llama2.c

原创 tensor.shape 已于 2024-03-01 17:39:06 修改 阅读量910 收藏 15 点赞数 15 版权

分类专栏： llama 文章标签： 深度学习 语言模型

llama 专栏收录该内容

1、下载模型

从Hugging Face下载中文微型Llama2基础模型，这是一个参数量115M左右的超微型小模型，采用Llama2架构。

2、将模型hf格式转换为bin格式

```
python export.py ./model/chinese-baby-llama2.bin --hf /mnt/workspace/llama2.c/model
```

model 文件夹 中命名一个文件chinese-baby-llama2.bin，chinese-baby-llama2压缩包和解压的都放在model文件夹的

3、debug export.py

修改arg中3代码

```
parser.add_argument("--filepath", type=str,default="/mnt/workspace/llama2.c/model/chinese-baby-llam
group = parser.add_mutually_exclusive_group()
group.add_argument("--hf", type=str,default="/mnt/workspace/llama2.c/model", help="huggingface mode
```

4、几个重点

4.1 model

```
 model = load_hf_model(args.hf)
```

输出的model

```
Transformer(
  (tok_embeddings): Embedding(32000, 768)
  (dropout): Dropout(p=0.0, inplace=False)
  (layers): ModuleList(
    (0): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (1): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (2): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (3): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (4): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (5): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (6): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (7): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (8): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (9): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (10): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
    (11): TransformerBlock(
      (attention): Attention(
        (wq): Linear(in_features=768, out_features=768, bias=False)
        (wk): Linear(in_features=768, out_features=768, bias=False)
        (wv): Linear(in_features=768, out_features=768, bias=False)
        (wo): Linear(in_features=768, out_features=768, bias=False)
        (attn_dropout): Dropout(p=0.0, inplace=False)
        (resid_dropout): Dropout(p=0.0, inplace=False)
      )
      (feed_forward): FeedForward(
        (w1): Linear(in_features=768, out_features=2268, bias=False)
        (w2): Linear(in_features=2268, out_features=768, bias=False)
        (w3): Linear(in_features=768, out_features=2268, bias=False)
        (dropout): Dropout(p=0.0, inplace=False)
      )
      (attention_norm): RMSNorm()
      (ffn_norm): RMSNorm()
    )
  )
  (norm): RMSNorm()
  (output): Linear(in_features=768, out_features=32000, bias=False)
)
```

4.2 legacy_export

用于将模型参数以特定格式保存到二进制文件 中

4.2.1

```
out_file.write(header)
```

构建文件头部信息，其中包括模型的一些参数，如隐藏层 维度、层数、注意力头数等。这些参数被打包成一个结构体，并写入到二进制文件中。

4.2.2 serialize_fp32(out_file, model.tok_embeddings.weight)

将模型的 token embeddings 权重写入二进制文件。

4.2.3 循环

```
for layer in model.layers:
        serialize_fp32(out_file, layer.attention_norm.weight)
    for layer in model.layers:
        serialize_fp32(out_file, layer.attention.wq.weight)
    for layer in model.layers:
        serialize_fp32(out_file, layer.attention.wk.weight)
    for layer in model.layers:
        serialize_fp32(out_file, layer.attention.wv.weight)
    for layer in model.layers:
        serialize_fp32(out_file, layer.attention.wo.weight)
```

几个循环分别处理模型的注意力层和前馈神经网络层的权重，并将它们写入二进制文件。其中，serialize_fp32用于张量以单精度浮点数的格式写入到二进制文件中。

最终生成chinese-baby-llama2.bin

5、run.c

5.1 文件结构

llama2.c》debug22》CMakeLists.txt

CMakeLists.txt如下：

```
cmake_minimum_required(VERSION 3.16)
project(llama2.c)
set(CMAKE_BUILD_TYPE debug) # Debug Release
set(CMAKE_MODULE_PATH ${CMAKE_MODULE_PATH} "${CMAKE_SOURCE_DIR}/")
set(CMAKE_CXX_STANDARD 14)
SET(CMAKE_C_ FLAGS "${ACMAE_C_FLASS} -O0 -ffast-math -manch=native -fopenmp -mavx2 -mfma -DEISEN_ST
SET(CNAKE_CXX_FLAGS "${ACNAKE_CXX_FLASS} -O0 -ffast-math -march=native -fopenmp -mavx2 -mfma -DEITC
add_executable(run /mnt/workspace/llama2.c/run.c)
target_link_libraries(run -lpthread -lm -ldl -m64 -lpthread)
```

5.2 run.c代码

```
int main(int argc, char *argv[]) {

    // default parameters
    char *checkpoint_path = NULL;  // e.g. out/model.bin
    float temperature = 1.0f;   // 0.0 = greedy deterministic. 1.0 = original. don't set higher
    float topp = 0.9f;          // top-p in nucleus sampling. 1.0 = off. 0.9 works well, but slower
    int steps = 256;            // number of steps to run for
    // char *prompt = "NULL";        // prompt string
    char *prompt = "今天是武林大会，我是武林盟主";        // prompt string
    unsigned long long rng_seed = 0; // seed rng with time by default
    char *mode = "generate";    // generate|chat
    char *system_prompt = NULL; // the (optional) system prompt to use in chat mode

    // poor man's C argparse so we can override the defaults above from the command line
    char *tokenizer_path = "/mnt/workspace/llama2.c/model/tokenizer.bin";
    if (argc >= 2) { checkpoint_path = argv[1]; } else { error_usage(); }
```

5.3 开始debug

```
cd /mnt/workspace/llama2.c/debug22
```

5.3.1 编译

```
cmake .
make
```

5.3.2 启动调试器

```
gdb ./run
```

5.3.3 设置断点

```
break main
```

5.3.3 set args

```
set args /mnt/workspace/llama2.c/model/chinese-baby-llama2.bin
```

5.3.4 run起来

```
run>next  ==  r>n
```

5.3.5 生成

```
今天是武林大会，我是武林盟主，也是少林掌门，我们还是拭目以待吧！”一行八人已经呈到击中点，急！这得是多么累的！
```

achieved tok/s: 2.399571

参考链接：
①https://zhuanlan.zhihu.com/p/674666408
②https://github.com/karpathy/llama2.c