DPO初步探索

卢艳峰

September 16, 2023

前言

- 研究 **DPO** 的开源实现，在中文基座模型 **Chinese-LLaMA-2-7B** 上进行中文微调研究。
- 阅读 **DPO** 论文，论文题目为：**Direct Preference Optimization: Your Language Model is Secretly a Reward Model**。
- 阅读 **Llama 2** 原论文（**77** 页，未完成）。

DPO

- 实验细节和效果：**http://10.4.7.189/llm/dpo.html**

计划

- 研究 **DPO** 算法效果不好的原因。
- 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**
- **DPO：https://arxiv.org/abs/2305.18290**

Thanks

分享人：卢艳峰