Source: https://www.forbes.com/sites/robtoews/2023/09/03/transformers-revolutionized-ai-what-will-replace-them/

Transformers Revolutionized AI. What Will Replace Them?

If modern artificial intelligence has a founding document, a sacred text, it is Google's 2017 research paper "Attention Is All You Need."

This paper introduced a new deep learning architecture known as the transformer, which has gone on to revolutionize the field of AI over the past half-decade.

The generative AI mania currently taking the world by storm can be traced directly to the invention of the transformer. Every major AI model and product in the headlines today—ChatGPT, GPT-4, Midjourney, Stable Diffusion, GitHub Copilot, and so on—is built using transformers.

Transformers are remarkably general-purpose: while they were initially developed for language translation specifically, they are now advancing the state of the art in domains ranging from computer vision to robotics to computational biology.

In short, transformers represent the undisputed gold standard for AI technology today.

But no technology remains dominant forever.

It may seem surprising or strange, with transformers at the height of their influence, to contemplate what will come next. But in the fast-moving world of AI, it is both fascinating and advantageous to seek to "see around corners" and glimpse what the future holds before it becomes obvious.

Transformers 101

In order to explore this question, we must first understand transformers more deeply.

The now-iconic transformer paper was co-authored by eight researchers working together at Google over the course of 2017: Aidan Gomez, Llion Jones, Lukasz Kaiser, Niki Parmar, Illia Polosukhin, Noam Shazeer, Jakob Uszkoreit and Ashish Vaswani.

An often-overlooked fact about the paper is that all eight authors are listed as equal contributors; the order in which the authors' names appear on the paper was randomly determined and has no significance. With that said, it is generally recognized that Uszkoreit provided the initial intellectual impetus for the transformer concept, while Vaswani and Shazeer were the two authors most deeply involved in every aspect of the work from beginning to end.

All eight authors have become luminaries in the world of AI thanks to their work on the paper. None of them still work at Google. Collectively, the group has gone on to found many of today's most important AI startups, including Cohere, Character.ai, Adept, Inceptive, Essential AI and Sakana AI.

Why, exactly, was the transformer such a massive breakthrough?

Before the "Attention Is All You Need" paper was published, the state of the art in language AI was a deep learning architecture known as recurrent neural networks (RNNs).

By definition, RNNs process data sequentially—that is, one word at a time, in the order in which the words appear.

But important relationships often exist between words even if they do not appear next to each other in a sequence. In order to better enable RNNs to account for these long-distance dependencies between words, a mechanism known as attention had recently become popular. (The invention of the attention mechanism is generally attributed to a 2014 paper from deep learning pioneer Yoshua Bengio.)

Attention enables a model to consider the relationships between words regardless of how far apart they are and to determine which words and phrases in a passage are most important to "pay attention to."

Before the transformer paper, researchers had only used attention as an add-on to the RNN architecture. The Google team's big leap was to do away with RNNs altogether and rely entirely on attention for language modeling. Hence the paper's title: Attention Is All You Need.

(A charming, little-known fact about the paper: according to co-author Llion Jones, its title is a nod to the Beatles song "All You Need Is Love.")

Transformers' fundamental innovation, made possible by the attention mechanism, is to make language processing parallelized, meaning that all the words in a given body of text are analyzed at the same time rather than in sequence.

As an interesting analogy, co-author Illia Polosukhin has compared the transformer architecture to the fictional alien language in the 2016 science fiction movie Arrival. Rather than generating strings of characters sequentially to form words and sentences (the way that humans do), the aliens in the film produce one complex symbol at a time, all at once, which conveys detailed meaning that the humans must interpret as a whole.

Transformers' parallelization gives them a more global and thus more accurate understanding of the texts that they read and write. It also makes them more computationally efficient and more scalable than RNNs. Transformers can be trained on much larger datasets and built with many more parameters than previous architectures, making them more powerful and generalizable. Indeed, a hallmark of today's leading transformer-based models is their scale.

