llama.cpp

使用

首先转换模型到gguf格式，参考 https://github.com/ggml-org/llama.cpp/blob/master/convert_hf_to_gguf.py#L4950

如 python convert_hf_to_gguf.py --outtype f16 --print-supported-models /mnt/models/opt-6.7b/

如果模型不支持，在这里添加模型，https://github.com/ggml-org/llama.cpp/blob/master/convert_hf_to_gguf_update.py

如添加opt {"name": "opt-6.7b","tokt": TOKENIZER_TYPE.BPE, "repo": "https://huggingface.co/facebook/opt-6.7b", } 还要添加模型结构应该是，https://github.com/ggml-org/llama.cpp/blob/master/convert_hf_to_gguf.py#L952

```
# Instructions:
#
# - Add a new model to the "models" list
# - Run the script with your huggingface token:
#
#   python3 convert_hf_to_gguf_update.py <huggingface_token>
#
# - The convert_hf_to_gguf.py script will have had its
get_vocab_base_pre() function updated
# - Update llama.cpp with the new pre-tokenizer if necessary
```

还没有太搞明白这里

然后编译llama.cpp，把cuda后端参数设置为on

编译后运行测试,一个简单的生成测试，https://github.com/ggml-org/llama.cpp/blob/master/examples/simple/README.md

结构分析

代码结构：

参考

- https://www.bilibili.com/video/BV1Ez4y1w7fc
- https://zhuanlan.zhihu.com/p/665027154