Detecting and Explaining Emotions in Video Advertisements

ABSTRACT

Understanding and explaining abstract concepts in videos such as emotions is an unsolved problem. There is a large body of work that tries to predict human emotion or activity from videos, however, this is not sufficient. This paper describes a prototype system which, given a video and a model, returns an explanation for the model prediction. The explanation is determined via a two-step hierarchical process and is displayed as masked pixel regions over the most important frame.

KEYWORDS

Neural networks, Video classification, Explainability

ACM Reference Format:

. 2018. Detecting and Explaining Emotions in Video Advertisements. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Creative advertisements are ubiquitous, every service online has some flavor of content promotion to its users []. However, creating complex advertisements is not a not straightforward task, especially videos. It is manual work and producing large number of video advertisements can be prohibitively expensive. A lot of elements go into a video advertisement and foremost is the emotion the creative designer aims to evoke in its viewer. Irrespective of the advertising segment, engaging and attention grabbing videos need a good script, story line encapsulating a primary emotion.

While creating videos is challenging, even understanding or predicting the underlying emotion of an ad is an unsolved problem. Existing work is limited to detecting emotions from facial expressions in videos or images [1, 4]. However, in an advertisement, emotion could be motivated by a series of actions of events not just by humans but also different objects or characters []. Detecting and explaining what emotion does a sequence of clips capture can be instrumental for a creative designer to determine whether the video meets her requirements and can be used to target the intended audience. There is very limited work on *detecting emotions* from video advertisements [] and *explaining the classifier decision with visual cues*. In this work, we demonstrate that such a system can be built and can support a creative designer in understanding which scenes in a video contribute to its corresponding emotion.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

In this work, we show how existing video classification models can be paired with local explanation models to explain abstract concept such as emotion that can span multiple scenes in a video. Our system takes a video as an input, predicts the underlying emotion and explains the decision with supported scenes or visual segments from the video. Our system is designed for creative designers who need to glean insights into emotions captured by existing video advertisements and understand what parts of the video elude to the detected emotion.

2 RELATED WORK

This section explores previous research conducted on relevant areas.

2.1 Emotion classification in videos

Existing work on emotion classification in videos has primarily focused on developing models that can accurately recognize emotions in videos, such as happiness, sadness, anger, fear, and surprise. Different approaches have been used to classify emotions in videos, including feature-based methods, deep learning-based methods, and multimodal fusion methods. The most basic strategies involve expression detection from faces in specific frames. More advanced methods such as feature-based methods rely on extracting low-level features to classify emotions. Deep learning-based methods, on the other hand, leverage convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to learn high-level features from raw data. Multimodal fusion methods combine information from different modalities, such as visual, audio, and text, to improve emotion classification accuracy.

2.2 Explainability

In many applications, an explanation indicating how a prediction was reached is crucial for ensuring trust and transparency [3]. Large classification model predictions are untrustworthy and the opaqueness of the algorithms put a question mark on their validity. The desire to understand so-called black-box model predictions, usually in order to increase trust in a system, is currently a very attractive area of research [6]. Explainability aims to clarify the inner workings of the learning model and has been explored in depth for text and image classification [7]. It has been proposed that explainability should even be treated as a non-functional requirement in order to mitigate a system's transparency [9]. Prior to our work, research focused solely on tabular, text and image data [10, 16]. Our proposed system will employ and expand a standard image and text explainability algorithm; LIME [12].

2.2.1 LIME. Local Interpretable Model-Agnostic Explanations, or LIME, is an algorithm which can faithfully explain predictions of any image or text classifier. LIME proposes that the global decision boundary in a black box model can be approximated locally in the neighbourhood we want to predict. LIME generates fake points around this neighbourhood by making small perturbations (adding noise or removing features) to the instance we want to explain. The sampled instances are then weighed by their similarities and

a surrogate linear regression model is trained. We assume that this learned explanation approximation is locally (but not globally) faithful.

3 CREATIVE EMOTION EXPLANATION SYSTEM

Figure 1 shows a schematic overview of our proposed system. It shows the principle components and the inner workings of determining an explanation. The system should return an explanation alongside the model prediction to hopefully give users confidence in the model. A given video is initially processed with the aim of locating the most important frames (**find key frames**). This is discussed further in Section 3.2. The most important frame is then analysed further in order to locate which specific pixel regions are the most crucial.We achieve this by abstracting LIME which we explained in Section 2.2.1. This process is discussed further in Section 3.3.

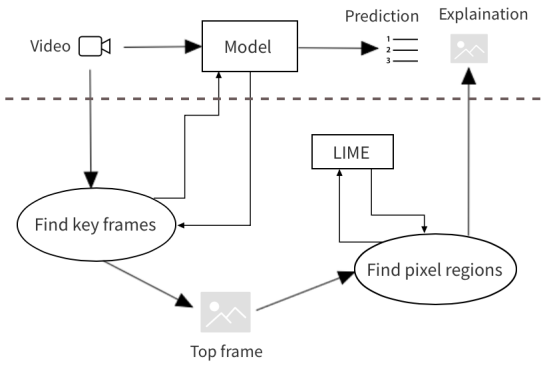

Figure 1: Overview of our system for explaining video classification outputs.

