CO-ATTENTION NETWORK AND LOW-RANK BILINEAR POOLING FOR ASPECT BASED SENTIMENT ANALYSIS

Peiran Zhang[†⋆], Hongbo Zhu[†], Tao Xiong[†], Yihui Yang[‡]

[†]Alibaba Group, Hangzhou, China, [‡]Afterpay, Melbourne, Australia

{peiran.zpr,xiaofeng.zhb,weilue.xt}@alibaba-inc.com, yihui.y@outlook.com

ABSTRACT

Aspect Based Sentiment Analysis (ABSA) is an important and challenging task in language understanding. It aims to assign the correct polarity to a given sentence considering the entity on which an opinion is expressed. Extant neural networks usually employ Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), or Attention networks to address the so-called *"target-sensitive sentiment"* problem, referring to the fact that sentence polarity is decided by aspect information and surrounding contexts jointly. However, those models usually are complicated and will incur tremendous training cost. Instead of using sophisticated sequential networks, we present a novel co-attention based network to capture the correlation between aspect and contexts. We evaluate our model on three public datasets and the results demonstrate a strong evidence of improved accuracy (up to 2.32% absolute improvement) and efficiency (model converges at least 4 times faster).

Index Terms— Aspect sentiment analysis, co-attention

1. INTRODUCTION

Aspect Based Sentiment Analysis(ABSA) is one of the most important and challenging tasks in language understanding. It aims to assign the correct polarity to a given sentence by identifying the entity on which an opinion is expressed [1]. The major difference between aspect sentiment analysis and traditional sentiment analysis is that the former is aspect or target oriented. [1] A significant amount of classification errors in traditional sentiment analysis models are caused by ignoring aspect information [2]. Aspect sentiment analysis usually involves two sub-tasks: (1) identifying target words and phrases indicating aspect term; and (2) identifying sentiment polarity based on inferred aspect and its surrounding contexts. In this work we focus on the sentiment classification task defined by SemEval 2016: *"for a predefined set of aspects in a given sentence, determine whether the polarity of each aspect term is positive, negative or neutral"* [1].

[1]For example, in a given review *"The food is good, but the service is terrible"*, two different aspects are mentioned: food and service, and the author assigned different sentiments to them respectively.

Traditional models on ABSA task adopt statistical learning approach based on discrete features such as semantic rules and sentiment lexicons [3, 4], these approaches rely heavily on the quality of handcrafted features and lack the ability to cultivate aspect information [5, 6]. In recent years, neutral network based models, such as Tree-based networks [7], Convolutional Neural Network (CNN) [8], and Recursive Neural Networks (RNN) [2, 7, 9], are becoming popular. It has been noticed that utilizing aspect information is critical for ABSA tasks, Wang et al. [10] defines it as the *"target-sensitive sentiment"* issue, indicating that the sentence polarity is jointly decided by the aspect information and its surrounding contexts. Extant models usually employ sequential networks such as Long Short-Term Memory (LSTM) [11, 5] and Gated Recurrent Unit (GRU) [12], or attention networks [13, 12] to take aspect information into consideration. However, RNN based networks and attention networks usually are complicated and are associated with high training costs, but still have limitations such as the inability to capture long sentence information and nonnegligible attention noises [5].

In this work we propose a novel co-attention [14] based network to cultivate aspect information as well as improve model efficiency. The major difference between co-attention network and traditional attention network is that co-attention network deploys a bi-directional attention mechanism to attend aspect information and context information simultaneously. Specifically, we reshape an input sentence into three parts based on its inferred aspect: the left context, aspect term and the right context [6], and we use a pre-trained word embedding space to generate feature maps for these three subsentences. Two co-attention networks are then deployed to attend left context and aspect term, and right context and aspect term to capture the correlation between aspect term and contexts. We also design a third co-attention network to attend left context and right context to capture any potential semantic or sentiment information that are located far from aspect term. The three co-attention networks reshape the three input subsentences into six feature vectors, instead of concatenating them directly for the final prediction, we use a Low-rank Bilinear Pooling (LRBP) method based on Hadamard product to build the final sentence feature vector of much lower dimension without losing discriminative power [15]. Our model

