Project Proposal

Overview

In a paper by François Chollet titled "On the Measure of Intelligence, [1]" the author offers a transformative perspective on evaluating and advancing AI systems. He claims that the current "task-specific" approach to measuring the intelligence of these machine-learning systems is inadequate. According to him, these approaches cannot capture the essence of true intelligence, which is significantly influenced by an entity's skill-acquisition efficiency across varying domains. Based on this, the author created the Abstraction and Reasoning Corpus (ARC), a benchmark to evaluate machine-learning systems on tasks that require human-like general fluid intelligence. The tasks in the ARC challenge emphasize the importance of innate priors and the ability to generalize from minimal data.

Based on Chollet's foundation, in this project, we aim to solve the ARC challenge by leveraging a novel approach that combines human insights with the power of large language models. The main goal of this project is to create a corpus consisting of human-annotated explanations that describe the reasoning processes involved in solving ARC problems. This corpus will then be used to fine-tune an open-source LLM to integrate it with an understanding of problem-solving strategies that resemble human thought processes.

Project Steps

1. Annotation Creation: We will start this project by creating detailed, step-by-step explanations for the training and evaluation sets of the ARC challenge. These annotations will explain the logical and abstract reasoning patterns that we use to navigate the complex problem space of the ARC.
2. Model Fine-Tuning: We will use these annotations to create a two-step task for fine-tuning the LLM model. First, for each input from the ARC, the model will generate a step-by-step guide on how to solve the given corpus, and then it will use this guide to predict the output of that ARC input. In LLM literature, such an approach to solving a problem is called chain-of-thought prompting and was first introduced in the paper titled "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. [2]"

   a. During the fine-tuning stage of the LLM, its training task will be to first generate explanations that closely match the human explanations in terms of the underlying logic. From there, it will be further tasked to generate the correct output by feeding it the human-annotated explanation for each example (an example of the teacher forcing method)
   b. During the testing/evaluation phase, the LLM will be presented with only the ARC inputs, and its task will be to first autonomously generate explanations on how to approach and solve each problem. Following the generation of these explanations, the model will then attempt to solve the problems based on its generated reasoning paths.
3. (Optional) With the recent popularity of multimodal LLMs, we are also thinking of fine-tuning a multimodal LLM like LLaVA on the ARC dataset. This is due to the fact that when we, as humans, solve the ARC inputs, we look at a visual representation of the data and not the numerical inputs that we feed to the ML models. The main reason this is listed as optional is due to some recent research that claims the reasoning capabilities of multimodal LLMs are worse than their vanilla counterparts. [3]
4. Finally, we will make a dashboard to present our results and host a live demo where we take a few inputs from ARC and compare the human-generated annotations with the annotations generated by the fine-tuned model.

Expected Outcomes and Impact

This project is an exploration into the feasibility of improving large language models' (LLMs) problem-solving capabilities through chain-of-thought prompting guided by human-annotated explanations. The main goal is to investigate whether an LLM can be fine-tuned to understand and generate human-like reasoning paths for solving complex problems and to evaluate its ability to apply these reasoning paths to produce correct solutions.

This project is essentially a stepping stone, aiming to show that it's possible to integrate human-like reasoning into large language models (LLMs) by carefully teaching them with detailed explanations. The findings from this project will help us identify the limitations, challenges, and potential areas for future research. It will also lay some groundwork for more advanced explorations of systems capable of human-like reasoning and problem-solving.

Technical Details

Abstraction and Reasoning Corpus

1. The dataset is currently being hosted on Kaggle.
2. It consists of 400 training examples, 400 evaluation examples and 100 testing examples.
3. Each example is stored as a JSON file. In each file, we have the main key called root that contains two other keys called train and test.
4. Each train section of the JSON file contains three examples that depict the reasoning task.
5. Each test section of the JSON file contains the reasoning task that we need to solve.
6. This dataset also comes with an interactive app that allows us to visualize each of the reasoning problems. That tool can be found here.

Annotation details

The annotations that we will generate will be stored in a CSV file consisting of two columns: the filename of the JSON file whose explanations are in that row and the human-annotated explanations. We will likely make a makeshift dashboard that modifies the code of the interactive app to add the option of creating the annotations directly within the app interface. This will streamline the annotation process, allowing annotators to interact with the reasoning tasks visually and enter their explanations in a more intuitive and efficient manner.

References

1. Chollet, François. "On the measure of intelligence." arXiv preprint arXiv:1911.01547 (2019).
2. Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837
3. Mitchell, Melanie, Alessandro B. Palmarini, and Arseny Moskvichev. "Comparing Humans, GPT-4, and GPT-4V on abstraction and reasoning tasks." *arXiv preprint arXiv:2311.09247* (2023).