Yet at Memotion 2.0 2022 : Hate Speech Detection Combining BiLSTM and Fully Connected Layers

Yan Zhuang[1], Yanru Zhang*[1,2]

[1]University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, UESTC

Abstract

People are increasingly willing to send emojis rather than plain text to express themselves. However, emoticons with bad feelings such as hatred are more subtle and harder to detect than text. The 'Memotion 2.0' task of the 'DE-FACTIFY' workshop aims to detect the hate speech in internet memes. In this paper, we present our approach for solving this task. The GloVe is applied to get the text embedding matrix and pre-trained VGG-16 is used to get the image representation. Instead of attention mechanism, we combine the bidirectional Long Short Term Memory (LSTM) and fully connected layers to promote the feature interactions. Compared to text-only models, our model showed better performance and helped us increase the baseline (43.4%) by 7.4%, which finally ranked 3rd in 'Sentiment Analysis' task.

Keywords

Multimodality, Hate Speech Detection, Meme Classification

1. Introduction

The development of social media has made it possible for people to express their feelings and opinions on things whenever and wherever they want. Internet memes are popular among internet users because of their rich content of images and complementary text descriptions and they are often based on experiences or cultures [1], or on popular trends, and therefore have more social and communication properties than pure text. However, online speech is devoid of real-time monitoring, which is why hate speech runs rampant in online media. And internet memes, especially text embedded within images, make it more difficult to detect what it is trying to convey as opposed to texts [2].

In this paper, we propose a multimodal model using text and images as input, where pre-trained GloVe and VGG-16 are used to obtain text and image features respectively [3, 4]. The model also contains bidirectional Long Short Term Memory (LSTM) layers to enhance the representation of textual context [5], and multiple fully connected layers to facilitate the interaction between the stitched multimodal features, which are finally classified by softmax. The experiment results showed that our model outperformed BERT and had good effectiveness [6], which helped us rank 3rd in 'Sentiment Analysis' task.

* *Corresponding author*

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

delecisz@gmail.com (Y. Zhuang); yanruzhang@uestc.edu.cn (Y. Zhang*)

The rest of the paper is organised as follows. In section 2, related works about meme classification is provided. The introduction of the task is detailed in section 3. Followed by the structure of our proposed models in section 4. The experiments results and analysis of different models are discussed in section 5. Finally, section 6 concludes the paper along with future directions.

2. Background

Lots of efforts have been put into internet memes classification, such as in Memotion Analysis task of SemEval-2020 [1], both text-only models and image-text models are proposed.

The text-only models show competitive effectiveness compared to the multimodal model. It shows that text-only model using Feed Forward Neural Network (FFNN) with Word2vec embeddings performs superior to BERT and image-text models like multimodal bitransformers (MMBT) in [7], while MMBT combines the ResNet-152 and BERT through mapping the image embedding into the text space [8, 9]. While the multi kernel convolution layers, bidirectional LSTM as well as attention mechanism have been adopted to obtain the higher-level features and long-term dependencies in [10], which also shows good performance.

The multimodal models are no less impressive and demonstrate good performance. There are various ways of fusing different modal features, such as ensembling the predictions of different modal model, and extracting multimodal features with different models and then fusing them before processing. [11] takes the former approach, where text-based models, such as XLNet [12], and image-based models, such as SENet [13], are used separately for training and prediction, and eventually the predictions from all models are ensembled. [14] takes another approach that it uses ALBERT [15] and VGG16 as feature extractors to extract features from texts and images respectively, and the features are concatenated together and put into two fully-connected layers. The overall process and idea of [16] is similar to the former work, the difference being in the feature extractors and the processing of the fused features. The latter takes GloVe to extract the text features and uses inception network to extract the image features, the fused features are then processed through attention, bi-LSTM and GRU layers [17].

There are also some recent models that focus more on the deep interaction of different modal features [18, 19], which may provide a new idea for a better solution to this problem.

3. Task setup

Memotion 2.0 is the second version of Memotion task, which is detailed in [1]. It contains 10k images and the corresponding OCRs, and the dataset is divided into the training set, validation set and test set at the ratio of 7:1.5:1.5 [20, 21]. Based on the images and OCRs, the Memotion 2.0 aims to solve the following three sub-tasks, 'Sentiment Analysis', 'Emotion Classification' and 'Scales of Emotion Classes'. Examples of the Memotion 2.0 dataset can be seen in Figure 3 And the label distributions (%) of each task can be seen in Table 1, Table 2 and Table 3. The details of the three sub-tasks are as follows:

- **Task A: Sentiment Analysis**: The task aims to classify the internet meme into three categories based on the expressed emotion.

