HW2

R11944074 汪宣甫

Q1: Model

- Model:

這次使用的是google的MT5模型，他是T5模型的multilingual版本，pretrain在 mc4包含101種語言的資料集上。其架構主要是由transformer的encoder跟 decoder組成。在text summarization這個task當中，input會是一整段文本（ maintext)，output會是一個sequence，當中包含每一個token的機率，在訓練的 時候這些機率可以跟target（title）去做cross entropy loss. 下面是模型細部設定：

```
"_name_or_path": "google/mt5-small",
"architectures": [
  "MT5ForConditionalGeneration"
],
"d_ff": 1024,
"d_kv": 64,
"d_model": 512,
"decoder_start_token_id": 0,
"dense_act_fn": "gelu_new",
"dropout_rate": 0.1,
"eos_token_id": 1,
"feed_forward_proj": "gated-gelu",
"initializer_factor": 1.0,
"is_encoder_decoder": true,
"is_gated_act": true,
"layer_norm_epsilon": 1e-06,
"model_type": "mt5",
"num_decoder_layers": 8,
"num_heads": 6,
"num_layers": 8,
"pad_token_id": 0,
"relative_attention_max_distance": 128
"relative_attention_num_buckets": 32,
"tie_word_embeddings": false,
"tokenizer_class": "T5Tokenizer",
"torch_dtype": "float32",
"transformers_version": "4.27.1",
"use_cache": true,
"vocab_size": 250112
```

- Preprocessing:

跟HW1使用的wordpeice不一樣，MT5使用的是sentencepiece演算法，結合了 byte pair encoding跟unigram language model。跟wordpeice比起來， sentencepiece會切出更長單位的詞。

Q2: Training

- Hyperparameters:
  - learning_rate: 3e-05
  - train_batch_size: 4
  - eval_batch_size: 8
  - seed: 42
  - gradient_accumulation_steps: 4
  - total_train_batch_size: 16
  - optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
  - lr_scheduler_type: linear
  - num_epochs: 10.0

- Learning curves
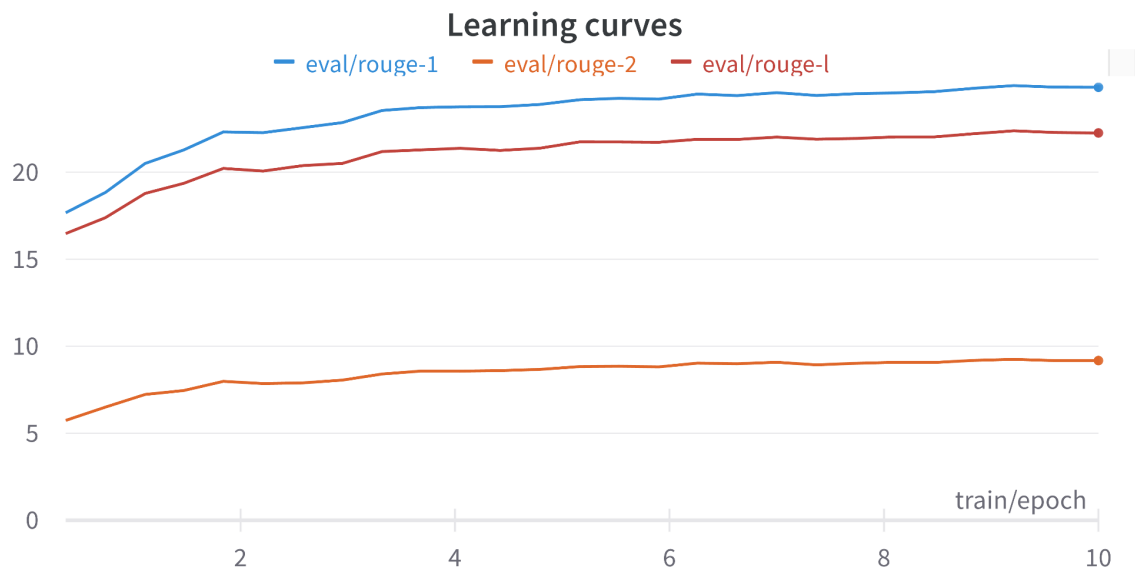
Q3: Generation strategies:
- Strategies
  - Greedy: 在每一個step用最高機率的token當作output
  - Beam Search: Greedy的改進，在每一個step決定output時都保持著n個候選人
  - Top-k Sampling: 在每一個step用前k個高的tokens去sample出output
  - Top-p Sampling: 在每一個step用由高至低的tokens的機率累積至p去sample出output
  - Temperature: Sample時，在softmax加入temperature去改變分佈的情況。t越大，分佈會越趨向均勻分布

- Hyperparameters
  - Greedy

	rouge-1 (f)	rouge-2 (f)	rouge-l (f)
greedy	24.88	9.176	22.257
sampling	15.029	4.321	13.238

- ○ Beam search

	rouge-1 (f)	rouge-2 (f)	rouge-l (f)
num_beams=2	26.007	10.126	23.236
num_beams=3	26.354	10.379	23.533

- ○ Top k

	rouge-1 (f)	rouge-2 (f)	rouge-l (f)
k=10	22.109	7.283	19.372
k=50	19.252	5.804	16.823

- ○ Top p

	rouge-1 (f)	rouge-2 (f)	rouge-l (f)
p=0.1	24.405	9.026	21.782
p=0.3	23.359	8.38	20.716

- ○ Temperature (constrain on top p = 0.1)

	rouge-1 (f)	rouge-2 (f)	rouge-l (f)
temp=0.8	24.777	9.183	22.154
temp=1.2	24.219	8.88	21.564