Capstone Project

Compulsory task 2

Read up on any innovative technology using NLP and write a brief summary about the technology, what it achieves/does and an overview of how it works (250-500 words).

The most prevalent NLP technology right now that is making a huge impact worldwide is the Transformer model, proposed by Vaswani et. Al. (2017), on which both open AI ChatGPT and Google BERT are based.

The transformer model is a type of neural network architecture presented in 2017 by a Google team that was developed to solve the problem of sequence transduction, which has to do with how to transform an input sequence to an output sequence without learning a general model of the transformation. It is widely used for various natural language processing applications, such as text summarization, question answering, text generation, and conversational response generation.

The transformer model has several advantages over previous models such as recurrent or convolutional neural networks. First, it can process the entire input sequence in parallel, which enables faster training. Second, it can capture long-range dependencies and contextual information using self-attention mechanisms that weight each input part differently. Third, it can be easily scaled up by increasing the number of layers or parameters.

The model architecture consists of an encoder and a decoder structure. Both of them are composed of multiple identical layers and each layer consists of two sub-layers, a self-attention and a feed-forward one. The encoder takes an input sequence, a vector of tokens and maps it to a sequence of continuous representations, a vector of numbers. The decoder accepts as input, the output of the encoder together with the output of the decoder at the previous time step and outputs a probability distribution over the next token in the sequence.

Together, the self-attention and feed-forward sublayers work in harmony to process the input sequence. The self-attention mechanism enables the model to capture dependencies and context by calculating the attention weights, which determine how much each token contributes to the sequence. On the other hand, the feed-forward sublayer captures complex patterns and relationships within the input sequence by applying non-linear transformations to the self-attention output, thus enabling it to better capture the semantics of the input sequence and enhance the representation's expressiveness. This combination allows the transformer model to effectively capture long-range dependencies and contextual information in the input sequence.

(1) Attention is all you need. https://arxiv.org/pdf/1706.03762.pdf.

(2) Transformer (machine learning model) - Wikipedia.
https://en.wikipedia.org/wiki/Transformer_%28machine_learning_model%29.

(3) Gentle Introduction to transduction in Machine Learning https://machinelearningmastery.com/transduction-in-machine-learning/

(4) What Is a Transformer Model? | NVIDIA Blogs. https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/.

(5) How transformers Work. https://towardsdatascience.com/transformers-141e32e69591.

(6) The Transformer Model. https://machinelearningmastery.com/the-transformer-model/

(7) The attention mechanism from scratch. https://machinelearningmastery.com/the-attention-mechanism-from-scratch