ADL2022 HW3 Report

R10246002 應數碩二 陳俊憲

Q1: Model (2%)

Model (1%)：Describe the model architecture and how it works on text summarization.

- The model used in this task is mT5, which is a sequence-to-sequence model that belongs to the family of Encoder-Decoder transformer. mT5 is pretrained on the mC4 corpus, covering 101 languages. The variant I use is **google/mT5-small,** it has 8 encoder layers, 8 decoder layers, 6 attention heads, size of feed forward layer is 1024, and size of the key, query, value projections are 64.

1. Encoder transforms the input sequence $x_{1:T}$ to $x^{'}_{1:T}$ through bert-liked encoder.
2. In each time step t, the decoder is given the previous generated y_{t-1} and $x^{'}_{1:T}$ to generate next token y_t, until </s> appears.
3. $y_{1:N}$ is mapped to the size of vocabulary and normalized using softmax.
4. Various of generating methods can be applied to generate the summary, discussed in detail in Q3.

Preprocessing (1%)：Describe your preprocessing (e.g. tokenization, data cleaning and etc.)

- Data cleaning：I tried to remove “\n” and URL，but the performance does not improve, so raw data is input without additional cleaning.

- Tokenization：maintext and title in training dataset are tokenized into inputs and labels by SentencePiece algorithm. The max length of input is set to be 256, and the max length of output is set to be 64, length of tokenized text larger than this setting will we truncated, otherwise will be padded. Tokens corresponding to padded labels will not be involved in the loss computation.

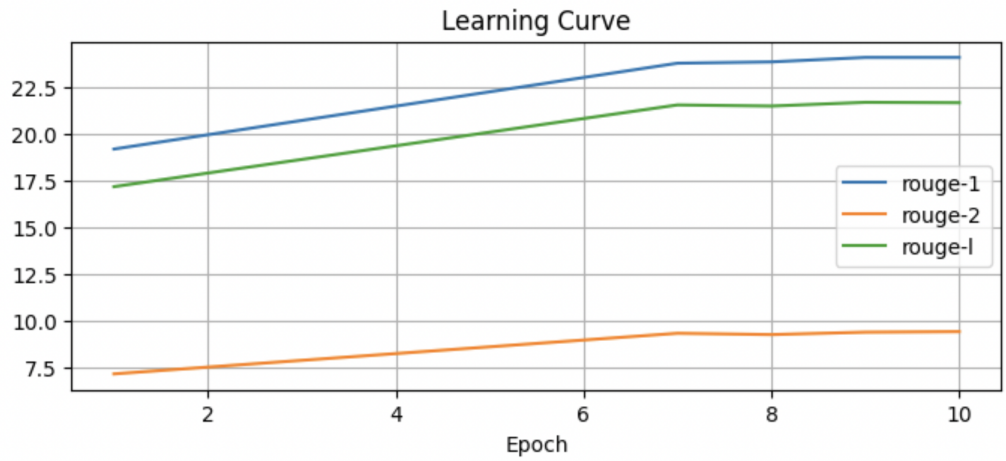
Q2: Training (2%)

Hyperparameter (1%)：Describe your hyperparameter you use and how you decide it.

- Max input/output length are chosen under the consideration of memory usage.
- Learning rate and batch size are chosen under the trade off of efficiency and performance, larger batch size and small learning rate updates model parameters less frequently, which may cause slowly convergence. So I set initial learning rate to be 1e-3, and it the result turns out to be acceptable.

Model	**Optimizer**	**Initial Learning Rate**	**Epoch**
google/mt5-small	AdamW	1e-3	15
Total train batch size	**Max input length**	**Max input length**	**Loss function**
64	256	64	CrossEntropyLoss

Learning Curves (1%)：Plot the learning curves (ROUGE versus training steps)



Q3: Generation Strategies(6%)

Stratgies (2%)：Describe the detail of the following generation strategies:

- Greedy

At each time stamp t, greedy search selects the word with the highest conditional probability as its next word. Generation ends when <EOS> (</s>) token is generated or max length is reached.

$$w_t = \arg\max_w P(w|w_{1:t-1})$$

- Beam Search

By keeping a B most likely hypothesis, beam search can avoid missing high probability word sequence. At each time step, beam search choose B candidates with highest probability, and eventually choose the sequence with highest probability. Notice that B is a hyperparameter to be chosen, and the procedure reduced to greedy search when $B = 1$.

