Nicholas Hamilton

Cognitive Generative AI Externship

May 25 2025

Transformers Overview

RNNs -> LSTMs -> Transformers

How AI models generate text, understand language, etc. Transformers make tasks like this much more efficient and easier. Some of the challenges with previous models like RNNs and CNNs are solved with transformers.

At the heart of these models is the self-attention mechanism. This allows the model to focus on the most important parts of a sequence, no matter where in the piece they appear. This boosts its ability to capture global relationships. They use positional encoding to add a sense of structure to the data, so they can understand the order of elements in a sequence. They can also process entire sequences at once, thanks ot parallelization. This makes training faster, and capable of handling massive datasets. The transformer is the foundation of models like GPT.

Recurrent Neural Networks use the output from the previous step in the input for the next step. Long Short Term Memory (LSTM) is a specialized Recurrent Neural Network that was designed to address issues with traditional RNNs. Mainly their ability to retain long term information. Key features include a long list of features. Their ability to handle long term dependencies for starters. Their core which is a memory cell, acting as a storage unit to retain information. They have three gates, a forget gate that determines what to forget, an input gate deciding what should be added, and an output gate that decides what to be sent out. They have sequential processing, maintaining short and long term context. Source for additional reading: Long Short-Term Memory (LSTM) | NVIDIA Developer.

For additional reading on Transformers: Attention is All you Need

ChatGPT, google language assistants, etc. The older models could not focus on a large context. Transformers use Attention. They assign points to different words in a sentence, to determine its worth. It is made of two components, the encoder and the decoder. The encoder has multiple layers, which each have two parts. The first part is the self-attention mechanism, which allows the encoder to understand how all the words in a sentence relate to each other. It decides which other words are most relevant to it. After that, it goes through a feed-forward network, the data, which is refining the understanding of a sentence. These layers are repeated multiple times. The decoder also has self-attention layers, but they are masked to ensure the model generates one word at a time. This makes sure the model does not cheat by looking at future words. It has another layer called encoder-decoder attention. This helps it focus on the input sentence while generating the output. If generating ‘the cat sat on the mat’ to another language, it ensures it knows what word translates to the word its generating.

The encoder processes the input sequence, the decoder uses the encoder's representation and previous outputs to generate the next word. This continues until the entire output sequence is generated.

Transformers are a neural network architecture designed for processing sequential data like text, speech, and images. They use self-attention mechanism to identify relationships across a sequence. They process sequences in parallel unlike RNNs and LSTMs.

For my assignment, I used GPT-2, a text generation model. With a max length of 50, a temperature of 0.7, and a prompt of 'In the future, education will', I received the following result:

“In the future, education will include the creation of more than 1,700 colleges, universities and training centers in the United States. These facilities will be staffed by more than 4,000 teachers and students.”

With something like GPT-2, we are taking a massive vocabulary, and at each moment, calculating the probability for each of these words being next in the sequence. We have a probability distribution. The temperature parameter controls how sharp or flat our probability distribution becomes. With a very low temperature, like 0.1, it becomes exceptionally difficult for anything other than the dominant word to become next. We are creating a ‘peaked’ distribution. With a very high temperature, 2.0 for instance, we flatten the distribution. We are mathematically creating randomness. 0.7 is a good balance for a model like this.

The more words we add, the more complex the attention matrix becomes. This is because the transformer model keeps track of all words. With my experiments, it is clear that the higher the temperature, the more inaccurate the model.

prompt 1, temperature 0.1

The impact of AI on the future of work is clear.

"The future of work is not just about the future of work, it's about the future of work itself," says Dr. D.S. Kaur, a professor of

prompt 2, temperature 0.1

The impact of AI on the future of work is clear.

"The future of work is not just about the future of work, it's about the future of work itself," says Dr. R.J. Dyer, a professor of

prompt 3, temperature 0.7

The impact of AI on the future of work is already clear.

In 2007, for example, the US Department of Labor reported that "the U.S. has the highest employment growth for any country in Asia in five years, surpassing

prompt 4, temperature 0.7

The impact of AI on the future of work, and its impact on productivity, will be extremely significant."

prompt 5, temperature 2.0

The impact of AI on the future of work is evident when it looks after people: jobs that humans were supposed to lose and others that are now being offered new ones. And it isn't merely "better" people but also what those job positions create

prompt 6, temperature 2.0

The impact of AI on the future of work, including my colleagues with CRI, tells one about the enormous scope of scientific advancement – it requires new tools so fundamental – those of all institutions may not need much attention once AI becomes integrated with every computer