Research Article

Multitask Learning for Aspect-Based Sentiment Classification

Chunhua Yao,¹ Xinyu Song ,² Xuelei Zhang ,² Weicheng Zhao ,² and Ao Feng ²

¹The 30th Research Institute of China Electronics Technology Group Corporation, Chengdu, China

²Chengdu University of Information Technology, Chengdu, China

Correspondence should be addressed to Xinyu Song; xinyu.s@foxmail.com

Received 20 July 2021; Revised 8 October 2021; Accepted 8 November 2021; Published 29 November 2021

Academic Editor: Sze-Teng Liong

Copyright © 2021 Chunhua Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aspect-level sentiment analysis identifies the sentiment polarity of aspect terms in complex sentences, which is useful in a wide range of applications. It is a highly challenging task and attracts the attention of many researchers in the natural language processing field. In order to obtain a better aspect representation, a wide range of existing methods design complex attention mechanisms to establish the connection between entity words and their context. With the limited size of data collections in aspect-level sentiment analysis, mainly because of the high annotation workload, the risk of overfitting is greatly increased. In this paper, we propose a Shared Multitask Learning Network (SMLN), which jointly trains auxiliary tasks that are highly related to aspect-level sentiment analysis. Specifically, we use opinion term extraction due to its high correlation with the main task. Through a custom-designed Cross Interaction Unit (CIU), effective information of the opinion term extraction task is passed to the main task, with performance improvement in both directions. Experimental results on SemEval-2014 and SemEval-2015 datasets demonstrate the competitive performance of SMLN in comparison to baseline methods.

1. Introduction

Sentiment analysis is one of the fundamental tasks in natural language processing and has received an increasing level of attention in recent years. Aspect-level sentiment classification focuses on fine-grained sentiment analysis and is widely applied to automatic processing tasks for online review text. The purpose of this task is to determine the emotional polarity of entities in each aspect of a review piece [1–4], with each entity consisting of one or multiple words. The number of aspect terms in a sentence is arbitrary [5–9], and each aspect may carry a different sentiment polarity. Within the sentence “I love this program, it is superior to windows movie maker” in Figure 1, “program” and “windows movie maker” are two separate aspect terms, but they carry positive and negative emotions, respectively. In the example above, “love” and “superior” are defined as opinion terms. It can be observed that the emotional polarity of aspect words comes from their corresponding opinion words.

Existing algorithms for aspect-level sentiment analysis are mainly divided into feature engineering methods and deep learning models. For the methods based on feature

engineering, the main idea is to design a series of handcraft features, and a traditional classifier is trained to achieve high emotion classification accuracy [5, 10, 11]. This class of methods consumes a lot of manpower, and the vocabulary dependency on individual scenarios makes it difficult to generalize.

Models based on deep neural networks have better potential in solving these issues. Through multilayer neural networks, low-dimensional vectors representing term semantics can be effectively trained without complex feature engineering process. These embedding representations become the input of downstream neural networks, in order to identify the emotional polarity of target words. Satisfactory results have been achieved through various target-dependent sentiment mechanisms [5–9]. These Long Short-Term Memory (LSTM)-based methods take static word vectors such as Word2Vec and GloVe as input and use the feature representation of entity words for sentiment classification. They simply fuse contextual information into the representation of the target word, without considering their semantic correlation. Recent research efforts apply the attention mechanism to consider interaction between the

FIGURE 1: An example of consumer review that contains aspect and opinion terms.

aspect words and their context. A variety of complex structures were designed to calculate attention weights between aspect words and their context [12–15]. As most aspect-level datasets do not come in large scale, the risk of overfitting greatly increases with the model size and complexity. As a result, methods based on the attention mechanism tend to make more mistakes when mining deeper features.

In recent years, multitask learning (MTL) becomes an active research area in machine learning, which improves the generalization performance of a task by jointly training other related tasks [16]. Due to the success of MTL [17], there are several NLP models based on neural networks that adopt this mechanism [18–20]. By using shared representations to learn semantically related tasks in parallel, MTL captures the correlation between tasks, improving the generalization ability of the model under certain circumstances. The multitasking architecture of these models contains shared lower layers to train their common features, and the remaining layers are customized to handle different tasks. For aspect-level sentiment classification, [21, 22] have made some attempts in MTL, and their study shows that joint training with document-level sentiment classification can significantly improve performance at the aspect level. Yu and Jiang [23] design an auxiliary task that is highly relevant to sentiment analysis, which predicts whether the input sentence contains positive or negative words.

