gpt-2

This is a C++ example running GPT-2 inference using the ggml library.

- 这是一个使用ggml库运行GPT-2推理的C++例子
- 这个例子中使用了ggml库
- 这个例子中用C++来运行GPT-2的推理

The program runs on the CPU - no video card is required.

- 这个程序运行在CPU上，不需要使用显卡

The Cerebras-GPT models are also supported.

- Cerebras-GPT模型也是可以被支持的
- 使用这个程序例子可以直接运行Cerebras-GPT模型

The example supports the following GPT-2 models:

- 这个示例支持下列的GPT-2 models

Model	Description	Disk Size
117M	Small model	240 MB
345M	Medium model	680 MB
774M	Large model	1.5 GB
1558M	XL model	3.0 GB

Sample performance on MacBook M1 Pro（MacBook M1 Pro 上的样本性能）:

Model	Size	Time / Token
GPT-2	117M	5 ms
GPT-2	345M	12 ms
GPT-2	774M	23 ms
GPT-2	1558M	42 ms

TODO: add tables for Cerebras-GPT models

Sample output:

```
$ ./bin/gpt-2 -h
usage: ./bin/gpt-2 [options]

options:
  -h, --help            show this help message and exit
```

```
  -s SEED, --seed SEED  RNG seed (default: -1)
  -t N, --threads N     number of threads to use during computation (default: 8)
  -p PROMPT, --prompt PROMPT
                        prompt to start generation with (default: random)
  -n N, --n_predict N   number of tokens to predict (default: 200)
  --top_k N             top-k sampling (default: 40)
  --top_p N             top-p sampling (default: 0.9)
  --temp N              temperature (default: 1.0)
  -b N, --batch_size N  batch size for prompt processing (default: 8)
  -m FNAME, --model FNAME
                        model path (default: models/gpt-2-117M/ggml-model.bin)

$ ./bin/gpt-2
gpt2_model_load: loading model from 'models/gpt-2-117M/ggml-model.bin'
gpt2_model_load: n_vocab = 50257
gpt2_model_load: n_ctx   = 1024
gpt2_model_load: n_embd  = 768
gpt2_model_load: n_head  = 12
gpt2_model_load: n_layer = 12
gpt2_model_load: f16     = 1
gpt2_model_load: ggml ctx size = 311.12 MB
gpt2_model_load: memory size =    72.00 MB, n_mem = 12288
gpt2_model_load: model size  =   239.08 MB
main: number of tokens in prompt = 1

So this is going to be the end of the line for us.

If the Dolphins continue to do their business, it's possible that the team could
make a bid to bring in new defensive coordinator Scott Linehan.

Linehan's job is a little daunting, but he's a great coach and an excellent coach.
I don't believe we're going to make the playoffs.

We're going to have to work hard to keep our heads down and get ready to go.
<|endoftext|>

main: mem per token =  2048612 bytes
main:     load time =   106.32 ms
main:   sample time =     7.10 ms
main:  predict time =   506.40 ms / 5.06 ms per token
main:    total time =   629.84 ms
```

Downloading and converting the original models (GPT-2)（下载和对original models进行转化）

You can download the original model files using the download-model.sh Bash script. The models are in Tensorflow format, so in order to use them with ggml, you need to convert them to appropriate format. This is done via the convert-ckpt-to-ggml.py python script.

- 可以使用download-model.sh脚本文件来下载original model files
- 经过download-model.sh脚本文件下载下来的original model files是tensorflow format

- 为了用ggml调用这些tensorflow format格式的文件，那么需要将这些tensorflow format格式的文件转化为合适的format
- 可以使用convert-ckpt-to-ggml.py脚本文件来进行转化

Here is the entire process for the GPT-2 117M model (download from official site + conversion):

- 1、从官方网站下载GPT-2 117M model
- 2、将下载下来的GPT-2 117M model转化格式

```
cd ggml/build
../examples/gpt-2/download-model.sh 117M

Downloading model 117M ...
models/gpt-2-117M/checkpoint                        100%
[============================>]      77  --.-KB/s    in 0s
models/gpt-2-117M/encoder.json                      100%
[============================>]   1018K  1.20MB/s    in 0.8s
models/gpt-2-117M/hparams.json                      100%
[============================>]      90  --.-KB/s    in 0s
models/gpt-2-117M/model.ckpt.data-00000-of-00001  100%
[============================>] 474.70M  1.21MB/s    in 8m 39s
models/gpt-2-117M/model.ckpt.index                  100%
[============================>]   5.09K  --.-KB/s    in 0s
models/gpt-2-117M/model.ckpt.meta                   100%
[============================>] 460.11K   806KB/s    in 0.6s
models/gpt-2-117M/vocab.bpe                         100%
[============================>] 445.62K   799KB/s    in 0.6s
Done! Model '117M' saved in 'models/gpt-2-117M/'

Run the convert-ckpt-to-ggml.py script to convert the model to ggml format.

  python /Users/john/ggml/examples/gpt-2/convert-ckpt-to-ggml.py models/gpt-2-
117M/ 1
```

This conversion requires that you have python and Tensorflow installed on your computer. Still, if you want to avoid this, you can download the already converted ggml models as described below.