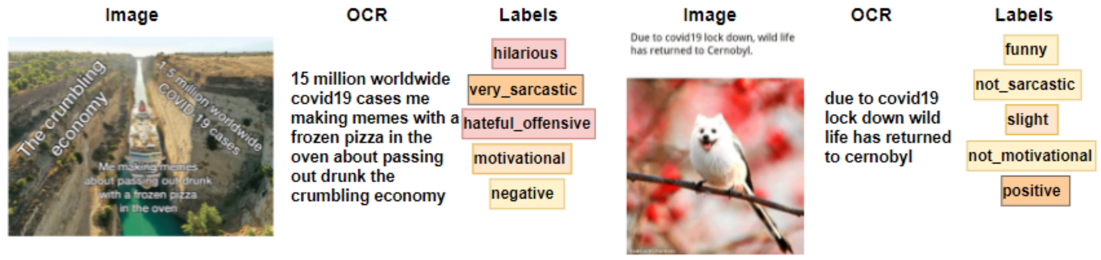
Image
OCR
Labels
The crumbling economy
1.5 million worldwide COVID 19 cases
Me making memes about passing out drunk with a frozen pizza in the oven
15 million worldwide covid19 cases me making memes with a frozen pizza in the oven about passing out drunk the crumbling economy
hilarious
very_sarcastic
hateful_offensive
motivational
negative
Image
OCR
Labels
Due to covid19 lock down, wild life has returned to Cernobyl.
due to covid19 lock down wild life has returned to cernobyl
funny
not_sarcastic
slight
not_motivational
positive

Figure 1: Each sample of the Memotion 2.0 dataset contains an image and the corresponding OCR. Each box in the 'Labels' column represents a classification task. The darker of the box color, the greater the degree of emotion. The first 4 classification tasks are the sub-tasks of the 'Scales of Emotion Classes' task, while the left one is of the 'Sentiment Analysis'.

- **Task B: Emotion Classification**:The images are classified into two categories in this task based on whether the images are humorous, sarcastic, offensive and motivational.
- **Task C: Scales of Emotion Classes**: Based on the intensity of different emotions the meme contains, the internet memes are classified into 4 classes.

Table 1
The label distributions (%) of Task A.

Sentiment Analysis	Train	Negative and Very Negative	0.139
		Neutral	0.644
		Positive and Very Positive	0.217
	Valid	Negative and Very Negative	0.133
		Neutral	0.65
		Positive and Very Positive	0.217
	Test	Negative and Very Negative	0.307
		Neutral	0.647
		Positive and Very Positive	0.052

Table 2
The label distributions (%) of Task B.

			Humorous	Sarcastic	Offensive	Motivational
Emotion Classification	Train	Negative	0.131	0.553	0.74	0.956
		Positive	0.867	0.447	0.259	0.043
	Valid	Negative	0.153	0.536	0.74	0.953
		Positive	0.847	0.464	0.26	0.046
	Test	Negative	0.041	0.123	0.629	0.987
		Positive	0.959	0.877	0.371	0.013

Table 3
The label distributions (%) of Task C.

			Humour	Sarcasm	Offense	Motivation
Scales of Emotion Classes	Train	Not	0.131	0.553	0.74	0.956
		Little	0.524	0.251	0.158	0.043
		Very	0.266	0.153	0.076	——
		Extremely	0.079	0.043	0.026	——
	Valid	Not	0.153	0.536	0.74	0.953
		Little	0.497	0.259	0.159	0.046
		Very	0.279	0.164	0.071	——
		Extremely	0.071	0.041	0.03	——
	Test	Not	0.041	0.123	0.629	0.987
		Little	0.595	0.165	0.305	0.013
		Very	0.265	0.595	0.058	——
		Extremely	0.099	0.117	0.009	——

From the above tables, the label distributions are not balanced especially in task 'Emotion Classification'. Almost all the images are motivational and very few are not humorous. The distribution of each label is relatively consistent on the training and validation sets, while on the test set, there are large deviations in the distribution of some labels, like 'Sarcastic' in task B and 'Sarcasm' in task C.

4. Model

The model can be divided into unimodal models and bimodal models. The former only uses the text or the images while the latter uses both. Both the images and the texts are adapted in our model.

4.1. The unimodal models

Unimodal models regard the 'Sentiment Analysis' task as the multi-class classification task, and the rest as multi-label classification task. In fact, the 'Emotion Classification' and 'Scales of Emotion Classes' task can also be seen as several sub-multi-class classification tasks. The unimodal models include text-only models and image-only models. And both models always extract the features from the pre-trained models, such as BERT, RoBERTa for texts and VGG-16, ResNet-50 for images, and then put the features into different neural networks for classification.