3.1 Classification model

Before we build our explanation system we briefly describe the video classification model we will use. We use previous work recently conducted as part of a Masters project [17]. The model aims to predict emotions in advertisement videos, specifically excitement and humour. This is achieved in three steps; feature extraction, sequence learning and finally, classification. Figure 2 shows an overview of the model.

To extract features from a video, our model uses CLIP (Contrastive Language Image Pretraining) [11]. CLIP is a primarily transformer-based model consisting of two encoders; one to embed text, the other to embed images. Our model will use the CLIP image encoder function to represent each frame in from video as a series of image patches. These are then split into a linearly embedded sequence of 512-dimensional patches [2]. Our model then sequentially learns features using Long Short-Term Memory (LSTM) networks. LSTMs contain memory block units in the recurrent hidden layer [14]. These blocks contain self-connecting memory cells storing the temporal state of the network. We use LSTMs as they retain

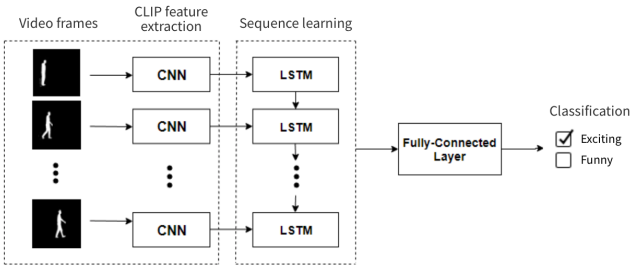

Figure 2: Overview of our video classification model.

past information longer making them more suitable for processing sequences of data [15].

3.2 Key frames

We find and order by importance a video's frames using a Leave One Feature Out (LOFO) approach [16]. LOFO calculates feature importance for any model and dataset by iterative removal of features to find global importance to the model. A simple approach would be to treat each frame in a given video as a feature. We would iteratively mask these out and by comparing model outputs we could rank each frame in order of importance. There are two main issues to this approach:

3.2.1 Efficiency. This approach is computationally expensive and inefficient. For a video k frames long, we require k passes of our model. To mitigate this, we will quantise the given video by splitting it into n frames. We utilise **skimage**'s structural similarity algorithm to find frames which are different from each other. Our hope is that the n frames returned are representative of the entire video. Generally, we found taking $n \approx 15$ for our short advertisement videos produced the best results.

3.2.2 Frame independence assumption. This approach assumes frames are independent of one another which is clearly not true. For any given frame in a video, closely surrounding frames are likely very similar and important. We take this into consideration by sliding a context window across each of the n key frames. Again, using **skimage**'s structural similarity, we compute the similarity of surrounding frames to the key frame in order to group together similar frames. We obtain frame sections (usually around 20 frames in length) around each key frame.

We are now able to treat each of these n frame sections as separate features, iteratively removing them and passing the resulting video through our model. By comparing the model outputs we rank the frame sections by order of importance for either excitement or humour. The top frame section can then be passed on to the next part of our system for further analysis.

3.3 Key pixels

Our aim here is to locate the most important pixel regions within a given frame section. Similarly to above, a naive approach could be taken by iteratively masking out groups of pixels and observing

the impact on our model predictions. This approach results in no change in model prediction and is very inefficient. Instead, we will expand LIME (detailed in Section 2.2.1) to be compatible with video inputs.

The first step is to create fake data by sampling around the frame section we want to explain. We accomplish this by masking out some pixels of our frame section. Rather than masking out stand alone pixels we will segment our frames into superpixel regions which, we hope, will hold some semantic and perceptual meaning. We utilise **skimage**'s Simple Linear Iterative Clustering (SLIC) algorithm to segment a given image into n superpixel segments [13]. To create a single fake data instance we mask out b of the n superpixel regions at random for each frame in the given frame section. The more sampled instances we create the more accurate our local approximation of the model is. With this in mind, we create 100 sampled instances.

To build our linear regression model, we need to extract features from our samples. As with our classification model, we will use CLIP (explained in Section 3.1). CLIP generates 512-dimensional vectors for each of our samples which we use to construct a surrogate model. The model utilises ridge regression to fit our data and locally approximate our black box model [5]. From this model we can extract which of the sampled instances have the biggest impact and rank our pixel segments accordingly. The final step is to overlay the most important segments over our key frame and return this as our explanation.

4 EXPERIMENTS

4.1 Dataset

To train and test our model and we use Pitts Dataset [8]. Pitts is a dataset created by Hussain et al. (2017) to support their research. It contains around 5000 video adverts with manually labelled annotations including sentiment and mood. The sentiment labels are exciting and funny.

4.2 Offline results

We show some example explanations returned by our proposed system.

