Sentiment Analysis with Weighted Graph Convolutional Networks

Fanyu Meng, Junlan Feng, Danping Yin, Si Chen, Min Hu
China Mobile Research Institute, Beijing, China
{mengfanyu,fengjunlan,yindanping,chensiyjy,humin}@chinamobile.com

Abstract

Syntactic information is essential for both sentiment analysis(SA) and aspect-based sentiment analysis(ABSA). Previous work has already achieved great progress utilizing Graph Convolutional Network(GCN) over dependency tree of a sentence. However, these models do not fully exploit the syntactic information obtained from dependency parsing such as the diversified types of dependency relations. The message passing process of GCN should be distinguished based on these syntactic information. To tackle this problem, we design a novel weighted graph convolutional network(WGCN) which can exploit rich syntactic information based on the feature combination. Furthermore, we utilize BERT instead of Bi-LSTM to generate contextualized representations as inputs for GCN and present an alignment method to keep word-level dependencies consistent with wordpiece unit of BERT. With our proposal, we are able to improve the state-of-the-art on four ABSA tasks out of six and two SA tasks out of three.

1 Introduction

Sentiment analysis(SA), also known as opinion mining, is the task of determining the polarity of a piece of text. Commonly the classification is whether the text is expressing a negative or positive attitude towards a topic or a product. Fine-grained sentiment analysis involves more than two sentiment classes (very negative, negative, neutral, positive and very positive). Aspect-based sentiment analysis(ABSA) is one step further by assigning sentiment polarities to specific aspects of an involved entity or a topic. For example, comment on a restaurant saying *"The restaurant was expensive, but the menu was great"* has *positive* and *negative* attitudes for two aspects *food* and *price*.

Much progress has been made recently to advance the state-of-the-art on shared SA and ABSA tasks. Contributions mainly come from two research directions.

One is to take advantage of the pre-trained language models such as ELMo(Peters et al., 2018), BERT(Devlin et al., 2018) and XLNet(Yang et al., 2019a), which are typically employed to extract contextual features of a piece of text for the final classifer. These models effectively alleviate the heavy effort of feature engineering of earlier work on SA and ABSA. Further inventions have been proposed to better fine-tune these models. For instance, a recent work (Sun et al., 2019a) converts ABSA to a sentence pair classification task, where an auxiliary sentence is generated. It then fine-tunes the pre-trained model from BERT for this new task. Promising experimental results are observed.

Second line of research is to exploit the syntactic structures of subjective sentences with a belief that interactions between words need to be considered in sentiment analysis, which however is not sufficiently captured by even the latest attention-based models. (Zhang et al., 2019) quotes a concrete example *"Its size is ideal and the weight is acceptable"*, where *acceptable* is often incorrectly identified by attention models as the most attentive word to *size*. Previous works in (Socher et al., 2011; Dong et al., 2015; Qian et al., 2015; Socher et al., 2013) propose a recursive tree-structured model to compose sentence representation from its constituent phrases. (Kim et al., 2018) presents a novel RvNN architecture to dynamically integrate comprehensive syntactic information derived from the sentence parsing structure and linguistic tags on word level. Models using a Graph Convolutional Network(GCN) over the dependency tree of a sentence have shown evident effectiveness in ABSA tasks. The argument is that GCN captures long-

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 586–595
November 16 - 20, 2020.

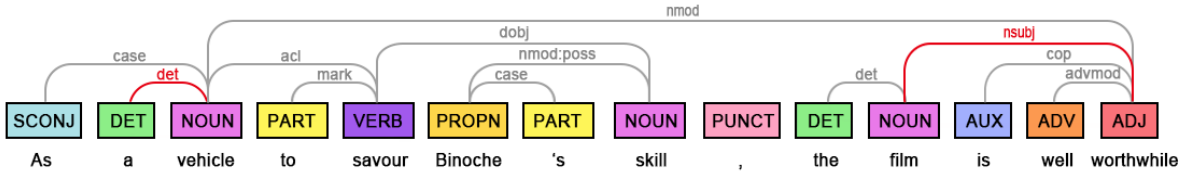

Figure 1: An example of a dependency tree noted with type of dependency relation and POS tag for each word.

range syntactic relations that are obscure from the surface(Sun et al., 2019b; Zhang et al., 2019; Zhao et al., 2019).

