ADL HW2 Report

B10705005 資工三 陳思如

1. Model

- Architecture:

  It inputs with text and also output the text. It is a seq2seq model. It is multilingual transformer that supports many different languages ad have also been pretrained on mC4 datasers. It is a transformers that contained both encode and decode stages. The decode strtegies also affects the result of the prediction.

  mT5 is the model based on the T5. It improves upon T5 by using GeGLU nonlinearities, scaling both dmodel and dff. It doesn't do the dropout. One of its great benefit is that mT5 can perform zero-shot learning.

- Text Summarization:

  Text summarization needs to condense the long piece text into a shorter but also meaninful text to represent the whole content. The model needs to understand and know the difference between the input words and generate a coherent summary. The mT5 has been pretrained on multi languages so it can easily know some of the common phrases or words. Then I finetuned with the model to make it have a better fit on the news article from udn.com.

2. Preprocess

- Tokenization:

  I use the t5 pretrained tokenizer. Itsplit character into groups and transform the group into index id based on the t5 tokenizer library. Since we have set the max_source_length and max_target_length, the tokenizer will also pad the encoded sequence with meaningless index to fill up the length or just truncate the exceeded length.

- Data Cleaning:

  I have removed other unrelated columns in the datasets. Only leaves the "maintext" and "title". Other columns are removed in the tokenization.

3. Hyperparameter

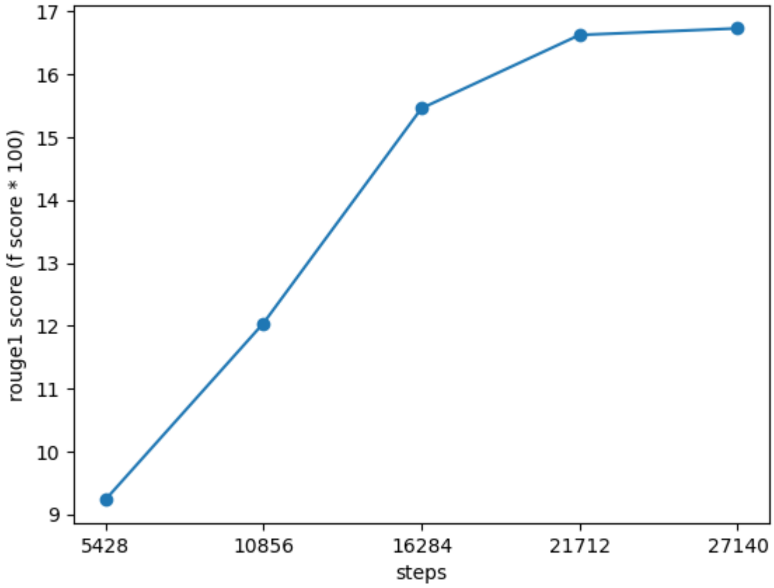
- batch size: 2*2
- max_source_length: 1024
- max_target_length: 128
- epochs: 5
- warmup steps: 300
- learning rate: 5e-4
- optimizer: AdamW
- lr_schedular: linear

The prediction evaluation on public.jsonl file is {rouge1: 0.25517, rouge2: 0.10395, rougeL: 0.23007}

