ADL HW2 Report

鄭恩庭 R12323031

October 24, 2024

1 Model

Model

In this assignment, I use `google/mt5-small` to do Chinese summarization. The mt5 model can be regarded as a multilingual version of T5 (Text-to-Text Transfer Transformer). T5 is an encoder-decoder based model that treats every NLP task as a text-to-text problem. Specifically, in this summarization task, the input is the context of the news, and the output is the predicted news title. During training, T5 learns to generate a shorter version of the input text by optimizing the likelihood of the correct news title, using a large dataset of text-summary pairs. At inference time, given a new piece of text, T5 generates a summary by predicting the most likely sequence of words that represent the main ideas of the input, utilizing its encoder-decoder architecture. The following table depicts the parameters in `google/mt5-small`:

feed_foward_proj	gated_gelu
dropout_rate	0.1
num_decoder_layers	8
num_heads	6
num_layers	8
dense_act_fn	gelu_new
d_ff	1024
d_kv	64
d_model	512
layer_norm_epsilon	1e-6

Table 1: Parameters of `google/mt5-small`

Preprocessing

In this assignment, I adopt `T5Tokenizer`, a SentencePiece model that hat supports a large vocabulary covering many languages. After splitting text into tokens, the tokenizer converts

these tokens into corresponding numerical IDs. Moreover, `T5Tokenizer` also manages special tokens:

- `<pad>`: padding token
- `<\s>`: end of sequence
- `<unk>`: unknown tokens

After the MT5 model generates output IDs, the tokenizer can convert these numerical IDs back into readable text. Furthermore, I set `max_source_length` to 256 and `max_target_length` to 64.

2 Training

Hyperparameter

The following table shows the hyperparameters I used:

batch_size	1
learning_rate	3e-4
weight_decay	1e-4
num_train_epochs	10
gradient_accumulation_steps	8
lr_scheduler_type	linear
num_warmup_steps	300

Table 2: Hyperparameters of `google/mt5-small`

Learning Curves

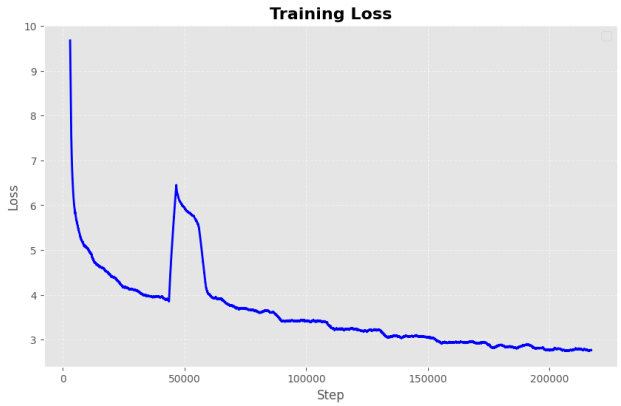

Figure 1: Learning Curve of `google/mt5-small`

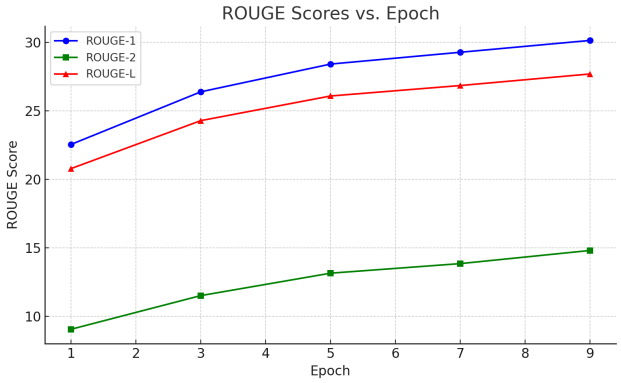

Figure 2: Rouge Scores v.s. Epoch

3 Generation Strategies

Strategies

- **Greedy**: Greedy is an intuitive decoding strategy that always select the most probable word conditional on all preceding words. Let w_i be the word at position i, then Greedy follows the formula below to predict the next word:

$$w_i = \arg\max_w P(w \mid w_1, \ldots, w_{i-1})$$

- **Beam Search**: Beam search keep track of k most probable sequences and find a better one. Specifically, $k = 1$ is exactly Greedy. Smaller beam size, k, can be ungrammatical or incorrect. Larger beam size reduces some above issues, but can be computationally expensive.

