HW3

40344032S 吳承哲

Q1: Model

1. Model

I use google mt5 model. It's a multilingual variant of T5 covering 101 languages. T5 (Text-to-Text Transfer Transformer) is a pretrained language model whose main objective is its use of a unified text-to-text format for all text-based NLP problems. T5 uses a basic encoder-decoder Transformer architecture.
When input data enters the model, it would predict the probability distribution of a set of next words. Then use decoder algorithm to generate the output texts.

2. Preprocessing

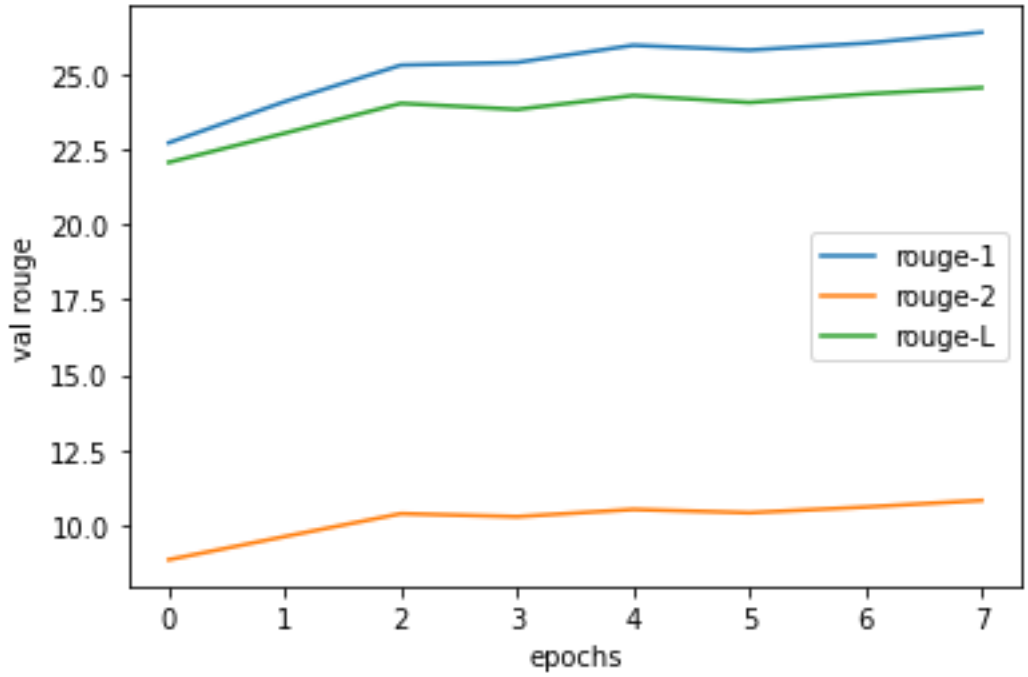
(1) Transfer original data to dataset format of hugging face.
(2) Using pretrained tokenizer, google/mt5-small, it could truncate the inputs and labels by our setting length.
(3) Using DataCollatorForSeq2Seq, it could pad inputs and labels to the longest length in a batch.

Q2: Training

1. Hyperparameter

Hyperparameter	**Value**	**Reason**
The max length of truncating for inputs	1024	If higher, it would be out of memory.
The max length of truncating for labels	128	If higher, it would be out of memory.
Learning rate	0.001	I tried 0.00001, 0.0001, 0.001, 0.01, and 0.001 was best setting.
Batch size	4	If higher, it would be out of memory.
Weight decay	0.01	From hugging face script
Epochs	8	If higher, it would be overfitting.
Gradient accumulation steps	32	Random choice

2. Learning Curve



Q3: Generation Strategies

1. Strategies

Greedy: Choose the word with the highest probability as its next word.

Beam Search: Depend on the number of beams. If number is 2, it would track the 2 most probable sequences and find a better one each step and finally choose the one has the overall highest probability.

Top-k Sampling: Sampling means selecting the next word according to its conditional probability distribution randomly. Top-k means that the k most likely next words are filtered and the probability mass is redistributed among only those k next words.

Top-p Sampling: Select from the smallest possible set of words whose cumulative probability goes beyond the probability p. Then, the probability mass is redistributed among this set of words. In this way, the number of words in the set can increase and decrease dynamically according to the next probability distribution.

Temperature: It is a hyperparameter for softmax. Higher temperature would make probability distribution more diversity.

2. Hyperparameters

Setting	Rouge-1	Rouge-2	Rouge-L
Beam number=1 Top-k=50 Top-p=1 Temperature=1	26.05	9.87	24.14
Beam number=2 Top-k=50 Top-p=1 Temperature=1	26.88	10.63	24.86

Final generation strategy:

Beam number=2

Top-k=50

Top-p=1

Temperature=1

The performance with public.jsonl dataset

- Rouge-1: 26.88, Rouge-2: 10.63, Rouge-L: 24.86