


"Innovating Healthcare: AI-Driven Drug Classification"

Maryam Zubair

Introduction




In this project, we leverage the advanced capabilities of OpenAI's GPT-3.5 to fine-tune a model for drug classification. Utilizing a carefully selected dataset of 2,000 drug examples from an Excel file, our aim is to enhance the accuracy of drug categorization, focusing specifically on their associated ailments.

The essence of this project lies in training an intelligent computer system to identify the diseases targeted by various medications. Initially, we compile comprehensive data about numerous medicines and the conditions they treat. This data is then systematically organized to facilitate the computer's learning process. Upon successful training, the computer applies this knowledge to accurately identify the disease each medication is intended to treat.

Design

Preparing the Data

- 1. Data Conversion:** We begin by transforming the XLSX data file into a JSONL format, suitable for model fine-tuning. This is achieved using Pandas and OpenAI tools, with the data being formatted to include drug names and corresponding illnesses in a prompt-completion style. Special care is taken to ensure that each completion begins with a whitespace.



Design (cont.)

Fine-Tuning Commands

1. **Data Analysis and Preparation:** We utilize OpenAI's tools, specifically the `fine_tunes.prepare_data` command, to process the data. This involves dividing the dataset into training and validation sets, which are crucial for effective model training.
2. **Model Training:** The training of the model is carried out using the `fine_tunes.create` command. We set parameters like the model type (ada) and classification metrics, and specify the training and validation data files to guide the training process.
3. **Monitoring Job Progress:** In case of disconnection during fine-tuning, a command is provided to check the progress of the job.
4. **Completion of Fine-Tuning:** Upon the completion of the fine-tuning job, an output is received confirming the completion cost and other pertinent details. The fine-tuned model is then ready for generating completions based on the trained data.

Code

```
import pandas as pd

# Number of rows to read
n = 2000

df = pd.read_excel('Medicine_description.xlsx', sheet_name='Sheet1',
                  header=0, nrows=n)

# Get the unique values in the 'Reason' column of the data frame,
# stores them in an array called reasons
reasons = df["Reason"].unique()
reasons_dict = {reason: i for i, reason in enumerate(reasons)}

df["Drug_Name"] = "Drug: " + df["Drug_Name"] + "\n" + "Malady:"
df["Reason"] = " " + df["Reason"].apply(lambda x: "" + str(reasons_dict[x]))
df.drop(["Description"], axis=1, inplace=True)
df.rename(columns={"Drug_Name": "prompt", "Reason": "completion"}, inplace=True)

# Convert the dataframe to jsonl format
jsonl = df.to_json(orient="records", indent=0, lines=True)

# Write the jsonl to a file

with open("drug_malady_data.jsonl", "w") as f:
    f.write(jsonl)
```

Implementation

1. **Setting Up the Environment:** To begin, activate or create a virtual environment.
2. **Installing Necessary Packages:** Install the required packages by executing `pip install pandas openpyxl openai==0.28`.
3. **Data Processing:** The next step involves processing the dataset in preparation for model fine-tuning.
4. **Data Preparation for Fine-Tuning:** Use the `openai tools` `fine_tunes.prepare_data` -f `drug_malady_data.jsonl` command to prepare the data for the fine-tuning process

```
(chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro HW_08 % cd homework02
(chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro homework02 % python3 fine_tune.py
(chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro homework02 % openai tools fine_tunes.prepare_data -f drug_malady_data.jsonl
Analyzing...

- Your file contains 2000 prompt-completion pairs
- Based on your data it seems like you're trying to fine-tune a model for classification
- For classification, we recommend you try one of the faster and cheaper models, such as `ada`
- For classification, you can estimate the expected model performance by keeping a held out dataset, which is not used for training

- All prompts end with suffix `\nMalady:`
- All prompts start with prefix `Drug: `

No remediations found.
- [Recommended] Would you like to split into training and validation set? [Y/n]: Y

Your data will be written to a new JSONL file. Proceed [Y/n]: Y

Wrote modified files to `drug_malady_data_prepared_train.jsonl` and `drug_malady_data_prepared_valid.jsonl`
Feel free to take a look!

Now use that file when fine-tuning:
> openai api fine_tunes.create -t "drug_malady_data_prepared_train.jsonl" -v "drug_malady_data_prepared_valid.jsonl" --compute_classification_metrics --classification_n_classes 7

After you've fine-tuned a model, remember that your prompt has to end with the indicator string `\nMalady:` for the model to start generating completions, rather than continuing with the prompt.
Once your model starts training, it'll approximately take 50.33 minutes to train a `curie` model, and less for `ada` and `babbage`.
Queue will approximately take half an hour per job ahead of you.
(chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro homework02 %
```

