）や、医療・環境・教育分野など、より良い社会につながる応用のためのAIに関する研究に参画している。

ベンジオ博士の詳しい業績については、6ページを参照のこと。

このように、ヨシュア ベンジオ博士は、人工知能分野における深層学習の基礎技術に関する先駆的研究とその応用に多大な貢献をされてきた。ここに大川賞を贈呈し、その功績を称えるものである。

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