

A Multi-Layer Attention Network based on Self-Attention Mechanism

Yifan Hu^{1,a}, Zhen Wang^{1,b}, Chengguo Lv^{1,c}

¹Department of Computer Science Heilongjiang University Harbin, Harbin, China.

^ahyf1009nlp@163.com, ^bwangzhennlp@163.com, ^ccgl1257@163.com

Abstract

The aspect-level sentiment analysis task aims to determine the sentiment polarity of different target objects in a sentence. The previous model for aspect-level sentiment analysis tasks used the RNN network or the Attention network modeling context to connect with the target. Although the RNN network can establish the long-term dependency of sentences, it is difficult to implement in parallel. However, a single self-attention mechanism cannot capture the position information of the sequence. For this reason, this paper proposes a multi-layer attention network based on self-attention, so that the model can obtain context information that is more closely related to the context. And through experiments and analysis, it is proved that our model can achieve good results when facing common aspect-level sentiment analysis data sets.

Keywords

Aspect-level Sentiment Analysis; Attention.

1. Introduction

Aspect-level sentiment analysis is a classic task in the field of NLP. It aims to infer the sentiment polarity of different objects in a sentence. It is a fine-grained sentiment analysis task. For example, "The clothes are very good, but the logistics of the merchant is a bit slow". In this sentence, the emotional polarity of "clothing" is positive, while the emotional polarity of "logistics" is negative. It can be seen that aspect-level sentiment analysis tasks have a very wide range of applications in life, so the research on aspect-level sentiment analysis will greatly improve the ability of machines to recognize complex human languages.

The key to the aspect-level sentiment analysis task is to make the model obtain more accurate and effective contextual information for the target word. Previous studies have tried, for example, RNN or attention methods to model aspect words and context. However, as we all know, although RNN can model the long-term dependence of sentences, it is difficult to perform parallel processing. The self-attention [1] mechanism solves the long-term dependency of sentences and has very good parallelism. The performance of the model based on the self-attention mechanism also exceeds the performance of the RNN model. However, the previous attention network is purely based on the self-attention mechanism and cannot capture the position information of the sentence. In the previous research, the Attention Over Attention (AOA) [2] network can calculate the importance weight of the context for the target word from the target to the context and the context to the target. Inspired by the AOA network, this article designs a multi-layer attention network based on the self-attention mechanism for aspect-level sentiment analysis. We abbreviate it as MLAN.

The main contributions of this work are as follows:

a) We designed a multi-layer attention network based on self-attention to model the internal connection between context and target.

b) Combining the features captured by self-attention and attention over attention, the model achieves better results.

2. Related Work

For sentiment analysis tasks, solutions include traditional machine learning methods and neural network methods. In recent years, with the introduction of transformers and BERT and their powerful capabilities, it is nothing new to apply them to different natural language processing tasks. Maintaining the Integrity of the Specifications.

Aspect-level sentiment analysis tasks are no exception. Devlin, Jacob, et al (2018) proposed BERT-SPC [3], which uses BERT to fine-tune aspect-level sentiment analysis tasks. Song, youwei, et al (2019) proposed AEN-BERT [4], which uses an attention-based coding method to model context and target words. Xu, Hu, et al (2019) proposed BERT-PT [5]. Rietzler, Alexander, et al (2019) proposed BERT-ADA [6], which uses a self-supervised domain-specific BERT model fine-tuning and supervised task-specific fine-tuning two-step process to deal with sentiment analysis.

