Understanding the Attention Model of Humans in Sarcastic Videos

Dipto Das, Md. Forhad Hossain, Anthony J. Clark
Department of Computer Science
Missouri State University
Springfield, Missouri, USA
Email: dipto175@live.missouristate.edu

***Abstract*—Sarcasm is a common part of human communication that has long been ignored by sentiment analysis researchers. Sarcasm is also an important aspect in entertainment industry for TV series, movies etc. Recently, some works have shown the applicability of multimodality (e.g., image and text) in sarcasm research from a sentiment analysis perspective instead of text only approaches. However, none of those studies harness video. We argue videos can be interesting to study to understand the nature of sarcasm on social media. We study how sarcastic videos can gain an individual's attention and popularity at large. We show how an AI agent can suggest areas that might gain a viewer's attention in a sarcastic video. Identification of both attention gaining areas (AGA) and objects contained in sarcastic videos can be compared with the AGAs and objects in previously successful/popular sarcastic videos. In this paper, we present an AI agent to identify the optimal AGAs and one empirical study of objects commonly shown in directed sarcastic video settings.**

I. Introduction

Sarcasm is a common part of human communication that requires deeper understanding of context, including verbal and non-verbal cues, during a conversation. Sarcasm is prevalent not only in human communication but also in popular TV series and movies. In this paper, we studied how artificial intelligence (AI) can be applied as the *guide* for understanding the attention model of humans in sarcastic videos.

Application of AI agents in recreating and understanding the attention model of humans in storytelling kind of contents is not a new concept [1], [2], [3]. However, to the best of our knowledge, no previous work has considered sarcasm as a mode of emotion in human attention model. In this work, we bridge the research gap in understanding attention model of affective videos by including sarcasm as a mode of emotion.

A recent qualitative study on how humans detect sarcasm [4] shows that an attention model exists for how human audience enjoy sarcastic videos. That means, an AI agent learns to identify the person or object that the audience look for in sarcastic videos, and they can serve as a *guide* for understanding the attention model of an audience.

We took a two-step approach. First, we tested if the agent can identify the focused region of videos instead of points. Second, we analyzed what are the objects that are mostly presented in well-directed sarcastic videos. Thus, our first approach is a deployable AI agent that can be validated with performance measures, whereas our second approach is an empirical study of sarcastic videos with the help of an AI agent.

The rest of the paper is organized as follows: section II discusses the relevant previous works and identifies research gap that we would try to address; section III describes the dataset preparation phase for this research; section IV focuses on the AI agents to recreate the attention models of sarcasm in point and region levels; section V discusses the empirical analysis of the sarcastic videos with respect to the objects shown in those. After that, we discussed our plan for future works followed by conclusion.

II. Related Works

Existing works related to this paper can be divided into two directions. First, works dedicated to understanding attention model of humans in affective contents (e.g., videos); second, works related to sarcasm dataset collection and utilizing them to train AI agents in order to achieve a particular objective.

A. Attention Model in Affective Contents

Understanding an attention model in affective contents is a well researched area from the affective computing and interaction design communities. Multimodal interaction has found its way into different aspects of our life. For example, TV advertisements (ads) or videos try to influence or engage viewers through catching their attention. Shukla et al. showed how videos in TV ads can be analyzed to understand their capability to emotionally influence viewers [3], [5]. They criticized previous works for not taking actual human labeling into consideration. They used a convolutional neural network (CNN) to identify emotions in a short TV ad [3]. They utilized gaze points recorded on a TV ad to understand the attention model of viewers [5]. They showed how visual context and attention, even without narratives, become the primary drivers of affect in videos. However, they did not consider sarcasm as a mode of emotion. Whereas Shukla et al. [3], [5] focused on identifying emotions with AI agents in human generated contents, Cardona-Rivera et al. [6] showed how systems can be used to generate narratives for multimedia contents with story telling approach comparable to humans. However, they also did not take sarcasm into account as a type of human emotion.

