SaliencyBERT: Recurrent Attention Network for Target-Oriented Multimodal Sentiment Classification

Jiawei Wang[1], Zhe Liu[1](✉), Victor Sheng[2], Yuqing Song[1], and Chenjian Qiu[1]

[1] Jiangsu University, Zhenjiang 212013, China
lzhe@ujs.edu.cn
[2] Texas Tech University, Lubbock, TX, USA

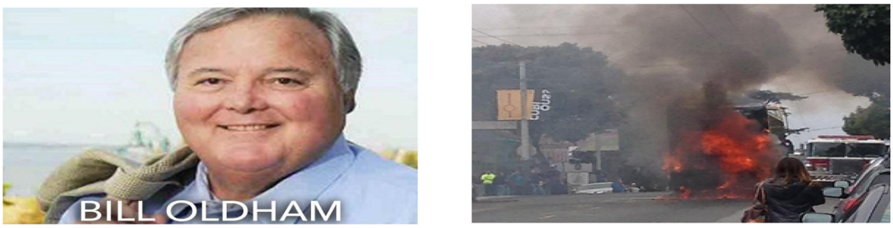
Abstract. As multimodal data become increasingly popular on social media platforms, it is desirable to enhance text-based approaches with other important data sources (e.g. images) for the Sentiment Classification of social media posts. However, existing approaches primarily rely on the textual content or are designed for the coarse-grained Multimodal Sentiment Classification. In this paper, we propose a recurrent attention network (called SaliencyBERT) over the BERT architecture for Target-oriented Multimodal Sentiment Classification (TMSC). Specifically, we first adopt BERT and ResNet to capture the intra-modality dynamics with the textual content and the visual information respectively. Then, we design a recurrent attention mechanism, which can derive target-sensitive visual representations, to capture the inter-modality dynamics. With recurrent attention, our model can progressively optimize the alignment of target-sensitive textual features and visual features and produce an output after a fixed number of time steps. Finally, we combine the loss of all-time steps for deep supervision to prevent converging slower and overfitting. Our empirical results show that the proposed model consistently outperforms single modal methods and achieves an indistinguishable or even better performance on several highly competitive methods on two multimodal datasets from Twitter.

Keywords: Target-oriented multimodal sentiment classification · BERT architecture · Recurrent attention

1 Introduction

With the increasing popularity of social media, a large number of multimodal posts containing images and text are generated by users on social media platforms such as Twitter, Facebook, and Flickr to express their attitudes or opinion. It is quite valuable to analyze such large-scale multimodal data to study the user's emotional orientation toward certain events or topics. Target-oriented Sentiment Classification (TSC) is a fine-grained sentiment classification task, which identifies sentiment polarities through individual opinion targets hidden inside each input sentence. For example, "More on this dramatic bus fire on Haight Street by @Evan", the polarity of the sentence towards the "Haight Street" is neutral while the polarity is negative in terms of "bus" in Fig. 1.b.

H. Ma et al. (Eds.): PRCV 2021, LNCS 13021, pp. 3–15, 2021.
https://doi.org/10.1007/978-3-030-88010-1_1



(a) Congratulations to Shelby County Sheriff [***Bill Oldham***]$_{positive}$ on his re-election.

(b) More on this dramatic [***bus***]$_{negative}$ *fire* on [***Haight Street***]$_{neutral}$ by @Evan

Fig. 1. Representative examples for Target-oriented Multimodal Sentiment Classification in our Twitter datasets. The target words and corresponding sentiment are highlighted and show that different target opinions in the same sentence may express different sentiment polarities.

Since TSC was proposed, this problem of fine-grained Sentiment Classification has been receiving the attention and research of the academic community. Early research uses statistical methods, such as support vector machines [1, 2], which require carefully designed manual features. In recent years, neural network models [3, 4] have been widely used to automatically learn the representation of target words and their context. Attention mechanisms [5, 6] have also been studied to strengthen the target characteristics' attention to important words in the context. However, most existing target-oriented sentiment classification methods are only based on text content and ignore other associated data sources. Multimodal posts usually come with images, and these images often provide valuable insights into users' sentiment (e.g., the smile of the man is a sign of positive sentiment in Fig. 1.a). In a word, due to the shortness and informality of the text in a post, the sentiment classification of a target sometimes depends largely on its associated images. Especially for sentences with multiple targets, the textual content often only expresses the subjective feelings of a certain target but ignores other targets. The introduction of related pictures can help supplement additional sentiment information. Therefore, Target-oriented Multimodal Sentiment Classification (TMSC) with text and images will be meaningful. A tensor fusion network [8] and a memory fusion network [9] was designed to better capture the interactions between different modalities. However, these methods are designed for the coarse-grained dialogue multimodal sentiment classification and do not explore the relationship between the individual opinion target and the multi-modal content.

