BROWALLIA at Memotion 2.0 2022 : Multimodal Memotion Analysis with Modified OGB Strategies

Baishan Duan[1], Yuesheng Zhu[1]

[1]*Peking University*

Abstract

Emotion analysis with social media is important for many social psychology tasks such as hatespeech detection. Internet memes comprises of various modalities including textual, visual and audio, multimodal information computational processing needs superior methods. Recently, Memotion 2.0 attracted widespread attention. In this paper, we propose a novel multimodal system to analyze textual- visual pair information. We choose LSTM and ResNet50 as backbone, then employee late-fusion to fusion two modalities. In addition, to train network well, we adopt modified offline-gradient-blending strategy to alleviate overfitting. Our approach achieves good performance in three sub-tasks, ranking 2^{nd} in Subtask A, 3^{rd} in Subtask B, 2^{nd} in Subtask C.

1. Introduction

Internet media has taken a large part in human life, and internet media platforms such as facebook, twitter, and weibo contain a large amount of information, which mostly consists of images and text. People like to express their emotions on social media, which may be humorous or contain sarcasm, may be offensive, harmful to individuals, governments, organizations, and races, or motivational. Sentiment analysis of social media can be a good solution to these problems.

Common text sentiment classification has been relatively improved, such as RNN, LSTM[1], BERT[2] et al. Multimodal sentiment classification faces a big challenge. The MEmotion task collects 10K annotated memes online, including images and text information obtained from image OCR. This data format will make the sentiment information mining more difficult, the text in the images will generate noise to the image content, and the OCR results are not guaranteed to be exactly the same as in image. In addition, the relationship between some picture contents and textual expressions is not very close, which brings trouble to the relationship mining between modalities. In addition to the semantic information of images, imbalanced dataset also brings challenges to the final results.

There are three subtasks in Memotion 2.0[3, 4]. Subtask A is overall sentiment (positive, negative, and neutral) analysis of memes, Subtask B is binary emotion classification from humor, sarcasm, offensive, and motivational. Subtask B is extended into Subtask C, which is Fine-grained sentiment classification for four categories. We consider Subtask A as the pretraining task and Subtask B and C as downsteam tasks.

baishanduan@stu.pku.edu.cn (B. Duan); zhuys@pku.edu.cn (Y. Zhu)

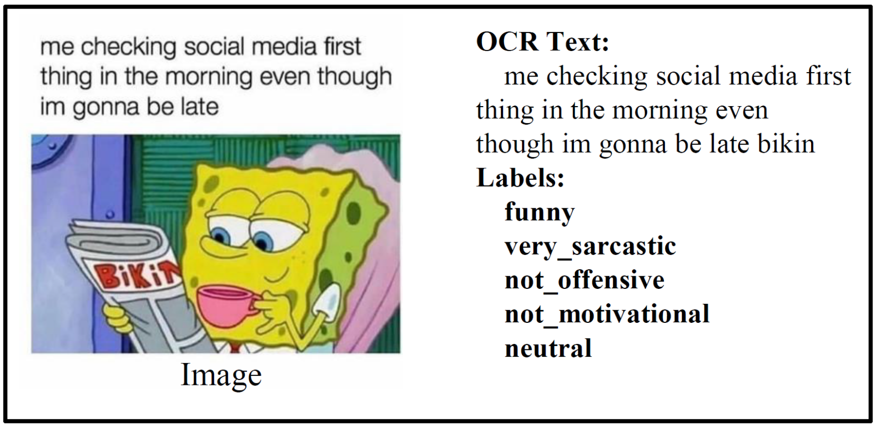
me checking social media first thing in the morning even though im gonna be late
BiKiN
Image
OCR Text:
me checking social media first thing in the morning even though im gonna be late bikin
Labels:
funny
very_sarcastic
not_offensive
not_motivational
neutral

Figure 1: An example of Memotion 2.0 dataset(Image, OCR Text, 5 labels)

In this paper, we propose a novel multimodal system to analyze textual- visual pair information. We use ResNet-50[5] and LSTM to be backbone and late fusion features extracted from backbone. We additionally adopt modified offline-gradient-blending(OGB)[6] to alleviate overfitting. Our contributions are as follows:

- We propose a multimodal information fusion network, which can extract text and image features well.
- We use modified OGB strategy to train multimodal networks, which alleviate overfitting.
- Experiments show that our model achieves good results in all three subtasks.

