

# 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

```
1 | 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

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

## 4、几个重点

### 4.1 model

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

输出的model

```
4 |
5 | Transformer(
6 |   (tok_embeddings): Embedding(32000, 768)
7 |   (dropout): Dropout(p=0.0, inplace=False)
8 |   (layers): ModuleList(
9 |     (0): TransformerBlock(
10 |       (attention): Attention(
11 |         (wq): Linear(in_features=768, out_features=768, bias=False)
12 |         (wk): Linear(in_features=768, out_features=768, bias=False)
13 |         (wv): Linear(in_features=768, out_features=768, bias=False)
```

```
14         (wo): Linear(in_features=768, out_features=768, bias=False)
15         (attn_dropout): Dropout(p=0.0, inplace=False)
16         (resid_dropout): Dropout(p=0.0, inplace=False)
17     )
18     (feed_forward): FeedForward(
19         (w1): Linear(in_features=768, out_features=2268, bias=False)
20         (w2): Linear(in_features=2268, out_features=768, bias=False)
21         (w3): Linear(in_features=768, out_features=2268, bias=False)
22         (dropout): Dropout(p=0.0, inplace=False)
23     )
24     (attention_norm): RMSNorm()
25     (ffn_norm): RMSNorm()
26 )
27 (1): TransformerBlock(
28     (attention): Attention(
29         (wq): Linear(in_features=768, out_features=768, bias=False)
30         (wk): Linear(in_features=768, out_features=768, bias=False)
31         (wv): Linear(in_features=768, out_features=768, bias=False)
32         (wo): Linear(in_features=768, out_features=768, bias=False)
33         (attn_dropout): Dropout(p=0.0, inplace=False)
34         (resid_dropout): Dropout(p=0.0, inplace=False)
35     )
36     (feed_forward): FeedForward(
37         (w1): Linear(in_features=768, out_features=2268, bias=False)
38         (w2): Linear(in_features=2268, out_features=768, bias=False)
39         (w3): Linear(in_features=768, out_features=2268, bias=False)
40         (dropout): Dropout(p=0.0, inplace=False)
41     )
42     (attention_norm): RMSNorm()
43     (ffn_norm): RMSNorm()
44 )
45 (2): TransformerBlock(
46     (attention): Attention(
47         (wq): Linear(in_features=768, out_features=768, bias=False)
48         (wk): Linear(in_features=768, out_features=768, bias=False)
49         (wv): Linear(in_features=768, out_features=768, bias=False)
50         (wo): Linear(in_features=768, out_features=768, bias=False)
51         (attn_dropout): Dropout(p=0.0, inplace=False)
52         (resid_dropout): Dropout(p=0.0, inplace=False)
53     )
54     (feed_forward): FeedForward(
55         (w1): Linear(in_features=768, out_features=2268, bias=False)
56         (w2): Linear(in_features=2268, out_features=768, bias=False)
57         (w3): Linear(in_features=768, out_features=2268, bias=False)
58         (dropout): Dropout(p=0.0, inplace=False)
59     )
60     (attention_norm): RMSNorm()
61     (ffn_norm): RMSNorm()
62 )
63 (3): TransformerBlock(
64     (attention): Attention(
```

```
65         (wq): Linear(in_features=768, out_features=768, bias=False)
66         (wk): Linear(in_features=768, out_features=768, bias=False)
67         (wv): Linear(in_features=768, out_features=768, bias=False)
68         (wo): Linear(in_features=768, out_features=768, bias=False)
69         (attn_dropout): Dropout(p=0.0, inplace=False)
70         (resid_dropout): Dropout(p=0.0, inplace=False)
71     )
72     (feed_forward): FeedForward(
73         (w1): Linear(in_features=768, out_features=2268, bias=False)
74         (w2): Linear(in_features=2268, out_features=768, bias=False)
75         (w3): Linear(in_features=768, out_features=2268, bias=False)
76         (dropout): Dropout(p=0.0, inplace=False)
77     )
78     (attention_norm): RMSNorm()
79     (ffn_norm): RMSNorm()
80 )
81 (4): TransformerBlock(
82     (attention): Attention(
83         (wq): Linear(in_features=768, out_features=768, bias=False)
84         (wk): Linear(in_features=768, out_features=768, bias=False)
85         (wv): Linear(in_features=768, out_features=768, bias=False)
86         (wo): Linear(in_features=768, out_features=768, bias=False)
87         (attn_dropout): Dropout(p=0.0, inplace=False)
88         (resid_dropout): Dropout(p=0.0, inplace=False)
89     )
90     (feed_forward): FeedForward(
91         (w1): Linear(in_features=768, out_features=2268, bias=False)
92         (w2): Linear(in_features=2268, out_features=768, bias=False)
93         (w3): Linear(in_features=768, out_features=2268, bias=False)
94         (dropout): Dropout(p=0.0, inplace=False)
95     )
96     (attention_norm): RMSNorm()
97     (ffn_norm): RMSNorm()
98 )
99 (5): TransformerBlock(
100     (attention): Attention(
101         (wq): Linear(in_features=768, out_features=768, bias=False)
102         (wk): Linear(in_features=768, out_features=768, bias=False)
103         (wv): Linear(in_features=768, out_features=768, bias=False)
104         (wo): Linear(in_features=768, out_features=768, bias=False)
105         (attn_dropout): Dropout(p=0.0, inplace=False)
106         (resid_dropout): Dropout(p=0.0, inplace=False)
107     )
108     (feed_forward): FeedForward(
109         (w1): Linear(in_features=768, out_features=2268, bias=False)
110         (w2): Linear(in_features=2268, out_features=768, bias=False)
111         (w3): Linear(in_features=768, out_features=2268, bias=False)
112         (dropout): Dropout(p=0.0, inplace=False)
113     )
114     (attention_norm): RMSNorm()
115     (ffn_norm): RMSNorm()
```