As the aforementioned previous causes are, in this paper, we propose to use a soft, sequential, top-down attention mechanism on top of the recent BERT [10] architecture. Through the enhancement of visual modality, we can more accurately capture sentiment polarities of individual opinion target in each input sentence. Specifically, a stand-alone BERT can be used at each time step to obtain rich target-sensitive textual representations. Then, the attention mechanism learns appropriate attention weights for different regions in the associated image to induce the alignment of target-sensitive textual representations and visual representations. Furthermore, we adopt a feed-forward network and two-layer norms with residual connections to obtain the output of the current time step. By deconvoluting the output of the current time step, the rich interactive information between target-sensitive textual features and visual features is propagated to the higher resolution layer as the input of the next time step. Through multiple time steps, to progressively optimize the alignment of the target-sensitive textual representation and the visual representation. It is sort of analogous to the cognition procedure of a person, who might first notice part of the important information at the beginning, then notices more as she reads through. Finally, we combine the loss of each time step for deep supervise to prevent converging slower and overfitting.

The main contributions of this paper can be concluded as follows: (1) We propose a recurrent attention network for Target-oriented Multimodal Sentiment Classification, which uses the BERT architecture. Its input consists of two modalities (i.e., text, image). (2) We further develop a soft, sequential, top-down attention mechanism to effectively capture the intra-modality and inter-modality dynamics. The core idea is to obtain the saliency feature of a certain modal through the enhancement of another modal. (3) We also present a deep supervision method to overcome the problems caused by the number of unrolling steps, which makes the back-propagation convergence rate slower and easy to overfit.

To investigate the performance of the proposed model, we conduct comprehensive experiments in a supervised setting on two benchmark multimodal datasets. The proposed model can achieve an indistinguishable or even better performance over several highly competitive multimodal approaches.

2 Related Work

Early sentiment classification was usually performed using machine learning [11] and lexical-based methods [12]. These technologies are inseparable from a lot of manual work, such as data preprocessing and manually designing a set of task-specific features. The preprocessing becomes difficult as the number of data increases. Deep Learning is a relatively new approach that has been employed to carry out sentiment analysis [13]. Deep Learning has been found to perform better than traditional machine learning or lexical-based approaches when an enormous amount of training data is available. Target-oriented Sentiment Classification (TSC) has been extensively studied in recent years [14]. Target-oriented sentiment classification is a branch of sentiment classification, which requires considering both the sentence and opinion target. Unlike the previous coarse-grained dialogue sentiment classification, target-oriented fine-grained sentiment classification [7, 15] is more challenging. Because different target words in the same sentence may express different sentiment polarities. Inspired by the advantages of attention mechanisms in capturing long-range context information in other NLP tasks [10, 16, 17], many recent studies have devised different attention mechanisms to model the interactions between the target entity and the context [18, 19].

With the increasing popularity of social media, a large number of multimodal posts are generated by users on social media platforms to express their attitudes or opinion. People began to study the use of information from different modalities (visual, auditory, etc.) to provide additional emotional information for traditional text features. Early work was designed for coarse-grained sentiment analysis for multimodal dialogue and focused on how integrating other relevant information with text features. Bertero et al. [20] proposed a hierarchical CNN method, which classifies the emotions and emotions of each utterance in the interactive speech dialogue system. But their work is designed for coarse-grained sentence-level sentiment analysis, whereas our work targets at fine-grained target-level sentiment analysis. In recent years, the majority of these studies learned to effectively model the interactions between the target entity, the textual context, and the associated image context. A tensor fusion network [8] and a memory fusion network [9] was designed to better capture the interactions between different modalities. Attention mechanisms [21, 23] are studied to enhance the influence of target opinion on the final representation for prediction. Yu et al. [22] proposed an entity-sensitive attention and fusion network, which uses the single attention mechanism to perform target-image matching to derive target-sensitive visual representations. However, these single-attention-based methods may hide the characteristics of each attended word when attending multiple scattered words with one attention.

