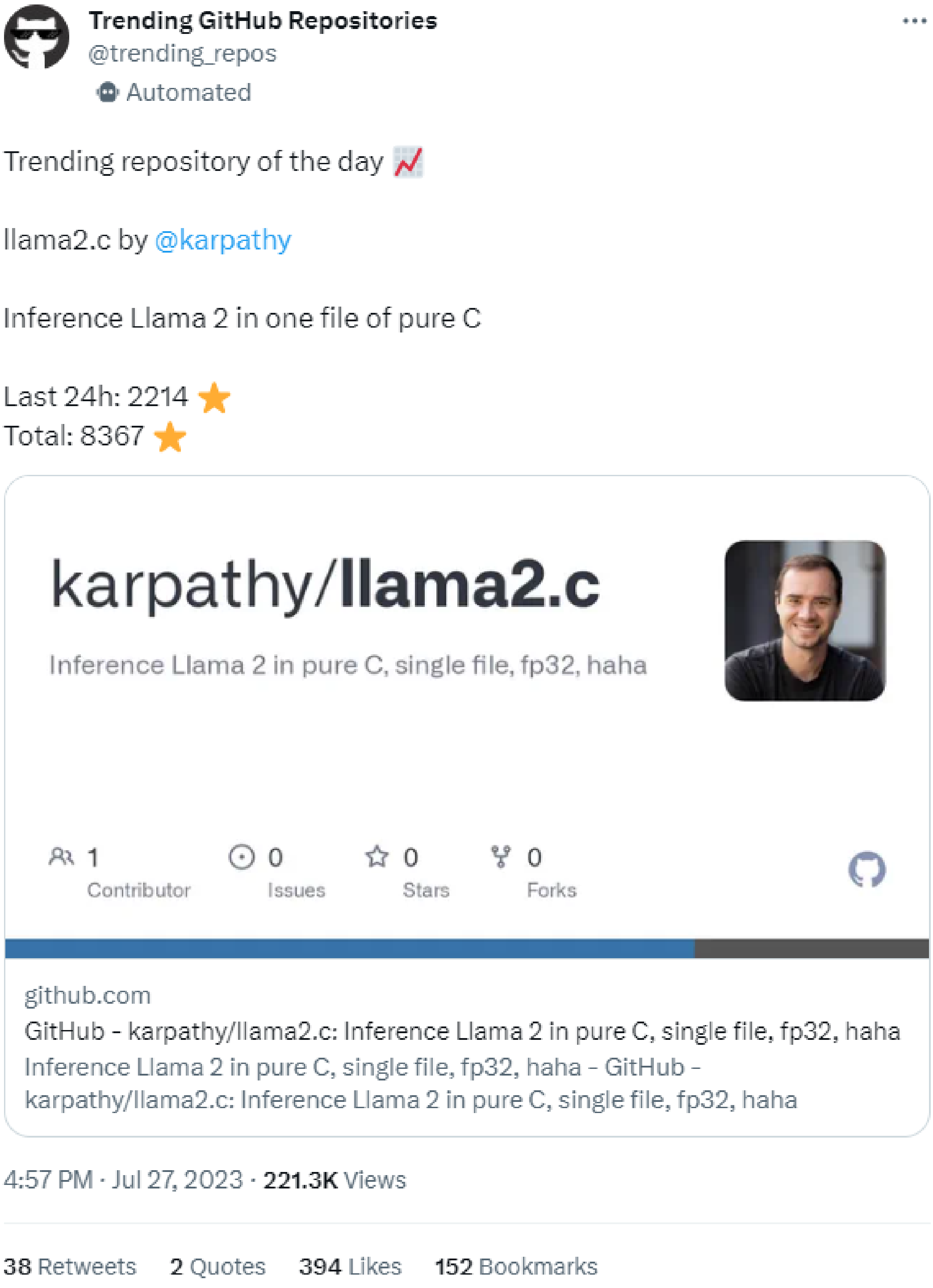
Llama2.c 学习笔记1： 概要&体验 原创

修改于 2023-08-04 00:45:45 2K 0 举报

文章被收录于专栏： 范传康的专栏

llama2.c 还登上了 Github 的热门趋势榜首，最近这周时间花了一点时间研究llama2.c，应景写一个学习笔记吧。



Trending GitHub Repositories
@trending_repos
Automated

Trending repository of the day 📈

llama2.c by @karpathy

Inference Llama 2 in one file of pure C

Last 24h: 2214 ⭐
Total: 8367 ⭐

karpathy/llama2.c

Inference Llama 2 in pure C, single file, fp32, haha

1 Contributor 0 Issues 0 Stars 0 Forks

github.com
GitHub - karpathy/llama2.c: Inference Llama 2 in pure C, single file, fp32, haha
Inference Llama 2 in pure C, single file, fp32, haha - GitHub - karpathy/llama2.c: Inference Llama 2 in pure C, single file, fp32, haha

4:57 PM · Jul 27, 2023 · 221.3K Views

38 Retweets 2 Quotes 394 Likes 152 Bookmarks

1. Why ？

这个repo大火原因有一下几点：

源自meta llama2 模型的大火；经热心群众评测是最接近gpt3.5的开源 LLM ；它是目前唯一可以做到function Call的Open LLM。