In this study, we propose a Shared Multitask Learning Network (SMLN), which employs opinion word extraction as an auxiliary task and trains it together with aspect-level sentiment analysis. This task is an upstream step that extracts key opinion terms, and its performance also affects the accuracy of the main task. The pretrained BERT model is used as the underlying structure shared by the two tasks. SMLN introduces a new feature sharing mechanism, Cross Interaction Unit (CIU), to facilitate the information exchange between the main task and the auxiliary task. Specifically, CIU consists of multiple groups of attention mechanisms, integrating the information of the two tasks from different viewing angles. With extensive experiments on the SemEval-2014 and SemEval-2015 datasets, the results indicate that our SMLN model outperforms other baseline methods in terms of classification accuracy. For a fair comparison, some of the baseline methods are built on the same BERT-based representation, so the performance boost is originated from the multitask setting and information sharing mechanism.

Main contributions of the work include the following: (1) a multitask learning method customized for fine-grained sentiment analysis, which utilizes additional opinion word information to improve the learning performance of the current task and reduce the risk of overfitting; (2) a multitask

sharing mechanism to accomplish multiview information transfer between tasks.

2. Related Work

In this section, relevant research on aspect-level sentiment analysis and multitask learning is reviewed, covering both traditional neural networks and more recent BERT-based model.

2.1. Conventional Neural Network. In aspect-level sentiment analysis, it has been agreed among researchers that context words have different influence on the sentiment polarity of multiple targets in the opinion sentence. When a learning system is built for sentiment classification, the most important task is to integrate the relationship between each target and its context. Vo and Zhang [7] divide the original sentence into the target words, the left context, and the right context and use different networks to extract their features. Tang et al. [8] propose a target-dependent LSTM model. In order to represent the characteristics of the aspect word more accurately, they use two LSTMs to encode the previous context and the next context including the target itself.

Previous studies try to establish the connection between the target words and their context, but the interaction information is hard to capture with canonical methods. Wang et al. [12] propose an LSTM network based on the attention mechanism, which is used for aspect-level sentiment classification. When different aspects are involved, this model automatically directs its attention to different parts of the sentence. Li et al. [13] design an end-to-end structure, constantly focusing on an aspect term and its context. Ma et al. [14] believe that entities and corresponding contexts can be reasonable, and calculate attention weights for target and context, respectively. Fan et al. [15] propose a multi-grained attention network. In addition to the attention calculation of the aspect words and the overall context, they also introduce fine-grained attention. The purpose is to describe the influence of an aspect on its context, or context on aspect words in the reverse direction. Tang et al. [24], Chen et al. [25], and Zhu and Qian [26] propose a readable and writable external memory module, which shows the contribution of each word to the final sentiment classification.

2.2. Multitask Learning. All methods mentioned above focus on a single task and obtain acceptable performance in aspect-level sentiment analysis. If related tasks are built on the same text representation and language modeling, it is possible to further improve the learning performance of the main task. In recent years, studies have adopted multitask learning methods to handle fine-grained sentiment analysis. Yu and Jiang [23] design an auxiliary task to determine if an input sentence contains positive or negative sentiment words. This auxiliary task is closely related to the main task of sentiment analysis. They propose to train the sentiment analysis task and the hidden feature representation task together, and the auxiliary task helps in generating more

representative embedding for sentiment analysis sentences. He et al. [21] propose a multitask learning method that jointly trained the document-level and aspect-level sentiment classification tasks. Leveraging the information at the document level, current aspect model's limitation is alleviated with the introduction of the larger datasets. He et al. [22] employ an information delivery mechanism so that identical implicit expressions can be shared among multiple tasks in an iterative manner. This multitask model can utilize more global knowledge to improve the accuracy of sentiment analysis.

2.3. BERT-Based Networks. Traditional neural networks include structures like LSTM or multilayer CNN as encoders and use word vectors generated by Word2Vec or GloVe. However, the performance of these models is limited by the static nature of word vectors, and the downstream networks cannot break the performance bottleneck imposed by the text representation. In order to solve this problem, recent research has focused on large-scale pretrained attention models like ELMo, GPT, and BERT, and gains have been observed in many text applications with the richer representation. Sun et al. [27] propose four methods of constructing auxiliary sentences with the aspect term, feeding auxiliary sentences together with the original sentence as the input to the BERT model. This method transforms aspect-level sentiment analysis into a sentence pair classification task. Xu et al. [28] learn domain knowledge based on large-scale pretraining with BERT in the same domain as the original dataset. Gao et al. [29] design an intuitive method to use the feature expression of the target word on aspect-level sentiment classification, with slight modification of the BERT model. Zhou et al. [30] use the graph convolutional network (SK-GCN) model of grammar and knowledge to enhance the representation of sentences for given aspects. Song et al. [31] propose a semantics perception and refinement network (SPRN) for sentiment analysis based on aspects. Local semantic features are extracted by multi-channel convolution operation. They use gated networks to enhance aspect and context connections while filtering noise.