Though these efforts have substantially pushed up the state-of-the-art accuracy of SA and ABSA, some challenges remain for sentiment classification. For example, the aforementioned GCN-based models are designed to encode the dependency tree of a sentence, where the adjacency matrix is binary with 1 representing if there is a dependency relationship between two corresponding words and 0 for others. However, types of dependency relations are diversified and the corresponding words of each relation may have different part-of-speech(POS) tags. These syntactic information should also influence the message passing process of GCN. As it is shown in Figure 1, the relationship(*"det(vehicle-3, a-2)"*) has less influence on polarity than the relationship(*"nsubj(worthwhile-14, film-11)"*) in the sentence *"As a vehicle to savour Binoche 's skill , the film is well worthwhile"*. Besides, as (Sethi and Bhattacharyya, 2017) points, pitfalls of SA and ABSA like *Sentiment Shifters* (such as *Negations, Double Negations and But clauses*) have not been well handled by current models.

In this paper, we are motivated to encode more syntactic features and leverage both the pre-trained models and the syntactic parsing in a compositional way. We believe these are complementary to tackle the long-standing challenges for SA and ABSA. More specifically, we propose a Weighted Graph Convolutional Network(WGCN) to work with BERT. WGCN improves on top of GCN to model rich syntactic information. The adjacency matrix in WGCN represents not only the binary representations of dependency relations, but also the types of dependency relations as well as the part-of-speech(POS) categories of the involved words. We argue that the POS tag of each word is the category assigned in accordance with its syntactic function , hence has influence on the overall sentiment of the sentence as well as sentiments of aspects. All weights and embeddings in WGCN are trainable. Details of this model will be provided later in this paper. WGCN reply on BERT to extract contextualized representations as inputs for the WGCN layers. One challenge is the inconsistency between the WordPiece unit of BERT, and the word-pairs considered in the dependency tree. We propose an alignment method to bridge this chasm.

Our contributions are summarized as follows:

- We propose a novel weighted GCN(WGCN) architecture over dependency tree which can exploit rich syntactic features by assigning trainable weights for adjacent matrix.
- We propose a framework to compositionally exploit the pre-trained language models(BERT) and WGCN for SA and ABSA. We refer to the whole architecture as BERT-WGCN.
- With our proposal, we are able to improve the state-of-the-art on four ABSA tasks out of six and two SA tasks out of three.

The rest of the paper is organized as follows. Section 2 gives a brief review of BERT and GCN. Section 3 elaborates on our proposed overall model architecture that integrates WGCN and BERT, as well as how the model is trained respectively for SA and ABSA tasks. Section 4 reports our experiments and analysis.

2 Review of GCN and BERT

Graph convolutional network(Kipf and Welling, 2016) is an adaptation of the convolutional neural network(LeCun et al., 1998) for encoding unstructured data. Given a graph with k nodes, we can obtain an adjacency matrix A where A_{ij} is obtained based on the connection between node i and node j. In an L-layer GCN where H^{l-1} represents the output feature matrix at $(l-1)$-th layer and H^l represents the output feature matrix at the l-th layer, a graph convolutional operation can be written as:

$$H^l = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{l-1}W^l) \quad (1)$$

$\tilde{A} = A + I_k$ is the adjacency matrix with self-loops, where I_k is the identity matrix. $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix. W^l is a linear transformation weight,and σ is a nonlinear function(e.g., ReLU). In each graph convolution, each node collects and processes information from its neighboring nodes.

BERT(Devlin et al., 2018) is one of the key innovations in the recent progress of contextualized representation learning inspired by Tranformer(Vaswani et al., 2017). Given a sentence $s = \{w_1, ..., w_n\}$, its tokenized sequential representation is $\{t_1, t_2, ..., t_k\}$. Transformer creates three vectors (query, key and value) for each sequence position, and then applies the attention mechanism for each position x_i, using the query vector for x_i as well as key and value vectors for all other positions. This computation can be presented as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Instead of performing a single attention function, (Vaswani et al., 2017) found it is beneficial to have multiple attention heads. Bert built on Transformers contains a number of layers(Transformer blocks) L . Each layer is identical with a fixed number of hidden units H and a fixed number of multi-threading self-attention heads A. Particularly we use the $BERT_{LARGE}$ model with $L = 24$, $H = 1024$ and $A = 16$ as hyper-parameters.