2. Background

Analysis of Memotion 2.0 is a part of sentiment analysis.In the field of text-only sentiment classification, there are many excellent methods, such as RNN, LSTM, Transformer[7] and BERT. Attention mechanisms are proposed to pay better attention to the contextual information of the text, and BERT with variants like RoBerta[8] and XLNet[9] are the state-of-the-art methods in text-only tasks . In the field of image classification, ResNet is the most used backbone, which is pretrained on ImageNet[10]. However, there is still much to be mined in multimodal sentiment classification. TFN[11] uses inner product to fuse the features of three modalities. LMF utilizes a low-rank matrix decomposition of the weights, where each mode is first individually linearly transformed before multi-dimensional dot product, which can be viewed as a sum of the results of multiple low-rank vectors, thus reducing the number of parameters in the model. MulT[12] attends to interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another.

In our work, considering the small dataset, we prefer LSTM to transformer-based models as textual backbone. ResNet is used to extract image features. Our architecture is light but effective.

3. Method

3.1. Overview

We proposed a multimodal network, based on pretrained ResNet-50 and LSTM. We regard TaskA as pretraining task and then finetune it to complete TaskB and TaskC. We also refer to uni-model architecture and compare the result to multi-model result. As a result, we found that multi-model can extract features well.

3.2. Uni-model

We experiment with text-only and image-only models for Memotion classification. The result analysis is justified in subsection 4.2.1

3.2.1. Visual

To extract image feature, we opted for the ResNet-50 architecture, which is the most common visual backbone. The residual block can avoid network degradation and alleviate gradient vanishing problems. In the experiment, we use pretrained ResNet-50 from timm[1].

3.2.2. Textual

To extract textual features, we refer to two basic language modeling network LSTM and BERT. LSTM is the variant of RNN, it is mainly to solve the gradient disappearance and gradient explosion during long sequence training. Bert is the variant of Transformer proposed by Google research. Through training on huge corpora, Bert achieved state of the art result on many NLP tasks. For LSTM, we realize it by pytorch, for Bert, we adopted the pretrained version by HuggingFace[2].

3.3. Multi-model

In order to better mining relationship between image and text, we combined two single feature extraction components into a unified network. For Multi-model design, we just add late fusion layer to fusion features extracted from visual and textual backbone. Note that we remove classification layer in ResNet-50 to obtain only convolutions used in extracting the features from the input image . In addition, we adopted modified offline-gradient-blending strategy to alleviate overfitting. For TaskB and TaskC we just add four FC layer based on pretrained TaskA multi-modal network.

3.4. OGB

We add offline-gradient-blending to alleviate overfitting. There are three branches in our multi-modal network, and overfitting of each branch will happen at different time as well as the