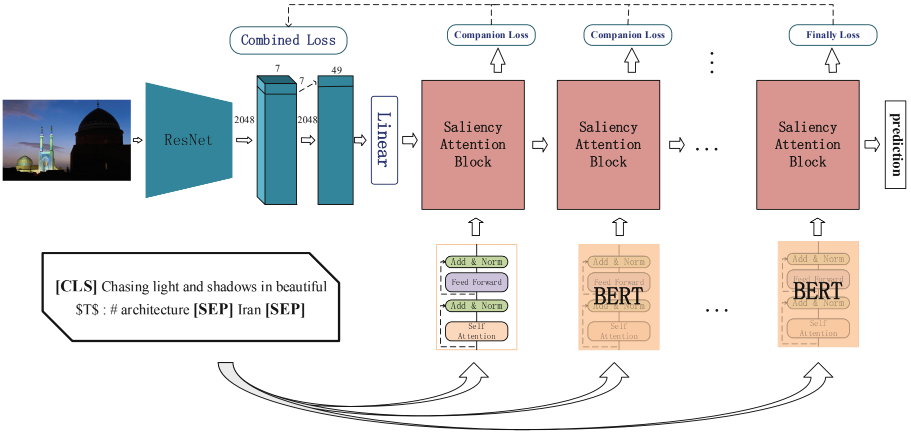


Fig. 2. Overview of our multimodal model for TMSC. The final decision of sentiment classification is obtained after a fixed time step.

3 Proposed Model

In this section, we first formulate our task and introduce visual encoder and textual encode respectively. Then, we will discuss our multimodal recurrent attention network in detail, which is end-to-end trainable.

3.1 Task Definition

Our task is to learn a target-oriented multimodal sentiment classifier so that it can use both textual modal data and visual modal data to predict the sentiment label of the opinion target in an unseen sample. Specifically, given a sentence $X = \{x_1, x_2, \ldots, x_n\}$ containing an opinion target T (a sub-sequence of words in X) and an associated image V. For the opinion target T, it has a sentiment label Y, which can be 2 for positive, 1 for negative, or 0 for neutral.

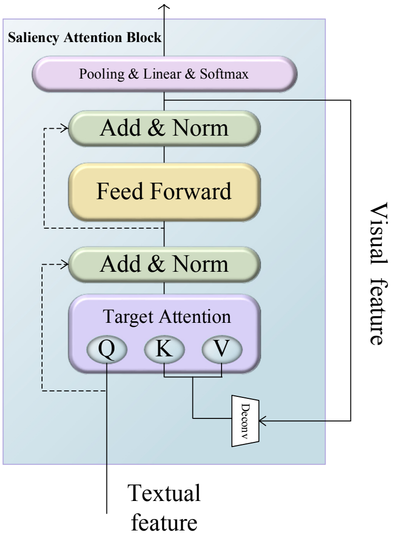


Fig. 3. Overview of Saliency Attention Block. By deconvoluting the output of the current time step, the rich interactive information between target-sensitive textual features and visual features is propagated to the higher resolution layer as the input of the next time step.

3.2 Recurrent Attention Network

Visual Feature Encode. Unlike other models that use only single modal data, information from visual modalities is leveraged to provide additional feature information to traditional textual features in our model. As illustrated in Fig. 2, for the associated image V, we first resize it to 224 × 224 pixels and adopt one of state-of-the-art pre-trained image recognition models ResNet-152 [24] to obtain the output of the last convolutional layer.

$$\mathrm{ResNet(V)} = \left\{R_i | R_i \in R^{2048}, i = 1, 2, \ldots, 49\right\} \tag{1}$$

Which essentially divides the original image into $7 \times 7 = 49$ regions and each region consists of the vector R_i of 2048 dimensions. Next, we adjust the output through linear transformation to project the visual features to the same space of textual features.

$$\mathrm{G} = W_v \mathrm{ResNet(V)} \tag{2}$$

Where $W_v \in R^{d \times 2048}$ is the learnable parameter and d is the latitude of the word vector. Finally, the visual features G $\in R^{d \times 49}$ are fed into the Saliency Attention Block.

