2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) | 978-1-6654-8474-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/AEMCSE55572.2022.00059

Multimodal sentiment analysis based on multi-head self-attention and convolutional block attention module

1st Feng Geng
School of Information Science and Electrical Engineering
Shandong Jiaotong University
Jinan, China
gf_sdjtu2022@163.com

2nd Hai Yang*
School of Information Science and Electrical Engineering
Shandong Jiaotong University
Jinan, China
yh_sdjtu@163.com

3rd Changde Wu
School of Information Science and Electrical Engineering
Shandong Jiaotong University
Jinan, China
wcd2680@163.com

4th Jinqiang Li
School of Information Science and Electrical Engineering
Shandong Jiaotong University
Jinan, China
ljq133558@163.com

***Abstract*—Sarcasm is a type of emotional expression. Sarcasm is commonly used on social media to express the inverse of what appears to be a statement and what is said. Previous automatic sarcasm detection mainly focused on text. With the rise of image sharing mode on social media platforms, text cannot fully reveal users' emotions, so people begin to study multimodal sentiment analysis by combining text and images. Previous researches on sarcasm detection have used Bidirectional Long Short-term Memory Network (Bi-LSTM) and Residual Network (ResNet) to extract text and image feature vectors, respectively. While Multi-Head Self-Attention (MH-SA) is added to the Bi-LSTM model to perform relation extraction, which can effectively avoid complex feature engineering in traditional tasks. In the process of image extraction, the channel attention module (CAM) and the spatial attention module (SAM) are used to weight different spatial and channel features and focus on different regions and features of the image. The two complement each other, greatly improving the network's ability to express features. On the Twitter dataset, our proposed model has a sarcasm detection accuracy of 87.55 %, which outperforms most models proposed in current papers.**

Keywords-Multimodal, sentiment analysis, sarcasm detection

I.INTRODUCTION

Social media is one of the most popular places for people to express themselves and share information. Sarcasm is widely used in social media to express different emotions. Social media such as Twitter show abundant sarcasm phenomena, which are very suitable for sarcasm detection research. Due to the complexity of multi-modal data and the ambiguity of sarcasm language, we need to combine images and texts to understand the true intent of the tweet. For example, without seeing the dark clouds in the image, we can't tell what the true meaning of "the weather is so nice" is. Therefore, detecting the combination of image and text data can help reveal or refute the sarcastic nature of Twitter content.

To study the multimodal sarcasm detection problem on Twitter data, this paper uses Bi-LSTM and MH-SA to extract text features and uses pre-trained Resnet50 and Convolutional Block Attention Module (CBAM) to extract image features. An overview of the CBAM and Multi-Head Self-Attention model (CMHA) proposed in this paper is shown in Figure 1. Our main contributions are summarized as follows.

1. In the process of text feature extraction, the Multi-head self-attention (MH-SA) is added to the Bi-LSTM algorithm for relation extraction, avoiding complex feature engineering in traditional tasks.

2. In the process of image extraction, CAM and SAM are used o weight different spatial and channel features, focus on different areas and features of the image, and improve image feature representation ability.

