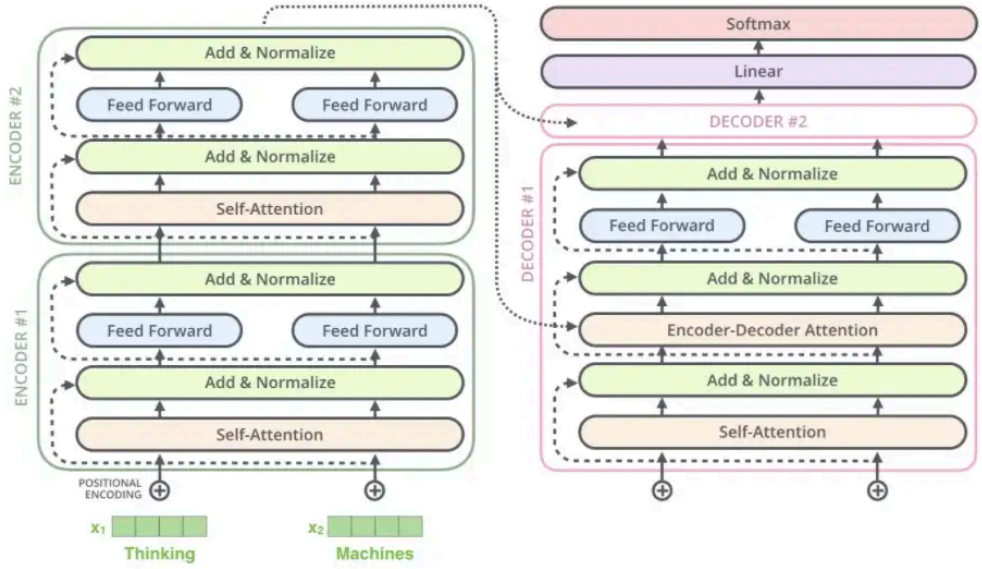
2022 fall ADL HW3 Report

R11922A15 張仲喆

Q1: Model

1.Describe the model architecture



T5 是encoder-decoder的架構，並利用span corruption (denoising)作為pretaining objective，且與BART輸出整句sentence不同，T5是將task設定為預測「被mask/noise」的部分。且T5是將所有task都以Seq2Seq的方式pretrain，故在fine-tune/實作時只需要在輸入起始點加入「summarize:」便可以進行summarization的任務。

2. Preprocessing

利用mT5自帶的tokenizer，將context tokenize後再padding至max length (1024)，並沒有做特別的data cleaning。

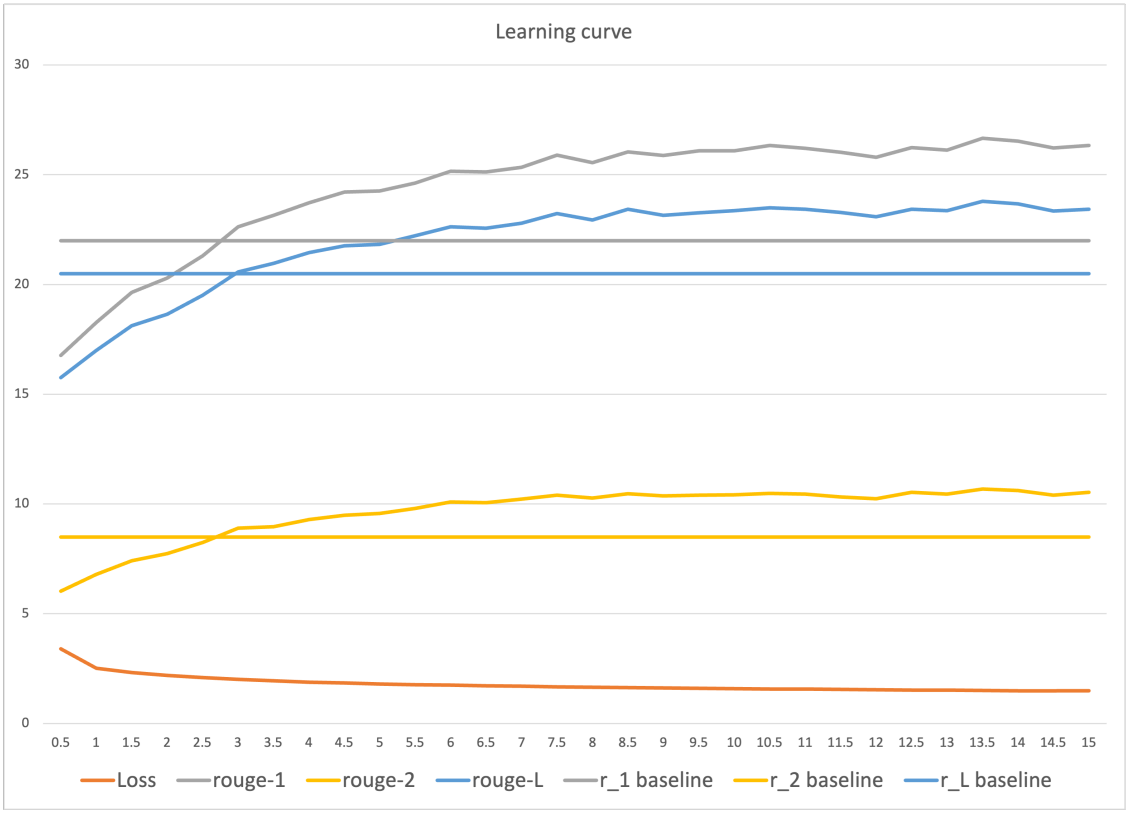
Q2: Training

a. Hyperparameter

- Learning rate = 5e-5
- Batch size = 8
- Max source length = 1024
- Max output length = 128
- Optimizer: AdamW
- Loss function: Cross Entropy loss
- Num-Beam = 16

b. Learning curve (15 epoch)

(batch size = 8, beam search w=4)



Q3: Generation Strategies

a. Strategies

- **Greedy**
  - 模型會在每個word預測時，選擇當下機率最高的word作為output，以此逐一輸出整個sentence。
- **Beam search**
  - 模型會保留機率前n個高(beam width) word，並各自進行下一次的預測，再次取前n個prediction繼續下一次的預測，最後再選出機率最高的sentence最為輸出。
- **Top-k Sampling**
  - 取出top-k個probability的word，並修正probability的總和為1後，利用其機率選出每個位置的output word。
- **Top-p Sampling**
  - 與top-k是接近的概念，選出機率前幾大的word，並將其機率加總需大於hyperparameter p∈(0, 1)，再從中依照機率隨機取出output word。
- **Temperature**
  - Temperature是利用在softmax中加入參數**τ**，使得word的diversity得以調整。
    - higher **τ**: more uniform, more diversity
    - Lower **τ**: more spiky, less diversity

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

b. Performance (other hyperparams same as Q2)
train for 5 epoch

	rouge-1	rouge-2	Rouge-l
Greedy	22.76	8.32	20.58
Beam search w=4	24.90	9.51	**21.70**
Beam search w=16	**24.91**	**9.68**	21.67
top_k = 2	22.19	7.81	20.00
top_k = 8	20.55	6.82	18.30
top_p = 0.5	21.45	7.60	19.26
top_p = 0.9	18.60	6.63	16.50
temperature = 0.3	20.99	7.07	18.79
temperature = 0.8	22.59	8.17	20.33

- **Final strategy: Beam search 4**
  - Good performance & faster inference time than beam16