HCILab at Memotion 2.0 2022: Analysis of Sentiment, Emotion and Intensity of Emotion Classes from Meme Images using Single and Multi Modalities

Thanh Tin Nguyen[1], Nhat Truong Pham[2,3], Ngoc Duy Nguyen[4], Hai Nguyen[5], Long H. Nguyen[6] and Yong-Guk Kim[1] (Corresponding author)

[1]*Human Computer Interaction Lab, Department of Computer Engineering, Sejong University, Seoul, Korea*
[2]*Division of Computational Mechatronics, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam*
[3]*Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam*
[4]*Institute for Intelligent Systems Research and Innovation, Deakin University, Victoria, Australian*
[5]*Khoury College of Computer Sciences, Northeastern University, Boston, USA*
[6]*Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam*

Abstract

Nowadays, memes found on internet are overwhelming. Although they are innocuous and sometimes entertaining, there exist memes that contain sarcasm, offensive, or motivational feelings. In this study, several approaches are proposed to solve the multiple modality problem in analysing the given meme dataset. The imbalance issue has been addressed by using a new Auto Augmentation method and the uncorrelation issue has been mitigated by adopting deep Canonical Correlation Analysis to find the most correlated projections of visual and textual feature embedding. In addition, both stacked attention and multi-hop attention network are employed to efficiently generate aggregated features. As a result, our team, i.e. HCILab, achieved a weighted F1 score of 0.4995 for sentiment analysis, 0.7414 for emotion classification, and 0.5301 for scale/intensity of emotion classes on the leaderboard. This results are obtained by using concatenation between image and text model and our code can be found at https://git.io/JMRa8.

Keywords

Meme analysis, attention models, correlation analysis, emotion classes, multimodality, vision and language

1. Introduction

The task of analyzing sentiments, emotions and their intensity has attracted a great deal of attention in research community, especially when it can help to subdue unnecessary damages. As the internet has spread worldwide, false information, hatred, or offensive language are also

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022, Vancouver, Canada.
nttin@sju.ac.kr (T. T. Nguyen); phamnhattruong.st@tdtu.edu.vn (N. T. Pham); n.nguyen@deakin.edu.au (N. D. Nguyen); hainguyen@ccs.neu.edu (H. Nguyen); hoanglong.fruitai@gmail.com (L. H. Nguyen); ykim@sejong.ac.kr (Y. Kim)
0000-0002-6798-9808 (T. T. Nguyen); 0000-0002-8086-6722 (N. T. Pham); 0000-0002-4052-5819 (N. D. Nguyen)

increasing tremendously. A common way to disseminate these threads is by using texts in meme images that vicious people can mitigate as a mean to agitate arguments, disputes, and social wars.

To mitigate harmful effects of toxic memes, machine learning [1] and deep learning [2, 3] are normally employed to tackle the problem. These techniques can detect and classify memes effectively, although it requires humans to label the data. Nevertheless, the results are promising and the algorithm can be integrated into social media platforms such as Facebook or Twitter to automatically detect and remove these memes completely.

Following the success of the Semeval 2020 challenge [4], in the Memotion2 challenge, the organizer provides a new dataset [5], [6] including memes and corresponding texts. The task includes three subtasks: (Subtask A) sentiment analysis which is to classify negative, neutral, and positive contents; (Subtask B) emotion classification which is to classify emotions of memes, there are four main categories including funny, sarcastic, offensive and motivational; (Subtask C) the last task is to seek for detail information of each emotion, for example, funny, sarcastic, and offensive emotions have four levels while the last emotion only has two levels. The weighted F1 score is used in this competition to evaluate each subtask, the final score is the average of three subscores.

In addition, the task has two important issues. Firstly, the data is imbalanced among different classes. Secondly, images and their corresponding texts are not well correlated to each other, because the texts and the images often point to different meanings in meme images, so there is a need to build an effective fusion technique to reduce a semantic gap between two modalities. In order to address the problem, our team proposed several models and achieved good results on the private leaderboard.

The remaining of this study is organized as follows. Section 2 introduces brief literature of this problem. Section 3 describes our methodology including unimodal, bimodal as well as auxiliary techniques. Results are summarized in section 5. Finally, we conclude our study and outline future research in Section 6.

2. Background

In the previous competition, different approaches were developed to tackle the problem. For instance, in [1, 2, 3], the authors introduce machine learning and deep learning models including Naive Bayes [7], BERT [8], Multimodal Transformer [9], and ResNet [10] to tackle the problem. Nevertheless, these are mainly divided into two types of models including unimodal and bimodal. The unimodal uses only one modality as an input which can be texts or images. The bimodal adopts fusion techniques to aggregate features from different modalities to obtain related information and achieve a better classification rate. Previous studies have employed state-of-the-art models for text and vision, but they did not consider the correlation of two modalities and how to preprocess the data to create a clean one.

In this study, EfficientNet-v2 [11] is employed as a visual extractor, while LSTM [12] and RoBERTa [13] are used in the bimodal to extract textual features. In addition, RoBERTa is also used for text model. With respect to fusion techniques, there are three methods to obtain aggregated features including traditional aggregated method, multi-hop attention [2], and

stacked attention [14]. These techniques are used to combine visual features and textual features from LSTM and RoBERTa. In general, we evaluated six different models during the competition.

Besides, we also adopt several techniques to improve the classification rate such as Auto Augmentation [15] and Deep Canonical Correlation Analysis [16]. Finally, to enhance the visual extraction, we remove texts on memes by using EAST [17] to detect texts within images and then remove them.

3. Methodology

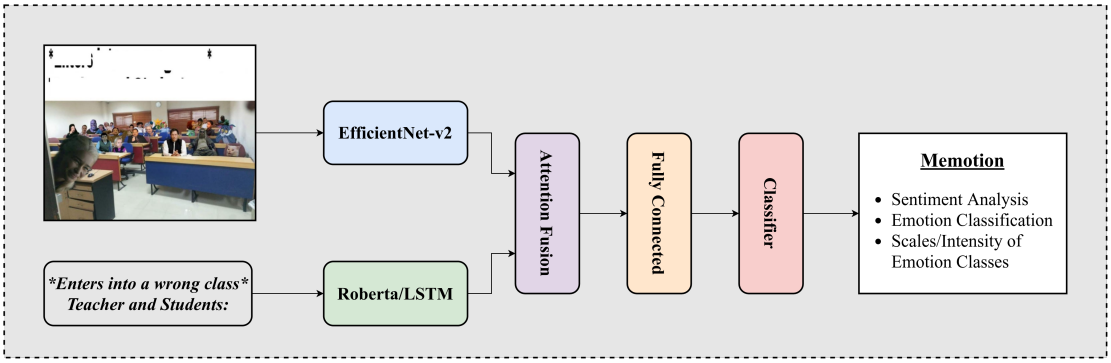

Figure 1: The proposed framework for multiple modalities that has two inputs: (1) an image after removing the text in it; and (2) an extracted text from the image. The image is processed by Efficientnet-v2, while a text is processed by either a Roberta or an LSTM. Then, these extracted features are aggregated by a fusion network which includes a combination of a multi-hop attention network and/or a stacked attention network. A fused vectors is used to predict the classes of corresponding inputs depending on each subtask.

We have evaluated many network architectures along with fusion techniques. In addition, we employ auxiliary methods such as Auto Augmentation and Canonical Correlation Analysis to enhance the efficiency. The proposed models are divided into Unimodal and Multimodal based on a vision backbone, i.e., EfficientNet-v2, [11] and LSTM [12] and RoBERTa [13] for text processing.

Figure 1 depicts the proposed framework for multiple modalities. It includes one branch for extracting features from the image and one for extracting features from the text. These features are concatenated by an attention-based fusion module before passing through a fully connected layer for final classification. The number of output nodes of this layer depends on each task. For the sentiment task, it will be 3 nodes denoting for Negative, Neutral and Positive classes. In terms of the emotion task, there will be 4 final linear classifiers, each one will two output nodes, because in this task, there are 4 types of emotions, each emotion has two classes 0 and 1. Lastly, for the intensity task, there are also 4 final linear classifiers, but each one will have different output nodes, for example, in the intensity of humour class, there will be 4 nodes denoting 4 levels of intensity. Meanwhile, that of motivation class will have only 2 output nodes.

3.1. Unimodal for Text

BERT [8] and its variant, e.g., RoBERTa [13], are widely used in Natural Language Processing (NLP) tasks and have demonstrated as efficient methods. In this competition, we employed them for three subtasks. In subtasks A and B, a RoBERTa [13] is used while in subtask C, four RoBERTa models are adopted so that every backbone corresponds classifying the intensity of each emotion.

3.2. Unimodal for Image

As a vision-based approach, EfficientNet-v2 [11] is a well-known backbone with respect to speedy inference and a low number of parameters. In three subtasks, EfficientNet-v2 is used as an extractor to create embeddings. Subtask A has one classification branch to deal with three sentiment types while subtasks B and C have four branches which each is responsible for classifying four types of emotions as well as their intensity.

3.3. Multimodal for Image and Text

Multi-modality is to aggregate vision and text to obtain correlated information. In this challenge, we build three different fusion models including concatenation, multi-hop attention [2], and stacked attention network [14].

3.3.1. Concatenation

Traditionally, concatenation of two feature vectors, a.k.a two modalities, has been a typical solution to obtain aggregated features. However, the method does not take into account the importance of each word that is within corresponding regions of the image.

3.3.2. Multi-hop Attention

Multi-hop attention is initially proposed by [2]. It focuses parts of a given image together with texts within it. The technique aims to emphasize dissimilar features between image regions and textual utterances by defining a relevant matrix R, which is the cosine distance between textual and visual features.

3.3.3. Stacked Attention

While a multi-hop attention network is used to learn attention maps between an image and texts within it, a stacked attention network introduced in [14] has a capability of learning an attention map in multiple times. Through such attention layers, interested regions are promoted through a referred concept within a given sentence.

3.4. Useful Techniques

3.4.1. Auto Augmentation

Augmentation is a simple but important technique to increase the size of a given dataset, leading to a better generalization of a training model. However, current data augmentation is based on a set of manually designed algorithms such as Crop, Rotation, and Resize. In our experiment, we adopt the Auto Augment technique [15] which uses reinforcement learning to automatically search for a better data augmentation strategy.

3.4.2. Canonical Correlation Analysis

The Canonical correlation analysis (CCA) was proposed by [18]. It is based on a well-established statistical technique that searches for a linear combination of input vectors by maximizing their correlations. Deep CCA [16] tries to utilize the power of both deep neural networks and CCA to overcome projection constraints of CCA . In this study, correlation scores obtained from Deep CCA is included to our loss function to maximize the correlation between two features, leading to a higher classification rate.

4. Experiment

4.1. Dataset

In this shared task of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection, MEMOTION 2.0 [5] was used which was a hate speech detection dataset. It included 7,000 samples for the training set and 1,500 samples for the validation set. This dataset was used for three subtasks in the MEMOTION 2.0 Challenge and labeled as follows:

- Sentiment analysis:
  - *Negative* and *Very Negative* are labeled 0;
  - *Neutral* is labeled 1;
  - *Positive* is labeled 2.
- Emotion classification:
  - *Not Humorous* is labeled 0, while *Humorous* is labeled 1 that includes funny, very funny, and hilarious;
  - *Not Sarcastic* is labeled 0, while *Sarcastic* is labeled 1 including little sarcastic, very sarcastic, and extremely sarcastic;
  - *Not Offensive* is labeled 0, while *Offensive* is labeled 1 that are slight, very offensive, and hateful offensive;
  - *Not Motivational* is labeled 0 and *Motivational* is labeled 1.
- Scale/intensity of emotion classes:
  - *Humour*: Not funny, funny, very funny, and hilarious are labeled 0, 1, 2, 3, respectively;

- *Sarcasm*: Not sarcastic, little sarcastic, very sarcastic, and extremely sarcastic are labeled 0, 1, 2, 3, respectively;
- *Offense*: Not offensive, slight, very offensive, and hateful offensive are labeled 0, 1, 2, 3, respectively;
- *Motivation*: Not motivational is labeled 0 and motivational is labeled 1.

4.2. Preprocessing

Although both textual and visual features are important for meme emotion analysis, there is little correlation between them in the MEMOTION 2.0 dataset. Besides, the caption is also provided as a part of the dataset. Therefore, in this study, the text is removed from the image before extracting and training the proposed model.

Based on the previous work [19] that summarized both traditional and deep learning approaches for text detection and recognition, we design a preprocessing scheme to remove texts from images as follows. First, we employ the EAST [17] module to detect all text regions in an image. Then these regions are removed from the image, and we use the output image as the input for EfficientNet-v2 in the proposed framework. Figure 2 visualizes the steps of the preprocessing scheme.

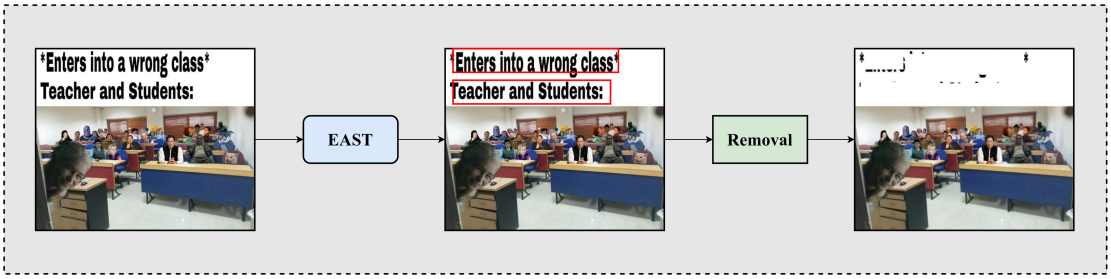

Figure 2: Preprocessing scheme: Given an image as input, we use the EAST [17] detector to detect the region of texts on the image and then remove them.

4.3. Experimental setup

All experiment was carried out using a Titan Xp GPUs station. The batch size is 10, the input image size is 256×256, the learning rate is 2e-5, the Adam [20] optimizer is used in this model with a weight decay of 1e-5. Moreover, the Cosine Annealing Warm Restarts [21] scheduler is used for scheduling the learning rate. We also use common augmentation techniques such as *Resize, CenterCrop, RandomFlip* with probability of 0.5, especially adopt the Auto Augmentation mentioned above, then take *Normalize* with mean and std are (0.485, 0.456, 0.406), (0.229, 0.224, 0.225), respectively. Finally, our models use a cross-entropy as the loss function except for single models for the texts that use binary cross-entropy instead.

5. Results

The evaluation metric of this competition is the Weighted F1 score, and the final score will be the average of three Weighted F1 scores of all subtasks. Table 1 summarizes our results in the public phase with different models. The results of the private phase are presented in Table 2. Among the best Weighted F1 scores of three subtasks, we achieved a score of 0.5124 for sentiment analysis, 0.7423 for emotion classification, and 0.5296 for scale/intensity of emotion classes, respectively.

Table 1
The Weighted F1 scores of three subtasks, namely, Sentiment, Emotion, and Intensity of Emotion in public phase. Note that these results are obtained with the validation data during the public phase, and SAN denotes for Stacked Attention Network.

Model	Sentiment	Emotion	Intensity	Average
Only Text	0.5145	0.7140	0.5781	0.6025
Only Image	0.5176	0.7033	0.5628	0.5946
Multihop Image + Text	**0.5316**	0.7107	0.5590	0.6004
SAN Image + Text	0.5138	0.7140	0.5745	0.6008
Concat Image + Text	0.5253	**0.7141**	0.5823	**0.6072**
SAN Image + Text	0.5200	0.7083	**0.584**	0.6041

Table 2
The Weighted F1 scores of three subtasks Sentiment, Emotion, and Intensity of Emotion in private phase.

Task	Sentiment	Emotion	Intensity	Average
Weighted F1 score	0.4995	0.7414	0.5301	**0.5903**

6. Conclusion and Future Work

For this study, we have integrated several attention models and the correlation analysis technique for the meme dataset analysis. To handle the imbalanced dataset, Auto Augmentation [15] is proposed and it is found that it provides a richer dataset for further processes. The visual and textual features extracted by attention models are projected into the most correlated directions by using DCCA [16] for the stable and generalized training. The best result of each subtask varies depending on combination of the used models. For the sentiment task, the multihop attention-based LSTM performs the best, whereas concatenation of CNN and BERT gives the highest result for the emotion task. The stacked attention network with CNN and BERT achieves the best for the intensity task.

In the future, an in-depth analysis shall be done by collecting or synthesizing more dataset as well as mitigating the semantic gap between the text and the image. The imbalance between classes has been a vitally important problem that can be tackled by data augmentation or formulating a new loss function that can put more weight on classes with fewer data. In

addition, since feature fusion is not always compulsory in the vision-language task, designing a noble network that can choose whether to use the fusion or not is to be investigated.

Acknowledgments

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2021-2016-0-00312) as well as a grant (IITP-2019-0-00231) supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP). In addition, the authors would like to thank the FruitLab team for useful ideas and discussion.

References

[1] 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.

[2] S. Pramanick, M. S. Akhtar, T. Chakraborty, Exercise? i thought you said'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis, arXiv preprint arXiv:2103.12377 (2021).

[3] Z. Li, Y. Zhang, B. Xu, T. Zhao, Cn-hit-mi. t at semeval-2020 task 8: Memotion analysis based on bert, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1100–1105.

[4] 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).

[5] 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.

[6] 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.

[7] I. Rish, et al., An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3, 2001, pp. 41–46.

[8] 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).

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

[10] 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.

[11] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, arXiv preprint arXiv:2104.00298 (2021).
[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[13] 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. corr abs/1907.11692 (2019), URL: http://arxiv. org/abs/1907.11692 (1907).
[14] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21–29.
[15] E. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. Le, Autoaugment: Learning augmentation policies from data. arxiv 2018, arXiv preprint arXiv:1805.09501 (2019).
[16] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International conference on machine learning, PMLR, 2013, pp. 1247–1255.
[17] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, East: an efficient and accurate scene text detector, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 5551–5560.
[18] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in statistics, Springer, 1992, pp. 162–190.
[19] S. Long, X. He, C. Yao, Scene text detection and recognition: The deep learning era, International Journal of Computer Vision 129 (2021) 161–184.
[20] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[21] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983 (2016).