- 使用convert-ckpt-to-ggml.py对model files进行转化
- 在本地环境中需要由python和tensorflow
- 如果不想进行转化步骤，那么直接下载使用已经转化号的ggml models

Downloading and converting the original models (Cerebras-GPT)

Clone the respective repository from here: https://huggingface.co/cerebras

Use the convert-cerebras-to-ggml.py script to convert the model to `ggml` format:

```
cd ggml/build
git clone https://huggingface.co/cerebras/Cerebras-GPT-111M models/
```

```
python ../examples/gpt-2/convert-cerebras-to-ggml.py models/Cerebras-GPT-111M/
```

Downloading the ggml model directly (GPT-2)

For convenience, I will be hosting the converted ggml model files in order to make it easier to run the examples. This way, you can directly download a single binary file and start using it. No python or Tensorflow is required.

- 为了方便这里使用已经经过转化的ggml model file
- 可以直接下载对应的文件并且进行使用

Here is how to get the 117M ggml model:

```
cd ggml/build
../examples/gpt-2/download-ggml-model.sh 117M

Downloading ggml model 117M ...
models/gpt-2-117M/ggml-model.bin         100%[===============================>]
239.58M  8.52MB/s    in 28s
Done! Model '117M' saved in 'models/gpt-2-117M/ggml-model.bin'
You can now use it like this:

  $ ./bin/gpt-2 -m models/gpt-2-117M/ggml-model.bin -p "This is an example"
```

At some point, I might decide to stop hosting these models. So in that case, simply revert to the manual process above.

Quantizing the models（量化models）

You can also try to quantize the `ggml` models via 4-bit integer quantization.（可以通过4-bit整数量化的方式进行量化） Keep in mind that for smaller models, this will render them completely useless.（对于小模型而言，完全是没有必要的，一般来说你想要量化更大的模型） You generally want to quantize larger models.

```
# quantize GPT-2 F16 to Q4_0 (faster but less precise)
./bin/gpt-2-quantize models/gpt-2-1558M/ggml-model-f16.bin models/gpt-2-
1558M/ggml-model-q4_0.bin 2
./bin/gpt-2 -m models/gpt-2-1558M/ggml-model-q4_0.bin -p "This is an example"

# quantize Cerebras F16 to Q4_1 (slower but more precise)
./bin/gpt-2-quantize models/Cerebras-GPT-6.7B/ggml-model-f16.bin models/Cerebras-
GPT-6.7B/ggml-model-q4_1.bin 3
./bin/gpt-2 -m models/Cerebras-GPT-6.7B/ggml-model-q4_1.bin -p "This is an
example"
```

Batched generation example（批量生成示例）

You can try the batched generation from a given prompt using the gpt-2-batched binary.（你可以使用 gpt-2-batched 二进制文件尝试根据给定的提示进行批量生成。）

Sample output:

```
$ gpt-2-batched -np 5 -m models/gpt-2-117M/ggml-model.bin -p "Hello my name is" -n
50

main: seed = 1697037431
gpt2_model_load: loading model from 'models/gpt-2-117M/ggml-model.bin'
gpt2_model_load: n_vocab = 50257
gpt2_model_load: n_ctx   = 1024
gpt2_model_load: n_embd  = 768
gpt2_model_load: n_head  = 12
gpt2_model_load: n_layer = 12
gpt2_model_load: ftype   = 1
gpt2_model_load: qntvr   = 0
gpt2_model_load: ggml tensor size    = 320 bytes
gpt2_model_load: backend buffer size = 312.72 MB
ggml_init_cublas: found 1 CUDA devices:
  Device 0: NVIDIA GeForce GTX 1660, compute capability 7.5
gpt2_model_load: using CPU backend
gpt2_model_load: memory size =    72.00 MB, n_mem = 12288
gpt2_model_load: model size  =   239.08 MB
extract_tests_from_file : No test file found.
test_gpt_tokenizer : 0 tests failed out of 0 tests.
main: compute buffer size: 3.26 MB


main: generating 5 sequences ...
main: prompt: 'Hello my name is'
main: number of tokens in prompt = 4, first 8 tokens: 15496 616 1438 318


sequence 0:

Hello my name is John. You can call me any way you want, if you want, but for my
very first date, I will be on the phone with you. We're both in our early 20s, but
I feel like it's all

sequence 1:

Hello my name is Robert, and I want to say that we're proud to have your company
here on the world's largest platform for sharing your stories with us. This is a
huge opportunity for our community. We have hundreds of people on this team and

sequence 2:

Hello my name is Jack. I'm the one who created you.
```

```
Jack is a boy with a big smile and a big heart. He is a handsome guy. He loves the
outdoors and loves the people he meets. He wants to be a

sequence 3:

Hello my name is John. I am a Canadian citizen with a large number of family in
Quebec and I am interested in studying. My aim is to take up a post in the Journal
of the International Academy of Sciences of Canada which I am currently finishing.

sequence 4:

Hello my name is Dan. I am an entrepreneur. I am a great father. I am a great
husband. I am a great husband. I am a great dad. And I am a great husband.

I love my life. I love



main:     load time =   880.80 ms
main:   sample time =    91.43 ms
main:  predict time =  2518.29 ms
main:    total time =  3544.32 ms
```