Vani Mittal | MT23102 | Assignment 4

Problem Overview:

The task involves generating summaries for textual reviews using the GPT-2 language model and evaluating the quality of the generated summaries using the ROUGE metric. Specifically, the goal is to create concise summaries of product reviews that capture the essential information conveyed in the original text.

Approach:

1. Data Preprocessing:

- The dataset used for this task is the Amazon Fine Food Reviews dataset, which contains reviews of various food products.
- Preprocessing steps include converting text to lowercase, removing numbers, punctuation, HTML tags, non-ASCII characters, and stopwords, and lemmatizing the words.

2. Model Training:

- The GPT-2 model is used for generating summaries. It is pretrained on a large corpus and fine-tuned on the review dataset.
- Training is performed using the AdamW optimizer with linear learning rate scheduling and warmup steps.

3. Summary Generation:

- The trained model is used to generate summaries for the test dataset by providing the review text as input.
- Beam search with a beam width of 5 is used for decoding the generated text.

4. Evaluation:

- The quality of the generated summaries is evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric.
- ROUGE measures the overlap between the generated summary and the reference summary in terms of precision, recall, and F1-score.

Methodologies:

- Data preprocessing involves cleaning the text data to remove noise and irrelevant information, ensuring that the model receives clean input for training and inference.
- The GPT-2 language model is fine-tuned on the review dataset to adapt it to the specific task of summary generation.
- Training is performed in batches using DataLoader to efficiently process the dataset and update the model parameters.
- Evaluation is conducted using the ROUGE metric, which provides insights into the effectiveness of the generated summaries compared to the reference summaries.

Assumptions:

- The quality of the generated summaries depends on various factors such as the size and quality of the training data, model architecture, hyperparameters, and training duration.
- It is assumed that the GPT-2 model, when fine-tuned on the review dataset, can effectively capture the salient features of the reviews and generate coherent summaries.
- The ROUGE metric is used as a proxy for assessing the quality of the summaries, assuming that higher ROUGE scores indicate better summary generation performance.

Results:

- The trained GPT-2 model generates summaries for the test dataset, which are then evaluated using the ROUGE metric.
- The ROUGE scores provide insights into the precision, recall, and F1-score of the generated summaries compared to the reference summaries.
- Based on the ROUGE scores, the effectiveness of the summary generation approach can be assessed, and further improvements or adjustments to the model or training process can be made if necessary.

Overall, the approach involves training a language model to generate summaries of textual reviews and evaluating the quality of the generated summaries using objective metrics like ROUGE. This enables the assessment of the model's performance in capturing the essence of the reviews and producing concise and informative summaries.

Overall, the performance of the summary generation model is mixed, with some generated summaries closely matching the given summaries and others failing to capture the essence of the reviews effectively. The quality of the generated summaries varies based on factors such as the complexity of the review text, the presence of repetitive phrases, and the ability of the model to generate coherent and concise summaries. Further refinement of the model architecture, training process, and hyperparameters may be necessary to improve the overall performance of the summary generation task.