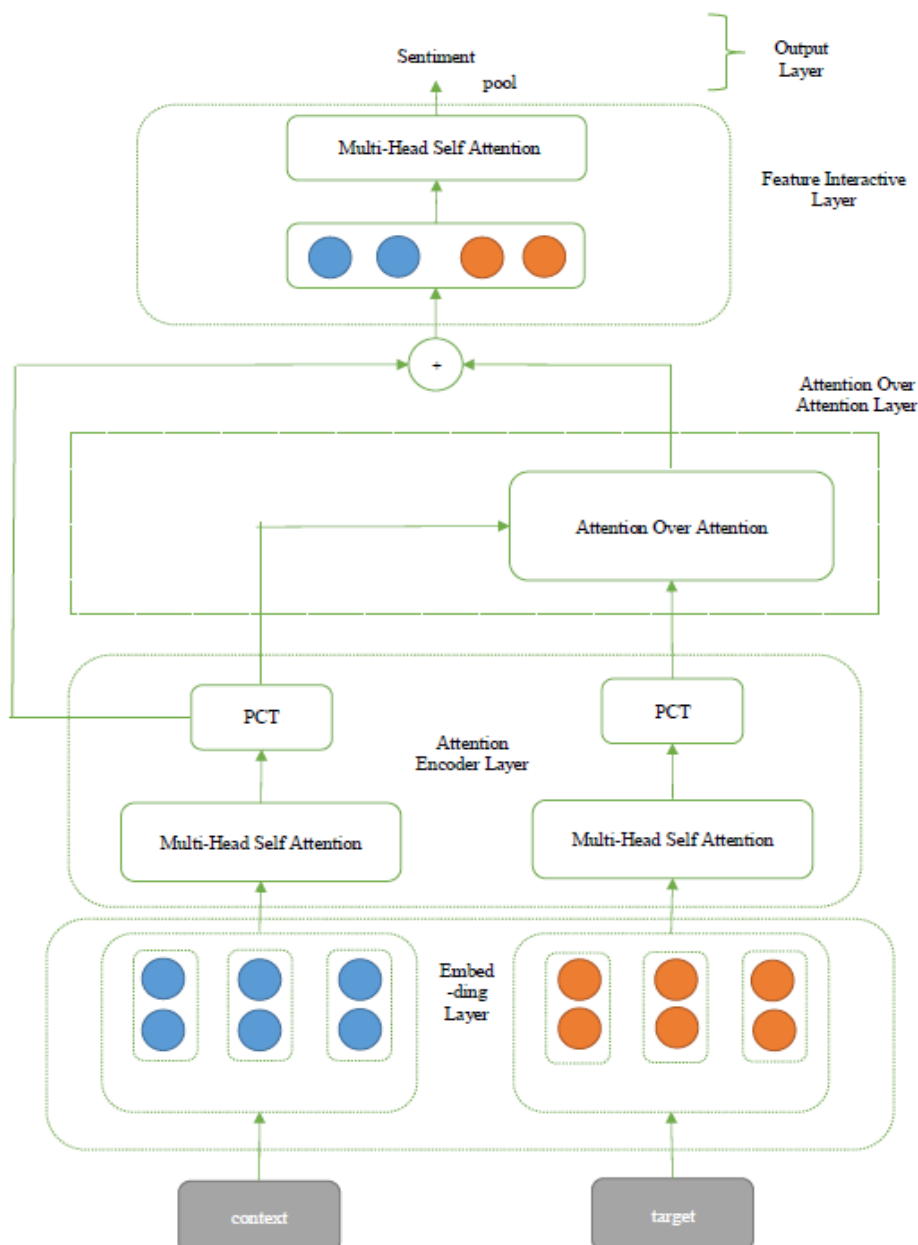


Figure 1. Overall architecture of the MLAN

3. Methods

The input of aspect-level sentiment analysis task usually consists of context sequence and aspect sequence. In this article, unlike other attention networks, we have designed multiple different attention layers to obtain more accurate contextual information related to the target word. Compared with previous attention-based models, our model shows better performance. Due to the advantages of Bert, we adopted the Bert model to model the input sentence sequence. Because the Attention-Over-Attention mechanism can not only draw attention from context to aspect, but also from aspect to context. We designed a feature interaction layer so that the model can integrate the weights learned by the attention encoding module and the attention-over-attention module. Our model has made some improvements to address the deficiencies in the Attention-Over-Attention and Multi-Head Attention mechanisms [1] so that the model can obtain more accurate context information. In the face of the semeval14 dataset [7], quite good results have been achieved. Figure 1 shows the overall structure of our proposed model.

3.1. Embedding Layer

Assuming that any sentence $s=\{w_1, w_2, \dots, w_i, \dots, w_j, \dots, w_n\}$ is an input sequence including a context sequence, this sequence consists of n words, $t=\{w_i, w_{i+1}, \dots, w_{i+m-1}\}$ is the input sequence containing aspect words, and t is the subsequence of s , including m words. Since the Bert model has very good language modeling capabilities, in our model, we obtain the vector representation of the input word sequence through the pre-trained Bert model. Among them, X_{BERT} is the Bert vector representation of the input sequence, and T_{BERT} is the BERT vector representation of the target sequence.

$$X_{BERT} = BERT(s) \quad (1)$$

$$T_{BERT} = BERT(t) \quad (2)$$

3.2. Attention Encoder Layer

The attention coding layer includes two modules: Multi-Head Self-Attention(MHA) module and Point-wise Convolution Transformation (PCT) module.