In one of those mutually beneficial, mutually reinforcing historical co-occurrences, the transformer's parallel architecture dovetailed with the rise of GPU hardware. GPUs are a type of computer chip that are themselves massively parallelized and thus ideally suited to support transformer-based computing workloads. (Nvidia, the world's leading producer of GPUs, has been perhaps the single biggest beneficiary of today's AI boom, recently surpassing a $1 trillion market capitalization amid staggering demand for its chips.)

The rest, as they say, is history. Thanks to these tremendous advantages, transformers have taken the world by storm in the six years since their invention, ushering in the era of generative AI.

Every popular "chatbot" today—OpenAI's ChatGPT, Google's Bard, Microsoft's Bing Chat, Anthropic's Claude, Inflection's Pi—is transformer-based. So is every AI tool that generates images or videos, from Midjourney to Stable Diffusion to Runway. (Text-to-image and text-to-video technology is powered by diffusion models; diffusion models make use of transformers.)

Transformers' influence reaches well beyond text and images. The most advanced robotics research today relies on transformers. Indeed, Google's most recent robotics work is actually named RT-2, where the T stands for "transformer." Similarly, one of the most promising new avenues of research in the field of autonomous vehicles is the use of vision transformers. Transformer-based models have unlocked breathtaking new possibilities in biology, including the ability to design customized proteins and nucleic acids that have never before existed in nature.

Transformer co-inventor Ashish Vaswani summed it up well: "The transformer is a way to capture interaction very quickly all at once between different parts of any input. It's a general method that captures interactions between pieces in a sentence, or the notes in music, or pixels in an image, or parts of a protein. It can be purposed for any task."

All Good Things Must End?

Yet despite its incredible strengths, the transformer is not without shortcomings. These shortcomings open the door for the possible emergence of new and improved architectures.

Chief among the transformer's shortcomings is its staggering computational cost.

As anyone familiar with the world of AI knows, one of the defining characteristics of today's AI models is their insatiable computing needs. Training a cutting-edge large language model today entails running thousands of GPUs around the clock for months at a time. The reason that OpenAI raised an eye-popping $10 billion earlier this year, for instance, was in order to foot the bill for the vast computing resources needed to build advanced AI models. As another example, eighteen-month-old startup Inflection recently raised over $1 billion in venture funding in order to build a massive GPU cluster to train its language models.

Transformer-based models are so compute-hungry, in fact, that the current AI boom has triggered a global supply shortage, with hardware manufacturers unable to produce AI chips fast enough to keep up with demand.

Why are transformers so computationally demanding?

One basic answer is that transformers' great strength also becomes a weakness: because they scale so much more effectively than previous architectures, transformers make it possible—and irresistible—to build models that are orders of magnitude larger than have previously existed. Such massive models require correspondingly massive compute.

But there is a more specific reason for transformers' computational cost: the transformer architecture scales quadratically with sequence length. Put simply, this means that as the length of a sequence processed by a transformer (say, the number of words in a passage or the size of an image) increases by a given amount, the compute required increases by that amount squared, quickly growing enormous.

There is an intuitive reason for this quadratic scaling, and it is inherent to the transformer's design.

Recall that attention makes it possible to understand relationships between words regardless of how far apart they are in a sequence. How does it do this? By comparing every single word in a sequence to every other word in that sequence. The consequence of this pairwise comparison is that as sequence length increases, the number of required computational steps grows quadratically rather than linearly. To give a concrete example, doubling sequence length from 32 tokens to 64 tokens does not merely double the computational cost for a transformer but rather quadruples it.

This quadratic scaling leads to a related drawback: transformers have a hard time handling very long sequences.

As sequences grow in length, feeding them into transformers eventually becomes intractable because memory and compute needs explode quadratically. Consider, for example, processing entire textbooks (with millions of tokens) or entire genomes (with billions of tokens).

Increasing the maximum sequence length that a model can be fed at one time, known as the model's "context window," is an active area of research for large language models today. The context window for the base GPT-4 model is 8,000 tokens. A few months ago, OpenAI released a souped-up version of GPT-4 with a 32,000-token context window. OpenAI competitor Anthropic then upped the ante, recently announcing a new model with a 100,000-token context window.