978-1-7281-4127-5/19/$31.00 ©2019 IEEE
DOI 10.1109/TransAI46475.2019.00022

B. AI Agents Learning Sarcasm

As we all understand, and as the existing literature suggest, context is important for detecting sarcasm [7], [8]. However, it is difficult to understand the context of a content, specially for the ones that come from unknown users on a platform like SNS or video sharing website that is common scenario for public posts.

Most studies utilize a unimodal approach to sarcasm detection. In fact, most use text as the only data in their machine learning models [9], [10], [11]. However, some recent works have shown how images can be used to train machine learning models to detect sarcasm [12], [13]. The use of multimodal data in sarcasm related research is a new concept. In fact, this idea was first mentioned without any implementation by Razali et al. [14] for the first time very recently. The first work showing superiority of this multimodal approach was done by Das et al. [15]. However, they did not include video data.

To the best of our knowledge, there is no work to date that considers sarcasm as a mode of emotion and computationally analyzes its nature in video data. In this paper, we try to bridge that research gap by analyzing attention model of humans in sarcastic videos and studying the applicability of video data in training an AI agent to recognize sarcastic cues. Thus, it will pave the way to build a system that can analyze and understand human attention model in sarcastic videos.

III. Dataset Preparation

We collected and prepared our own video-based dataset. Then we recorded gaze points and generated the corresponding gaze videos.

A. Video Data Collection

We collected 50 short sarcastic video clips from popular TV series: Friends (20), Silicon Valley (10), The Big Bang Theory (10), and Two and a Half Men (10). The numbers inside parentheses denote how many clips were taken from that particular TV series. All videos were collected from YouTube. We used self-annotation to label our dataset, as done by [16], [10], [9]. Videos were tagged with "sarcasm", when the authors agreed that the video contained sarcasm. To abide by the limitation imposed by the gaze recording software (discussed later in this section), we only used video clips up to 1 minute in length. However, we also ensured that the sarcastic incident in the video had enough context information. This led to a variation in the lengths of our videos ranging from 45 seconds to 1 minute.

B. Gaze Labeling of the Data

Like many supervised learning approach-based studies, we needed to label our dataset. In our case, these labels were gaze points of a person viewing the videos as determined by Gaze Recorder [17]. This software package allows users to configure several settings, including degrees of field of view (FOV), resolutions, gender, adaptive/non-adaptive extend time during static scenes, and how much change in frames would be enough to consider a frame as a new one. Figure 1 shows an example pair of frames from original and gaze labeled videos.

Fig. 1. Example pair of frames from original and gaze labeled videos. The original frame is from the TV series "The Big Bang Theory" where the female character is being sarcastic about the male character's enthusiasm about a movie franchise.

C. Locating the Gaze Point

We located gaze points (labels of our experiment) in the frames of gaze labeled videos by subtraction. For any frame after the first one, subtraction result of each frame and the previous one of that frame gives the gaze point. Though this approach works well for most cases, this faces issues when there is sudden and drastic changes in video frames. We call the result frame obtained in this way as "subtracted gaze frame", as shown in Figure 2. In heat map, Red denotes the region where the spectator gazed for a longer time, and green for short period of time.

D. Calculating the Gaze Point Coordinates

Fig. 2. Examples of differently shaped gaze areas.

After RGB value calibration, we understand that all the images contain mainly shades of three basic (RGB) colors. We devised a simple approach called first and last point finding. We scan each image row by row and keep the first and last point of the desired color. Then using those two points as endpoints of a diameter of a circle, we draw the circle. Then, we saved the center of such circle as the gaze point coordinates for an individual frame. In case of a frame having more than one gaze area, only the coordinates of the center of the region with largest area is saved.

E. Preparing Final Dataset

At this stage, the original videos and the gaze point videos had different number of frames due to Adaptive FPS where (a) there was not enough frame change from one frame to the next, and (b) ambient lighting. We discarded frames from the videos with larger fps keeping the ratio between the numbers of read frames from original and gaze video equal to 1.0. In end dataset, we have 31,307 frames in total. Each frame has a size of 1536 x 864.