源自作者：Andrej Karpathy；他是open ai chatgpt核心成员，热心于 深度学习 AI布道，擅长于深入浅出通过手搓tiny model讲清楚大模型的原理；比如“Let's build GPT: from scratch, in code, spelled out.”有288万的播放量，通过构建一个tiny gpt阐述gpt的基本原理，寓教于行（可运行，完整的架构）。

构建模型、训练、推理、微调一体的框架；相比llama.cpp 专注cpp实现模型的推理；llama2.c有利用基本transformer块构建model参考代码、有tinystory的数据集以及预处理token的代码以及训练的代码；有c实现的推理引擎代码；在学习层面更有价值，并且更具有实际应用的扩展性。

可以转换llama模型为用；llama2.c提供一个脚本可以把meta llama-2模型转为自己能够运行的格式，一方面打开了格局，可以引入外部的模型；另外在模型存储格式转换、运行引擎本质打开了大公司专有的缺口

代码量小；目前基本看就Andrej一个维护者，才几十个commit，代码量不大，可以深入理解，也是学习LLM机制的道路。

目标

Llama2.c涉及LLM微调、模型构建、推理端末部署（量化、硬件加速）等众多方面，是学习研究Open LLM的很好切入点，计划如下：

1）Setup&体验：拉下代码，根据README跑一遍流程，写一个“概要&体验”；

2）Code Understand：打算从模型构建、train过程、模型转换、推理引擎几个方面深入阅读理解；

3）使用其他场景+数据重复过程，加深理解。

希望下半年自由时间更多点。

2.Feel the magic

2.1 编译run.c为推理引擎

```
gcc -O3 -o run run.c -lm
```

以上为ubuntu 22环境，如果是centos环境，

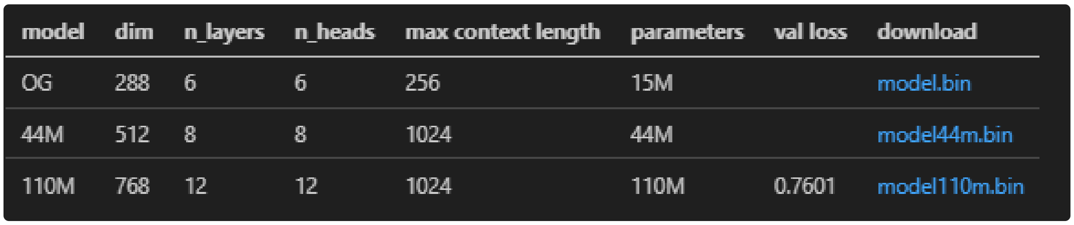
```
gcc -O3 -std=c11 -o run run.c -lm
```

