Day4 - 從nanoGPT開始 (3)

15th鐵人賽

jjchen1
團隊 我在鐵人賽烙賽、也在外木山裸泳◑ω◐

2023-09-05 23:25:10
308 瀏覽

開始使用NanoGPT專案來做Training，我研究了一下專案的內容，發現雖然NanoGPT是以教學為導向的但是似乎還是有點大，不知道單張顯卡跑不跑得動最小的模型。

不過由於作者專門為那些只是想體驗一下自己訓練GPT的過程的人們提供了一個選擇，那就是像前面的Bigram Language Model訓練一樣，拿Shakespeare文章集來訓練一個字元級的GPT。

> If you are not a deep learning professional and you just want to feel the magic and get your feet wet, the fastest way to get started is to train a character-level GPT on the works of Shakespeare.

- 首先安裝需要的Package

```
pip install torch numpy transformers datasets tiktoken wandb tqdm
```

- 資料前處理

```
python data/shakespeare_char/prepare.py
```

- 開始訓練作者準備的玩具級GPT

> If you peek inside it, you'll see that we're training a GPT with a context size of up to 256 characters, 384 feature channels, and it is a 6-layer Transformer with 6 heads in each layer.

```
python train.py config/train_shakespeare_char.py
```

- 一下就訓練完可以看結果了，雖然是字元單位的模型且只訓練了5000個iteration，但就算內容怪怪的但字大多都是拼對的，錯字數量已經頗少了。

```
python sample.py --out_dir=out-shakespeare-char
```

明天除了試著用OpenWebText訓練以外（NanoGPT正規使用的文本內容），我也想做個練習用的實驗，自己產生一個訓練集，字典內容是 `0123456789.()+-*/=` ，因為可以無限生成數據，看GPT是否能夠學會讀文字做加減乘除並且考慮到括號以及先乘除後加減的規則。不過今天有點忙來不及弄，明天再弄。