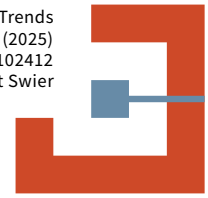
Trends

JALTCALL Trends
ISSN 2982-3749, 1, 102412 (2025)
https://doi.org/10.29140/jct.v1n1.102412



Castledown

A technical background on artificial intelligence and intelligent language models

Robert Swier
Kindai University
robert.swier@lac.kindai.ac.jp

Stunning advancements in artificial intelligence (AI) over the last several years have undoubtedly opened new possibilities and challenges for the field of second language learning. Of course, AI is not new. For decades, it has attracted the interests and imaginations of philosophers, computer scientists, and the public. Though the performance of AI systems had typically been underwhelming until recently, the results of the current systems have made AI relevant for nearly everyone. Much of the focus in language education has been on exploring how these new systems can be used by teachers, researchers, and especially learners. This paper provides a brief technical primer on the discoveries that have helped make the AI revolution possible, including algorithmic discoveries in deep learning such as backpropagation and the attention mechanism of transformers, and hardware improvements that have made it feasible to train such massive models. Understanding the background and underlying principles behind AI may benefit language teachers and researchers who are interested in demystifying the technology and understanding its long-term potential.

Keywords: artificial intelligence, backpropagation, transformers

Introduction

The story of the development of artificial intelligence (AI) may one day become one of the most enduring tales of imagination and perseverance in human history, similar perhaps to achieving flight or walking on the moon. The possibility of intelligent machines was recognized very early in the history of computing. Alan Turing, the British mathematician who is widely viewed as the father of computer science, began to lay the early groundwork for artificial intelligence in his 1936 paper, "On computable numbers, with an application to the Entscheidungsproblem", exploring the power of mathematical logic nearly a decade before the first programmable computers had even been built. By 1950, Turing was explicitly talking about intelligent machines in his seminal paper "Computing machinery and intelligence". By 1955, the term "artificial intelligence" itself was coined by John McCarthy when he

was proposing the now famous Dartmouth summer workshop "to find out how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy, et al., 1955). Optimism was high, and the workshop was intended to make significant progress on several problems related to intelligent machines.

Despite some initial progress, it did not take long for researchers to discover how difficult it would be to achieve anything close to human-level performance on tasks involving language or reasoning (Crevier, 1993). Problems were everywhere. Logical, mathematical, hand-written symbolic representations of knowledge worked well for simple tasks but proved unable to capture the nuanced, real-world knowledge required for understanding language and common-sense reasoning. Rule-based systems were inflexible and had trouble handling all but the most common situations. Tasks involving planning or decision-making faced the problem of combinatorial explosion, where the number of possibilities that require evaluation grows exponentially and quickly becomes intractable after only a few steps.

Over the decades, lack of progress led to several periods of reduced funding and interest in AI, now called "AI winters" (Hendler, 2008). It was not until the 2010s that progress had finally been made in enough areas that AI started to become the commercially valuable and practically useful product that it is today. Here, I briefly review three major advancements that made this possible.

Advancements in AI

Artificial neural networks and backpropagation

The idea of building intelligent computer systems with idealized, artificial "neurons" was considered almost immediately. After all, since the brain was already known to be an example of an intelligent object, mimicking the brain seemed to be a sensible place to start. McCulloch and Pitts (1943) were the first to describe how such a system could work, and Rosenblatt (1957) was the first to produce a working implementation that was able to learn. However, single layers of neurons were later shown to be incapable of learning to recognize nonlinearly separable patterns (where two classes of data cannot be separated by a simple straight line), making them seemingly unsuitable for many tasks (Minsky & Papert, 1969). Though this could be mitigated by using multiple layers of neurons, there was no known method that could effectively train such networks. For years, mainstream AI researchers had very little interest in neural networks, and almost all research investigated other approaches – such as expert systems – with little emphasis on learning from data.

It was not until 1986 when Rumelhart, Hinton, and Williams developed and applied the backpropagation algorithm that multilayer neural networks could first be trained effectively. Multilayer networks use groups of interconnected neurons where each connection is assigned a weight (multiplier) that indicates the strength of that connection. With backpropagation, when the network produces an error during training, the error is propagated backwards through the network, with the connection weights being slightly adjusted up or down to make the correct output more

likely. Critically, the connections that have contributed the most to the error receive the most adjustment.

The immediate effect of backpropagation was to renew interest in neural networks and to help launch the field of deep learning (where "deep" simply refers to the use of networks with multiple layers). Still, however, neural network models failed to become the dominant approach to machine learning, as a competing technique known as a support vector machine produced superior results throughout the 1990s and 2000s, given the hardware and dataset size limitations at the time (Cortes & Vapnik, 1995; Cristianini & Shawe-Taylor, 2000).

Better hardware, more data

The twin engines of AI's progress have been the exponential growth in the amount of available data and the power of computing hardware. The internet, social media, e-commerce, digital imagery, connected devices, and other systems have generated vast amounts of digital data that are now available for training machine learning systems. Enormous datasets, such as ImageNet (Deng et al., 2009) for image classification and Common Crawl (n.d.) for language models, are now easily available and have dramatically changed the resources available for training AI models. However, storing and training on vast amounts of data is only possible with enormous computing resources.

