Chinese Text Summarization Based on Multi-Layer Attention

Jiecheng Jiang
Software School, Southeast University, Nanjing, China

Keywords: Text Summarization, Sequence to Sequence, Long Short-Term Memory, Deep Learning, Neural Network.

Abstract: The increasing volume of textual data on the internet leads individuals to spend more time sifting through and identifying crucial information within texts. Automatic summarization technology emerges as a method to extract key information from lengthy texts, reducing the time required for information retrieval in the age of information overload, thus garnering increased attention from researchers. Automatic summarization technology can be categorized into extractive summarization, which relies solely on the original text content and has its limitations, and generative summarization, which offers greater flexibility. However, challenges persist in maintaining sufficient information integrity during text initialization and ensuring the generation of high-quality summaries in Chinese. This paper proposes a Multi-Layer attention model to solve Chinese text summarization. This model obtains a 39.51 ROUGE-1 score and 37.25 ROUGE-L on the LCSTS validation dataset. In addition, a model acceleration method is proposed, which uses 1x1 convolution kernel to replace the linear layer in encoder-decoder to reduce size of the neural network and improve speed of summary generation.

1 INTRODUCTION

Text summarization involves condensing the essential information from a text to create a brief and logical summary while maintaining the core concepts and essence of the original text. It involves condensing a longer document, such as an article, research paper, or news story, into a shorter version, typically by selecting and rephrasing key sentences or paragraphs. Text summarization can be approached through two primary methods: Extractive Summarization (Dorr et al 2003) and Abstractive Summarization (Chopra et al 2016).

In Chinese text summarization, some of the technologies are based on TF-IDF (Tao and Chen 2020), some of the technologies are based on Word2vec (Chengzhang and Dan 2018), and most of technologies are based on attentional encoder-decoder models (Li 2020). Most technologies include only single-layer attention. The author proposes a new Multi-Layer attention model to solve Chinese text summarization. There is a small uptick in the parameter count and the training time remains almost unchanged. There has been a visible improvement in the performance of the model. This model obtains a 39.51 ROUGE-1 score and 37.25 ROUGE-L on LCSTS validation dataset. The author also proposes another optional solution, in which the author uses a 1x1 convolutional kernel to replace the linear layers in the encoder and decoder. The use of convolutional kernel can make the model smaller, but according to the experiment, the performance has hardly decreased.

2 METHODS

2.1 Word Segmentation

Chinese word segmentation is more challenging compared to English word segmentation because the segmentation of words in English involves directly utilizing spaces and punctuation marks. Using Jieba library can solve this problem (Xianwei et al 2019). The word segmentation process is as follows.

1) Employing a prefix dictionary for streamlined scanning of word graphs, which produces a directed acyclic graph (DAG) encapsulating various potential word formations from Chinese characters within a sentence.
2) Employing dynamic programming to systematically seek out the most probable path, determining the optimal segmentation combination by considering word frequencies.

Jiang, J.
Chinese Text Summarization Based on Multi-Layer Attention.
DOI: 10.5220/0012839000004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 565-568
ISBN: 978-989-758-690-3

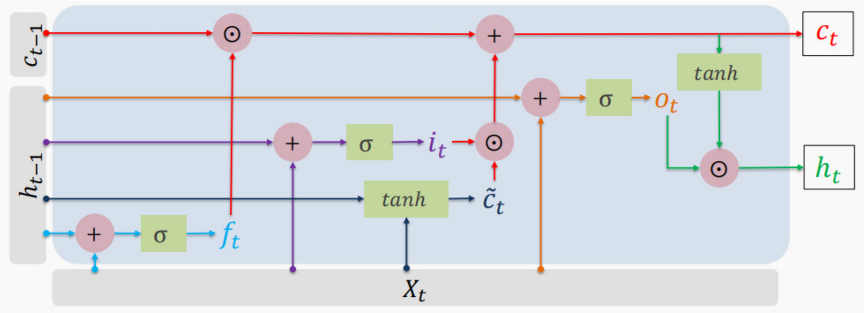

Figure 1: LSTM'S structure (Picture credit: Original).

3) In cases of unrecognized words, implementing an HMM model that leverages the intrinsic capability of Chinese characters to form words, with the application of the Viterbi algorithm.

2.2 Neural Network Model

The neural network model is based on the encoder-decoder network (Sutskever et al 2014 & Neubig and Graham 2017). Each unit is a bi-directional LSTM unit (Staudemeyer et al 2019). The LSTM structure is shown in Figure 1.

