使用文档

大模型下载和部署使用指南

下载大模型

1. 安装依赖

使用以下命令安装 Hugging Face 的依赖库：

```
pip install -U huggingface_hub
```

2. 设置环境变量

Linux

在终端中执行以下命令：

```
export HF_ENDPOINT=https://hf-mirror.com
```

建议将上面这一行写入 `~/.bashrc` 文件，以便每次终端启动时自动设置环境变量。

Windows Powershell

在 Powershell 中执行以下命令：

```
$env:HF_ENDPOINT = "https://hf-mirror.com"
```

3.1 下载模型

使用以下命令下载模型（如gpt2）：

```
huggingface-cli download --resume-download gpt2 --local-dir gpt2
```

3.2 下载数据集

使用以下命令下载数据集（如Wikitext数据集）：

```
huggingface-cli download --repo-type dataset --resume-download wikitext --
local-dir wikitext
```

可以添加 `--local-dir-use-symlinks False` 参数禁用文件软链接，这样下载路径下所见即所得。

部署 OPENAI 接口

1. 创建虚拟环境

使用 Conda 创建并激活虚拟环境：

```
conda create -n fastchat python==3.10.6 -y
conda activate fastchat
```

2. 克隆代码

克隆 FastChat 仓库并进入目录：

```
git clone https://github.com/lm-sys/FastChat.git
cd FastChat
```

3. 安装依赖库

升级 pip 并安装依赖库：

```
pip install --upgrade pip  # enable PEP 660 support
pip install -e .
pip install transformers_stream_generator
pip install cpm_kernels
```

4. 启动服务

启动 Controller

```
python3 -m fastchat.serve.controller
```

启动 Model Worker(s)

```
python3 -m fastchat.serve.model_worker --model-names "gpt-3.5-turbo,text-davinci-003,text-embedding-ada-002" --model-path THUDM/chatglm2-6b
```

启动 RESTful API Server

```
python3 -m fastchat.serve.openai_api_server --host localhost --port 8000
```

5. 设置 OpenAI 接口

设置 OpenAI Base URL

```
export OPENAI_API_BASE=http://localhost:8000/v1
```

设置 OpenAI API Key

```
export OPENAI_API_KEY=EMPTY
```

采用LLaMA Factory 微调步骤

安装 LLaMA Factory

```
git clone --depth 1 https://github.com/hiyouga/LLaMA-Factory.git
cd LLaMA-Factory
pip install -e ".[torch,metrics]"
```

可选的额外依赖项：torch、torch-npu、metrics、deepspeed、bitsandbytes、hqq、eetq、gptq、awq、aqlm、vllm、galore、badam、qwen、modelscope、quality

快速开始

下面三行命令分别对如 Llama3-8B-Instruct 模型进行 LoRA **微调**、**推理**和**合并**。

使用数据集时，更新 `data/dataset_info.json` 文件。

```
llamafactory-cli train examples/train_lora/llama3_lora_sft.yaml
llamafactory-cli chat examples/inference/llama3_lora_sft.yaml
llamafactory-cli export examples/merge_lora/llama3_lora_sft.yaml
```

高级用法请参考**LLaMA Factory**的 examples/README_zh.md（包括多 GPU 微调）。

程序运行及测试

1. 运行文件和结果文件

不同文档下对应不同模型的运行文件。其中 `results` 文件夹包含数据结果文件和评分结果文件，包括 GPT、Llama2 和 XuanYuan 模型。

跟目录下的 `data` 文件夹包含数据集文件和测试数据文件。

2. 设置 API Key 和 URL

OpenAI 可以在注册个key，然后填入 `OPENAI_API_KEY` ，直接在程序中修改 KEY 和 URL 进行调用。Llama2 和 XuanYuan 等大模型需要从 Hugging Face 上进行下载。同时需要从官网注册 `SERPAPI_API_KEY` 并填入到代码的环境变量中以供调用。

注意： `SERPAPI_API_KEY` 的使用有次数限制，不能无限制使用，注意使用次数

3. 运行评分系统

每个模型的运行文件最后会生成评测结果的 CSV 文件和评测分的 JSON 文件，并保存在 `results` 文件夹中。

也可以单独运行评分系统文件 `result.py` 以修改指定的结果csv文件，并单独运行测试评分。