Although progress in AI has not always been smooth, improvements in computing capacity have been exponential, particularly with respect to data storage and computation. While the central processing units (CPUs) on computers had, for several decades, been doubling their transistor capacity every two years, even this massive increase in capacity would still present a bottleneck for training AI models. CPUs are designed to process information sequentially, one instruction at a time. Even if this can be done at a very high speed, the unique computational requirements of neural networks means that it is far more effective to process information in parallel – that is, performing multiple independent calculations at once. As it happens, graphics processing units (GPUs) – designed initially for fast processing of graphics information for applications such as computer gaming – make ideal processors for neural network training. This has resulted in leading GPU companies, such as NVIDIA, refocusing their businesses on producing specialized hardware for training AI models.

The final pivotal development in hardware has been the rise of cloud computing, where enormous data centers with high-performance computing hardware are available on demand, allowing firms which are training large AI models to shorten the necessary training time by distributing the necessary computation across many individual nodes and even across multiple data centers.

Transformers and the attention mechanism

The third key advancement that has helped create the current boom in AI was the development of a type of neural network architecture known as a "transformer" by a group of researchers working at Google (Vaswani et al., 2017). As research progressed on the design of neural networks, several different architectures emerged that were

suitable for different types of learning tasks. For instance, recurrent neural networks were developed for use on sequential data – such as natural language – where the order of the elements in the data is of key importance, and convolutional neural networks were developed for spatial or visual data, such as object or facial recognition in images. Both models, however, struggle to capture long-range dependencies – cases where different input elements have important relationships but are not close together in the input. The transformer architecture addresses this through use of an "attention mechanism" in which a relationship score is calculated for every pair of tokens in the input. With linguistic data, each input token corresponds to a word or part of a word. This technique thus allows the model to focus on the relationship between words regardless of how far apart they are in the input. Since each of these relationships is independent, these calculations may be easily done in parallel, making transformers ideal for taking advantage of the processing capabilities of GPUs.

The transformer architecture has led to a significant improvement in the performance of neural networks for learning from linguistic data (Brown et al., 2020). Scaled-up versions of transformers now form the basis of large language models, often referred to as LLMs. Well-known LLMs include GPT by OpenAI, BERT and Gemini by Google, LLaMA from Meta, and Claude from Anthropic. These models contain billions of neurons and are capable of understanding, generating, analyzing, and summarizing human language. Long-range context sensitivity has led to extraordinary improvements in answering questions, writing essays, summarizing texts, translating languages, generating computer code, and engaging in conversation.

Conclusion

AI has progressed from a distant aspiration to a powerful and widely distributed technology that will have profound effects on the way we teach, learn, and communicate. The three key advancements discussed here – the development of backpropagation for multilayer neural networks, improvements in hardware and data, and the invention of transformers – have led to the development of large language models that can indeed now meet McCarthy's 1955 goal for machines to "solve the kinds of problems now reserved for humans".

The journey was certainly longer than early researchers expected, and many challenges remain. Two in particular relate to data. AI systems are trained on data generated by humans, and humans are known to exhibit behavior that is biased and unfair. While humans have a right to their own prejudices, these are generally not desirable qualities in a machine. Ensuring that AI models do not inadvertently learn biases and are otherwise aligned with desirable human values remains an unsolved issue (Bender et al., 2021). Data privacy is another concern, particularly for teachers, who may expose the personal information of their students to AI models maintained by third parties. Keeping such data private and secure as AI models are called on to complete an increasing number of routine tasks must be a priority for users and developers alike.

For both teachers and students, recognizing that the current systems are the result of a long process of development may serve as a reminder that they are neither magical nor perfect, and that the inputs and outputs to AI systems must be carefully

considered and evaluated before being fully integrated into educational practice. Understanding the historical trajectory of this technology empowers educators to engage critically and creatively with these tools as they shape the future of learning.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. doi: 10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 159:1–159:25). Curran. doi: 10.48550/arXiv.2005.14165

Common Crawl. (n.d.). *Common crawl corpus* [Data set]. https://commoncrawl.org

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20,* 273–297. doi: 10.1007/BF00994018

Crevier, D. (1993). *AI: The tumultuous history for the search for artificial intelligence.* Basic Books.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press. doi: 10.1017/CBO9780511801389

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 248–255. doi: 10.1109/CVPR.2009.5206848

Hendler, J. (2008). Avoiding another AI winter. *IEEE Intelligent Systems, 23*(2), 2–4. doi: 10.1109/MIS.2008.19

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence.* Stanford University. https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McCulloch W. S., & Pitts W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics. 5*(4): 115–133. doi: 10.1007/BF02478259

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* MIT Press.

Rosenblatt, F. (1957). *The perceptron – a perceiving and recognizing automaton.* Report 85-460-1. Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. doi: 10.1038/323533a0

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2*(42), 230–265. doi: 10.1112/plms/s2-42.1.230

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. doi: 10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.* doi: 10.48550/arXiv.1706.03762

JALTCALL Trends
Vol. 1
2024 JALTCALL Conference Papers Special Issue

Author biodata

Robert Swier is a member of the Faculty of Literature, Arts, and Cultural Studies at Kindai University. He has degrees in computer science (focusing on AI and computational linguistics) from the University of Rochester and the University of Toronto. He has a PhD in second language learning from Kyoto University, focusing on virtual worlds as platforms for learner interaction.