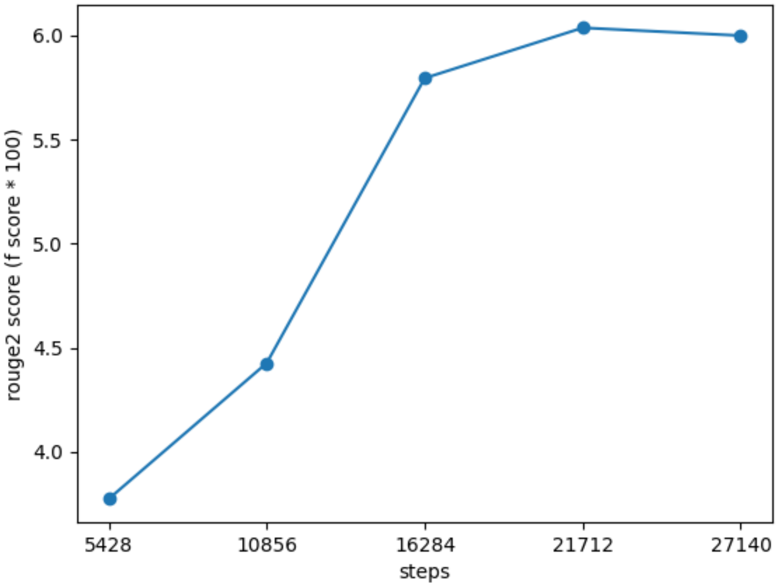
4. Learning Curve

I plotted these rouge scores based on the `evaluate("rouge")` metric score which is quite different with the `tw_rouge` score.

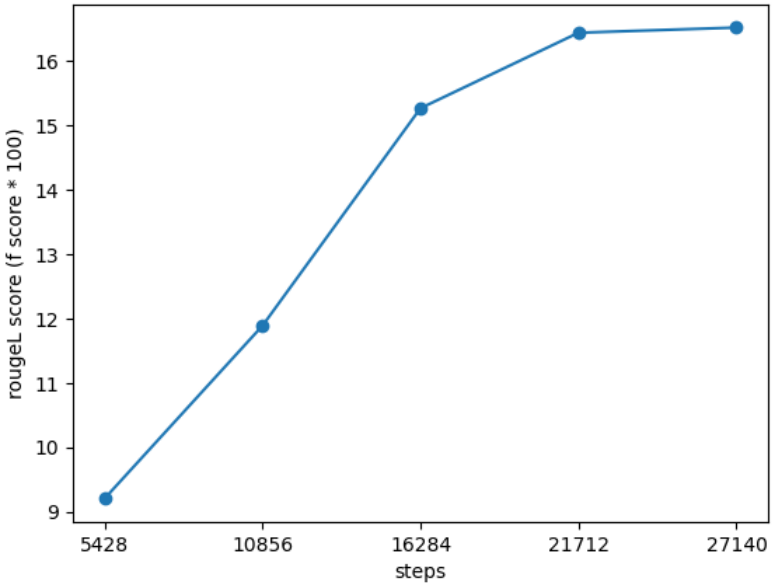
Rouge 1 score curve:



Rouge 2 score curve:



Rouge 3 score curve:



5. Strategies

- Greedy

  Greedy strategy selected **one** word with the **highest probability** at each step.

- Beam Search (beam size: n)

  Beam search strategy keeps the **top n candidates** with the highest probability in each step. After all the steps, it chooses the sequence with the **highest probability**. The probability is calculated by multiplying the probability of each word in the sequence. So the beam search results in high probability of short length output.

- Top-K (size: k)

  Top-K also selcts one word in each step. However, it chooses the world by **sampling** from the **top k words** with the high probability.

- Top-P (probability: p)

  Similar to the top k strategy. Top-P also **samples** one word in each step. The difference is that it samples from the top n words that have the **probability higher than p**.

- Temperature

  It is used with the sampling strategies. It makes the sampling task not so randomly. It **increases the probability** of the higher potential one and decreases the probability of the lower potential one.

6. Generation Strategy Hyperparameters

The text summarization is more based on the original text. Also, it is actually already determined with lower fraction of creativity. That means, the sampling strategy can't help much with our prediction.

In this case, I choose the beam search decode strategy. Trying with different beam size, Ifound out that **num_beam=5** is the best for my model.

These results are predicted with the public.jsonl file and the model with $5e-5$ learning rate.

	rouge1	rouge2	rougeL
greedy	24.532	9.538	22.085
beam (n=3)	**24.642**	**9.766**	**22.103**
beam (n=5)	24.508	9.721	21.958
beam (n=7)	23.043	8.333	20.625
top k (k=10)	20.281	6.333	17.913
top k (k=20)	19.099	5.888	16.875
top p (p=0.8)	19.589	6.269	17.393
top p (p=0.9)	18.978	6.059	15.815
top p (p=0.8) + tempertature (0.5)	22.549	8.076	20.179
top p (p=0.8) + tempertature (1.0)	19.843	6.506	17.715