CN-HIT-MI.T at SemEval-2020 Task 8: Memotion Analysis Based on BERT

Zhen Li, Yaojie Zhang
Harbin Institute of Technology
linklizhen@163.com
yjzhang@hit-mtlab.net

Bing Xu, Tiejun Zhao
Harbin Institute of Technology
hitxb@hit.edu.cn
tjzhao@hit.edu.cn

Abstract

Internet memes emotion recognition is focused by many researchers. In this paper, we adopt BERT and ResNet for evaluation of detecting the emotions of Internet memes. We focus on solving the problem of data imbalance and data contains noise. We use RandAugment to enhance the data of the picture, and use Training Signal Annealing (TSA) to solve the impact of the imbalance of the label. At the same time, a new loss function is designed to ensure that the model is not affected by input noise which will improve the robustness of the model. We participated in sub-task a and our model based on BERT obtains 34.58% macro F1 score, ranking 10/32.

1 Introduction

Memes are from people's culture or some social activities in daily life, usually composed of one or two forms of image, video, gif, and text (Park, 2020). Memes are active in people's social media, but with the number of memes increasing, offensive memes are also increasing (Williams et al., 2016). So for many social media companies, how to identify the type of meme is very important. Using machine learning to identify the type of meme is a very important solution, and has shown promising Performance.

In SemEval-2020 Task 8: Memotion Analysis (Sharma et al., 2020), the organizer collecting the Internet memes and then using OCR recognition it and correct manually the text on the pictures. The task is divided into three sub-tasks: a) sub-task a detects the emotions of the Internet memes, which are divided into positive, negative and neutral; b) sub-task b detects the types of humor expressed by Internet memes, which are divided into sarcastic, humorous, offensive and motivation type, a meme can belong to multiple categories; c) sub-task c is a semantic level classification of the various humor types of task b, each type is divided into 4 categories: not, slightly, mildly and very. The task uses macro F1 as the evaluation standard for sub-task a, and average macro F1 as the evaluation standard for sub-tasks b and c.

There are three challenges for this task. The first, how to fuse picture and text features, because whether it is a separate text or picture, the meaning expressed is lacking, and it is necessary to fuse the picture features and text semantics to obtain the best classification results. The second, the distribution of train data is imbalanced, which will increase the learn difficulty of the model. At last, the text information provided by the training data contains too much noise, which affects the semantic understanding of the text by the model and it will need to clean the data. Because we didn't find a proper way to fuse text and picture feature, we used BERT (Devlin et al., 2018) for text feature extraction, and ResNet (He et al., 2015) for image feature extraction. Through experiments, we found that the macro F1 score of BERT was higher than ResNet, so we submitted the predicted result of BERT in the final submission result.

In the rest of this article, we organize the content as follows: section 2 introduces the background of this task; section 3 mainly introduces the model we used and some details; section 4 introduces the data and preprocessing; section 5 shows the experimental results and make some analysis. Finally, we discussed our work and future work directions.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1100–1105
Barcelona, Spain (Online), December 12, 2020.

2 Background

Prez Rosas et al. (2013) use the Spanish video collected from YouTube to fuse the audio, video and text at feature level, and then perform sentiment classification on the video dataset. Dobrisek et al. (2013) use decision-level fusion to merge audio and video. Designed a multimodal emotion recognition system. Wollmer et al. (2013)focus on the task of automatically analyzing a speaker's sentiment in on-line videos contain movie reviews, using a mixed feature model to merge audio, video and text. Truong and Lauw (2019) use pictures to extract effective text information for sentiment classification. Xu et al. (2019) researched aspect-level sentiment classification of images and texts. Cai et al. (2019) extract image attributes features, which help the multimodal models for photo-text sarcasm detection. Zadeh et al. (2017) mix single-mode, dual-mode and triple-modal feature information to identify emotion in the comment video. However, none of these research work uses a pre-trained model based on Transformers, but the pre-trained model has proved its effectiveness in many works, so we decided to use BERT instead of LSTM this time.

We participated in sub-task a, which is to judge the emotions of memes, and we only used the data provided by the competition organizer (Sharma et al., 2020). In order to effectively extract the image features and text features, we considered using the single-modal model to extract the features of the image and the text, and then the multimodal model is designed on the basis of the single-modal model. However, the multimodal model is found to be less effective than the single-modal model. It will discuss in section 5.

3 System Overview

In this task, we mainly used the ResNet-101 pre-training model and the BERT pre-training model. We refer to VistaNet (Truong and Lauw, 2019), Hierarchical Fusion Model (Cai et al., 2019) and Tensor Fusion Network (Zadeh et al., 2017) respectively. We also refer to Liu (2019), and extract the effective features of the picture through the sentences in the text, and then merge the text features and the effective features of the picture. But we found that the prediction results of the multimodal model are not as good as the single-modal model, so we did not use multimodal model in the final submission results, see section 5 for details.