```
116 )
117 (6): TransformerBlock(
118     (attention): Attention(
119         (wq): Linear(in_features=768, out_features=768, bias=False)
120         (wk): Linear(in_features=768, out_features=768, bias=False)
121         (wv): Linear(in_features=768, out_features=768, bias=False)
122         (wo): Linear(in_features=768, out_features=768, bias=False)
123         (attn_dropout): Dropout(p=0.0, inplace=False)
124         (resid_dropout): Dropout(p=0.0, inplace=False)
125     )
126     (feed_forward): FeedForward(
127         (w1): Linear(in_features=768, out_features=2268, bias=False)
128         (w2): Linear(in_features=2268, out_features=768, bias=False)
129         (w3): Linear(in_features=768, out_features=2268, bias=False)
130         (dropout): Dropout(p=0.0, inplace=False)
131     )
132     (attention_norm): RMSNorm()
133     (ffn_norm): RMSNorm()
134 )
135 (7): TransformerBlock(
136     (attention): Attention(
137         (wq): Linear(in_features=768, out_features=768, bias=False)
138         (wk): Linear(in_features=768, out_features=768, bias=False)
139         (wv): Linear(in_features=768, out_features=768, bias=False)
140         (wo): Linear(in_features=768, out_features=768, bias=False)
141         (attn_dropout): Dropout(p=0.0, inplace=False)
142         (resid_dropout): Dropout(p=0.0, inplace=False)
143     )
144     (feed_forward): FeedForward(
145         (w1): Linear(in_features=768, out_features=2268, bias=False)
146         (w2): Linear(in_features=2268, out_features=768, bias=False)
147         (w3): Linear(in_features=768, out_features=2268, bias=False)
148         (dropout): Dropout(p=0.0, inplace=False)
149     )
150     (attention_norm): RMSNorm()
151     (ffn_norm): RMSNorm()
152 )
153 (8): TransformerBlock(
154     (attention): Attention(
155         (wq): Linear(in_features=768, out_features=768, bias=False)
156         (wk): Linear(in_features=768, out_features=768, bias=False)
157         (wv): Linear(in_features=768, out_features=768, bias=False)
158         (wo): Linear(in_features=768, out_features=768, bias=False)
159         (attn_dropout): Dropout(p=0.0, inplace=False)
160         (resid_dropout): Dropout(p=0.0, inplace=False)
161     )
162     (feed_forward): FeedForward(
163         (w1): Linear(in_features=768, out_features=2268, bias=False)
164         (w2): Linear(in_features=2268, out_features=768, bias=False)
165         (w3): Linear(in_features=768, out_features=2268, bias=False)
166         (dropout): Dropout(p=0.0, inplace=False)
```

