Text Summarization

Introduction:

This report outlines the details of the text summarization project developing a language model capable of generating summaries from input text. The project is based on the T5 model architecture and involves training the model on the CNN and daily Mail dataset of text mails as inputs and mails summary as targets. The trained model demonstrated the ability to generate concise summaries.

Data Collection:

We used the CNN and Daily Mail dataset from the datasets library provided by the Hugging Face Transformers library. The dataset size with over 200k observations.

The CNN/Daily Mail dataset is a collection of news articles paired with their corresponding summaries. The dataset contains articles from two major news sources: CNN and Daily Mail. The goal of using this dataset is to train models that can generate concise summaries of news articles.

Each article is paired with one or more human-written summaries. These summaries are intended to capture the main points of the article in a concise form.
Handling the variability in article and summary lengths can be challenging. Models need to learn how to deal with different lengths and generate summaries that capture the main content.

Data Preprocessing:

The collected data underwent preprocessing to prepare it for training. Tokenization was applied to segment the text into smaller units, such as words or subwords. Encoding techniques were used to convert the tokenized text into numerical representations that could be processed by the model.The data preprocessing steps included tokenization, encoding, and handling of special characters or symbols.

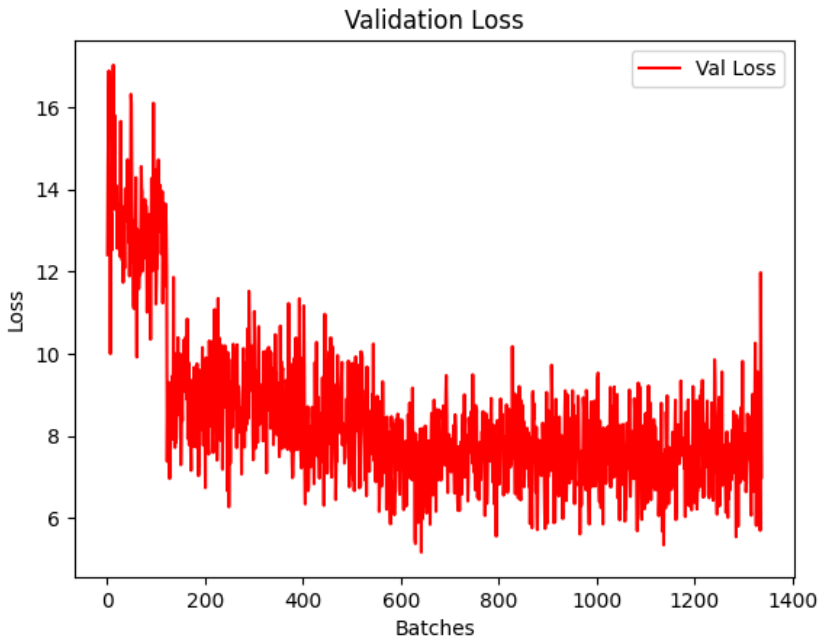
Model Construction:

The project utilized the T5 (Text-To-Text Transfer Transformer) model architecture, a state-of-the-art transformer-based model designed for a wide range of natural language processing tasks. The T5 model was trained on a large corpus of text data, which was preprocessed to prepare it for training.

The training process involved optimizing the model's parameters using a suitable optimizer of Adam and a loss function of CrossEntropyLoss. Hyperparameters such as learning rate, batch size, and sequence length are chosen based on model permanence and the memory utilization based on Google Colab V100 Gpu instances. The training duration was 6-8 hours, the number of epochs is set to 1 due to the large data points.

Model Evaluation:

We used the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric to evaluate our model, which is a metric specially made for Automatic Text Summarization.
ROUGE-L replaces the N-gram counting by the LCS (Longest Common Subsequence), which is much more robust for a variety of cases.
This metric measured the quality of the generated summaries by comparing them to human-generated reference summaries.



Model Inference:

Input:

"Islington Council said the move was necessary to improve air quality in the borough. The authority said pollutants in diesel exhausts had been linked to heart and lung disease. But a motoring group said drivers were confused by the penalising of one fuel over another as today's diesel cars were the cleanest ever. Mike Hawes, from the Society of Motor Manufacturers and Traders, said: Bans and parking taxes on diesel vehicles therefore make no sense from an environmental point of view. The allegations against diesel cars made in recent months threaten to misguide policy-making and undermine public confidence in diesel. It's time to put the record straight. The surcharge, which will be imposed by Islington Council from Monday, coincides with an increase in its parking permits. The cost of an Islington resident's permit depends on the emission or engine size of their vehicle with the highest priced at £444 for a year from Monday. This was found to be the highest charge for some drivers in the capital, according to a recent survey carried out by Churchill Car Insurance. Claudia Webbe, the council's executive member for transport and environment, said diesel fumes were the major cause of pollution. She added: Pollutants in diesel exhausts have been linked to heart and lung diseases, which are major causes of serious and long-term health issues and even death in Islington, and the surcharge will encourage a move away from diesel. In 2014 the council threatened to hand out £20 fines to drivers who refused to switch off their diesel engines while parked"

Generated Summary:

“Islington Council said the move was necessary to improve air quality. Pollutants in diesel exhausts have been linked to heart and lung disease. Motoring group said drivers confused by penalising one fuel over another.”

The evaluation results demonstrated that the trained model achieved promising performance in generating accurate and coherent summaries. The ROUGE scores indicated a significant overlap with the reference summaries, indicating the model's ability to capture the key information and context from the input text.

Conclusion:

The project successfully developed a language model capable of generating accurate and meaningful summaries from input text. The utilization of the T5 model architecture, along with appropriate preprocessing techniques and training strategies, enabled the model to capture the essential information and generate coherent summaries.

The trained model exhibited promising performance in summarizing text, demonstrating its potential for various applications such as news summarization, document

References:
[1] https://medium.com/analytics-vidhya/text-summarization-using-nlp-3e85ad0c6349
[2]https://medium.com/besedo-engineering/text-summarization-part-1-a-gentle-introduction-to-automatic-text-summarization-31f14b1f9e53