This arms race will no doubt continue. Yet there are limits to how big OpenAI, Anthropic or any other company can make its models' context windows if they stick with the transformer architecture.

Various attempts have been made to build modified versions of transformers that still use attention but are better equipped to handle long sequences. Yet these modified transformer architectures—with names like Longformer, Reformer, Performer, Linformer and Big Bird—generally sacrifice on performance and so have failed to gain adoption.

Challengers to the Throne

This leads us to perhaps the most fertile area of research today in the effort to create a replacement for transformers. The guiding principle for this school of research is to replace attention with a new function that scales sub-quadratically. Sub-quadratic scaling would unlock AI models that are (1) less computationally intensive and (2) better able to process long sequences compared to transformers. The challenge, of course, is to do this while still matching transformers' overall capabilities.

A 2021 research effort named S4 out of Chris Ré's lab at Stanford laid the foundations for this avenue of research. A handful of promising subquadratic architectures based on S4 have followed.

One of the most intriguing new architectures in the S4 family is Hyena, published a few months ago by a powerhouse team that includes Ré and Yoshua Bengio.

In place of attention, Hyena uses two other operations: long convolutions and element-wise multiplication.

Convolutions are one of the oldest existing methods in machine learning, first conceived of by Yann LeCun back in the 1980s. Hyena's fresh take on this venerable architecture is to stretch and vary the size of the convolution filter based on the sequence length in order to boost computational efficiency.

Hyena's initial results are promising. The model achieves new state-of-the-art performance for a non-attention-based language model. It matches transformers' performance in certain settings while using significantly less compute. Importantly, Hyena's efficiency gains relative to transformers become more dramatic as sequence length increases, underscoring their advantages for very long inputs: at an 8,000-token sequence length, Hyena operators are twice as fast as attention, whereas at a 64,000-token length they are one hundred times faster.

As the Hyena authors put it: "Breaking the quadratic barrier is a key step towards new possibilities for deep learning, such as using entire textbooks as context, generating long-form music or processing gigapixel scale images."

With at least a hint of snark, the authors add: "Our promising results at the sub-billion parameter scale suggest that attention may not be all we need."

One compelling early application of the Hyena architecture is HyenaDNA, a new foundation model for genomics out of Stanford. Capitalizing on Hyena's superior ability to handle long sequences, HyenaDNA has a whopping 1-million-token context window. The human genome is one of the longest (not to mention one of the most important) datasets in existence: each human's DNA contains 3.2 billion nucleotides. This makes it an ideal use case for a model architecture like Hyena that excels at capturing long-range dependencies.

The HyenaDNA authors offer a tantalizing hint of what this technology might unlock in the future: "Imagine being able to prompt ChatGPT with an entire human genome - wouldn't it be neat to ask questions about likely diseases, predict drug reactions, or guide treatment options based on your specific genetic code?"

An important caveat here is that the initial Hyena work was carried out at relatively small scales. The largest Hyena model has 1.3 billion parameters, compared to GPT-3's 175 billion parameters and GPT-4's (rumored) 1.8 trillion parameters. A key test for the Hyena architecture will be whether it continues to demonstrate strong performance and efficiency gains as it is scaled up to the size of today's transformer models.

Other novel deep learning architectures in this family include Monarch Mixer (also from Chris Ré's lab at Stanford), BiGS (from Cornell and DeepMind) and MEGA (from Meta).

Like Hyena, all of these models feature subquadratic scaling, meaning that they are more computationally efficient and better equipped to handle long sequences than are transformers. And like Hyena, they are all promising but unproven: it remains to be seen whether any of them can maintain strong performance at the scales at which today's transformer models operate.

Stepping back, computational efficiency and long-range dependencies are not the only two weaknesses of transformers that new architectures aim to improve on.

An additional limitation of transformer models is their inability to learn continuously. Today's transformer models have static parameters. When a model is trained, its weights (the strength of the connections between its neurons) are set; these weights do not update based on new information that the model encounters as it is deployed in the world.

