Self-Instruct: Aligning LM with Self Generated Instructions

This repository contains code and data for the Self-Instruct paper, a method for aligning pretrained language models with instructions.

Introduction

Self-Instruct is a framework that helps language models improve their ability to follow natural language instructions. It does this by using the model's own generations to create a large collection of instructional data. With Self-Instruct, it is possible to improve the instruction-following capabilities of language models without relying on extensive manual annotation.

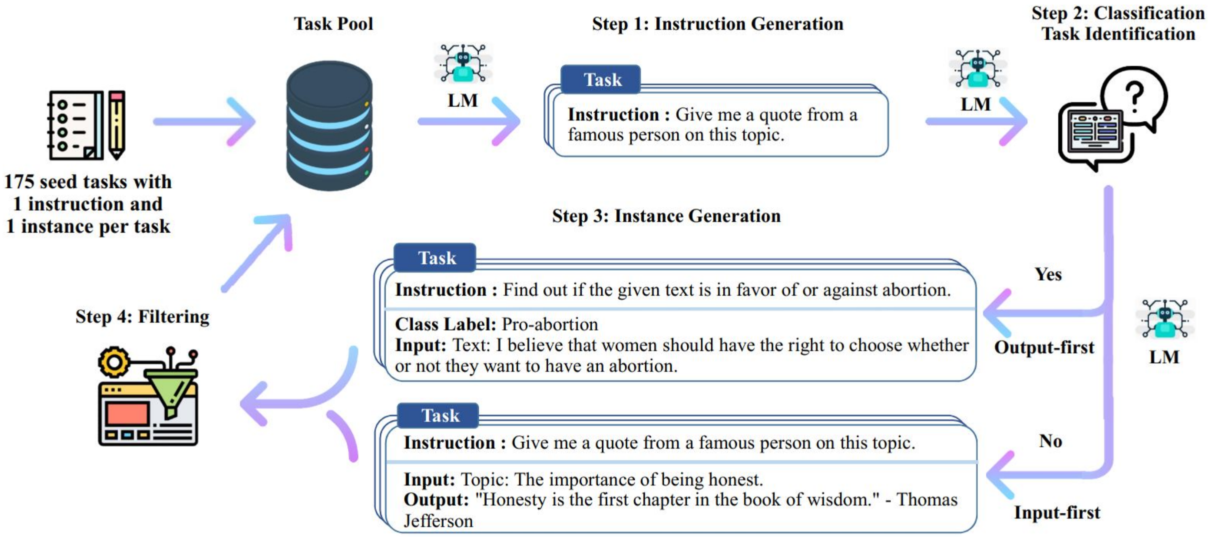
Background

In recent years, there has been a growing interest in building models that can follow natural language instructions to perform a wide range of tasks. These models, known as "instruction-tuned" language models, have demonstrated the ability to generalize to new tasks. However, their performance is heavily dependent on the quality and quantity of the human-written instruction data used to train them, which can be limited in diversity and creativity. To overcome these limitations, it is important to develop alternative approaches for supervising instruction-tuned models and improving their instruction-following capabilities.

How Self-Instruct works?

The Self-Instruct process is an iterative bootstrapping algorithm that starts with a seed set of manually-written instructions and uses them to prompt the language model to generate new instructions and corresponding input-output instances. These generations are then filtered to remove low-quality or similar ones, and the resulting data is added back to the task pool. This process can be repeated multiple times, resulting in a large collection of instructional data that can be used to fine-tune the language model to follow instructions more effectively.

Here is an overview of Self-Instruct:


Task Pool
Step 1: Instruction Generation
Step 2: Classification Task Identification
LM
LM
175 seed tasks with 1 instruction and 1 instance per task
Task
Instruction : Give me a quote from a famous person on this topic.
Step 3: Instance Generation
Task
Instruction : Find out if the given text is in favor of or against abortion.
Class Label: Pro-abortion
Input: Text: I believe that women should have the right to choose whether or not they want to have an abortion.
Yes
Output-first
LM
Step 4: Filtering
Task
Instruction : Give me a quote from a famous person on this topic.
Input: Topic: The importance of being honest.
Output: "Honesty is the first chapter in the book of wisdom." - Thomas Jefferson
No
Input-first

Usage

*** This work is still in progress. We may update the code and data as we make progress. Please be cautious about the version control.**

Instruction-tuning using our Self-Instruct data

We release a dataset that contains 52k instructions, paired with 82K instance inputs and outputs. This instruction data can be used to conduct instruction-tuning for language models and make the language model follow instruction better. The entire model-generated data can be accessed in `data/gpt3-generations/batch_221203/all_instances_82K.jsonl`. This data (+ the 175 seed tasks) reformatted in clean GPT3-finetuning format (prompt + completion) is put in `data/finetuning/self_instruct_221203`. You can use the script in `./scripts/finetune_gpt3.sh` to finetune GPT3 on this data.

Note: This data is generated by a language model (GPT3) and inevitably contains some errors or biases. We analyzed the data quality on 200 random instructions in our paper, and found that 46% of the data points may have problems. We encourage users to use this data with caution and propose new methods to filter or improve the imperfections.

Evaluating instruction-following capabilities

We also release a new set of 252 expert-written tasks and their instructions motivated by user-oriented applications (rather than well-studied NLP tasks). This data is used in the human evaluation section of the self-instruct paper. Please refer to the human evaluation README for more details.

Generating Self-Instruct data from scratch

To generate Self-Instruct data using your own seed tasks or other models, we open-source our scripts for the entire pipeline here. Our current code is only tested on the GPT3 model accessible via the OpenAI API.

Here are the scripts for generating the data:

```
# 1. Generate instructions from the seed tasks
./scripts/generate_instructions.sh

# 2. Identify whether the instruction represents a classification task or not
./scripts/is_clf_or_not.sh

# 3. Generate instances for each instruction
./scripts/generate_instances.sh

# 4. Filtering, processing, and reformatting
./scripts/prepare_for_finetuning.sh
```

Citation

If you use the Self-Instruct framework or data, feel free to cite us.

```
@misc{selfinstruct,
  title={Self-Instruct: Aligning Language Model with Self Generated Instructions},
  author={Wang, Yizhong and Kordi, Yeganeh and Mishra, Swaroop and Liu, Alisa and Smith, Noah A. and Khashabi, Daniel and Hajishirzi, Hannaneh},
  journal={arXiv preprint arXiv:2212.10560},
  year={2022}
}
```