AI Based Healthcare Chatbot System

Tapan Mahata, Diganta Diasi, Tarun Kumar

August 12, 2024

GitHub Repository
Supervisor: Dr. Chiranjib Sur

Abstract

This project focuses on developing a medical chatbot by fine-tuning the llama2-7b language model with a custom dataset sourced from the Gale Encyclopedia on alternative medicine. Leveraging Hugging Face's Transformer library, LoRA (Low Rank Adaptation) and QLoRA (Quantized Low Rank Adaptation) techniques are employed for parameter-efficient fine-tuning. LoRA enhances the fine-tuning process by optimizing smaller matrices that approximate the weight matrix of the pretrained model, while QLoRA further improves memory efficiency by utilizing quantized 4-bit weights. The Hugging Face PEFT (Parameter Efficient Fine-Tuning) and TRL (Transformer Reinforcement Learning) libraries are utilized for implementation, facilitating seamless integration and efficient training. The project aims to create a coherent and effective medical chatbot capable of providing accurate information and assistance in the realm of alternative medicine.

Keywords: Artificial Intelligence, NLP, Hugging Face, Fine Tune , Inferences, LLM, Llama 2, Lora, QLora

1 Introduction

Recent advancements in Natural Language Processing (NLP) have enabled programs to converse with users in a manner similar to human interaction. This technological progress has led to the widespread adoption of chatbots across various industries, significantly enhancing user experiences. In healthcare, AI-driven chatbots have emerged as powerful tools, revolutionizing the way individuals access medical information and support.

Health and fitness chatbots, such as Health Tap's Messenger bot, exemplify the potential of NLP-driven solutions to provide convenient and reliable medical advice to users. The integration of NLP into healthcare services promises transformative impacts on health management, offering personalized and accessible resources to individuals seeking medical guidance.

2 Problem Statement

The AI-based Healthcare Chatbot System addresses several challenges prevalent in accessing healthcare information:

Information Accuracy and Accessibility: Many individuals struggle to find trustworthy health information online due to medical jargon and complex language. The chatbot aims to provide a user-friendly platform with accurate and clear medical explanations, overcoming barriers to understanding. Interactive Healthcare Resources: Traditional methods of accessing healthcare information can be passive, lacking interactivity. The chatbot offers an interactive platform where users can ask questions in natural language and receive clear, concise answers, promoting active engagement with healthcare resources. Limitations of Current Chatbots: Existing healthcare chatbots may have limitations in information retrieval and natural language understanding. The AI-based chatbot leverages large language models to overcome these limitations, providing a comprehensive and informative user experience.

3 Related Work

Research has shown the potential of large language models (LLMs) in providing mental health support and assisting patients with chronic conditions. Studies have explored the development of chatbots using

LLMs to assess user symptoms, offer medication reminders, and facilitate communication with healthcare providers.

For instance, "The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support" investigates how people use LLMs for mental health support, highlighting the implications of this use. Additionally, LLMs have been utilized to develop chatbots that provide educational resources, self-care options, and recommendations for further evaluation by medical professionals.

4 What is Llama 2?

Llama 2 is a family of pre-trained and fine-tuned large language models (LLMs) released by Meta AI in 2023. Released free of charge for research and commercial use, Llama 2 AI models are capable of a variety of natural language processing (NLP) tasks, from text generation to programming code.

The Llama 2 model family, offered as both base foundation models and fine-tuned "chat" models, serves as the successor to the original LLaMa 1 models, which were released in 2022 under a noncommercial license granting access on a case-by-case basis exclusively to research institutions. Unlike their predecessors, Llama 2 models are available free of charge for both AI research and commercial use.

In addition to making its code and model weights freely available, the Llama project has focused advancing the performance capabilities of smaller models, rather than through increasing parameter count. Whereas most prominent closed-source models have hundreds of billions of parameters, Llama 2 models are offered with seven billion (7B), 13 billion (13B) or 70 billion parameters (70B).

5 Llama 2 chat models

Llama-2-chat models are fine-tuned for dialogue-driven use cases, similar to the specific GPT model versions used in ChatGPT.

Supervised fine tuning (SFT) was used to prime the pre-trained Llama 2 base model to generate responses in the format expected by users in a chatbot or virtual agent setting. In a series of supervised learning tasks, labeled pairs of dialogue-style exchanges, annotated as (prompt, response), are used to train the model to minimize the divergence between its own response for a given prompt and the example response provided by the labeled data. The model thus learns, for example, that the proper response to a prompt of "teach me to bake cookies" is to provide actual instructions to bake cookies, rather than merely complete the sentence.