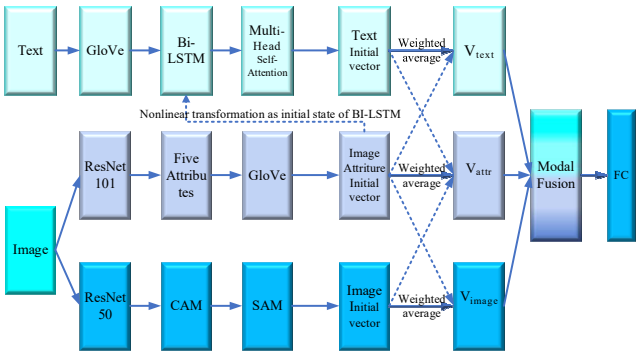

Figure 1. Overview of the CMHA Model

978-1-6654-8474-9/22/$31.00 ©2022 IEEE
DOI 10.1109/AEMCSE55572.2022.00059

II.RELATED WORK

The majority of early sarcasm detection researches are on the textual side, and most methods rely on hand-labeling sentiment words. When there is no sarcasm word or the sarcasm word is not obvious, Aditya Joshi et al. proposed to use of machine learning algorithms and GLoVe word embedding to obtain the vector representation of words for sarcasm detection [1]. The use of contextual information by Akshay Khatri et al. has been shown to help better identify sarcasm [2]. N Babanejad et al. extended BERT's architecture by combining sentiment and contextual features and proposed two new sarcasm detection models, based on sentiment and contextual embeddings [3]. At present, it's not enough to rely solely on textual information for sarcasm detection, and other multimodal information provides an important supplement for sarcasm detection. Multi-modal data are richer than unimodal data because they describe objects from different angles, usually complimentary, overlapping or opposite in content.

Based on the BERT (Bidirectional Encoder Representation from Transformers), H Pan et al. proposed a multi-modal sarcasm detection model. The model is based on the self-attention mechanism, which creates cross-modal attention to capture cross-modal inconsistency and then combines intra-modal and inter-modal inconsistency information for prediction [4]. B Liang et al. proposed to determine the emotional inconsistency within a modality and between different modalities by determining the incongruity relationship between text and images [5]. Y Wu et al. proposed the incongruity-aware attention network (IWAN), which is a scoring mechanism that assigns greater weights to words with incongruity modality [6].

III.MULTI-HEAD SELF-ATTENTION MODEL AND CBAM

A.Text Feature Representation

In this paper, Bi-LSTM and MH-SA are used to construct a deep neural network to extract text features. Bi-LSTM is made up of two layers: a forward LSTM layer and a backward LSTM layer. The forward LSTM layer obtains historical information about the sequence, while the backward LSTM layer obtains future information about the sequence. The LSTM algorithm can combine current and past input information to update the hidden layer state without gradient vanishing, descending, or exploding problems [7].

In order to ensure that each word will be concerned by many factors to represent the complete meaning of the whole sentence. In this paper, multiple attention is used to concentrate the information of different subspaces in different locations. First, use GloVe to perform N-word embeddings, mapping each word into a vector space. The parameter $W_{k1}(W_{k1} \in R^{g\times 2p})$ is multiplied by the word encoded hidden state $H(H \in R^{N\times 2p})$ as input, and the output is passed to the tanh function to obtain Y. Y is multiplied by the parameter $W_{k2}(W_{k2} \in R^{g\times 2p})$ and sent to softmax for normalization, which involves calculating the weights of the normalized importance of different heads (q), yielding the weight vector Z.

When the hidden state H is multiplied by the weight vector Z, the sentence embedding matrix M is obtained. The calculation formula for this part is as follows.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(w_t, \overrightarrow{h_{t-1}}) \tag{1}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(w_t, \overleftarrow{h_{t-1}}) \tag{2}$$

$$H = [\vec{h} + \overleftarrow{h}] = (h_1, h_2, \cdots h_n) \tag{3}$$

$$Y = \tanh(W_{k1}H^T) \tag{4}$$

$$Z = soft\max(K_{k2}Y) \tag{5}$$

$$M = Z \otimes H \tag{6}$$

Where x_t, h_t, w_t denotes the input state, hidden state, and vector representation at the time t, and p is the size of $\vec{h}$ to $\overleftarrow{h}$. The two-layer network function parameters that the hidden state H passes through are W_{k1} and W_{k2}, and the number of hidden units is g. $\otimes$ denotes element-wise multiplication.

B.Image Feature Representation

This equation is an exception, which stipulates that ResNet is regarded as a classic image feature extraction algorithm. Adding CAM and SAM to the residual block structure of the ResNet50 algorithm can effectively improve the feature extraction capability of the network model. The input photo is first resized to 448x448, then divided into 14x14 region vectors and fed into the ResNet-CBAM model. The intermediate feature map $F(F \in R^{C\times H\times W})$ is traversed by CAM and SAM in turn. Figure 2 depicts the processing of CAM and SAM [8].

