Q1

Model

```json
{
  "_name_or_path": "google/mt5-small",
  "architectures": [
    "MT5ForConditionalGeneration"
  ],
  "d_ff": 1024,
  "d_kv": 64,
  "d_model": 512,
  "decoder_start_token_id": 0,
  "dropout_rate": 0.1,
  "eos_token_id": 1,
  "feed_forward_proj": "gated-gelu",
  "initializer_factor": 1.0,
  "is_encoder_decoder": 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": "4.18.0",
  "use_cache": true,
  "vocab_size": 250100
}
```

Preprocessing

padding to batch's max length

max_source_length: 1024

max_target_length: 128

沒有 prefix

eos_token": "</s>", "unk_token": "<unk>", "pad_token": "-100"

use b T5 tokenizer base on SentencePiece

Q2

Hyperparameter

Batch size: 32 (real_batch_size=2 with gradient accumulation)

Lr: 1e-4

Lr_scheduler_type: linear

Loss: CrossEntrophy
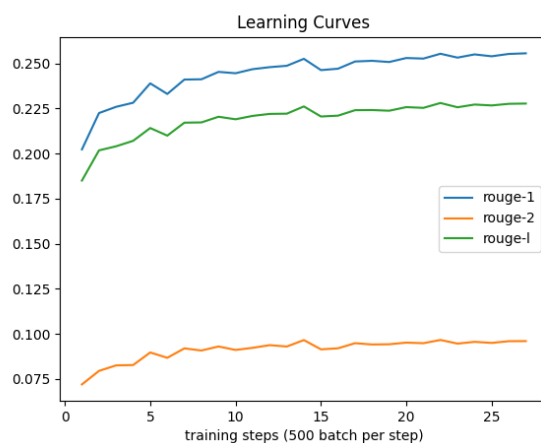
_n_gpu=2,

Optimizer: ADAMW_HF

    adam_beta1=0.9,

    adam_beta2=0.999,

    adam_epsilon=1e-08,

Epoch: 20

Learning Curves

Q3

Stratgies

- Greedy
  每個單字都選擇機率最高的的來當成 prediction，直到跑出 EOS
- Beam Search
  每次保留機率最高的前幾個來拓展，最終選出整句綜合最高的
- Top-k Sampling
  從直接選擇機率最高的變成使用機率 sample，但只從前 k 個最高的

```
sample
```

- Top-p Sampling

  從直接選擇機率最高的變成使用機率 `sample`，但只從前幾高機率加起來大於預設值的那些來 `sample`
  - Temperature
    在 softmex 的每一項分母加上一個 temperature value，以此將預測變得更 sharp 或 flat，可以配合 top_p 使用

Hyperparameters

共試了 8 種配置
發現 top_p 會導致太多選項而 sample 到不好的結果
需要用 t 才會好一點
反而因為我 top_k 的 k 只設成 10 導致結果較好
發現 num_beam_groups diversity 多寡不影響結果
而且 8beams 就夠了
最終採用 8 beam search

1. greedy

```
predictions/greedy.jsonl
{
  "rouge-1": {
    "r": 0.24444503774093412,
    "p": 0.28705515331592407,
    "f": 0.25572891895523026
  },
  "rouge-2": {
    "r": 0.09365675241056474,
    "p": 0.10512441948190304,
    "f": 0.09590826111460411
  },
  "rouge-l": {
    "r": 0.2177051854733419,
    "p": 0.2559669145907668,
    "f": 0.2277059630869554
  }
}
```

2. top_p=0.9, do_sample=True

```
predictions/nucleus_no_t.jsonl
{
  "rouge-1": {
    "r": 0.18812370463133024,
    "p": 0.19544246255075787,
    "f": 0.18651745473997194
  },
  "rouge-2": {
    "r": 0.06297100482384783,
    "p": 0.06429732215428768,
    "f": 0.06165239206839407
  },
  "rouge-l": {
    "r": 0.16765737878851653,
    "p": 0.17433047112355893,
    "f": 0.1661952697840628
  }
}
```

3. top_p=0.9, do_sample=True, temperature=0.5

```
predictions/nucleus_t.jsonl
{
  "rouge-1": {
    "r": 0.2379087723695437,
    "p": 0.27311805357883634,
    "f": 0.24663706863136348
  },
  "rouge-2": {
    "r": 0.08978649798918949,
    "p": 0.09891318717790719,
    "f": 0.09118265371796673
  },
  "rouge-l": {
    "r": 0.2118662122088088,
    "p": 0.24329195740197693,
    "f": 0.21951508963974045
  }
}
```

4. top_k=10, do_sample=True

```
predictions/top_k_no_t.jsonl
{
  "rouge-1": {
    "r": 0.2224980618556608,
    "p": 0.247000338605494,
    "f": 0.22747666631500965
  },
  "rouge-2": {
    "r": 0.07759266364150832,
    "p": 0.0835344080292519,
    "f": 0.07805444748484959
  },
  "rouge-l": {
    "r": 0.19612078176366682,
    "p": 0.2182209355795164,
    "f": 0.20063778900869397
  }
}
```

5. top_k=10, top_p=0.9, do_sample=True, temperature=1.2

```
predictions/mix.jsonl
{
  "rouge-1": {
    "r": 0.22224158980946146,
    "p": 0.24419802356683157,
    "f": 0.2259512020786834
  },
  "rouge-2": {
    "r": 0.07739589728530492,
    "p": 0.08166323222344861,
    "f": 0.07695494342833831
  },
  "rouge-l": {
    "r": 0.19622950205432016,
    "p": 0.21599617508904584,
    "f": 0.1995279936425481
  }
}
```

6. num_beams=8

```
predictions/8beams.jsonl
{
  "rouge-1": {
    "r": 0.25359171182681567,
    "p": 0.2890390115265507,
    "f": 0.26283478120297565
  },
  "rouge-2": {
    "r": 0.10334370270986401,
    "p": 0.11628085874971701,
    "f": 0.1064138183409635
  },
  "rouge-l": {
    "r": 0.2258343699931318,
    "p": 0.25766931125905235,
    "f": 0.234058818036066
  }
}
```

7. num_beams=8, num_beam_groups=2

```
predictions/8beams_group.jsonl
{
  "rouge-1": {
    "r": 0.25341658122988314,
    "p": 0.29113493722027517,
    "f": 0.263504231512911
  },
  "rouge-2": {
    "r": 0.10300092576668728,
    "p": 0.11603467398588703,
    "f": 0.10604794966782535
  },
  "rouge-l": {
    "r": 0.22569341855966013,
    "p": 0.2593013140581411,
    "f": 0.23456423579179767
  }
}
```

8. num_beams=16, num_beam_groups=4

```
predictions/16beams_group.jsonl
{
  "rouge-1": {
    "r": 0.25341658122988314,
    "p": 0.29113493722027517,
    "f": 0.263504231512911
  },
  "rouge-2": {
    "r": 0.10300092576668728,
    "p": 0.11603467398588703,
    "f": 0.10604794966782535
  },
  "rouge-l": {
    "r": 0.22569341855966013,
    "p": 0.2593013140581411,
    "f": 0.23456423579179767
  }
}
```