Project Report

BioGPT Based Medical Chat Bot: Design and Technical Explanation

Group-10

1. Naman Pathak – 100899717
2. Shobhit Panwar – 100904075
3. Somya Sachan – 100901887
4. Paramvir Singh Bali – 100843502

Introduction

Pre-trained language models have demonstrated remarkable success in various natural language processing tasks. In the biomedical domain, while models like BERT and its variants have been extensively explored for discriminative tasks, such as BioBERT and PubMedBERT, there is a need for domain-specific generative models. In this context, we introduce BioGPT, a domain-specific generative Transformer language model pre-trained on a large-scale biomedical literature corpus. This report presents the application of Microsoft's BioGPT Large model in building a medical chat bot for assisting patients and doctors in medical conversations.

Data Collection

The dataset used for training and evaluating the medical chat bot consists of three components:

1. ****HealthCareMagic-100k**:** This dataset includes 100,000 real conversations between patients and doctors sourced from HealthCareMagic.com.

2. ****icliniq-10k**:** Comprising of 10,000 real conversations between patients and doctors, this dataset is collected from icliniq.com.

3. ****GenMedGPT-5k and disease database**:** This dataset includes 5,000 generated conversations between patients and physicians using ChatGPT, along with a disease database.

The combination of real conversations and generated content provides a diverse set of examples to train and test the medical chat bot.

Methodology

The medical chat bot is built using the BioGPT Large model, which is fine-tuned for medical conversations using the ChatDoctor-200K dataset. The following steps outline the methodology:

1. **Model Initialization:** The BioGPT Large model is initialized using the model identifier "Narrativaai/BioGPT-Large-finetuned-chatdoctor."

2. **Tokenization:** The AutoTokenizer is used to tokenize input text, preparing it for model input.

3. **Generation Configuration:** A GenerationConfig is defined, specifying parameters like temperature, top-p, top-k, and num_beams. These parameters influence the creativity and diversity of the generated responses.

4. **Generation:** The model generates responses based on the provided prompt using the configured parameters.

5. **Post-Processing:** The generated response is post-processed to extract the actual response content from the model's output.

Results: The model achieves a **44.98%** F1 score on the BC5CDR end-to-end relation extraction task.

Code Components

Model and Tokenizer Setup:

- The model used is Narrativaai/BioGPT-Large-finetuned-chatdoctor, which is designed for medical question answering.
- The tokenizer is based on microsoft/BioGPT-Large, specialized for medical text.

answer_question Function:

- This function generates a response given a prompt/question.
- It takes model, tokenizer, prompt, and various generation parameters as inputs.
- The function tokenizes the input prompt, sets up generation parameters, and generates a response using the model.
- The response is decoded using the tokenizer and returned.

generate_response Function:

- This function prepares an example prompt with the given question, based on a predefined template.
- It calls the answer_question function with the example prompt and returns the generated response.

Gradio Interface (iface):

- Gradio is used to create an interactive web interface for the model.
- The interface takes user input in the form of a medical question and provides the generated response.
- Examples of input questions are provided for users to understand the interface.

Technical Explanation

1. The model and tokenizer are initialized with their respective pre-trained configurations.
2. The answer_question function takes the question prompt and converts it into model-readable format (input IDs and attention mask).
3. Generation parameters like temperature, top-k, and num-beams are set up using the GenerationConfig.
4. The model generates a response based on the input and generation parameters.
5. The response sequence is decoded using the tokenizer, and the actual response is extracted.
6. The generate_response function prepares a prompt with the input question and passes it to answer_question.
7. The Gradio interface is set up to use the generate_response function for generating responses.
8. Users can input medical questions through the interface and get AI-generated responses.

Evaluation Metrics

1. **Perplexity:** Measures how well the model predicts a sample. Lower perplexity indicates better performance.
2. **BLEU Score:** Compares generated text to reference text, evaluating fluency and adequacy.
3. **ROUGE Score:** Measures overlap between generated text and reference text.
4. **Human Evaluation:** Actual medical professionals can assess the accuracy and relevance of generated responses.

Limitations

1. **Bias and Errors:** The model's responses might contain biases or errors, impacting the quality of medical advice.
2. **Lack of Context:** The model might misunderstand context or fail to consider important information in complex medical questions.
3. **Over-Reliance on Training Data:** The model's responses are based on patterns in training data and might not reflect the latest medical knowledge.
4. **Legal and Ethical Concerns:** AI-generated medical advice should not replace professional medical consultations due to legal and ethical considerations.

Ethical Considerations

1. **Professional Oversight:** The AI-generated responses should be used as supplementary information, not a substitute for professional medical advice.
2. **Transparency:** Users interacting with AI should be aware that they are receiving responses from a machine, not a human.
3. **Data Privacy:** User input should be treated with care to ensure patient privacy and compliance with data protection laws.
4. **Bias Mitigation:** Efforts should be made to minimize bias in AI-generated medical advice to ensure equitable responses for all patients.
5. **Continual Monitoring and Updates:** The model's responses should be regularly reviewed and updated to incorporate the latest medical knowledge.

Conclusion

The product is an AI-based medical question responder using a pre-trained language model fine-tuned for medical questions. While the system offers an interactive interface, it comes with limitations related to bias, context understanding, and data accuracy. Ethical considerations include transparency, professional oversight, and privacy. Regular monitoring and updates are essential to ensure safe and accurate medical assistance.