Report for 582 Final Project

Professor:Wenpeng Yin

May 5th, 2023

Sirui Qi

For this model, the problem is that this model is hard to generate the words which didn't show. The dataset still needs lots of data in there for increasing the scores. Otherside, it needs to spend lots of time on the prediction for workable results.

In this project, I decided to do the summarization of TED talks using the TED talk dataset, which was obtained from Kaggle. It is surprising that no one has attempted this before with the existing dataset. It's not like typical text summarization. TED talk transcripts were noted as speech with timestamps and broken-up sentences. We must also think about the length. Indeed, most TED talks have over 2000 words, making this task more challenging than conventional summarization.

For the dataset acquired from Kaggle, it has hundreds of attributes and more than 2000 speeches, but we opted to use just the transcripts for input and headlines for output. All information came from actual TED talks available to the public. We first removed all numbers and punctuation marks before starting the tasks.

Because pre-trained language models like GPT4 are banned and would have been the ideal choice, we utilized the seq2seq model with GRU layers instead. This model has two components: an encoder and a decoder. The encoder has an input layer, an embedding layer, and three GRU layers. The decoder features an input layer, an embedding layer, a GRU layer, and a Timing fully connected layer with Relu activation functions.

We achieved a BLEU score of approximately 35%, according to the outcome, indicating a reasonably good and understandable summary. Although we didn't reach over 60%, it demonstrates that this model is still functional. I believe that it have enough space to improve.

One issue with this model is its difficulty in generating words that didn't appear before. To improve the scores, the dataset requires more data. It causes the rate that has Out of Vocabulary to increase a lot and reduces the text generation performance. On the other hand, obtaining practical results necessitates spending a significant amount of time on making predictions. Also, some of the adpositions like a, the etc. influence the generation environment. These words are mostly meanless but make the BLEU score higher because most titles should include these words.

The summarization of the TED talk is meaningful but still needs to be improved task. The pre-trained language model still works better with this kind of work because of larger datasets and more useful data.