The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Sentiment Lexicon Enhanced Attention-Based LSTM for Sentiment Classification

Zeyang Lei, Yujiu Yang *
Graduate School at Shenzhen,
Tsinghua University, Shenzhen, P. R. China
leizy16@mails.tsinghua.edu.cn and yangyj@gmail.com

Min Yang
Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences, Shenzhen, P.R.China
min.yang1129@gmail.com

Abstract

Deep neural networks have gained great success recently for sentiment classification. However, these approaches do not fully exploit the linguistic knowledge. In this paper, we propose a novel sentiment lexicon enhanced attention-based LSTM (SLEA-LSTM) model to improve the performance of sentence-level sentiment classification. Our method successfully integrates sentiment lexicon into deep neural networks via single-head or multi-head attention mechanisms. We conduct extensive experiments on MR and SST datasets. The experimental results show that our model achieved comparable or better performance than the state-of-the-art methods.

Introduction

In general, sentiment classification is the problem of classifying the sentiment polarity of a text as positive, negative or neutral. Most existing work establishes sentiment classifiers using supervised machine learning approaches, such as support vector machine (SVM), convolutional neural network (CNN) (Lei, Barzilay, and Jaakkola 2015), long short-term memory (LSTM) (Tai, Socher, and Manning 2015).

Despite the effectiveness of previous studies, sentiment classification still remains a challenge in real-world. A comprehensive and high quality sentiment lexicon plays a crucial role in traditional sentiment classification approaches. Despite its usefulness, to date, the sentiment lexicon has received little attention in recent neural network models (e.g., CNN and LSTM) that achieve the state-of-the-art in generic sentiment classification.

To address the aforementioned limitation, we propose a novel sentiment lexicon enhanced attention-based LSTM (SLEA-LSTM) model. More specifically, our model consists of two independent LSTM networks with additional attention layers on the top. The two LSTM networks are used to learn the hidden representations of context and sentiment words of input, respectively. In addition, we explore two types of attention mechanisms: single-head and multi-head attention methods (Vaswani et al. 2017). The single-head attention computes attention weights of different word locations according to their intent importance associated with sentiment lexicon, which usually only focuses on specific parts of the input. However, the multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. The experiment results show that our model is efficient and achieves state-of-the-art results.

*Corresponding author. This work was supported in part by the Research Fund for the development of strategic emerging industries by ShenZhen city (No.JCYJ20160301151844537 and No. JCYJ20170412170118573).

Our Model

We first use two independent LSTM layers to convert the sequence of input word embeddings into two sequences of hidden states $H^c = [h_1^c, ..., h_m^c]$ and $H^s = [h_1^s, ..., h_n^s]$ for context words and sentiment words, where m represents the length of the sequence and n represent the number of sentiment words. Then, the representation of sentiment words can be computed by applying a mean-pooling operation: $z_s = \sum_{i=1}^{n} h_i^s / n$.

Single-head Attention model

Most existing attention approaches compute an attention weight vector for each input, which we call single-head attention mechanism. With single-head attention, the final lexicon enhanced sentence representation is a weighted sum of context hidden states:

$$\mathbf{o} = \sum_{i=1}^{m} a_i h_i^c, \text{ with } a_i = \frac{\exp(\sigma([h_i^c; z_s]))}{\sum_{i=1}^{m} \exp(\sigma([h_i^c; z_s]))} \quad (1)$$

where a_i indicates the importance of the i-th word in the context, and σ is a score function that calculates the importance of h_i^c in the context. The score function σ is defined as:

$$\sigma([h_i^c, z_s]) = u_{s_1}^T \tanh(W_{s_1}[h_i^c; z_s]) \quad (2)$$

where $[h_i^c; z_s]$ denotes the concatenation of h_i^c and z_s, W_{s_1} and u_{s_1} are parameters to be learned.

