The Magic of Large Language Models: Unpacking the Technology Behind Chatbots

Imagine stumbling upon a short movie script that describes a scene between a person and their AI assistant. The script has what the person asks the AI, but the AI's response has been torn off. Now, imagine having a powerful magical machine that can take any text and provide a sensible prediction of what word comes next. This is essentially what's happening when you interact with a chatbot. A large language model is a sophisticated mathematical function that predicts what word comes next for any piece of text.

How Large Language Models Work

Instead of predicting one word with certainty, a large language model assigns a probability to all possible next words. To build a chatbot, you lay out some text that describes an interaction between a user and a hypothetical AI assistant, add on whatever the user types in as the first part of the interaction, and then have the model repeatedly predict the next word that such a hypothetical AI assistant would say in response. This is what's presented to the user.

To make the output look more natural, it's essential to allow the model to select less likely words along the way at random. This means that even though the model itself is deterministic, a given prompt typically gives a different answer each time it's run.

Training Large Language Models

Models learn how to make these predictions by processing an enormous amount of text, typically pulled from the internet. For a standard human to read the amount of text that was used to train GPT-3, for example, if they read non-stop 24-7, it would take over **2600 years**. Larger models since then train on much, much more.

You can think of training a little bit like tuning the dials on a big machine. The way that a language model behaves is entirely determined by these many different continuous values, usually called *parameters* or *weights*. Changing those parameters will change the probabilities that the model gives for the next word on a given input.

What puts the large in large language model is how they can have **hundreds of billions of these parameters**. No human ever deliberately sets those parameters. Instead, they begin at random, meaning the model just outputs gibberish, but they're repeatedly refined based on many example pieces of text.

The Training Process

One of these training examples could be just a handful of words, or it could be thousands, but in either case, the way this works is to pass in all but the last word from that example into the model and compare the prediction that it makes with the true last word from the example.

An algorithm called *backpropagation* is used to tweak all of the parameters in such a way that it makes the model a little more likely to choose the true last word and a little less likely to choose all the others. When you do this for many, many trillions of examples, not only does the model start to give more accurate predictions on the training data, but it also starts to make more reasonable predictions on text that it's never seen before.

The Scale of Computation Involved

Given the huge number of parameters and the enormous amount of training data, the scale of computation involved in training a large language model is mind-boggling. To illustrate, imagine that you could perform one billion additions and multiplications every single second. How long do you think it would take for you to do all of the operations involved in training the largest language models?

The answer is actually much more than **100 million years**. This is only part of the story, though. This whole process is called *pre-training*. The goal of auto-completing a random passage of text from the internet is very different from the goal of being a good AI assistant.

Reinforcement Learning with Human Feedback

To address this, chatbots undergo another type of training, just as important, called *reinforcement learning with human feedback*. Workers flag unhelpful or problematic predictions, and their corrections further change the model's parameters, making them more likely to give predictions that users prefer.

The Role of GPUs and Transformers

This staggering amount of computation is only made possible by using special computer chips that are optimized for running many operations in parallel, known as *GPUs*. However, not all language models can be easily parallelized. Prior to 2017, most language models would process text one word at a time, but then a team of researchers at Google introduced a new model known as the *transformer*.

Transformers don't read text from the start to the finish; they soak it all in at once, in parallel. The very first step inside a transformer, and most other language models for that matter, is to associate each word with a long list of numbers.

Attention and Feed-Forward Neural Networks

What makes transformers unique is their reliance on a special operation known as *attention*. This operation gives all of these lists of numbers a chance to talk to one another and refine the meanings they encode based on the context around, all done in parallel.

Transformers typically also include a second type of operation known as a *feed-forward neural network*, and this gives the model extra capacity to store more patterns about language learned during training.

Conclusion

All of this data repeatedly flows through many different iterations of these two fundamental operations, and as it does so, the hope is that each list of numbers is enriched to encode whatever information might be needed to make an accurate prediction of what word follows in the passage.

When you use large language model predictions to autocomplete a prompt, the words that it generates are uncannily fluent, fascinating, and even useful. If you're a new viewer and you're curious about more details on how transformers and attention work, there are resources available to help you dive deeper into the topic.

- Check out the series on deep learning for a detailed explanation of attention and transformers.
- Watch a talk on the topic of large language models and their applications.

Large language models have revolutionized the way we interact with machines, enabling more natural and intuitive communication. As the technology continues to evolve, we can expect to see even more innovative applications of large language models in the future.
