UNIVERSITÄT ZU LÜBECK

Hello World

Hallo Welt

Masterarbeit

verfasst am
Institut für Informationssysteme

im Rahmen des Studiengangs
Informatik
der Universität zu Lübeck

vorgelegt von
Jakob Horbank

ausgegeben und betreut von
Prof. Dr. Ralf Möller

Lübeck, den 15. April 2024

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

_____

Jakob Horbank

Zusammenfassung

Es geht darum, der Welt »Hallo« zu sagen.

Abstract

It is about saying “hello” to the world.

Contents

1

Introduction

This an introduction like not other. Information retrieval hard an stuff. Would be nice to have a chatbot that answer natural language questions and gives sources from research database and even web.

1.1 Contributions of this Thesis

Hopefully the above.

1.2 Related Work

There are different attempts at creating LLM outputs with sources. Perplexity AI hosts a question answering service, that gives source from websites.

Open Source LLM Agents

- AutoGPT
- babyagi

Information Retrieval Applications

In a variety of application contexts, the answers of an assistent have to be correct. This is espciacally true in reasearch contexts. A model that hallucinates isn't feasable in this case. But while hallucination can be reducing with fine-tuning, it can not be competely eliminated. Perplexity AI is an online service that leverages a language model to provide a search that generates an answer from different sources in the internet. The content of the answer is then linked to the found sources, so the user can see and verify the result.

As this is a propierty closed source product accessible through a web interface, this is not a useful to create our own research assistant. The provided API simply hosts different LLMs and allows for prompting them. There is no possibilty of of fine-tuning.

1.3 Structure of this Thesis

I do this then that and then that.

2

Backgrounds

Different branches of research were used to build upon in this work. These topics and how their are connected to my work are explained in more detail in this section.

2.1 Information Retrieval

Handling large databases filled with information of different is a common problem. There has been extensive reasearch on database architectures and indexing algorithms.

2.2 Agents

Although the notion of agent is not new, it recently gained attention in combination with the rise of generative language models.

2.3 Large Language Models

Natural language processing is a long studied research area of computer science. In recent years delevopments have significantly sped up, with the introduction of deep learning into NLP. The transformer archticture has been the basis for all advancements in recent years.

Transformer

Until 2017, the dominating strategy to train models for language tasks revolved around recurrent structures. Every sentence token was represented as a hidden state that resultet from all the previous tokens. While this approach has a reasonable motivation, the sequential nature contraints computation speed. There is no way to compute the recurrent architectures in parallel.

The transformer model [1] solely relys on these attention mechanisms. In particular, they employ self-attention and multi-headed self-attention layers

Attention mechanism allow learning dependencies between tokens in sentences without regard to their distance. Their non-sequential nature allow for massive parallelization.

Self-Attention is an attention mechanism that computes relations between different positions of the same sequence with goal of finding a representation of the sequence.

Like a typical sequence modeling architecture, a transformer consists of an enconder and a decoder. The encoder has two parts that are stacked on each other multiple times. The first part is a mulit-head self-attention block, the second part is a classic fully connected feed foward block.

BERT

The BERT model family consists of encoder only transfomers. They are trained to generating rich embeddengs of tokens for use in a range of different downstream tasks.

GPT

GPT models are decoder only transformer models. They excel at text generation tasks. they are trained to predict the next token of a sequence.

InstructGPT

Instruct models are fine-tuned version of language models. Capturing the intent of the user is a key challenge for language models. This process is called *alignment*. Even though LLMs are trained on huge datasets, they are not tailored towards human users by default. A popular approach to the alignment problem is reinforcement learning with human feedback (RLHF). Handcrafted prompts are used to fine-tune GPT-3. The outputs of the model are collected into a set and ranked by humans. This set is then used to train a reward model. With this reward model, the language model is further fine-tuned. The resulting model is called *InstructGPT* and performs better than the baseline GPT-3 model.

Large language models are trained to predict the next token of a sequence, not to follow the instruction of the user. This leads to some unwanted results, such as toxic, harmful answers or fabricated information that is not true.

LLM Agents

AutoGPT is an open source project that tries to leverage LLMs to function as an agent controller. GPT models are extended with different modules to creat an agent that receives a goal and then acts towards reaching it. To reach the goal AutoGPT plans, has a memory, and reflects on past actions.

AutoGPT is an example implementation of an agent. The project tries to become a framework to build agents with different architectures. AutoGPT employs the key components of an agent outlined in [2]. The profile module allows the agent to assume different roles, such as expert, teacher or coder. Profile are LLM specific as each model responds best for different prompting styles. Memory is implemented through a database.

Currently, AutoGPT is only compatible with OpenAI models. Switching to a different model is not that hard, because only the API format would have to be changed. The challenge is to switch between different prompting styles, as every model needs to be prompted differently. For example OpenAI models benefit from profile sentences like ”You are an expert in computer science”, while Anthropics Claude does not…

Other open source llm agent systems

– miniagi miniagi
– babyagi miniagi

3

Handcrafted Prompts as a Research Chatbot

- First, only crafting prompt for default GPT-3.5
- What could be questions about hadith corpus?
  - Ask for classifaction outputs
    These would be the same as the BERT classifactions, just as question sentences. Questions about contained persons, locations, dates, ….
  - Further questions
    Information probably not contained in corpus. LLM might be able to give some reasoning learned in pre-training.
- How to design prompts?
  Few-shot with examples or Zero Shot? Include example labelings from corpus.
- How to include corpus?
  The corpus can be large. Large than the context length of LLMs. So first try handling a single document
- How to craft testing examples?
  For the labeling questions create examples for amended data.

4

Custom Fine-Tuned Language Model as a Research Chatbot

- Here, explore fine-tuning a LLM
- Identify problems of non fine-tuned Chatbot
- How to create fine-tuned dataset

  Again, for labeling questions from amended data, but with focus on weaknesses.

5

Conclusion

Saying hello world is quite easy.

Bibliography

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. *A Survey on Large Language Model based Autonomous Agents*. Aug. 2023. DOI: 10.48550/ARXIV.2308.11432. arXiv: 2308.11432 [cs.AI].