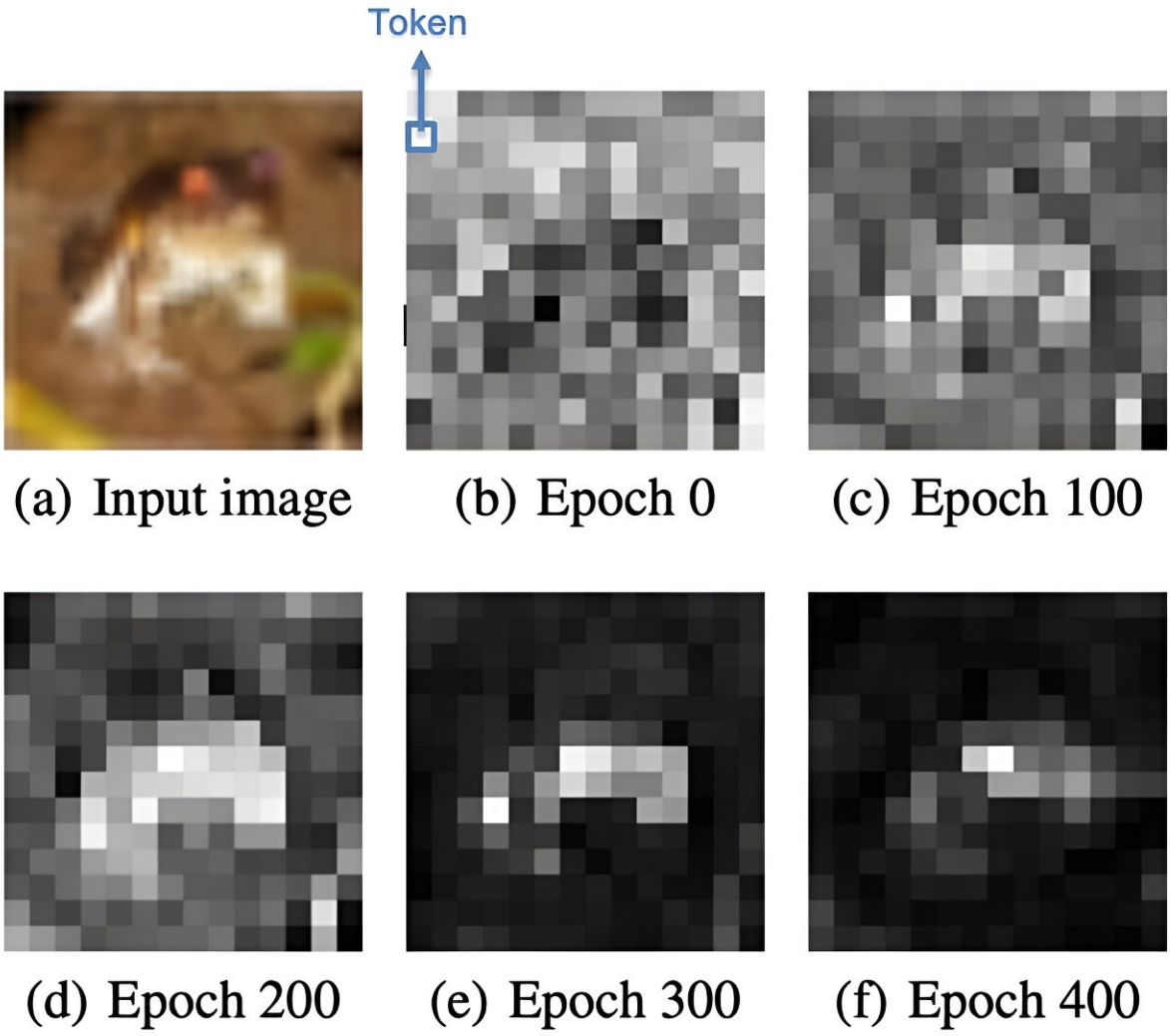
Understanding attention in large language models

December 13 2023, by Zach Robertson



In this test case from the study, a transformer model homes in on the relevant part of the image—the frog—over hundreds of rounds of training (epochs). At first, attention is directed randomly around the image. But with training, the

model learns to ignore the parts of the image that aren't the frog. Credit: Davoud Ataee Tarzanagh and Xuechen Zhang, SOTA Lab, University of Michigan

Chatbot users often recommend treating a series of prompts like a conversation, but how does the chatbot know what you're referring back to? A new study reveals the mechanism used by transformer models—like those driving modern chatbots—to decide what to pay attention to.

"Let's say you have some text which is very long, and you are asking the chatbot to identify key topics, to aggregate and summarize them. In order to do this, you need to be able to focus on exactly the right kinds of details," said Samet Oymak, assistant professor of electrical and computer engineering at the University of Michigan, who supervised the study to be presented at the Neural Information Processing Systems Conference on Wednesday, Dec. 13.

"We have mathematically shown for the first time how transformers learn to do this," he said.

Transformer architectures, first proposed in 2017, revolutionized natural language processing because they are so good at consuming very long strings of text—GPT-4 can handle whole books. Transformers break the text up into smaller pieces, called tokens, that are processed in parallel yet hang onto the context around each word. The GPT large language model spent years digesting text from the internet before springing onto the scene with a chatbot so conversant it could pass the bar exam.

The key to transformers is the attention mechanism: They decide what information is most relevant. What Oymak's team found is that part of a transformer's method for doing this is pretty old-school—they're

basically using support vector machines invented 30 years ago. An SVM sets a boundary so that the data falls into one of two categories. For instance, they're used to identify positive and negative sentiment in customer reviews. It turns out that transformers are doing something similar in deciding what to pay attention to—and what to ignore.

For all it sounds like you're talking to a person, ChatGPT is actually doing multidimensional math. Each token of text becomes a string of numbers called a vector. The first time you enter a prompt, ChatGPT uses its mathematical attention mechanism to attach weights to each vector, and hence each word and word combination, to decide which to take into account as it formulates its response. It's a word prediction algorithm, so it starts by predicting the first word that might begin a good response, then the next and the next, until the response is complete.

Then when you enter the next prompt, it feels like a continuation of the conversation to you, but ChatGPT is actually reading the whole conversation from the start, assigning new weights to each token, and then formulating a response based on this new appraisal. This is how it gives the impression of being able to recall something said earlier. It's also why, if you give it the first hundred lines of Romeo and Juliet and ask it to explain the problem between the Montagues and Capulets, it can summarize the most relevant interactions.

This much was already known about how transformer neural networks operate. However, transformer architectures aren't designed with an explicit threshold for what to pay attention to and what not to. That's what the SVM-like mechanism is for.

"We don't understand what these black box models are doing, and they are going mainstream," Oymak said. "This is one of the first studies to clearly show how the attention mechanism can find and retrieve a needle of useful information in a haystack of text."

The team intends to use this knowledge to make large language models more efficient and easier to interpret, and they anticipate that it will be useful for others working on aspects of AI where attention is important, such as perception, image processing and audio processing.

A second paper, digging more deeply into the topic, will be presented at the Mathematics of Modern Machine Learning workshop at NeurIPS 2023: Transformers as support vector machines.

More information: Max-margin token selection in attention mechanism (2023).

Provided by University of Michigan