Textual Feature Encode. We input the personal opinion target words and the remaining context words as two sentences into stand-alone BERT at each time step to obtain target-sensitive textual representations. For example, the BERT input is given in Fig. 2. The preprocessing method is to convert each sentence X into two sub-sentences: the opinion target and the remaining context and connect them as the input sequence of BERT [25]. Different from the most of existing recurrent models [27], the data are encoded once, and the same underlying features are input, our model uses a stand-alone textual encoder in each time step t. Using stand-alone textual encode can produce a target-sensitive textual representation in each step to allow the target to attend to different parts of the sentence during each pass. Because when paying attention to multiple scattered words at once, the characteristics of each word of attention may be hidden. Finally, we will obtain rich target-sensitive textual representations S for the recurrent attention mechanism.

$$S = BERT_t(\mathrm{X}) \tag{3}$$

Where $S \in R^{d \times N}$, d is the vector dimensions and N is the maximum length of the sentences.

Saliency Attention Block. The recurrent attention mechanism contains two key components: the Saliency Attention Block and the sequential nature of the model. As illustrated in Fig. 3, the Saliency Attention Block, which we improve over the BERT architecture, can select the regions where the input sequence is closely related to the target, and other irrelevant regions are ignored. Specifically, we regard visual feature G as a sequence of 49 items, each of which is a vector of d-dimensions. Then visual feature G is used as the input sequence, and the textual feature is used as the target. Different from the existing model [25], which only inputs a single target word for matching, our method makes full use of the target word and its context. Because we believe that the single target word without its context cannot provide good textual in-formation for the visual representations.

Following the BERT architecture, we use the m-head target attention mechanism to match the target-sensitive textual representations and the image to obtain a visual representation that is sensitive to the target-sensitive textual representations. Specifically, we process the S as the target to generate the query Q and the G is used as the input sequence to generate the key K and value V. So as to use the target to guide the model to align it with the appropriate area, which is the image areas closely related to the target-sensitive textual representations, and to assign high attention weights. Then, the i-th head target attention ATT_i is defined as follow.

$$ATT_i(G, S) = softmax\left(\left[W_{Qi}S\right]^T [W_{Ki}G]/\sqrt{d/m}\right) \times [W_{Vi}G]^T \tag{4}$$

Where $\{W_{Qi}, W_{Ki}, W_{Vi}\} \in R^{d/d \times m}$ are learnable parameters corresponding to queries, keys, and values respectively. After that, we adopt the same structure as the standard BERT. The outputs of the m attention mechanisms (MATT) are concatenated together followed by a linear transformation. Then, using the feedforward network (MLP) and the two-layer norms (LN) with residual connections to obtain the target-sensitive visual output TI.

$$\mathrm{MATT(G,\ S)} = W_m[ATT_1(G, S), \ldots, ATT_m(G, S)]^T \tag{5}$$

$$\mathrm{TI(G, S) = LN(S + MLP(LN(S + MATT(G, S)))} \tag{6}$$

Where $W_m \in R^{d \times d}$ is the learnable parameter. Then, we stack such TI(G, S) to obtain the final target-sensitive visual output H, where H $\in R^{d \times m}$ and the first token of H is essentially a weighted sum of the textual features and the 49 regions image features. Furthermore, We provide a pool function to obtain the first token: $\mathrm{O} = H_0$, and then feed the softmax function to the classification at time step t.

$$\mathrm{p}(\mathrm{y}|O) = \mathrm{softmax}\left(W^T O\right) \tag{7}$$

Where $\mathrm{W} \in R^{d \times 3}$ is the learnable parameter. Similar to the RNN, the sequential nature of the model can connect the information H_{t-1} of the previous time step to the current time step for learning, and actively process the interactive information related to the multimodal data at each time step to refine its estimate of the correct label. Specifically, the rich interactive information between target-sensitive textual features and visual features is propagated to the higher resolution layer as the input of the next time step by deconvoluting the output of the current time step. We feed the H_{t-1} to a deconvolution to get the input H_t at the next time.

$$H_t = \mathrm{Deconv1D}(H_{t-1}) \tag{8}$$

Finally, Due to the number of unrolling steps, the model may have more and more parameters, which makes the back-propagation convergence rate slower and easy to overfit. To overcome this problem, we introduce the method of deep supervision where auxiliary classifiers are added at all Saliency Attention Blocks and their companion losses are added to the loss of the final layer. At training time, we use the standard cross-entropy loss function to obtain companion losses after auxiliary classifiers.