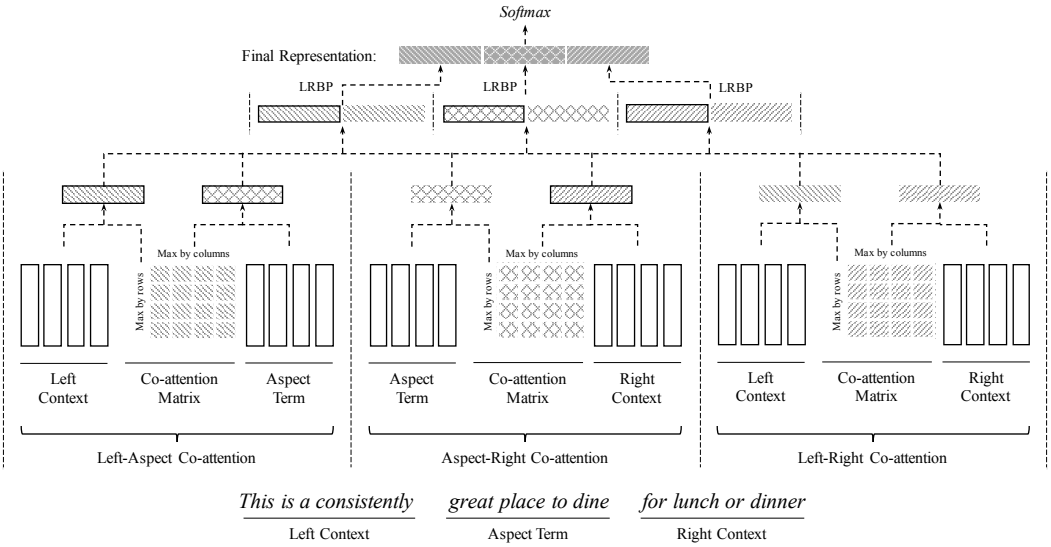

Fig. 1. Model architecture. We adopt three co-attention networks to attend left context and aspect term, aspect term and right context, and left context and right context respectively.

is mainly based on matrix computations and is not using any sequential components such as LSTM or GRU to extract features, therefore, parallelized training can be easily achieved to improve training efficiency.

We evaluate our model on three publicly available datasets: *Laptop*, *Restaurant* and *Twitter*. Experiment results demonstrate that our model outperforms most of extant neural networks on all three datasets, and we improve the state-of-the-art performance on *Restaurant* and *Twitter* datasets. In addition, we notice that our model dominates extant models using one-directional LSTM networks or attention networks. Training time comparison also demonstrates a strong evidence that our model improves training efficiency significantly. We believe this model can be applied to handle large scale datasets with incurring limited training costs.

We organize this paper as follows. In section 2 we describe our proposed model in detail. In section 3 we show our experimental settings and compare experiment results with baselines. We summarize this paper in section 4.

2. MODEL DESCRIPTION

In this section we describe our proposed model in detail. Model architecture is depicted in Figure 1.

2.1. Word Embedding and Sentence Representation

We start by training a word embedding space $\mathbf{E} \in \mathbb{R}^{V \times k}$, where V is the size of vocabulary and k is the dimension for each word vector. For any word w in our vocabulary, we assign feature vector $\mathbf{e}_w \in \mathbb{R}^k$ as w's feature vector, $\mathbf{e}_w$ is the corresponding row of elements in $\mathbf{E}$.