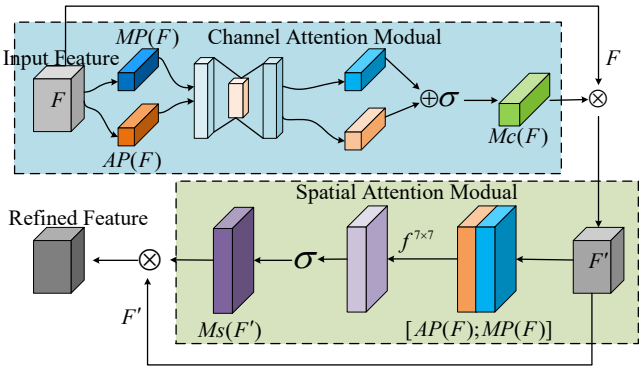

Figure 2. CBAM model (CAM+SAM)

CAM focuses on important parts of the image while ignoring irrelevant information. First, compress F into a one-dimensional vector in the Channel dimension, and then compute two matrices of all channel weights using Average Poling (AP) and Max Pooling (MP). The two-weight matrices are then fed into the same Multi-Layer Perceptron (MLP) to learn and optimize the weights. Finally, the output features are combined

using element-wise summation. This part of the calculation can be expressed as

$$\begin{aligned} Mc(F) &= \sigma(MLP(AP(F)) + MLP(MP(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \end{aligned} \tag{7}$$

$$F' = Mc(F) \otimes F \tag{8}$$

Where $Mc(Mc \in R^{C\times 1\times 1})$ denotes the attention extraction operation in the channel dimension, $W_0(W_0 \in R^{C/r\times C})$ and $W_1(W_1 \in R^{C\times C/r})$ denote the two-layer parameters in the multilayer perceptron model. F^c_{avg} and $F^c_{\max}$ denote the average pooling and maximum pooling operations, respectively. $\otimes$ denotes element-wise multiplication, σ the sigmoid function, and F' the features obtained after channel attention transfer.

SAM is an important complement of CAM in that it calculates spatial attention at the spatial level, allowing the network to notice which parts of F should have higher responses. To begin, use AP and MP to compress to obtain two 2D features. It is then convolved with a convolutional layer of kernel size $f^{7\times 7}$ to produce 2D spatial attention. The following formula expresses the mathematical processing of this section.

$$\begin{aligned} Ms(F) &= \sigma(f^{7\times 7}([AP(F); MP(F)])) \\ &= \sigma(f^{7\times 7}([F^s_{avg}; F^s_{\max}])) \end{aligned} \tag{9}$$

$$F'' = Mg(F') \otimes F' \tag{10}$$

Where $Ms(Ms \in R^{H\times W})$ represents the attention extraction operation in the spatial dimension, F^s_{avg} and $F^s_{\max}$ represents the average and maximum pooling operations, respectively, and F'' represents the vector obtained after passing through spatial attention.

C. Attribute Feature Representtation

Use image properties as a bridge between images and text. First predicted labels for 1000 image attributes trained with ResNet101 and COCO (image captioning dataset). Then, each Twitter image is processed by the ResNet-101 model, which predicts 5 image attributes $a_i(i = 1, 2, \cdots, 5)$,and uses GloVe for word embedding to obtain the image attribute vector $\mathrm{E}(\mathrm{a}_i)$. Then, $\mathrm{E}(\mathrm{a}_i)$ gets the attention weight A_i through a two-layer neural network. Finally, the attention weight A_i is multiplied by the image attribute vector $\mathrm{E}(\mathrm{a}_i)$ to obtain the vector representation of the image attribute. This part of the calculation can be expressed as.

$$A_i = W_2 \bullet \tanh(W_1 \bullet E(a_i) + b_1) + b_2 \tag{11}$$

$$A = soft\max(A_i) \tag{12}$$

$$V_A = \sum_{i=1}^{5} A_i \bullet E(\mathrm{a}_i) \tag{13}$$

Where W_1 and W_2 represent weights, b_1 and b_2 represent bias.