2.3 Encoder Module

At each decoding step t, e_{ti} is declared as the attention score of the hidden input state (Niu et al 2021). e_{ti} should be based on the hidden input state and the decoder state at each decoding step. While Romain Paulus (Paulus et al 2017) decides to use a bilinear function, the author chooses to use a gate-like unit to compute the attention score of the encoder attention score. The attention score can have multiple layers sharing some weights. The author conducted many experiments and found that when the W_s between two layers are set to be the same, the new model can reduce training parameters while ensuring good results. The first and second layers of the network are calculated as follows: (1) (2).

$$e_{ti}^{1} = V^{T}\tanh\left(W_{h1}h_{i}^{e} + W_{s}h_{t}^{d} + b_{1}\right) \qquad (1)$$

$$e_{ti}^{2} = V^{T}\tanh\left(W_{h2}h_{i}^{e} + W_{s}h_{t}^{d} + b_{2}\right) \qquad (2)$$

Where W_{h1} and W_{h2} are different weights to capture different information in the encoder. For example, the first level of attention concentrates more on the importance of word levels. The second level of attention concentrates on the importance of sentences or paragraphs. In the Chinese text summarization, it is important to learn from different lawyers. Two layers also share the same weight V for the purpose of simplicity. Add two layer's attention to get the attention associated with the input token.

$$e_{ti} = e_{ti}^{1} + e_{ti}^{2} \qquad (3)$$

In the experiment, the Chinese text summarization always has many repetitions, reducing readability. In order to punish the high-score tokens to avoid repetitions (Paulus et al 2017), new temporal scores are defined:

$$e'_{ti} = \begin{cases} exp(e_{ti}) & if\ t = 1 \\ \dfrac{exp(e_{ti})}{\sum_{j=1}^{n} exp(e_{ti})} & otherwise \end{cases} \qquad (4)$$

Then compute the normalized attention scores and get the context vector $c^{encoder}$.

2.4 Decoder Module

In order to capture more information in the previous decoding steps, multiple layers mechanism should also be used in the decoder attention to improve the quality of text summaries. For each decoding step t, use the following equations to compute the first layer attention. Similar to the encoder, the network layer is defined as follows: (5) (6).

$$e_{tj}^{1} = V^{T}\tanh\left(W_{h1}h_{j}^{d} + W_{s}h_{t}^{d} + b_{1}\right) \qquad (5)$$

$$e_{tj}^{2} = V^{T}\tanh\left(W_{h2}h_{j}^{d} + W_{s}h_{t}^{d} + b_{2}\right) \qquad (6)$$

Where W_{h1} and W_{h2} are different weights to capture different information in the encoder. Two layers also share the same weight V and W_s for the purpose of simplicity and effectiveness. Note that V, W, b in the decoder attention and the encoder attention are not the same. Integrate attention mechanisms from two additional layers to compute the attention score for the hidden output state.

$$\mathrm{E_{tj}} = \mathrm{e_{tj}^1} + \mathrm{e_{tj}^2} \qquad (7)$$

Using softmax function to normalize the focus score of the already generated output token at each moment and get the context vector $\mathrm{c^{decoder}}$.

2.5 1x1 Convolutional Module

A 1x1 convolution, as described in reference (Badrinarayanan et al 2017), possesses unique characteristics allowing it to serve purposes such as reducing dimensionality, creating efficient low-dimensional embeddings, and applying non-linear transformations subsequent to convolutions. In the Chinese text summarization, the encoder and the decoder attention are linear layers. 1 x 1 convolutional kernel can be used to replace linear layers. To reduce parameters and apply non-linearity in the model, when using 1 x 1 convolutional kernel, only the first layer attention should be used.

The next token mainly depends on $\mathrm{c^{encoder}}$, $\mathrm{c^{decoder}}$ and $\mathrm{h_t^d}$. The probability distribution is calculated by the following equation.

$$\mathrm{softmax}\left(\mathrm{W}\left[\mathrm{h_t^d}\,\middle|\mathrm{c^{encoder}}\middle|\, c^{decoder}\right] + b\right) \qquad (8)$$

3 EXPERIMENT

3.1 Dataset

The author assesses the model's performance using LCSTS, a dataset for large-scale Chinese short text summarization derived from Sina-Weibo, a Chinese microblogging platform. LCSTS comprises more than 2 million authentic Chinese short texts, each accompanied by a brief summary provided by the respective author. The dataset is curated by the Intelligent Computing Research Center at the Shenzhen Graduate School of Harbin Institute of Technology.

The LCSTS has been divided into three parts. The training dataset has 1048575 short text, summary pairs. The validation dataset has 10666 short text, summary pairs which is human labeled. The test dataset has 1103 short text, summary pairs.