[1] https://github.com/rwightman/pytorch-image-models

[2] https://github.com/huggingface/transformers

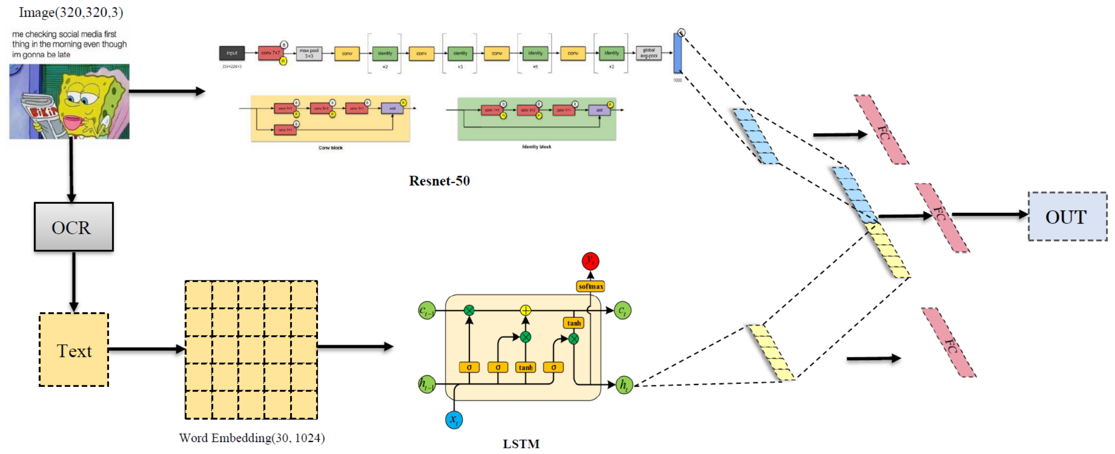

Figure 2: Overview of multimodal network, including three branches representing fusion branch, visual branch and textual branch. Test using only the output of fusion branch.

overffiting ratio is also different. Overfiting-to-Generalization Ratio is proposed to measure overfitting as follows

$$OGR = |\frac{\Delta O_{N,n}}{\Delta G_{N,n}}| = |\frac{O_{N,n} - O_N}{L_N^* - L_{N+n}^*}| \tag{1}$$

Where L^* indicates the validation loss of N epoch, O indicates the gap between validation loss and train loss.

Then compute the weight of each branch loss:

$$\{w_i^*\}_{i=1}^{k+1} = \frac{1}{Z}\frac{G^i}{{O^i}^2} \tag{2}$$

3.5. Objective Function

We sum three branch losses with the weight computed by OGB:

$$L_{blend} = \sum_{i=1}^{k+1} w_i L_i \tag{3}$$

4. Experiment

4.1. Experiment setup

4.1.1. Dataset and preprocessing

Our experiments are conducted on the official dataset of Memotion 2.0. The training dataset consists of 7k human annotated internet memes. Each sample is composed of image and text extracted from image by OCR as shown in Fig 1. Besides, the class distribution is imbalanced shown in Table 1, thus we adopt focal loss in our experiments.

For image we use transformers such as random flip and random crop, finally the input image is resized to 320x320. For text, we padding it to the fixed length 30, and tokenize it by BertTokenizer additionally in Bert Model.

Table 1
Distribution of data in Subtask A

positive	negative	neutual
1517	973	4510

4.1.2. Implementation

We conducted our model by pytorch 1.7.1, we use ResNet-50 from timm and Bert-base-uncased from Huggingface. We set the hyper-parameters manually, the batch-size, epoch, learning rate, weight decay are 32, 40, 1e-5, 5e-3. The Optimizer is Adam with linear lr decay. We train all models on a single 2080Ti GPU on Ubuntu system.

4.2. Result

4.2.1. Uni-model and Multi-model

Table 2 contains the comparison between uni-model and multi-model. As expected, multi-model performance is better than uni-model. The difference between LSTM and BERT is not remarkable. We conjecture this is due to the small datasets. So in the rest of our experiment, we choose LSTM to be textual backbone because of limited computation resource.

Table 2
Result of uni-model and multi-model

Modality	Model	Marcro-F1
Text-Only	LSTM	0.2915
	BERT	0.2920
Image-Only	ResNet-50	0.3247
Image-Text	LSTM-ResNet	**0.3344**

4.2.2. Modified Online Gradient Blending

Table 3 contains the result of ogb strategy training. At first, we compute the weight by Eq .2, the weight for fusion branch, visual branch and textual branch weight are 0.22, 0.17, 0.61. The result shows worse performance than 1: 1: 1. We visualize the loss as shown in Fig 3.

