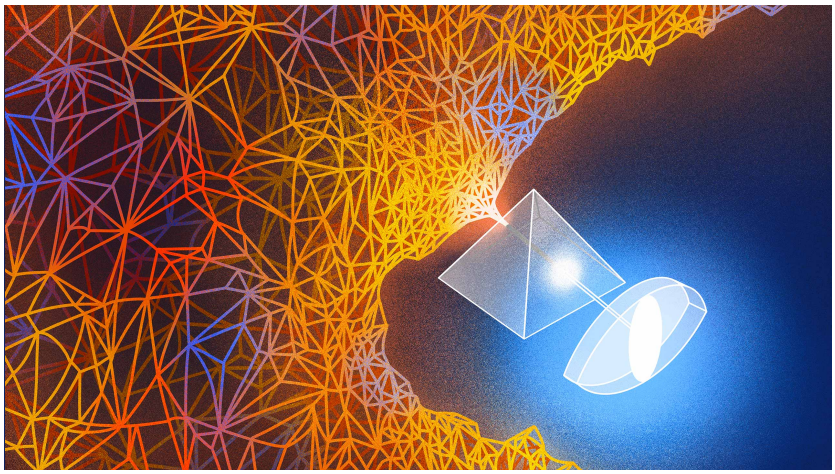
Quantamagazine

Will Transformers Take Over Artificial Intelligence?

By **Stephen Ornes**

March 10, 2022

A simple algorithm that revolutionized how neural networks approach language is now taking on vision as well. It may not stop there.



Avalon Nuovo for Quanta Magazine

Imagine going to your local hardware store and seeing a new kind of hammer on the shelf. You've heard about this hammer: It pounds faster and more accurately than others, and in the last few years it's rendered many other hammers obsolete, at least for most uses. And there's more! With a few tweaks — an attachment here, a twist there — the tool changes into a saw that can cut at least as fast and as accurately as any other option out there. In fact, some experts at the frontiers of tool development say this

hammer might just herald the convergence of all tools into a single device.

A similar story is playing out among the tools of artificial intelligence. That versatile new hammer is a kind of artificial neural network — a network of nodes that "learn" how to do some task by training on existing data — called a transformer. It was originally designed to handle language, but has recently begun impacting other AI domains.

The transformer first appeared in 2017 in a paper that cryptically declared that "Attention Is All You Need." In other approaches to AI, the system would first focus on local patches of input data and then build up to the whole. In a language model, for example, nearby words would first get grouped together. The transformer, by contrast, runs processes so that every element in the input data connects, or pays attention, to every other element. Researchers refer to this as "self-attention." This means that as soon as it starts training, the transformer can see traces of the entire data set.

Before transformers came along, progress on AI language tasks largely lagged behind developments in other areas. "In this deep learning revolution that happened in the past 10 years or so, natural language processing was sort of a latecomer," said the computer scientist Anna Rumshisky of the University of Massachusetts, Lowell. "So NLP was, in a sense, behind computer vision. Transformers changed that."

Transformers quickly became the front-runner for applications like word recognition that focus on analyzing and predicting text. It led to a wave of tools, like OpenAI's Generative Pre-trained Transformer 3 (GPT-3), which trains on hundreds of billions of words and generates consistent new text to an unsettling degree.

The success of transformers prompted the AI crowd to ask what else they

could do. The answer is unfolding now, as researchers report that transformers are proving surprisingly versatile. In some vision tasks, like image classification, neural nets that use transformers have become faster and more accurate than those that don't. Emerging work in other AI areas — like processing multiple kinds of input at once, or planning tasks — suggests transformers can handle even more.

"Transformers seem to really be quite transformational across many problems in machine learning, including computer vision," said Vladimir Haltakov, who works on computer vision related to self-driving cars at BMW in Munich.

Just 10 years ago, disparate subfields of AI had little to say to each other. But the arrival of transformers suggests the possibility of a convergence. "I think the transformer is so popular because it implies the potential to become universal," said the computer scientist Atlas Wang of the University of Texas, Austin. "We have good reason to want to try transformers for the entire spectrum" of AI tasks.

