ADL HW2

B10915024 王程煜

Q1: Model (2%)

Model (1%)

T5 模型是一個經過大量預訓練資料訓練而成的 encoder-decoder 模型。對於每次生成的字詞，它使用<BOS>或前一個生成的字詞來進行下一個字詞的機率預測，可以透過選取最高機率或使用其他技巧實現。在 text summarization 的任務中，我們有一個明確的輸入 x = document 和輸出 y = summary，因此我們只需 finetune T5 模型，以使其能夠生成 summary。此次作業所使用的 mT5 是 T5 模型的一個變種，經過多語言的訓練，可以應對多種語言的生成任務。

Preprocessing (1%)

在一開始得到的資料中有 date_publish、title、source_domain、maintext、split 共五個欄位，我先將對於訓練沒有任何幫助的 date_publish、source_domain、split 三欄刪去才餵入模型。送入模型前還會先經過 mt5 自帶的 tokenizer 進行 tokenization。

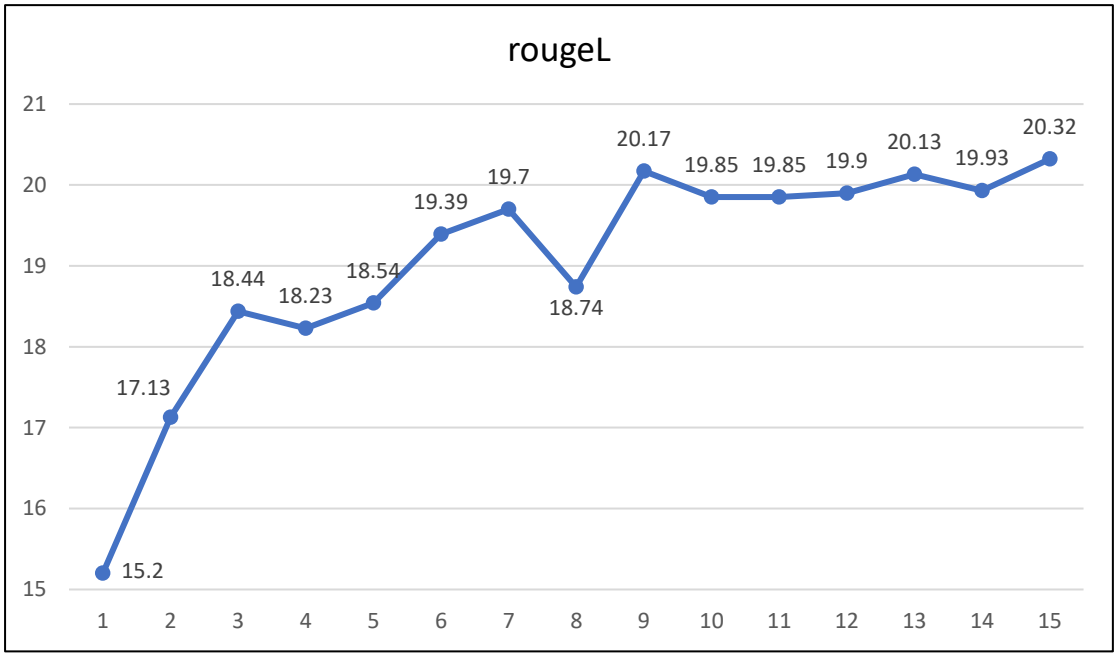
Q2: Training (2%)

Hyperparameter (1%)

- fp16、AdaFactor、learning_rate、lr_scheduler
  因為 T5 模型的 fintune 比較偏向大範圍的方向，且參考了 hugging face 有關 T5 頁面的說明及論壇討論結果，最後將模型套用較低精度的 fp16、可自適應 lr 的 Adafactor、learning_rate 調整成 3e-4、將 lr_scheduler 移除。
  論壇頁面：T5 Finetuning Tips - Models - Hugging Face Forums
- max_source_length、max_target_length
  依據助教提供的 baseline 提示分別使用 256 及 64 長度。
- per_device_train_batch_size、gradient_accumulation_steps
  因為從 hugging face 的範例 code 可知預設 batch size 為 8，所以我透過調整這兩項參數分別為 4 和 2 來達到 8 的大小，不突破顯卡記憶體上限的同時保留效能。
- num_train_epochs 15
  在上述的參數設置下，所能在四小時內跑完的最大 epoch。
- num_beams 4
  eval 時產生字詞使用的 beams search 長度。

Learning Curves (1%)

Epoch = 15



Q3: Generation Strategies(6%)

Strategies (2%)

- Greedy
  永遠選擇最高機率的字當作下一個輸出。
- Beam Search
  使用時會傳入一個值 k，代表會從生成樹往下走幾步，拿取總共的機率最高的那一個路徑，解決掉 Greedy 可能出現無意義字詞的情況。
- Top-k Sampling
  使用時會傳入一個值 k，代表從生成字候選單內拿取前 k 高機率的字來 sample。
- Top-p Sampling
  使用時會傳入一個值 p，代表從最高開始累加候選字的機率值，直到不超過 p 值。
- Temperature
  用於調整隨機程度的數字。

Hyperparameters (4%)

Beam_search	2		3		4	
top_p	0.7	0.85	0.7	0.85	0.7	0.85
Rouge-1	25.65	25.64	25.92	25.92	25.98	25.98
Rouge-2	9.87	9.87	10.06	10.06	10.15	10.15
Rouge-l	23.00	23.01	23.22	23.22	23.22	23.22

由上表的交叉比對後可知，beam size 為 4 且 top_p 為 0.85 為最佳狀態。

最後參數：

max_length = 64, beam_size = 4, top_k = 20 --top_p = 0.85