According to Eq .2. G represents the degree to which the validation loss decreases. The greater the G, the greater the weight, but in this data set, the validation loss has been increasing. Because of the long-tailed problem, the accuracy has been decreasing, so this formula is not

Table 3
Result of different OGB weight

Method	OGB(f : v: l)			Macro-F1
LSTM-ResNet-50	1	1	1	0.3450
	0.22	0.17	0.61	0.3261
	0.50	0.30	0.20	0.3533
LSTM-ResNet-50 (focal loss)	0.50	0.30	0.20	**0.3649**

suitable. What we need to do is to suppress the growth of validation loss and decrease the gap between validation loss and train loss. After the analysis, We tried to modify the formula to

$$\{w_i*\}_{i=1}^{k+1} = \frac{1}{Z}\frac{1}{{O^{i}}^{2}G^{i}} \tag{4}$$

which means that the weight is inversely proportional to the degree of increase in the validation loss, and the calculated weights are 0.5, 0.3, 0.2, and have relatively good results. Additionally, for the imbalanced dataset, we use focal loss and get the best F1 score in validation set.

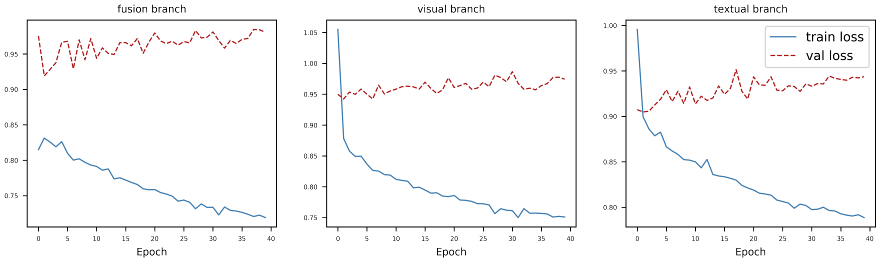

Figure 3: validation loss(red) and train loss(blue) of three branches

4.2.3. Final Result

Result is shown in Table 4 , ⋆ indicates our model. Our method gets decent performance compared with baseline and ranking2nd in Subtask A, 3rd in Subtask B, 2nd in Subtask C.

Table 4
Leaderboard of Memotion 2.0

Ranking	Subtask A	Subtask B	Subtask C
1	**0.5318**	**0.8299**	**0.5564**
2	0.5255⋆	0.8059	0.5453⋆
3	0.5088	0.767⋆	0.5443
4	0.5081	0.7690	0.5301
baseline	0.434	0.7358	0.5105

5. Conclusion

In this paper, we proposed a multimodal architecture with modified OGB strategy for Memotion 2.0 tasks. Experiments show that our methods outperform strong baseline in three subtasks. In addition our score ranks high in the test sets.

For future work, we will try more powerful backbones like ResNext[13], DenseNet[14] in visual and RoBERTa, XLNet in textual. Strong backbone allows for better extraction of modal features. For multimodal architecture, we will try joint end-to-end model such as VLBert[15], VisualBert[16]. At the same time, we find that imbalenced dataset cause low score in subtask B and subtask C, maybe some strategy to solve long-tailed problem make sence such as loss function[17, 18] and new architecture like BBN[19].

References

[1] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[2] 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).

[3] 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.

[4] 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.

[5] 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.

[6] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv (2017).

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach (2019).

[9] 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).

[10] D. Jia, D. Wei, R. Socher, L. J. Li, L. Kai, F. F. Li, Imagenet: A large-scale hierarchical image database, Proc of IEEE Computer Vision Pattern Recognition (2009) 248–255.

[11] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. P. Morency, Tensor fusion network for multimodal sentiment analysis, arXiv (2017).

[12] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2019, NIH Public Access, 2019, p. 6558.

[13] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, CoRR abs/1611.05431 (2016). URL: http://arxiv.org/abs/1611.05431. arXiv:1611.05431.

[14] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, CoRR abs/1608.06993 (2016). URL: http://arxiv.org/abs/1608.06993. arXiv:1608.06993.

[15] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: pre-training of generic visual-linguistic representations, CoRR abs/1908.08530 (2019). URL: http://arxiv.org/abs/1908.08530. arXiv:1908.08530.

[16] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).

[17] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9268–9277.

[18] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, arXiv preprint arXiv:1906.07413 (2019).

[19] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9719–9728.