gururise / **AlpacaDataCleaned** Public

Alpaca dataset from Stanford, cleaned and curated

Apache-2.0 license

366 stars **31** forks

Star Watch

Code Issues 3 Pull requests Actions Projects Security Insights

main

gururise Merge pull request #17 from hzj5790/main ... 5 hours ago **68**

View code

README.md

Cleaned Alpaca Dataset

Welcome to the Cleaned Alpaca Dataset repository! This repository hosts a cleaned and curated version of a dataset used to train the Alpaca LLM (Large Language Model). The original dataset had several issues that are addressed in this cleaned version.

Dataset Quality and its Impact on Model Performance

One possibility behind the lack of a significant improvement in performance from fine-tuning the 7B Alpaca model to the 13B model is the quality of the original dataset. The original dataset used to train the Alpaca model was generated with GPT-3, which itself may have had limitations due to data quality. More evidence pointing to poor data quality is that fine-tuning on the original dataset resulted in poor loss curves.

The quality of the dataset plays a crucial role in determining the performance of the natural language processing models trained on it. A dataset that is noisy, inconsistent, or incomplete can result in poor performance even with the most advanced models. In contrast, a high-quality dataset can enable a model to perform well with smaller parameters.

Therefore, it is possible that with better data, we could improve the performance of the models more than what would be gained by simply increasing model size.

Data Cleaning and Curation

Alpaca is a fine-tuned version of LLAMA that was trained using an Instruct Dataset generated by GPT-3. The generated dataset was designed to be diverse; however, recent analysis indicates it is very US centric. The original dataset used to train the Alpaca LLM was found to have many issues that impacts its quality and usefulness for training a machine learning model.

Issues with the Original Dataset

The original dataset had several issues that likely affect the final performance of the fine-tuned model. These issues include:

1. **Hallucinations:** Many instructions in the original dataset had instructions referencing data on the internet, which just caused GPT3 to hallucinate an answer.

```
"instruction":"Summarize the given article in 200 Words.",
"input": "https://www.bbc.com/news/world-51461830",
"output": "The recent protests in Belarus are part of a growing wave of
discontent against the rule of Alexander Lukashenko, who has led the
country since 1994. Demonstrations against his government have been
ongoing since August, and come in the wake of an election widely seen as
rigged. As police and security forces have cracked down on protesters, the
movement has only grown, with hundreds of thousands of people taking to
the streets to call for new leadership. Human rights activists have
accused the government of using extreme violence in order prosecute
protesters, leading to an international outcry and sanctions from the
European Union. As the situation in Belarus continues to unfold, the
international community is monitoring the situation closely - hoping for a
peaceful resolution, while also standing firm against the oppressive
tactics employed by the Belarusian state."

"instruction": "Classify the following shape according to simple
geometry.",
"input": "<Picture Attached>",
"output": "Triangle"

"instruction": "Given a recent news article, write a headline for it.",
"input": "https://www.news.example.com/article-1",
```

```
"output": "\"NASA Launches a Mars 2020 Mission to Search for Signs of
Ancient Life\""
```

2. **Merged Instructions:** There were many instructions that were merged together in the original dataset for some reason.

```
"instruction": "Should I invest in stocks?",
"input": "",
"output": "It depends on your individual situation. Investing in stocks
carries a certain amount of risk but can also offer potential rewards.
Make sure to research the different types of stocks, understand the risks
and rewards, and develop a plan that works for you.\n##\n18. Instruction:
Identify the noun in the following sentence:\n18. Input:\nThe large apple
was delicious.\n18. Output:\nApple"
```

3. **Empty outputs:** Some entries in the original dataset had empty outputs.

```
"instruction": "Calculate the area of the following shape in square
centimeters.",
"input": "rectangle of size 4 cm x 5 cm",
"output": ""
```

4. **Empty code examples:** Some descriptions in the original dataset were missing code examples, making it difficult to understand the intended behavior of the code.
5. **Instructions to generate images:** Some descriptions in the original dataset included instructions to generate images, something obviously not possible.

```
"instruction": "Create a graphic or logo that visually represents the word
\"courage\".",
"input": "",
"output": "<No Output>"
```

6. **N/A outputs:** Some code snippets in the original dataset had N/A outputs.
7. **Inconsistent input field:** The original dataset had inconsistent usage of the input field when it was supposed to be empty.

```
"input":"<no input>"
"input":"No input"
```

```
"input":"noinput"
"input":"<noinput>"
```

8. **Wrong answers:** Some instructions/questions in the original dataset had incorrect answers. About 80% of the math problems are estimated to have incorrect answers.

```
"instruction": "Calculate the median of the following data set.",
"input": "1, 2, 4, 5, 8, 9",
"output": "5"

"instruction": "Convert 25m to km.",
"input": "",
"output": "25km"
```

9. **Non-Sensical/Unclear instructions:** Many instructions are unclear, we try to clarify (or re-write) if instructions are non-sensical. Instructions that are slightly unclear, but where one could deduce the meaning are not altered.

```
"instruction": "Freeze the following sample of yogurt for 10 minutes.",
"input": "Yogurt sample",
"output": "<noinput>"

"instruction": "Increase the font size to 12 points.",
"input": "",
"output": "The font size has been increased to 12 points."
```

10. **Extraneous escape and control characters:** The original dataset had several entries with extraneous escape and control characters.

Hugging Face Hub

The cleaned dataset is also available on the Hugging Face Hub.

Contributions

With over 52k entries, several issues still exist. Please help out by submitting a pull-request.

Goals

The primary goal of this project is to provide a cleaned and curated version of the Alpaca dataset that will improve the performance of natural language processing models trained on this data. By removing errors and inconsistencies, the goal is to improve performance of the fine-tuned llama models and reduce the likelihood of hallucinations.

Acknowledgments

The original version of the Alpaca dataset was sourced from tatsu-lab's github repository. We would like to thank the original creators of these datasets for making their data available to the public. We would also like to thank the team at Meta AI for their work in developing Llama.

Releases

No releases published

Packages

No packages published

Contributors 7

Languages

- **Python** 100.0%