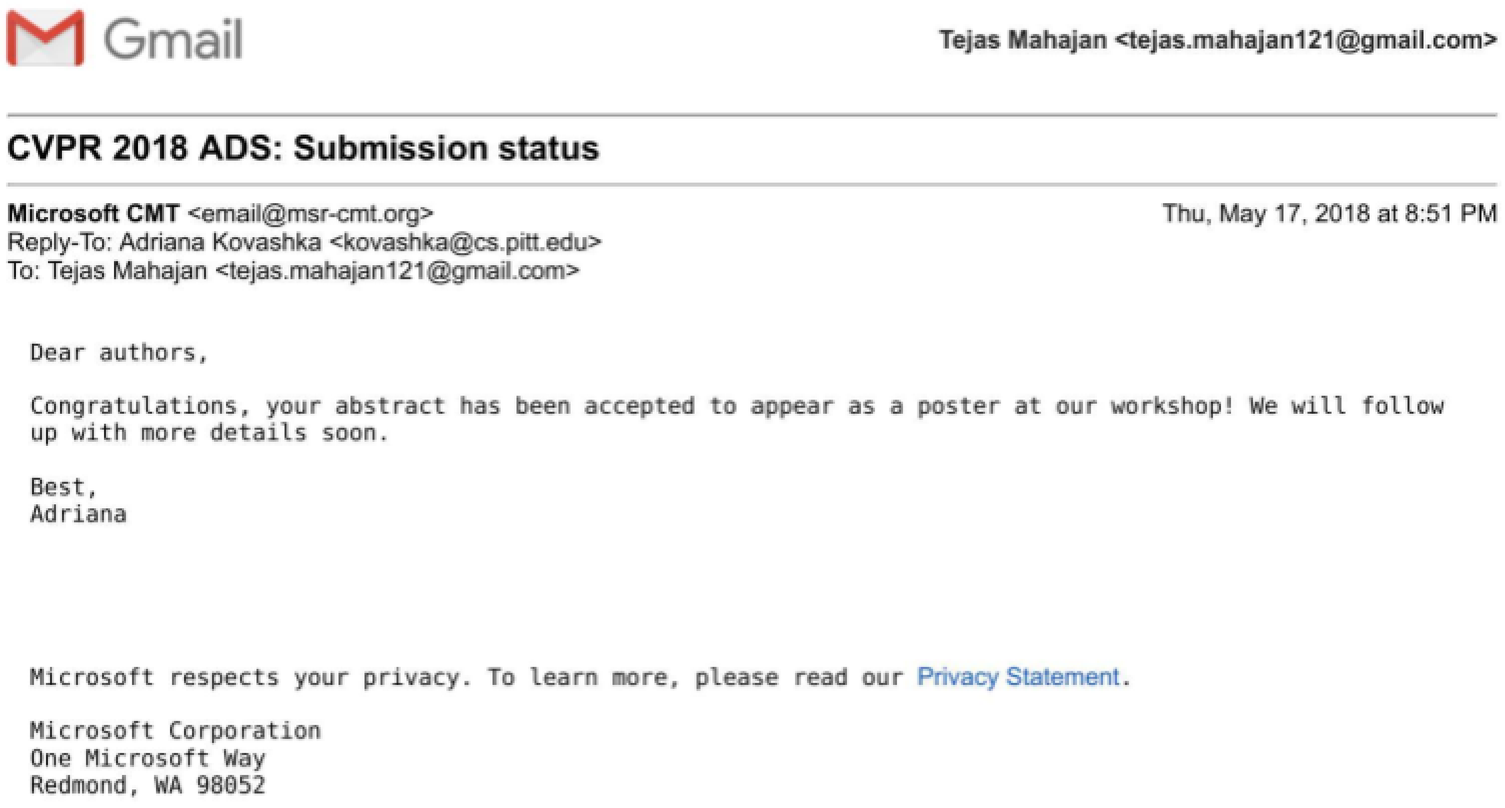

Tejas Mahajan <tejas.mahajan121@gmail.com>

CVPR 2018 ADS: Submission status

Microsoft CMT <email@msr-cmt.org>
Reply-To: Adriana Kovashka <kovashka@cs.pitt.edu>
To: Tejas Mahajan <tejas.mahajan121@gmail.com>

Thu, May 17, 2018 at 8:51 PM

Dear authors,

Congratulations, your abstract has been accepted to appear as a poster at our workshop! We will follow up with more details soon.