3 Approach

Figure 2 gives an overview of the whole architecture. Our model consists of 3 main components. First, the input sequence of text is parsed into word-based syntactic features as inputs for WGCN. At the same time, the text is also directly fed into BERT for wordpiece contextualized representations. One challenge here is the inconsistency between the wordpiece unit of BERT and word-based syntactic features for WGCN. The second part is the reform of GCN to exploit rich syntactic features. The third component is the sentiment classifer for SA and ABSA. The components will be introduced separately in the rest of the section.

3.1 Token Alignment towards BERT

Traditional GCN-based approaches over dependency tree use Bi-LSTM to get contextualized representations as initialized inputs for GCN (Zhang et al., 2018a,b). Recently pre-trained models have proved the effectiveness of capturing contextual information. Thus we first feed input sentences into BERT model to generate contextualized representations. This BERT contextualization layer is trained jointly with the rest of the network. One challenge to have BERT work with WGCN as shown in Figure 2 is the tokenization inconsistency between them. Bert tokenizes input into wordpiece units, instead of keeping word boundaries as they are.

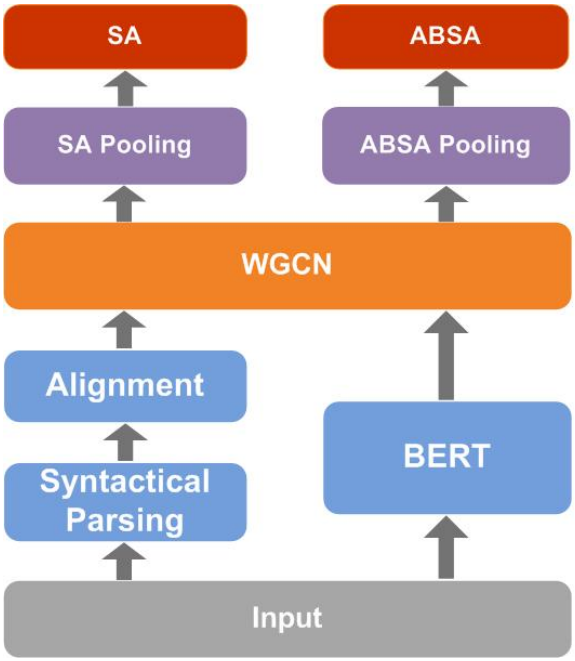

Figure 2: Overview of proposed architecture for SA and ABSA.

To resolve this issue, we propose an alignment procedure to map the word-level sequence from the parser to the wordpiece sequence in BERT. Dependency relations and POS tags are then accordingly aligned. The procedure is as follows:

Given a piece of text s, the parser tokenizes it into a n word-level sequence: $s = \{w_1, ..., w_i, ..., w_n\}$ and BERT processes it into a k wordpiece sequence: $s_t = \{t_1, ..., t_m, ..., t_n, ..., t_k\}$. For any given w_i in s, there is a corresponding subsequence of wordpiece tokens $seg_i = \{t_m, ..., t_n\}$, where $1 \leq m \leq n \leq k$. We apply two alignment rules to map parsing results into a new form:

- Rule 1: If w_i is labeled by a POS Tagger as p_i, then all tokens in seg_i are assigned the same tag p_i.
- Rule 2: If there is a dependency relation r_{ij} between w_i and w_j, then we assign the same

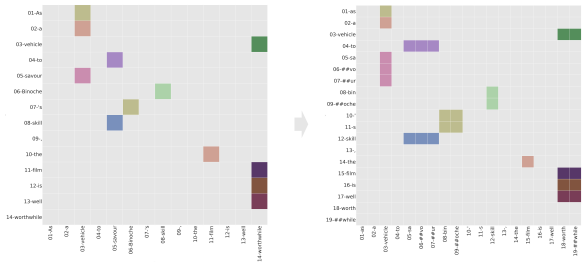

Figure 3: Alignment from word-based adjacency matrix to wordpiece adjacency matrix.

dependency relation r_{ij} between any token in seg_i and any token in seg_j.

With this alignment, given an adjacency matrix A where $A_{ij} = 1$ if node i is connected with node j, we can obtain a new adjacency matrix A^{align} where $A^{align}_{xy} = 1$ for any token x in seg_i and any token y in seg_j. We plot one example in Figure 3. For a better illustration, we show what the adjacency matrix looks like before and after the alignment. The left side shows the dependency matrix between the 14 words for the sentence *"As a vehicle to savour Binoche 's skill , the film is well worthwhile"*. Each color represents a particular relation type. The right side shows the dependency matrix on wordpiece sentence *"as a vehicle to sa ##vo ##ur bin ##oche ' s skill , the film is well worth ##while"* after we run alignment with the above procedure. It's worth noting that we present directed graphs in Figure 3 for clarity. As GCNs generally do not consider directions, we use un-directional graph in our model.

