Encoder Extract Transformer Text Summarization

Bumjin Park

April 7, 2021

Abstract

This paper is about the result of the EET(Encoder Extract Transformer) for text summarization. Similar to the copy mechanism, we extract the word distribution from the encoder and apply it to find the first word in the decoder. The source code is available in Github [1]

1 Experiment

The EET(Encoder Extract Transformer) model has an additional path from the encoder to the first prediction $\hat{y}_1$. After encoding the document X using transformer encoder, we get the encoded vectors $e_1, e_2, \cdots e_D$. There are two assumptions for the modeling. First, each word of the document may has different preference of the first predicted word $\hat{y}_1$ which is the most important prediction because we are using sequential decoding. Second, by controlling the first word in the decoder, we can get totally different summary.

The model is trained with CNN-DM dataset[1] and baseline model is the pretrained BART[2] offered by Fairseq[3]. The learning rate is $3 \cdot 10^{-5}$ and the warm up is 500 steps. We trained the model for 5 epoch with the freezed encoder and the decoder of the BART. To add mean pooled word distribution to the first prediction y_1, we trained new linear layer for the prediction of $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_S$.

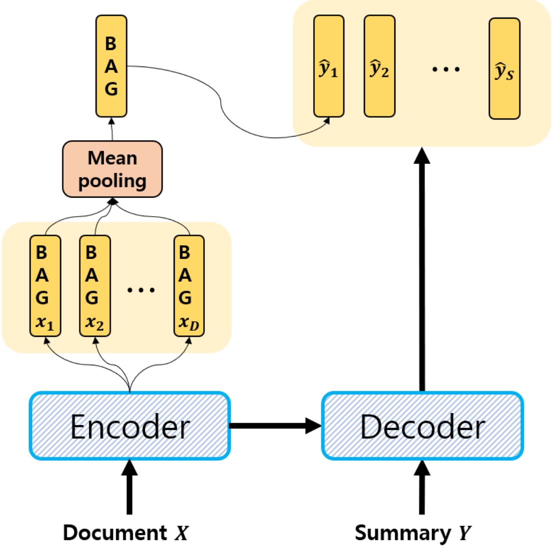

Figure 1: Encoder Extract Modeling. The Blue Rectangle means freezing while fine tuning. The only trained parts are extract encoder part and linear probing part

[1] https //github.com/fxnnxc/text_summarization

2 Result

Model	ROUGE1	ROUGE2	ROUGE3	NOVEL1	NOVEL2	NOVEL3
BART	43.45	20.62	**40.35**	1.62	11.83	21.78
EET	**43.63**	**20.81**	40.30	**1.00**	**7.86**	**15.32**

ROUGE[4] score measures how close the generated summary and the true summary in n-gram method. Therefore the higher the ROUGE score, the higher the accuracy. NOVEL score measures how close the document and the generated summary in n-gram method. Therefore the lower the NOVEL score, the more the generated summary copies the document.

EET model has similar performance in ROUGE score or even better in the case of ROUGE1 and ROUGE2. However the impact is not that significant. And it is not clear whether EET model has some difference with vanilla transformer model because the gradient vanishing problem to the additional path is possbile and in that case, the EET model is just same with the transformer. However the NOVEL score shows that EET model generate a summary by using the words in the document more than the baseline BART.

3 Conclusion

Since the CNN-DM dataset achieves high ROUGE score even by taking the lead sentences as a summary, it is natrual that both models have low NOVEL scores. The gap of NOVEL scores between two models may suggests that by controlling the impact of additional path to the first word, we may control the degree of summary.

References

[1] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.

[2] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv e-prints*, page arXiv:1910.13461, October 2019.

[3] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[4] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.