Best,
Adriana

Microsoft respects your privacy. To learn more, please read our Privacy Statement.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

While our paper was successfully accepted to appear at the workshop (Workshop Link), we couldn’t get a VISA to travel to the USA from India due to a long waitlist for visa interview, we couldn’t show up for the conference.

Understanding Emotional and Effective Components of Advertisements

Tejas Bhole[1]* Tejas Mahajan[1]* Nihar Gajare[1]* Niraj Pandkar[1]* Niranjan Pedanekar[2]
Shilpa Paygude[1]
[1]Department of Computer Engineering.
[1]Maharashtra Institute of Technology, Pune.
[2]Tata Research Development and Design Center, Pune.

Abstract

Digital advertising is a dominant form of advertising for any enterprise. Expressed in audio visual medium these ads use different types of visual rhetoric to convey their message, namely: common-sense reasoning, symbolism, and recognition of non-literal visual cues. These strategies make the task of analyzing and understanding ads using computer vision algorithms challenging. In this paper, we focus on two tasks in particular: prediction of effectiveness of an ad and, prediction of emotional components of the ad in terms of arousal and valence. We use the similarity between human generated text annotations to obtain an effectiveness score. Based on these effectiveness values, we present baseline results for effectiveness prediction using a novel Convolutional Neural Network (CNN) architecture. Similarly, we obtain arousal and valence scores from the human-annotated dataset and present a baseline prediction model for these scores.

1. Introduction

Advertisements, the strongest form of marketing communication are categorized in a variety of ways which includes style, target audience, medium, purpose and geographic scope. Advertisements existing in both visual and audio form drive a major portion of revenue for large Internet companies like Google and Facebook[3]. Furthermore, advertisements also help communicate messages of societal problems (e.g. ads highlighting issues like global warming and domestic violence) and awareness programs (e.g. ads portraying messages like smoking is bad for health).

Despite the advertising industry being one of the most cash-rich industries, it is also regarded as one of the least effective mainly because of the high rate of ineffective ads being created every day. Unlike many other businesses in the world which are governed by feedback loops, advertisements receive sparse and unreliable feedback, which can be misleading and biased to reviewers' opinions. This is primarily because advertisements are implicitly hard to understand and decode, because they use different forms of external knowledge to communicate the message.

This makes the task of understanding, decoding and evaluating advertisements very challenging and goes beyond the tasks of object detection, image captioning and simple visual question answering systems. Since people buying products for emotional reasons is far greater than people buying for logical reasons, we explore the emotional and effectiveness components of advertisements.

In this paper, we attempt to understand advertisements by utilizing the visual features of image and video ads to predict an effectiveness score and an arousal-valence score. On the one hand, we approach the task of modeling the sentiment portrayed in an advertisement in the form a valence and arousal score, while on the other hand, we attempt to determine which visual cues make an advertisement effective.

2. Related Work

The task of gaining a deeper understanding of advertisements is a challenging problem. It goes beyond just listing present objects or even producing a sentence about the image because ads are as much about how objects are portrayed and why they are portrayed so, as about what objects are portrayed.

Zaeem Hussain et al. in [7] analyzed the visual rhetoric of images. Semantic visual attributes describe images beyond labeling the objects in them. Zaeem Hussain et al. [7] presented a dataset dedicated to decoding visual rhetoric. The dataset consisted of $64,832$ image ads and $3,477$ video ads, and was collected using Amazon Mechanical Turk. Using these annotations, they presented baseline results for question-answering, topic classification and sentiment classification tasks for both image and video ads. They also

*Equal Contribution

reported results for funny and exciting annotation tags for video ads.