并且run.c修改如下

```
// #include <time.h> for centos
#include <linux/time.h>
```

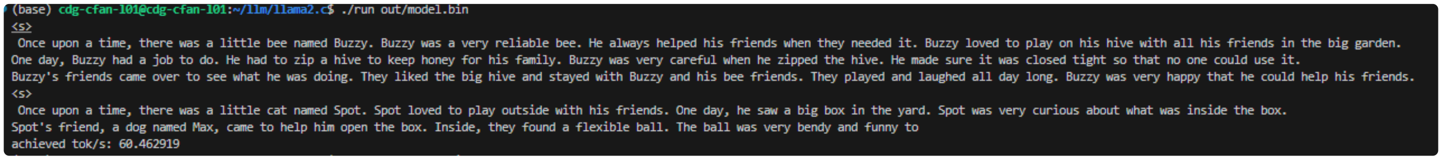
2.2 拉取模型推理

提供OG/15M，44M，110M的基础模型。



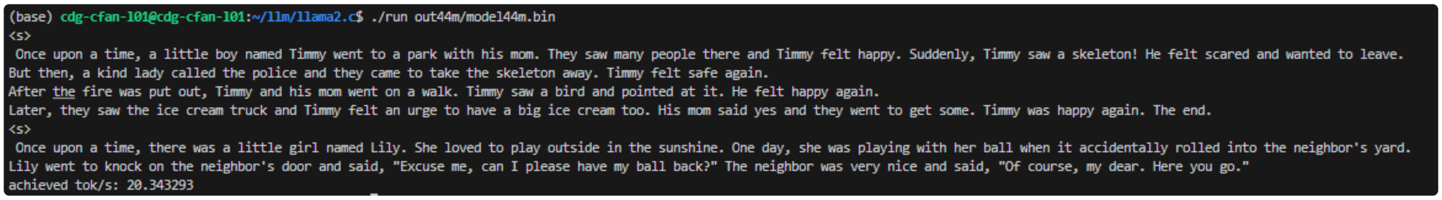
model	dim	n_layers	n_heads	max context length	parameters	val loss	download
OG	288	6	6	256	15M		model.bin
44M	512	8	8	1024	44M		model44m.bin
110M	768	12	12	1024	110M	0.7601	model110m.bin

1 ） 15M Model run



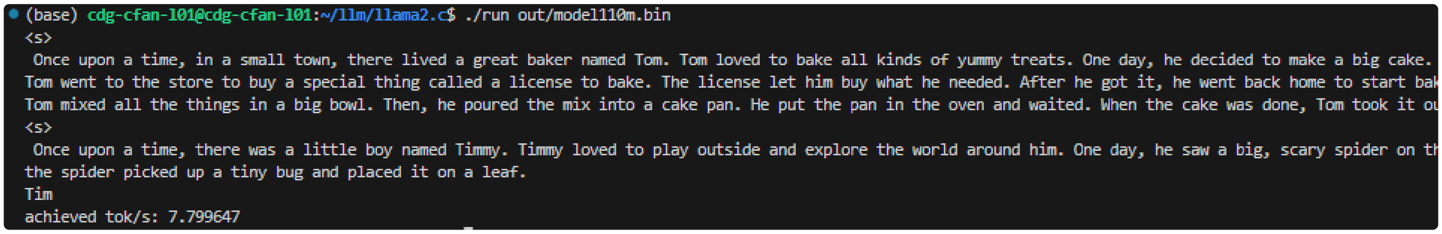
```
(base) cdg-cfan-101@cdg-cfan-101:~/llm/llama2.c$ ./run out/model.bin
<s>
 Once upon a time, there was a little bee named Buzzy. Buzzy was a very reliable bee. He always helped his friends when they needed it. Buzzy loved to play on his hive with all his friends in the big garden.
One day, Buzzy had a job to do. He had to zip a hive to keep honey for his family. Buzzy was very careful when he zipped the hive. He made sure it was closed tight so that no one could use it.
Buzzy's friends came over to see what he was doing. They liked the big hive and stayed with Buzzy and his bee friends. They played and laughed all day long. Buzzy was very happy that he could help his friends.
<s>
 Once upon a time, there was a little cat named Spot. Spot loved to play outside with his friends. One day, he saw a big box in the yard. Spot was very curious about what was inside the box.
Spot's friend, a dog named Max, came to help him open the box. Inside, they found a flexible ball. The ball was very bendy and funny to
achieved tok/s: 60.462919
```

