Abstractive Text Summarization

Project Report

Author(s): Hemangi Kinger, Varun .R. Shah, Ishan Kuchroo

12/11/2022

Table of Contents:

1. Abstract
2. Introduction
3. Data Overview
4. Methods and Procedures
5. Fine Tuning
6. Metrics
7. SHAP Explainer
8. Results
9. Conclusion
10. Appendix and References

Final Term Project – NLP

Abstract

Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP can be used in different areas, Sentiment Analysis (Text Classification), Text Editing, Language Translation, Text Summarization, and various other use cases.

In this project, the goal is to summarize news articles using methods and procedures developed for NLP. Text summarization is the problem of creating a short, accurate, and fluent summary of a longer text document. The benefits of text summarization could be many, such as reducing reading time, extracting the most important information from a lengthy document and is helpful in researching documents.

We implement the Text Summarization task using different models built on Transformer architecture such as BART, Text-To-Text Transfer Transformer (T5) and GPT-2 fine-tuned for better results. This is implemented using PyTorch.

The project includes various fine tuning techniques used on the pretrained models and the evaluation metric used here is ROUGE, (Recall-Oriented Understudy for Gisting Evaluation),

We have also used SHAP Explainer for visually showcase how the summarization model works.

Introduction

The goal of this project is Automatic Text Summarization using Natural Language Processing. Text Summarization creates a shorter version of a document or an article that captures all the important information.

Text summarization techniques can be grouped into two:
Extractive Summarization (ES): Extract the most relevant information from a document.
Abstractive Summarization (AS): Generate new text that captures the most relevant information

Extractive Summarization (ES) is typically faster and easier than AS, since it extracts verbatim most salient sentences from text. However, there are two main problems with ES. First, the textual coherence is not guaranteed as resolving anaphora resolution is not paid attention in this approach. Second, redundant phrases still exist in the summary.

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. AS is relied on Natural Language Processing (NLP) techniques to copy-paste sentence fragments from the input document and maybe combine the selected content with extra linguistic information to generate the final summary. Abstractive Summarization can solve this problem by carrying out NLP techniques to post-process the output of ES such as sentence truncation, aggregation, generalization, reference adjustment and rewording.

In this project, we use various Transformer models for Abstractive text summarization. We implement this by fine tuning models such BART, T5 and GPT-2 and compare the results from each model. The metric used here is Rouge Metric.

Data Overview

Overview:

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.

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

Snapshot:

Source Document
(@entity0) wanted : film director , must be eager to shoot footage of golden lassos and invisible jets . <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie (the hollywood reporter first broke the story) . <eos> @entity5 was announced as director of the movie in november . <eos> @entity0 obtained a statement from @entity13 that says , " given creative differences , @entity13 and @entity5 have decided not to move forward with plans to develop and direct ' @entity9 ' together . <eos> " (@entity0 and @entity13 are both owned by @entity16 . <eos>) the movie , starring @entity18 in the title role of the @entity21 princess , is still set for release on june 00 , 0000 . <eos> it 's the first theatrical movie centering around the most popular female superhero . <eos> @entity18 will appear beforehand in " @entity25 v. @entity26 : @entity27 , " due out march 00 , 0000 . <eos> in the meantime , @entity13 will need to find someone new for the director 's chair . <eos>
Ground truth Summary
@entity5 is no longer set to direct the first " @entity9 " theatrical movie <eos> @entity5 left the project over " creative differences " <eos> movie is currently set for 0000

Methods and Procedures

For this project, we model abstractive text summarization using models built on Transformer architecture. We implement this by fine-tuning the below pretrained models:

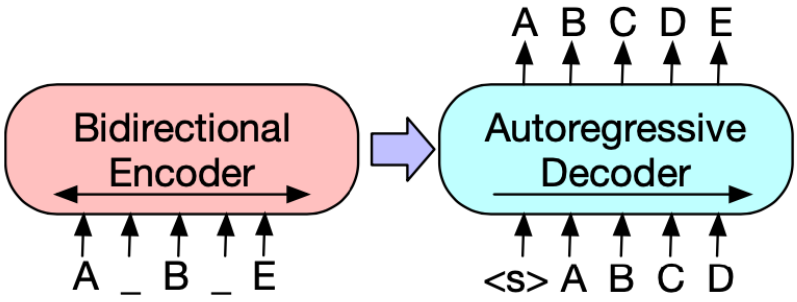
BART

BART is a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Transformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes.

This means that a fine-tuned BART model can take a text sequence (for example, English) as input and produce a different text sequence at the output (for example, French).

The pretraining task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token.

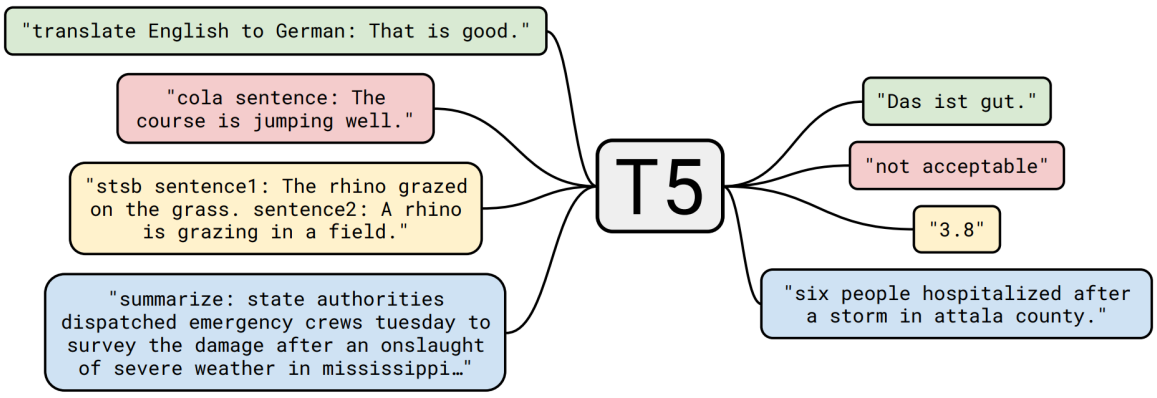
This type of model is relevant for machine translation, question-answering, text summarization, or sequence classification (categorizing input text sentences or tokens).