The self-attention mechanism can obtain the global and local relationship of the sentence, while reducing the dependence on external information, and has a good effect in capturing the internal correlation of data or features. Multi-head attention can perform multiple attention functions based on self-attention. After obtaining the Bert vector representation of the word, we hope to obtain the connection within the context through the multi-head attention mechanism. For this reason, we designed a multi-headed self-attention layer to calculate the attention weight of context words.

$$Attention(Q, K, V) = \frac{Softmax(Q \cdot K^T)}{\sqrt{dk}} \cdot V \quad (3)$$

$$MHSA(X_{BERT}) = Tanh(\{H_0; H_1; \dots; H_h\} \cdot W^{MH}) \quad (4)$$

Among them, h represents the number of attention heads, H_i is the vector representation learned by each attention head, and ";" means vector cascade.

The input representation of the PCT layer is the output representation of the MHSA encoder.

$$PCT(O_{mhsa}) = ReLU(O_{mhsa}) \cdot w + b \quad (5)$$

$$O_s = PCT(O_{mhsa}^{s-embed}) \quad (6)$$

$$O_t = PCT(O_{mhsa}^{t-embed}) \quad (7)$$

We use O_s and O_t to denote the context and target sequence obtained through the Attention Encoder layer, respectively.

3.3. Attention Over Attention Layer

In the AOA module, we calculate the word importance weights from context to target word and target word to context. According to the original text of AOA, we first calculate the interaction matrix between the context representation and the target representation. Then, we do softmax on the matrix I by row and column respectively and get the attention α from the target to the context and the attention β from the context to the sentence respectively.

$$I = O_s \cdot O_t^T \quad (8)$$

$$\alpha = \text{Softmax}_i(I_{ij}) \quad (9)$$

$$\beta = \text{Softmax}_j(I_{ij}) \quad (10)$$

$$\bar{\beta}_j = \frac{1}{n} \sum_i \beta_{ij} \quad (11)$$

$$\text{Attention}_{AOA} = \alpha \cdot \bar{\beta}^T \quad (12)$$

We weight the obtained attention weights on the input sequence to obtain the final output O_{AOA} of the AOA module.

$$O_{AOA} = O_s^T \cdot \text{Attention}_{AOA} \quad (13)$$

3.4. Feature Interactive Layer

In the original text of AOA, the author chose to feed the obtained importance weights as the final extracted sentence features into the linear layer. And we designed an interactive feature learning layer so that the model can interactively learn the weights learned by the attention-over-attention module and the weights learned by the Multi-Head Self-Attention.

The feature learning layer connects O_{AOA} and O_s and then encodes with MHSA.

$$O_{concat} = \{O_{AOA}; O_s\} \quad (14)$$

$$O_{FIL} = \text{MHSA}(O_{concat}) \quad (15)$$

3.5. Output Layer

The output layer feeds the features learned by the feature interaction layer into the pooling layer and predicts the emotional polarity of the target through the softmax function.

$$X_{pool} = \text{POOL}(O_{FIL}) \quad (16)$$

$$Y = \text{Softmax}(X_{pool}) \quad (17)$$

4. Experiment

4.1. Datasets and Hyperparameters

In our work, we used two datasets from SemEval14 for experiments, namely the Laptop dataset and the Restaurant dataset. Among them, the Laptop data set has more implicit sentiment sentences than the restaurant data set, the number is smaller, and the difficulty is greater. The dataset includes three emotional polarities: positive, neutral, and negative.

In this research, we use Accuracy and Macro-F1 indicators to evaluate the performance of the model. In the experiments, we fine-tuned the pre-trained BERT, the number of pre-trained bert layers is set to 768, and the number of hidden layers is set to 300. Besides, the initialization of our model weight refers to Glorot initialization (Glorot and Bengio, 2010) when the L2 regularization term is set to 10^{-5} , dropout is set to 0.3, and the Adam optimizer is used to update the parameters, the effect of the model reaches the best state. After several fine adjustments, we recorded the number of iterations and batch-size that made the model state the best.

Table 1. The emotional polarity distribution of the data set used in this work table shows the number of training and testing examples with different emotional polarities in the two datasets

Polarity	Restaurant		Laptop	
	Train	Test	Train	Test
Positive	2164	728	994	341
Neutral	637	196	464	169
Negative	807	196	870	128

Table 2. Accuracy and macro-f1 of laptop data set prediction results under different training parameters

Accuracy of prediction	Macro-F1	Epochs	Batch-size
0.8025	0.7545	20	16
0.8182	0.7840	25	16
0.7821	0.7330	30	16
0.7947	0.7541	25	32
0.8166	0.7828	30	32
0.7994	0.7543	35	32

Table 3. Accuracy and macro-f1 of restaurant data set prediction results under different training parameters

Accuracy of prediction	Macro-F1	Epochs	Batch-size
0.8527	0.7789	10	16
0.8509	0.7917	15	16
0.8598	0.7868	20	16
0.8464	0.7717	10	32
0.8607	0.7885	15	32
0.8491	0.7846	20	32

4.2. Comparison Models

In order to evaluate the performance of our model, we used two data sets to test the model and compared it with multiple baseline models. The results show that our model has a good predictive ability for aspect sentiment.

