Little Flower at Memotion 2.0 2022 : Ensemble of Multi-Modal Model using Attention Mechanism in MEMOTION Analysis

Kim Ngan Phan[1], Guee-Sang Lee[1], Hyung-Jeong Yang[1] and Soo-Hyung Kim[1]

[1]*Dept. of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea*

Abstract

In modern society with the explosion of multimedia, the content and quality of information need strong attention. On social platforms such as Facebook, Instagram, and Twitter, the memes in the hate speech are sarcastic, threatening, and hateful. Natural language processing and computer vision are applied to the study of multi-modality social media such as visual and textual. In Memotion 2.0 2022, we were provided with 8,500 annotated memes with an emotion classification task. The tasks of the challenge included sentiment analysis (Task A), emotion classification (Task B), and intensity classification of meme emotions (Task C). In this paper, we propose multi-modal architecture for textual and visual modalities. Our architecture applies attention mechanisms and residual learning for VGG16 and BiLSTM to extract textual and visual representation respectively. Our approach on test set achieves 82.29% weighted average F1 score and ranks 1^{st} for Task B. In addition, we get 50.81% weighted average F1 score and ranks 4^{th} for Task A of Memotion 2.0 2022.

Keywords

memes, hate speech detection, emotion classification, natural language processing, attention mechanism, deep learning.

1. Introduction

Along with the development of the Internet and smart devices, the connection between people regardless of geographical distance has become more convenient and necessary. Social media can deliver news very quickly and become an effective instrument to receive new information. Besides that, anonymity and freedom of expression in social media raise negative issues. The social media is used to propagate hate speech and promote reactionary organizations [1] [2] [3]. In [4], hate speech is understood as "any communication that disparages a person or a group based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" . The Facebook and Twitter companies have been implementing manual moderation solutions to solve these problems [5] [6]. But with the rapid development of social networks, hate speech remains unresolved and is abundantly expressed with various templates in memes on social media. Internet memes are a form of communication and conveying information. Offensive memes have information from various media formats

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

kimngan260997@gmail.com (K. N. Phan); gslee@jnu.ac.kr (G. Lee); hjyang@jnu.ac.kr (H. Yang); shkim@jnu.ac.kr (S. Kim)

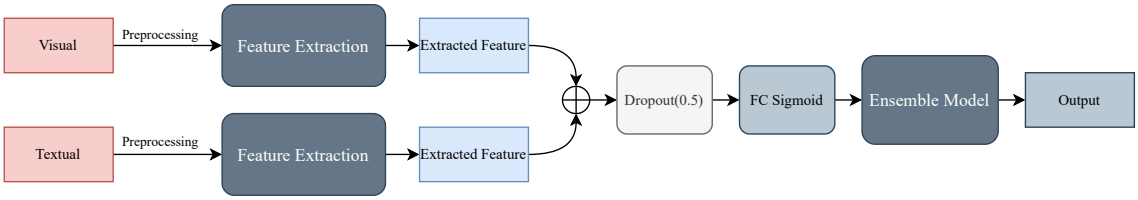

Figure 1: The Proposed System Architecture. We first extract representations from unimodal models of visual and textual. The combination of representations is fed into the dropout layer followed by dense layer with sigmoid activation. Finally, we utilize the ensemble model and get the final prediction.

such as visual and textual so detecting offensive memes is more difficult than detecting offensive text. In recent years, a few researchers have proposed many automatic approaches to Internet memes. The SemEval-2020 Task 8 [7] brings the opportunity for researchers to construct classification systems for memotion analysis. In [8], the authors propose multi-modal multi-task approach with employing ALBERT and VGG-16 for text and image representation respectively. In [9], they perform 5-fold cross-validation and ensemble five different representations including Bi-GRU, BERT, and ELMo for text extraction, Resnet50 for image extraction, and fusion features of text and images. In Memotion 2.0 2022 competition [10], we have a condition to propose an approach using visual and textual modalities for memes. We perform two unimodal models to get representation for visual and textual modalities. For visual modality, we propose the VGG16 pre-trained model and the multi-head attention [11] to extract the features. For textual modality, we utilize simply Bahdanau attention [12] to extract the context vector. Besides, unimodal models employ residual learning [13] after using attention mechanisms to connect information between the dense layer and the previous layer of attention mechanisms. The representation features are fused and fed into dense layer with the sigmoid activation to get probability classes. Finally, we perform an ensemble model for the final prediction. Our approach achieves 82.29% weighted average F1 score with ranks 1^{st} for task B of Memotion 2.0 2022. [14]. Additionally, our approach achieves 50.81 % weighted average F1 score and ranks 4^{th} rating for Task A. The main content of the paper is utilized for task B and the additional results of task A will be discussed further in the results section.

