CSECU_KDE_MA at SemEval-2020 Task 8: A Neural Attention Model for Memotion Analysis

Abu Nowshed Chy[1], Umme Aymun Siddiqua[2], and Masaki Aono[3]
Department of Computer Science and Engineering
[1]University of Chittagong, Chattogram-4331, Bangladesh
[2]Asian University for Women, Chattogram-4000, Bangladesh
[3]Toyohashi University of Technology, Aichi, 441-8580, Japan.
nowshed@cu.ac.bd, umme.siddiqua@auw.edu.bd, and aono@tut.jp

Abstract

A meme is a pictorial representation of an idea or theme. In the age of emerging volume of social media platforms, memes are spreading rapidly from person to person and becoming a trending ways of opinion expression. However, due to the multimodal characteristics of meme contents, detecting and analyzing the underlying emotion of a meme is a formidable task. In this paper, we present our approach for detecting the emotion of a meme defined in the SemEval-2020 Task 8. Our team CSECU_KDE_MA employs an attention-based neural network model to tackle the problem. Upon extracting the text contents from a meme using an optical character reader (OCR), we represent it using the distributed representation of words. Next, we perform the convolution based on multiple kernel sizes to obtain the higher-level feature sequences. The feature sequences are then fed into the attentive time-distributed bidirectional LSTM model to learn the long-term dependencies effectively. Experimental results show that our proposed neural model obtained competitive performance among the participants' systems.

1 Introduction

Nowadays, social media platforms such as Facebook, Instagram, and Twitter become the most popular information sharing medium among the people due to its convenient features and realtime behavior. People usually use the various modalities of information such as textual, visual, and audio to express their views, opinions, breaking news, and ideas here. Due to the robust feature of social media, researchers and companies are trying to distill various kinds of information from its contents. But most of the previous studies address only one modality i.e. image information extraction addressed by the computer vision community and textual information extraction by the natural language processing community. However, with the growing ubiquity of Internet memes on social media, it is important to extract information from memes. We need to employ a hybrid approach in this regard since a meme comprises of an image with textual information.

Memotion analysis (Sharma et al., 2020) is commonly defined as the process of detecting and analyzing the underlying emotion of a meme. It might have a significant impact on addressing various issues related to social media. For example, evil-minded people nowadays use the meme contents to propagate anti-social behavior including online harassment, cyber-bullying, and hate speech. Therefore, memotion analysis might help to limit these anti-social behaviors.

To address the challenges of memotion analysis on social media contents, (Sharma et al., 2020) proposed the task 8 at SemEval-2020. The task focuses on three related subtasks. Task A defines a sentiment classification problem where a system needs to predict whether a meme content is positive, negative, or neutral. Whereas task B defines the multilabel multiclass humour classification problem where a system needs to identify the types of humor expressed by a meme. The categories are sarcastic, humorous, offensive, and motivation meme. Task C defines a quantification of semantic class problem where a system needs to quantify the extent of each humour class (defined in Task B) expressed by a meme.

The rest of the paper is structured as follows: **Section 2** provides a brief overview of prior research. In **Section 3**, we introduce our proposed neural attention model. **Section 4** includes experiments and evaluations as well as the analysis of our proposed method. Some concluded remarks and future directions of our work are described in **Section 5**.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1106–1111
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

Early studies on memes focused mainly on automatic meme generation from various sources. Among several prominent works, Peirson et al. (2018) employed a pre-trained Inception-v3 network for image embedding and passed the embedding features to an attention-based deep-layer LSTM model for generating meme caption. Vyalla et al. (2020) utilized the encoder-decoder based transformer model for meme caption generation. Oliveira et al. (2016) designed a rule-based classifier that relied on linguistic triggers for the generation of memes focused on news headlines.

Besides, some studies shed light on other applications of meme analysis. Beskow et al. (2020) proposed a multi-modal deep learning model to identify and characterize the political memes. Williams et al. (2016) studied the racial microaggressions and perceptions of Internet memes. A few researchers (Amalia et al., 2018; Verma et al., 2020) have tried to distill the inherent sentiment of the meme contents.