Rather than using millions of labeled examples, the paper states that results were improved by using "fewer but higher-quality examples," noting that Meta AI collected 27,540 annotated samples.

Following SFT, Meta used reinforcement learning with human feedback (RLHF) to further align the chat models' behavior with human preferences and instructions. In RLHF, direct human feedback is used to train a "reward model" to learn patterns of the kind of responses humans prefer. By translating the reward model's predictions (regarding whether a given response would be preferred by humans) into a scalar reward signal, the reward model is then used to further train Llama-2-chat via reinforcement learning.

6 Llama 2 Prompt

In case of Llama 2, the following prompt template is used for the chat models System Prompt (optional) to guide the model

User prompt (required) to give the instruction

Model Answer (required)

¡s¿[INST] ¡¡SYS¿¿ System prompt¡¡/SYS¿¿

User prompt [/INST] Model answer ¡/s¿

7 Dataset Description

The Gale Encyclopedia of Alternative Medicine serves as a comprehensive resource covering various aspects of complementary and alternative medicine. It offers information on prevalent conditions and diseases, therapies, herbs/plants, foods, and individuals in the field. The encyclopedia provides a vast

amount of data points on alternative medical practices, making it suitable for training the chatbot to understand a broad spectrum of user queries related to alternative medicine.

By leveraging this dataset, the AI-based Healthcare Chatbot System can offer users accurate and relevant information on alternative medical practices, empowering them to make informed decisions about their health and well-being.

Gale Encyclopedia on alternative medicine

8 Dataset Preparation

The dataset preparation procedure begins with the implementation of a Python function called `format_text`, designed to merge instructions and responses into a structured format adhering to the Llama-2 prompt specifications. This function encapsulates the instruction within `[INST]` and `[/INST]` tags before appending it to the response. Paths for both the input CSV file containing the instruction-response pairs (`input_csv_file`) and the output TXT file (`output_txt_file`) are defined to facilitate data handling. Using the `csv.DictReader` method, the code iterates through each row of the CSV file, extracting instructions and responses, formatting them via the `format_text` function, and writing the resulting text to the output TXT file. Upon completion, a confirmation message signals the successful conversion. Subsequently, the formatted TXT file is read, and its contents are loaded into a dataset utilizing the `Dataset.from_dict` method from the `datasets` library. This process ensures that the instruction-response pairs are appropriately formatted and poised for integration into the chatbot training pipeline, aligned with the specified Llama-2 prompt structure.

9 Fine-Tuning, LoRA and QLoRA

9.1 Fine-Tuning

In the realm of language models, fine tuning an existing language model to perform a specific task on specific data is a common practice. This involves adding a task-specific head, if necessary, and updating the weights of the neural network through backpropagation during the training process. It is important to note the distinction between this finetuning process and training from scratch. In the latter scenario, the model's weights are randomly initialized, while in finetuning, the weights are already optimized to a certain extent during the pre-training phase. The decision of which weights to optimize or update, and which ones to keep frozen, depends on the chosen technique. Full finetuning involves optimizing or training all layers of the neural network. While this approach typically yields the best results, it is also the most resource-intensive and time-consuming. Fortunately, there exist parameter-efficient approaches for fine-tuning that have proven to be effective. Although most such approaches have yielded less performance, Low Rank Adaptation (LoRA) has bucked this trend by even outperforming full finetuning in some cases, as a consequence of avoiding catastrophic forgetting (a phenomenon which occurs when the knowledge of the pretrained model is lost during the fine-tuning process).

9.2 LoRA

LoRA is an improved finetuning method where instead of finetuning all the weights that constitute the weight matrix of the pre-trained large language model, two smaller matrices that approximate this larger matrix are fine-tuned. These matrices constitute the LoRA adapter. This fine-tuned adapter is then loaded to the pretrained model and used for inference.

9.3 QLoRA

QLoRA is an even more memory efficient version of LoRA where the pretrained model is loaded to GPU memory as quantized 4-bit weights (compared to 8-bits in the case of LoRA), while preserving similar effectiveness to LoRA. Probing this method, comparing the two methods when necessary, and figuring out the best combination of QLoRA hyperparameters to achieve optimal performance with the quickest training time will be the focus here. LoRA is implemented in the Hugging Face Parameter Efficient Fine-Tuning (PEFT) library, offering ease of use and QLoRA can be leveraged by using bitsandbytes and PEFT together. HuggingFace Transformer Reinforcement Learning (TRL) library offers a convenient trainer for supervised finetuning with seamless integration for LoRA. These three libraries will provide

the necessary tools to finetune the chosen pretrained model to generate coherent and convincing product descriptions once prompted with an instruction indicating the desired attributes.

