

Applied Deep Learning HW2

林宇萱 r13922159

Q1: Model (2%)

● Model

我使用 google/mt5-small (Multilingual T5) model 來完成這次的 Summarization task。T5 (Text-T-Text Transfer Transformer) pre-trained model，架構主要由 Encoder 和 Decoder 組成，其中每一層包含 self-attention 和 cross-attention 機制。在 summarization task 中，將欲做 summarization 的 text 作為 input 通過 Encoder，encoder 會生成包含上下文關係的 representation，再經過 Decoder 的 cross-attention 機制讓 model 可以參考 input text 中的資訊，再逐字生成 output。

● Preprocessing

- 將原 training data 分為 training data 18454 筆(85%)與 validation data 3256 筆(15%)
- 將 maintext 與 title 做 tokenization，並分別設定 truncated to length=1024 和 128
- 準備 compute_metrics function，用來在 training 過程中做 evaluation 時評估當下 model performance，其中使用助教提供的 tw_rouge 來計算 rouge score

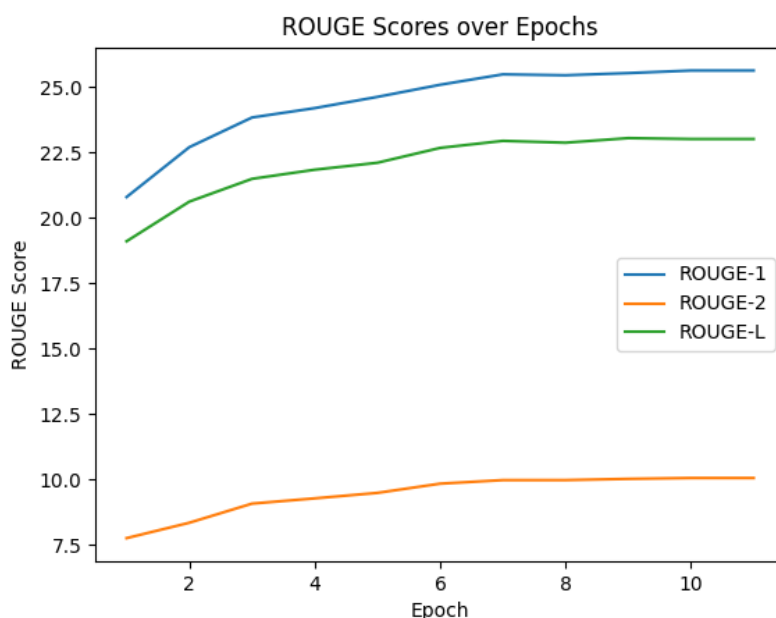
Q2: Training (2%)

● Hyperparameter

- Optimizer: AdamW
- Learning rate: 5e-5
- Weight decay: 0.01
- Batch size: 16 (gradient_accumulation_steps:8 ; per_device_train_batch_size: 2)
- Epochs: 10

● Learning Curves

每個 epoch 進行一次 evaluation，使用 greedy 作為 generation strategy，計算出來的 rouge score 採計 $f1\text{-score} \times 100$ 記錄。



Q3: Generation Strategies(6%)

● Strategies (2%)

- Greedy：每次選機率最高的 token，直到整個 sequence 生成完成。簡單快速，但生成的可能不是 global optimal，容易卡在 local optimal。
- Beam search：每次考慮 num_beams 個最佳選擇，直到最後選擇機率最高的那條完成 sequence。比起 Greedy，比較有機會找到 global optimal，但所需運算資源較高。
- Top-k sampling：每次選 k 個機率最高的 token，並從當中隨機選一個。可以使生成的 sequence 較有變化，避免每次只選最高機率的 token。增加 k 值可以增加生成的多樣性，減少 k 值可以增加生成的通用性。
- Top-p sampling：設定 token 累積的機率達到 p 時，才有機會被隨機選擇並生成。可以使生成有變化又避免選到機率非常低的 token。
- Temperature：控制隨機性、多樣性的參數，較低的 temperature 使生成更包保守、傾向選高機率的 token，而較高的 temperature 可以增加生成的多樣性，讓低機率的 token 更有可能被選擇。

- Hyperparameters:

- Try at least 2 settings of each strategies and compare the result.

- ◆ rouge score 採計 f1-score*100 記錄

- ◆ on public dataset

Strategy	Rouge-1	Rouge-2	Rouge-l
Greedy	23.72	8.83	21.37
Beam Search (num_beams=3)	24.91	9.85	22.40
Beam Search (num_beams=5)	25.1	10.05	22.49
Top-k sampling (k=5)	21.92	7.52	19.51
Top-k sampling (k=10)	20.99	7.05	18.67
Top-p sampling (p=0.3)	22.01	7.94	19.7
Top-p sampling (p=0.5)	20.20	7.09	18.05
Temperature (temperature=0.9)	15.88	4.94	14.22
Temperature (temperature=1.1)	11.04	3.00	9.96

- What is your final generation strategy? (you can combine any of them)
根據以上 results，最後我選擇 Beam search with num beams=5 作為 generation strategy。

Reference：

<https://huggingface.co/learn/nlp-course/en/chapter7/5>

<https://github.com/deankuo/ADL24-HW2>