3. The Approach

The SMLN architecture is shown in Figure 2. In this section, we will elaborate on the details of the SMLN structure for aspect-level sentiment classification. It starts with the definition of the main task and auxiliary tasks, together with the necessary notations. Then, the BERT-based representation is included as a shared layer. CIU, which is the unit that establishes interaction between the main and auxiliary tasks, is introduced after that.

3.1. Problem Definition and Notations. s_i denotes a sentence from the training dataset, which consists of a sequence of tokens:

$$s_i = [w_1, w_2, \dots, w_i, \dots, w_n]. \quad (1)$$

Sentence s_i includes target words t that need to be annotated with their polarity of sentiment, and opinion words o that carry the corresponding emotion information. A target

$$t = [w_i, w_{i+1}, \dots, w_{i+m-1}], \quad (2)$$

contains m words; the opinion terms

$$o = [w_j, w_{j+1}, \dots, w_{j+k-1}], \quad (3)$$

include k words; and t and o are both subsequences of s . For the main task, aspect-level sentiment classification (ASC), its goal is to determine the emotional polarity of the target word t in the sentence s_i . Available tags include “positive,” “negative,” and “neutral.” For the auxiliary task, opinion terms extraction (OTE), its aim is to extract all the opinion terms appearing in a sentence. For simplicity, we treat OTE as a sequence tagging problem, with the BIO tagging scheme. Specifically, we use three categories of tags: $Y^{\text{OTE}} = \{B, I, O\}$ indicating the beginning and interior of the opinion term, and other words, respectively. For example, for the sentence “The screen is large and crystal clear with amazing colors”, its opinion extract label is shown in Table 1.

3.2. Shared Layer. The input embedding layer maps the original text representation into a high-dimensional vector space. The pretrained BERT model is employed to obtain embedding representation of each word with fine-tuning capability in the original Transformer network. BERT [32] is one of the leading language representation models, which uses a bidirectional Transformer [33] network to pretrain a language model on a large text corpus, and the pretrained representation can be fine-tuned on other tasks. The task-specific BERT design is able to represent either a single sentence or a pair of sentences as a consecutive array of tokens. For a given token, its input representation is constructed by adding up its corresponding token, segment, and position embeddings. For a typical classification task, the first word of the sequence is identified with a unique token [CLS], and a fully connected layer is attached at the [CLS] position of the last encoder layer. The last layer is usually softmax which completes the classification task.

BERT has two parameter intensive settings:

BERT_{base}: The number of Transformer blocks is 12, the hidden layer size is 768, the number of self-attention heads is 12, and the total number of parameters for the pretrained model is 110M.

BERT_{large}: The number of Transformer blocks is 24, the hidden layer size is 1024, the number of self-attention heads is 16, and the total number of parameters for the pretrained model is 340M.

The BERT_{large} model requires considerably more memory than BERT_{base}. As a result, the maximal batch size for BERT_{large} is so small on a single GPU with limited memory that it actually hurts the model accuracy, regardless of the learning rate [32]. Therefore, we use BERT_{base} as our baseline model, with modifications that do not significantly increase the model size.

FIGURE 2: Framework of the proposed SMLN.

TABLE 1: An auxiliary task training instance with gold opinion extract labels.

Input	The	Screen	Is	Large	And	Crystal	Clear	With	Amazing	Colors
Label	O	O	O	B	O	B	I	O	B	O

Following annotations in the previous section, a sentence with size n contains a target/aspect that is composed of m terms. BERT uses WordPiece [34] as its tokenizer. After the multilayer bidirectional Transformer network, the word vector matrix Sr of the sentence S is represented by the hidden status of the last layer.

$$Sr = [x_0, x_1, x_2, \dots, x_i, \dots, x_n, x_{n+1}], \quad (4)$$

$Sr \in R^{(n+2) \times d}$, where d is the dimension of hidden state. x_0 is the vector of the sentence classification mark [CLS], and $x_{(n+1)}$ is the word vector of the sentence separator [SEP]. Then, we use two Bi-LSTM networks to decompose Sr . The outputs of the two networks are denoted as S_p and S_o , which focus on ASC and OTE, respectively.