In a follow-up paper to the dataset published in [7], Kovashka and Ye in [11] used the symbolic mappings to predict the messaging in advertisements. They proposed a joint image-text embedding using external knowledge and presented a model called ADVISE (ADs VIsual Semantic Embedding) regarded as the current state-of-the-art for retrieving statements that describe an ad.

For predicting success of an online image ad before it is published, Michael and Jonathon[4] evaluated the potential of deep learning algorithms. They presented a dataset of over 260,000 ad images in a diverse set of categories and focused on smaller subset of automotive industry ads to build a regression model for click rate prediction. These tasks require a model to understand the base context that the given advertisement presents. Such understanding lines up with our task of predicting effectiveness of advertisements.

Kaliouby et al. in [9] presented a facial response dataset to ads. They also modeled the relationship of facial responses of users with ad liking and changes in purchase intent. They presented statistical machine learning algorithms for modeling such behavior. In this paper, we take a deep learning approach towards understanding ads. Moreover, they present results around narrow and short-term intent categories, in contrast to our paper.

Another task that requires contextual awareness in an ad is that of identifying candidate advertisement insertion points. Yadati et al. present the CAVVA model in [10] for this task. Similar to our work, they used arousal and valence to represent affect and to decide the transition between scenes and ad insertions.

We present our analysis and baseline results on this dataset as our tasks of understanding the sentiments evoked by ads and predicting their effectiveness scores is aligned with their tasks and the provided annotations.

3. Experimentation

3.1. Arousal Valence

Emotions play an extremely important role for the creation and conduction of successful advertising campaigns. Past studies have shown that emotionally charged events create powerful memories in peoples minds and these strong emotions drive people in making product purchases or contribute and support some social causes in some form. Therefore modeling emotions is a very important aspect for successful advertising.

On studying emotion classification in other domains, we found that classification of emotions into a set of fixed classes is not comprehensive because there are many advertisements capable of portraying contrasting emotions. Therefore, we explored two scientific concepts called "arousal" and "valence" that helped distinguishing one emotion from another. Here, arousal refers to the intensity of an emotion (how calming or exciting it is), while valence identifies the positive or negative character of the emotion. An ad with high or positive valence is supposed to signify emotions like joy, love, or pride; while low or negative valence ads portray concepts like death, anger, and violence. On similar lines, the more exciting, inspiring, or infuriating something is, the higher the arousal. Information that is soothing or calming produces low arousal.

3.1.1 Images

As we said earlier that we are utilizing the Ads Dataset published in [7], we realized that we did not have annotations for each advertisement with its corresponding valence and arousal score. To tackle this task we utilized two things; firstly the tagged sentiment annotation of every ad and secondly an emotion dictionary consisting of valence and arousal scores of adjectives obtained from analyzing large amounts of textual data. We mapped these two parts with a custom function to calculate a ground truth valence and arousal for each advertisement. 1.

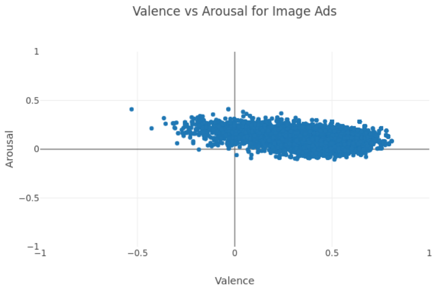

Figure 1. Valence vs Arousal distribution for Images

The arousal valence prediction for images uses two models; one for arousal prediction and other for valence prediction. Each of them is based on DenseNet [6] architecture initialized with ImageNet[8] weights 1. It is trained to find a mapping between the visual characteristics of the advertisement to a real valued score. While analyzing our results, we found that arousal models were easier to train than valence models since even for humans it's much harder to conclusively tell whether the ad is portraying a negative or positive character.

We achieved a squared error of 0.02 for valence prediction and 0.005 for arousal prediction on the validation set.