3.2 Weighted Graph Convolutional Networks over Syntactic Information

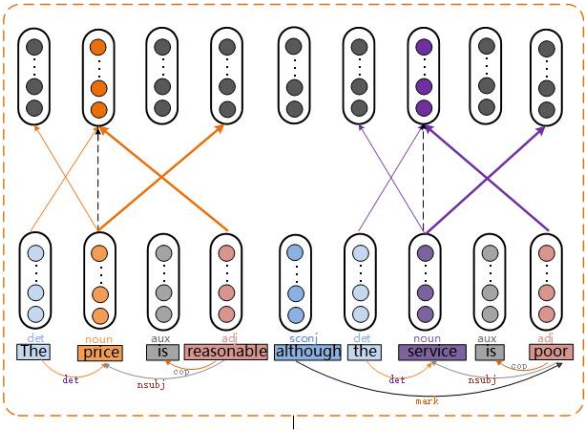

Figure 4: An overview of WGCN. We only show the detailed graph convolution computation for the aspect words *price* and *service* for clarity.

We aim to extend GCN to model rich syntactic information. To this end, we propose WGCN, which is depicted in Figure 4. Following the same strategy in (Sun et al., 2019b; Zhang et al., 2018b,a), WGCN also considers the adjacency matrix obtained from dependency tree as input. Different from their approaches, WGCN assigns trainable weights to the adjacency matrix. Each weight is compositionally determined by syntactic information including the type of dependency relation and the corresponding POS tags of the word-pairs.

Our hypothesis is that the type of dependency relation and POS tags of the word-pairs should have combined impacts on the process of aggregating information from neighbours in GCN. We follow the procedure proposed by (Guo et al., 2017) for Factorization Machines(FM) to cast pairwise feature interactions as inner product of the latent vectors, which has shown very promising results on many tasks. Let W_{type} be a matrix of $R^{d \times N_{type}}$, where

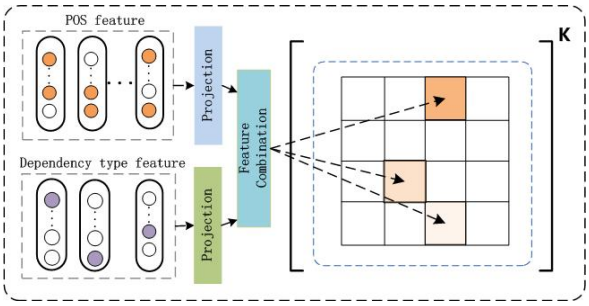

Figure 5: Computation of the adjacency matrix in WGCN.

d is the dimension of the embedding space which is fixed hyper-parameter, and N_{type} is the number of types of dependency relations. Let W_{pos} be a matrix of $R^{d \times N_{pos}}$ and N_{pos} is the number of combinations of POS tags of all word-pairs appeared in dependency relations. The feature combination weight over the dependency relation from node x to node y in adjacency matrix can be presented as:

$$\alpha_{xy} = f(r_{xy})g((p_x, p_y)) \quad (3)$$

r_{xy} is the type of dependency relation from node x to node y, p_x and p_y are the POS tags in the sentence for node x and node y. The function $f()$ maps the one-hot type vector into the corresponding column of W_{type}. The function $g()$ maps the two-hot POS vector into the corresponding column of W_{pos}.

Let $\widehat{A}$ be the final adjacent matrix for WGCN,

Tasks	ABSA tasks					SA tasks		
Datasets	**SEM14(LAP)**	**SEM14(Rest)**	**Rest15**	**Rest16**	**Twit**	**SST2**	**SST5**	**SE13**
Train	2282	3608	1204	1748	6051	6920	8544	6021
Dev	-	-	-	-	-	872	1101	890
Test	632	1119	542	616	677	1821	2210	2376
ofClass	3	3	3	3	3	2	5	3

Table 1: Dataset statistics of aspect-based sentiment analysis(ABSA) and sentiment analysis(SA)

then each value of $\widehat{A}$ can be computed as:

$$\widehat{A}_{xy} = \alpha_{xy} A^{align}_{xy} \tag{4}$$

where α_{xy} is computed from Equation (3) and A^{align}_{xy} is obtained by alignment rules. The process of obtaining $\widehat{A}$ is shown in Figure 5.