3.3. Cross Interaction Unit. When generating the representation with two independent Bi-LSTMs, the information of ASC and that of OTE, as two individual tasks, are separated from each other. However, the reality is that the two parts are closely related. For example, when ‘love’ appears around an aspect term, its polarity is likely to be positive. The Cross Interaction Unit (CIU) is designed to exchange information between these tasks, mining opinion terms and identifying aspect sentiment polarity in a cooptimization manner. The CIU architecture is shown in Figure 3.

A basic CIU is composed of a pair of attention modules: polarity attention and opinion attention. We define the output $S_p = \{p_0, p_1, \dots, p_n, p_{n+1}\}$ and $S_o = \{o_0, o_1, \dots, o_n, o_{n+1}\}$ of two Bi-LSTM networks to represent the distribution of sentiment feature and opinion feature, respectively, where $p_i \in R^{2d}$ and $o_i \in R^{2d}$ are representations of the i -th token w_i . P and O are input to the emotional attention module, in which

we first calculate the composition vector $\alpha_{ij}^p \in R^K$ between P and O through a tensor operation:

$$\alpha_{ij}^p = f_p(p_i, o_j) = \tanh((p_i)^T G_p o_j), \quad (5)$$

where $G_p \in R^{K \times 2d \times 2d}$ is a 3-dimensional tensor. A tensor operator can be viewed as multiple bilinear terms that are capable of modeling more complicated compositions between two vectors [35]. K is a hyperparameter representing the number of G_p channels. Each channel of G_p is a bilinear term that can extract specific information. A larger number of K represents the complicated intrinsic correlation between sentiment classification features and opinion word extraction features. As the value of K increases, more information is extracted together with higher complexity.

After obtaining the composition vectors, the attention score e_{ij}^p for token w_i is calculated as

$$e_{ij}^p = v^p \alpha_{ij}^p. \quad (6)$$

Here, $v^p \in R^K$ can be seen as a weight vector to measure each value of the composition vector. e_{ij}^p is a scalar value that composes a matrix E_p . A higher score for e_{ij}^p indicates that the current sentiment feature of the i -th word captures more information from the opinion expression of the j -th word. Finally, we fuse the sentiment feature P and opinion feature O generated by the original Bi-LSTM as follows:

$$\begin{aligned} S'_p &= S_p + \text{softmax}_r(E_p) S_o, \\ S'_p &= \{p'_0, p'_1, \dots, p'_n, p'_{n+1}\}, \end{aligned} \quad (7)$$

where softmax_r is a row-based softmax function. S'_p represents the final sentiment expression of the sentence. Similarly, we can get the final expression of the opinion

FIGURE 3: The Cross Interaction Unit.

feature S'_0 . With the crossover operations in CIU, the accuracy of sentiment classification and opinion extraction can be improved simultaneously.

The vector of the target word obtained from S'_p is

$$Tr = [p'_i, p'_{i+1}, \dots, p'_{i+m-1}], \quad (8)$$

where $Tr \in R^{m \times 2d}$ and m represents the length of the target word. A max-pooling operation is performed on the target word vector, and the most important features at each position are selected from different words.

$$V = \max\{Tr, \dim = 0\}, V \in R^{1 \times d}. \quad (9)$$

Finally, V is fed into a fully connected layer for classification. For the opinion extraction task, we use a dense layer plus a softmax operation to generate the final opinion tags.

3.4. Joint Learning. Output from the previous step contains representation of the original text for two purposes: one is polarity labeling, and the other is opinion term labeling. These tasks require different forms of output, so it is necessary to apply gradient descent training that better fit their respective application.

For the sentiment classification branch, V represents the polarity characteristics of the target word. It passes through the fully connected softmax layer to obtain probability values representing emotion polarities.

$$p(a) = \text{softmax}(W_p V_a + b_p). \quad (10)$$

After that, we use the standard cross-entropy loss as the cost function:

$$\text{loss}_p = - \sum_{a \in D} \sum_{c=1}^C P_c^g(a) \cdot \log(P_c(a)), \quad (11)$$

where a represents an aspect term appearing in training data D . C represents the number of categories of sentiment classification. $P_c(a)$ is the probability of predicting s as class c from the softmax layer, and $P_c^g(a)$ indicates whether class c is the correct sentiment category, with value 1 or 0.