3.1.2 Videos

Every video advertisement in the dataset [7] has sentiment and adjectives annotations in the form of freeform text. We preprocess these text annotations to get a set of 296 adjectives. On similar lines, we mapped these adjectives to their corresponding valence and arousal scores in the emotion dictionary and used a custom function to calculate a ground truth valence and arousal score for each video advertisement 2.

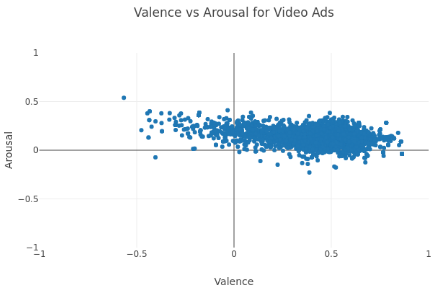

Figure 2. Valence vs Arousal distribution for Videos

Similar to images, the arousal and valence prediction for videos uses two models; one for arousal prediction and other for valence prediction. Both models use a DenseNet [6] based architecture that is combined with recurrent layers of LSTM [5] units for taking advantage of presence of temporal and spatial aspects in video ads. The model finds a mapping between the video ad sequence and the real valued arousal and valence scores. Similar to the observation in training image models, video arousal models were easier to train than valence models. This difference is attributed to the explanation that gauging excitement is easier for humans and likewise affects the annotations. For this task we achieved a mean squared error of 0.04 for valence prediction and 0.009 for arousal prediction for video ads. This final score is obtained by averaging the scores of all the individual sequence evaluations of a video ad.

3.2. Effectiveness

3.2.1 Images

Effectiveness Scores Since our dataset lacks direct effectiveness scores for image ads, we needed to find another metric which could be used as a proxy for them. For each image in the dataset, we have user annotated answers to two questions: 1) "What does this ad tell you to do?" and 2) "Why does it tell you to do it?". Thus, we had textual descriptions of the message given in the ad as rated by 3-5 human raters on Amazon Mechanical Turk. We use the reason annotation to generate a similarity score which could be used as a proxy for effectiveness of the ad. Because of the way this score is calculated, it favors images having a clear and concise message and gives low score to images that are confusing or unclear and verbose.
To calculate the similarity scores of sentences, we use fastText[1] vectors. These vectors use sub-word embeddings to deal with OOV words. Here, each word is represented as a bag of character n-grams. A vector representation is associated to each character n-gram; words being represented as the sum of these representations. This means that even for out-of-vocabulary words, we get a good approximation of a vector representation which is an important property for this task as the user annotations include brand names and catch phrases along with spelling mistakes.

For calculating similarity using word vectors, we use the cosine similarity formula:

$$similarity(A, B) = \frac{A \cdot B}{||A|| * ||B||}$$

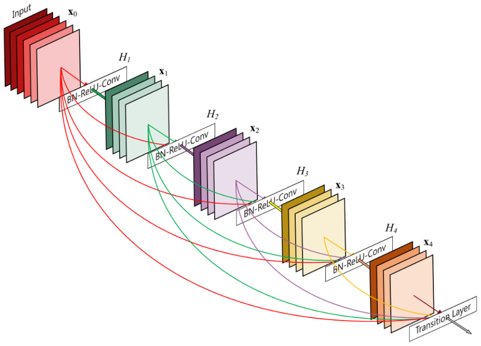

Figure 3. Dense Block Architecture from [6]

Model

DenseNet For extracting the visual features from image advertisements, we use a DenseNet [6]. It contains a series of dense blocks and transition layers. In dense blocks, each layer uses the feature maps of all the layers before it as inputs and its feature maps are used by all the successive layers as their inputs. The transition layers are connected between the dense blocks and are used to reduce the size of feature maps. DenseNets allow better feature propagation and reuse due to the densely connected convolutional layers and reduce the number of parameters. In our experiments, we use a DenseNet-121 implementation from Keras[2].

