BUPT-OS-LLM

> BUPT 2023 autumn operating system course big assignment - LLM local deployment project

项目描述

这个项目是 ChatGLM2 的本地化部署。

环境检查

请确保你的本地环境符合本项目的要求，具体如下：

- OS: Linux 建议 Ubuntu22.04
- CUDA Toolkit: 11.8+ 驱动版本请注意匹配
- 内存: 大小建议大于 16GB
- CPU: 不建议使用 AMD 的显卡或 CPU，可能会发生一些错误
- 硬盘: 请确保本地至少有 12GB 的空间
- GPU: 越强越好，但一定是 NVIDIA

> 以上要求都不是强制，但如果不符合，可能会出现一些意料之外的错误

开始使用

1. 克隆项目到本地：

```
git clone git@github.com:Conqueror712/BUPT-OS-LLM.git
```

2. 进入项目目录：

```
cd BUPT-OS-LLM
```

3. 安装依赖（建议在一个新的 conda 环境中进行）：

```
pip install -r requirements.txt
```

4. 下载 LFS 文件（这可能需要一段时间，如果遇到网络问题请使用镜像 https://aliendao.cn/models/THUDM/chatglm2-6b ）

```
git lfs install
git clone https://huggingface.co/THUDM/chatglm2-6b
```

5. 启动：

```
# 网页 Version
python web.py

# 终端 Version
python terminal.py
```

```
Loading checkpoint shards: 100%|████████████████████████████████████| 7/7 [00:10<00:00,  1.56s/it]
/root/BUPT-OS-LLM/web_demo.py:93: GradioDeprecationWarning: The `style` method is deprecated. Please set these arguments in the construct
or instead.
  user_input = gr.Textbox(show_label=False, placeholder="Input...", lines=10).style(
Running on local URL:  http://127.0.0.1:7860

To create a public link, set `share=True` in `launch()`.
```

使用示例

> 注意，如果你的硬件配置不够，可能网页 Version 不能正确运行，请使用终端 Version 进行对话。

网页 Version：

终端 Version：

```
Loading checkpoint shards: 100%|████████████████████████████████████| 7/7 [00:10<00:00,  1.50s/it]
欢迎使用 ChatGLM2-6B 模型，输入内容即可进行对话，clear 清空对话历史，stop 终止程序

用户：你好

ChatGLM：你好👋！我是人工智能助手 ChatGLM2-6B，很高兴见到你，欢迎问我任何问题。
```

贡献

如果你发现了bug，或者有任何改进意见或建议，请提交issue。欢迎参与项目的开发和贡献！

许可证

这个项目基于 MIT license 开源。