For each input sentence, we separate it into three components based on its inferred aspect[6]: left context, aspect term and right context, and we generate feature maps for them based on individual word embeddings. Specifically, for a left context sentence containing l words, we define its feature map $\mathbf{L} = [\mathbf{e}^{\mathrm{T}}_{w_1^L}, \mathbf{e}^{\mathrm{T}}_{w_2^L}, ..., \mathbf{e}^{\mathrm{T}}_{w_l^L}]$, $\mathbf{L} \in \mathbb{R}^{k \times l}$, $\mathbf{e}_{w_i^L}$ is the word embedding vector for word i contained in left context. Similarly, we calculate feature map $\mathbf{A}$, $\mathbf{A} \in \mathbb{R}^{k \times m}$ for aspect term containing m words, and $\mathbf{R}$, $\mathbf{R} \in \mathbb{R}^{k \times r}$ for right context containing r words.

2.2. Co-attention

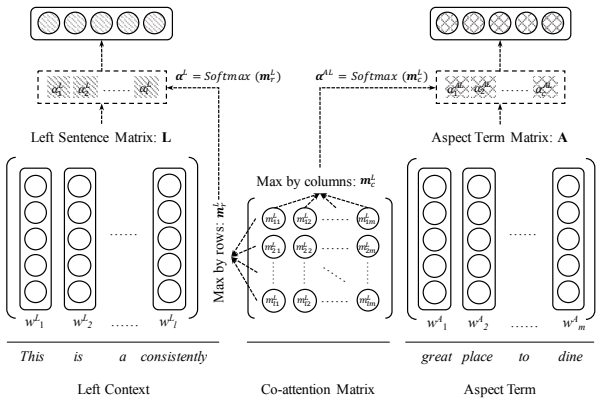

Fig. 2. Co-attention mechanism

We employ co-attention mechanism to attend the aspect term and its surrounding contexts simultaneously. The detail of a co-attention network is depicted in Figure 2. For

the left-aspect co-attention network, we compute two vectors: $\boldsymbol{a}^L$, $\boldsymbol{a}^L \in \mathbb{R}^l$, and $\boldsymbol{a}^{AL}$, $\boldsymbol{a}^{AL} \in \mathbb{R}^m$, as attention weights to reshape sentence matrix and aspect matrix. We define a new feature vector $\mathbf{z}^L = \mathbf{L} \cdot \boldsymbol{a}^L$, $\mathbf{z}^L \in \mathbb{R}^k$ and vector $\mathbf{z}^{AL} = \mathbf{A} \cdot \boldsymbol{a}^{AL}$, $\mathbf{z}^{AL} \in \mathbb{R}^k$, to represent attended left context and aspect term respectively.

To calculate weights in $\boldsymbol{a}^L$ and $\boldsymbol{a}^{AL}$ and allow them to be adjusted simultaneously, we define an affinity matrix $\mathbf{M}^L \in \mathbb{R}^{l \times m}$, which is calculated by:

$$\mathbf{M}^L = tanh(\mathbf{L}^{\mathrm{T}} \cdot \mathbf{W}_b^L \cdot \mathbf{A}) \tag{1}$$

$\mathbf{W}_b^L \in \mathbb{R}^{k \times k}$ is a parameter matrix. We follow the maximization activation approach proposed by Lu et al. [14] to calculate attention weights. We calculate two intermediate vectors $\boldsymbol{m}_r^L$, $\boldsymbol{m}_r^L \in \mathrm{R}^l$ and $\boldsymbol{m}_c^L$, $\boldsymbol{m}_c^L \in \mathrm{R}^m$ by maximizing elements in $\mathbf{M}$ by rows and by columns respectively:

$$\boldsymbol{m}_r^L = max_{row}(\mathbf{M}_{r,i}^L), i = 1, 2, ..., m. \tag{2}$$

$$\boldsymbol{m}_c^L = max_{column}(\mathbf{M}_{j,c}^L), j = 1, 2, ..., l. \tag{3}$$

We normalize $\boldsymbol{m}_r^L$ and $\boldsymbol{m}_c^L$ by inputing them into the *Softmax* function and the outputs are used as co-attention weights: $\boldsymbol{a}^L = Softmax(\boldsymbol{m}_r^L)$, $\boldsymbol{a}^{AL} = Softmax(\boldsymbol{m}_c^L)$.

We repeat the same procedure proposed above to construct the aspect-right co-attention network and left-right co-attention network and calculate another four feature vectors for the three sub-sentences: $\mathbf{z}^{AR}$ and $\mathbf{z}^R$ are generated from aspect-right co-attention network, representing aspect term and right context respectively; and $\mathbf{z}^{L'}$ and $\mathbf{z}^{R'}$ are generated from left-right co-attention network, representing left context and right context respectively. $\mathbf{z}^{AR}, \mathbf{z}^R, \mathbf{z}^{L'}, \mathbf{z}^{R'} \in \mathbb{R}^k$. Parameters need to be learned in these two co-attention networks are $\mathbf{W}_b^R$ and $\mathbf{W}_b^{LR}$, $\mathbf{W}_b^R, \mathbf{W}_b^{LR} \in \mathbb{R}^{k \times k}$.

2.3. Low-rank Bilinear Pooling

The co-attention network generates two feature vectors for each input sub-sentence, these two vectors may carry redundant information and it is unnecessary to input all of them into the final prediction layer. In this case, we employ a Low-rank Bilinear Pooling (LRBP) method based on Hadamard product to reduce the dimension of the final input vector without losing discriminative power [15].

The process of LRBP is straightforward: given two input feature vectors of left context: $\mathbf{z}^L$ and $\mathbf{z}^{L'}$, we compute a projection feature vector $\boldsymbol{f}^L = \mathbf{U}^{\mathrm{T}} \cdot \mathbf{z}^L \circ \mathbf{V}^{\mathrm{T}} \cdot \mathbf{z}^{L'} + \boldsymbol{g}$, $\boldsymbol{f}^L \in \mathbb{R}^c$, and $\circ$ represents the Hadamard product. $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{k \times c}$ are parameters and $\boldsymbol{g} \in \mathbb{R}^c$ is bias, c is a hyperparameter and $c < 2k$. Similarly, we compute projection feature vectors $\boldsymbol{f}^A$ and $\boldsymbol{f}^R$ for aspect term and right context respectively.

We concatenate $\boldsymbol{f}^L, \boldsymbol{f}^A$ and $\boldsymbol{f}^R$ to generate the final feature vector for the input sentence: $\mathbf{z} = \boldsymbol{f}^L \oplus \boldsymbol{f}^A \oplus \boldsymbol{f}^R$ and we input $\mathbf{z}$ into a Softmax classification layer to make the final prediction: $\boldsymbol{p}_i = Softmax(\mathbf{W} \cdot \mathbf{z} + \boldsymbol{b})$, where $\boldsymbol{p}_i \in \mathbb{R}^3$, and $\mathbf{W} \in \mathbb{R}^{3 \times 3c}$, $\boldsymbol{b} \in \mathbb{R}^3$ are parameters to be learned.

		Positive	Negative	Neutral
Laptop	Train	994	870	464
	Test	341	128	169
Restaurant	Train	2,164	807	637
	Test	728	196	196
Twitter	Train	1,561	1,560	3,127
	Test	173	173	346

Table 1. Descriptive statistics

2.4. Training Objective

For the final prediction, we optimize a multi-class focal loss function proposed by Lin et al. [16]. The loss function for one training instance is defined as:

$$J(\theta) = -\sum_{i=1}^{N} y_i \cdot (1 - p_i)^{\gamma} \cdot log(p_i) + \lambda \cdot R \tag{4}$$

where N is number of classes, y_i is the true label for sentence i. γ is a hyperparameter controlling the extent to which we focus on ambiguous cases in the training dataset. $\theta = \{\mathbf{W}_b^L, \mathbf{W}_b^R, \mathbf{W}_b^{LR}, \mathbf{W}, \boldsymbol{b}, \mathbf{U}, \mathbf{V}, \boldsymbol{g}\}$ are model parameters[2]. $R = \|\theta\|_{L2}$ is the L2 regularization term and λ is a hyperparameter measuring the weight of regularization term.

3. EXPERIMENTS

3.1. Datasets

We use three datasets to evaluate our model: the *Laptop* and *Restaurant* datasets are from the 2014 SemEval ABSA task [17] containing reviews on laptops and restaurants; and the *Twitter* dataset is provided by Dong et al. [2], containing twitter posts. We follow previous studies by removing reviews with "conflict" labels to make experiment results comparable [12, 5]. Descriptive statistics are reported in Table 1.

3.2. Baseline Models

We compare our model with following models.

- **Simple SVM**: proposed by Kiritchenko et al. [3], is a traditional support vector machine approach based on discrete feature engineering such as sentiment lexicons.
- **TD-LSTM**: Tang et al. [18] proposed an enhanced LSTM network by incorporating aspect/target information into LSTM network to boost model accuracy.
- **AE-LSTM/ATAE-LSTM**: Wang et al. [11] proposed a LSTM based model with aspect information embedded (AE-LSTM) as an improvement to traditional LSTM methods. They also employ an attention mechanism to further strengthen model performance(ATAE-LSTM).

[2]Six parameter matrices are used in the LRBP process, for simplicity, we just use $\mathbf{U}$ and $\mathbf{V}$ to represent them.

- **MemNet**: unlike sequential neural networks such as LSTM, Tang et al. [19] proposed a deep memory network consisting of multiple layers of attentions to better capture aspect information.

- **IAN**: stands for the Interactive Attention Network proposed by Ma et al. [13], is also an attention based network with context representation and aspect representation learned separately but interactively.

- **RAM**: Chen et al. [12] proposed a Recurrent Attention on Memory network with LSTM to build memory network and multiple attention mechanisms to solve the multiple words attention problem.

3.3. Experimental Settings

We download a GloVe word embedding space with dimension $k = 300$. [3] For the training process, we fix the maximum number of words in left and right context $l = r = 20$ and the maximum number of words in the aspect term $m = 4$.[4] For the hyperparameter in the LRBP process, we set $c = 300$. We use the RMSProp optimizer to optimize model parameters with batch size of 16 and maximum epochs of 100. We set learning rate equals to 0.01. In loss function, $\gamma = 2$ and L2 regularization is set to 10^{-4}.

3.4. Experiment Results

We follow previous works and use *Accuracy* to evaluate model performance [19, 12, 8]. Results are reported in Table 2. Statistics demonstrate that our model is among the top two models on all the datasets, we outperform traditional one-directional LSTM networks (e.g., TD-LSTM and AE-LSTM/ATAE-LSTM) and attention based networks (e.g. MemNet and IAN). On the *Laptop* dataset, only RAM has a slightly higher accuracy than ours; On the *Restaurant* dataset, the difference between our model and the best of extant models are close, we improve accuracy by 0.12%. On the *Twitter* dataset, we significantly improve the best performance by 2.32%. RAM use bi-directional LSTM networks to extract features from original input sentence to better capture sequential information [12]. However, our model employs no sophisticated sequential neural components and simply use co-attention networks to capture the dynamics between aspect term and contexts. The three co-attention networks have no interdependencies and can be trained simultaneously.

[3]Pre-trained GloVe word embedding result can be downloaded at https://nlp.stanford.edu/projects/glove/. We randomly assign elements in [-0.1, 0.1] to words that are not contained in the pre-trained embedding space.

[4]The lengths of different terms are aligned with Zheng and Xia [6]. For a given sentence, we identify the target words/terms as the aspect term, target words/terms are pre-defined, and words to the left of the aspect term and to the right of the aspect term are identified as the left context and right context respectively.