IAN [8] uses two LSTM networks to model sentences and aspect words.

RAM [9], a multi-attention mechanism is used to capture the connections between words in a longer sentence, and the results of multiple attentions are combined non-linearly with recurrent neural networks, which enhances the model's ability to handle problems in parallel.

AOA [2] adds another layer of Attention to the original Attention to show the importance of each attention.

TNET [10] uses the Bi-RNN layer to generate a representation of words containing contextual information, and then convert these representations based on the aspect, and finally use CNN as a feature extractor to give the classification results.

MemNet [11], a memory module composed of a recursive unit and a gating unit is proposed. Because the memory module has a gating mechanism, it can retain the required information.

BERT-SPC [3] is a fine-tuning of the BERT model for aspect-level sentiment analysis tasks.

BERT-PT [5] based on BERT, a new post-training method is explored to enhance the fine-tuning performance of BERT for aspect sentiment analysis.

AEN-BERT [4] based on the connection between the context of attention coding and aspect words, the problem of label unreliability is raised, and label smoothing regularization is introduced.

BERT-ADA [6] using self-supervised domain-specific fine-tuning of the BERT language model and supervised specific task fine-tuning to deal with aspects of sentiment analysis tasks.

Table 4. Forecast accuracy of several models

	Models	Restaurant		Laptop	
		Accuracy	Macro-F1	Accuracy	Macro-F1
RNN baselines	IAN	0.7860	-	0.7210	-
	RAM	0.8023	0.7080	0.7449	0.7135
	AOA	0.8120	-	0.7450	-
	Tnet	0.8079	-	0.7601	-
Non-RNN baselines	Memnet	0.8095	-	0.7221	-
Bert baseline	BERT-SPC	0.8446	0.7698	0.7899	0.7503
	BERT-PT	0.8495	-	0.7807	-
	AEN-BERT	0.8312	0.7376	0.7993	0.7631
	BERT-ADA	0.8789	-	0.8023	-
Ours	MLAN	0.8607	0.7885	0.8182	0.7840

5. Conclusion

In our article, we designed a multi-layer attention network based on self-attention mechanism and proved that the combination of AOA network and Multi-Head Self-Attention mechanism will enable the model to obtain more effective context information, and the effect is better than a single Multi-head self-attention or AOA network. Finally, experiments and analysis prove the validity of the model.

Acknowledgments

This work is due to the great help and support of the tutor and seniors, thank them for their guidance.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, no. Nips, pp. 5999–6009, 2017.
- [2] Binxuan Huang, Tanglan Qu, and Kathleen M. Caley, "Aspect Level Sentiment Classification with Attention-over-Attention Neural Network," 2018.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [4] Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao, "Attentional Encoder Network for Targeted Sentiment Classification," 2019.
- [5] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu, "Bert Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis," *arXiv preprint arXiv:1904.02232*, 2019.
- [6] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl, "Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification," *arXiv preprint arXiv:1908.11860*, 2019.
- [7] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.
- [8] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang, "Interactive Attention Networks for Aspect-Level Sentiment Classification," *arXiv preprint arXiv:1709.00893*, 2017.

- [9] Peng Chen, Zhongqian Sun, Lidong Bing and Wei Yang, "Recurrent Attention Network on Memory for Aspect Sentiment Analysis," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.
- [10] Xin Li, Lidong Bing, Wai Lam, and Bei Shi, "Transformation Networks for Target-Oriented Sentiment Classification," arXiv preprint arXiv:1805.01086, 2018.
- [11] Duyu Tang, Bing Qin, and Ting Liu, "Aspect Level Sentiment Classification with Deep Memory Network," Conference on Empirical Methods in Natural Language Processing 2016:214224, 2016.
- [12] Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han, "LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification", 2019.
- [13] Feifan Fan, Yansong Feng, and Dongyan Zhao, "Multi-grained Attention Network for Aspect-Level Sentiment Classification", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.
- [14] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang, "Pre-trained Models for Natural Language Processing: A Survey" arXiv preprint arXiv:2003.08271, 2020.