Text-To-Text Transfer Transformer (T5):

T5, or Text-to-Text Transfer Transformer, is a Transformer based architecture that uses a text-to-text approach. Every task including translation, question answering, and classification is cast as feeding the model text as input and training it to generate some target text.

Google has released the pre-trained T5 text-to-text framework models which are trained on the unlabelled large text corpus called C4 (Colossal Clean Crawled Corpus) using deep learning. C4 is the web extract text of 800Gb cleaned data. The cleaning process involves deduplication, discarding incomplete sentences, and removing offensive or noisy content.



Every task we consider uses text as input to the model, which is trained to generate some target text. This allows us to use the same model, loss function, and hyperparameters across our diverse set of tasks including translation (green), linguistic acceptability (red), sentence similarity (yellow), and document summarization (blue).

Fine Tuning

For Fine Tuning, there were different options such as running Seq2Seq model or training the model on our code. We proceeded with implementing a code which gave us better visibility on the Training Loss and Validation per Epoch, and also give us additional fine-tuning option such as using a Scheduler for Learning Rate with Warm-up steps. We ran the models with different Learning Rates, Batch Size, Maximum Input Length, Maximum Target Length, Changes in Training Size and different optimizers.

BART:

	1	2	3	4	5 increased the trian size to 5000
Optimizer	AdamW	Adam	Adam	Adam	Adam
Epoch	5	5	5	5	5
Learning Rate	0.00001	0.00001	0.00001	0.00001	0.00001
Btch Size	3	3	3	3	3
Max Length Input	264	264	264	264	264
Target Max length	64	64	80	90	90
rouge1	19.2038	19.2696	19.2645	19.2326	19.7834
rouge2	7.3176	7.2924	7.3135	7.3823	7.9065
rougeL	16.1014	16.11	16.2004	16.3171	16.7397
rougeLsum	16.0577	16.0609	16.1978	16.292	16.7586

T5:

Epochs	Optimizer	LR	Batch Size	Max Input Le	Max Output Length	Training RougeL (Best)	Test RougeL	Test RougeLSum
5	AdamW	0.00001	32	256	64	5.4377 (Epoch 5)	6.8921	
5	AdamW	0.0001	32	256	64	20.2347 (Epoch 5)	15.729	
5	AdamW	0.0001	25	256	64	20.2347 (Epoch 5)	15.729	
5	AdamW	0.0001	25	512	64	19.7691 (Epoch 5)	15.71	
5	**AdamW**	**0.001**	**25**	**256**	**64**	**20.1369 (Epoch 4)**	**15.8649**	15.866
5	Adam	0.0001	25	256	64	19.8974 (Epoch 3)	15.72	

Metrics

For Text Summarization, we used Rouge Metric to evaluate our model.

ROUGE stands for *Recall-Oriented Understudy for Gisting Evaluation*. It includes a measure to automatically determine the quality of a summary comparing it with human generated summaries. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

ROUGE-1

In ROUGE-1, Precision and Recall compare the similarity of **uni-grams** between reference and candidate summaries. By uni-grams, we simply mean each token of comparison is a single word.

ROUGE-2

In ROUGE-2, Precision and Recall compare the similarity of **bi-grams** between reference and candidate summaries. By bi-grams, we mean each token of comparison is 2 consecutive words from the reference and candidate summaries.