Multi-head Attention model

Instead of using attention weight vectors for single-head attention, multi-head attention produces attention weight matrix. With multi-head attention, the final sentence representation takes the following form:

$$\mathbf{o} = flatten(AH^c), \text{with} A = \frac{\exp(\rho([h_i^c; z_s]))}{\sum_{i=1}^{m} \exp(\rho([h_i^c; z_s]))} \quad (3)$$

where A denotes attention matrix which indicates the importance of the i-th word in multiple hops of attention, and $flatten$ is an operation that will flatten matrix into vector form. ρ is a score function that calculates the importance of h_i^c in multiple hops of attention:

$$\rho([h_i^c, z_s]) = {U_{s_2}}^T \tanh(W_{s1}[h_i^c; z_s]) \quad (4)$$

where U_{s_2} and W_{s_2} are projection parameters to be learned.

Finally, we feed the output vector $\mathbf{o}$ to a softmax layer to predict the sentiment distribution. The training objective is to minimize the cross-entropy error of the predicted and true class distributions.

Experiments

Datasets and sentiment lexicon

Movie Review (MR) This dataset (Pang and Lee 2005) consists of 5,331 positive and 5,331 negative samples. We use 80% samples for training, 10% samples for validation, and the remaining are used for testing.

Stanford Sentiment Treebank (SST) This dataset (Socher et al. 2013) contains 8545 training samples, 1101 validation samples, 2210 test samples, where each sample is annotated as *very negative*, *negative*, *neutral*, *positive*, or *very positive*.

In this paper, our sentiment lexicon combines the sentiment words from both (Qian et al. 2017) and our manual collection, which totally contains 11,017 sentiment words.

Baseline methods

We compare our models with several state-of-the-art baseline methods, including RNTN (Socher et al. 2013), LSTM/BiLSTM, Tree-LSTM (Tai, Socher, and Manning 2015), CNN, CNN-Tensor (Lei, Barzilay, and Jaakkola 2015), DAN (Iyyer et al. 2015), NCSL (Teng, Vo, and Zhang 2016), LR-LSTM and LR-Bi-LSTM (Qian et al. 2017).

Implementation details

In the experiments, we use 300-dimensional GloVe[1] vectors to initialize the word embeddings for context and sentiment words. We initialize the recurrent weight matrices as random orthogonal matrices and all the bias vectors are initialized to zero. Both LSTM and attention layer have 50 units each. We conduct mini-batch (with size 64) training using RMSprop optimization algorithm to train the model. The learning rate is 0.001, and the dropout rate is 0.5.

Experimental results

In our experiments, the evaluation metric is classification accuracy. We summarize the experimental results in Table 1. Compared to the baseline methods, our models achieve better or comparable results. For example, the classification accuracy increases from 82.9% to 84.0% on MR dataset. This verifies the effectiveness of the proposed approaches that integrates sentiment lexicon into the deep neural networks via attention mechanism. As expected, our multi-head attention model performs better than the single-head attention model. Our model also benefits from the information from different representation subspaces at different positions.

[1]http://nlp.stanford.edu/projects/glove

Method	MR	SST(sent.-level)
RNTN	75.9%	43.4%
LSTM	77.4%	45.6%
BiLSTM	79.3%	46.5%
Tree-LSTM	80.7%	48.1%
CNN	81.5%	46.9%
CNN-Tensor	-	50.6%
DAN	-	47.7%
NSCL	82.9%	47.1%
LR-LSTM	81.5%	48.3%
LR-Bi-LSTM	82.1%	48.6%
SLEA-LSTM (Single-head)	**82.9**%	**48.9**%
SLEA-LSTM (Multi-head)	**84.0**%	**49.3**%

Table 1: Evaluation results

Conclusion

In this paper, we propose a novel sentiment lexicon enhanced attention-based LSTM model which integrates sentiment lexicon into deep neural network via attention mechanisms. Experimental results showed that our method achieves better or comparable results.

References

Iyyer, M.; Manjunatha, V.; Boyd-Graber, J.; and Daumé III, H. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of ACL 2015*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2015. Molding cnns for text: non-linear, non-consecutive convolutions. In *Proceedings of EMNLP 2015*.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL 2005*.

Qian, Q.; Huang, M.; Lei, J.; and Zhu, X. 2017. Linguistically regularized LSTM for sentiment classification. In *Proceedings of ACL 2017*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP 2013*.

Tai, K. S.; Socher, R.; and Manning, C. D. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of ACL 2015*.

Teng, Z.; Vo, D.-T.; and Zhang, Y. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of EMNLP 2016*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *CoRR* abs/1706.03762.