3 Proposed Neural Attention Framework

In this section, we describe the details of our proposed neural attention framework. The goal of our proposed approach is to identify several emotion orientations of internet memes including sentiment, humors, and scales of semantic classes. Figure 1 depicts an overview of our proposed model.

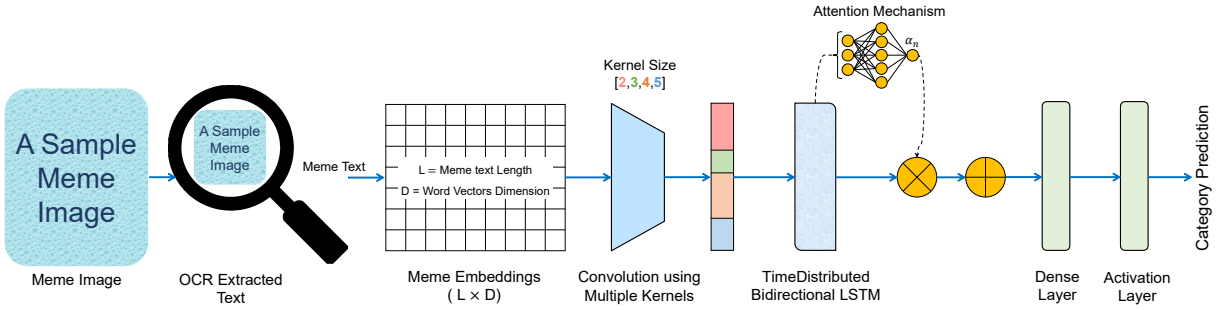

Figure 1: Proposed memotion analysis framework.

In our proposed architecture, we utilize the OCR (optical character recognition) extracted text contents to identify the emotional orientation of a meme. After extracting meme text, we employ a pre-trained word embedding model to obtain the high-quality distributed vector representations. Next, we apply the multi-kernel convolution (MKC) and time-distributed bidirectional long short-term memory (Bi-LSTM) models to extract the higher-level feature sequences with sequential information from the meme text embeddings. An attention mechanism is employed to amplify the contribution of important elements in the obtained feature representation. The generated output feature sequences are then sent to the fully-connected prediction module to determine the final category label. Next, we describe each component elaborately.

3.1 Embedding Layer

Word embedding is considered as the most popular representations of documents vocabulary. It can capture the context of a word within a text document while considering the semantic similarity and relation with other words (Mikolov et al., 2013; Bojanowski et al., 2017). In our proposed framework, we employ a pre-trained fastText (Bojanowski et al., 2017) word embedding model to capture the distributed vector representations of meme texts. The embedding matrix dimension will be $L \times D$, where L is the meme text length, and D is the word-vector dimension.

3.2 Multi-kernel Convolution

We perform the convolution operation on top of the embedding matrix obtained from the embedding layer to extract the higher-level features. Previous studies already demonstrated the efficacy of using multiple kernels based convolution compared to the single one (Kim, 2014; Zhang and Wallace, 2015; Wang et al.,

2017). In our multi-kernel convolution, we use four different kernel sizes: 2, 3, 4, and 5 to extract the different kinds of effective features.

3.3 Bidirectional Long Short-Term Memory (Bi-LSTM)

To learn sequential correlations from higher-level feature representations obtained from multi-kernel convolution, we employ the time-distributed bidirectional LSTM (Bi-LSTM) (Liu et al., 2020) model in our proposed framework. The bidirectional model runs feature representations in two ways, one from past to future and one from future to past. The difference between this approach from unidirectional is that the LSTM that gone through backward can preserve information from the future. Therefore, at any point in time, the combination of forward and backward LSTM enables the bidirectional LSTM to preserve information from both past and future. Bi-LSTMs showed very good results as they can understand the context better compared to the unidirectional RNN and LSTM.

3.4 Attention Mechanism

Recently, the attention mechanism has been widely used in the neural network frameworks to address the long-term dependencies effectively. This mechanism helps the model to learn what to attend or focus based on the input text (Vaswani et al., 2017; Fotso et al., 2018). To amplify the contribution of important elements in the final representation of time distributed bi-directional LSTM module, we employ a similar kind of attention mechanism used in DeepMoji architecture (Felbo et al., 2017). DeepMoji employed an approach based on the idea of (Bahdanau et al., 2015; Yang et al., 2016) to aggregate all the hidden states according to their relative importance weight.