For the opinion term extraction branch, all possible outputs of the tag sequence are defined as array Y . Y_{real} is the true label sequence. From the SMLN, the feature value of

each location is converted into a probability value through softmax, and the formula is as follows:

$$p(Y_{\text{real}}|X) = \frac{e^{S(X, Y_{\text{real}})}}{\sum_{y \in Y} e^{S(X, y)}}. \quad (12)$$

The goal of model optimization is to increase the probability of the appearance of the true label and ultimately reduce the value of the loss function. The objective loss function of the opinion word recognition model is defined as follows:

$$\text{loss}_o = -\log(p(Y_{\text{real}}|X)) = -S(X, Y_{\text{real}}) + \log\left(\sum_{y \in Y} e^{S(X, y)}\right). \quad (13)$$

Losses of the main sentiment classification task and the auxiliary opinion word extraction task are aggregated to form the total loss $\mathcal{F}(\theta)$ of the framework.

$$\mathcal{F}(\theta) = \text{Loss}_p + \text{Loss}_o. \quad (14)$$

4. Experiments

4.1. Datasets. Our framework is evaluated on three benchmark datasets from SemEval-2014 [4] and SemEval-2015 [36]. Statistics of the datasets are shown in Table 2. For simplicity, we use 14Lap, 14Res, and 15Res to denote SemEval-2014 Laptops, SemEval-2014 Restaurants, and SemEval-2015 Restaurants, respectively. There are four emotional labels in the entities in datasets, which are "positive," "negative," "natural," and "conflict." "Conflict" means that there are more than two emotions in the same entity. Labels on opinion terms are provided by Wang et al. [35, 37].

4.2. Experiment Settings. The pretrained uncased BERT-base model is used for fine-tuning. The number of Bi-LSTM hidden units is set to 300, and the output dimension of Bi-LSTM is 600. The hyperparameter K is set to 5. In the fine-tuning process, the same parameter settings as the BERT model are kept to ensure comparable results to other baseline models. To avoid overfitting, the dropout

TABLE 2: Statistics of the experiment datasets.

Dataset	Positive	Negative	Neutral	Conflict	Total	Opinion
14Lap-Train	987	866	460	45	2358	2504
14Lap-Test	341	128	169	16	654	674
14Res-Train	2164	805	633	91	3693	3484
14Res-Test	728	196	196	14	1134	1008
15Res-Train	902	252	34	—	1315	1210
15Res-Test	319	319	27	—	685	510

probability value is set as 0.5 for the opinion extract. The model is implemented with the PyTorch library and runs on a single Nvidia 2080 Ti GPU. Other hyperparameters are shown in Table 3.

5. Results

5.1. Baseline Approaches. Following the convention of related work, the average accuracy metric is used to measure the overall performance of sentiment classification models. To show the effectiveness of our model, several mainstream models for aspect-based sentiment analysis are used for comparison, including the following:

TD-LSTM [8] uses two LSTM networks to model the correlation between the target word and its context, and it concatenates the last hidden state of the two parts to predict the sentiment polarity of the target.

ATAE-LSTM [12] applies a typical attention-based LSTM structure to capture the key part of the sentence in response to a given aspect.

MemNet [24] is a deep memory network that applies multiple attention layers to capture the importance of each context word and predicts sentiments based on the sentence representation at the top level.

RAM [25] has a multilayer architecture where each layer consists of an attention-based aggregation of word features and a GRU cell to strengthen the expressive power of MemNet.

IAN [14] contains two LSTMs to encode target words and context information independently and completes the interaction of the two parts of information through the attention mechanism.

TNET [38] proposes a transformation unit for target representation, so that word coding can fully capture the key information of the target. In addition, the authors use a context feature preservation mechanism to better obtain useful information from the context.

TG-SAN [39] includes two core units. One is Structured Context extraction Unit (SCU), which undertakes the task of encoding semantic groups and extracts context fragments related to objects. The second is Context Fusion Unit (CFU), with the purpose of

TABLE 3: Hyperparameters used in the experiment.

Parameter	Value
Dropout rate	0.5
Batch size	32
Learning rate	$2e-5$
Max epoch	20
Max sequence length	128
Optimizer	Adam

learning the contribution of the extracted context to the object.

IMN [22] designs an end-to-end interactive multitask learning network for a variety of fine-grained sentiment analyses. General word vectors and domain-specific word vectors from [40] are concatenated as input. In the model, a special information transfer mechanism is implemented to help the model transfer information between the token level and the document level.