2 ） 44M Model run



```
(base) cdg-cfan-101@cdg-cfan-101:~/llm/llama2.c$ ./run out44m/model44m.bin
<s>
 Once upon a time, a little boy named Timmy went to a park with his mom. They saw many people there and Timmy felt happy. Suddenly, Timmy saw a skeleton! He felt scared and wanted to leave.
But then, a kind lady called the police and they came to take the skeleton away. Timmy felt safe again.
After the fire was put out, Timmy and his mom went on a walk. Timmy saw a bird and pointed at it. He felt happy again.
Later, they saw the ice cream truck and Timmy felt an urge to have a big ice cream too. His mom said yes and they went to get some. Timmy was happy again. The end.
<s>
 Once upon a time, there was a little girl named Lily. She loved to play outside in the sunshine. One day, she was playing with her ball when it accidentally rolled into the neighbor's yard.
Lily went to knock on the neighbor's door and said, "Excuse me, can I please have my ball back?" The neighbor was very nice and said, "Of course, my dear. Here you go."
achieved tok/s: 20.343293
```

3 ） 110M Model run



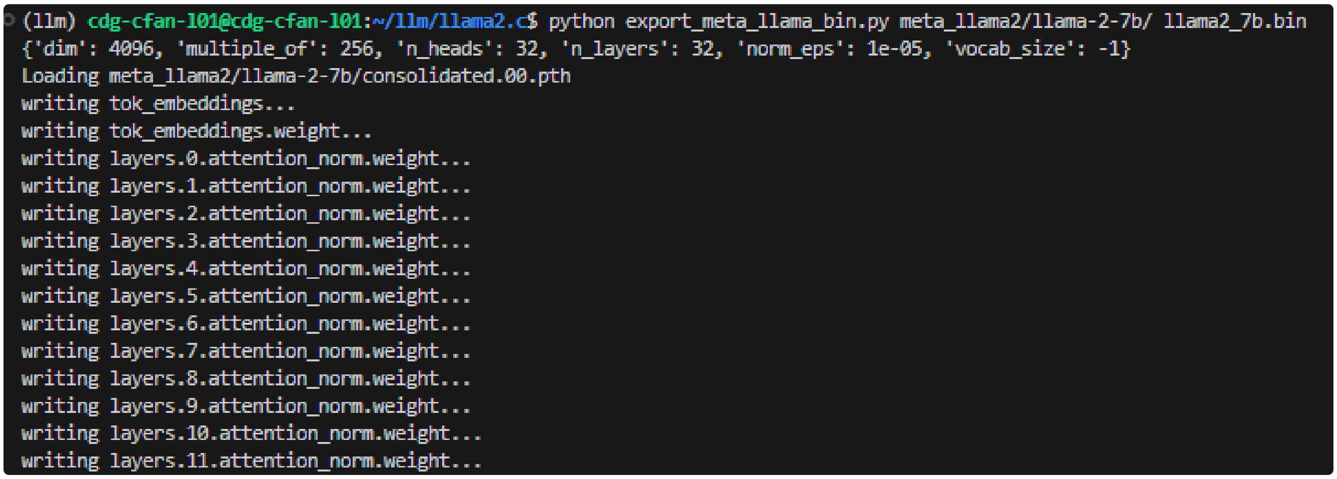
```
(base) cdg-cfan-101@cdg-cfan-101:~/llm/llama2.c$ ./run out/model110m.bin
<s>
 Once upon a time, in a small town, there lived a great baker named Tom. Tom loved to bake all kinds of yummy treats. One day, he decided to make a big cake.
Tom went to the store to buy a special thing called a license to bake. The license let him buy what he needed. After he got it, he went back home to start bak
Tom mixed all the things in a big bowl. Then, he poured the mix into a cake pan. He put the pan in the oven and waited. When the cake was done, Tom took it ou
<s>
 Once upon a time, there was a little boy named Timmy. Timmy loved to play outside and explore the world around him. One day, he saw a big, scary spider on th
the spider picked up a tiny bug and placed it on a leaf.
Tim
achieved tok/s: 7.799647
```

总结

基本都在普通笔记本程跑了出来，token/second 从60到7；这是没有大的优化，以及单线程C跑的结果，未来可期。

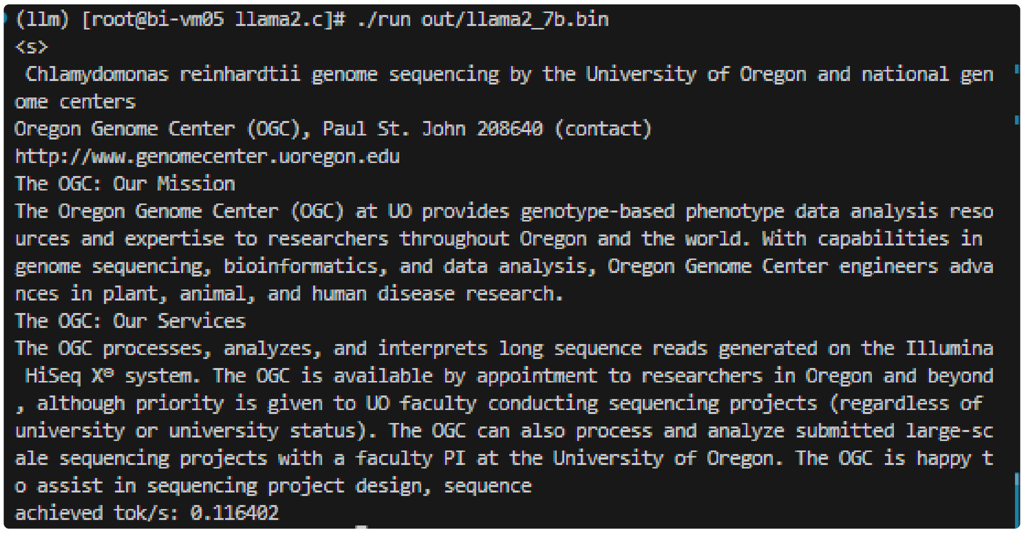
3. 转换模型

转换meta提供的 llama-2-7B模型，可以通过脚本完成



```
(llm) cdg-cfan-101@cdg-cfan-101:~/llm/llama2.c$ python export_meta_llama_bin.py meta_llama2/llama-2-7b/ llama2_7b.bin
{'dim': 4096, 'multiple_of': 256, 'n_heads': 32, 'n_layers': 32, 'norm_eps': 1e-05, 'vocab_size': -1}
Loading meta_llama2/llama-2-7b/consolidated.00.pth
writing tok_embeddings...
writing tok_embeddings.weight...
writing layers.0.attention_norm.weight...
writing layers.1.attention_norm.weight...
writing layers.2.attention_norm.weight...
writing layers.3.attention_norm.weight...
writing layers.4.attention_norm.weight...
writing layers.5.attention_norm.weight...
writing layers.6.attention_norm.weight...
writing layers.7.attention_norm.weight...
writing layers.8.attention_norm.weight...
writing layers.9.attention_norm.weight...
writing layers.10.attention_norm.weight...
writing layers.11.attention_norm.weight...
```

运行（非常慢，大概一分钟一个token，后来换了台强大点机器10s左右一个token）



```
(llm) [root@bi-vm05 llama2.c]# ./run out/llama2_7b.bin
<s>
 Chlamydomonas reinhardtii genome sequencing by the University of Oregon and national gen
ome centers
Oregon Genome Center (OGC), Paul St. John 208640 (contact)
http://www.genomecenter.uoregon.edu
The OGC: Our Mission
The Oregon Genome Center (OGC) at UO provides genotype-based phenotype data analysis reso
urces and expertise to researchers throughout Oregon and the world. With capabilities in
genome sequencing, bioinformatics, and data analysis, Oregon Genome Center engineers adva
nces in plant, animal, and human disease research.
The OGC: Our Services
The OGC processes, analyzes, and interprets long sequence reads generated on the Illumina
 HiSeq X® system. The OGC is available by appointment to researchers in Oregon and beyond
, although priority is given to UO faculty conducting sequencing projects (regardless of
university or university status). The OGC can also process and analyze submitted large-sc
ale sequencing projects with a faculty PI at the University of Oregon. The OGC is happy t
o assist in sequencing project design, sequence
achieved tok/s: 0.116402
```

4.其他

训练比较麻烦，开始python 3.11不支持torch compile，转为python 3.9；然后没有gpu不支持cuda，又mem比较小；目前还没有完整跑出来一个。

资源

1）Let's build GPT: from scratch, in code, spelled out. - YouTube

2）karpathy/llama2.c: Inference Llama 2 in one file of pure C (github.com)