In this paper, section 2 represents the proposed method. We propose uni-models to extract features for modalities, the ensemble method, and the loss function. Section 3 describes the data and experiments. Finally, we summarize our approach and future works in the conclusion.

2. Proposed Method

In this work, memes are multi-modal data consisting of visual and corresponding embedded textual. Faced with this problem, we propose fusing the unimodal model of individual modality for the emotional annotation of memes. Figure 1 describes the entire process of our approach.

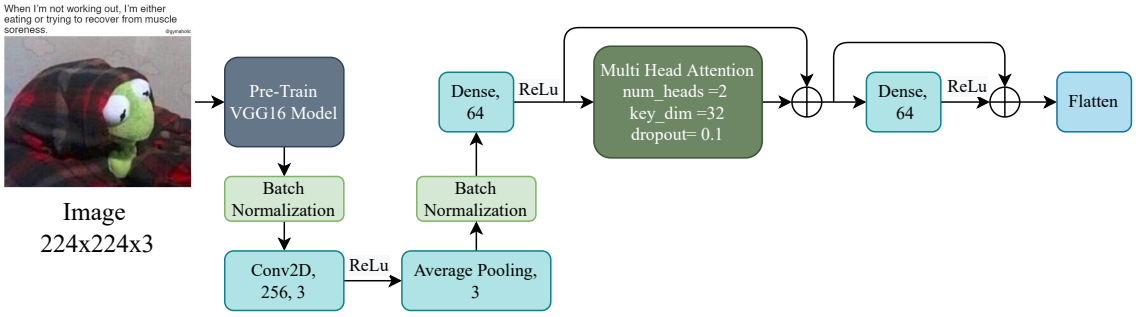

Figure 2: Pipeline for the unimodal model using visual modality. The visual features are extracted using the pre-trained VGG16 model, multi-head attention, and residual learning.

2.1. Visual Feature Extraction

VGG16 is a famous and widely available model using the convolutional neural network that has 16 layers. This pre-traineded network trained on more than a million images from the ImageNet database. In this work, we used this powerful model to extract high-level features. The high-level features of the VGG16 model represent the character of the image. The original images are synchronously resized into 224x224x3 and role as input to the VGG16 model that removes the last dense layers of the network. The output shape of the feature map is 7x7x512. The feature map is standardized by batch normalization [15] and further apply to the convolution layer and ReLu activation. Then, they are batch normalized and fed into dense layer to apply the multi-head attention module [11] with two head attentions. We implement blocks based on residual learning to connect information before and after using the attention module. We continue to implement this through the dense layer and get the final representation through flatten. Figure 2 depicts unimodal model to extract features based on visual modality.

2.2. Textual Feature Extraction

With a given textual, we need to perform pre-processing to normalize the textual before it was tokenized and embedded. Pre-processing is the process of filtering noise to reduce noise from the raw data. The words are uniformly converted to be lowercase to avoid the distinction between uppercase and lowercase as their meanings are not different. In addition, we filter out special characters and 'top words'. Next, we perform vectorization of a textual corpus. For each textual, we implement a tokenizer to assign an integer string. These integers are role as the token in our dictionary. If the words are not in the original dictionary, they are added and replaced out-of-vocabulary words. To improve efficiency in the classification, we use Bahdanau attention [12] for the unimodal model of the textual. In Bahdanau attention, we employ the weighted sum of attention weights and the encoder hidden states for the context vector to keep useful information. The attention weights represent the weight of influence for each word of the input sequence. In this work, the bidirectional LSTM (BiLSTM) is used as the encoder class. The BiLSTM learns two LSTM networks: forward directional from left to right and backward directional from right to left. It helps to increase the amount of encoding information and

