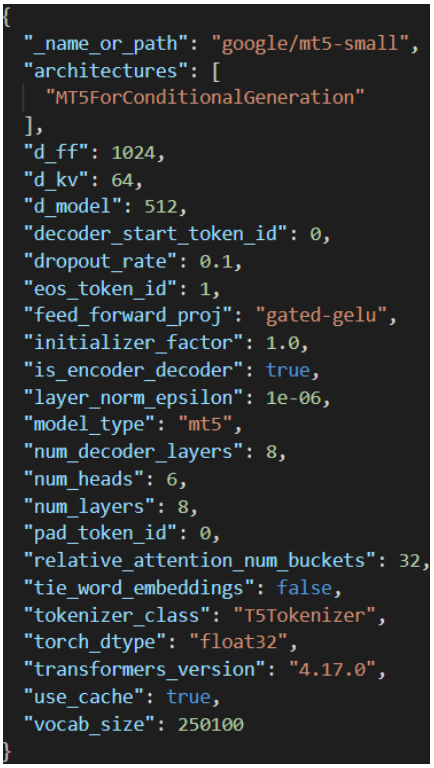
ADL HW3

Collaborator: r10525012 涂皓瑋

Q1: Model

- Model



  T5 trains with the same objective as that of BERT's which is the Masked Language Model with a little modification to it. Masked Language Models are Bidirectional models, at any time t the representation of the word is derived from both left and the right context of it. The subtle difference that T5 employs is to replace multiple consecutive tokens with a single Mask keyword, unlike, BERT that uses Mask token for each word. The Original text is transformed into Input and Output pairs by adding perturbations to it. Since the final objective is to have trained a model that inputs text and outputs text, the targets were designed to produce a sequence, unlike BERT, that tries to output one word (itself) through final feed-forward and softmax at the output level.

- Preprocessing

  T5 tokenizer is a SentencePiece tokenizer, which consists of byte pair encoding and unigram algorithm and applies Viterbi algorithm to construct the appropriate vocabulary.

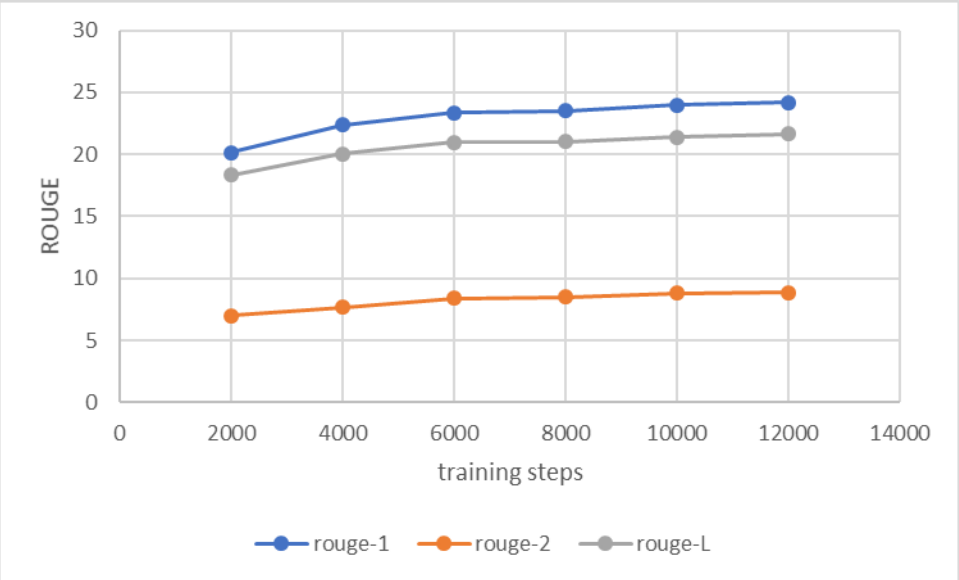
Q2: Training

- Hyperparameter

Model	google/mt5-small
Epoch	5
Loss function	CrossEntropyLoss
Optimizer	AdamW
Learning rate	5e-5
Scheduler	linear
Warmup ratio	0.1
Batch size	8

I set all parameters to the default, but the rouge score is shy of the baseline. I set warmup ratio of 0.1 to find a good optimization point and increase the training epoch to surpass the baseline.

- Learning Curves



Q3: Generation Strategies

- Strategies
  - Greedy Search simply selects the word with the highest probability as its next word.
  - Beam Search reduces the risk of missing hidden high probability word sequences by keeping the most likely number of beams of hypotheses at each time step and eventually choosing the hypothesis that has the overall highest probability.
  - Top-k Sampling: the K most likely next words are filtered and the probability mass is redistributed among only those K next words.
  - Top-p Sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p. The probability mass

is then redistributed among this set of words.

- Temperature is a hyperparameter used to control the randomness of predictions by scaling the logits before applying softmax. When the temperature is 1, we compute the softmax directly on the logits

- Hyperparameters
  - Greedy

strategy	Greedy	Sample
do_sample	False	True
max_length	128	
num_beams	1	
temperature	1	
top_k	50	
top_p	1	
rouge-1	23.65	17.62
rouge-2	8.76	5.46
rouge-L	21.31	15.62

Greedy is better than sampling since it always selects the word with the highest probability.

  - Beam Search

num_beams	1	3	5
max_length	128		
do_sample	False		
temperature	1		
top_k	50		
top_p	1		
rouge-1	23.65	24.97	25.03
rouge-2	8.76	9.96	10.17
rouge-L	21.31	22.48	22.51

Beam search can avoid choosing the bad word at the early stage. The more beams to search, the more option we get. Therefore, bigger beam number usually perform better.

  - Top-k Sampling

top_k	0	10	50
max_length	128		
do_sample	True		
num_beams	1		
temperature	1		

top_p	1		
rouge-1	13.28	20.77	17.64
rouge-2	3.78	6.87	5.52
rouge-L	11.91	18.43	15.70

The k should be selected appropriately. Too big or too small k will lower the performance.

■ Top-p Sampling

top_p	0.3	0.7	1
max_length	128		
do_sample	True		
num_beams	1		
temperature	1		
top_k	50		
rouge-1	23.16	21.10	17.62
rouge-2	8.57	7.34	5.46
rouge-L	20.77	18.86	15.62

Small p performs better.

■ Temperature

temperature	0.3	0.7	1
max_length	128		
do_sample	True		
num_beams	1		
top_k	0		
top_p	1		
rouge-1	23.37	19.78	13.28
rouge-2	8.66	6.76	3.78
rouge-L	21.05	17.73	11.91

Small temperature performs better.

■ Final generation strategy

strategy	Greedy
do_sample	False
max_length	128
num_beams	5
temperature	0.3
top_k	10
top_p	0.3
rouge-1	25.03

rouge-2	10.17
rouge-L	22.51

Bonus: Applied RL on SummarizationConfiguration (bert-base-chinese)

- Algorithm
- Compare to Supervised Learning

Reference

1. Understanding T5 Model : Text to Text Transfer Transformer Model | by Prakhar Mishra | Towards Data Science
2. How to generate text: using different decoding methods for language generation with Transformers (huggingface.co)
3. What is Temperature in LSTM? - Quora