3.1 Single-modal for Photo

ResNet He et al. (2015) proposed a residual nets (ResNet) .The residual nets solves the problem of network degradation in deep networks by adding the shallow output of the model to the deep output of the model, and successfully increased the network depth to 152 layers. We would like to use ResNet as a single-modal model to train the image. Through experiments, we decided to use ResNet-101. The pre-training model we use is from the torchvision module of pytorch[1].

RandAugment The Google research team released RandAugment (Cubuk et al., 2019), which is used to automatically data augmentation with a reduced search space . RandAugment has two parameters N and M, where N represents the number of augmentation transformations to apply sequentially and M represents the magnitude for all the transformations. Because we found that the training data is small, which may lead to the model can not be fully trained, so we consider using RandAugment to enhance the data. In this task, we set N = 14, M = 11, and in the submission result F1 (macro) obtained a score of 0.334.

3.2 Single-modal for Text

BERT The Google research team released the pre-trained model BERT (Devlin et al., 2018) on the basis of Transformer. Through pre-training on a huge number of corpora, BERT has achieved state of the art results on many NLP tasks. Considering that the dataset of SemEval-2020 Task8 (Sharma et al., 2020) is not large, we adopted the BERT-base[2] version. Because Google did not provide the pytorch version of

[1] https://pytorch.org/docs/stable/torchvision/models.html\#classification

[2] https://github.com/google-research/bert

positive		negative		neutral
very positive	positive	very negative	negative	neutral
1033	3127	151	480	2201

Table 1: The first table is the distribution of data in sub-task a.

the BERT model code, we used the BERT model loading module provided by HuggingFace (Wolf et al., 2019).

TSA Google mentioned a training technique Training Signal Annealing (TSA) in Unsupervised Data Augmentation (Xie et al., 2019). TSA can gradually release labeled training data as the training progresses. Because we consider the imbalance of the dataset of SemEval-2020 Task8 (Sharma et al., 2020), we learn from TSA's skills. During the training process, if the prediction accuracy of a certain type of data reaches the threshold, then remove it during training.

KL-divergence Because we found that the text data provided by SemEval-2020 Task8 (Sharma et al., 2020) contains a lot of text fragments that are not related to the task, we cleaned the text data, see section 4.2 for details. Because we only cleaned the training data, but not the test data. , Resulting in inconsistency between training data and test data. To solve this problem, we redefined a new loss as follows:

$$O_c = F_{bert}(x_c) \tag{1}$$
$$O_d = F_{bert}(x_d) \tag{2}$$
$$KL_{loss} = Loss_{kl}(O_c, O_d) \tag{3}$$
$$CE_{loss} = Loss_{ce}(O_c, y_t) \tag{4}$$
$$loss = KL_{loss} + CE_{loss} \tag{5}$$

Where $F_{bert}(\cdot)$ is the model based on BERT, O_c and O_d represent the probability distribution of the cleaned and uncleaned text obtained by the model. $Loss_{kl}$ and $Loss_{ce}$ is the kullback-leibler divergenceloss and cross-entropy loss function.

4 Experimental Setup

4.1 Data Setup

Memotion Dataset 7k Dataset (Sharma et al., 2020) is collected for memotion analysis. This task judges the various emotions expressed by the meme based on the meme picture and the text in the picture. Table 1 lists the number of memes of each category in the sub-tasks a. We divide the data into five parts, four of which are training set and one is development set.

4.2 Preprocessing

We found that the text in the data set mainly contains three kinds of errors, a) contains meaningless text, such as URL; b) lack of punctuation, resulting in two sentences becoming one sentence; c) the order of the sentences is disordered or multiple sentences are randomly mixed. Therefore, we have corrected the text with the above errors.

The training objective is cross-entropy and KL divergence, and Adam (Kingma and Ba, 2015) optimizer is adopted to compute and update all the training parameters. Learning rate is set to 2e5 for model, respectively. We also use gradual warmup (Goyal et al., 2017) and cosine annealing schedule for learning rate.

5 Results

The evaluation metric of sub-task a is macro-F1. We divide the dataset into training (5595 meme data) and development (1397 meme data) subsets. We initialize the model with different random seeds for many

Input Type	Model Type	System	Dev-F1	Test-F1
text-picture	multimodal	HFM	0.3682	0.2062
text-picture	multimodal	TFNet	0.3608	0.3267
text-picture	multimodal	TFNet+TSA	0.3658	0.3329
picture	single-modal	ResNet	0.3694	0.3294
text	single-modal	BERT	0.3822	0.3203
picture	single-modal	ResNet+TSA+RA	0.3728	0.3336
text	single-modal	BERT+TSA+KL+200	0.3746	0.3458

Table 2: Dev-F1 means the prediction result on the development set, Test-F1 means the prediction result on the test set. TSA means training signal annealing, RA means RandAugment, KL means kullback-leibler divergence loss.