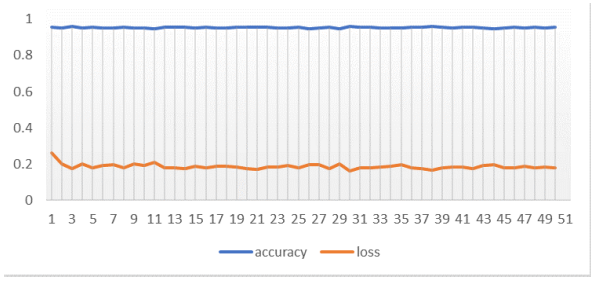


Fig. 3. Performance graph of semantic segmentation approach for recreating the attention model of sarcasm in videos.

IV. Methodologies

In this section, we discuss an AI agent to identify human attention locations in sarcastic videos. We used a semantic segmentation-based approach. A regression-based point level implementation of such an agent is described in [4]. Then we conducted an empirical study to see what types of objects are often close to human attention points in sarcastic videos.

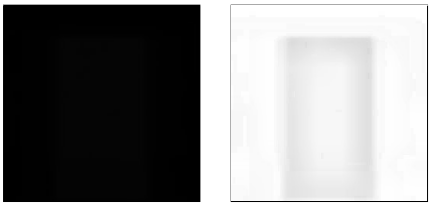
A. Semantic Segmentation Based Approach

While locating gaze points by subtracting subsequent frames, we removed the non-gazed areas of frames as background. Examples of the result of this subtraction is shown in Figure 2. The gazed regions of frames can be thought as one segment, and the non-gazed regions as another. In this way, we can model the problem of finding the attention gaining area (AGA) in sarcastic videos into a binary segmentation problem. We used black and white colors for binary labeling of pixels to have binary segmented images.

We used 25,308 frames from first 40 videos in our dataset for training and 5,999 frames from the rest 10 videos for testing. Besides, we also used standard data augmentation methods.

We used a U-net for the semantic segmentation task [18]. The inputs to the network were original video frames and outputs were binary colored segmented video frames. We optimized the network using Adam optimizer. We used binary cross entropy loss function and accuracy as metric for training the network. With 50 epochs, our performance metrics results were loss=0.1781 and accuracy=0.9542. Figure 3 shows the gradual improvement of these through training.

Though metric-wise these results seem promising, when we looked at the prediction of the network as images, shown in Figure 4(b), we can see the network identifies a larger area as the attention area that includes both the original attention gaining region and also some non-gazed or non-attention gaining background regions. The performance with this approach improved when we increased the epoch count from 20 to 50. Thus, increasing epoch count and the size of the dataset can benefit the semantic segment-based approach to improve its performance and the resultant AI agent to identify the AGA more precisely, i.e., narrow down the output AGA to smaller and more focused region.



(a) (b)

Fig. 4. Performance of semantic segmentation approach as images (a) original output (b) inverted output with increased contrast for better view.

B. Object Location and Distance Based Approach

Though the semantic segmentation based approach achieved numerically promising result, it still has room for improvement. At this stage, we modeled the experiment in a different way. We wanted to see what objects are often looked for in a sarcastic video. In other words, what are the objects that are often close to the gaze points on the sarcastic videos.

We passed the original frames in the dataset to a YOLO object detection model [19]. We identified the objects in the frames with their corresponding locations. YOLO can detect 80 different objects. If there is no object in a frame that can be detected by a YOLO network, we used a default object "unidentified" with default location (0, 0). If the frame had multiple objects that the YOLO model can identify, we saved names of all of those objects and their corresponding locations.

We calculated the center coordinates of gaze points as described earlier. Then, we calculated which object's location is the closest to the center of gaze point for a particular frame. We showed the top five objects that appear near the gaze center points in sarcastic videos, in Table I.

