Implementation of Learning to Ask Paper

Md Abdur Razzaq Riyadh

May 7, 2024

1 Methodology of the Paper

1.1 Problem

I have selected the paper titled 'Learning to Ask: Neural Question Generation for Reading Comprehension' to implement Du et al. (2017). This paper tackles the task of question generation from text. More specifically, given an input sequence x, generate output sequence y that is a question and the answer must be found in x. Note that the length of x and y does not have to be equal. Figure 1 shows examples of this sequence to sequence problem.

Sentence	Question
any person that hath occasion for the said engines may apply themselves to the patentee at his house near st thomas apostle london or to mr. nicholas wall at the workshoppe near saddlers wells at islington or to mr. william tillcar , turner , his agent at his house in woodtree next door to the sun tavern london	what tavern did william tillcar live adjacent to ?
in the ancient egyptian era of atenism , possibly the earliest recorded monotheistic religion , this deity was called aten , premised on being the one " true " supreme being and creator of the universe .	what was the first monotheistic religion ?

Table 1: Examples from the processed SQuAD Dataset

1.2 Methodology

They have used RNN-based encoder-decoder architecture (Bahdanau et al., 2014),(Cho et al., 2014) to conditionally generate the question. During decoding step, they have also used attention (Luong et al., 2015) to condition the output.

The encoder is a bi-directional RNN that produces sentence encoding from the entire $[x_1, x_2, ..., x_s]$ sequence. The decoder is a uni-directional RNN that takes a single token y_t and produce decoder representation. The attention score is calculated using Luong et al. (2015) which determines by how much sentence tokens attends to the question token, y_t. The attention score is then used to calculate the context vector by multiplying the encoder representation. This is considered as taking the weighted average of encoder representation. Finally, to generate the next token's logits, decoder representation and context vector is concatenated and passed through linear layers.

The paper used the SQuAD dataset to process and prepare the dataset for question generation task. SQuAD have paragraphs and multiple questions and answers for each paragraphs. Du et al. (2017) selected questions and the sentence from the paragraph containing the answer to create the parallel dataset they have used for this task.

The paper reported experiments with and without pre-trained word embedding (Glove 840B, 300D), with and without paragraph on top of the sentence. I have reimplemented with and without word embedding experiment only.

2 Implementation Details and Challenges

The paper originally was implemented using Lua and Torch. I have implemented in Python and PyTorch. For the dataset, I have used already pre-processed and tokenized data provided by the authors.

The primary challenge in the implementation was understanding the attention score equation (see Du et al., 2017, eq 5). Luong et al. (2015) was helpful in clarifying this issue. I also used Robertson (2024) whenever had any confusion regarding the implementation.

Another challenge was the unstable training because of the high learning rate. Most of the time, the binary cross entropy loss would start to increase from 10 to 115 and jump around 100. After cancelling and re-running the training would solve this problem and both loss and perplexity on the validation split would decrease gradually.

3 Result Comparison

1. not good, insert tables

4 Personal Contribution

1. describe nucleus sampling 2. implementation details

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Robertson, S. (2024). NLP From Scratch: Translation with a Sequence to Sequence Network and Attention PyTorch Tutorials 2.3.0+cu121 documentation — pytorch.org. `https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html`. [Accessed 07-05-2024].