The vector V_A of the five image attributes is used as the initial state of Bi-LSTM after nonlinear transformation, which can promote Bi-LSTM's understanding of text modalities and improve the representation ability of the model.

D. Modality Fusion

We computed the raw feature vectors for text, image, and image attributes in the preceding sections. The weighted average of the reconstructed eigenvector V_m is calculated using a two-layer neural network and nonlinear variation. Set V_m to a fixed-length V'_m, then use a double-layer feedforward neural network to calculate the attention weight $\tilde{A}_m$ of each model m, and then use it as the weighted average of the transformed feature vector V'_m, yielding a fixed-length vector V_{fused}. This part of the calculation can be expressed as.

$$\widetilde{A}_m = W_{m2} \bullet \tanh(W_{m1} \bullet V_m + b_{m1}) + b_{m2} \tag{14}$$

$$\widetilde{A} = soft\max(\widetilde{A}_m) \tag{15}$$

$$V'_m = \tanh(W_{m3} \bullet V_m + b_{m3}) \tag{16}$$

$$V_{fused} = \sum_{m\in\{text, image, attribute\}} \widetilde{A}_m V'_m \tag{17}$$

Where W_{m1}, W_{m2}, W_{m3} represent weights, b_{m1}, b_{m2}, b_{m3} represent bias.

E. Classification Layer

The two-layer fully connected neural network is used in the classification layer. The activation functions of hidden layer and output layer are the ReLu function and the sigmoid function, respectively. The loss function is the cross entropy, and the optimizer is Adam.

IV. Dataset And Preprocessing

To evaluate the CMHA multimodal sarcasm detection model proposed in this paper, we use the public dataset Twitter. Y Cai et al. released the Hierarchical Fusion Model (HFM)and Twitter datasets in 2019 ACL (Annual Meeting of the Association for Computational Linguistics).

Table 1. Conflguration of our dataset

	Training	Development	Test
Positive	8642	959	959
Negative	11174	1451	1450

The Twitter dataset removes emotional words like sarcasm, sarcastic, irony, and ironic, as well as tweets that do not contain URL links. Using sarcasm as a positive example and non-sarcasm as a negative example. Use NLTK Toolkit to separate words, emojis, and tags. Tags are separated from each other by the label symbol <#>, and uppercase letters are replaced with lowercase letters. Replace emotions mentioned in tweets with the symbol <user>, and words that do not appear or appear only once in the training set with the symbol <unk>. The data is divided into a training set, a development set, and a test set, with the ratios being 80%, 10%, and 10%, respectively [9]. Table 1 depicts the data set's statistics.

V.EXPERIMENTAL RESULTS

Figure 3 shows the comparison results between the proposed model and the baseline model. Under the premise of using the same Twitter data, F1 score, accuracy, precision, and recall taken as the important basis to measure the sarcasm detection results of each model, in order to reflect the effectiveness of the model improvement proposed in this paper.

Attribute, Image and **Text (Bi-LSTM).** Use information from a single modality for sarcasm detection. A single modal vector representation is obtained and input to a two-layer fully connected neural network for prediction.

HFM. (Hierarchical Fusion Model）Cai et al. proposed the HFM model in the 2019 ACL, proposed the HFM model in the paper and released the Twitter dataset used in this paper.

BERT. (Bidirectional Encoder Representations from Transformers) Pan et al. proposed the improved BERT model proposed by EMNLP in 2020. In this paper, it is proposed to detect multimodal sarcasm by using Intra and Inter-modality Incongruity.

