Final Term Project – NLP

# Abstractive Text Summarization

## Individual Project Report

Author(s): Varun .R. Shah

12/11/2022

## Overview of the Project

Abstractive text summarization is the task of generating a headline or a short summary consisting of a few sentences that captures the salient ideas of an article or a passage. We use the adjective 'abstractive' to denote a summary that is not a mere selection of a few existing passages or sentences extracted from the source, but a compressed paraphrasing of the main contents of the document, potentially using vocabulary unseen in the source document

For this project, we use the CNN/Daily Mail Dataset. This is a dataset that includes news articles from DailyMail and CNN websites and includes human generated abstractive summary bullets. The authors released the scripts that crawl, extract, and generate pairs of passages and questions from these websites.

In all, the corpus has 286,817 training pairs, 13,368 validation pairs and 11,487 test pairs, as defined by their scripts. The source documents in the training set have 766 words spanning 29.74 sentences on an average while the summaries consist of 53 words and 3.72 sentences. The dataset is inspired from the one used in the Association for Computational Linguistics (ACL) 2017 paper Get To The Point: Summarization with Pointer-Generator Networks.

## Roles and Responsibilities

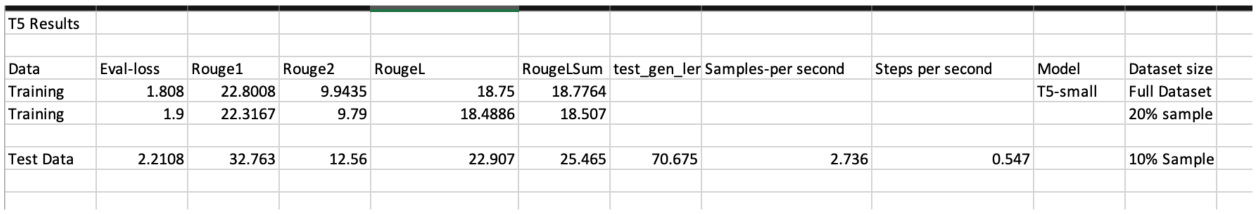
| Team Member | Area of Work | Shared Responsibility |
|---|---|---|
| Varun Shah | Data Preprocessing and T5 Model | Fine-tuning |
| Hemangi Kinger | Model interpretation and BART Model | Fine-tuning |
| Ishan Kuchroo | Build own trainer, GPT-2, PL-BART, and other models | Fine-tuning |

In this project, my role was preprocessing the data. The original corpus had news articles from DailyMail and CNN as individual file, with each file including the original article along with the human generated highlights on a single text file. This had to pulled and stored in a dataframe which was then converted to a csv file with 'Stories' and 'Highlights' as separate columns.

In addition to this, I tested the T5 Seq2Seq model available from HuggingFace with small changes and fine tuning to the available code.

The limitation I had with my model was the visibility of the Training and Validation Loss after each epoch and the efficiency of the code. It took over 12+ hours to run the model on the entire dataset, and eventually the results I got was only on 10% of the sample.

**Results:**



| T5 Results | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| Data | Eval-loss | Rouge1 | Rouge2 | RougeL | RougeLSum | test_gen_ler | Samples-per second | Steps per second | Model | Dataset size |
| Training | 1.808 | 22.8008 | 9.9435 | 18.75 | 18.7764 | | | | T5-small | Full Dataset |
| Training | 1.9 | 22.3167 | 9.79 | 18.4886 | 18.507 | | | | | 20% sample |
| | | | | | | | | | | |
| Test Data | 2.2108 | 32.763 | 12.56 | 22.907 | 25.465 | 70.675 | 2.736 | 0.547 | | 10% Sample |

## <u>Fine Tuning</u>

For Fine Tuning, I ran the models with different Learning Rates, Batch Size, Maximum Input Length, Maximum Target Length, Changes in Training Size and different optimizers.

## Conclusion

In conclusion, the preprocessing simplified a lot of the tasks when we were reading the data and run it in our model. The model from HuggingFace provided a very well written steps and procedures to fine tune the model and make changes based on the data we had. T5 model was the slowest among the other models we ran and took a long time to train the model and eventually run the test model.

## Appendix

GitHub Link - https://github.com/IshanKuchroo/Text-Summarization-for-CNN-and-DailyMail

## Datasource:

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

https://cs.nyu.edu/~kcho/DMQA/

## References

https://huggingface.co/course/chapter7/5?fw=pt

Rouge: A Package for Automatic Evaluation Summaries (https://aclanthology.org/W04-1013.pdf)

Text-To- Text Transfer Transformer (T5) (https://arxiv.org/abs/1910.10683)

https://huggingface.co/docs/transformers/model_doc/t5