To adapt with trainable adjacency matrix, we reform the custom GCN. Inspired by (Zhang et al., 2018c), we use K-th power of adjacency matrix to aggregate information from K-hop neighbours. Since nodes never connect to themselves in a dependency relation, following the idea of self-looping(Kipf and Welling, 2016), we add a matrix I^{align} which is transformed by an identity matrix with proposed alignment method to carry over information.

Let H^b be the final output of BERT layer, WGCN can be presented as :

$$H^{GCN} = \sigma(C_i((\widehat{A})^K + I^{align})H^b W) \tag{5}$$

$C_i(\cdot)$ is a clip function for the matrix. W is the parameter matrix for WGCN and σ is the nonlinear $ReLU$ function.

3.3 Model Training for SA and ABSA

Sentiment analysis considers the polarity of the whole sequence. In our framework, we use an average pooling to aggregate the whole sequence. Let $H^{GCN} = \{h^{GCN}_1, ..., h^{GCN}_k\}$ be the final output of WGCN, $Avg(\cdot)$ be the average pooling function. The pooling process can be presented as:

$$h_{SA} = Avg\left(\{h^{GCN}_1, ..., h^{GCN}_k\}\right) \tag{6}$$

Aspect-Based Sentiment Analysis considers the polarity of several aspect words given in a current sentence. The BERT model and WGCN allow embeddings for aspect tokens to respectively aggregate contextual tokens and neighbouring tokens in a dependency tree, providing supervisory signals for the aspect-based classification task. Different from sentiment analysis, we use an average pooling to aggregate only the aspect words. Given a sentence pair (a, s), where a is a sub-sequence of s as aspect tokens. The final outputs of WGCN are $\{h^{GCN}_1, ..., h^{GCN}_{a_s}, ..., h^{GCN}_{a_e}, ..., h^{GCN}_k\}$ where a_s and a_e are indexes an aspect starts from and ends at. The pooling process can be presented as:

$$h_{ABSA} = Avg\left(\{h^{GCN}_{a_s}, ..., h^{GCN}_{a_e}\}\right) \tag{7}$$

h^{SA} or h^{ABSA} is then fed into a linear layer followed by a softmax operation to obtain a probability distribution over polarities. For training we use Adam algrithm(Kingma and Ba) with the cross-entropy loss and L2-regularization.

4 Experiments

4.1 Datasets and Experimental Settings

We conduct our experiments on five aspect-based sentiment analysis datasets and three sentiment analysis datasets:

- TWITTER dataset for ABSA, was originally built by (Li et al., 2014) containing thousands of twitter posts. Annotations are sentiment labels(negative, neutral and positive) for given keywords or topics such as *"taylor swift","xbox"*.

- LAP14, REST14, REST15, REST16 datasets for ABSA are respectively from SemEval 2014 task 4(pontiki et al., 2014), SemEval 2015 task 12(Pontiki et al., 2015) and SemEval 2016 task 5(Pontiki et al., 2016), consisting of data from two categories, i.e. laptop and restaurant.

- SST(SST2, SST5) is a dataset for sentiment analysis on movie reviews, which are anno-

Datasets	SEM14(LAP)		SEM14(REST)		SEM14(AVG)		REST15		REST16		Twitter	
Model	ACC.	F1	ACC.	F1	ACC.	F1	ACC.	F1	ACC.	F1	ACC.	F1
ASGCN-DG	75.55	71.05	80.77	72.02	78.16	71.54	79.89	61.89	88.99	67.48	72.15	70.40
CDT	77.19	72.99	82.30	74.02	75.09	73.51	-	-	85.58	69.93	74.66	73.66
BERT-PT	78.07	75.08	84.95	76.96	81.51	76.02	-	-	-	-	-	-
SDGCN	**81.35**	**78.34**	83.57	76.47	82.46	77.41	-	-	-	-	-	-
TNET	76.54	71.75	80.69	71.27	78.62	71.51	-	-	-	-	74.97	73.60
BERT-ADA Rest	79.14	74.93	**87.89**	**81.05**	83.52	77.99	-	-	-	-	-	-
BERT-ADA Lapt	80.23	75.77	86.22	79.79	83.22	77.78	-	-	-	-	-	-
BERT(comp)	78.26	73.35	83.50	73.33	80.88	73.34	81.20	60.11	88.25	72.06	71.09	70.81
BERT-GCN(comp)	80.03	75.79	85.32	78.05	82.68	76.92	85.30	66.01	90.91	**75.31**	73.98	71.62
BERT-WGCN	80.96	76.95	86.71	79.12	**83.84**	**78.03**	**85.39**	**66.26**	**91.35**	75.19	**75.89**	**73.82**

