

CPSC 452 Final Project Proposal

Long Document Summarization with Efficient Transformers

Xiangru Tang, Wenxin Xu, Yizhi Zhao, Chengqian Jin

1. What is the problem that you will be investigating? Why is it interesting?

Transformer models such as BERT, and other variants have achieved state-of-the-art results on many challenging Natural Language Processing (NLP) tasks. Sequence to sequence (seq2seq) models have been successfully used for a variety of NLP tasks, including text summarization, machine translation and question answering. Our project is designed to build efficient transformers to better handle long sequences, thus we evaluate it on the long document summarization task, which tends to have long source sequences as the input.

2. What are the challenges of this project?

While successful, Transformer-based models have limits on the length of input sequence. When the input is long, the learning degrades particularly for tasks that require a comprehensive understanding of the entire paragraph or document. One of the main learning challenges for seq2seq models is that the decoder needs to attend to token level representations from the encoder to predict the next token, while at the same time it must learn from a large context.

3. What dataset are you using? How will you obtain it?

We will utilize the existing dataset, the PubMed and arXiv datasets, which contain scientific articles from PubMed and arXiv respectively, and will use the abstract of the articles as the target summary.

4. What deep learning approach will you use/develop for this project?

We propose a retrieval augmented framework for long document summarization. Introducing retrieved exemplars into summarization can bring not only guidance of writing format but also additional background knowledge. Motivated by the success of the dense passage retriever (DPR), we design our retrieval as a semantic matching process. Additionally, by grasping meta-knowledge in an unsupervised fashion within and across different information sources, we incorporate domain knowledge in the model.

5. How will you evaluate your results?

For the evaluation of our experiments, we will use classic evaluation metrics of ROUGE-1, ROUGE-2, and ROUGE-L. File2rouge will be used to evaluate Rouge score for all the results.