Models	*Laptop*	*Restaurant*	*Twitter*
Simple SVM	70.49	80.16	63.40
TD-LSTM	71.83	78.00	66.62
AE-LSTM	68.90	76.60	-
ATAE-LSTM	68.70	77.20	-
MemNet	70.33	78.16	68.50
IAN	72.10	78.60	-
RAM	**74.49**	80.23	69.36
Our Approach	73.20	**80.35**	**71.68**

Table 2. Experiment Results. Accuracy is reported, results of SVM, MemNet and RAM are from Chen et al. [12]; results of TD-LSTM, AE-LSTM/ATAE-LSTM, IAN, and TNet is from Li et al. [5].

Models	One Epoch	Convergence
TD-LSTM	1.21	42.35
RAM	6.20	186.00
Our Approach	**0.07**	**7.35**

Table 3. Training time for models in seconds. Training time on *Restaurant* dataset is reported. We set the batch size for all models to be 200 and other hyperparameters are set as reported in original papers.

Therefore, our model achieves much faster training efficiency than RAM.[5]

To evaluate our training efficiency, we compare the training time of our model with TD-LSTM, which is a one-directional LSTM based network, and RAM, which is a bi-directional LSTM network. We train all models on a server with a GTX 1080Ti GPU and training times for one epoch and for models to converge are recorded. Experiment results are reported in Table 3. Compared with TD-LSTM, our model reduces the training time of one epoch by 94% and training time for model to converge by 83%. Compared with RAM, our model reduces the training time for one epoch and model to converge by 99% and 96% respectively.

4. CONCLUSION

In this work we propose a novel co-attention based network to capture the relationship between aspect term and its surrounding contexts, which helps address the *"target-sensitive sentiment"* issue in the ABSA task. We conduct extensive experiments on three publicly available datasets, experiment results demonstrate that our model outperforms most of extant ABSA models on all three datasets and we even achieve the best model accuracy on two datasets. In addition, our model involves only matrix computations and it could achieve parallelized training to improve training efficiency. We believe this model can be applied to address large scale datasets with incurring limited training costs.

[5]We repeat the training process for multiple times and the results are not significantly different, detail results can be provided upon request.

5. REFERENCES

[1] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al., "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 19–30.

[2] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu, "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)*, 2014, vol. 2, pp. 49–54.

[3] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad, "Nrc-canada-2014: Detecting aspects and sentiment in customer reviews," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 437–442.

[4] Duy-Tin Vo and Yue Zhang, "Target-dependent twitter sentiment classification with rich automatic features.," in *IJCAI*, 2015, pp. 1347–1353.

[5] Xin Li, Lidong Bing, Wai Lam, and Bei Shi, "Transformation networks for target-oriented sentiment classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018, pp. 946–956, Association for Computational Linguistics.

[6] Shiliang Zheng and Rui Xia, "Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention," *arXiv preprint arXiv:1802.00892*, 2018.

[7] Kai Sheng Tai, Richard Socher, and Christopher D Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[8] Wei Xue and Tao Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018*, 2018, pp. 2514–2523.

[9] Meishan Zhang, Yue Zhang, and Duy-Tin Vo, "Gated neural networks for targeted sentiment analysis.," in *AAAI*, 2016, pp. 3087–3093.

[10] Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang, "Target-sensitive memory networks for aspect sentiment classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, vol. 1, pp. 957–967.

[11] Yequan Wang, Minlie Huang, Li Zhao, et al., "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[12] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 452–461.

[13] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2017, IJCAI'17, pp. 4068–4074, AAAI Press.

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[15] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang, "Hadamard product for low-rank bilinear pooling," in *Proceedings of the 2017 International Conference on Learning Representation*, 2017.

[16] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[17] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al., "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 10th international workshop on semantic evaluation*, 2014, pp. 27–35.

[18] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu, "Effective lstms for target-dependent sentiment classification," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 3298–3307.

[19] Duyu Tang, Bing Qin, and Ting Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 214–224, Association for Computational Linguistics.