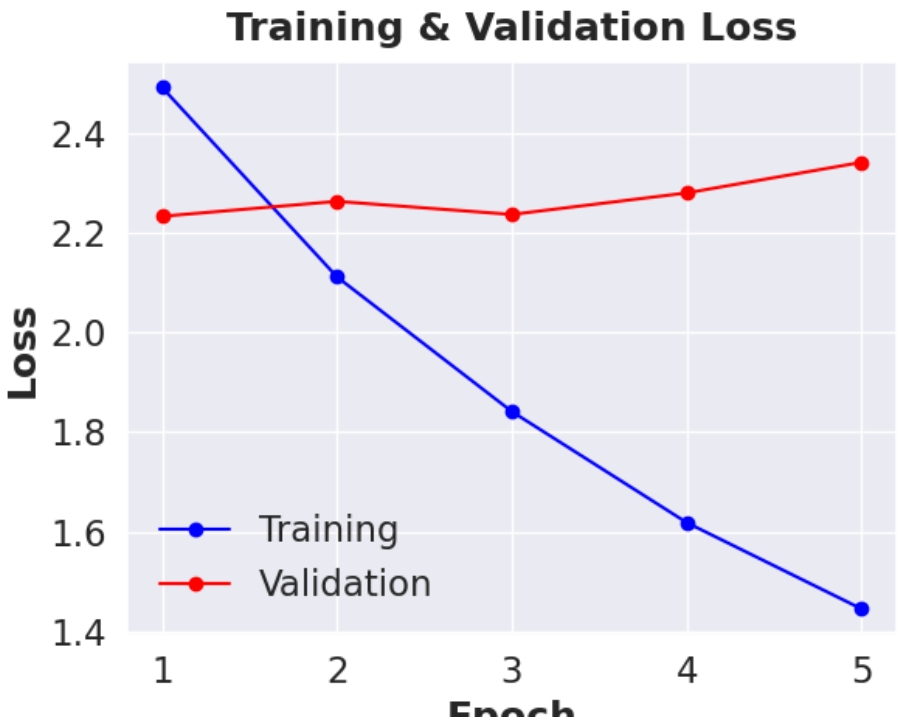
ROUGE-L

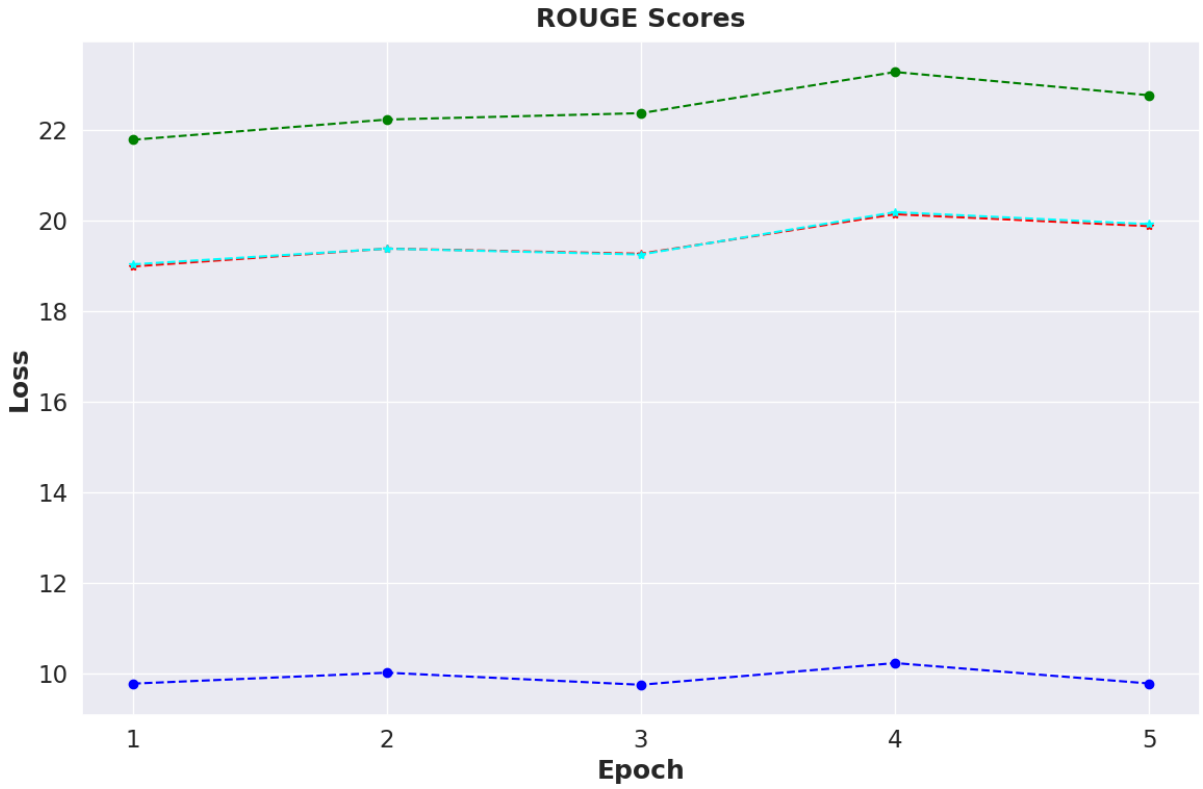
In ROUGE-L, Precision and Recall measures the **Longest Common Subsequence (LCS)** words between reference and candidate summaries. By LCS, we refer to word tokens that are **in sequence**, but **not necessarily consecutive**.

Results:

T5 Model:

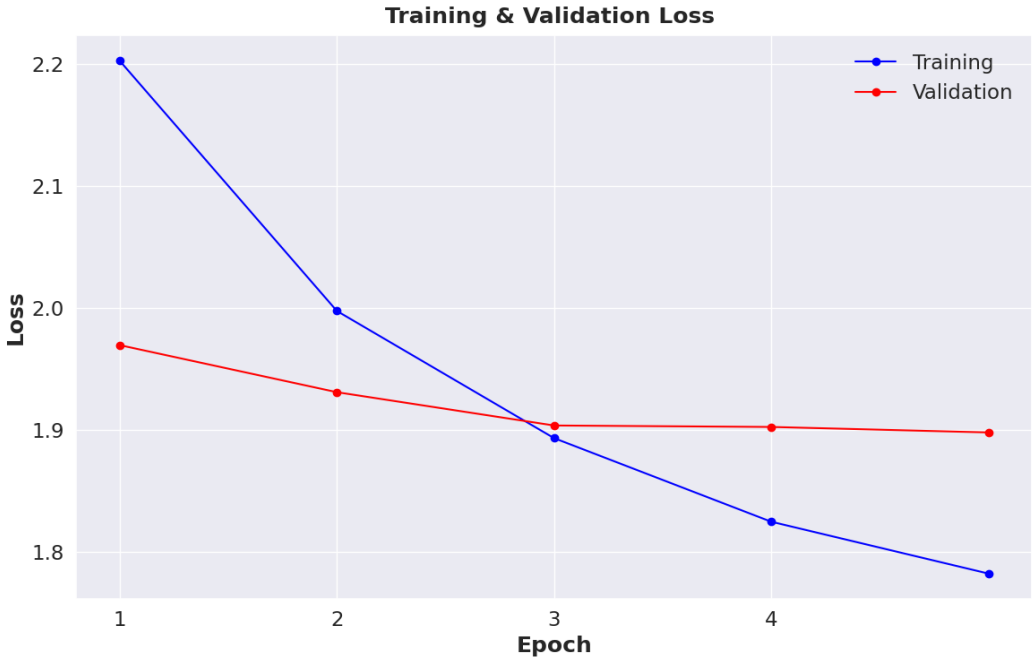
Epochs	Optimizer	LR	Batch Size	Max Input Length	Max Output Length	Training RougeL (Best)	Test RougeL
5	**AdamW**	**0.001**	**32**	**256**	**64**	**21.2265**	**16.6188**