From Language to Vision

One of the most promising steps toward expanding the range of transformers began just months after the release of "Attention Is All You Need." Alexey Dosovitskiy, a computer scientist then at Google Brain Berlin, was working on computer vision, the AI subfield that focuses on teaching computers how to process and classify images. Like almost everyone else in the field, he worked with convolutional neural networks (CNNs), which for years had propelled all major leaps forward in deep learning and especially in computer vision.

The computer scientist Alexey Dosovitskiy helped create a neural network called the Vision Transformer, which applied the power of a transformer to visual recognition tasks.

Courtesy of Alexey Dosovitskiy

CNNs work by repeatedly applying filters to the pixels in an image to build up a recognition of features. It's because of convolutions that photo apps can organize your library by faces or tell an avocado apart from a cloud. CNNs were considered indispensable to vision tasks.

Dosovitskiy was working on one of the biggest challenges in the field, which

was to scale up CNNs to train on ever-larger data sets representing images of ever-higher resolution without piling on the processing time. But then he watched transformers displace the previous go-to tools for nearly every AI task related to language. "We were clearly inspired by what was going on," he said. "They were getting all these amazing results. We started wondering if we could do something similar in vision." The idea made a certain kind of sense — after all, if transformers could handle big data sets of words, why not pictures?

The eventual result was a network dubbed the Vision Transformer, or ViT, which the researchers presented at a conference in May 2021. The architecture of the model was nearly identical to that of the first transformer proposed in 2017, with only minor changes allowing it to analyze images instead of words. "Language tends to be discrete," said Rumshisky, "so a lot of adaptations have to discretize the image."

The ViT team knew they couldn't exactly mimic the language approach since self-attention on every pixel would be prohibitively expensive in computing time. Instead, they divided the larger image into square units, or tokens. The size is arbitrary, as the tokens could be made larger or smaller depending on the resolution of the original image (the default is 16 pixels on a side). But by processing pixels in groups, and applying self-attention to each, the ViT could quickly churn through enormous training data sets, spitting out increasingly accurate classifications.

The transformer classified images with over 90% accuracy — a far better result than anything Dosovitskiy expected — propelling it quickly to the top of the pack at the ImageNet classification challenge, a seminal image recognition contest. ViT's success suggested that maybe convolutions aren't as fundamental to computer vision as researchers believed.

"I think it is quite likely that CNNs will be replaced by vision transformers or derivatives thereof in the midterm future," said Neil Houlsby of Google Brain Zurich, who worked with Dosovitskiy to develop ViT. Those future models may be pure transformers, he said, or approaches that add self-attention to existing models.

Additional results bolster these predictions. Researchers routinely test their models for image classification on the ImageNet database, and at the start of 2022, an updated version of ViT was second only to a newer approach that combines CNNs with transformers. CNNs without transformers, the longtime champs, barely reached the top 10.