PRET + MULT [21] uses document-level knowledge to improve the performance of aspect-level sentiment analysis. PRET represents the use of documents to train the weight of LSTM, and MULT implies the use of multitask learning methods to complete document-level and aspect-level sentiment analysis tasks.

BERT-FC is the vanilla model built on BERT. The base BERT model is fine-tuned on the target task, and information is extracted at the placeholder [CLS] token for sentiment analysis.

TD-BERT [29] is also based on the BERT fine-tuning model. Instead of the BERT default token [CLS], the vector corresponding to the position of the target term is fed into downstream pipeline. The output is also processed by softmax to get the final emotion category.

AEN-BERT [9] proposes an Attentional Encoder Network (AEN) which does not use the traditional recurrent structure. It employs attention-based encoders for the modeling between a target and its context. They raise the label unreliability issue and introduce label smoothing regularization.

BERT-PT [28] explores a posttraining method on the BERT model with related datasets, with an expectation that the introduction of additional data will improve fine-tuning performance of BERT for sentiment classification.

BERT-pair-QA-M [27] constructs an auxiliary question from the target and uses it together with the original sentence as input. It converts the sentiment classification task into a special QA problem. Since the original paper is applied to task 2 of SemEval-2014, which is not the same as ours, reproduced performance data from [29] is taken as the result.

SK-GCN-BERT [30] proposes a new syntax- and knowledge-based graph convolutional network (SK-GCN) model which leverages the syntactic dependency tree and commonsense knowledge via GCN. In particular, to enhance the representation of the sentence

towards the given aspect, it develops SK-GCN to combine the syntactic dependency tree and commonsense knowledge graph.

SPRN-BERT [31] proposes a semantics perception and refinement network (SPRN) for sentiment analysis with aspects. Local semantic features and global context information are extracted by multichannel convolution and SA, respectively. Then, gated network (DRG) is used to enhance the connection between aspect and context while filtering noise.

5.2. Results and Analysis. Table 4 shows the performance of our model together with previous methods described above. Models in the first part of the table use traditional neural network methods. These methods rely on well-designed attention and LSTM to process static word embeddings. These pretrained word embeddings are generated on large-scale generic corpus or domain-related datasets [40] through the Word2Vec [41] or GloVe [42] method. The second part includes previous multitask learning methods on aspect-level sentiment analysis. The third part shows other models based on the BERT representation. For better performance, they also include customized revisions for the fine-grained sentiment analysis task. Compared with the methods above, our SMLN model shows clear performance improvement over the baseline methods on three datasets from SemEval-2014 and SemEval-2015. This result benefits from the multitask learning setting as well as the information exchange mechanism in CIU.

For the ABSA task, BERT-based models have achieved significant accuracy improvement in comparison to the original static word embeddings. The multilayer Transformer stack structure of BERT has the clear advantage of representing the intrinsic semantics of terms in the context. BERT is a better choice as the shared representation layer in the model, in comparison to static embeddings which lack flexibility in word semantics. When compared to other BERT-based methods, our model still shows significant improvements, about 5% over the baseline BERT-FC model. Our analysis shows that the improvement is mainly from the multitask learning framework. In this work, we introduce opinion terms extraction (OTE) as an auxiliary task. OTE and ABSA tasks are closely related, but there are also clear differences between them. Complementary information from similar but different applications can be used as regularization items between tasks. It effectively improves the generalization performance of the model. In order to further improve the transmission of complementary information between different tasks, we design the CIU module based on an improved self-attention mechanism.

5.3. Ablation Study. In order to study the effects of different components, we gradually add auxiliary tasks and CIU modules starting from the vanilla model. Vanilla model represents the combination of TD-BERT and LSTM network. The experimental results are shown in Table 5, in

TABLE 4: Performance comparison with previous models, with average accuracy as the evaluation metric; the best result for each dataset is boldfaced.