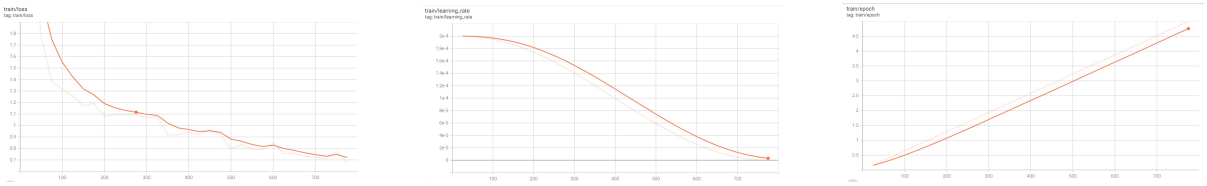
10 Training Procedure

The training procedure for fine-tuning the Llama 2–7b model employs parameter-efficient techniques to accommodate the limited resources, particularly the 15GB Graphics Card available in Google Colab. Given the constraints, full fine-tuning is not feasible, necessitating the utilization of techniques like LoRA or QLoRA to optimize model performance. In this instance, QLoRA is implemented to drastically reduce VRAM usage by fine-tuning the model in 4-bit precision.

The process begins with loading the tokenizer and model with QLoRA configuration, ensuring compatibility with the GPU's capabilities, especially with bfloat16 support if available. The base model is then loaded using the AutoModelForCausalLM function, incorporating the specified quantization configuration and device mapping. Additionally, the LoRA configuration is applied to optimize fine-tuning for the $\mathrm{CAUSAL}_L Mtasktype$.

Training parameters are set using TrainingArguments, defining key aspects such as output directory, number of epochs, batch size, learning rate, and optimization techniques. Supervised fine-tuning is performed using the SFTTrainer, where the model, dataset, tokenizer, and training arguments are integrated. The training process typically takes approximately 30 minutes, comprising 5 epochs and 775 steps, during which the loss gradually decreases from 2.7226 to 0.6808.

Overall, the training procedure prioritizes parameter-efficient techniques to optimize model performance within the constraints of available resources, resulting in a trained model capable of delivering accurate medical information with enhanced usability and efficiency.



(a) Loss vs Traning Steps (b) Learning Rate vs Steps (c) Epoach vs Steps

Figure 1: Results of Fine Tune Model

11 Setup Instructions

1. Open Google Colab:
2. Navigate to Google Colab in your web browser.
3. Upload Notebook:
4. Click on File -¿ Upload Notebook.
5. Select the "Fine_tune_Llama_2_colab_final.ipynb" file from your local system and upload it.
6. Select a GPU:
7. Navigate to the notebook menu and click on Runtime.
8. Then, select Change runtime type.
9. In the dropdown menu for Hardware accelerator, choose GPU.
10. Click Save.
11. Run the Notebook:
12. Simply run the notebook cells by clicking the play button on each code cell, or use Ctrl + Enter to run the currently selected cell.
13. This will execute the code within the notebook.

12 Conclusion

The development of an AI-based Healthcare chatbot system utilizing Natural Language Processing holds significant promise in providing accurate medical information in an accessible and user-friendly manner. By fine-tuning the llama2-7b language model with a custom dataset from the Gale Encyclopedia on alternative medicine and employing techniques like LoRA and QLoRA, the system aims to deliver enhanced usability and security. Leveraging the Hugging Face model and libraries facilitates seamless integration and efficient training. Ultimately, this project endeavors to offer a reliable and intuitive platform for accessing medical information and support in the realm of alternative medicine, thereby contributing to improved healthcare accessibility and empowerment.

Our fine-tuned chatbot model link

References

1. On the Effectiveness of Parameter-Efficient Fine-Tuning `https://arxiv.org/abs/2211.15583`.
2. Llama 2: Open Foundation and Fine-Tuned Chat Models `https://arxiv.org/abs/2307.09288`.
3. LoRA: Low-Rank Adaptation of Large Language Models `https://arxiv.org/abs/2106.09685`
4. QLoRA: Efficient Finetuning of Quantized LLMs `https://arxiv.org/abs/2305.14314`.