- **Top-k Sampling**: Top-k sampling predicts the next word with randomness, which samples the word via distribution but restricted to the top-k probable words. Smaller k leads to safer outputs, and larger k leads to more diverse outputs.
- **Top-p Sampling**: Top-p sampling samples from a subset of vocabulary, V^p, with the most probability mass. Mathematically,

$$V^p = \sup_{V' \subseteq V} \sum_{w \in V'} P(w \mid w_1 \cdots w_{i-1}) \geq p$$

- **Temperature**: Intuitively the hyperparameter temperature τ controls the diversity of the predicted next word. Mathematically,

$$P(w) = \frac{\exp(s_w/\tau)}{\sum_{w' \in V} \exp(s_{w'}/\tau)}$$

  Higher temperature leads to more diversity, and lower temperature leads to less diversity.

Hyperparameters

- **Greedy**

	rouge-1	rouge-2	rouge-L
Greedy	21.9472	7.4134	19.7027

Table 3: Performance of Greedy

- **Beam Search**

k	rouge-1	rouge-2	rouge-L
3	23.6902	8.8812	21.3372
5	23.7573	9.2221	21.4632

Table 4: Performance of Beam Search

- **Top-k Sampling**
  The beam size is fixed to 5 in the following table.

k	rouge-1	rouge-2	rouge-L
10	18.7323	5.4239	16.5907
100	15.1211	3.9949	13.5096

Table 5: Performance of Top-k Sampling

- **Top-p Sampling**
  The beam size is fixed to 5 in the following table.

p	rouge-1	rouge-2	rouge-L
0.1	22.0178	7.4144	19.7729
0.9	17.2962	4.9098	15.2938

Table 6: Performance Top-p Sampling

- **Temperature**
  The beam size is fixed to 5 in the following table.

τ	rouge-1	rouge-2	rouge-L
0.1	23.7517	9.2206	21.4616
0.9	23.3508	8.9106	21.0756

Table 7: Decoding Performance with different temperature

4 Bonus : Applied GPT-2 on Summarization

Model

The model I used in this section is `ckiplab/gpt2-base-chinese`. GPT-2, or Generative Pretrained Transformer-2, is a decoder-only model that leverages the decoder block of transformer. The following table shows the parameters in `ckiplab/gpt2-base-chinese`:

atten_pdrop	0.1
layer_norm_epsilon	1e-5
n_ctx	1024
n_embd	768
n_head	12
n_layer	12
n_positions	1024
activation_function	gelu_new

Table 8: Parameters of `ckiplab/gpt2-base-chinese`

The table below describes the hyperparameters I used in training:

max_source_length	256
max_target_length	256
batch_size	2
learning_rate	3e-4
weight_decay	1e-4
num_train_epochs	10
gradient_accumulation_steps	8
lr_scheduler_type	linear
num_warmup_steps	300

Table 9: Hyperparameters of `ckiplab/gpt2-base-chinese` during Training

The following table shows the hyperparameters I used in inference:

beam_size	7
top_k	50
temperature	0.1
repetition_penalty	20
max_new_tokens	16

Table 10: Hyperparameters of `ckiplab/gpt2-base-chinese` during Inference

Compare to t5 model

Figure 3: GPT-2 Learning Curve

rouge-1	rouge-2	rouge-L
5.6856	0.2559	0.4699

Table 11: Performance of `ckiplab/gpt2-base-chinese` on public set

There are three main differences between `ckiplab/gpt2-base-chinese` and `google/mt5-small`

- GPT-2 is a decoder-only model that predates mt5, so its performance is expected to be inferior to that of mt5.
- Comparing Table 11 with Table 4, I found that mt5 indeed works much better than GPT-2.
- The output below shows that the text generated by GPT-2 is ungrammatical and contains unnecessary spaces between characters.

Ground Truth	Anker 新款眞無線藍牙耳機 Liberty Air 2 Pro 引進台灣市場
mt5	Anker 推出眞無線藍牙耳機 Liberty Air 2 Pro
GPT-2	在 此 次 2021 中 ，宣 布 以 旗 下 雲 、算 能 網 推 連 證 業 體 的 盟 合 洞 安 據

Table 12: output text of public dataset id 21710

Ground Truth	藍染、客家美食、舊山線自行車「苗栗一日遊」超人氣美食美景
mt5	苗栗「七姊八弟山城小店」懶人包! 全台最熱門「鐵道自行車」懶人包
GPT-2	來 到 了 苗 栗 旅 遊 ，除 了 有 超 人 氣 點 雲 」復 百 藏 臥 風 手 子 氛 高 華 用 還 統

Table 13: output text of public dataset id 21711

Ground Truth	華碩打造對應軍規防護與 2 in 1 設計的 15.6 吋 Chromebook
mt5	華碩推出換上 Intel 第 11 代 Core 處理器的 Chromebook Flip Q5
GPT-2	如 同 其 他 品 牌 選 擇 在 2021 公 布 新 全 配 能 筆 第 抄 裝 設 更 時 亮 ? 薄 操 式 面

Table 14: output text of public dataset id 21712