	Method	14Lap	14Res	15Res
Conventional network	TD-LSTM (2016)	68.13	75.63	76.39
	ATAE-LSTM (2016)	68.70	77.20	78.48
	MemNet (2016)	70.33	78.16	77.31
	TNET (2016)	74.65	80.05	78.47
	IAN (2017)	72.10	78.60	78.58
	RAM (2017)	74.49	80.23	79.98
	TG-SAN (2020)	75.27	81.66	—
Multitask learning	PRET + MULT (2018)	71.15	79.11	81.30
	IMN (2019)	75.36	83.89	85.64
BERT	BERT-FC (2018)	76.54	81.28	81.52
	BERT-pair-QA-M (2019)	77.93	85.12	81.89
	AEN-BERT (2019)	78.35	81.46	—
	BERT-PT (2019)	78.07	84.95	—
	TD-BERT (2019)	78.87	85.10	—
	SK-GCN-BERT (2020)	79.00	83.48	83.20
	SPRN-BERT (2021)	79.31	85.03	85.30
Our model	SMLN	80.09	85.67	86.31

TABLE 5: Performance of BERT model with different components, with average accuracy reported over 5 runs.

Model variants	14Lap	14Res
Vanilla model	78.92	85.15
+auxiliary task	79.31	85.33
+CIU	80.09	85.67

which each variant adds a new module based on the previous model. With only the auxiliary task to form a multitask framework, the model achieves a small but noticeable improvement for both the 14Lap and 14Res dataset. It can be considered as the generalization performance improvement brought by the multitask method. At this time, the model cannot benefit from the emotional information provided by the auxiliary task. When the CIU module is added, the additional improvement is about twice that of the previous step. With the CIU, the emotional knowledge has been successfully transferred to the ABSA task.

6. Discussion

For the aspect-based sentiment classification task, we design an SMLN based on multitask learning and attention mechanism. This network can better utilize the rich emotional information in the context and related information among similar tasks at the same time. It tries to solve the problems of sentiment classification and opinion word extraction in an end-to-end manner. In this model, text information is first converted into a vector representation by the BERT preprocessing model. This representation is a common feature in the shared layer that applies to all downstream tasks, and output of the shared layer enters two

independent Bi-LSTM networks to learn the unique features of each task. In particular, this article designs an information interaction unit between two independent representations. This module accomplishes the function of information transfer between the two parts based on the attention mechanism. On publicly available sentiment analysis datasets, its performance is compared to many existing ABSA methods, including some recent work that claims to be state-of-the-art. On all three datasets, the SMLN model achieves competitive results in aspect-based sentiment classification. Its classification accuracy reaches 80.09%, 85.67%, and 86.31% on the 14Lap, 14Res, and 15Res datasets, respectively. To verify the value of its two main components, auxiliary tasks and CIU module, an ablation test is carried out; it shows the step-by-step performance improvement when each individual component is added. The results demonstrate the effectiveness of the SMLN, together with detailed analysis for each component.

In the NLP literature, attention mechanism has been widely used because it can better learn long-range sequence knowledge. However, the latest research shows that the pure self-attention network (SAN) without skip connection and multilayer perceptron (MLP) loses certain expression ability. The loss of feature extraction ability is related to the network depth in double exponential order. Specifically, the researchers prove that the network output converges to a rank-1 matrix at the rate of cubic convergence [43]. Thus, we are currently focusing on the following extensions to the proposed method. First of all, we try to design multilayer attention units in CIU module to obtain stronger feature fusion ability, which is helpful to understand and infer the implied semantics in sentences. Further research aims to explore the changes of internal attention matrix in the process of model reverse updating. Second, we plan to introduce more subtasks into our multitask learning framework, such as entity extraction. The addition of related tasks helps to improve the generalization performance of each task. Finally, we are exploring the effectiveness of our method for other NLP tasks, such as relationship extraction. Overfitting is a common issue for NLP tasks, especially when the model complexity exceeds data size. Multitask learning is an effectively way to improve the generalization ability of a complex model, but understanding the internal correlation between these tasks is more important than blindly stacking more tasks.

Abbreviations

SMLN: Shared Multitask Learning Network
 CIU: Cross Interaction Unit
 LSTM: Long Short-Term Memory
 MTL: Multitask learning
 NLP: Natural language processing
 ASC: Aspect-level sentiment classification
 OTE: Opinion terms extraction
 BERT: Bidirectional encoder representations from transformers.

Data Availability

Restrictions apply to the availability of these data. The datasets SemEval-2014 and SemEval-2015 were taken from <http://alt.qcri.org/semeval2014/task4> and <https://alt.qcri.org/semeval2015>, respectively.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank Chengdu University of Information Technology for providing the GPU computing power. This research was funded by National Key Research and Development Program of China under grant no. 2017YFC0820700 and Key R&D Projects in Sichuan Province under grant no. 2020YFG0168.

