Ostbayerische Technische Hochschule Amberg-Weiden
Fakultät Elektrotechnik, Medien und Informatik

Prof. Dr. Ulrich Schäfer

Studiengang Künstliche Intelligenz

Projektarbeit Natural Language Processing & Information Retrieval

Ulrich Schäfer

14. November 2024

Inhaltsverzeichnis

1 Aufgabenstellung

Ziel: Vergleich von fine-tuned Transformer-Modellen und pretrained (aber nicht fine-tuned) LLMs (LLAMA2, ChatGPT) für die Aufgabe des Question Answering.

Sie sollen in dieser Studienarbeit einen Vergleich durchführen, wie gut verschiedene Modelle die Aufgabe des Question Answering beherrschen. Bis zur Veröffentlichung von ChatGPT war die typische Vorgehensweise dafür, ein Transformer-Modell für den spezifischen Datensatz zu fine-tunen. Dieses Modell erhält die Frage sowie einen Kontext, aus dem sich (meistens) die Frage beantworten lässt, als Input.

Mit dem Erfolg von ChatGPT und weiteren LLMs ist eine weitere Strategie möglich geworden: Anhand eines geschickt gewählten Prompts erhält das Modell die Aufgabe, die Frage mittels des Kontext zu beantworten.

Sie sollen in dieser Arbeit beide Ansätze vergleichen.

Verwenden Sie für den ersten Ansatz folgende Modelle:

- DistilBERT, `https://huggingface.co/docs/transformers/model_doc/distilbert#transformers.DistilBertForQuestionAnswering`
- T5, `https://huggingface.co/docs/transformers/model_doc/t5#transformers.T5ForQuestionAnswering`

Diese besitzen eine verlinkte Q&A Klasse in der transformers Bibliothek, die die Verwendung vereinfacht.

2 Datensatz: SQuAD

Verwenden Sie SQuAD als Datensatz (jedoch ohne unbeantwortbare Fragen, also in Version 1). Dieser enthält rund 100.000 Fragen.

`https://rajpurkar.github.io/SQuAD-explorer/`

`https://huggingface.co/datasets/squad`

Sie finden im Internet mehrere Anleitungen zum Fine-Tuning auf diesem Datensatz. Beispielsweise:

- `https://medium.com/@ajazturki10/simplifying-language-understanding-a-beginners-g`
- `https://www.youtube.com/watch?v=ZIRmXkHp0-c`
- `https://colab.research.google.com/drive/1M43YmOX4jp1aMAfinK5c3RlfBj1li5hL`
- `https://colab.research.google.com/drive/1WxGxCFE_1cESJ02baaBY-HBHmGjSlxJx`

Nur zur Info: Übersicht QA data sets [1].

3 Evaluierung

Die Evaluierung dieses Datensatzes geschieht über den F1 Score und den Exact Match Score.

Da der Datensatz als eine Challenge veröffentlicht wurde, ist der Test-Datensatz geheim. Behandeln Sie daher den veröffentlichten Validierungsdatensatz als Test-Datensatz.

Die beiden fine-tuned Modelle evaluieren Sie auf diesem gesamten Test-Datensatz (TD-A).

Erzeugen Sie zusätzlich durch zufälliges Ziehen einen kleineren Test-Datensatz (TD-B), der 100 Fragen enthält.

Als LLMs verwenden Sie ChatGPT und LLAMA 2 7b Chat (`https://huggingface.co/meta-llama/Llama-2-7b-chat-hf`).

Diese testen Sie beide auf TD-B. Testen Sie die beiden fine-tuned Modelle ebenfalls auf TD-B, sodass der Vergleich zu den LLMs möglich wird.

Beide LLMs werden mittels eines Prompts aufgefordert, die Antwort zur Frage basierend auf dem Kontext zu geben. Erzeugen Sie ein Prompt-Template, das Sie programmatisch mit Kontexten und Antworten aus TD-B füllen können.

LLAMA 2 können Sie anschließend auf unserem Server verwenden. ChatGPT müssen Sie über das Web-Interface aufrufen: `https://chat.openai.com`

Für die Evaluierung von ChatGPT ist es also nötig, dass Sie die einzelnen, ausgefüllten Prompts per Copy & Paste an ChatGPT geben. Speichern Sie die Eingaben und Ausgaben ab.

Bei beiden LLMs kann es passieren, dass die Modelle nicht ausschließlich die Antwort ausgeben, sondern zusätzlich beispielsweise einen einleitenden Text (z.B. „Sure, here is the answer to the question based on the context. . . “) oder eine Erklärung beinhalten. Dies müssen Sie vor der Evaluierung entfernen. Versuchen Sie, dieses Verhalten schon mit ihrem Prompt-Template zu verhindern.

Stellen Sie abschließend alle Modelle gegenüber. Wie verhalten sich die fine-tuned Modelle auf TD-A und TD-B? Ist TD-B ausreichend groß, um Rückschlüsse auf das Verhalten bei TD-A zu ziehen? Wie schneiden die LLMs im Vergleich zu den fine-tuned Modellen auf TD-A ab?

Literatur

[1] Wang, Zhen: *Modern Question Answering Datasets and Benchmarks: A Survey*. 2022