Review Summarization using GPT2

ANKUR TIWARI

1. Introduction

In this assignment, we aimed to summarize reviews using the GPT-2 model, focusing on the Amazon Fine Food Reviews dataset. The task involved cleaning and preprocessing the text data, training a GPT-2 model to generate summaries, and evaluating its performance using ROUGE scores.

2. Dataset Preparation

We started by loading the Amazon Fine Food Reviews dataset and selecting a subset for experimentation. The dataset contained reviews along with corresponding summaries. We performed basic cleaning and preprocessing to remove any irrelevant information and prepare the data for training.

3. Model Training

Tokenization and Model Initialization: We used the Hugging Face library to initialize a GPT-2 tokenizer and model. The tokenizer was customized to include special tokens for the beginning and end of the summary.

Dataset Preparation: We implemented a custom dataset class to prepare the data for training. Each review was paired with its corresponding summary, tokenized, and converted into input-output pairs suitable for training the GPT-2 model.

Fine-tuning: The GPT-2 model was fine-tuned on the review dataset using different hyperparameters such as learning rate, batch size, and number of epochs. We experimented with various configurations to optimize the model's performance.

4. Evaluation

ROUGE Scores: After training the model, we evaluated its performance using ROUGE scores on the test set. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate the quality of summaries by comparing them to reference summaries.

Example Evaluation: We provided an example of how ROUGE scores were computed for a generated summary compared to the actual summary. This involved calculating precision, recall, and F1-score for unigram, bigram, and longest common subsequence matches.

5. Results and Discussion

ROUGE Scores: The ROUGE scores indicated the model's performance in terms of precision, recall, and F1-score for different n-gram matches and the longest common subsequence.

Example Discussion: We discussed the significance of the example evaluation results, highlighting areas where the model performed well and areas that could be improved.

Hyperparameter Tuning: We discussed the impact of hyperparameters on the model's performance and provided insights into potential optimizations for future experiments.

6. Conclusion

In conclusion, our assignment demonstrated the effectiveness of using the GPT-2 model for review summarization tasks. By fine-tuning the model on the Amazon Fine Food Reviews dataset and evaluating its performance using ROUGE scores, we were able to assess the quality of the generated summaries. Further research and experimentation could focus on exploring advanced techniques for improving summarization quality and scalability.

Results

Step	Training Loss	Validation Loss
200	4.045100	4.006210
400	3.876900	3.929599
600	3.660800	3.918465
800	3.689700	3.899529
1000	3.656000	3.898855
1200	3.531700	3.898726
1400	3.491900	3.894540

```
TrainOutput(global_step=1410, training_loss=4.008358707833797, metrics={'train_runtime': 1069.2385, 'train_samples_per_secon
d': 10.522, 'train_steps_per_second': 1.319, 'total_flos': 2939535360000000.0, 'train_loss': 4.008358707833797, 'epoch': 5.0})
```

```
ROUGE Scores: [{'rouge-1': {'r': 0.75, 'p': 0.18181818181818182, 'f': 0.29268292368828075}, 'rouge-2': {'r': 0.285714285714285
7, 'p': 0.058823529411764705, 'f': 0.09756097277810835}, 'rouge-l': {'r': 0.625, 'p': 0.15151515151515152, 'f': 0.243902435883
40277}}]
```