References

- [1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, ACM, Seattle, Washington D. C., USA, August 2004.
- [2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [3] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [4] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 27–35, Dublin, Ireland, August 2014.
- [5] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 151–160, Portland, Oregon, USA, June 2011.
- [6] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 49–54, Baltimore, Maryland, June 2014.
- [7] D.-T. Vo and Y. Zhang, "Target-dependent twitter sentiment classification with rich automatic features," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, June 2015.
- [8] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective lstms for target-dependent sentiment classification," 2015, <https://arxiv.org/abs/1512.01100>.
- [9] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Attentional encoder network for targeted sentiment classification," 2019, <https://arxiv.org/abs/1902.09314>.
- [10] J. Wagner, P. Arora, S. Cortes et al., "Dcu: aspect-based polarity classification for semeval task 4," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 223–229, Dublin, Ireland, August 2014.

- [11] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, "Nrc-Canada-2014: detecting aspects and sentiment in customer reviews," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442, Dublin, Ireland, August 2014.
- [12] Y. Wang, M. Huang, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 606–615, <https://aclanthology.org/people/x/xiaoyan-zhu/>, Austin, TX, USA, November 2016.
- [13] C. Li, X. Guo, and Q. Mei, "Deep memory networks for attitude identification," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 671–680, Cambridge, UK, February 2017.
- [14] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," 2017, <https://arxiv.org/abs/1709.00893>.
- [15] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3433–3442, Brussels, Belgium, October–November 2018.
- [16] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, <https://arxiv.org/abs/1706.05098>.
- [17] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [18] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167, Helsinki, Finland, July 2008.
- [19] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, May 2015.
- [20] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," 2016, <https://arxiv.org/abs/1605.05101>.
- [21] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Exploiting document knowledge for aspect-level sentiment classification," 2018, <https://arxiv.org/abs/1806.04346>.
- [22] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An interactive multi-task learning network for end-to-end aspect-based sentiment analysis," 2019, <https://arxiv.org/abs/1906.06906>.
- [23] J. Yu and J. Jiang, "Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, November 2016.
- [24] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," 2016, <https://arxiv.org/abs/1605.08900>.
- [25] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 452–461, Copenhagen, Denmark, September 2017.
- [26] P. Zhu and T. Qian, "Enhanced aspect level sentiment classification with auxiliary memory," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1077–1087, Santa Fe, New Mexico, USA, August 2018.
- [27] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, <https://arxiv.org/abs/1904.02232>.
- [28] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," 2019, <https://arxiv.org/abs/1904.02232>.
- [29] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with bert," *IEEE Access*, vol. 7, Article ID 154299, 2019.
- [30] J. Zhou, J. X. Huang, Q. V. Hu, and L. He, "Sk-gcn: modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification," *Knowledge-Based Systems*, vol. 205, Article ID 106292, 2020.
- [31] W. Song, Z. Wen, Z. Xiao, and S. C. Park, "Semantics perception and refinement network for aspect-based sentiment analysis," *Knowledge-Based Systems*, vol. 214, Article ID 106755, 2021.
- [32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," 2018, <https://arxiv.org/abs/1810.04805>.
- [33] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," pp. 5998–6008, 2017, <https://arxiv.org/abs/1706.03762>.
- [34] Y. Wu, M. Schuster, Z. Chen et al., "Google's neural machine translation system: bridging the gap between human and machine translation," 2016, <https://arxiv.org/abs/1609.08144>.
- [35] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, Hilton San Francisco, CA, USA, February 2017.
- [36] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: aspect based sentiment analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 486–495, Denver, Colorado, June 2015.
- [37] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Recursive neural conditional random fields for aspect-based sentiment analysis," 2016, <https://arxiv.org/abs/1603.06679>.
- [38] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," 2018, <https://arxiv.org/abs/1805.01086>.
- [39] J. Zhang, C. Chen, P. Liu, C. He, and C. W.-K. Leung, "Target-guided structured attention network for target-dependent sentiment analysis," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 172–182, 2020, <https://aclanthology.org/2020.tacl-1.12>.
- [40] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double embeddings and cnn-based sequence labeling for aspect extraction," 2018, <https://arxiv.org/abs/1805.04601>.
- [41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, <https://arxiv.org/abs/1301.3781>.
- [42] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.
- [43] Y. Dong, J. B. Cordonnier, and A. Loukas, "Attention is not all you need: pure attention loses rank doubly exponentially with depth," 2021, <https://arxiv.org/abs/2103.03404>.