Let us consider h_t representation of a word at some time t and w_a corresponds to the weight matrix at the attention layer. The attention scores a_t are estimated by multiplying h_t and w_a and perform normalizing to obtain the probability distribution. Finally, the attentional representation is obtained by a weighted summation over all the time steps as follows:

$$e_t = h_t w_a$$

$$a_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)}$$

$$v = \sum_{i=1}^{T} a_i h_i$$

3.5 Prediction Module and Model Training

After obtaining the high-level representation from the attentive bi-directional LSTM module, we pass it to a fully connected softmax layer for category prediction. We consider cross-entropy as the loss function and train the model by minimizing the error, which is defined as:

$$E(x^{(i)}, y^{(i)}) = \sum_{j=1}^{k} 1\{y^{(i)} = j\} \log(\tilde{y_j^{(i)}})$$

where $x^{(i)}$ is the training sample with its true label $y^{(i)}$. $\tilde{y_j^{(i)}}$ is the estimated probability in $[0, 1]$ for each label j. $1\{condition\}$ is an indicator which is 1 if true and 0 otherwise. We use the stochastic gradient descent (SGD) to learn the model parameter and adopt the Adam optimizer (Kingma and Ba, 2014).

4 Experiments and Evaluations

4.1 Dataset Collection and Evaluation Strategy

The organizer of the memotion analysis task 8 at SemEval-2020 (Sharma et al., 2020) provided a benchmark dataset to evaluate the performance of the participants' systems. The training dataset contained around 6992 annotated memes along with the OCR extracted text contents and the test dataset contained 1878 annotated memes, respectively.

To evaluate the performance of the system, the organizers used different strategies for the task A, B, and C (Sharma et al., 2020). For the task A, macro average F1-score was applied to estimate the performance of a system. However, for the task B and C, at first, macro average F1-score of each subtask is estimated and their average is considered as the final evaluation measure.

4.2 Model Configuration

In the following, we describe the set of parameters that we have used to design our proposed neural network model. We used the OCR extracted text provided by the task organizers and used the Tensorflow (Abadi et al., 2016) framework to design our neural model. Our model is trained on a GPU (Owens et al., 2008) to utilize the benefits of tensor computations parallelly. We used a simple grid search to select the optimal hyper-parameters. At the embedding layer, we employed the pre-trained fastText embedding model (Bojanowski et al., 2017) for the vector representation of meme texts. We used 600 filters in our multiple kernels based convolution and used a single layer Bi-LSTM model. We trained our model using 30 epochs and set the initial learning rate of 0.001 with Adam optimizer. Besides, we set the $L2$ regularization factor 0.01 in the softmax layer. Unless otherwise stated, default settings were used for the other parameters.

4.3 Experimental Results

We now evaluate the performance of our proposed method. The comparative results with top-5 performing systems (Sharma et al., 2020) along with the baseline system for task A, task B, and task C are presented in Table 1, Table 2, and Table 3, respectively. The systems are ranked based on the primary evaluation measure macro average F1 score.

Table 1: (Task A) Comparative result with other selected participants.

Team Name	Macro Avg. F1-Score	Micro Avg. F1-Score
CSECU_KDE_MA (Our Proposed)	0.323011174	0.502662407
Top 5 Participants Team based on Macro Avg. F1-Score (Sharma et al., 2020)		
vkeswani_IITK	0.354658157	0.487220447
guoym_guoym	0.351973013	0.501064963
aihaihara	0.350165731	0.470181044
souryaDiptadas	0.348853120	0.502129925
IrinaBejan	0.347551594	0.445686901
Baseline	0.217648922	0.307774228

Table 2: (Task B) Comparative result with other selected participants.

