IE 7500: Question/Answering using Encoder and Decoder Transformers

Samani Karan: samani.k@northeastern.edu

Kota Mohit: kota.mo@northeastern.edu

Desai Saharsh: desai.sah@northeastern.edu

Objective:

- The objective is to Perform "Question/Answering" using state of art transformers.
- The problem we selected for this project is related to how different architectures of encoders and decoders in the transformers perform Q/A on general data and biomedical data.
- The aim is to see how a transformer model with both encoder and decoder is better than just an encoder stacked transformer model in summarizing the predicted answer.
- Transformer model uses self-attention as a technique to perform inferencing and this kind of mechanism helps to improve the downstream Question/Answering task.

Approach and Current State of Art:

Transformer Model
The transformer model is a neural network that learns context and meaning by tracking relationships in sequential data such as the words in sentences. These models apply self-attention mathematical techniques, to detect subtle ways even distant data elements in a series influence and depend on each other.

Encoder:
The task of the encoder is to map an input sequence to a sequence of continuous representations, which is then fed into a decoder.

Decoder:
The task of the decoder is to receive the output of the encoder together with the decoder output at the previous time step, to generate an output sequence.

BERT and BioBERT:

- BERT is a transformer model pretrained on a large corpus of English data in a self-supervised fashion.
- BioBERT is a transformer model pretrained on large medical corpus known as "PMC" and "PubMed".

- It makes use of Transformer, an attention mechanism that learns contextual relations between words in a text. It is considered bidirectional. These models are a stack of transformer encoders.

T-5 Tranformer:

- The T-5 transformer is a slightly different transformer than the previous encoder transformers.
- Where the previous encoder transformers needed the start and end tokens as the target, here we don't need starting or ending tokens.
- We just need the text of the answer as the target.
- Hence, here we need only three things, Context, Question, Answer.
- T5 is the simplest and more powerful of all the three transformers as it has both encoder and decoder and is also known as a multi- purpose transformer as it can be used to perform numerous downstream tasks.

Corpus to be used:

The open-source datasets used are:

- Squad Dataset will be used. Link ⇒ https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset
- PubMed dataset. Link ⇒ https://huggingface.co/datasets/pubmed

Deliverables and Progress Timeline:

- Final code and results: 10th April
- Final research paper:15th April
- Final PPT: 15th April