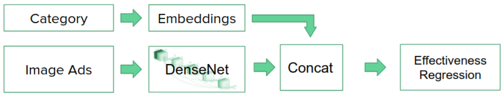

Figure 4. Representation of the Architecture for Effectiveness Regression

The effectiveness regression model for image advertisement uses the features extracted by the densenet concate-

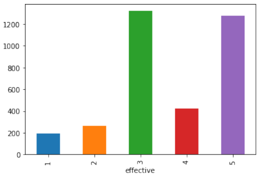

Figure 5. class distribution for videos

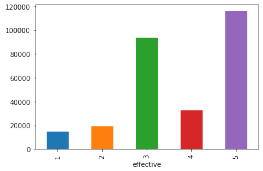

Figure 6. class distribution for extracted video samples

nated with the output of the category embedding layer as shown in 4. The final mse loss for this architecture is 0.160.

3.2.2 Videos

Effectiveness Scores The dataset contains direct annotations for video ad effectiveness. Each video has been rate on a scale of 1 to 5, where 1 indicates that the ad is not effective and 5 indicates that the ad is highly effective. Every video has been annotated by 5 human raters. The class distribution for videos as shown in 5 brings out the inherent class imbalance. This distribution after sampling the sequences from videos is depicted in 6. Since the set consists of multiple sequences per video, we tackle this skewed distribution by varying the samples collected from each video.

As the effectiveness data is annotated in 5 discrete classes, we handle this as a Classification task. For this we use softmax activation for final predictions.

Model Since video data has both spatial as well as temporal aspects, both static and dynamic features are extracted from the frames. It is necessary to take into consideration the sequences of the frames in a video in addition to the visual content of the frame itself.

We experiment with architectures that combine 2-dimensional convolutional neural networks and recurrent layers. This allows us to capture both spatial and the temporal aspect that videos inherently exhibit. Two architectures are experimented with for this task. The first architecture trains a fixed number of summarized frames from the videos. It uses a VGG-based Convolutional Neural Network with a combination of LSTM[5]. It achieves an accuracy of 38.90%.

The second architecture makes use of all alternative frames that are closely sequenced thus preserving the temporal aspect. This allows us to extract more data from the videos compared to video summarization. This architecture makes use of conventional convolutional networks as well as Convolutional LSTMs wherein the input transformations and recurrent transformations are both convolutional. This model predicts the effectiveness of an ad with 20.51% accuracy. This model learns the temporal aspect better because of the presence of closely sequenced frames.

4. Conclusion

In this paper, we present a new emotion recognition system leveraging arousal and valence scores for advertisements and how effectiveness as a metric is modeled for categorization of low and high effectiveness ads. We show how category embeddings helped improve results in detecting effectiveness of image advertisements. We also discover a much lower error on the arousal task compared to valence prediction. This can be explained by the fact that human annotators find it easier to comprehend arousal than valence as a single image or a video can produce both positive and negative emotions in the viewer. The architecture using continuous sequences of frames retains better temporal aspect than summarized frames that represent key ideas of the video ads when applied to recurrent layers. Since this is a work in progress, we are experimenting now on using symbolism detections and external knowledge for enhancing the models in better effectiveness prediction and emotion recognition systems.

References

[1] P. Bojanowskia, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 3

[2] F. Chollet et al. Keras. `https://keras.io`, 2015. 3

[3] J. D'Onfro. Google and facebook extend their lead in online ads, and that's reason for investors to be cautious. *https://www.cnbc.com/2017/12/20/google-facebook-digital-ad-marketshare-growth-pivotal.html*, 2017. 1

[4] M. Fire and J. Schler. Exploring online ad images using a deep convolutional neural network approach. *The 3rd IEEE International Conference on Smart Data, Exeter, UK, June 2017.*, 2017. 2

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 3, 4

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[7] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[8] D. Jia, D. Wei, S. Richard, L. Li-Jia, L. Kai, and F.-F. Li. Imagenet: A large-scale hierarchical image database. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[9] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6:223–235, 2015. 2

[10] K. Yadati, H. Katti, and M. Kankanhalli. Cavva: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 16(1):15–23, 2014. 2

[11] K. Ye and A. Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. *arXiv eprint arXiv:1711.06666*, 2017. 2