EXTRACTIVE SUMMARIZATION WITH VERY DEEP PRETRAINED LANGUAGE MODEL

Yang Gu[1] and Yanke Hu[2]

[1]Suning USA, Palo Alto, California, USA
[2]Humana, Irving, Texas, USA

ABSTRACT

Recent development of generative pretrained language models has been proven very successful on a wide range of NLP tasks, such as text classification, question answering, textual entailment and so on. In this work, we present a two-phase encoder decoder architecture based on Bidirectional Encoding Representation from Transformers(BERT) for extractive summarization task. We evaluated our model by both automatic metrics and human annotators, and demonstrated that the architecture achieves the state-of-the-art comparable result on large scale corpus – CNN/Daily Mail[1] . As the best of our knowledge, this is the first work that applies BERT based architecture to a text summarization task and achieved the state-of-the-art comparable result.

KEYWORDS

BERT, AI, Deep Learning, Summarization

1. INTRODUCTION

Document summarization is a widely investigated problem in natural language processing, and mainly classified into two categories: extractive summarization and abstractive summarization. Extractive approaches work in the way of extracting existing words or sentences from the original text and organizing into the summary, while abstractive approaches focus more on generating the inherently same summary that is closer to the way human expresses. We will focus on extractive summarization in this paper.

Traditional extractive summarization approaches can be generally classified into two ways like: greedy approaches [1] and graph-based approaches [2]. Recently, deep learning techniques have been proven successful in this domain. Kageback et al. 2014 [3] utilized continuous vector representations in the recurrent neural network for the first time, and achieved the best result on Opinosis dataset [4]. Yin et al. 2015 [5] developed an unsupervised convolutional neural network for learning the sentence representations, and then applied a special sentence selection algorithm to balance sentence prestige and diversity. Cao et al. 2016 [6] applied the attention mechanism in a joint neural network model that can learn query relevance ranking and sentence saliency ranking simultaneously,and achieved competitive performance on DUC query-focused summarization benchmark datasets[2]. Cheng et al. 2016 [7] developed a hierarchical document

[1]https://github.com/deepmind/rc-data
[2]https://www-nlpir.nist.gov/projects/duc/

DOI: 10.5121/ijaia.2019.10203

encoder and an attention-based extractor,and achieved results comparable to the state of the art on CNN/Daily Mail corpus [8] without linguistic annotation.

Recent development of GPT [9] and BERT [10] has proven the effectiveness of a generative pre-trained language model on a wide range of different tasks, such as text classification, question answering, textual entailment, etc, but neither of these approaches has been tested on the text summarization task.

In this paper, we introduced a two-phase encoder decoder architecture based on BERT. We fine-tuned this model on the CNN/Daily Mail corpus for single document summarization task. The result demonstrated that our model has the state-of-the-art comparable performance by both automatic metrics (in terms of ROUGE [11]) and human assessors. As the best of our knowledge, this is the first work that applies BERT based architecture to a text summarization task.

2. TRAINING DATASET

CNN/Daily Mail [8] dataset is the most widely used large scale dataset for summarization and reading comprehension. The training set contains 287226 samples. The validation set contains 13368 samples. The test set contains 11490 samples. Each sample contains an article and a referenced summary. After tokenization, an article contains 800 tokens on average, and a corresponding summary contains 56 tokens on average.

3. SUMMARIZATION MODEL

In this section, we propose the summarization model that efficiently utilizes BERT [10] as the text encoder. The architecture is shown in Figure 1 and consists of two main modules, BERT encoder and sentence classification.

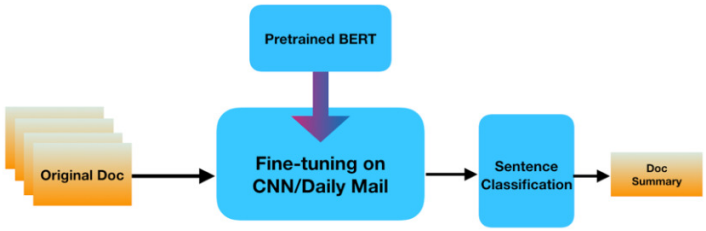

Figure 1: Architecture of the Summarization Model

The sentences of the original document will be feed into the BERT encoder sequentially, and be classified as if or not to be included in the summary. BERT is essentially a multi-layer bidirectional Transformer [19] encoder. Each layer consists of a multi-head self-attention sub-layer and a following linear affine sub-layer with the residual connection. The BERT encoding process utilizes query matrix W^q, key matrix W^k of dimension d_k, and value matrix $W^v$of dimension d_v, and it can be expressed as Eq.(1).

$$BERT(W^q, W^k, W^v) = Softmax(\frac{W^Q(W^K)^T}{\sqrt{d_k}})W^V \quad (1)$$

Given the input document as $X = \{x_1, \dots, x_m\}$ where x_i denotes one source token. The BERT encoder's output can be expressed as Eq. (2).

$$H = BERT\left(W^q, W^k, W^v\right)(x_1, \dots, x_m) \quad (2)$$

In CNN/Daily Mail dataset, the ground truth is the referenced summaries without sentence labels, so we follow Nallapati et al. [20] to convert the abstractive referenced summaries to extractive labels. The idea is to add one sentence each time incrementally to the candidate summary, so that the Rouge score of the current set of selected sentences is increasing in regard to the referenced summary. We stopped when the remaining candidate sentences cannot promote the Rouge score with respect to the referenced summary. With this approach, we convert a text summarization task to a sentence classification task.

