CSE508 Information Retrieval

Winter 2024

Assignment-4

Name: Aniketh 15th April Roll No: 2020360

1. Project Overview

Develop a model that can generate concise summaries of customer reviews, aiding in quicker comprehension of the content. The model should reliably reflect the main points of the reviews while maintaining the context and sentiment expressed by the original text.

2. Dataset

Source: The dataset used in this project is the Amazon Fine Food Reviews, sourced from the Stanford Network Analysis Project. It has over 500,000 reviews of its food products from Amazon.

Structure: The dataset's 'Text' and 'Summary' columns were primarily utilized for this project. Each entry in 'Text' contains the body of a review, and 'Summary' provides a concise headline of the review written by the customer.

Preprocessing:

- Cleaning: Initial steps included removing HTML tags, special characters, and numbers to standardize the text.
- Tokenization: The GPT-2 tokenizer was employed to split the text into manageable tokens.
- Punctuation Removal and Lowercasing: These steps were initially tested but were eventually excluded from the preprocessing pipeline as they did not significantly impact model performance.
- NaN Handling and Reindexing: Non-valuable entries were dropped, and the DataFrame was reindexed to ensure data integrity.

3. Model

Model Selection: The project utilized the pre-trained GPT-2 model from Hugging Face's Transformers library, renowned for its effectiveness in generating coherent and contextually relevant text.

Initialization: The GPT-2 model and tokenizer were initialized with their default configurations as provided by Hugging Face.

Fine-tuning:

- The model was fine-tuned specifically on the Amazon reviews dataset to adapt its capabilities to the summarization task.
- Adjustments were made to the training loop to accommodate the particularities of managing input and output sequence formats, focusing on generating succinct summaries from detailed reviews.

4. Training

Data Split: The dataset was divided into training (75%) and testing (25%) sets, with a focus on a subset of 10,000 reviews, selected based on data cleanliness and diversity.

Training Environment: Training was conducted on a high-performance computing environment, leveraging GPUs for accelerated computing. Specific software included Python 3.8 and PyTorch with the Transformers library.

Hyperparameters:

- Learning Rate: Tested various learning rates, with 3e-4 and 2e-5 providing the best results. Finally trained with 5e-5 for first 10 epochs and then slowly reduced learning rate for better convergence. Also added weight decay parameter later.
- Batch Size: Experimented with sizes of 8, 16, and 32 to balance between computational efficiency and memory constraints. Used the largest batch size which could be accommodated in memory.
- Epochs: Models were trained for 3, 5, and 10 epochs to determine the optimal stopping point based on validation loss and performance for small subset of data to understand convergence. But resorted to over 50 epochs on a subset data of 10k reviews.

5. Results and Discussion

Performance Analysis: The model exhibited strong capabilities in producing accurate and pertinent summaries, particularly excelling with longer text inputs where more contextual cues were available.

Summary Examples:

- Original Review: "I have bought several of the Vitality canned dog food products and have found them all to be of good quality."
- Generated Summary: "High-quality dog food."

Performance on Different Data Sets:

- Test Set Results:
    - ROUGE-1: 0.65
    - ROUGE-2: 0.58
    - ROUGE-L: 0.65
- Subset of 20 Examples:
    - ROUGE-1: 0.67
    - ROUGE-2: 0.61
    - ROUGE-L: 0.67

Sample Outputs from the Subset:

- Correct Summaries: ['Good Quality Dog Food', 'Not as Advertised', '"Delight" says it all', 'Cough Medicine', 'Great taffy', 'Nice Taffy', 'Great! Just as good as the expensive brands!', 'Wonderful, tasty taffy', 'Yay Barley', 'Healthy Dog Food']
- Predicted Summaries: ['Good Quality Dog Food Quality Food', 'Not as Advertised any.', '', 'Cough Medicine Sassr', 'Great taffy a.', '', 'Great! Just as good as the expensive brands! Very!', 'Wonderful, tasty taffy fullied', 'Yay Barley for', 'Healthy Dog Food Good Dog']

Qualitative Analysis and Improvement Suggestions:

The model struggles with consistency in generating complete and contextually accurate summaries. Errors include missing text, irrelevant additions, and slight alterations of the intended meaning. These issues could be due to insufficient model training on diverse data or lack of understanding of nuanced language. Improvements can be achieved by:

- Enhancing the training dataset with a broader range of text complexities.
- Implementing advanced natural language processing techniques to better capture the essence of the text.

- Fine-tuning the model parameters to better handle edge cases and subtleties in the language.