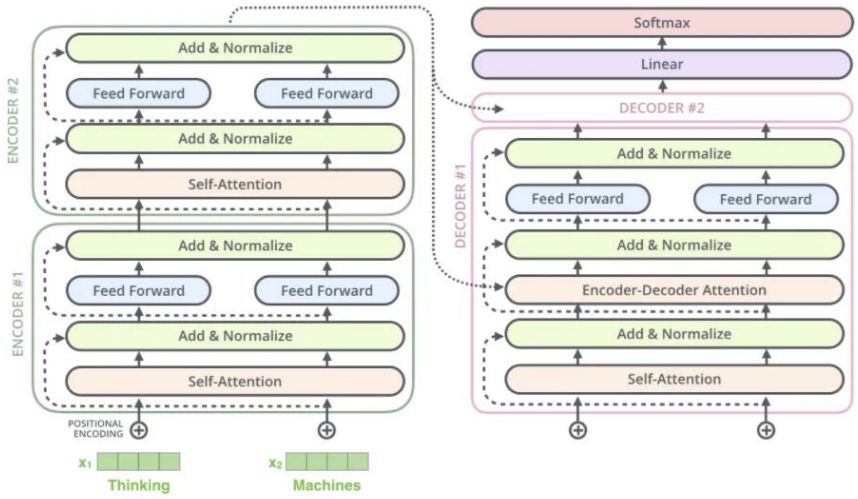
Q1. Model

(a) Model

Multilingual T5 由 [mT5: A massively multilingual pre-trained text-to-text transformer]此 paper 提出。mT5 為一多國語言 seq2seq 的模型，其架構與 T5 一樣，如下圖:



(source: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51)

mT5 預訓練在 multilingual C4 dataset 上，可以使用在 101 種不同的語言上。因為 mt5 是 text-to-text 的模型，可以用在 text-to-text 的 generation 上，而 summarization 可以視為一種 text generation。

(b) Preprocessing

Tokenization 使用的是 huggingface 的 mt5 Tokenizer。Tokenizer 會將中文轉成 token 然後再用 vocab 轉成對應的 id，也會加上<UNK>，<EOS>與<PAD>特殊 token 到 sequence 裡。

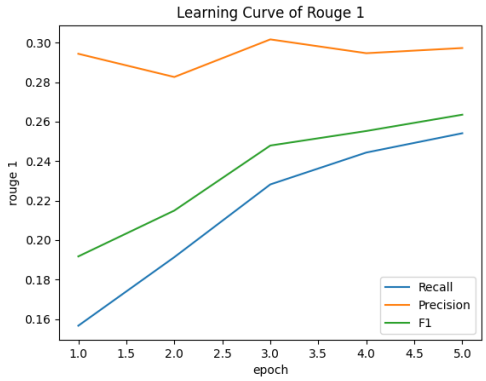
Q2. Training

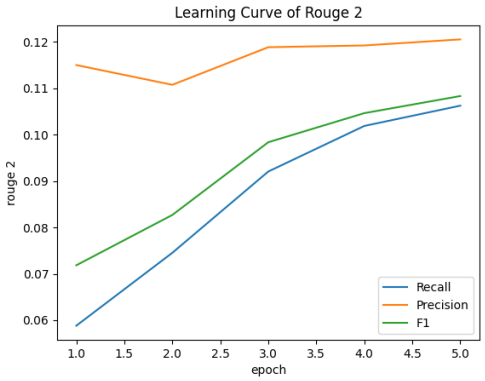
(a) Hyperparameter

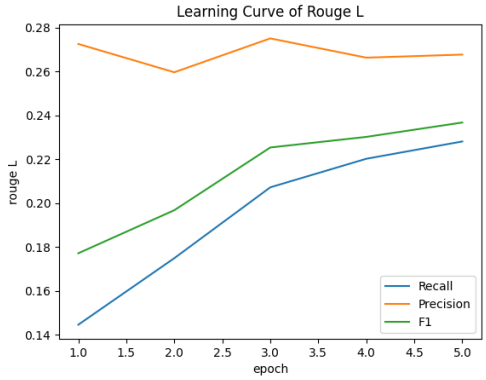
- Batch size : 1
- Gradient accumulate step : 2
- Learning rate: 5e-5
- Epoch: 5

當 batch size 與 step 再提升時，GPU 可能會有 out of memory 的狀況。Learning Rate = 5e-5 與 Epoch = 5 可以在四個小時內 Train 完。

(b) Learning Curve







Q3. Generation Strategy

(a) Strategies

- Greedy

  Greedy 選擇下一個機率最高的文字生成。

- Beam Search

  Beam Search 每次更新文字記下 N 個最有可能的路線，最後選擇整體機率最高的路線做生成。

- Top-K

  Top-K 會從 K 個最高的機率的文字中做取樣，生成下一個可能的文字。

- Top-P

  Top-P 會從 N 個最高的機率的文字加起來為 P 的機率中做取樣，生成下一個可能的文字。

- Temperature

  Temperature 生成下個文字的機率如下:

  $$P(x_i|x_{1:i-1}) = \frac{\exp(u_i/t)}{\sum_j \exp(u_j/t)}$$

  Temperature 會提升比較有可能生成的文字的機率，降低比較不可能文字的機率

(b) Hyperparameter

測試的資料集為 public 中的前 50 項，表中 Rouge 的數值為 F1 score。Beam-5 有最高的 F1 score，最後用於 submission 的 strategy 決定為 Beam=5。

Strategy	Parameter	Rouge1	Rouge2	RougeL
Greedy	NA	0.259218	0.117328	0.233027
Beam	3	0.292166	0.150378	0.271376
Beam	5	0.297612	0.154679	0.274955
Beam	7	0.297038	0.154987	0.272383
Top-k	4	0.245757	0.099630	0.213248
Top-k	10	0.231155	0.082439	0.199145
Top-p	0.8	0.232316	0.105679	0.214997
Top-p	0.9	0.2313013	0.102031	0.211419
Temperature	0.7	0.243220	0.111148	0.229546
Temperature	0.9	0.234919	0.111196	0.213808