Cady Baltz (cmb180010)
4/8/23
CS 4395.001 Human Language Technologies
Reading ACL Papers

Title: Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network [1]

Author List:
- Bin Liang, Harbin Institute of Technology, Shenzhen, China
- Chenwei Lou, Harbin Institute of Technology, Shenzhen, China
- Xiang Li, Harbin Institute of Technology, Shenzhen, China
- Min Yang, Chinese Academy of Sciences, Shenzhen, China
- Lin Gui, University of Warwick, UK
- Yulan He, University of Warwick, UK
- Wenjie Pei, Harbin Institute of Technology, Shenzhen, China
- Ruifeng Xu, Harbin Institute of Technology, Shenzhen, China

Problem Summary:
The problem addressed by this paper is detecting satirical and ironic expressions in “multimodal messages.” By “multimodal message”, the paper is referring to images paired with a text description. An example of a sarcastic multimodal image would be a picture of a thunderstorm paired with the caption “What wonderful weather!”. This problem is particularly important today due to the prevalence of sarcasm on social media platforms. Having the ability to detect sarcasm would help improve the performance of sentiment analysis and opinion mining online. What makes this problem challenging to solve is that you must have an understanding of the information presented in different modalities, as well as how these two distinct modalities are connected to each other. Otherwise, you risk misunderstanding the user’s intent. For example, if you only perform sentiment analysis on the text in the example I provided previously, then you would think that the user is happy with the weather, instead of upset with it. After all, detecting sarcasm from text and images only is a challenging problem for even humans to solve, so solving it via machine learning is even more difficult.

Prior Work Summary:
Prior research had already addressed the problem of detecting multi-modal sarcasm in text and image data. One study approached the problem by concatenating the textual and visual features to fuse the information [2]. Several other studies attempted to instead implicitly fuse these features based on external knowledge [3], [4], [5]. Finally, the most recent approach was to build interactive graphs that model the relationships between the information from multiple modalities [6].

However, these prior works had several limitations. First, this paper pointed out how prior work only considered the image as a whole, rather than separating the key visual objects related to the text from the irrelevant ones. The researchers claimed that this would limit the ability of a model to learn visual information. Furthermore, the prior work does not consider how visual

information related to a sarcastic cue may be scattered across an image. Thus, there was no existing way to draw “intricate sentiment connections between text and image modalities” [1].

Unique Contributions:
In comparison with the prior work, this research is unique in that it tracks visual information in bounding boxes as opposed to considering the image as a whole. The paper includes several images providing examples of these bounding boxes and their corresponding labels. For example, in one of these images a tree was labeled as “bare tree” while the sky above it was labeled as “blue sky”.

This paper’s algorithm discriminates key visual objects from irrelevant ones with the hopes of improving learning of visual sarcastic cues. To accomplish this goal, the researchers developed a “cross-modal graph convolutional neural network” [1]. By using a cross-modal graph, they were able to explicitly link important visual information with the textual tokens that describe them. Creating this graph was a multi-step process.

First, their algorithm detects the important visual regions in the image. To accomplish this step, the researchers were able to build off of prior work to generate a list of visual attribute-object pairs [7]. Then, they assign edges linking these attribute-object pairs to object descriptors in the text using WordNet [8]. These edges can be assigned varying degrees of importance. Finally, they assign a sentiment relation to each of these edges using external knowledge from SenticNet, which they applied to attribute descriptors (most commonly adjectives) as well as textual words [9]. These sentiment relations can then be used to detect “incongruities of the cross-modal nodes in the graph” [1]. By this, it means that they are able to find when two connected nodes in their graph have opposing sentiments, which indicates the presence of sarcasm.

Evaluation:
To evaluate their work, the researchers used an existing multi-modal sarcasm detection benchmark dataset [3]. Each instance in this dataset has an English tweet with an image and a caption, as well as a label of whether or not it expresses sarcasm. They divided the data into a training and test set, with 19,816 instances in the training set and 2,409 instances in the test set. They included two layers in their experimental network, and set it to use the Adam optimizer. To prevent overfitting, they used a dropout rate of 0.1. Their batch size was 32, and their learning rate was 0.00002. Finally, they averaged their results over 10 runs with different random seeds to ensure they were statistically stable.

Then, they measured the accuracy, precision, recall, and F1-score of their results.These were the same metrics that prior researchers used with the same benchmark dataset [3]. They compared their metrics with twelve previous algorithms, and found that they outperformed all of them. For example, their accuracy was 87.55%, and the next highest-performing algorithm was 86.10%. Similarly, their F1-score was 84.16%, while the runner-up was 82.84%. Specifically, they compared their results with two image-based algorithms, which both produced accuracies under 70%. They also compared their results with five text-based algorithms, which all had

accuracies under 84%. Finally, they compared their results with five other multimodal algorithms, which had accuracies in the range of 83% to 86%.

Author Information:
Since the paper's publication in 2021, the authors have received 13 citations on Google Scholar [10]. While this is not an incredibly long list of citations, I believe that their work was important because it has many real-world applications, particularly in the realm of social media analysis. For example, a paper recently published in 2022 that cites this paper is called "Review on Sentiment Analysis and Polarity Classification of Sarcastic Sentences using Deep Learning in Social Media" [11]. I think it is exciting that other researchers were able to leverage this paper for polarity classification, as identifying polarizing issues on social media is important to understanding key issues in our world, especially in the realm of politics. Clearly, being able to detect sarcasm in images with text is essential to sentiment analysis in the online world today.

Of the seven researchers credited on this paper, Yulan He had the most citations on Google Scholar [12]. She has 11,463 citations in total, across 254 papers dating back to 2006. She is currently a professor at King's College in London focusing on sentiment analysis, and she has done extensive research on sentiment analysis on Twitter.

Conclusion:
I chose to read this paper because I have noticed in my own experience that sarcasm is often misunderstood online, especially by non-native speakers of a language. I was curious to learn more about how we may be able to use machine learning and natural language processing to aid in detecting sarcasm. I feel that this is an important problem to solve due to the prevalence of misinformation on social media. For this reason, I believe that the authors' contributions are important and incredibly relevant today.

References

[1] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

[2] Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1136–1145.

[3] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

[4] Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

[5] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and intermodality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

[6] Bin Liang, Hang Su, Rongdi Yin, Lin Gui, Min Yang, Qin Zhao, Xiaoqi Yu, and Ruifeng Xu. 2021b. Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 208–218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18-22, 2018, pages 6077–6086. IEEE Computer Society.

[8] George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

[9] Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 105–114. ACM.

[10] Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. Google Scholar.

[11] Kumar Bhadra, A., Shaila, S.G., Banga, M.K. 2022. Review on Sentiment Analysis and Polarity Classification of Sarcastic Sentences using Deep Learning in Social Media. In *Data Engineering and Intelligent Computing. Lecture Notes in Networks and Systems*, vol 446. Springer, Singapore.

[12] Yulan He. Google Scholar.