CSE508 Winter 2024 A4 Report

Om Mehroliya
Roll Number: 2021404

April 22, 2024

Abstract

This report outlines the application of a GPT-2 model to summarize text from the Amazon Fine Food Reviews dataset. The project's goal is to automate the generation of concise summaries for reviews, facilitating quicker understanding and analysis of customer feedback.

1 Introduction

The project leverages the transformer-based GPT-2 model, pre-trained by OpenAI, to generate text summaries. This model is fine-tuned on a subset of the Amazon Fine Food Reviews dataset, which comprises over 500,000 reviews but is limited to 10,000 for training purposes to manage computational constraints.

2 Methodology

2.1 Data Preprocessing

Textual data often contains HTML tags, special characters, and case inconsistencies, which were cleaned and standardized. The GPT-2 tokenizer then processes this cleaned data, encoding the texts and summaries into tensors that the model can interpret.

2.2 Model Training

The GPT-2 model, configured with custom hyperparameters, is trained using the AdamW optimizer. Training involves minimizing the loss calculated between the predicted and actual summaries, adjusting model weights to better fit the dataset.

3 Assumptions

- The quality of the summaries is dependent on the initial weights of the pre-trained model and the representativeness of the training subset.

- Reviews and summaries are assumed to be truthful and accurately reflect the product.

4 Results

The model's performance was evaluated using the ROUGE metric, which assesses the overlap between the generated summaries and human-written reference summaries. While specific scores are pending final evaluation, preliminary results suggest the model can effectively capture key information.

5 Code Implementation Review

5.1 Approach

The implemented script automates the summarization of product reviews using a pre-trained GPT-2 model. It includes steps for data loading, preprocessing, summary generation, and performance evaluation.

5.2 Methodologies

The script processes data with text cleaning techniques, adjusts for device compatibility, and uses advanced parameters in text generation to improve the quality of output. It uses DataLoader for efficient batch processing during training.

5.3 Assumptions

The success of the text summarization heavily relies on the assumptions that the processed data contains valid mappings from reviews to summaries and that the summaries provided are a good representation of the content.

5.4 Results

Preliminary testing on a subset of the data shows promising results with the generation of coherent and contextually appropriate summaries. Detailed metrics such as ROUGE scores are computed to quantitatively evaluate the quality of the generated summaries, showing the model's potential to generalize from training to unseen data.

6 Conclusion

Fine-tuning GPT-2 on the Amazon Fine Food Reviews dataset demonstrates potential for automated summary generation in e-commerce, providing a foundation for further research into text summarization applications.