Table 2: Model comparison results for ABSA tasks. The state-of-the-art performance with each dataset is in bold. We list average scores on SemEval2014 on accuracy and F1 to evaluate generalization of different models.

tated with two or five labels(Socher et al., 2013).

- SemEval13 is a dataset of Semeval-2013 task 2 (Nakov et al., 2013) for sentiment analysis, consisting of tweets with three sentiment labels(positive, negative and neutral).

The statistics of datasets are reported in Table 1. The datasets are parsed by Stanford parser(v3.6.0) for dependency relation and spacy(2.2.3) for POS tag. We use a learning rate of 0.0001 and a batch size of 32. We set the number of WGCN layers to 3 and the dimension of syntactic feature to 20, which are the best-performing settings in pilot studies. Experiments and benchmarks are run with a single GPU server with 4 V100 GPU cards and 8Gb of RAM. All models are implemented with Tensorflow 1.13 using Cuda 10.1.

The experimental results are obtained by averaging 5 runs with random initialization, where Accuracy and Macro-Averaged F1 are adopted as the evaluation metrics.

4.2 Model for Comparison

To evaluate the effectiveness of our model(BERT-WGCN), we compare our performance with a range of baselines and state-of-the-art models, as listed below:

- CDT(Sun et al., 2019b) is a dependency graph convolutional network integrated with a Bi-LSTM model.

- ASGCN-DG(Zhang et al., 2019) utilizes aspect-aware attention on a dependency graph convolutional network.

- BERT-PT(Xu et al., 2019) transforms ABSA tasks to machine reading comprehension (MRC) and uses a post-training approach on BERT for ABSA tasks..

- SDGCN(Zhao et al., 2019) employs GCN to model the sentiment dependencies between different aspects in one sentence.

- TNET(Li et al., 2018) employs CNN as the feature extractor and uses target specific transformation component to better integrate target information into the word representation.

- BERT-ADA (Rietzler et al., 2019) uses self-supervised domain-specific BERT language model for tuning, followed by supervised task-specific fine-tuning.

- BCN+CoVe(Brahma, 2018) utilizes prefix and suffix of each token in a sentence, which is encoded in both forward and reverse directions to capture long range dependencies for classification tasks.

- SSAN (Ambartsoumian and Popowich, 2018) is a simple multiple self-attention network with positional-encoding for sentiment analysis.

- XLNet (Yang et al., 2019b) is an unsupervised language representation learning method based on a novel generalized permutation language modeling objective and employs Transformer-XL as the backbone model.

- BERT-GCN(comp) (Rietzler et al., 2019) is a model for comparison which connects GCN after BERT-LARGE model with our way of

Model	Aspect	Weight visualization	Prediction	Label
BERT	food	great food but the service was dreadful !	pos	pos
	service	great food but the service was dreadful !	pos	neg
	staff	Our waiter was friendly and it is a shame that he didn't have a supportive staff to work with .	pos	neg
BERT-GCN	food	great food but the service was dreadful !	pos	pos
	service	great food but the service was dreadful !	neg	neg
	staff	Our waiter was friendly and it is a shame that he didn't have a supportive staff to work with .	pos	neg
BERT-WGCN	food	great food but the service was dreadful !	pos	pos
	service	great food but the service was dreadful !	neg	neg
	staff	Our waiter was friendly and it is a shame that he didn't have a supportive staff to work with .	neg	neg

Table 3: The weight visualization on aspect sentiment analysis tasks for BERT(comp), BERT-GCN(comp) and BERT-WGCN with corresponding labels.

alignment and the size of parameters is in the same order of magnitude with our BERT-WGCN.

- BERT(comp) (Rietzler et al., 2019) is a model for comparison which is based on BERT-LARGE and the size of parameters is in the same order of magnitude with our BERT-WGCN.