Our BERT encoder is based on Google's TensorFlow[3] implementation (TensorFlow version >= 1.11.0). We used the BERT-Base model(uncased, 12-layer,768-hidden,12-heads, 110M parameters) and then fine tune the model on training set of CNN/Daily Main corpus. All inputs to the reader are padded to 384 tokens; the learning rate is set to 3x10-5, and other settings are by default.

4. EXPERIMENT RESULT

We evaluate ROUGE-1, ROUGE-2, and ROUGE-L [11] of our proposed model on CNN/Daily Mail test dataset. This automatic metric measures the overlap of 1-gram (R-1), bigrams (R-2) and the longest common subsequence between the model generated summaries and the reference summaries. We also listed the previously reported ROUGE scores on CNN/Daily Mail in Table 1 for comparison. As Table 1 shows, the ROUGE score of BERT summarization is comparable to the start-of-the-art models.

Table 1: Comparative evaluation of BERT Summarization with recently reported summarization systems

Models	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Pointer Generator [12]	36.44	15.66	33.42
ML + Intra-Attention [13]	38.30	14.81	35.49
Saliency + Entailment reward [14]	40.43	18.00	37.10
Key information guide network [15]	38.95	17.12	35.68
Inconsistency loss [16]	40.68	17.97	37.13
Sentence Rewriting [17]	40.88	17.80	38.54
Bottom-Up Summarization [18]	41.22	18.68	38.34
BERT Summarization (Ours)	37.30	17.05	34.76

As per Schluter [21], extractive summarization performance tends to be undervalued with respect to ROUGE standard, so we also conduct human assessment on Amazon Mechanical Turk (MTurk)[4] for the relevance and readability between Bottom-Up Summarization model (best

[3]https://github.com/google-research/bert
[4]https://www.mturk.com/

reported score on CNN/Daily Mail) and our BERT Summarization model. We selected 3 human annotators who had an approval rate over than 95% (at least 1000 HITs), and showed them 100 samples from the test dataset of CNN/Daily Mail, including the input article, the reference summary, and the outputs of the two candidate models. We asked them to choose the better one between the two candidate models' outputs, or choose "tie" if both outputs were equally good or bad. The relevance score is mainly based on if the output summary is informative and redundancy free. The Readability score is mainly based on if the output summary is coherent and grammatically correct. As shown in Table 2, our BERT Summarization model achieves higher scores than the best reported Bottom-Up Summarization model by human assessment.

Table 2: Human assessment: pairwise comparison of relevance and readability between Bottom-Up Summarization [18] and BERT Summarization

Models	**Relevance**	**Readability**	**Total**
Bottom-Up Summarization	42	38	80
BERT Summarization (Ours)	49	46	95
Tie	9	16	25

5. CONCLUSIONS

In this paper, we present a two-phase encoder decoder architecture based on BERT for extractive summarization task. We demonstrated that our model has the state-of-the-art comparable performance on CNN/Daily Mail dataset by both automatic metrics and human assessors. As the best of our knowledge, this is the first work that applies BERT based architecture to a text summarization task. In the future, we will test employing better decoding and reinforcement learning approaches to extend it to abstractive summarization task.

REFERENCES

[1] Carbonell, J., and Goldstein, J. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 335–336. ACM.

[2] Radev, D., and Erkan, G. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 457–479.

[3] Kageback, M.; Mogren, O.; Tahmasebi, N.; and Dubhashi, D. 2014. Extractive summarization using continuous vector space models. 31–39.

[4] Ganesan, K.; Zhai, C.; and Han, J. 2010. Opinosis: a graph based approach to abstractive summarization of highly redundant opinions. In Proceedings of the 23rd international conference on computational linguistics, 340–348. Association for Computational Linguistics.

[5] Yin, W., and Pei, Y. 2015. Optimizing sentence modeling and selection for document summarization. In Proceedings of the 24th International Conference on Artificial Intelligence, 1383–1389. AAAI Press.

[6] Cao, Z.; Li, W.; Li, S.; and Wei, F. 2016. Attsum: Joint learning of focusing and summarization with neural attention. arXiv preprint arXiv:1604.00125.

[7] Cheng, J., and Lapata, M. 2016. Neural summarization by extracting sentences and words. 54th Annual Meeting of the Association for Computational Linguistics.

[8] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Proceedings of Neural Information Processing Systems (NIPS).

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever.2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

[11] Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In Proceedings of ACL Workshop on Text Summarization Branches Out.

[12] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

[13] Romain Paulus, CaimingXiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.

[14] RamakanthPasunuru and Mohit Bansal. 2018. Multireward reinforced summarization with saliency and entailment. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), volume 2, pages 646–653.

[15] Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding generation for abstractive text summarization based on key information guide network. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), volume 2, pages 55–60.

[16] Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. arXiv preprint arXiv:1805.06266.

[17] Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. arXiv preprint arXiv:1805.11080.

[18] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. Bottom-up abstractive summarization. arXiv preprint arXiv:1808.10792, 2018.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and IlliaPolosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[20] Ramesh Nallapati, FeifeiZhai, and Bowen Zhou. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In AAAI, pp. 3075–3081. AAAI Press, 2017.

[21] Natalie Schluter. 2017. The limits of automatic summarization according to rouge. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers. Valencia, Spain, pages 41–45.

Authors

Yang Gu
Software Engineer at Suning USA AILab. 2 years research and development experience in computer vision and natural language processing.

Yanke Hu
Senior Cognitive/Machine Learning Engineer at Humana, responsible for deep learning research and AI platform development. 9 years research and development experience in software industry at companies majoring in navigation (TeleNav), business analytics (IBM), finance(Fintopia), retail (AiFi) and healthcare (Humana).