Project Proposal

Yihui Liu(yl1353), Sirui Wang(sw1245), Rae Zhang(yz961)

Project Goal

With the emerging science and technology, a significant amount of digital data are generated everyday by every walks of lives. Nowadays, our ways of communication and obtaining information are heavily depending on our digital devices, which led us to wonder if there is a shortcut to efficiently gain information. We are proposing our project idea of single document abstractive text summarization using BERT based model, HuggingFace, and Rouge score evaluation metric to test our model.

Our goal is to build a text summarization model which can generate meaningful and accurate summary of an entire text. Except from regular evaluation metric Rouge, we will also build a text classification model to make sure the model generated summary delivers the same topic as the original text does.

Method and Data Source

First, we plan to build a text classification model. We plan to use AG news datasets. This dataset contains 4 largest classes ("World", "Sports", "Business", "Sci/Tech"), and 30,000 training and 1,900 test samples per class. We plan to train a BERT-ITPT-FiT model on AG news dataset and use the trained model to generate classification classes for the article body and highlight in CNN_DailyMail datasets. This dataset contains 287,113 training, 13,368 validating and 11,490 testing news and highlight samples. The highlight could be used as a summarization of the news body.

Second, we plan to train a BERTSUM text summarization model on CNN_DailyMail datasets, and use this model to generate summarization of the news body.

Lastly, we will generate classification classes for our model generated summary, and compare the results with the classification results of the original news body and highlight.

We plan to evaluate our model using ROUGE score, and confusion matrix from BERT based model classification results.

For this final project, we will use Google Colab with the GPU hardware accelerator.

Expected outcomes

We would like to expect our text summarization model to have a high accuracy on certain categories after training and hyperparameter tuning. Meanwhile, we would like to adjust the minimum and maximum length of text summary outcomes without decreasing the model accuracy.

Live demonstrations and oral presentations will be presented as the result of our work.

Challenges and limitations

We are aware of certain challenges and limitations of our potential project. Some of the uncertainties include failure to retrieve a significant amount of text data within a particular category, the unforeseeable limitations of model evaluation methods, human errors in regards to cleaning and piping text data, limited computational power, not advanced algorithms or limited knowledge of hyperparameter tuning.

Reference

Community, T. H. F. D. (n.d.). *Ag_news · datasets at hugging face*. ag_news · Datasets at Hugging Face. Retrieved November 8, 2022, from https://huggingface.co/datasets/ag_news

Papers with code - how to fine-tune bert for text classification? How to Fine-Tune BERT for Text Classification? | Papers With Code. (n.d.). Retrieved November 8, 2022, from https://paperswithcode.com/paper/how-to-fine-tune-bert-for-text-classification

Leveraging pre-trained checkpoints for sequence generation tasks. Papers With Code. (n.d.). Retrieved November 8, 2022, from https://paperswithcode.com/paper/leveraging-pre-trained-checkpoints-for