4.3 Experimental Results

Table 2 shows the performance of our model on accuracy and macro-F1 on ABSA tasks. Our BERT-WGCN outperforms most of the compared models on REST15, REST16 and TWITTER datasets, and achieves competitive results on SEM14(LAP) and SEM14(REST) datasets compared with SDGCN and BERT-ADA. Notably, our model achieves highest average accuary and F1 on SEM14(LAP) and SEM14(REST) dataset combined. The results demonstrate the effectiveness of BERT-WGCN.

For ablation study, we compare our GCN-based models with BERT(comp) with same number of parameters. BERT-GCN(comp) and BERT-WGCN can consistently show improvements. It implies that the syntactic structure is helpful for ABSA tasks. Compared to BERT-GCN(comp), BERT-WGCN is able to gain better performance for almost all ABSA datasets. It proves that WGCN factorizing dependency relations and POS tags is better at utlizing syntactic information than the traditional GCN architecture. For the slight F1 degradation on the REST16 dataset, the reason might be that the size of REST16 datasets is relatively small. Another important observation is that all architectures that achieve the state-of-the-art results utilize pre-trained model. SDGCN-BERT initializes the word embeddings with pre-trained BERT token embeddings and uses self-attention network for training. BERT-ADA uses domain-specific dataset for model pre-training. Thus we believe that the contextualized information is essential for ABSA tasks.

Model	SST-2	SST-5	SE13
BCN+CoVe	-	56.2	-
XLNet	**96.8**	-	-
SSAN	84.2	48.1	72.2
BERT(comp)	94.3	54.8	74.9
BERT-GCN(comp)	94.3	55.0	75.2
BERT-WGCN	94.9	**56.5**	**77.3**

Table 4: Model comparison results for SA tasks. The state-of-the-art performance with each dataset is in bold.

For SA task, as it is shown in Table 4, the message is complex. For SST-2 dataset, our proposed model has no improvement. For SST-5 and SemEval2013, as far as we know, we achieve the new state-of-the-art performance. For ablation study, BERT-GCN(comp) and BERT(comp) get almost the same performance. We believe the main reason is that the importance of sentence structure in SA tasks is not as important as that

in ABSA tasks. BERT-WGCN gets better performance mainly based on the additional feature combinations.

4.4 Case Analysis

In this section we compare BERT-WGCN with two baseline models on case examples. To this end we present visualizations showing the weights extracting from the whole sentence by aspect tokens on ABSA tasks. To show the effectiveness of our model, we expect the aspect tokens can attend to tokens which can influence the sentiment correctly.

As it is shown in Table 3, the first example *"great food but the service was dreadful!"* has two aspects within one sentence. The BERT model is able to detect the polarity for the first aspect *"food"* but fails to infer sentiment polarities for aspect *"service"*. Our hypothesis is that the distance between aspect token and adjunct token is important for attention-based model. The GCN-based model can address this connection correctly because they are directly related on the dependency tree. The second example *"Our waiter was friendly and it is a shame that he didn't have a supportive staff to work with ."* shows the importance of feature combination of dependency relation and POS tags on Negatives. These results suggest the advantage of our model against attention-based model and traditional GCN-based models.

4.5 Investigation on the Combination of Syntactic Features

High Importance		Low Importance	
Relation	**POS-pairs**	**Relation**	**POS-pairs**
amod	(NOUN, ADJ)	*cc*	(CCONJ, CCONJ)
nsubj	(NOUN, ADJ)	*nsubj*	(DET, AUX)
advmod	(ADV,VERB)	*prt*	(ADP,VERB)
advmod	(ADV, ADJ)	*det*	(SCONJ,SCONJ)
cc	(VERB,CCONJ)	*pobj*	(ADP,NOUN)
csubj	(AUX,VERB)	*amod*	(ADJ,ADJ)
advcl	(VERB,AUX)	*amod*	(ADV,ADV)
prep	(SCONJ,VERB)	*det*	(DET,DET)

Table 5: Importance of different Feature Combination on SST-5 task.

To evaluate the influence of feature combination of dependency relation and POS tags of word-pairs, we present several combinations of different importance in WGCN based on the weight score in adjacency matrix. As we use clip function in training, the combinations in column is not ordered. As it is shown in Table 5, relations of adjectival modifier(*"amod"*) or nominal subject(*"nsubj"*) from *"ADJ"* to from *"NOUN"* outweighs relation of determiner(*"det"*) in SA tasks. Another observation is that dependency type and POS tags jointly determine the importance. Same dependency relation may have different importance according to the corresponding POS tags.

