RLHF初步探索

卢艳峰

September 9, 2023

前言

- 深入使用微软的 **DeepSpeed-Chat** 后，决定放弃使用它作为后续工作的研究工具。
- 研究其他 **RLHF** 的开源实现，决定使用中文基座模型 **Chinese-LLaMA-2-7B** 作为后续工作的预训练模型。
- 在上述学习过程中发现新的 **RLHF** 方法（**DPO**）不需要第 **2** 步的奖励模型的训练，决定采用 **DPO** 作为后续的研究方法。
- 阅读 **Policy Gradient** 原论文。
- 阅读 **Llama 2** 原论文（**77** 页，未完成）。

DeepSpeed-Chat

- **facebook/opt-1.3b**，英文模型。
- **facebook/opt-350m**，英文模型。
- **bigscience/bloom-1b1**，词表导致显存消耗太大，且中文效果不好。
- **Langboat/bloom-1b4-zh**，中文效果不好。
- **IDEA-CCNL/Wenzhong2.0-GPT2-3.5B-chinese**，无法成功导入模型进行训练。
- **FlagAlpha/Atom-7B**，**zero_stage = 3** 时，**DeepSpeed-Chat** 官方实现的 **save_zero_three_model** 函数无法成功保存训练完成的模型。

DeepSpeed-Chat

- 英文比较数据：**Dahoas/rm-static**
- 中文比较数据：**zwh9029/rm-static-m2m100-zh-jianti**
- 中文问答数据：**wangrui6/Zhihu-KOL，Hello-SimpleAI/HC3-Chinese**

Chinese-LLaMA-2-7B

对比项	中文LLaMA-2	中文Alpaca-2
模型类型	**基座模型**	**指令/Chat模型（类ChatGPT）**
已开源大小	7B、13B	7B、13B
训练类型	Causal-LM (CLM)	指令精调
训练方式	LoRA + 全量emb/lm-head	LoRA + 全量emb/lm-head
基于什么模型训练	原版Llama-2（非chat版）	中文LLaMA-2
训练语料	无标注通用语料（120G纯文本）	有标注指令数据（500万条）
词表大小[1]	55,296	55,296
上下文长度[2]	标准版：4K（12K-18K） 长上下文版：16K（24K-32K）	标准版：4K（12K-18K） 长上下文版：16K（24K-32K）
输入模板	不需要	需要套用特定模板[3]，类似Llama-2-Chat
适用场景	文本续写：给定上文，让模型生成下文	指令理解：问答、写作、聊天、交互等
不适用场景	指令理解 、多轮聊天等	文本无限制自由生成

https://github.com/ymcui/Chinese-LLaMA-Alpaca-2

DPO

- **ziqingyang/chinese-llama-2-7b** 于 **09/08/2023 21:01:07** 完成 **DPO** 在 **GPT-4 Generated Data (zh)** 的训练（第一阶段：大约 **5** 小时，第二阶段：大约 **3** 小时）。
- 由于时间关系，还未对结果进行详细评测。
- 以上数据取自 **1** 张 **NVIDIA GeForce RTX 4090** 的训练结果。

计划

- 阅读 **DPO** 论文，加深对 **RLHF** 的认识。
- 用一个中文预训练模型跑完 **RLHF** 流程。
- 将中文通用的问答数据和师妹用 **ChatGPT** 生成的数据混合作为我们的训练集。
- 用微调后的模型（第一步）生成的回复作为 **rejected** 回复。

参考

- **Llama 2：https://arxiv.org/abs/2307.09288**
- **ziqingyang/chinese-llama-2-7b：https://huggingface.co/ziqingyang/chinese-llama-2-7b**
- **hiyouga/LLaMA-Efficient-Tuning：https://github.com/hiyouga/LLaMA-Efficient-Tuning**
- **stack_llama/scripts：https://github.com/huggingface/trl/tree/main/examples/research_projects/stack_llama/scripts**
- **stack_llama_2/scripts：https://github.com/huggingface/trl/tree/main/examples/research_projects/stack_llama_2/scripts**
- **ymcui/Chinese-LLaMA-Alpaca-2：https://github.com/ymcui/Chinese-LLaMA-Alpaca-2**
- **https://github.com/microsoft/DeepSpeedExamples/tree/master/applications/DeepSpeed-Chat**

参考

- **Policy Gradient Methods：https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html**
- **DPO：https://arxiv.org/abs/2305.18290**

Thanks

分享人：卢艳峰