Large Language Models such as ChatGPT in a Nutshell

Suppose given one word, you can predict the next word with some accuracy. And given two words, you can predict the next word with even better accuracy. And given three words… You get the idea. How could we design a system to do this? Well, let's assume every word has a number of dimensions such as part of speech, singular or plural, referring to living things, etc. But, let's get more abstract and assume each word has, say, 1000 dimensions. Now, we cannot say what these are nor can we visualize them and it is better not to try. Instead, let's assume they are mathematical abstractions and each word has a value on each of these dimensions and these values are coordinates that specify the word's unique location in this 1000-dimension space. This is the same idea as latitude/longitude specifying your home's unique location in two-dimensional space there are simply more dimensions. Further, let's assume there is a correlation or relation between a word's coordinates (location) in the 1000-dimension space and the next word. In other words, if we know the coordinates of word n, we can predict word n+1. That is, given word n, we can assign a probability for each word in our vocabulary being word n+1. Finally, if we can include context information such as the word's position in the sentence, our predictions get even better. This is the essence of Large Language Models (LLM).

How could we build such a system? One way would be to look at the billions of sentences in written documents, for each sentence take three words at a time, and work through the sentence predicting the middle word using the other two words and using the results of each prediction to guide the next prediction. We would be wrong just about all of the time at first but, eventually, after millions or trials, we would get pretty good. That is, over millions if not billions of trials, the system would learn the patterns. This sort of thing would take massive computing power and storage. But, once these resources became available over the last twenty years or so, a system like this was built and is the heart of LLM today.

Let's fill in a few details. To build this system we need to assign a number to each word and we need some sort of algorithm to predict a word, feed those results back and adjust the algorithm for the next prediction. Assigning a number (integer) to each word is pretty straightforward, i.e., each word in the vocabulary is assigned a unique integer. The second part is not quite so simple and consists of a computer program called a neural network with back propagation. The output of this network is a set of 1000 numbers for each word where each number can range from -1 to +1. These numbers represent that word's location in the 1000-diminsion space and can also be thought of as the word's strength or loading on each of the 1000 dimensions. These sets of numbers are called vectors, one vector for each word.

And now comes the magic. The neural network embeds semantics into the vectors. This embedding is not accomplished by our somehow assigning meaning to the word vectors. Instead, from the millions of examples of word usage that are passed through the mathematical model, the model learns by example using nothing more than linear algebra, a little calculus, and a bit of non-linear scaling. To illustrate that meaning is indeed

embedded in these vectors, If we subtract the vector for “male” from the vector for “king” and then add the vector for “female”, the resulting vector is very similar to the vector for “queen.” By similar, I mean the location of the point in 1000-dimension space for the resulting vector is very close to the point in space for the “queen” vector. Similarly, if the vector for “Capital” is added to the vector for “Alabama” the result is a vector very close to the vector for “Montgomery.”

And how is all of this used to generate text as if it came from a human? It works like this. After you type your question, the text is analyzed as discussed above and the most likely next word is appended to your text. Then all of that is passed through again and the next most likely word is appended to your text plus the first appended word and so forth. Eventually the process comes to an end and with the result being your question, a question mark, and an answer.

Two last details and we will sum it up. Above, I mentioned that our predictions are better if we know something about the context or word’s position in the sentence. For example, the word “bank” could refer to a river bank or a financial institution. The system handles such instances by building multiple vectors for each word using what are called transformers but the general idea is the same as discussed above. Secondly, I have used “word” as the basic unit of analysis. That is not quite correct. The model’s first step is to break the text into tokens. A token may be a word, a word ending such as “ing”, punctuation, etc. In this fashion, word roots with different endings can be accommodated. This detail does not change the preceding; in fact, you can substitute “token” for “word” and the meaning is the same just a bit more accurate.

So, we have built a system that seems to understand language and respond appropriately using existing written text, well-known mathematics and enormous computing power. In other words, instead of writing code to instruct the computer model as some of the earlier AI approaches attempted, we have simply given the computer millions of examples and programmed it to find patterns. With these results in hand, we then present the computer with new text, and the response is human like and suggesting cognition.

Hopefully, this brief description provides some relief from the mystery and even fear of AI that is portrayed in the media. It has taken humanity thousands of years to get to this point and the advances in the last few years are breathtaking. In the last ten or so years we have gone from a model that could rather slowly interpret human text and provide some crude responses to a tool that can provide information and even analysis in a fraction of the time it would take a human. We are likely on the threshold of advances that most of us cannot fathom. Rather than attempting to curtail, tax and regulate this technology it would seem far more productive to facilitate its security, fidelity, and growth and to ensure the United States maintains its leadership role.