ANLY-580 Natural Language Processing

Project Proposal: Text Summarization System

Group Members: Hanshen Jing | Peijin Li | Zihang Weng | Zixuan Wang

NetID: hj288 | pl724 | zw301 | zw306

Project Description

Text summarization involves compressing a document and preserving key content and meaning. In light of the time and labor involved in manual text summarization, automating the process is becoming more popular. Text summarization has a wide range of uses in NLP, such as text classification, question answering, legal text summarization, news summarization, and headline generation. The goal of this project is to reduce the length of the document in many areas. Our approach can be done in either an extractive or abstractive manner. While an extractive summarization model extracts salient fragments from the source document to form a summary. An abstractive summarization system aims to generate a semantically coherent and linguistically fluent summary by conditioning on the document (Pang et al., 2022). Motivated by concrete problems in summarization, we will propose novel models and show whether they can provide additional improvement in performance. In the final part, we propose a new dataset for summarization into multiple sentences and establish benchmarks.

Method

Modeling Approach

The seq2seq models with CNNs, RNNs, or transformers will be used in our project to address text summarization tasks. CNN, in general, is more computationally inexpensive and is easier to tune as it has a simpler architecture. In contrast, RNN has higher interpretability but takes more time to train as it processes sequentially. One of the models we will use in our project is DynamicConv, which is a lightweight CNN model with high performance in the self-attention results. The other one is words-lvt2k-temp-att, an Attentional Encoder-Decoder RNNs model, which will be used to process some critical tasks in text summarization, like crucial word modeling.

As methods that widely used across many NLP tasks, transformers-based models such as Bidirectional Encoder Representations from Transformers(BERT; Devlin et al. 2019) and GPT-3 will also be adopted and be fine-tuned. They achieved state-of-the-art performance in terms of the readable and natural summaries they can generate.

We will also apply BERT framwork in the project, we plan to inrtroduce a document-level encoder based on BERT. It is able to express the semantics of a document and obtain representation for its sentences.

Data

CNN-DailyMail News is an English-language dataset containing just over 300k unique news articles written by journalists at CNN and the Daily Mail.
Amazon Fine Food Review contains reviews of fine food from Amazon. It has around 500K reviews, including some basic information about the product and user, the ratings, and the text reviews.
We will use both of the datasets to conduct our project. We also have a supplement source that can be used to extend the functionality of our system in terms of performance after we get our system working. Moreover, we plan to use a new dataset from this source to evaluate the model.

Model Evaluation

The main focus of our project is text summarization. Therefore, it is important to check how well the system summarizes the original text. Here we propose several evaluation criteria.

1. Perform sentiment analysis over the original text and summarized text. Then we can calculate the accuracy in summarizing texts in terms of positive, negative, or neutral sentiments.
2. Perform text summarization on a topic the system did not see before. Then we can evaluate the system using the full-length Rouge F1 metric.
3. Combine the original documents and the summarization. Then implement LDA, LSA, together and compare the cosine similarity between the long and short texts.

Computational/Hardware Considerations

The modeling approach we employ in the project--CNNs, RNNs and transformers are memory intensive and computationally intensive, especially with backpropagation through time. In our case, we plan to use GPU, since its speed is faster than CPU when running massively parallel operations.

Concerns

A key challenge in summarization is to optimally compress the original document in a lossy manner such that the key concepts in the original document are preserved. A further challenge is to process long-form texts up to hundreds of pages or over 100,000 words.

Presentations

We plan to give a presentation during class to demonstrate our projects with a live demo to showcase the functionality of the system. For the live demo, we will implement a web-based UI and perform text summarization on selected documents to showcase the system.

References

Pang, B. *et al.* (2022) *Long document summarization with top-down and bottom-up inference*, *arXiv.org*. Available at: https://arxiv.org/abs/2203.07586 (Accessed: November 10, 2022).

Jacob Devlin. *et al.* (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota.

Useful Link:

https://paperswithcode.com/dataset/cnn-daily-mail-1

https://paperswithcode.com/datasets?task=text-summarization

https://blog.paperspace.com/implement-seq2seq-for-text-summarization-keras/

https://360digitmg.com/gpt-vs-bert