4.6 Impact of GCN Layers

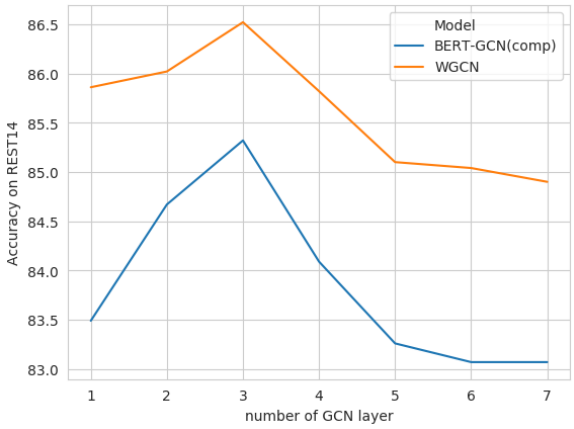

Figure 6: Accuracy curves for BERT-GCN(comp) and BERT-WGCN on the Rest14 dataset.

The number of GCN layers K indicates that we can obtain K-hop neighborhood matrix. We vary the number of layers in $\{1, 2, 3, 4, 5, 6, 7\}$ and check the corresponding accuracy of BERT-GCN(comp) and BERT-WGCN on the REST14 dataset. The results are shown in Figure 6. In particular, the performances of both models increase in first 3 layers. The performance becomes unstable after that. With the increase of number of layers, the model becomes more difficult to train and the performance begins to fall.

5 Conclusion

In this paper we propose a novel weighted graph convolutional networks(WGCN) to work with BERT on sentiment analysis and aspect-based sentiment analysis tasks. WGCN improves on top of GCN to model rich syntactic information including dependency relations as well as POS tags. BERT is used as a powerful tool to extract contextual representations, which are then used as inputs to WGCN to derive the final vectors for classification. We propose an alignment approach to solve the token inconsistency issue between WGCN and BERT. Our experimental results with visualizations show the success of our proposal comparing to the baseline and previous approaches in the literature.

References

Artaches Ambartsoumian and Fred Popowich. 2018. Self-attention: A better building block for sentiment analysis neural network classifiers. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 130–139, Brussels, Belgium. Association for Computational Linguistics.

Siddhartha Brahma. 2018. Improved sentence modeling using suffix bidirectional lstm. *arXiv preprint arXiv:1805.07340*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Furu Wei, Ke Xu, Shixia Liu, and Ming Zhou. 2015. Adaptive multi-compositionality for recursive neural network models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):422–431.

Huifeng Guo, Ruiming TANG, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: A factorization-machine based neural network for ctr prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1725–1731.

Taeuk Kim, Jihun Choi, Daniel Edmiston, Sanghwan Bae, and Sang-goo Lee. 2018. Dynamic compositionality in recursive neural networks with structure-aware tag representations. *CoRR*, abs/1809.02286.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Dong Li, Furu Wei, Chuanqi Tan, Duyu Tang, and Xu Ke. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, De Orphée Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, V. Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, María Jiménez Salud Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. *SemEval@NAACL-HLT*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

maria pontiki, dimitris galanis, john pavlopoulos, harris papageorgiou, ion androutsopoulos, and suresh manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *SemEval@COLING*.

Qiao Qian, Bo Tian, Minlie Huang, Yang Liu, Xuan Zhu, and Xiaoyan Zhu. 2015. Learning tag embeddings and tag-specific composition functions in recursive neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1374.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. *CoRR*, abs/1908.11860.

Abhishek Sethi and Pushpak Bhattacharyya. 2017. Aspect based sentiment analysis-a survey.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in neural information processing systems*, pages 801–809.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on*

empirical methods in natural language processing, pages 1631–1642.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *CoRR*, abs/1903.09588.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019b. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5679–5688, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019a. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *CoRR*, abs/1909.03477.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018a. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018b. Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*.

Zhengchao Zhang, Meng Li, Xi Lin, Yinhai Wang, and Fang He. 2018c. Multistep speed prediction on traffic networks: A graph convolutional sequence-to-sequence learning approach with attention mechanism. *CoRR*, abs/1810.10237.

Pinlong Zhao, Linlin Hou, and Ou Wu. 2019. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *CoRR*, abs/1906.04501.