CSE508 Information Retrieval

Winter 2024

Assignment-4

Konam Akhil Vamshi

2020513

Preprocessing:

- Remove HTML tags
- Expand contractions
- Remove non-ASCII characters
- Convert to lowercase
- Remove special characters/punctuation
- Replace multiple spaces with a single space

Data preparation:

- Only the text and summary columns from the dataset were used to train the model.

Model Initialization:

- Utilised the GPT2LMHeadModel from the Hugging Face transformers library.
- Customised the tokeniser to treat the end-of-string token as the padding token, accommodating the fixed input size requirement.

Dataset and DataLoader Implementation:

- Defined a custom Dataset class, ReviewSummaryDataset, for handling tokenisation and encoding of text and summary data.
- Split the dataset into training and validation subsets using a 75:25 ratio.

Training Setup:

- Configured training parameters, including a number of epochs, batch size, and weight decay, using the TrainingArguments class.
- Initialised a Trainer object with the model, training arguments, and datasets to handle the training loop.

Model Training:

- Executed the model training over epochs, logging the progress and saving the model upon completion.

Model Inference and Evaluation:

- Implemented a function, generate_summary, to predict summaries using the fine-tuned model, with parameters set for maximum length and beam search.
- Utilized the rouge_score library to compute ROUGE metrics, providing a quantitative measure of model performance against actual summaries.

ROUGE Score Calculation:

- Defined a function to calculate and print ROUGE-1, ROUGE-2, and ROUGE-L scores for generated text against reference summaries, illustrating model effectiveness in capturing pertinent information.

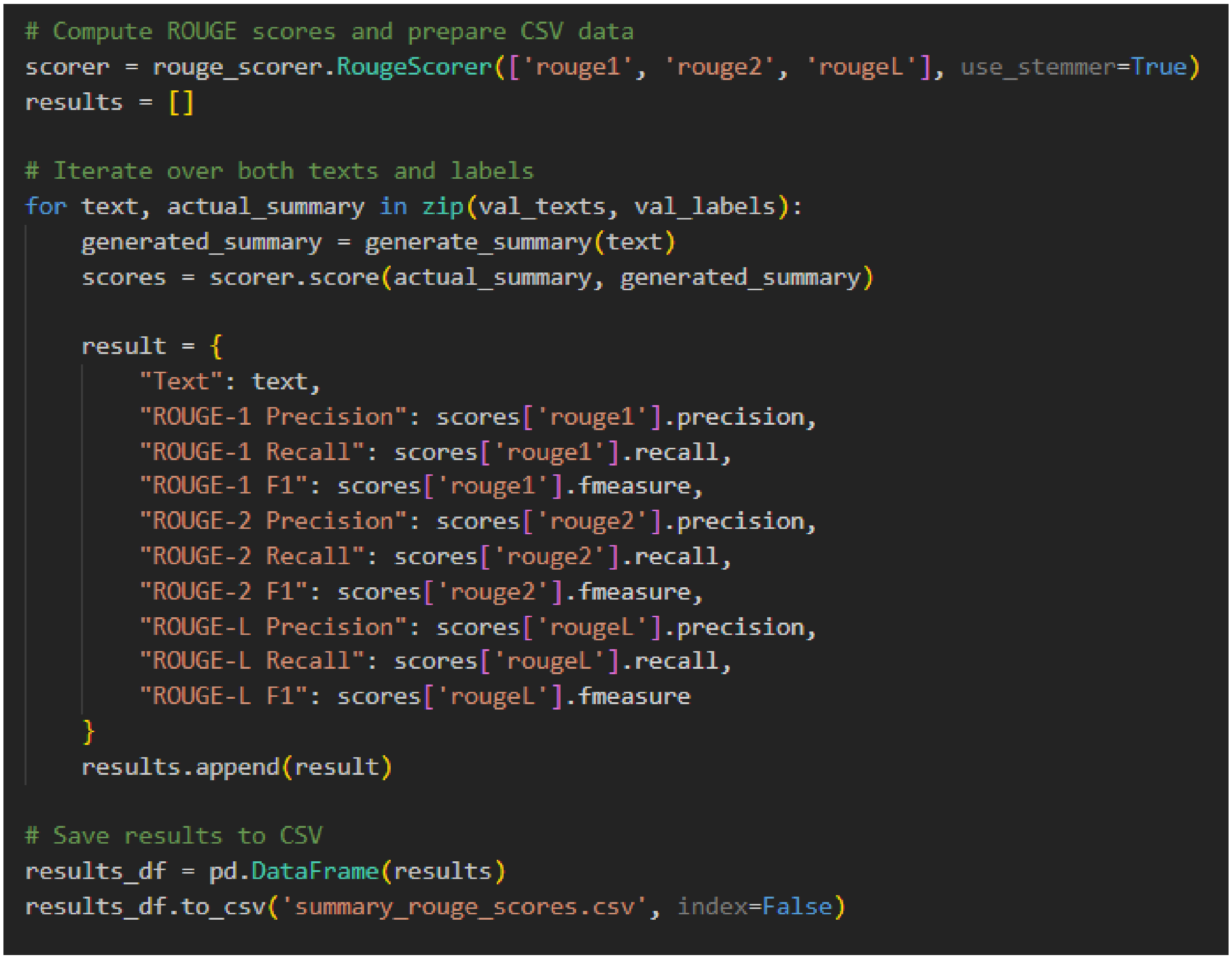
Additional details:

Training Environment: Kaggle T4*2 GPU
Training Time: 14 hours
Epochs: 10
Batch Size: 8
Subset Samples: 50k

(Archived very low results because of computational Issues)
(Kaggle notebook getting crashed if any additional or more parameters were used. Could have achieved better results if the computation power was not an issue)

Evaluation:

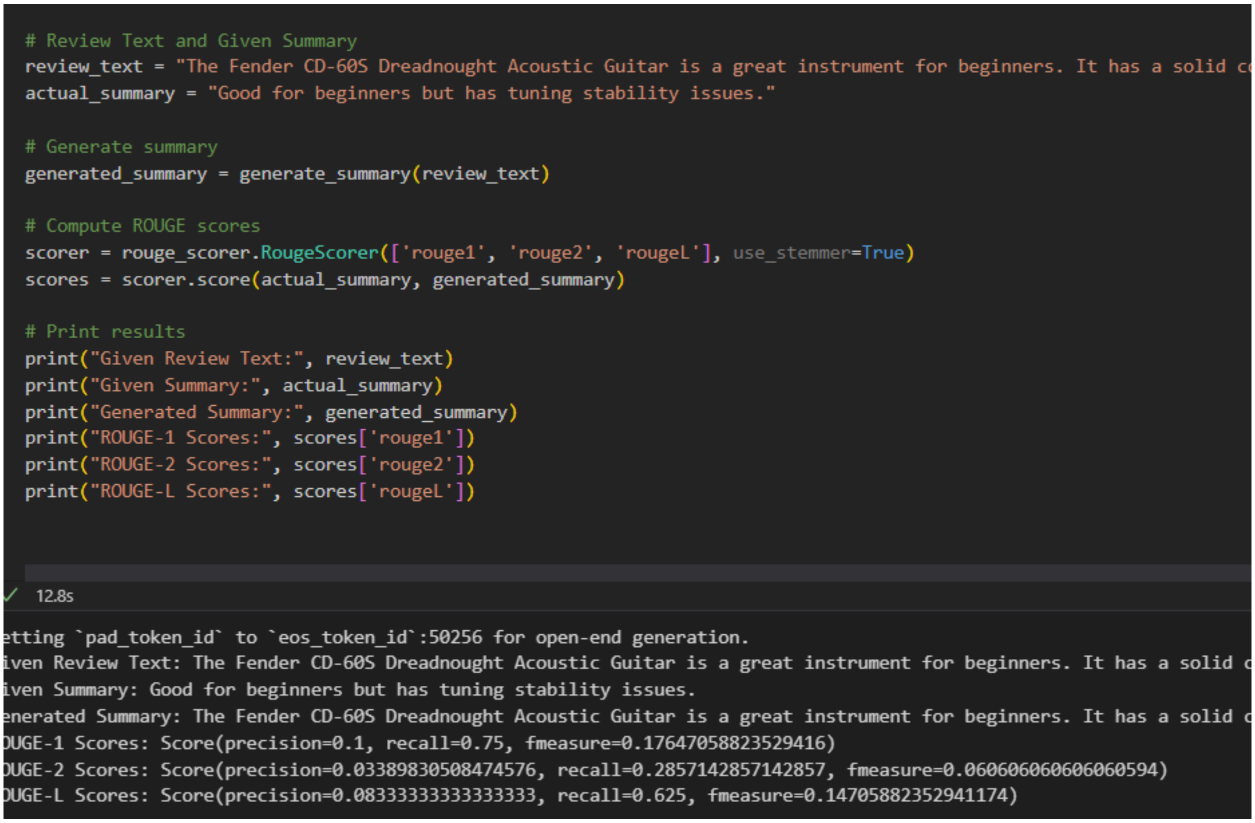
- Stored a CSV file that contains text and corresponding ROUGE scores for my validation set.



```python
# Compute ROUGE scores and prepare CSV data
scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2', 'rougeL'], use_stemmer=True)
results = []

# Iterate over both texts and labels
for text, actual_summary in zip(val_texts, val_labels):
    generated_summary = generate_summary(text)
    scores = scorer.score(actual_summary, generated_summary)

    result = {
        "Text": text,
        "ROUGE-1 Precision": scores['rouge1'].precision,
        "ROUGE-1 Recall": scores['rouge1'].recall,
        "ROUGE-1 F1": scores['rouge1'].fmeasure,
        "ROUGE-2 Precision": scores['rouge2'].precision,
        "ROUGE-2 Recall": scores['rouge2'].recall,
        "ROUGE-2 F1": scores['rouge2'].fmeasure,
        "ROUGE-L Precision": scores['rougeL'].precision,
        "ROUGE-L Recall": scores['rougeL'].recall,
        "ROUGE-L F1": scores['rougeL'].fmeasure
    }
    results.append(result)

# Save results to CSV
results_df = pd.DataFrame(results)
results_df.to_csv('summary_rouge_scores.csv', index=False)
```