Implementation (cont.)

5. Configuring OpenAI API Key: Set up your OpenAI API key with the command `export`

`OPENAI_API_KEY="your_api_key_here"`

6. Model Fine-Tuning: Fine-tune the model using the command `openai api fine_tunes.create`, specifying training data, validation data, classification metrics, the number of classes, model type, and suffix.

```
(Ctrl-C will interrupt the stream, but not cancel the fine-tune)
[2023-11-22 13:44:09] Created fine-tune: ft-eSSlbfl55HL1FqDNfMU61GMq
[2023-11-22 13:44:15] Fine-tune costs $0.05
[2023-11-22 13:44:15] Fine-tune enqueued. Queue number: 0
[2023-11-22 13:44:20] Fine-tune started
```


Implementation (cont.)

7. Monitoring the Fine-Tuning Job: Follow the progress of the fine-tuning job with `openai api fine_tunes.follow -i <JOB ID>`, where the job ID is provided in the final step of the setup.

```
(chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro homework02 % openai api fine_tunes.follow -i ft-eSSlbf155HL1FqDNfMU61GMq
[2023-11-22 13:44:09] Created fine-tune: ft-eSSlbf155HL1FqDNfMU61GMq
[2023-11-22 13:44:15] Fine-tune costs $0.05
[2023-11-22 13:44:15] Fine-tune enqueued. Queue number: 0
[2023-11-22 13:44:20] Fine-tune started
[2023-11-22 13:49:37] Completed epoch 1/4
[2023-11-22 14:00:03] Completed epoch 3/4
[2023-11-22 14:05:47] Uploaded model: ada:ft-learninggpt:drug-malady-data-2023-11-22-22-05-46
[2023-11-22 14:05:48] Uploaded result file: file-2MCQU1cnce695qa3R6smfXTj
[2023-11-22 14:05:48] Fine-tune succeeded
```

Job complete! Status: succeeded 🎉
Try out your fine-tuned model:

```
openai api completions.create -m ada:ft-learninggpt:drug-malady-data-2023-11-22-22-05-46 -p <YOUR_PROMPT>
```


Implementation (cont.)

8. Executing the Test Script: Finally, run the test script using python3 test.py.

```
● (chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro homework02 % python3 test.py
What is 'A CN Gel(Topical) 20gmA CN Soap 75gm' used for? is used for Acne

What is 'Addnok Tablet 20'S' used for? is used for Adhd

What is 'ABICET M Tablet 10's' used for? is used for Allergies

○ (chatgptforpythondevelopers) maryamz@Maryams-MacBook-Pro homework02 % □
```



Test

This section details the testing procedures employed to evaluate the effectiveness of the fine-tuned model. It involves running specific scripts and analyzing the outcomes to ensure the model's accuracy and reliability in drug classification.



