Long Document Abstractive Summarization

By: Andrew Kiruluta, Andreas Lemos, and Eric Lundy

Abstractive summarization has made significant strides since the introduction of the Transformer based model in NLP. However, the quadratic computational and memory complexities of large Transformers have limited their scalability for long document summarization as the token length for standard Transformer is limited to 512 tokens. One can try extractive summarization to reduce the length of the document into chunks followed by an abstractive approach on the reduced document. In the extractive step, the top k sentences are chosen per chunk to reduce the size of the total count to the token limits of the Transformer model.
Another way is to use successive abstractive summarization to summarize the document into chunks in iterative fashion. This is computationally very expensive. We will try the chuunk extractive method described above followed by an abstractive summarization of the individual chunks as a baseline model.
More recently, a group at Google has introduced a new implementation that replaces the self-attention module in Transformers with a Fourier transform that does not suffer from this quadratic computation penalty. We propose to extend this architecture to the long document summarization problem and compare the results to the baseline implementation.

Sources of training data: Gigaword dataset (https://catalog.ldc.upenn.edu/LDC2012T21) , CNN Mail Dataset (https://github.com/abisee/cnn-dailymail) and NIST Document Understanding Conferences (https://www-nlpir.nist.gov/projects/duc/data.html). Where human generated summaries exist, we intend to add a document similarity scoring module that automatically compares the abstractive summary output to the human annotated one. This will be done in addition to the standard ROUGE scores.

1. M. Zhong, P. Liu, Y. Chen. D. Wang, X. Qui and X. Huang, “Extractive Summarization as Text Matching”, ACL 2020. https://www.aclweb.org/anthology/2020.acl-main.552
2. M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang and A. Ahmed, “Big Bird: Transformers for Longer Sequences”, NeurIPS 2020. https://proceedings.neurips.cc//paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf
3. J. L.Thorp, J. Ainslie, I. Eckstein and S. Ontanon, “FNet: Mixing Tokens with Fourier Transforms, 05/2021. https://arxiv.org/abs/2105.03824v1
4. J. Zhang, Y. Zhao, M. Saleh and P. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”, ICML 2020. https://arxiv.org/pdf/1912.08777v2.pdf