Another commonly referenced limitation is transformers' lack of explainability. Transformer-based models are "black boxes": their internal workings are too complex and opaque for humans to understand exactly why they behave the way they do. This can be a real problem for safety-critical or highly regulated applications, for instance in healthcare.

Liquid neural networks, another buzzy new AI architecture seeking to challenge the transformer, claims to tackle both of these shortcomings.

Created at MIT by a research team led by Ramin Hasani and Daniela Rus, liquid neural networks are inspired by biology: in particular, by how the C. elegans worm's brain works. The "liquid" in the name refers to the fact that the model's weights are probabilistic rather than constant, allowing them to vary fluidly depending on the inputs the model is exposed to.

Liquid neural networks are also much smaller than today's transformer models. In one recent proof of concept, the MIT team built an autonomous vehicle system that was able to successfully drive on public roads with a mere 19 neurons and 253 parameters.

"Everyone talks about scaling up their network," said Hasani. "We want to scale down, to have fewer but richer nodes."

In addition to computational efficiency, this smaller architecture means that liquid neural networks are more transparent and human-readable than transformers. After all, it is more practicable for a human observer to interpret what is happening in a network with 253 connections than in one with 175 billion connections.

Rus is one of the world's leading roboticists, and liquid neural networks appear to be particularly well-suited for robotics applications, including autonomous vehicles and drones. They only work with time-series data (i.e., data with a time dimension to it), meaning that they cannot be applied to images or other static data modalities.

One final effort to build "what comes after the transformer" is worth mentioning. Llion Jones—one of the eight "Attention Is All You Need" co-authors—recently left Google to launch a new startup named Sakana AI alongside former Stability AI head of research David Ha.

Sakana's mission is to improve upon transformers with a nature-inspired approach to intelligence grounded in evolutionary principles. Key to the team's vision is the notion of collective or swarm intelligence, with a system of many small models acting collaboratively rather than one monolithic model.

"Learning always wins," said Jones. "The history of AI reflects the reality that it always works better to have a model learn something for itself rather than have a human hand-engineer it. The deep learning revolution itself was an example of this, as we went from building feature detectors by hand to letting neural networks learn their own features.

This is going to be a core philosophy for us at Sakana AI, and we will draw on ideas from nature including evolution to explore this space."

Distant Horizons

The transformer is an exceptionally powerful AI architecture.

Transformers have become the foundation of modern artificial intelligence. Virtually every advanced AI system is based on transformers; every AI researcher is accustomed to working with them. Transformers have been optimized by thousands of researchers building on one another's work over the past several years.

This gives them a powerful incumbency advantage that will make them formidable to dislodge.

Yet, outside the limelight, away from the echo chambers of AI hype, promising work is underway to develop next-generation AI architectures that are superior to transformers in different ways.

This work is still early and unproven. It remains far from certain whether these new architectures will succeed in replacing the transformer. But if they do, the implications for the world of AI will be enormous.

Before the transformer era, different AI architectures were predominant for different use cases: recurrent neural networks were used for language, convolutional neural networks were used for computer vision, reinforcement learning was used for game-playing, and so on.

It has been remarkable to witness the progressive unification of AI methodology in recent years as transformers have proven themselves state-of-the-art in one domain after the other, from language to vision to robotics to biology.

Yet it is not preordained that this trend toward unification—toward "one AI architecture to rule them all"—will continue indefinitely.

It is conceivable that a different version of the future will play out: that as the frontiers of AI research advance in the years ahead, new architectures are developed that prove themselves better suited for particular domains. Perhaps, for instance, transformers continue to dominate the field of language processing for years to come, while a novel architecture soon displaces transformers as state-of-the-art in robotics.

Or perhaps a new AI approach is developed that outperforms and rapidly replaces transformers across the board.

One thing is certain: the field of artificial intelligence is today so fast-moving and dynamic that we should expect change to come uncomfortably quickly, we should take nothing for granted, and we should prepare to be surprised by what the future holds.