How Transformers Work

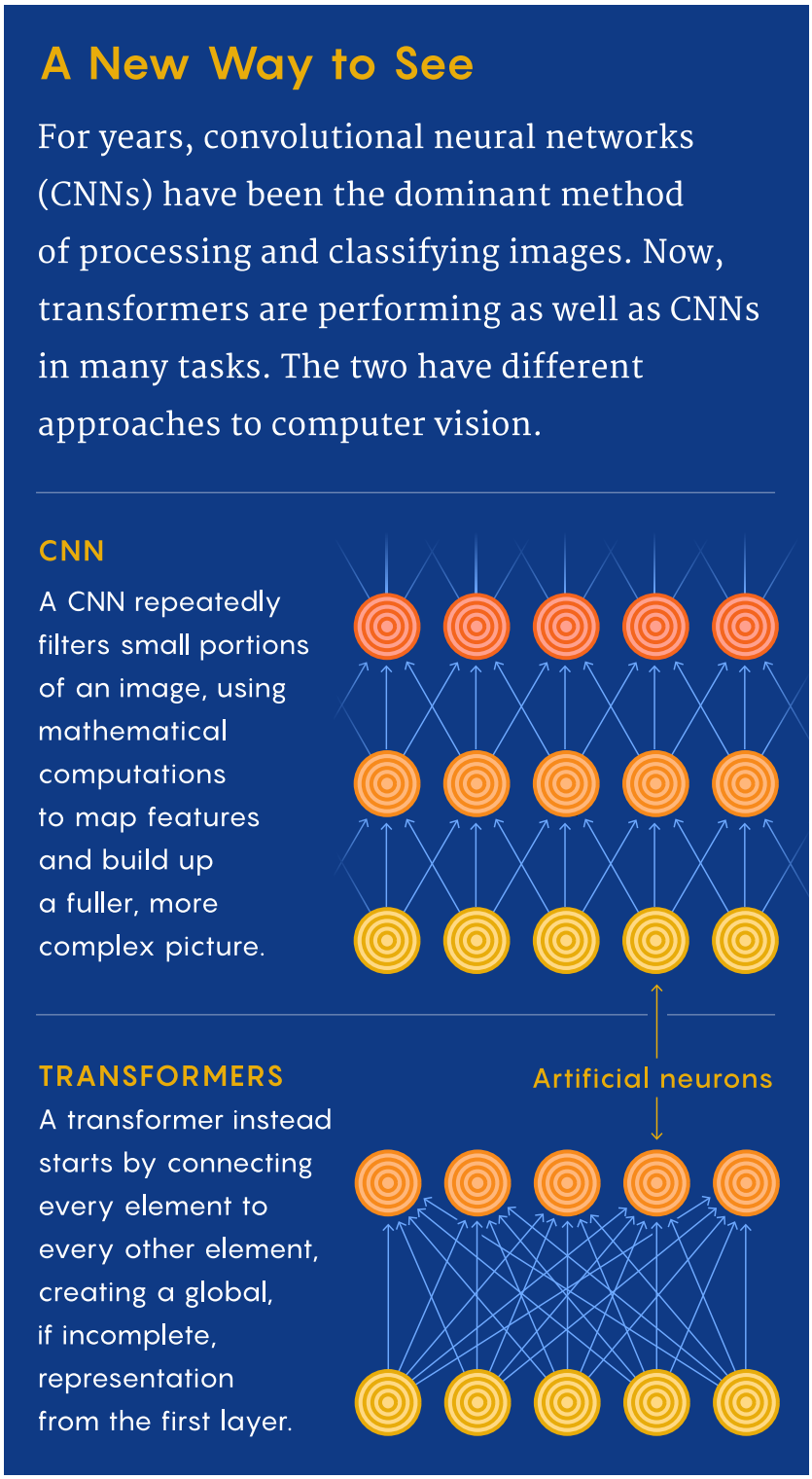
The ImageNet results demonstrated that transformers could compete with leading CNNs. But Maithra Raghu, a computer scientist at Google Brain's Mountain View office in California, wanted to know if they "see" images the same way CNNs do. Neural nets are notorious for being indecipherable black boxes, but there are ways to peek inside — such as by examining the net's input and output, layer by layer, to see how the training data flows through. Raghu's group did essentially this, picking ViT apart.

Maithra Raghu, a computer scientist at Google Brain, analyzed the Vision Transformer to determine exactly how it "sees" images. Unlike convolutional neural networks, which first focus on small portions to find details like edges or colors, transformers can capture the whole image from the beginning.

Arun Chaganty

Her group identified ways in which self-attention leads to a different means of perception within the algorithm. Ultimately, a transformer's power comes from the way it processes the encoded data of an image. "In CNNs, you start off being very local and slowly get a global perspective," said Raghu. A CNN recognizes an image pixel by pixel, identifying features like corners or lines by building its way up from the local to the global. But in transformers, with self-attention, even the very first layer of information processing makes connections between distant image locations (just as with language). If a CNN's approach is like starting at a single pixel and zooming out, a transformer slowly brings the whole fuzzy image into focus.

This difference is simpler to understand in the realm of language, where transformers were first conceived. Consider these sentences: “The owl spied a squirrel. It tried to grab it with its talons but only got the end of its tail.” The structure of the second sentence is confusing: What do those “it”s refer to? A CNN that focuses only on the words immediately around the “it”s would struggle, but a transformer connecting every word to every other word could discern that the owl did the grabbing, and the squirrel lost part of its tail.