exploit the context of neighbors for each word. It is the motivation for us to use a combination of BiLSTM and context vector of Bahdanau attention. For the input sequence with length T, the network takes only forward hidden states $\overrightarrow{h_j}$ and backward hidden states $\overleftarrow{h_j}$, $j = 1, ..T$. For each word, the annotation $h_j = [\overrightarrow{h_j}, \overleftarrow{h_j}]$ summarizes the information of the words before and after the jth word. In Bahdanau attention [12], the context vector is given by the formula:

$$c = \sum_{j=1}^{T} \alpha_j h_j \tag{1}$$

The attention weight α_j is the probabilities of softmax activation function:

$$\alpha_j = \frac{exp(e_j)}{\sum_{j=1}^{T} exp(e_j)} \tag{2}$$

where alignment model $e_j = a(o_j, h_j)$, a as a feedforward neural network. In this work, o_j is concatenating of last forward hidden state of and last backward hidden state. The context vector is a linear combination of annotations and alignment probabilities α_j in the input sequence. We use the context vector as the textual feature of the input sequence. To get more information, we continue to implement other blocks based on residual learning [13] between the context vector and dense layer. The feature is the flattened output of residual learning [13]. Figure 3 describes our pipeline for textual feature extraction.

2.3. Ensemble method

We suggest the ensemble model using K fold cross-validation to improve performance. The data is stratified folds into K groups but keeps the labels ratio to ensure fairness in the training process. We utilize the ensemble model for probabilities of folds. For each fold, the model trains on with different weights volume. Therefore, the predicted probabilities are different. Let p_k be the predicted probability of the kth fold classifier, the ensemble output is the average of the predicted probabilities. In this work, we choose a threshold of 0.5 for ensemble output, and the final prediction $\hat{y}$ is :

$$y_{ens} = \frac{1}{K} \sum_{k=1}^{K} p_k \tag{3}$$

$$\hat{y} = \begin{cases} 1, & \text{if } y_{ens} \geq 0.5. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

2.4. Loss function

For each emotion named k, we let N is number of samples, $N_{pos}^{(k)}$ and $N_{neg}^{(k)}$ are number samples that presence and absent of emotion, respectively. The weight $W_{pos}^{(k)} = \frac{N}{2*N_{pos}^{(k)}}$ and $W_{neg}^{(k)} =$