$$\mathcal{J} = -\frac{1}{|\mathrm{D}|}\sum_{j=1}^{|\mathrm{D}|} \log\left(y^{(j)} \middle| O^{(j)}\right) \tag{9}$$

Then, we optimize a loss function that is a sum of the final loss and companion losses with all Saliency Attention Blocks.

$$\mathcal{L} = \sum_{1}^{\mathrm{t}} \mathcal{J}_t \tag{10}$$

4 Experiments

In this section, the data set, baseline method, and experimental setup are described. Then, we empirically studied the performance of SaliencyBERT on several multimodal data sets and discussed important parameters.

4.1 Experiment Settings

Datasets. To evaluate the effect of SaliencyBERT, we adopt two multimodal publicly available named entity recognition datasets TWITTER-15 and TWITTER-17 respectively collected by Zhang et al. [14] and Lu et al. [26]. However, they only provide annotated target opinions, textual contents, and their associated images in each Twitter. Yu et al. [25] annotated the sentiment (positive, negative, and neutral) towards each target by three domain experts and all the entities belong to four types: Person, Location, Organization, and Miscellaneous. Finally, a total of 5288 tweets in TWITTER-15 and 5972 tweets in TWITTER-17 are retained. Then, we randomly separate all image-text pairs in each dataset into a training set, a development set, and a test set with the proportion of 60%, 20%, and 20% respectively. Each sentence has an average of 1.3 targets at TWITTER-15 and 1.4 targets at TWITTER-17.

Baselines. In this paper, we will investigate the performance of our model by comparing it with baseline models. The baseline models can be categorized into three groups: models using only the visual modality, models using only the text modality, and models with multiple modalities. The models are listed as follows: ResNet-Target: a pre-trained image recognition model and concatenating the target word; AE-LSTM [22]: incorporating aspect embeddings and target-specific attention mechanism; MGAN [20]: building up a multi-grained attention network for fusing the target and the context; BERT [10]: adding a pooling layer and the softmax function on top of BERT$_{\text{base}}$; Res-MGAN-TFN [8]: using Tensor Fusion Network (TFN) to fuse the textual and visual representations; Res-BERT: replacing the textual encoder in Res-MGAN-TFN with BERT; ESAFN [23]: fusing the entity-sensitive textual representations and the entity-sensitive visual representations with a bilinear interaction layer; mPBERT [26]: a multimodal BERT architecture that directly concatenates the image features with the final hidden states of the input sequence, followed by multimodal attention mechanism; Propose Model: For convenience, we denote SaliencyBERT-k to be our model that is unrolled for k-time steps. We select three representatives Proposed Model-1, Proposed Model-3, and Proposed Model-6 to compare with baseline models.

Detailed Parameters. We build a textual encoder model on top of the pre-trained cased BERT$_{\text{base}}$ model, and the parameters are both initialized from the pre-trained opinion model. The images with size 224 × 224 and channel RGB are used as the visual input, and pre-trained ResNet-152 are used to encode the visual features. Finally, like the BERT$_{\text{base}}$ model, we set the learning rate as 5e−5, the number of attention heads m = 12, and the dropout rate as 0.1 for our SaliencyBERT. All the models are fine-tuned between 15 and 45 epochs, depending on the number of unrolling steps.

Table 1. Experimental results on our two Twitter datasets for TMSC.

Modality	Method	TWITTER-15		TWITTER-17	
		ACC	Mac-F1	ACC	Mac-F1
Visual	ResNet-Target	59.88	46.48	58.59	53.98
Text	AE-LSTM	70.30	63.43	61.67	57.97
	MGAN	71.17	64.21	64.75	64.46
	BERT	74.02	68.86	67.74	65.66
Text+Visual	Res-MGAN-TFN	70.30	64.14	64.10	59.13
	Res-BERT	75.89	69.00	67.82	65.26
	ESAFN	73.38	67.37	67.83	64.22
	mPBERT	75.79	71.07	68.80	67.06
	SaliencyBERT-1	75.60	69.88	67.66	65.54
	SaliencyBERT-3	**77.03**(76.08)	**72.36**(71.00)	**69.69**(68.23)	**67.19**(65.55)
	SaliencyBERT-6	73.76(63.19)	65.36(60.56)	68.62(52.37)	65.51(48.64)

