Homework 3 Report

> r10922132 資工碩一 吳峻銘

Expand all
回到頂部
移至底部

Q1 : Model

Model

The model in this task is mT5. mT5 structure is same as T5 except for training data. mT5 train pn multilingual data. The mT5 structure is an encoder-decoder transformer.The input sequence first pass th the T5 encoder which has attention mechanism can capture contextual information. The detail of attention is the block consist of self-attention layer and feed forward layer and batch normalize layer. The decoder is similiar to encoder. However, the decoder self-attention is a little bit different from encoder which adopt autoregressive and can attend the token before present decode token. The last decoder layer then feed into dense layer to get the output. In summarize task, we need to abstract important information in main text and output a little summary. Hence, encoder-decoder Transformer model(mT5) can extract useful pattern in text and combine meaningful output. The pre-trained weight is from google/mt5-small.

Preprocessing

I use T5 tokenizer to tokinize input and output first. Then I truncate and pad all input to 256 words according to TA hint. Also, I truncate and pad all input to 64 words.

Q2 : Training

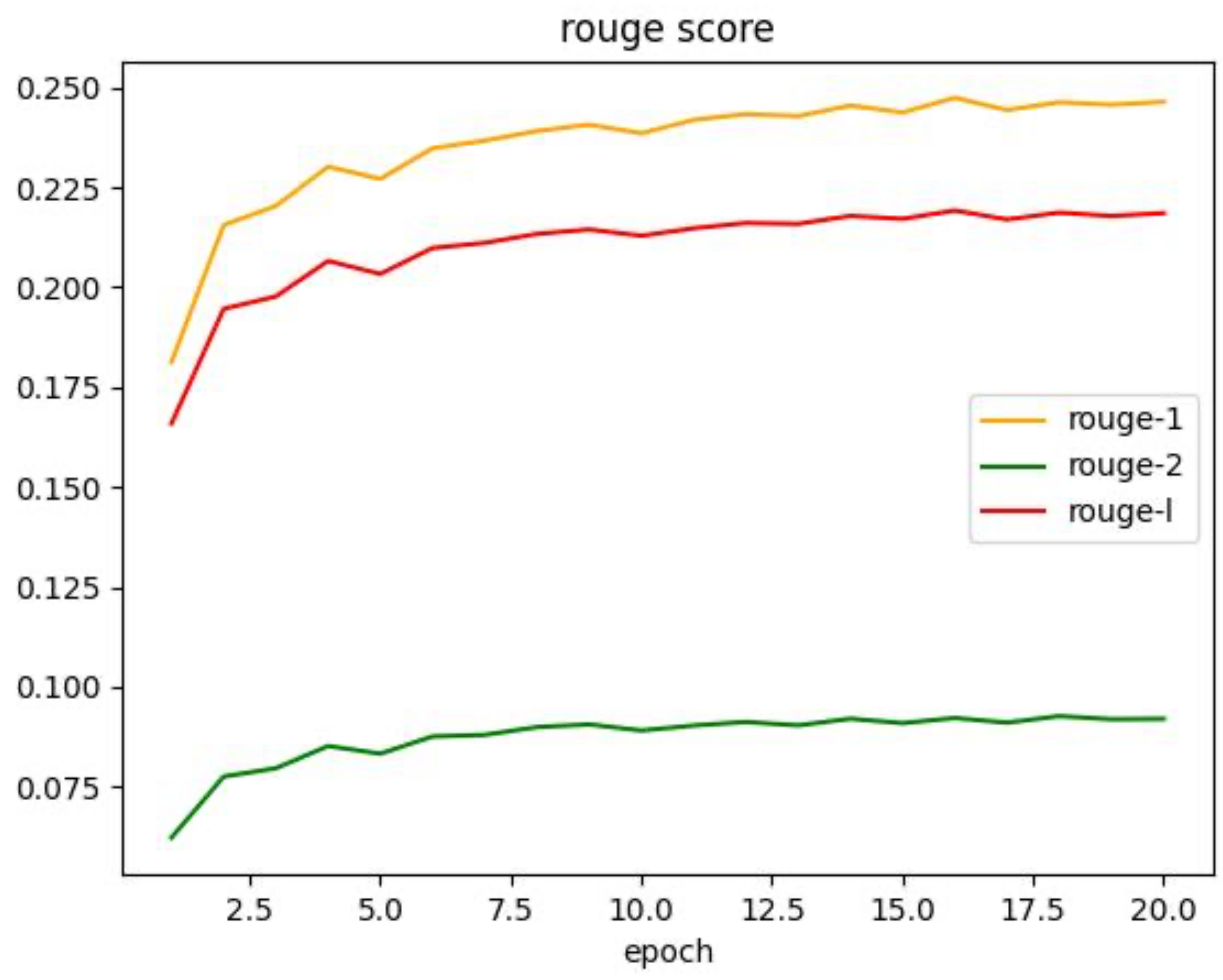
Hyperparameter

Because the model is too big, the learning rate, weight decay and batch size are select from small number to let training process more stable.

1. learning rate : 1e-4
2. weight decay : 5e-5
3. random seed : 123
4. fp16 : activate
5. epoch : 20
6. batch size : 4
7. accumulate step : 4

Learning Curves

The evaluation is perform every epoch and the strategy is greedy.



Q3 : Generation Strategies

Strategies

Greedy : greedy search the next word with highest probability
Beam Search : maitain the top beam numbers sequences with top probability
Top-k Sampling : sample from the top k words
Top-p Sampling : sample from the top tokens where the total probability of token pool is larger than p
Temperature : temperature is a hyperparameter applied to logits to affect the final probabilities from the softmax. temperature can control the diversity of the outcome from sampling

Hyperparameters

strategy	rouge 1	rouge 2	rouge l
Greedy	0.2482	0.0940	0.2220
Beam(3)	0.2612	0.1048	0.2349
Beam(5)	0.2645	0.1076	0.2371
Top-k(10)	0.2271	0.0784	0.1996
Top-k(20)	0.2191	0.0744	0.1932
Top-p(0.5)	0.2438	0.0905	0.2172
Top-p(0.25)	0.2498	0.0941	0.2233
Temperature(0.5)	0.2452	0.0913	0.2185
Temperature(0.7)	0.2346	0.0847	0.2082

I decide to choose beam(5) as my final strategy because it has highest rouge score.