4.2. The bimodal models

The bimodal models often use different strategies to fuse the features obtained separately from the unimodal models, like concatenation, adding and using attention mechanism. Here we use the concatenation, and the structure can be seen in Figure 4.2.

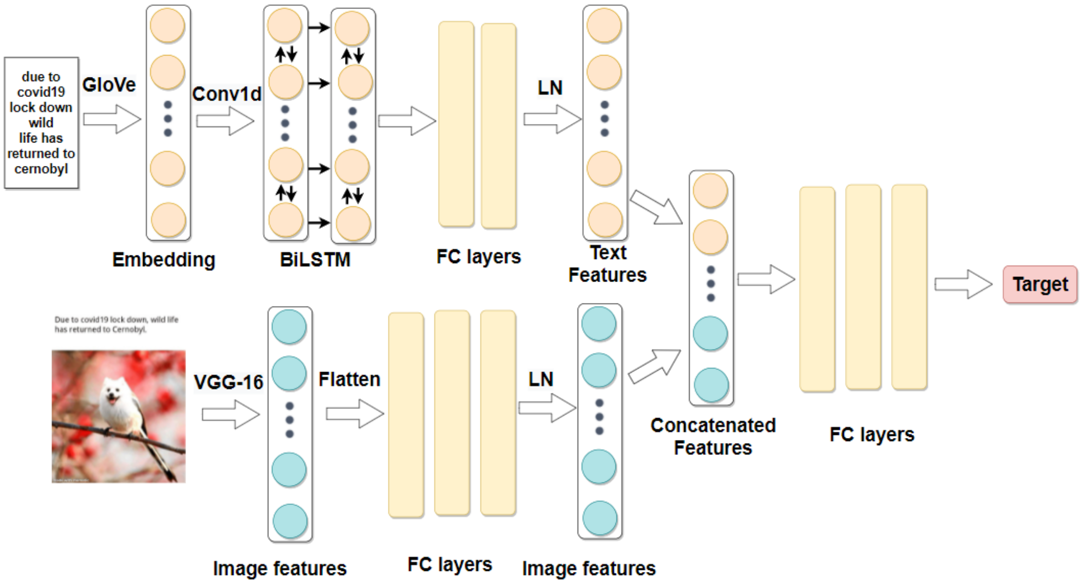

Figure 2: Network structure of the proposed model. The GloVe is used to extract the word embedding, then the embedding is converted and fed into the BiLSTM. After layer normalization, the final text features come out. The image features are first extracted from the pre-trained VGG-16, then they are faltened and put into three fully connected layers and a normalization layer. The final outputs are concatenated with the text features and are then put into three fully connected layers for classification. The 'FC layers' and 'LN' denotes Fully Connected layers and 'Layer Normalization' respectively.

For OCRs, GloVe is applied to get the embedding matrix. Then the embedding matrix is processed by three convolution layers, of which the filters are all 2 while the kernel size are separately 4, 3, 2. After that, the new output are fed into the two BiLSTM layers and corresponding layer normalization layers to better catch the semantic information of the whole sentence. The newly obtained features are passed through two fully connected layers to make the features more fully accessible to different information, followed by a normalisation layer to prevent overfitting of the model and to obtain the final 256-dimensional text features.

Compared to the acquisition of textual features, the acquisition of image features is a bit simpler. We first extract image features using pre-trained VGG-16, then falten it as well as use three fully-connected layers and a normalised layer to get the final 256-dimensional image features. Once the text and image features are concatenated together and put into three fully connected layers, the final prediction comes out.

In order to prevent overfitting, here we use the stochastic gradient descent optimizer with the learning rate of 2e-5 and weight decay of 1e-6. Besides, each fully connected layer is followed by a dropout layer with dropout rate of 0.1.

5. Experiments and evaluations

Both unimodal and bimodal models are applied in our experiments. We choose the BERT as the text-only model, of which the hyperparameters are all default and choose our proposed model as the image-text model. The overall performance of these two models on different tasks can be seen in Table 4. Our proposed model helped us increase the baseline (43.4%) by 7.4% and ranked 3rd in the 'Sentiment Analysis' task.

As the OCR texts extracted from images are usually not complete sentences, features from the images are needed to complement them. This lack of information makes text models like BERT lose their discriminative effect. In the course of our experiments, we found that neither the adoption of data augmentation nor the adoption of some balanced samples could improve the performance of the model. In addition to this, when predictions were made for almost all tasks, the obtained labels were extremely unbalanced, i.e. the predicted values were basically concentrated on only two labels regardless of the training, and 90% of them were concentrated on one of the categories.