- Top-k Sampling

At each time t, top-k sampling selects K words with highest K probabilities, and the probability mass is redistributed to form a new distribution among these words, then the next word is sampled based on this distribution. Notice that the parameter K remains constant over time step t.

- Top-p Sampling

Instead of fixing the population to a constant, top-p sampling scheme choose the minimum number of words such that the sum of their conditional probability mass exceeds a threshold p.

- Temperature

Temperature sampling introduce a parameter $0 < T \leq 1$, to control the conditional probability of next word. As the formulation below, where v_i is the i^{th} element in the vocab V, and $h_{t,i}$ is the i^{th} entry of output at state t. Notice that with higher temperature, it is more likely to sample low probability tokens, and it is the normal softmax function when $T = 1$.

$$P(w_t = v_i|w_{1:t-1}) = \frac{exp(h_{t,i}/T)}{\sum_{v_j \in V} exp(h_{t,j}/T)}$$

Hyper parameters (4%)

- Try at least 2 settings of each strategies and compare the result.

method	rouge-1	rouge-2	rouge-l	Batch_Size
Greedy	22.47	7.89	20.21	64

method	rouge-1	rouge-2	rouge-l	Batch_Size
Beam Search (=3)	24.43	9.41	21.95	32
Beam Search (=5)	**24.61**	9.64	**22.08**	32
Beam Search (=10)	**24.37**	**9.65**	**21.98**	32
Top_K (K=10)	20.05	6.31	17.7	64
Top_K (K=100)	16.07	4.51	14.2	64
Top_K (K=10) Temperature (t=0.3)	22.37	7.82	20.1	64
Top_P (P=0.25)	21.47	7.39	19.17	64
Top_P (P=0.2) Temperature (t=0.1)	22.47	7.89	20.21	64
Top_P (P=0.5) Temperature (t=0.5)	22.54	7.9	20.22	64
Temperature (t=0.35)	22.06	7.66	19.81	64
Temperature (t=0.7)	18.98	6.09	16.93	64

- What is your final generation strategy? (you can combine any of them)

As illustrated in the experiments above, beam search has the best performance among all methods, but it is time and memory consuming. We can see that using temperature scaling with small temperature can slightly improve the performance of sampling methods, but still not pass the baseline. So the final generation strategy I decided to use is beam search (=5) .

Bonus: Applied RL on Summarization (2%)

Algorithm (1%)

Describe your RL algorithms, reward function, and hyperparameters.

- $R(\tau^n) = w^T r_n / w^T r_b, \text{where } w \in \mathbb{R}^3, w^T r_b$ is the baseline .
  - $r = [\text{rouge-1}, \text{rouge-2}, \text{rouge-l}]^T$, $r_b = [0.22, 0.085, 0.2]^T$
  - Here, I choose $w = [0.3, 0.4, 0.3]^T$

- $R(\theta) = -\frac{1}{N}\sum_{n=1}^{N}(R(\tau^n) + 0.5) \times CE(\tau^n|\theta)$, and we do gradient assent over $R(\theta)$ iteratively updating θ.

  (Equivalently, gradient decent over $-R(\theta)$)

Compare to Supervised Learning (1%)

- The training steps of rl begins from the last checkpoint of supervised learning with the following config.
  - Epoch：5
  - Learning rate：3e-5
  - Other hyperparameters are same as supervised one.
- Both generating method is beam search with beam size 5. (on public.jsonl)

Best Rouge	rouge-1	rouge-2	rouge-l
Supervised	24.61	9.64	22.08
RL	25.5	10.26	22.92

- Comparison of generations examples (public.jsonl)

Ground truth	Supervised	
最多千金股的高價族群！台積電後的第二護國神山 這2檔被忽略的股票	台積電成為台股第二座神山	台積電最強護國神山 聯發科股價逼近千元
英特爾下單台積電！半導體果真「護國神山」 連陸都怕對我過度依賴	英特爾下單台積電一事「有影」了!	英特爾下單台積電 供電腦第2代獨立顯示卡「DG2」打入PC遊戲市場
迎合中國市場喜好！運動品牌農曆新年系列服飾、鞋款風格好接地氣	New Balance攜手Kiwi李函、Jun邱文駿打造新年新希望	聯合新聞網
哈登、厄文都得重修「武德」 籃網咩噗耗子尾汁	火箭快刀斬亂麻交易送走哈登 哈登四隊交易	火箭快刀斬亂麻交易送走哈登 四隊交易