Report: Review Summarization using GPT-2

Objective:

The goal was to fine-tune a GPT-2 model using the Amazon Fine Food Reviews dataset to generate concise summaries of reviews. This included preprocessing the data, training the model, and evaluating it using ROUGE scores.

---

1. Data Preprocessing

Dataset:

- **Source:** Amazon Fine Food Reviews dataset.
- **Columns Used:**
  - `Text` (review text)
  - `Summary` (provided summary).
- **Initial Dataset Size:** 568,401 entries.

Cleaning Steps:

1. **Removing duplicates:** Used `drop_duplicates()`.
2. **Handling null values:** Removed rows with `NaN` in `Text` or `Summary`.
3. **Preprocessing Functions:**
   - Removed HTML tags using `BeautifulSoup`.
   - Converted text to lowercase.
   - Removed punctuation and numbers.
   - Stripped whitespace.

Final Preprocessed Dataset:

- Saved as a pickle file (`Preprocessed_Reviews.pkl`).

---

2. Model Training

Model and Tokenizer:

- **Model:** GPT-2 (via Hugging Face).
- **Tokenizer:** GPT-2 tokenizer with EOS token added.

Dataset Preparation:

- Combined `Text` and `Summary` into a single column using "TL;DR" as a separator.
- Split into **training** (75%) and **testing** (25%) sets.
- Average sequence length: 84 words. Maximum length was set to 100 tokens.

Custom Dataset Class:

A PyTorch dataset class was created to:

1. Tokenize and encode input data.
2. Pad or truncate reviews to ensure uniform input size.

Training Configuration:

- **Batch Size:** 32.
- **Learning Rate:** 3e-4 (AdamW optimizer).
- **Epochs:** 1 (for demonstration).
- **Hardware:** GPU (if available).

Observations:

- Initial loss: ~8.23
- Loss reduced significantly to ~2.38 during training.

Model Saving:

- Model and tokenizer saved to `fine_tune_gpt2_model1`.

---

3. Evaluation

Inference:

- Custom `model_infer` function:
  - Generates summaries using top-k sampling for diversity.
  - Stops when max length or EOS token is reached.

Metrics:

- **ROUGE Scores:**
  - Precision, Recall, F1-Score for ROUGE-1, ROUGE-2, and ROUGE-L.
  - Example results:
    - **ROUGE-1:** Precision: 0.75, Recall: 0.80, F1-Score: 0.77.
    - **ROUGE-2:** Precision: 0.50, Recall: 0.67, F1-Score: 0.57.
    - **ROUGE-L:** Precision: 0.67, Recall: 0.75, F1-Score: 0.71.

User Input Testing:

For user-provided reviews, the model generates summaries, and ROUGE scores are computed. Example:

- **Review:** `"i'm italian and i lived in italy for years. i used to buy these cookies for my everyday breakfast with an italian espresso."`
- **Actual Summary:** `"great cookies"`

- **Generated Summary:** `"iced mocha cookies"`
- **ROUGE-1 Score:** Precision: 0.33, Recall: 0.50, F1-Score: 0.40.

---

4. Challenges and Recommendations

Challenges:

- **Computational Resources:** Limited fine-tuning due to GPU/CPU constraints.
- **Dataset Size:** Large dataset led to subsampling.
- **Evaluation Quality:** ROUGE scores showed variability depending on review length and complexity.

Recommendations:

1. **Hyperparameter Tuning:** Experiment with batch sizes, learning rates, and epochs for better performance.
2. **Increase Epochs:** Train for multiple epochs to improve generalization.
3. **Data Augmentation:** Use paraphrased summaries to enhance training diversity.
4. **Evaluation Metrics:** Complement ROUGE scores with human evaluations.

---

5. Conclusion

The project demonstrated the feasibility of fine-tuning GPT-2 for review summarization. Despite computational limitations, the model achieved reasonable performance. Further optimization can yield more accurate summaries and better evaluation metrics.