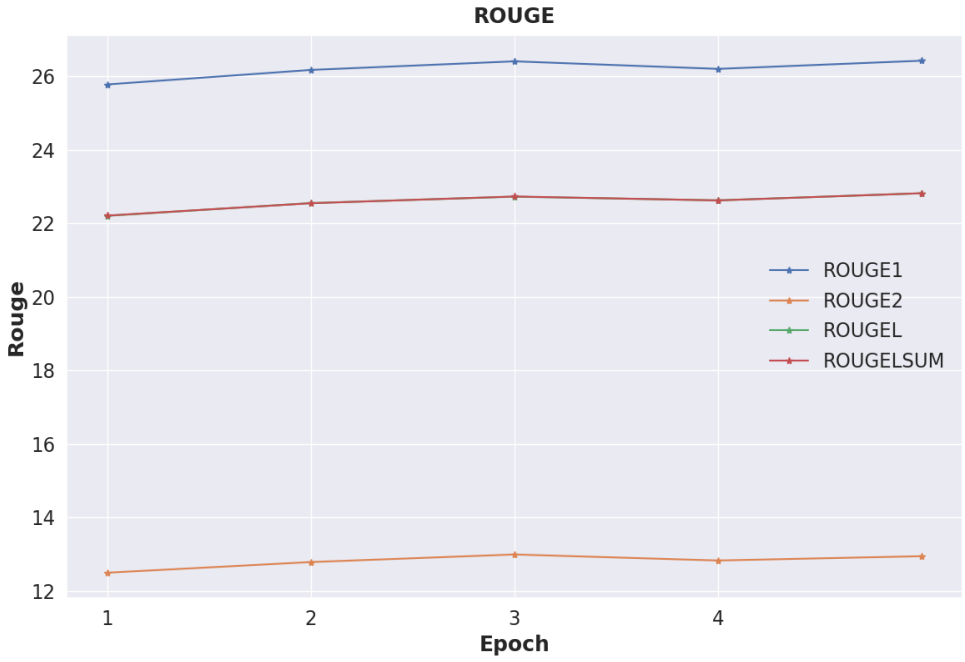




BART Scores:

Epochs	Optimizer	LR	Batch Size	Max Input Length	Max Output Length	Training RougeL (Best)	Test RougeL
5	**Adam**	**0.0001**	**3**	**264**	**90**	**17.654**	**16.739**


Training & Validation Loss
Training
Validation
Loss
2.2
2.1
2.0
1.9
1.8
1
2
3
4
Epoch

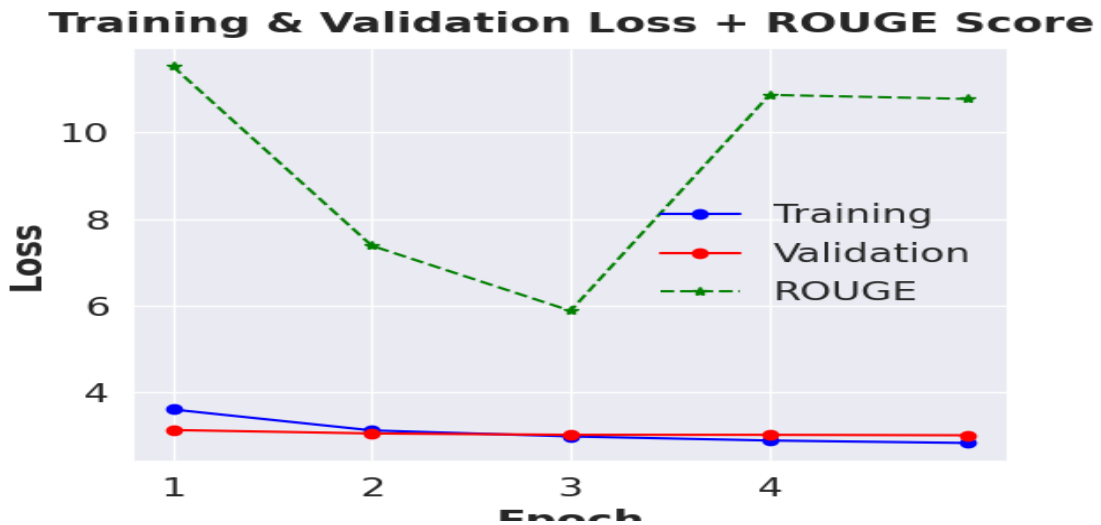

ROUGE
ROUGE1
ROUGE2
ROUGEL
ROUGELSUM
Rouge
26
24
22
20
18
16
14
12
1
2
3
4
Epoch

Comparison with other Models:

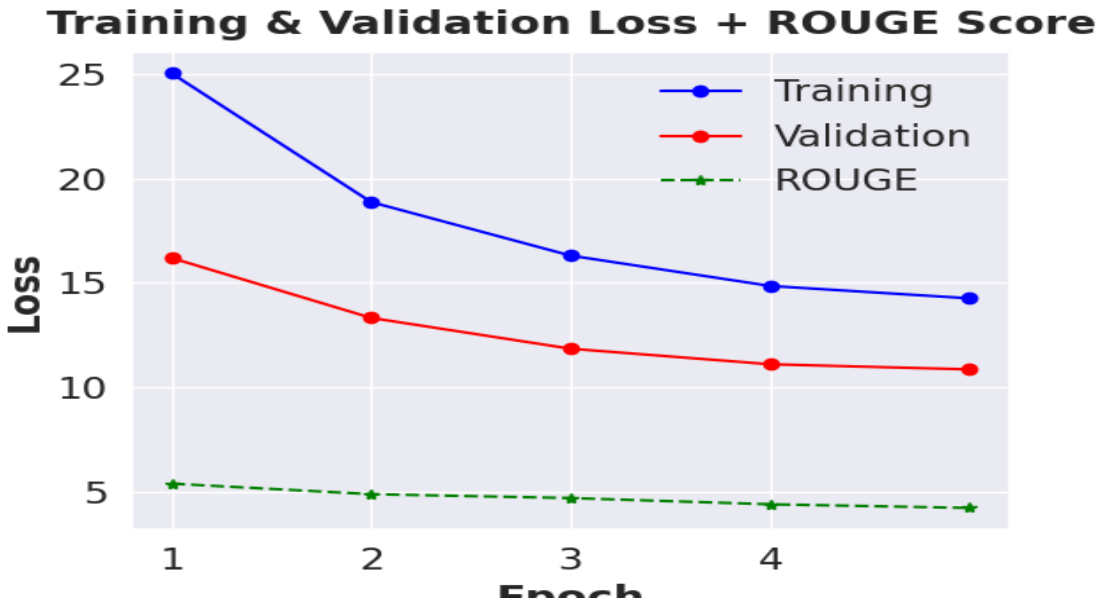
Case	Model	Configuration	Test Result
1	**PL-BART**	Epoch=5 Optimizer=Adam Batch=3 LR = 0.00001 Input Length = 256 Output Length = 64	{'rouge1': 9.7446, 'rouge2': 2.7701, 'rougeL': 8.4327, 'rougeLsum': 8.453}
2	**MT5-Small**	Epoch=5 Optimizer=Adam Batch=3 LR = 0.00001 Input Length = 256 Output Length = 64	{'rouge1': 3.1787, 'rouge2': 0.4494, 'rougeL': 3.0539, 'rougeLsum': 3.078}
3	**Prophetnet-large-uncased**	Epoch=5 Optimizer=Adam Batch=3 LR = 0.00001 Input Length = 256 Output Length = 64	{'rouge1': 8.8431, 'rouge2': 1.1545, 'rougeL': 8.232, 'rougeLsum': 8.2912}

4	**Tiny-Mbart**	Epoch=5 Optimizer=Adam Batch=3 LR = 0.00001 Input Length = 256 Output Length = 64	{'rouge1': 0.0, 'rouge2': 0.0, 'rougeL': 0.0, 'rougeLsum': 0.0}
5	**BlenderBot-Small**	Epoch=5 Optimizer=Adam Batch=3 LR = 0.00001 Input Length = 256 Output Length = 64	{'rouge1': 19.2891, 'rouge2': 6.8296, 'rougeL': 13.4452, 'rougeLsum': 13.4553}
6	**GPT-2**	Epoch=5 Optimizer=Adam Batch=3 LR = 0.00001 Input Length = 256 Output Length = 64	{'rouge1': 22.7762, 'rouge2': 9.2689, 'rougeL': 16.3148, 'rougeLsum': 16.3412}

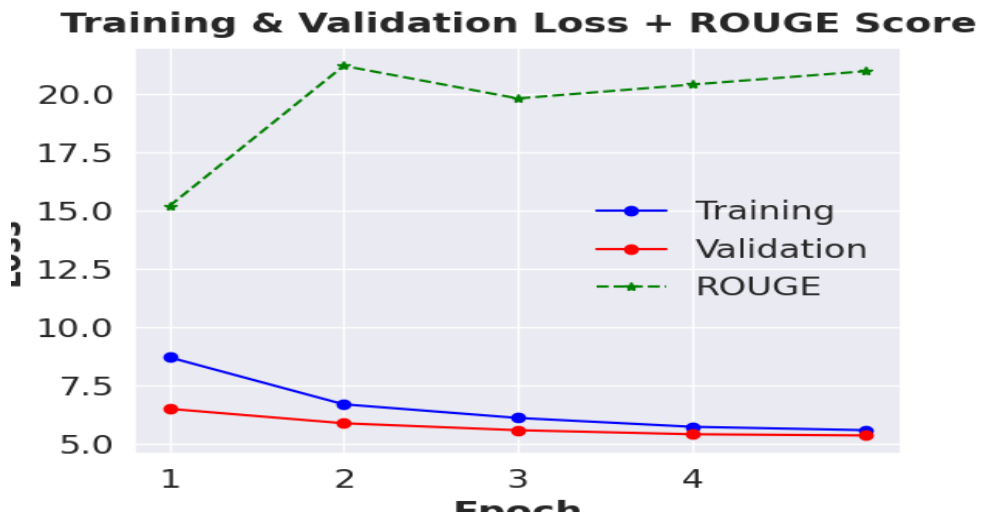
PLBART:



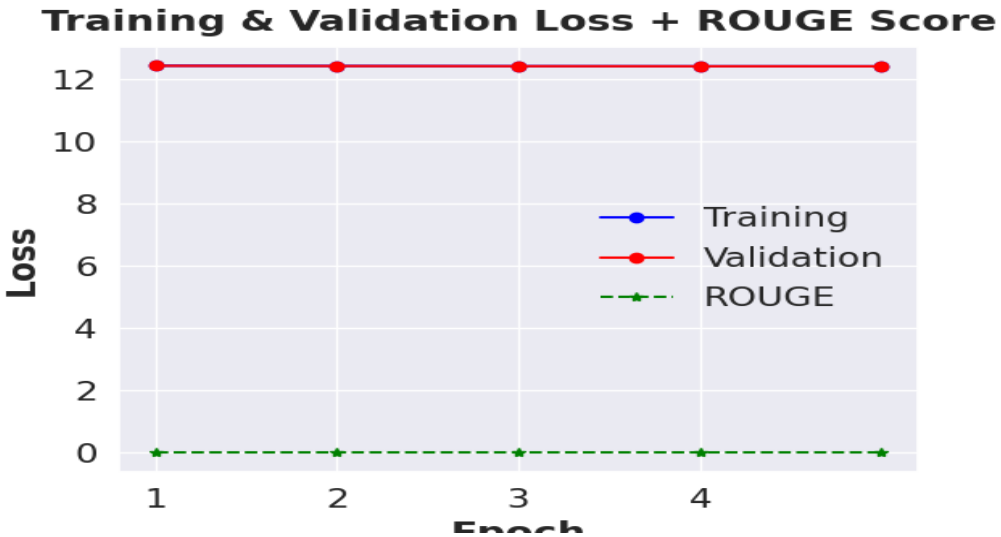
MT5-Small



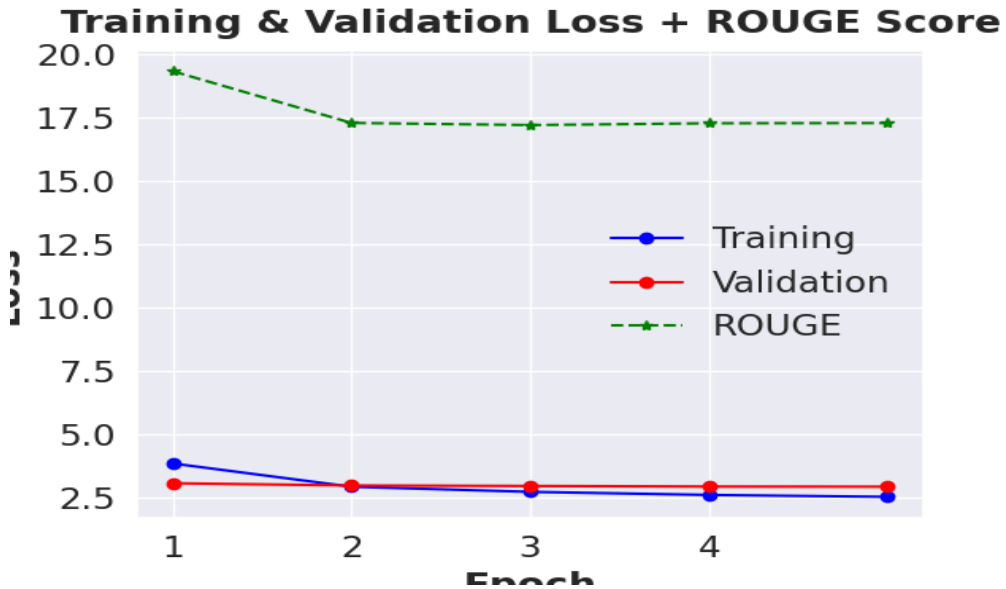
Prophetnet-large-uncased:



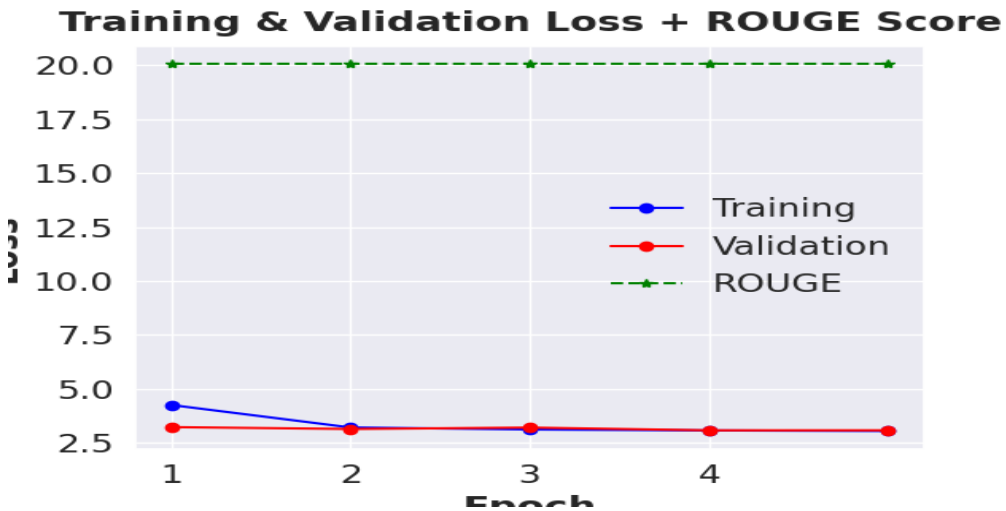
Tiny-Mbart:



BlenderBot-Small:



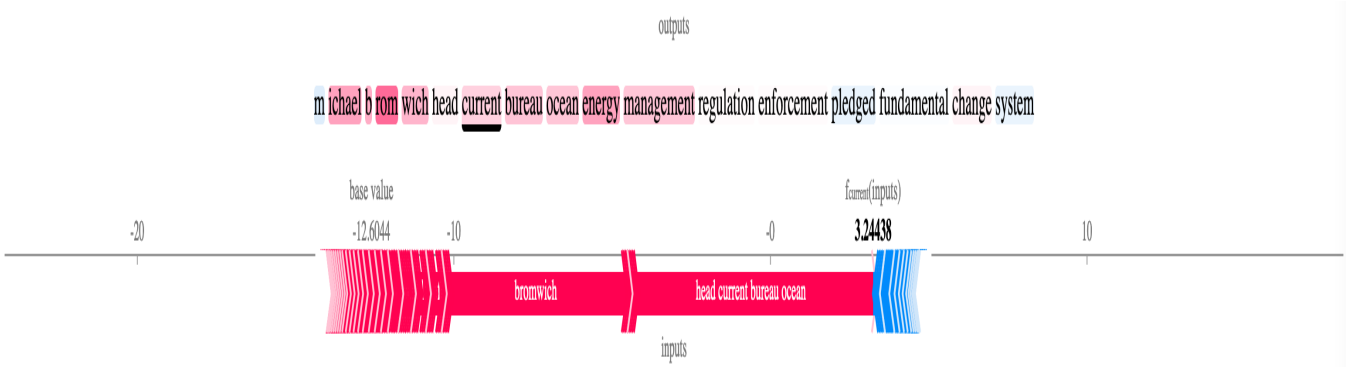
GPT-2:



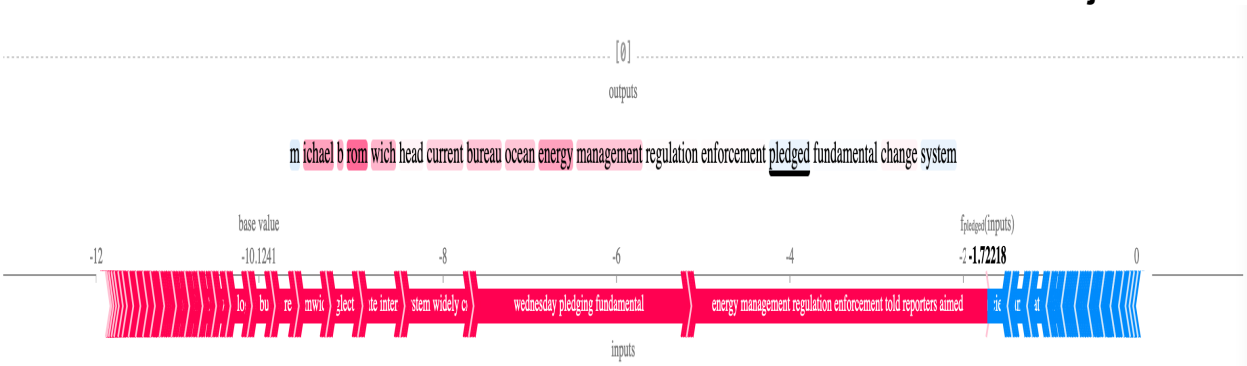
SHAP Explainer

SHAP or SHAPley Additive exPlanations is a visualization tool that can be used for making a machine learning model more explainable by visualizing its output. It can be used for explaining the prediction of any model by computing the contribution of each feature to the prediction. It is a combination of various tools like lime, SHAPely sampling values, DeepLift, QII, and many more.

When we pass a single instance to the text plot we get the importance of each token overlayed on the original text that corresponds to that token. Red regions correspond to parts of the text that increase the output of the model when they are included, while blue regions decrease the output of the model when they are included. In the context of the summarization model here red corresponds to a more positive impact and blue a more negative impact.