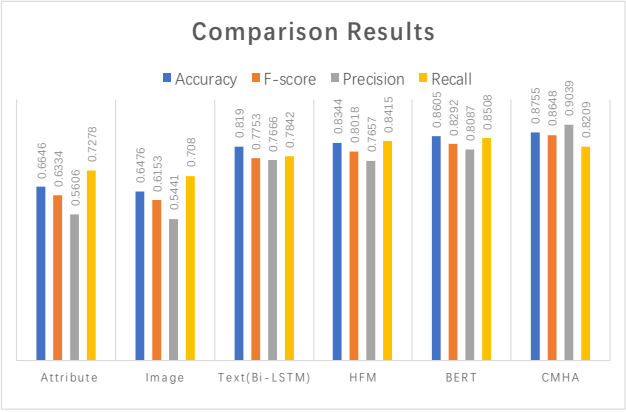

Figure3. Baseline Models Comparison Results

As shown in Figure 3, as a result of sarcasm detection using a single modality (text, image, or image attribute) as input, image modality and image attribute modality are average. The text modality is the best, with accuracy and F1 score value of 81.9% and 77.53%, respectively. The experimental results of the three modalities fully demonstrate the importance of text modalities for multimodal sarcasm detection. The CMHA model proposed in this paper is based on the structure of ResNet and Bi-LSTM, and the ability to obtain feature vectors is greatly improved by adding CBAM model and MH-SA model. Through continuous training and optimization of the CMHA model, the accuracy of test set can reach 87.55%, and the F1 score is 86.48%, which is 4.11% higher than that of the 2019 HFM model, and the F1 score is 6.3% higher. Compared with the 2020 BERT model, the accuracy of the CMHA model is increased by 1.5%, and the F1 score is increased by 3.56%. The experimental results fully demonstrate the effectiveness of the CMHA model.

VI.CONCLUSION AND FUTURE WORK

In this paper, Multi-Head Attention and CBAM models are added to the structure of the hierarchical fusion model, which greatly improves the prediction accuracy of the model. By comparing the accuracy of experimental results with other data, the effectiveness of our model improvement is proved. In the process of using the ResNet101 model to predict attribute labels, due to factors such as the large amount of calculation of image features and the small number of attribute labels, there are problems such as low image attribute prediction accuracy, which needs to be further strengthened in the later work. The way to extract attribute labels from images can be understood as converting image modality into text modality, and many papers have similar applications, such as image captions, video descriptions, text-to-image synthesis, and so on. The generated network GAN is used to enhance the data, expand the total number of attribute labels, increase the predicted number of image attributes, and express the image feature vectors more completely. My next project will continue to study how to use the fusion of text features and image attribute features for prediction.

ACKNOWLEDGMENTS

This paper is funded and supported by the Doctoral Research Foundation project of Shandong Jiaotong University (BS201902027).

REFERENCES

[1] Joshi A, Tripathi V, Patel K, et al. Are word embedding-based features useful for sarcasm detection?[J]. arXiv preprint arXiv:1610.00883, 2016.

[2] Khatri A. Sarcasm detection in tweets with BERT and GloVe embeddings[J]. arXiv preprint arXiv:2006.11512, 2020.

[3] Babanejad N, Davoudi H, An A, et al. Affective and contextual embedding for sarcasm detection[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 225-243.

[4] Pan H, Lin Z, Fu P, et al. Modeling Intra and inter-modality incongruity for multi-modal sarcasm detection[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1383-1392.

[5] Liang B, Lou C, Li X, et al. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4707-4715.

[6] Wu Y, Zhao Y, Lu X, et al. Modeling incongruity between modalities for multimodal sarcasm detection[J]. IEEE MultiMedia, 2021, 28(2): 86-95.

[7] Diao Y, Lin H, Yang L, et al. A multi-dimension question answering network for sarcasm detection[J]. IEEE Access, 2020, 8: 135152-135161.

[8] Clausen H, Grov G, Aspinall D. CBAM: A Contextual Model for Network Anomaly Detection[J]. Computers, 2021, 10(6): 79.

[9] Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in Twitter with hierarchical fusion model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2506-2515.