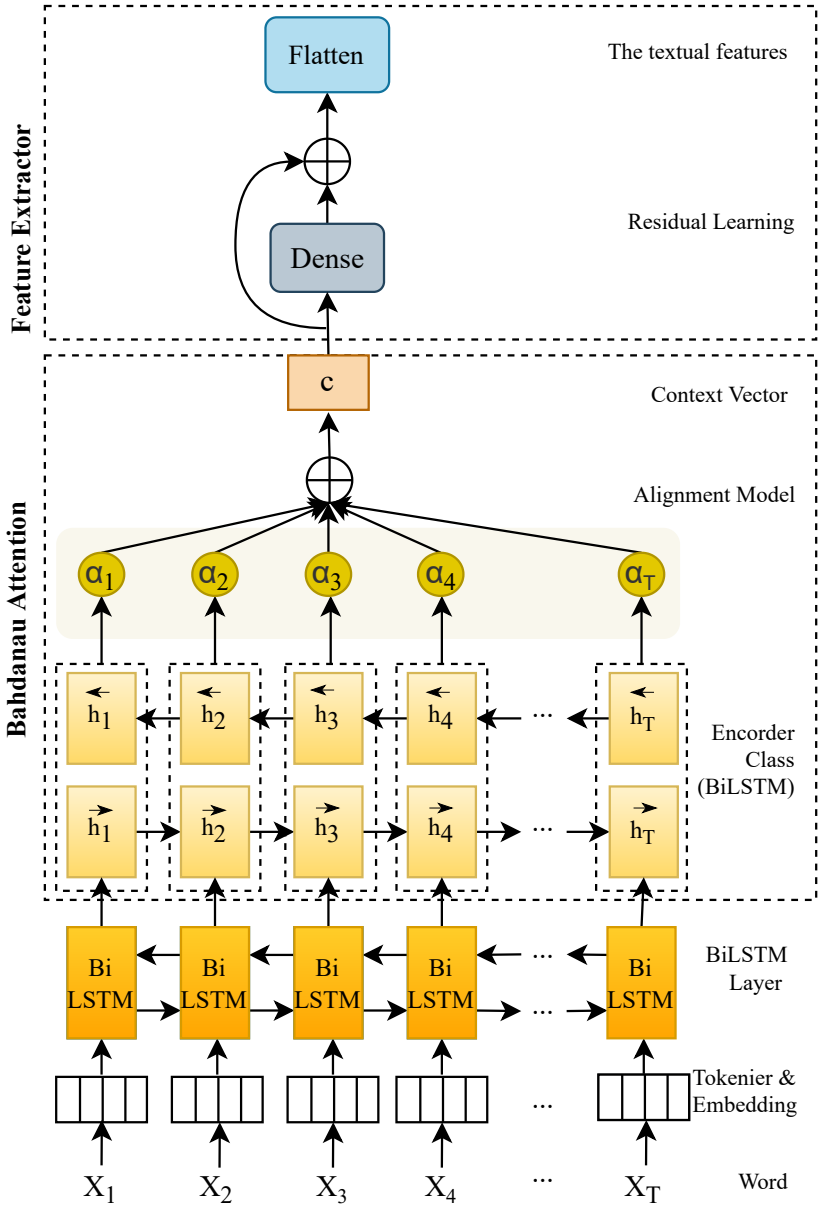

Figure 3: Pipeline for the unimodal model using textual modality. The textual features are extracted using the context vector of Bahdanau attention and residual learning. In this work, we perform BiLSTM with neurons of 64 and dense layer with neurons of 128.

$\frac{N}{2*N_{neg}^{(k)}}$ are utilized for unbalanced data. The loss function of each emotion is calculated using a weighted binary cross entropy function:

$$L^{(k)} = -\sum_{i=1}^{N} \left(W_{pos}^{(k)} y_i^{(k)} log(p(y_i)^{(k)} + W_{neg}^{(k)} (1 - y_i^{(k)}) log(1 - p(y_i)^{(k)}) \right) \quad (5)$$

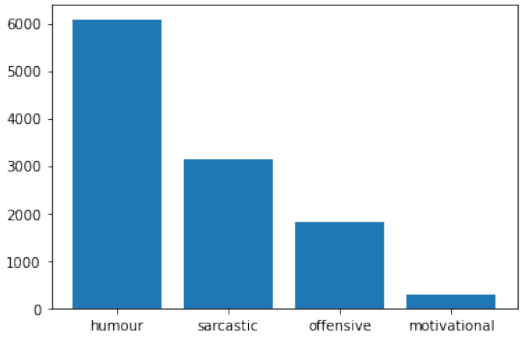

Figure 4: Memotion 2.0 distribution for emotion classification on training set.

The loss function of overall emotion:

$$L = L^{(humour)} + L^{(sarcastic)} + L^{(offensive)} + L^{(motivational)} \tag{6}$$

3. Experiments

3.1. The Memotion 2.0 2022 Task and Dataset

Memotion 2.0 [14] consists of three tasks:

- Task A - Sentiment classification is to classify positive, negative, or neutral.
- Task B - Emotion classification is to identify more than one emotion label that contains: humorous, sarcastic, offensive, and motivational.
- Task C - Intensity of emotion classification is to quantify the intensity of emotion that contains: humorous, sarcastic, offensive, and motivational.

Dataset of Memotion 2.0 [14] is part of the De-Factify workshop in AAAI-21. Data has released 8,500 annotated memes, each meme contains visuals with embedded textual contents by the English language. The meme data has 7000 as the training set, 1500 as the validation set, and 1500 samples as the test set. The main content of the paper is utilized for task B. Figure 4 and figure 5 describe the distribution of training set and test set for the emotion classification task, respectively. It is unbalanced data.

3.2. Training Details

All network processes are trained on Keras in Tensorflow version 2.7. We use a batch size of 32 with 5-fold cross-validation on the training set. The Adam optimization [16] is utilized to optimize the loss function. To evaluate performance, we perform a weighted F1 score for each emotion. The average of scores is used to evaluate the performance of the proposed model.