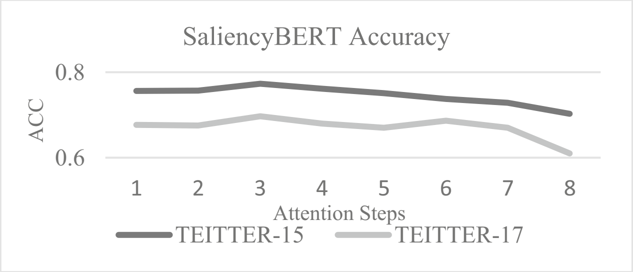
Table 2. Breakdown of accuracy with single opinion target and multiple opinion targets.

Method	TWITTER-15		TWITTER-17	
	Target = 1	Target ≥ 2	Target = 1	Target ≥ 2
BERT	77.40	73.42	68.64	66.25
ESAFN	73.14	73.76	68.16	67.53
mPBERT	76.50	73.88	**68.88**	68.37
SaliencyBERT-3	**79.27**	**75.64**	68.81	**69.89**

4.2 Results and Analysis

Main Metrics. We report the accuracy (ACC) and Macro-F1 score of the single modal methods and the highly competitive multimodal methods on all the two multimodal datasets in Table 1. The results show that the proposed model outperforms all of the above baseline methods. We can observe that the single visual modal method (ResNet-Target) is the worst among all the methods. Since the single visual model classifies sentiment with only visual data, this means that related images cannot handle goal-oriented emotion prediction independently. The simple method (e.g. Res-MG-TFN) of splicing visual modalities and textual modalities is inferior to BERT models that only use the textual modality. These results show that the introduction of multimodal data brings more abundant features, but also produces a lot of redundant information and noise. Simple fusion methods have difficulty in directly capturing the interactions between the two modalities. Worse still, it may negatively affect the final results. By observing BERT, Res-BERT, and mPBERT, we can know that the BERT model pays attention to the text representation in fine granularity and achieves good results, but the image can enhance the text-based method on the whole. Finally, when we use SaliencyBERT-3 with deep supervision, the model consistently achieves the best results on the two datasets. As time steps increase, the results without the deep supervision method get worse. These results appear to be in alignment with our hypothesis that using Saliency Attention Block in multiple time steps can well capture intra-modality and inter-modality dynamics. However, the number of unrolling steps makes the back-propagation convergence rate slower and easy to overfit.

Table 3. The accuracy rates for all models discussed so far.



Auxiliary Metrics. To further analyze the advantages of SaliencyBERT over other strong baseline methods, we divided the test set of the two datasets into a single target and multiple targets. Our experimental results are shown in Table 2. From Table 2, we can see that our model performs better than BERT, ESAFN, and mPBERT in dealing with multiple target sentences. These results are consistent with what we observed before. For single target sentences, our model can progressively optimize target-text encoding and the alignment of target-sensitive textual features and the visual features. through multiple time steps, and for multiple target sentences, our model can capture accurate and diverse target-sensitive textual features and progressively optimize the alignment of the target-sensitive textual features and the visual features.

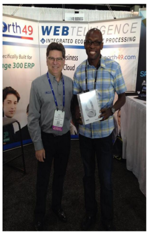
Associated Image	Input Sentence & Predicted Label	Associated Image	Input Sentence & Predicted Label
	1. [**Harold Russell**]$_{positive}$ has SUCH Swagger in [**THE BEST YEARS OF OUR LIVES**]$_{positive}$. BERT :(positive, neutral) ESAFN :(positive, neutral) mPBERT :(positive, neutral) SaliencyBERT :(positive, neutral)		2. Congratulations to our second draw winner Bulaire Leber of [**ADSS Global**]$_{positive}$, [**Haiti**]$_{neutral}$. Thanks for participating, [**Bulaire**]$_{positive}$. BERT :(neutral, neutral, positive) ESAFN :(neutral, neutral, positive) mPBERT :(neutral, neutral, positive) SaliencyBERT :(positive, neutral, positive)

Fig. 4. Representative examples for Target-oriented Multimodal Sentiment Classification in our Twitter datasets.

Moreover, Table 3 shows the accuracy rates for all models discussed so far. The results show that more attention steps weaken accuracy rates during our task. Considering that we have no more than 6 targets in the same sentence, too many attention steps make it difficult for the model to optimize.

