Report on Training a Chatbot with "Guyton and Hall Textbook of Medical Physiology

Muhammad Mubashir Hassan

I21-1764

DS(M)

Introduction

This report details the process of training a chatbot to answer questions using the content from the "Guyton and Hall Textbook of Medical Physiology." The objective was to develop a sophisticated chatbot capable of providing accurate and detailed responses to queries related to medical physiology.

Model Selection and Challenges

Initially, a transformer-based BERT question-answering (Q&A) model was employed. However, this model proved to be resource-intensive and inefficient, requiring substantial computational power and time. As a result, the focus shifted towards exploring more efficient models that could handle complex medical queries effectively.

Next, the GPT-2 model was tested. While it offered some basic assistance with simpler prompts, it lacked the depth required for more complex and in-depth medical questions from the textbook.

Finally, the decision was made to implement Retrieval-Augmented Generation (RAG) models—specifically Llama 3 and Mistral. These models are designed to combine the benefits of retrieval-based systems with generation capabilities, enabling the chatbot to retrieve relevant information from a large corpus (such as the textbook) and then generate detailed responses based on that information.

Training Process

The training process involved several key steps:

1. **Data Preparation**: The textbook content was extracted and prepared for training. This involved using tools to load, preprocess, and split the content into manageable chunks suitable for training the chatbot.
2. **Embedding Extraction**: Embeddings (representations of text) were generated from the textual data. This step is crucial for encoding the textual information into a format that the model can understand and use for retrieval and generation tasks.
3. **Vector Database Construction**: The text chunks, along with their corresponding embeddings, were stored in a vector database. This database allows for efficient retrieval of relevant text based on user queries.
4. **RAG Model Training**: The selected RAG model (in this case, Llama 3 or Mistral) was trained using the vector database. The model learns to retrieve relevant information from the database and generate responses based on the retrieved context.

Implementation Details

The implementation involved setting up a pipeline for the chatbot:

- **Retrieval Component**: This component retrieves relevant information from the vector database based on user queries. It uses the RAG model's retrieval capabilities to identify relevant text passages.
- **Generation Component**: Once relevant context is retrieved, the RAG model generates responses to user queries based on this context. The model is fine-tuned to produce accurate and contextually appropriate answers.

User Interaction

To interact with the chatbot:

1. Users input their queries or questions related to medical physiology topics.
2. The chatbot's retrieval component searches the vector database for relevant context.
3. The RAG model generates responses based on the retrieved context and the user's query.
4. The chatbot then presents the generated response to the user.

Conclusion

In conclusion, training a chatbot to interact with the "Guyton and Hall Textbook of Medical Physiology" involved selecting and training suitable models capable of handling complex medical queries. The use of RAG models proved effective in combining retrieval and generation tasks, allowing the chatbot to provide informative and contextually relevant responses to user queries. Future work could focus on further refining the model's performance and integrating additional features to enhance user interaction and usability.