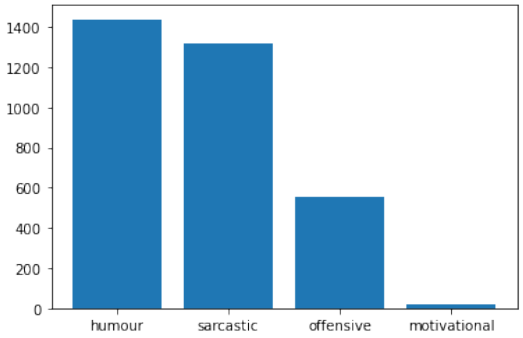

Figure 5: Memotion 2.0 distribution for emotion classification on test set.

Table 1
Task B - Emotion classification performance (%) on the test set with baseline.

Team	Method	Humour	Sarcastic	Offensive	Motivational	Average
-	Baseline[14]	78.78	64.43	55.17	95.95	73.58
Our Approach	Our Approach	93.84	81.90	55.40	98.00	82.29

Table 2
Our results on the test set in Memotion 2.0.

Approach	Task A	Task B
No residual learning	**50.81**	81.35
Residual learning	48.31	**82.28**

3.3. Results

The results using textual and visual modalities are shown in Table 1. There are predictions on humour, sarcastic, offensive, and motivational emotion labels. In [14], results show the effective performance using the combination of image and text. In baseline, the authors use ResNet-50 and BERT model to extract features from image and text respectively. Their model achieves 73.58% for the weighted average F1 score. In our approach, we apply attention mechanisms for VGG16 and BiLSTM model for image and text respectively. The residual learning block is performed to get information on the dense layer and previous layers. The combination of image and text features is fed into the dense layer with sigmoid activation and ensemble model for classification. Our model achieves 82.29% on the test set and wins for task B in Memotion 2.0 2022 [14].

Furthermore, we also report additional results for submissions in Memotion 2.0 in table 2. We try cutting out the last residual learning block to remove the information through the dense layer. This helps us only improve the performance of task A with cross entropy loss function and get 4th in the rankings. Table 3 compares model performances of the participating teams in Memotion 2.0 2022 [14].

Table 3
Comparison of performance (%) on the test set for task A and task B in Memotion 2.0.

Ranking	Team Name	Task A	Team Name	Task B
1	BLUE	53.18	**Little Flower (Our team)**	**82.29**
2	BROWALLIA	52.55	BLUE	80.59
3	Yet	50.88	BROWALLIA	76.70
4	**Little Flower (Our team)**	**50.81**	Amazon PARS	76.09
5	Greeny	50.37	HCILab	74.14
6	Amazon PARS	50.25	BASELINE	73.58
7	HCILab	49.95	weipengfei	69.15
8	weipengfei	48.87	Yet	61.06
9	BASELINE	43.40	Greeny	61.06

4. Conclusion

In this paper, we propose an approach for the sentiment classification and the emotion classification respectively task A and task B in Memotion 2.0 2022. We extract visual features using the pre-trained model VGG16 and multi-head attention. We use BiLSTM encoder for Bahdanau attention to extracting the context vector of the textual. Furthermore, we propose the ensemble model for the final prediction. Our team achieves 50.81% and 82.29% weighted average F1 score for task A and task B, respectively. However, our approach only simply combines the representation of textual and visual but does not impress for combining textual representation corresponding to visual representation with the same meaning. In the future, we will continue to overcome the disadvantages and improve the performance of the model on other meme datasets.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A3A04036408). The corresponding author is Soo-Hyung Kim.

References

[1] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[2] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & internet 7 (2015) 223–242.

[3] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, Semantic Web 10 (2019) 925–945.

[4] J. T. Nockleby, Hate speech, Encyclopedia of the American constitution 3 (2000) 1277–1279.
[5] N. Lomas, Facebook, google, twitter commit to hate speech action in germany, Last accessed: July (2017).
[6] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.
[7] 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).
[8] 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).
[9] Y. Guo, J. Huang, Y. Dong, M. Xu, Guoym at semeval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1120–1125.
[10] 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.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[12] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[13] 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.
[14] S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, et al., Memotion 2: Dataset on sentiment and emotion analysis of memes (2021).
[15] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR, 2015, pp. 448–456.
[16] I. K. M. Jais, A. R. Ismail, S. Q. Nisa, Adam optimization algorithm for wide and deep neural network, Knowledge Engineering and Data Science 2 (2019) 41–46.