Report on Generative Chatbot with the Cornell Movie Dialog Corpus

Using DialoGPT & GPT-4

By Anova Youngers

Abstract

This report details the development of a generative chatbot using the Cornell Movie Dialog Corpus, employing both **DialoGPT** (small and large) and **GPT-4** models. It outlines key challenges encountered during the project, rationale for model choices, evaluation metrics, and future improvements. The chatbot focused on handling informal movie dialogues, achieving significant conversational fluency. The project compares both models to provide insights into their relative performances and scalability.

Introduction

The objective of this project was to develop a chatbot capable of engaging in human-like conversations using movie dialogues. The Cornell Movie Dialog Corpus provided a dataset rich with informal human conversations, making it ideal for training dialogue models. I implemented two generative models:

- **DialoGPT (Small and Large)**: Fine-tuned models designed for conversational AI.

- **GPT-4**: A state-of-the-art model from OpenAI, used for comparison.

The project aimed to evaluate these models' conversational capabilities and compare their performance, fluency, and scalability.

Challenges Faced and Solutions Implemented

1. **Dataset Preprocessing**:
    - **Challenge**: Preprocessing the Cornell Movie Dialog Corpus definitely had some challenges such as long conversations, redundant stopwords, and metadata like character names that distorted dialogue flow quite a bit and needed to be cleaned.
    - **Solution**: For the DialoGPT model preparation I performed tokenization, padding, and truncation to standardize conversation lengths. I did decide to keep stopwords as I thought it important to maintain natural conversation flow. Metadata like character names and email addresses were removed to clean the dataset. For the GPT-4 model I did not incorporate tokenization as that model specifically thrives off of raw text.
2. **Model Training**:
    - **Challenge**: Running the training on Google Colab with a T4 GPU led to exorbitantly long runtimes, especially after encountering compute limitations with an A100 GPU. Then when running GPT-4 I hit training times around 300 min.

   - **Solution**: I had to switch programs entirely and ended up running the entire project on my VS Studio Code platform's local host. Regarding the exorbitant run time with GPT-4, I realized rather late, that I had forgotten to batch the data before feeding it into the model. So the model was looping back through each conversation separately instead of in faster batches.I
3. **Handling Conversational Context**:
   - **Challenge**: The DialoGPT model initially struggled to maintain conversational context across multiple turns.
   - **Solution**: I adjusted hyperparameters, including learning rate and sequence length, and used larger context windows during inference to ensure that the chatbot retained more relevant information from prior interactions.

Model Architecture and Rationale

DialoGPT (Small and Large)

DialoGPT is a transformer-based model fine-tuned specifically on conversational data:

- **Transformer-based architecture**: DialoGPT's self-attention mechanism allows it to understand long-term dependencies and generate relevant responses.
- **Pretrained on dialogue**: Its pretraining on dialogue-specific data allows faster convergence with minimal fine-tuning.

- **Contextual Understanding**: The model captures multi-turn dialogue flow, enabling it to handle informal, dynamic conversations typical in movie scripts.

Training Process:

- **DialoGPT-Small**: Trained using tokenization and padding, it showed moderate performance, with improvements in coherence after hyperparameter adjustments.
- **DialoGPT-Large**: A larger version of the model, fine-tuned using **gradient accumulation** to handle memory limitations. Gradient accumulation enabled training with smaller batches, which prevented out-of-memory errors and stabilized the learning process.

GPT-4

- **Advanced Contextual Understanding**: **GPT-4** offers enhanced comprehension of long, complex sequences, making it ideal for extended conversations.
- **Raw Text Processing**: Tokenization was not required for this model, which simplifies preprocessing. Truncation was applied to limit input lengths to 512 tokens for efficiency.
- **Comparison to DialoGPT**: GPT-4 offers richer, more accurate responses due to its larger architecture, but requires more computational resources.

Evaluation Results

DialoGPT-Small

- **Final Loss**: 1.19
- **Perplexity**: 2.46
- **BLEU Score**: 0.777
- **ROUGE-L**: 0.916

- **Accuracy**: 0.758

DialoGPT-Large

- **Final Loss**: 0.896 after training
- **BLEU Score**: Pending
- **ROUGE-L**: Pending
- **Accuracy**: Pending

GPT-4

- **BLEU Score**: 0.107 (initial evaluation)
- **ROUGE-L**: 0.423
- **Truncation**: Applied to limit input to 512 tokens.

Fine-tuning for GPT-4

- **Tokenization**: Not applied due to GPT-4's ability to process raw text, which streamlines preprocessing.
- **Truncation**: Implemented to prevent input data from exceeding token limits.
- **Batch Processing**: Introduced to optimize input handling and reduce runtime.

Future Improvements and Scalability

Contextual Memory Enhancements

- **Solution**: Future iterations could include memory networks or attention mechanisms to better retain context across longer conversation histories.

Domain-Specific Fine-Tuning

- **Solution**: Fine-tuning the models on specific datasets (e.g., customer service or education dialogues) would improve performance in specialized applications.

Scalability

- **Solution**: Scaling this model to larger datasets or using architectures like GPT-4 for deployment in real-world applications is feasible. For deployment at scale, cloud infrastructure (AWS Lambda, Google Cloud) can be leveraged to ensure real-time interaction.

Conclusion

This project successfully developed a generative chatbot using both **DialoGPT** and **GPT-4** models. The models performed well in handling conversational flow and informal movie dialogues, though some challenges related to runtime and conversational context were encountered. **DialoGPT** offered solid conversational performance, while **GPT-4** showed promise for more detailed and context-aware conversations.

Future work will focus on improving the model's memory capabilities, domain-specific training, and scaling to more extensive real-world deployments.

References

1. **Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020).** Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165
2. **Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019).** Language models are unsupervised multitask learners. *OpenAI*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
3. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017).** Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762 .
4. **Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... & Dolan, B. (2020).** Dialogpt: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2706-2715. https://doi.org/10.18653/v1/2020.acl-main.237
5. **Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020).** Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
6. **Jurafsky, D., & Martin, J. H. (2021).** *Speech and language processing* (3rd ed.). Pearson. https://web.stanford.edu/~jurafsky/slp3/