Team Name	Macro Avg. F1-Score	Micro Avg. F1-Score
CSECU_KDE_MA (Our Proposed)	0.493774206	0.650692226
Top 5 Participants Team based on Macro Avg. F1-Score (Sharma et al., 2020)		
george.vlad_eduardgzaharia_UPB	0.518339963	0.614483493
guoym_guoym	0.514633361	0.623136315
SouvikMishra_Kraken	0.509508227	0.608359957
prhlt-upv	0.509336714	0.607827476
mayukh_memebusters	0.508620655	0.612353568
Baseline	0.500205042	0.568690096

Table 3: (Task C) Comparative result with other selected participants.

Team Name	Macro Avg. F1-Score	Micro Avg. F1-Score
CSECU_KDE_MA (Our Proposed)	0.311347462	0.389110756
Top 5 Participants Team based on Macro Avg. F1-Score (Sharma et al., 2020)		
guoym_guoym	0.322460492	0.377928647
HonoMi_Hitachi	0.318839518	0.398828541
george.vlad_eduardgzaharia_UPB	0.317155246	0.405484558
jy930_rippleai	0.316371366	0.383120341
vkeswani_IITK	0.314542273	0.365149095
Baseline	0.300893044	0.332800852

Here, we see that we obtained the competitive performance while comparing with the top-performing systems. However, we think that our system lacks of taking advantage from the image information. Because our model only depends on the OCR extracted text contents. We believe that incorporating the image information might have a significant impact on memotion analysis as well as improve the performance of our model. There is a long thread of research (Islam and Zhang, 2016; Fengjiao and Aono, 2018) that used various techniques to distill the emotion of an image. Employing such techniques might be beneficial to extract the image information for this task.

5 Conclusion

In this paper, we presented our approach to the SemEval-2020 Task 8: Memotion analysis. We tackled the problem by employing an attention-based neural network model. Though we achieved the competitive performance, there is much room left to improve the performance of our method. We only exploit the information from text content. However, the information extracted from the image also necessary in this context. In the future, we have a plan to address this scenario and introducing several sophisticated deep learning techniques.

Acknowledgements

The part of this research is supported by MEXT KAKENHI, Grant-in-Aid for Scientific Research (B), Grant Number 17H01746.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, pages 265–283. USENIX Association.

Amalia Amalia, Arner Sharif, Fikri Haisar, Dani Gunawan, and Benny B Nasution. 2018. Meme opinion categorization by using optical character recognition (ocr) and naïve bayes algorithm. In *2018 Third International Conference on Informatics and Computing (ICIC)*, pages 1–5. IEEE.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

David M Beskow, Sumeet Kumar, and Kathleen M Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1615–1625. ACL.

Wang Fengjiao and Masaki Aono. 2018. Visual sentiment prediction by merging hand-craft and cnn features. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 66–71. IEEE.

Stephane Fotso, Philip Spanoudes, Benjamin C Ponedel, Brian Reynoso, and Janet Ko. 2018. Attention fusion networks: Combining behavior and e-mail content to improve customer support. *arXiv preprint arXiv:1811.03169*.

Jyoti Islam and Yanqing Zhang. 2016. Visual sentiment analysis for social images using transfer learning approach. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, pages 124–130. IEEE.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. ACL.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sisi Liu, Kyungmie Lee, and Ickjai Lee. 2020. Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. *Knowledge-Based Systems*, page 105918.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3119.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. 2016. One does not simply produce funny memes!–explorations on the automatic generation of internet humor. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016). Paris, France*.

John D Owens, Mike Houston, David Luebke, Simon Green, John E Stone, and James C Phillips. 2008. Gpu computing. *Proceedings of the IEEE*, 96(5):879–899.

V Peirson, L Abel, and E Meltem Tolunay. 2018. Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Devika Verma, Rohit Chandiramani, Pranay Jain, Chinmay Chaudhari, Anmol Khandelwal, Krishnanjan Bhattacharjee, S ShivaKarthik, Swathi Mithran, Swati Mehta, and Ajai Kumar. 2020. Sentiment extraction from image-based memes using natural language processing and machine learning. In *ICT Analysis and Applications*, pages 285–293. Springer.

Suryatej Reddy Vyalla and Vishaal Udandarao. 2020. Memeify: A large-scale meme generation system. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 307–311.

Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. 2017. Combining knowledge with deep convolutional neural networks for short text classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2915–2921. 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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.