

Q&A using Traditional Chinese Large Language model

學號 : 311356003

學生 : 陳澤昕

Dataset

- 使用原本Kaggle 提供的資料外，額外擴增了兩筆資料
 - <https://github.com/ntunlplab/traditional-chinese-alpaca/tree/main/data>
 - <https://github.com/A-baoYang/alpaca-7b-chinese/tree/main/data/general>

```
{
  "instruction": "Answer the given multiple choice question.",
  "input": "6月1號開始，負責總統府維安勤務的憲兵全面換裝，憲兵211營營長黃厚泰表示，新的勤務",
  "output": "1"
},
{
  "instruction": "Answer the given multiple choice question.",
  "input": "6月1號開始，負責總統府維安勤務的憲兵全面換裝，憲兵211營營長黃厚泰表示，新的勤務",
  "output": "3"
},
{
  "instruction": "Answer the given multiple choice question.",
  "input": "國人愛吃海鮮，去年整年進口食用貝類高達8500公噸，為加強食用貝類輸入源頭管理，食",
  "output": "2"
},
{
```

Git 上的檔名: alpaca-tw_en-align.json

Code

- `datapre.ipynb` : 進行資料處理
- `lora_finetune.py` : 進行Finetune
- `lora_generate2.py` : 產生結果
- `Bash.py` : 執行後就會跑完整個流程並產生結果(使用四卡進行運算)
- `Adapter_model.bin` : 放在tag(Final_model)裡面

Finetune Hyperparameter

```
base_model: str = 'decapoda-research/llama-7b-hf', # the only required argument
data_path: str = "/data/tzeshinchen/deep_learning/kaggle/lora/alpaca-lora/Origin.json",
output_dir: str = ("./lora-alpaca_extended"),
batch_size: int = 256,
micro_batch_size: int = 4,
num_epochs: int = 3,
learning_rate: float = 2e-4,
cutoff_len: int = 100,
val_set_size: int = 100,
# lora hyperparams
```

```
lora_r: int = 8,
lora_alpha: int = 16,
lora_dropout: float = 0.05,
lora_target_modules: List[str] = ["q_proj", "v_proj", "k_proj", "o_proj"],
train_on_inputs: bool = True, # if False, masks out inputs in loss
add_eos_token: bool = False,
group_by_length: bool = True, # wandb params
resume_from_checkpoint: str = ("/data/tzeshinchen/deep_learning/kaggle/lora/alpaca-lora/lora-alpaca_extended")
```

調整了 單次batch_size = 256

Epoch = 3

lora_target_modules :

使用“q_proj” “v_proj” “k_proj” “o_proj”

用台大先pretrain 好的中文模型

Generator

Alpaca-LoRA

Alpaca-LoRA is a 7B-parameter LLaMA model finetuned to follow instructions. It is trained on the [Stanford Alpaca](#) dataset and makes use of the Huggingface LLaMA in

Instruction

根據以下敘述回答問題選擇一個選項:
尹雪艷非常迷人, 名氣大了, 難免招忌。一些同行的姊妹看了眼紅, 就到處放話: 尹雪艷的八字帶著重煞, 犯了白虎, 沾上的人, 輕者敗家, 重者人亡。但是這一點反而增加了她的神秘感, 讓她名氣更響亮。
問題:
請問根據此段文章, 下面哪一個選項與「尹雪艷八字犯重煞」無關? (1)增加尹雪艷神秘感(2)同行因忌妒而放話 (3)會讓親近者家敗人亡(4)忠孝仁愛信義和平

Input

none

Temperature 0.1

Top p 0.75

Top k 3

Beams 4

Max tokens 129

☐ Stream output

Clear Submit

```
def evaluate(  
    instruction,  
    input=None,  
    temperature=0.1,  
    top_p=0.75,  
    top_k=2,  
    num_beams=4,  
    max_new_tokens=1,  
    stream_output=False,  
    **kwargs,  
):
```

Flag

控制生成的多樣性。值越大則生成的多樣性越大, 但可能會降低生成文本的質量。

輸出文本中概率累計超過 **top_p** 的最小概率值的單詞被剔除, 保留剩下的單詞作為生成文本的候選集。值越大多樣性越小, 但準確率比較高

限制生成時從所有可能的token中選擇的數量, 保留最有可能的**top_k**個token。值越小則生成的多樣性越小, 但可能會提高生成文本的質量。

Beam search size

生成文本的最大長度限制。

Loss

```
warnings.warn('Matplotlib: Inputs will be cast from {A.dtype} to float64')
{'loss': 1.7864, 'learning_rate': 2e-05, 'epoch': 0.19}
{'loss': 1.4509, 'learning_rate': 4e-05, 'epoch': 0.38}
{'loss': 1.3631, 'learning_rate': 6e-05, 'epoch': 0.57}
{'loss': 1.3183, 'learning_rate': 8e-05, 'epoch': 0.76}
{'loss': 1.2894, 'learning_rate': 0.0001, 'epoch': 0.95}
{'loss': 1.2577, 'learning_rate': 0.00012, 'epoch': 1.14}
{'loss': 1.2339, 'learning_rate': 0.00014, 'epoch': 1.33}
{'loss': 1.2033, 'learning_rate': 0.00016, 'epoch': 1.52}
{'loss': 1.1849, 'learning_rate': 0.00018, 'epoch': 1.71}
{'loss': 1.1452, 'learning_rate': 0.0002, 'epoch': 1.9}
{'loss': 1.1176, 'learning_rate': 0.00016428571428571428, 'epoch': 2.09}
{'loss': 1.1036, 'learning_rate': 0.00012857142857142858, 'epoch': 2.28}
{'loss': 1.0722, 'learning_rate': 9.285714285714286e-05, 'epoch': 2.47}
{'loss': 1.0616, 'learning_rate': 5.714285714285714e-05, 'epoch': 2.66}
{'loss': 1.0486, 'learning_rate': 2.1428571428571428e-05, 'epoch': 2.85}
{'train_runtime': 10215.1039, 'train_samples_per_second': 3.95, 'train_steps': 10000}
```

還沒完全收斂，應該可以在繼續往下train

Result

15

311356003



0.33333

8

9d



Your Best Entry!

Your submission scored 0.01333, which is not an improvement of your previous score. Keep trying!

16

▼ 1

311356003



0.36285

8

9d

Conclusion

- 我認為這次的準確率只有33%，是因為這次使用的運算資源非常大量，為了要滿足硬體的Ram 所以會將cutoff 設定到128，但是她一條句字最長1548，會導致準確率下降。
- 另外依照learning rate的下降我認為模型還沒有充分收斂握，eppoch就到了應該還可以再讓他繼續跑下去，進一步提高準確率。