Text summarization problem
DSTI : Deep Learning with Python

Mayank Bhandari
Ouc-Houang Fogoum Philippe Jacques

November 23, 2023

1 Introduction

Text summarization is a crucial aspect of natural language processing, aiming to condense lengthy texts while retaining essential information. In this project, the focus was on utilizing the google/pegasus cnn dailymail model for abstractive summarization.

2 Objective

The primary goal of the project was to automatically generate concise and coherent summaries for given texts using the abstractive summarization capabilities of the selected model.

3 Methodology

3.1 Model Selection

The google/pegasus-cnn dailymail model was chosen for its specialization in abstractive summarization. This model has been pre-trained on a diverse dataset, including the CNN/Daily Mail dataset, making it suitable for a wide range of summarization tasks.

4 Tokenization

we employed a fine-tuned Pegasus model tokenizer to break down input text into meaningful units. This tailored tokenizer, designed for optimal compatibility with the Pegasus model, enhances the text summarization pipeline's effectiveness, ensuring coherent and meaningful abstractive summarie

4.1 Evaluation Metrics

The model's performance was evaluated using standard ROUGE scores, including "rouge1," "rouge2," "rougeL," and "rougeLsum." These metrics provided a comprehensive assessment of the quality of the generated summaries compared to human-authored reference summaries.

5 Results

5.1 Model Performance Evaluation

For the model evaluation, we are using ROUGE score.

Rouge Names	Result
rouge1	0.015465
rouge2	0.000297
rougeL	0.015503
rougeLsum	0.015514

```
Your max_length is set to 128, but your input_length is only 122. Since this is a summarization task,
where outputs shorter than the input are typically wanted, you might consider decreasing max_length manually,
e.g. summarizer('...', max_length=61)
Dialogue:
Hannah: Hey, do you have Betty's number?
Amanda: Lemme check
Hannah: <file_gif>
Amanda: Sorry, can't find it.
Amanda: Ask Larry
Amanda: He called her last time we were at the park together
Hannah: I don't know him well
Hannah: <file_gif>
Amanda: Don't be shy, he's very nice
Hannah: If you say so..
Hannah: I'd rather you texted him
Amanda: Just text him 😉
Hannah: Urgh.. Alright
Hannah: Bye
Amanda: Bye bye

Reference Summary:
Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.

Model Summary:
Hannah is looking for Betty's number. Amanda can't find it. Larry called Betty last time they were at the park together.
Hannah would rather she text him.
```

Figure 1: Text Summary

5.2 Key Findings

The google/pegasus-cnn dailymail model consistently produced coherent and relevant summaries across a diverse set of input texts, as evidenced by the robust ROUGE scores.

6 Conclusion

n conclusion, the integration of the google/pegasus-cnn dailymail model, fine-tuned and evaluated using ROUGE scores, proved successful in addressing the text summarization task. The abstractive summarization approach enhances the model's ability to distill essential information from input texts, making it a valuable tool for a variety of applications.

7 Project Repository

- Mayank Bhandari
- Ouc-Houang Fogoum Philippe Jacques