TABLE I
Top five objects that were closest to the gaze center points

Object	Counts of being the closest to gaze center
Person	25,662
Unidentified	1,247
Refrigerator	704
Sofa	535
Tie	412

As we can see, people are looked at most frequently in sarcastic videos. This contradicts with the finding of image-based sarcasm detection work by Das et al. [12] who found that people often appear in non-sarcastic images and non-human objects appear often appear in sarcastic images. This raises the question whether the persons being most looked for in sarcastic videos is specific to the sarcastic nature of the video or it is usual for any video. However, since the scope of this paper is limited to studying the attention model of humans in sarcastic videos and does not concern with distinguishing sarcastic and non-sarcastic video contents, collection of non-sarcastic videos and analysis of human attention model in those is out of scope.

V. Conclusion

We have collected a video-based sarcasm dataset, labeled the dataset with human viewers' gaze recordings, and conducted a two-step study on human attention model for sarcastic videos. Our first approach aims to identify attention gaining areas in videos with semantic segmentation. It achieved promising result given the small size of the dataset used to train the networks. Second approach conducts an empirical study on the contents of sarcastic videos of our dataset. Both dataset and the models are publicly available at http://bit.ly/sarcasm-attention. Since this is the first work on attention model in sarcastic videos, this work can serve as the benchmark for both dataset collection and model training in future work.

References

[1] C. Martens and R. E. Cardona-Rivera, "Generating abstract comics," in *International Conference on Interactive Digital Storytelling*, pp. 168–175, Springer, 2016.

[2] R. E. Cardona-Rivera and B. Li, "Plotshot: Generating discourse-constrained stories around photos," in *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2016.

[3] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian, "Evaluating content-centric vs. user-centric ad affect recognition," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 402–410, ACM, 2017.

[4] D. Das, "A multimodal approach to sarcasm detection on social media," Master's thesis, Missouri State University, Springfield, Missouri, USA, 7 2019.

[5] A. Shukla, H. Katti, M. Kankanhalli, and R. Subramanian, "Looking beyond a clever narrative: Visual context and attention are primary drivers of affect in video advertisements," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 210–219, ACM, 2018.

[6] R. E. Cardona-Rivera, "Cognitively-grounded procedural content generation," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[7] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter.," in *International AAAI Conference on Web and Social Media (ICWSM)*, pp. 574–577, 2015.

[8] B. C. Wallace, L. Kertz, E. Charniak, *et al.*, "Humans require context to infer ironic intent (so computers probably do, too)," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 512–516, 2014.

[9] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in twitter," *Language resources and evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[10] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714, 2013.

[11] J. Tepperman, D. Traum, and S. Narayanan, "" yeah right": Sarcasm recognition for spoken dialogue systems," in *Ninth International Conference on Spoken Language Processing*, 2006.

[12] D. Das and A. J. Clark, "Sarcasm detection on flickr using a cnn," in *2018 International Conference on Computing and Big Data (ICCBD)*, (Charleston, South Carolina, USA), 9 2018.

[13] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 1136–1145, ACM, 2016.

[14] M. S. Razali, A. A. Halin, N. M. Norowi, and S. C. Doraisamy, "The importance of multimodality in sarcasm detection for sentiment analysis," in *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pp. 56–60, IEEE, 2017.

[15] D. Das and A. J. Clark, "Sarcasm detection on facebook: A supervised learning approach," in *20th ACM International Conference on Multimodal Interaction (ICMI)*, (Boulder, Colorado, USA), 10 2018.

[16] M. Khodak, N. Saunshi, and K. Vodrahalli, "A large self-annotated corpus for sarcasm," *arXiv preprint arXiv:1704.05579*, 2017.

[17] Szymon Deja, "Gaze Recorder." https://sourceforge.net/projects/gazerecorder/, n.a. Online; accessed 15 April 2019.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.