Our proposed model, on the other hand, relatively alleviates the above problem by enabling features of different modalities to interact through multiple fully connected layers, but the results are still less promising for some tasks. Just as shown in Table 5 and Table 6, compared with BERT, our model show competitive performance or even better performance in all tasks. Especially in the prediction of 'Sarcastic' label in 'Emotion Classification' task, our model improves performance by almost 20%.

Apart from this, the performance of our model on task B and task C is not outstanding, and is even a little worse than baseline. This is partly due to the lack of applicability of our model, as it is proposed and optimised only for task A. Secondly, due to the large number of fully connected layers in our model, the errors and noise introduced by the convolution of the embedding matrix and the faltten process of the image features are amplified by the propagation of multiple fully connected layers, which has an impact on the prediction. Finally, because the data distribution of some labels is not consistent between the training and test sets, the model loses its discriminative ability.

Acquiring the features of text and images in the same representation space and performing the same processing together may reduce the variability due to different models, while making better use of the complementary roles of text and images, for example through better interaction strategies and mechanisms, would enable better resolution of these tasks.

Table 4
The overall performance of different models. The metric is the weighted F1 score.

	Sentiment Analysis	Emotion Classification	Scales of Emotion Classes
Ours	**0.5089**	0.6579	0.51
BERT	0.5037	0.6106	0.484
Baseline	0.434	**0.7358**	**0.5105**

Table 5
The performance (F1 score) of different models on Task A and Task B.

	Sentiment Analysis	Emotion Classification			
	Sentiment	Humorous	Sarcastic	Offensive	Motivational
Ours	0.5089	0.9405	0.2239	0.4891	0.98
BERT	0.5037	0.9384	0.0386	0.4853	0.98

Table 6
The performance (F1 score) of different models on Task C.

	Scales of Emotion Classes			
	Humour	Sarcasm	Offense	Motivation
Ours	0.4686	0.1231	0.4961	0.98
BERT	0.4435	0.0271	0.4853	0.98

6. Conclusion

In this paper, we propose a text-image model based on bidirectional Long Short Term Memory (LSTM) and fully connected layers to solve the 'Sentiment Analysis' task in De-Factify workshop. The major challenge of this task derives from alignment and supplementation of incomplete OCR sentences and images, and existing approaches showed unsatisfactory performance. To address this problem, we used GloVe and VGG-16 to acquire text and image features respectively, followed by multiple LSTM layers and fully connected layers for feature interaction. The final model outperformed BERT and showed competitive performance, which helped us stand 3rd in 'Sentiment Analysis' task. Unified representation of text and image features in the same feature space and better multi-modal feature fusion and interaction strategies are conducive to the better solving this challenge.

References

[1] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, B. Gamback, Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor!, arXiv preprint arXiv:2008.03781 (2020).

[2] A. Williams, C. Oliver, K. Aumer, C. Meyers, Racial microaggressions and perceptions of internet memes, Computers in Human Behavior 63 (2016) 424–432.

[3] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[4] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[5] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm

network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, 2015, pp. 802–810.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] V. Keswani, S. Singh, S. Agarwal, A. Modi, Iitk at semeval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1135–1140.

[8] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, D. Testuggine, Supervised multimodal bitransformers for classifying images and text, arXiv preprint arXiv:1909.02950 (2019).

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] A. N. Chy, U. A. Siddiqua, M. Aono, Csecu_kde_ma at semeval-2020 task 8: A neural attention model for memotion analysis, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1106–1111.

[11] T. Morishita, G. Morio, S. Horiguchi, H. Ozaki, T. Miyoshi, Hitachi at semeval-2020 task 8: Simple but effective modality ensemble for meme emotion recognition, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1126–1134.

[12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[13] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[14] G.-A. Vlad, G.-E. Zaharia, D.-C. Cercel, C.-G. Chiru, S. Trausan-Matu, Upb at semeval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis, arXiv preprint arXiv:2009.02779 (2020).

[15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).

[16] M. Sharma, I. Kandasamy, W. Vasantha, Memebusters at semeval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1163–1171.

[17] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[18] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, arXiv preprint arXiv:2102.03334 (2021).

[19] W. Wang, H. Bao, L. Dong, F. Wei, Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, arXiv preprint arXiv:2111.02358 (2021).

[20] S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Memotion 2: Dataset on sentiment and emotion analysis of memes, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.

[21] P. Patwa, S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, A. Das,

T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Findings of memotion 2: Sentiment and emotion analysis of memes, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.