```
167         )
168         (attention_norm): RMSNorm()
169         (ffn_norm): RMSNorm()
170     )
171     (9): TransformerBlock(
172         (attention): Attention(
173             (wq): Linear(in_features=768, out_features=768, bias=False)
174             (wk): Linear(in_features=768, out_features=768, bias=False)
175             (wv): Linear(in_features=768, out_features=768, bias=False)
176             (wo): Linear(in_features=768, out_features=768, bias=False)
177             (attn_dropout): Dropout(p=0.0, inplace=False)
178             (resid_dropout): Dropout(p=0.0, inplace=False)
179         )
180         (feed_forward): FeedForward(
181             (w1): Linear(in_features=768, out_features=2268, bias=False)
182             (w2): Linear(in_features=2268, out_features=768, bias=False)
183             (w3): Linear(in_features=768, out_features=2268, bias=False)
184             (dropout): Dropout(p=0.0, inplace=False)
185         )
186         (attention_norm): RMSNorm()
187         (ffn_norm): RMSNorm()
188     )
189     (10): TransformerBlock(
190         (attention): Attention(
191             (wq): Linear(in_features=768, out_features=768, bias=False)
192             (wk): Linear(in_features=768, out_features=768, bias=False)
193             (wv): Linear(in_features=768, out_features=768, bias=False)
194             (wo): Linear(in_features=768, out_features=768, bias=False)
195             (attn_dropout): Dropout(p=0.0, inplace=False)
196             (resid_dropout): Dropout(p=0.0, inplace=False)
197         )
198         (feed_forward): FeedForward(
199             (w1): Linear(in_features=768, out_features=2268, bias=False)
200             (w2): Linear(in_features=2268, out_features=768, bias=False)
201             (w3): Linear(in_features=768, out_features=2268, bias=False)
202             (dropout): Dropout(p=0.0, inplace=False)
203         )
204         (attention_norm): RMSNorm()
205         (ffn_norm): RMSNorm()
206     )
207     (11): TransformerBlock(
208         (attention): Attention(
209             (wq): Linear(in_features=768, out_features=768, bias=False)
210             (wk): Linear(in_features=768, out_features=768, bias=False)
211             (wv): Linear(in_features=768, out_features=768, bias=False)
212             (wo): Linear(in_features=768, out_features=768, bias=False)
213             (attn_dropout): Dropout(p=0.0, inplace=False)
214             (resid_dropout): Dropout(p=0.0, inplace=False)
215         )
216         (feed_forward): FeedForward(
217             (w1): Linear(in_features=768, out_features=2268, bias=False)
```

```

218         (w2): Linear(in_features=2268, out_features=768, bias=False)
219         (w3): Linear(in_features=768, out_features=2268, bias=False)
220         (dropout): Dropout(p=0.0, inplace=False)
221     )
222     (attention_norm): RMSNorm()
223     (ffn_norm): RMSNorm()
224 )
    )
    (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

```

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

```

1 | for layer in model.layers:
2 |     serialize_fp32(out_file, layer.attention_norm.weight)
3 |     for layer in model.layers:
4 |         serialize_fp32(out_file, layer.attention.wq.weight)
5 |         for layer in model.layers:
6 |             serialize_fp32(out_file, layer.attention.wk.weight)
7 |             for layer in model.layers:
8 |                 serialize_fp32(out_file, layer.attention.wv.weight)
9 |                 for layer in model.layers:
10 |                     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：

```

1 | cmake_minimum_required(VERSION 3.16)
2 | project(llama2.c)
3 | set(CMAKE_BUILD_TYPE debug) # Debug Release
4 | set(CMAKE_MODULE_PATH ${CMAKE_MODULE_PATH} "${CMAKE_SOURCE_DIR}/")
5 | set(CMAKE_CXX_STANDARD 14)
6 | SET(CMAKE_C_FLAGS "${ACMAE_C_FLASS} -O0 -ffast-math -march=native -fopenmp -mavx2 -mfma -DEISEN_S1
7 | SET(CNAKE_CXX_FLAGS "${ACNAKE_CXX_FLASS} -O0 -ffast-math -march=native -fopenmp -mavx2 -mfma -DEITC
8 | add_executable(run /mnt/workspace/llama2.c/run.c)
9 | target_link_libraries(run -lpthread -lm -ldl -m64 -lpthread)

```

## 5.2 run.c代码

```

1 | int main(int argc, char *argv[]) {
2 |
3 |     // default parameters
4 |     char *checkpoint_path = NULL; // e.g. out/model.bin
5 |     float temperature = 1.0f; // 0.0 = greedy deterministic. 1.0 = original. don't set higher
6 |     float topp = 0.9f; // top-p in nucleus sampling. 1.0 = off. 0.9 works well, but slower
7 |     int steps = 256; // number of steps to run for
8 |     // char *prompt = "NULL"; // prompt string
9 |     char *prompt = "今天是武林大会，我是武林盟主"; // prompt string
10 |     unsigned long long rng_seed = 0; // seed rng with time by default
11 |     char *mode = "generate"; // generate/chat
12 |     char *system_prompt = NULL; // the (optional) system prompt to use in chat mode
13 |
14 |     // poor man's C argparse so we can override the defaults above from the command line
15 |     char *tokenizer_path = "/mnt/workspace/llama2.c/model/tokenizer.bin";
16 |     if (argc >= 2) { checkpoint_path = argv[1]; } else { error_usage(); }

```

## 5.3 开始debug

```

1 | cd /mnt/workspace/llama2.c/debug22

```

### 5.3.1 编译

```

1 | cmake .
2 | make

```

### 5.3.2 启动调试器

```

1 | gdb ./run

```

### 5.3.3 设置断点

```
1 | break main
```

### 5.3.3 set args

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

### 5.3.4 run起来

```
1 | run>next == r>n
```

### 5.3.5 生成

```
1 | 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：

①<https://zhuanlan.zhihu.com/p/674666408>

②<https://github.com/karpathy/llama2.c>