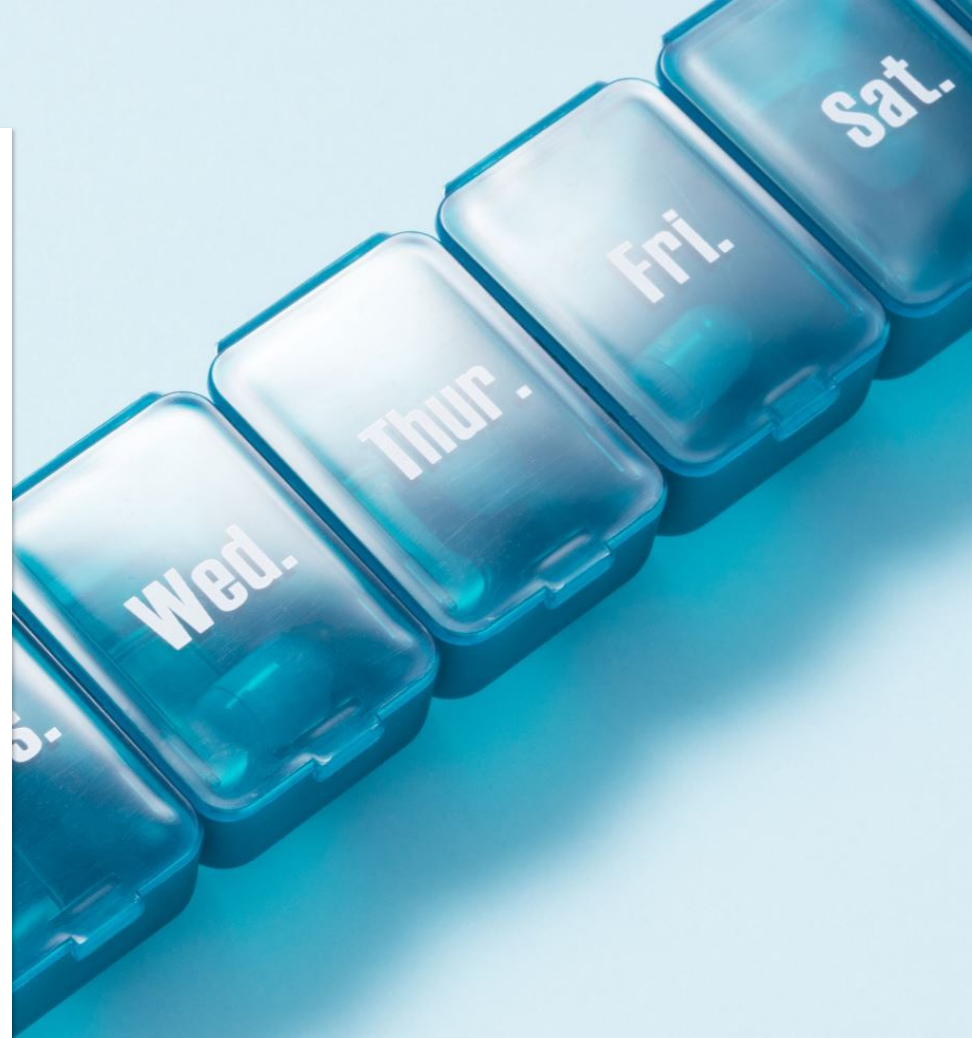
Enhancement

Enhancement


1. **Advanced Feature Integration:** We aim to augment the model's classification precision by incorporating additional elements such as dosage information, patient demographics, and potential side effects. This would expand the model's scope and enhance its accuracy.
2. **Real-Time Classification:** Developing a system capable of real-time drug classification is another objective. This would provide immediate insights and responses, particularly beneficial in dynamic healthcare scenarios.
3. **Continuous Model Refinement:** An ongoing process of model enhancement is envisioned, where periodic updates are made. These updates will include new drug information and maladies, ensuring the model remains up-to-date and effective.

Conclusion

This project has successfully harnessed the power of OpenAI's GPT-3.5 to develop a model specialized in drug classification, using a dataset of 2,000 drug examples. The primary goal was to train the system to proficiently categorize medicines based on the illnesses they are designed to treat. This initiative demonstrates the vast potential of advanced machine learning in the healthcare sector, setting a precedent for future developments in efficient and refined drug classification systems.



GitHub Link



https://github.com/Maryam-Zubair/MachineLearning_Assignment/tree/main/ChatGPT/Fine_Tuning/Drug_Example