3.2 Results

The author evaluates the Convolutional-Layer mechanism in the encoder attention and in both encoder and decoder attention and the Multi-Layer attention mechanism on the dataset.

After doing supervised learning, the experiment chooses a relatively good model (evaluate on validation set) and use the reinforcement learning to improve the performance of the model. Romain Paulus (Paulus et al 2017) define a mixed objective function.

$$\mathrm{L_{mixed}} = \gamma\, \mathrm{L_{rl}} + \left(1 - \gamma\right)\mathrm{L_{ml}} \qquad (9)$$

Where choose γ=0.75 and it can receive relatively good results. The overall results after doing ML+RL are shown in the following two tables. The evaluation indicators act on the validation and testing sets.

As shown in Table 1, the Multi-Layer attention achieves better ROUGE scores on the validation dataset compared to other three models, especially in ROUGE-1. Multi-Layer attention can reach 39.51 in ROUGE-1. The convolution-Layer mechanism can make model smaller, but its performance is only slightly inferior to the original linear-layer model.

Table 1: Quantitative results for 4 models on validation dataset.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Linear-Layer	34.06	16.46	33.83
Convolutional-Layer in encoder	32.99	**17.15**	31.81
Convolutional-Layer in both	33.28	15.82	31.66
Multi-Layer attention	**39.51**	**17.59**	**37.25**

Table 2: Quantitative results for 4 models on test dataset.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Linear-Layer	31.87	15.47	30.93
Convolutional-Layer in encoder	31.09	13.07	30.55
Convolutional-Layer in both	35.18	**17.39**	**34.19**
Multi-Layer attention	**36.81**	16.39	33.74

As shown in the Table 2, the Multi-Layer attention achieves better ROUGE-1 scores. When it comes to ROUGE-2 and ROUGE-L, the Convolutional-Layer mechanism in both the encoder and the decoder attention achieves better scores compared to the linear layer.

The Multi-Layer attention performed well on the ROUGE-1 both on the validation dataset and test dataset. And the improvement is relatively visible. The training time of the Multi-Layer attention model has hardly increased, which means this method has practical application significance.

4 CONCLUSION

This paper proposes a new Multi-Layer attention model to solve Chinese text summarization. There is a small uptick in the parameter count and the training time remains almost unchanged. But there has been a visible improvement in the performance of the model. This model obtains a 39.51 ROUGE-1 score and 37.25 ROUGE-L on LCSTS validation dataset. Attentions in different layers can have different learning task in the input sequence, while sharing the same weight in the decoder’s hidden state. The author also proposes another optional solution, in which the author uses a 1x1 convolutional kernel to replace the linear layers in the encoder and decoder. The use of convolutional kernel can make the model smaller, according to the experiment, the performance has hardly decreased.

REFERENCES

B. Dorr, D. Zajic and R. Schwartz, “Hedge trimmer: A parse-and-trim approach to headline generation,” In *Proceedings of the HLT-NAACL 03 on Text summarization workshop* (2003), pp. 1–8.

S. Chopra, M. Auli, A. M Rush and SEAS Harvard, “Abstractive sentence summarization with attentive recurrent neural networks,” in *Proceedings of NAACL-HLT16* (2016). pp. 93–98.

Z. Tao and C. Chen. "Research on automatic text summarization method based on tf-idf," Advances in Intelligent Systems and Interactive Applications: Proceedings of the 4th International Conference on Intelligent, Interactive Systems and Applications (IISA2019) 4. Springer International Publishing, 206-212 (2020).

X. Chengzhang, and L. Dan, Journal of Physics: Conference Series. 976(1), 12006 (2018).

Li, Zhixin, "Text summarization method based on double attention pointer network," in *IEEE Access*. 11279-11288 (2020).

Z. Xianwei, et al, Journal of Physics: Conference Series. 1302(2), 22010 (2019).

Sutskever, Ilya, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems 27* (2014).

Neubig, Graham, "Neural machine translation and sequence-to-sequence models: A tutorial," arXiv preprint arXiv:1703.01619 (2017).

Staudemeyer, C. Ralf, and E. R. Morris, "Understanding LSTM--a tutorial into long short-term memory recurrent neural networks," arXiv preprint arXiv:1909.09586 (2019).

Z Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," in *Neurocomputing 452* (2021). pp. 48-62.

Paulus, Romain, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," arXiv preprint arXiv:1705.04304 (2017).

Badrinarayanan, Vijay, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in I*EEE transactions on pattern analysis and machine intelligence* (2017). pp. 2481-2495.