(1) Vivo minions advert. Our model labels this video as exciting and funny. Figure 3 shows the explanation returned for the exciting label, likewise Figure 4 shows the explanation returned for the funny label. The outputs make sense; minions dancing together seems exciting and a minion falling over a chair is funny. Finally, Figure 5 shows the output of our system with adjusted parameters.

(2) Animal planet advert. Our model labels this advert as exciting and not funny. Hence, our system returns just one explanation. Figure 6 shows the most exciting key frame with the most important key pixel regions outlined. This seems to be a good output as an elephant holding a toothbrush is quite exciting.

Figure 3: Explanation for exciting label.

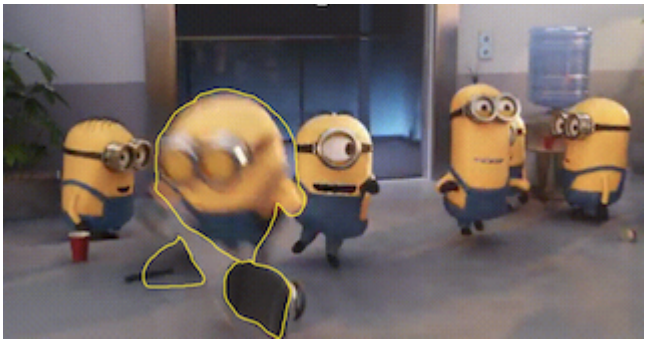

Figure 4: Explanation for funny label.

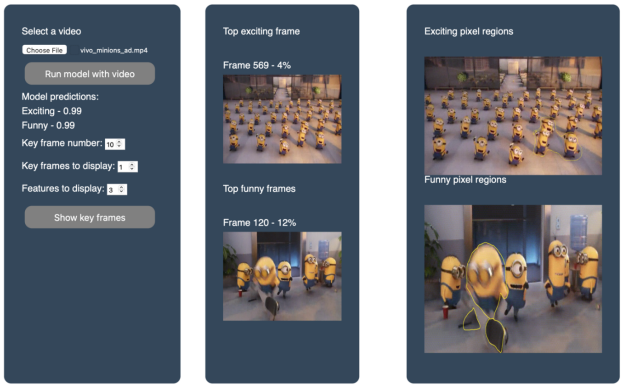

Figure 5: System output for minions advert.

5 CONCLUSION

Overall, our proposed system shows promise on standard video advertisement datasets. Each label and explanation returned makes sense from a human perspective aiding the user in understanding the inner workings of a model. Future work could include creating a synthetic video dataset with more detailed annotations in order to publish a standard way of evaluating video classification explanations. Additionally, text and audio clues could be used in parallel with out current system to give stronger explanations.

Figure 6: Explanation for exciting label.

REFERENCES

[1] Lei Chen, Su-Youn Yoon, Chee Wee Leong, Michelle Martin, and Min Ma. 2014. An Initial Analysis of Structured Video Interviews by Using Multimodal Emotion Detection. In *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems* (Istanbul, Turkey) *(ERM4HCI '14)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/2668056.2668057

[2] Hugging Face. 2020. BWorld Robot Control Software. https://huggingface.co/docs/transformers/model_doc/clip. [Online; accessed 08-February-2023].

[3] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. 2021. Explainable AI: current status and future directions. *CoRR* abs/2107.07045 (2021). arXiv:2107.07045 https://arxiv.org/abs/2107.07045

[4] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2594–2604. https://doi.org/10.18653/v1/D18-1280

[5] Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 1 (1970), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[6] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2022. *Explainable AI Methods - A Brief Overview*. Springer International Publishing, Cham, 13–38. https://doi.org/10.1007/978-3-031-04083-2_2

[7] Fatima Hussain, Rasheed Hussain, and Ekram Hossain. 2021. Explainable Artificial Intelligence (XAI): An Engineering Perspective. *CoRR* abs/2101.03613 (2021). arXiv:2101.03613 https://arxiv.org/abs/2101.03613

[8] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1705–1715.

[9] Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. 2019. Explainability as a Non-Functional Requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*. 363–368. https://doi.org/10.1109/RE.2019.00046

[10] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[12] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938 http://arxiv.org/abs/1602.04938

[13] scikit image. 2018. Segmentation.slic. https://scikit-image.org/docs/dev/api/skimage.segmentation.html. [Online; accessed 12-February-2023].

[14] Alex Sherstinsky. 2020. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306. https://doi.org/10.1016/j.physd.2019.132306

[15] Ralf C Staudemeyer and Eric Rothstein Morris. 2019. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586* (2019).

[16] Ugur Unal, Işıl Yenidoğan, Hasan Dag, and Aykut Cayir. 2020. Use Case Study: Data Science Application for Microsoft Malware Prediction Competition on Kaggle.

[17] Jin Wang. 2021. *Classifying Emotion Types in Advertisement Videos by a Deep Learning Approach*. Master's thesis. University of Glasgow.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009