Samuel Velasco/Quanta Magazine. Source: Dive into Deep Learning

Now that it was clear transformers processed images fundamentally

differently from convolutional networks, researchers only grew more excited. The transformer's versatility in converting data from a one-dimensional string, like a sentence, into a two-dimensional array, like an image, suggests that such a model could handle data of many other flavors. Wang, for example, thinks the transformer may be a big step toward achieving a kind of convergence of neural net architectures, resulting in a universal approach to computer vision — and perhaps to other AI tasks as well. "There are limitations to making it really happen, of course," he said, "but if there is a model that can universalize, where you can put all kinds of data in one machine, then certainly that's very fancy."

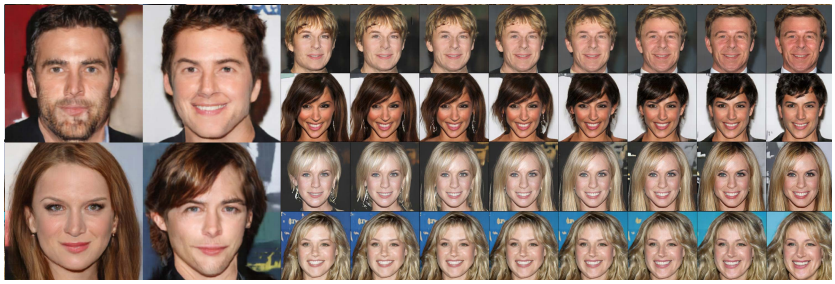
Convergence Coming

Now researchers want to apply transformers to an even harder task: inventing new images. Language tools such as GPT-3 can generate new text based on their training data. In a paper presented last year, Wang combined two transformer models in an effort to do the same for images, a much harder problem. When the double transformer network trained on the faces of more than 200,000 celebrities, it synthesized new facial images at moderate resolution. The invented celebrities are impressively realistic and at least as convincing as those created by CNNs, according to the inception score, a standard way of evaluating images generated by a neural net.

Wang argues that the transformer's success in generating images is even more surprising than ViT's prowess in image classification. "A generative model needs to synthesize, needs to be able to add information to look plausible," he said. And as with classification, the transformer approach is replacing convolutional networks.

Raghu and Wang see potential for new uses of transformers in multimodal

processing — a model that can simultaneously handle multiple types of data, like raw images, video and language. “It was trickier to do before,” Raghu said, because of that siloed approach where each type of data had its own specialized model. But transformers suggest a way to combine multiple input sources. “There’s a whole realm of interesting applications, combining some of these different types of data and images.” For example, multimodal networks might power a system that reads a person’s lips in addition to listening to their voice. “You could have a rich representation of both language and image information,” Raghu said, “and in a much deeper way than was possible before.”



These faces were created by a transformer-based network after training on a data set of more than 200,000 celebrity faces.

Courtesy of Atlas Wang

Emerging work suggests a spectrum of new uses for transformers in other AI domains, including teaching robots to recognize human body movements, training machines to discern emotions in speech, and detecting stress levels in electrocardiograms. Another program with transformer components is AlphaFold, which made headlines last year for its ability to quickly predict protein structures — a task that used to require a decade of intensive analysis.

The Trade-Off

Even if transformers can help unite and improve the tools of AI, emerging technologies often come at a steep cost, and this one is no different. A transformer requires a higher outlay of computational power in the pre-training phase before it can beat the accuracy of its conventional competitors.

That could be a problem. “People are always getting more and more interested in high-resolution images,” Wang said. That training expense could be a drawback to widespread implementation of transformers. However, Raghu sees the training hurdle as one that can be overcome simply enough with sophisticated filters and other tools.

Wang also points out that even though visual transformers have ignited new efforts to push AI forward — including his own — many of the new models still incorporate the best parts of convolutions. That means future models are more likely to use both than to abandon CNNs entirely, he says.

It also suggests the tantalizing prospect of some hybrid architecture that draws on the strengths of transformers in ways that today’s researchers can’t predict. “Perhaps we shouldn’t rush to the conclusion that the transformer will be the final model,” Wang said. But it’s increasingly likely that the transformer will at least be a part of whatever new super-tool comes to an AI shop near you.