Case Study. We can see that the test sample of Fig. 4 shows the input sentence, associated image, and predicted label of different methods. First, in Fig. 4, because the user-posted by the correlated image contains a smiley face, the multimodal approach completely correctly predicts the sentiment of the three target views. However, in Fig. 4, we can see that the multimodal approach makes the same incorrect prediction as the unimodal approach, although it utilizes the relevant images. This may be due to the fact that the relevant images posted by the users failed to provide valid information. These examples further confirm that our multimodal approach combining images and text can somewhat enhance the validity of the text-based approach, but relies on the relevant images posted by the users.

5 Conclusion

In this paper, we studied Target-oriented Multimodal Sentiment Classification (TMSC) and proposed a multimodal BERT architecture to effectively capture the intra-modality and inter-modality dynamics and progressively optimize the alignment of target-sensitive textual features and the visual features. Extensive evaluations on two datasets for TMSC demonstrated the effectiveness of our model in detecting the sentiment polarity for individual opinion targets.

References

1. Wagner, J., 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 (2014)
2. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: 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 (2014)
3. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: 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), pp. 49–54 (2014)
4. Nguyen, T. H., & Shirai, K.: PhraseRNN: phrase recursive neural network for aspect-based sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2509–2514 (2015)
5. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 4068–4074 (2017)
6. Li, C., Guo, X., Mei, Q.: Deep memory networks for attitude identification. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 671–680 (2017)

7. Xue, W., Li, T.: Aspect based sentiment analysis with gated convolutional networks. In: Proceedings Annual Meeting Association for Computational Linguistics, pp. 2514–2523 (2018)
8. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Empirical Methods in Natural Language Processing, pp. 1103–1114 (2017)
9. Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P.: Memory fusion network for multi-view sequential learning. In: AAAI, pp. 5634–5641 (2018)
10. Vaswani, A., et al.: Attention is all you need. In: Proceedings Neural Information Processing System, pp. 5998–6008 (2017)
11. Li, J., Qiu, L.: A Sentiment Analysis Method of Short Texts in Microblog. A Sentiment Analysis Method of Short Texts in Microblog. IEEE Computer Society (2017)
12. Fan, X., Li, X., Du, F., Li, X., Wei, M.: Apply word vectors for sentiment analysis of APP reviews. In: 2016 3rd International Conference on Systems and Informatics, ICSAI 2016, 2017, no. Icsai, pp. 1062–1066 (2016)
13. Tang, D., Wei, F., Qin, B., Liu, T., Zhou, M.: Coooolll: a deep learning system for twitter sentiment classification. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 208–212 (2014)
14. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. WIREs Data Mining Knowl. Discov. **8**(4), e1253 (2018)
15. Tang, D., Qin, B., Feng, X., and Liu, T.: Effective LSTMs for target-dependent sentiment classification. In: Computer Conference, pp. 3298–3307 (2015)
16. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings International Conference Learning Representation, pp. 1–15 (2014)
17. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings Conference North American Chapter Association Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
18. Majumder, N., Poria, S., Gelbukh, A., Akhtar, M.S., Ekbal, A.: IARM: inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3402–3411 (2018)
19. Fan, F., Feng, Y., Zhao, D.: Multi-grained attention network for aspect level sentiment classification. In: Proc. Conf. Empir. Methods Nat. Lang. Process, pp. 3433–3442 (2018)
20. Bertero, D., Siddique, F.B., Wu, C.S., Wan, Y., Chan, R.H.Y., Fung, P.: Real-time speech emotion and sentiment recognition for interactive dialogue systems. In: Proceedings of the 2016 Conference on Empirical Methods in NLP, pp. 1042–1047 (2016)
21. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615 (2016)
22. Yu, J., Jiang, J., Xia, R.: Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 429–439 (2019)
23. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I.: Ask me anything: dynamic memory networks for natural language processing. arXiv:1506.07285 (2015)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
25. Yu, J., Jiang, J.: Adapting BERT for target-oriented multimodal sentiment classification. In: Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19 (2019)

26. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: Visual attention model for name tagging in multimodal social media. In: The Association for Computational Linguistics, pp. 1990–1999 (2018)
27. Zoran, D., Chrzanowski, M., Huang, P.S., Gowal, S., Mott, A., Kohli, P.: Towards robust image classification using sequential attention models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9483–9492 (2020)