cnn interior department officials announced extensive restructuring agencies oversee offshore oil drilling wednesday pledging fundamental change system widely criticized worst oil spill us history expected moves split sincerenamed minerals management service separate agencies one responsible approving offshore leases another enforce safety environmental laws michael bromwich head current bureau ocean energy management regulation enforcement told reporters aimed beefing safety decades neglect conflicts interest among regulators much moving boxes around bromwich said comprehensive review fundamental change way agencies operate interior department said plans newly created bureau ocean energy management bureau safety environmental enforcement running october 1 separate agency collect revenues leases spun mms 2010 announcement came eight days presidential commission investigated last april sinking drill rig deepwater horizon sharply criticized regulators passivity finding outmatched underfunded conflicting prevented effective oversight offshore oil industry sinking killed 11 workers uncapped undersea gusher spewed 200 million gallons crude gulf mexico contained june efforts reshape mms began shortly explosion disaster internal watchdog reports criticized mms inspectors accepting meals tickets sporting events oil companies monitored gulf found regulators agency colorado office received improper gifts engaged illegal drug use inappropriate sexual relationships energy industry interior department also setting committee top scientists engineers recommend improvements offshore safety well containment procedures tom hunter former head sandia national laboratories lead committee department announced think become abundantly clear various including report deepwater horizon commission industry moved deeper deeper nation waters interior secretary ken salazar said technology oil gas drilling waters developed quickly technologies come along handinhand simply



cnn interior department officials announced extensive restructuring agencies oversee offshore oil drilling wednesday pledging fundamental change system widely criticized worst oil spill us history expected moves split sincerenamed minerals management service separate agencies one responsible approving offshore leases another enforce safety environmental laws michael bromwich head current bureau ocean energy management regulation enforcement told reporters aimed beefing safety decades neglect conflicts interest among regulators much moving boxes around bromwich said comprehensive review fundamental change way agencies operate interior department said plans newly created bureau ocean energy management bureau safety environmental enforcement running october 1 separate agency collect revenues leases spun mms 2010 announcement came eight days presidential commission investigated last april sinking drill rig deepwater horizon sharply criticized regulators passivity finding outmatched underfunded conflicting prevented effective oversight offshore oil industry sinking killed 11 workers uncapped undersea gusher spewed 200 million gallons crude gulf mexico contained june efforts reshape mms began shortly explosion disaster internal watchdog reports criticized mms inspectors accepting meals tickets sporting events oil companies monitored gulf found regulators agency colorado office received improper gifts engaged illegal drug use inappropriate sexual relationships energy industry interior department also setting committee top scientists engineers recommend improvements offshore safety well containment procedures tom hunter former head sandia national laboratories lead committee department announced think become abundantly clear various including report deepwater horizon commission industry moved deeper deeper nation waters interior secretary ken salazar said technology oil gas drilling waters developed quickly technologies come along handinhand simply

Results:

Original Text	Predicted Summary
(CNN) -- Interior Department officials announced an extensive restructuring of the agencies that oversee offshore oil drilling Wednesday, pledging "fundamental change" for a system widely criticized after the worst oil spill in U.S. history.\n\nThe expected moves split the since-renamed Minerals Management Service into separate agencies, with one responsible for approving offshore leases and another to enforce safety and environmental laws. Michael Bromwich, the head of the current Bureau of Ocean Energy Management, Regulation, and Enforcement, told reporters the reorganization is aimed at beefing up safety after "decades of neglect" and conflicts of interest among regulators.\n\n"This reorganization is much more than just moving boxes around," Bromwich said. "It is about a comprehensive review and a fundamental change in the way that these agencies operate."\n\nThe Interior Department said it plans to have its newly created Bureau of Ocean Energy Management and Bureau of Safety and Environmental Enforcement up and running by October 1. A separate agency to collect revenues from leases was spun off from MMS in 2010.\n\nThe announcement came eight days after the presidential commission that investigated last April\'s sinking of the drill rig Deepwater Horizon sharply criticized regulators for their passivity, finding they were outmatched, underfunded and had conflicting responsibilities that prevented them from effective oversight of the offshore oil industry. The sinking killed 11 workers and uncapped an undersea gusher that spewed more than 200 million gallons of crude into…	T5: the interior department announces an extensive restructuring of the agencies that oversee offshore oil drilling. the agency is split into separate agencies, with one responsible for approving offshore leases and another to enforce safety and environmental laws. the reorganization is aimed at beefing up safety after "decades of neglect" and conflicts of interest among regulators. BART: 'NEW: Former head of Sandia National Laboratories to lead committee on safety, well containment. "This reorganization is much more than just moving boxes around," Michael Bromwich says. Management Service will be split into separate agencies. The agency is responsible for approving offshore leases and enforcing safety and environmental laws.'

Conclusion

In conclusion, we implemented Abstractive Text Summarization using a few pretrained models such as Text-to-Text Transformer (T5) and BART. We fine-tuned the pretrained models making various changes to the Batch Size, Learning Rate, Optimization algorithms. With all the changes, BART provided us with the best results.

For metrics, we used ROUGE metric to evaluate our models and confirmed BART having the best score. We implemented SHAP for explainibility to visually check how our model shows the weights for each attribute.

For improvements, and learning, overall the project gave us satisfying results however it could not handle large dataset, hence we had to reduce the sample size and then run our models. We learnt that the T5 models takes a significant amount of time to train the model and generate the summarized text compared to BART.

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

https://shap.readthedocs.io/en/latest/example_notebooks/text_examples/summarization/Abstractive%20Summarization%20Explanation%20Demo.html

https://medium.com/analytics-vidhya/text-summarization-using-bert-gpt2-xlnet-5ee80608e961

https://huggingface.co/course/chapter3/4?fw=tf#the-training-loop

https://www.kaggle.com/code/sumantindurkhya/text-summarization-seq2seq-pytorch/notebook

https://huggingface.co/docs/transformers/training#train-in-native-pytorch

https://blog.paperspace.com/generating-text-summaries-gpt-2/

http://reyfarhan.com/posts/easy-gpt2-finetuning-huggingface/

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

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (https://arxiv.org/abs/1910.13461)

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

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