Inference:

Generating summary and ROUGE scores for a single Input text and Input summary.



```
# Review Text and Given Summary
review_text = "The Fender CD-60S Dreadnought Acoustic Guitar is a great instrument for beginners. It has a solid c
actual_summary = "Good for beginners but has tuning stability issues."

# Generate summary
generated_summary = generate_summary(review_text)

# Compute ROUGE scores
scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2', 'rougeL'], use_stemmer=True)
scores = scorer.score(actual_summary, generated_summary)

# Print results
print("Given Review Text:", review_text)
print("Given Summary:", actual_summary)
print("Generated Summary:", generated_summary)
print("ROUGE-1 Scores:", scores['rouge1'])
print("ROUGE-2 Scores:", scores['rouge2'])
print("ROUGE-L Scores:", scores['rougeL'])
```

✓ 12.8s

```
etting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
iven Review Text: The Fender CD-60S Dreadnought Acoustic Guitar is a great instrument for beginners. It has a solid c
iven Summary: Good for beginners but has tuning stability issues.
enerated Summary: The Fender CD-60S Dreadnought Acoustic Guitar is a great instrument for beginners. It has a solid c
OUGE-1 Scores: Score(precision=0.1, recall=0.75, fmeasure=0.17647058823529416)
OUGE-2 Scores: Score(precision=0.03389830508474576, recall=0.2857142857142857, fmeasure=0.060606060606060594)
OUGE-L Scores: Score(precision=0.08333333333333333, recall=0.625, fmeasure=0.14705882352941174)
```