NTU-ADL-HW3-Report (R11922A16 資工碩一 柳宇澤)

tags: NTU ADL

Q1. Model

Model (1%)

mT5 is a multilingual variant of T5 (Text-to-Text Transfer Transformer) that was pre-trained on a new Common Crawl-based dataset covering 101 languages. A classical Transformer sturcture is consist of a Encoder & Decoder. Encoder transfers a input (An existing information) into a vector in a latent space. Then, decoder iteratively transfer the vector into a text based on preview vector.

Preprocessing (1%)

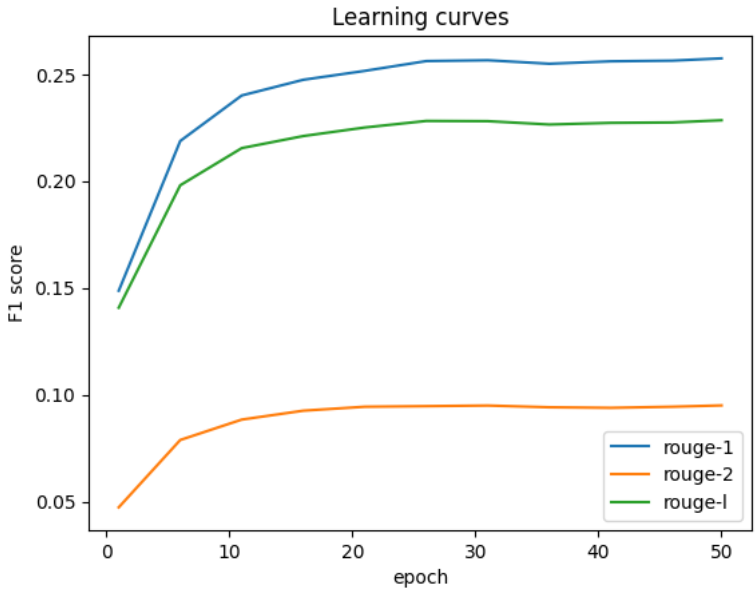
I transfer `jsonl` to `json` for loading datasets. Secondly, I added a prefix, `"summarize: "`, to every `maintext`. Then tokenizing `maintext` and padding sentences to `max_len=1024`.

Q2: Training

Hyperparameter (1%)

I used Accelerator so I set `per_device_XXX_batch_size` as large as possible, which is `6` in my server.
For `learning rate`, I set it to relatively small value, `5e-5`, as NLP generation transformer model could be hard to train. What's more, I also set `num_warmup_steps` to `20`, but this value could be too small for my training epochs.
Initially, I tried to train `5` epochs and observed the model didn't be converged. Hence, I trained it for `50` epochs. Finally, the model was successfully converged. However, I could try higher learning rate next time to save training time.
Last but not least, I set `max_len` to `512` which means sequences will be padded to 512 length.

Learning Curves (1%)



Q3: Generation Strategies

Stratgies (2%)

- Greedy Search
  - Greedy search simply selects the word with the highest probability as its next word.
- Beam Search
  - Beam search reduces the risk of missing hidden high probability word sequences by keeping the most likely num_beams of hypotheses at each time

step and eventually choosing the hypothesis that has the overall highest probability.

- Top-k Sampling
  - Sampling means randomly picking the next word according to its conditional probability distribution. In 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
  - Instead of sampling only from the most likely K words, in 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
  - A trick is to make the distribution sharper (increasing the likelihood of high probability words and decreasing the likelihood of low probability words) by lowering the so-called temperature of the softmax. While applying temperature can make a distribution less random, in its limit, when setting temperature to 0, temperature scaled sampling becomes equal to greedy decoding

Hyperparameters (4%)

My strategy is `Beam = 4`.
Belows are F1-scores under different settings.

Hyperparm.	Rouge-1	Rouge-2	Rouge-l
Greedy	0.2574	0.0950	0.2285
* **Beam = 4**	**0.2704**	**0.1075**	**0.2400**
Beam = 10	0.2698	0.1091	0.2400
Top-K = 10	0.2307	0.0781	0.2020
Top-K = 50	0.2124	0.0670	0.1861
Top-P = 0.94	0.1782	0.0558	0.1575
Top-P = 0.98	0.1810	0.0574	0.1594
Temperature = 0.2	0.2555	0.0942	0.2255
Temperature = 0.7	0.2317	0.0818	0.2050