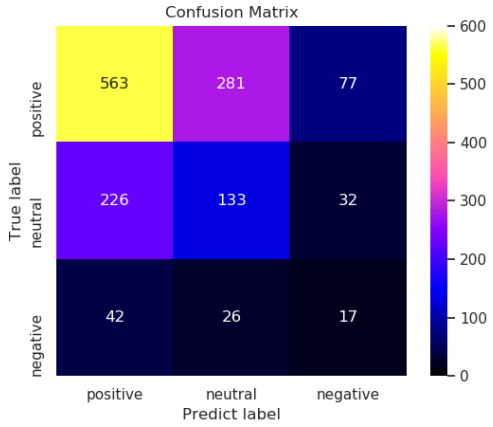

Figure 1: The confusion matrix of the classification results in Sub-task A

times. Various models are used for training. At the same time, we use a variety of models for training, and the prediction results are shown in Table 2.

From the results, we can see that the prediction results of the multimodal model HFM and TFNet are not as good as the single-modal model. We tried VistaNet (Truong and Lauw, 2019), HFM (Cai et al., 2019) and TFNet (Zadeh et al., 2017). VistaNet's prediction result in this task is the worst. We think VistaNet's main purpose is to filter important text through multiple pictures. It focuses is not just the fusion of text and picture information, so we gave it up early. However, the prediction results of HFM and TFNet are not as good as the single-modal model too. By observing the data set , We found that the meme dataset (Sharma et al., 2020) of SemEval-2020 Task8 has a feature, that is the understanding of meme is very dependent on the alignment of text segment and image regions. But neither HFM nor TFNet focuses on it. So we later designed a model to align the picture and the text. We divide the text into sentences and get the features of sentences by BERT (Liu and Lapata, 2019). Then we get the 7x7 image regions through ResNet-101. At last, performing attention operations on the images regions by sentences. But the actual effect is not good, which is worse than HFM and TFNet. We found that in many cases, a text is only divided into one sentence, so we consider that it may be the sentence level division is too rough. We guess that the use of phrase level or word level division may achieve better results, but we have not designed a suitable model. So in the end we did not use the multimodal model.

we can see that the result of using text alone is better than using pictures alone, it should be the picture information is more complicated, especially some pictures contain multiple sub-pictures. At the same time, we found that after adding TSA and RA, the ResNet's result is improved. Training Signal Annealing (TSA), which can gradually releases training data to model. Specifically, if the models predicted probability for the category positive is higher than a threshold η, we remove that data from

this training step (Liu, 2019). As for RandAugment, data augmentation has the potential to significantly improve the generalization of deep learning models (Cubuk et al., 2019). Then, when TSA and KL are added, the result of BERT is improved greatly. KL can be referred to 3.2 . In the end, the best result we submitted, the highest score of macro-F1 was 0.3458, which is about 0.0088 lower than the first place 0.3546 in sub-task a.

The confusion matrix is shown in Figure 1. From the figure, we can see that because of the problem of label imbalance in the dataset, the precision of the negative label is only 20%, and the precision of the positive label reaches 61.13%. This also leads to this macro- F1 is relatively low.

Conclusion

Is very important for social network to recognize the sentiment of Internet memes, but the main difficulty is how to integrate the information of pictures and texts, Which is a very challenging job. Although our multimodal model has not been get very good results, our single-modal model got the 10th . Considering that our multimode model may not align the text features and image features correctly, we hope to use a suitable way to align text features and picture features. For example, use a phrase-level way to divide text, and then align the features of the picture. At the same time, we found that the problem of label imbalance in the dataset led to the low value of the final macro-F1, so in the future we will consider using some other strategies to reduce the impact of label imbalance.

References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2506–2515. Association for Computational Linguistics.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. Randaugment: Practical automated data augmentation with a reduced search space.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Simon Dobrisek, Rok Gajek, France Mihelic, Nikola Pavesic, and Vitomir truc. 2013. Towards efficient multi-modal emotion recognition. *The International Journal of Advanced Robotic Systems*, 10, 01.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China, November. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

S. K. Park. 2020. Understanding usage of memes over social medias through semantics: A survey. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, Feb.

V. Prez Rosas, R. Mihalcea, and L. Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, May.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Quoc-Tuan Truong and Hady W. Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 305–312. AAAI Press.

Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Computers in Human Behavior*, 63:424 – 432.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Martin Wollmer, Felix Weninger, Tobias Knaup, Bjrn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *Intelligent Systems, IEEE*, 28:46–53, 05.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 371–378. AAAI Press.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics.