Multimodal Content Analysis for Effective Advertisements on YouTube

Nikhita Vedula*, Wei Sun*, Hyunhwan Lee†, Harsh Gupta*, Mitsunori Ogihara‡§, Joseph Johnson†§, Gang Ren§ and Srinivasan Parthasarathy*
*Dept. of Computer Science and Engineering, Ohio State University; †Dept. of Marketing, University of Miami;
‡Dept. of Computer Science, University of Miami; §Center for Computational Science, University of Miami
Email: {vedula.5, sun.1868, gupta.749, parthasarathy.2}@osu.edu, {aidenhlee, mogihara, jjohnson, gxr467}@miami.edu

***Abstract*—The recent advancement of web-scale digital advertising saw a paradigm shift from the conventional focus of digital advertisement distribution towards integrating digital processes and methodologies and forming a seamless workflow of advertisement design, production, distribution, and effectiveness monitoring. In this work, we implemented a computational framework for the predictive analysis of the content-based features extracted from advertisement video files and various effectiveness metrics to aid the design and production processes of commercial advertisements. Our proposed predictive analysis framework extracts multi-dimensional temporal patterns from the content of advertisement videos using multimedia signal processing and natural language processing tools. The pattern analysis part employs an architecture of cross modality feature learning where data streams from different feature dimensions are employed to train separate neural network models and then these models are fused together to learn a shared representation. Subsequently, a neural network model trained on this joint representation is utilized as a classifier for predicting advertisement effectiveness. Based on the predictive patterns identified between the content features and the effectiveness metrics of advertisements, we have elicited a useful set of auditory, visual and textual patterns that is strongly correlated with the proposed effectiveness metrics while can be readily implemented in the design and production processes of commercial advertisements. We validate our approach using subjective ratings from a dedicated user study, the text sentiment strength of online viewer comments, and a viewer opinion metric of the likes/views ratio of each advertisement from *YouTube* video-sharing website.**

I. Introduction

The recent integration of e-commerce infrastructures and web-scale multimedia distribution platforms has greatly increased the online presence of commercial advertisements and their impact on our society, while stimulating the development and deployment of innovative multimedia processing tools, content distribution schemes, and marketing behavioral models for digitally creating and disseminating persuasive advertisements with enhanced audience acceptance. Advertising along with product development, pricing and distribution forms the mix of marketing actions that managers take to sell products and services. It is not enough to merely design, manufacture, price and distribute a product. Managers must communicate, convince and persuade consumers of the competitive superiority of their product for successful sales. In this paper, we implemented a computational framework for the predictive analysis of the dependency patterns between content features and advertisement effectiveness metrics. This framework is aimed at creating a "formal grammar" between the advertisement content production and the effectiveness metrics to facilitate rational design and production approaches of effective commercial advertisements.

Video advertisements airing on television and social media are a crucial link of attracting customers towards a product. However, the factors contributing to advertisement effectiveness are rather complex and are a focus of study in marketing science and consumer psychology. In a landmark study, Lodish et al. [1] examined the sales effects of 389 commercials and found that in a number of cases advertising had no significant impact on sales. Many reasons can explain this finding. First, good advertising ideas are rare. Second, people often avoid or skip watching advertisements. Finally, even when an advertisement manages to hold a viewers interest it may not work because viewers may not pay close enough attention to the message embedded in it. All these factors make designing effective advertisements extremely challenging. Many ideas of how to create effective advertisements also come from the psychology literature [2], [3]. Positive or negative framing of an advertisement, the mix of reason and emotion, the synergistic interactions between music and narrative speech, the time arrangement of video shoots and the spatial organization, the type of message being delivered, the frequency of brand mentions, and the popularity of the endorser seen in the advertisement, all go into making an effective advertisement. But how these factors are combined to develop effective advertisements still remains a heuristic process.

The availability of large repositories of digital commercials, the advances made in content-based multimedia information retrieval, and the proliferation of user feedback/interaction mechanisms in social media, such as comments and like/dislike ratings provide us a new computational avenue for investigating the "success formula" of effective advertising. In this work, each advertisement clip is first divided into a sequence of short segments, from which we extract multimedia visual and auditory features to model the temporal timeline of the content. We also employ word-vector embeddings based on the text transcriptions

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.149

of the online advertisements as the "semantic" textural features. These individual feature dimensions are utilized to train separate neural networks to produce high level embeddings in their respective feature spaces, followed by a model fusing stage that learns a multimodal joint embedding for each advertisement. This multimodal joint embedding model is the basis for a binary classifier which predicts advertisement effectiveness metrics based on advertisement content features. We also analyzed the predictive patterns of advertisement effectiveness and have elicited a representative set of dependency patterns as a preliminary "effectiveness grammar". The novel methodological contributions of this work lie in the feature engineering and neural network architectural design. The primary, applied contributions of this work shed light on key questions of advertisement effectiveness from the related fields of multimedia, psychology, marketing, advertising, and television/film production.

II. Related Work

Previous work in targeted advertisement recommendation [4], [5] have utilized the visual and textual features of an advertisement along with user profile and click-through behavior for robust estimations of advertisement similarities and advertisement-viewer matching patterns. Content-based multimedia feature analysis is also crucial in the design and production of video commercials [6]. Besides signal processing and web content distribution applications, the use of temporal features is effective in media studies [7], movies [8] and music [9], because temporal shapes are easier to recognize and memorize for manual studies and are intuitive for computer aided explorations.

Our proposed framwork employs Long Short-Term Memory (LSTM) [10] and Deep Boltzmann Machine (DBM) [11] architectures to model the temporal structures in the content feature sequences, utilizing the flexible modeling capabilities of these neural network architectures for identifying patterns from hierarchical temporal resolutions and for modeling cross-dimensional dependencies. Related application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been proposed in [12] to generate a vector representation for videos and "decode" it using an LSTM model. Sutskever et al. applied a similar approach in the task of machine translation [13]. Venugopalan et al. [14] used an LSTM model to fuse video and text data from a natural language corpus to generate text descriptions for videos. Deep Boltzmann Machines have also been used to model multimodal data in various fields such as speech and language processing, image processing and medical research [11], [15].

III. Methodology

The multimedia temporal features employed in our proposed framework are extracted from the video, audio, and text transcriptions of commercial advertisements. These content dimensions are selected due to their easy integrability into existing workflows of video post-production and script creation. The feature extraction stage is followed by learning a multimodal embedding for predictive modeling.

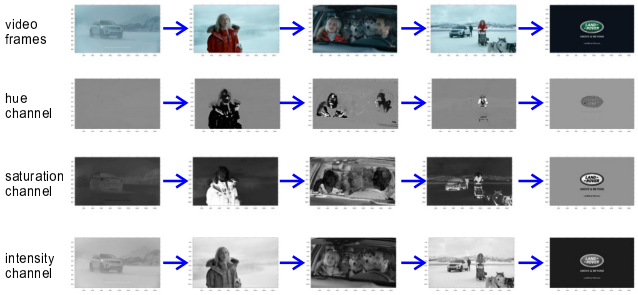

Figure 1. Multimedia timeline analysis of three video signal dimensions.

A. Feature Extraction

Visual/video Features: Video features of content timelines are extracted from the image features from sampled video frames as illustrated in Fig. 1. To speed up the signal processing algorithms, we sample one in ten video frames.We measure the hue, saturation and brightness values of each pixel in the sampled frames. The feature descriptors for each frame include the mean value and spatial distribution descriptors of the hue-saturation-brightness values of the constituent pixels. For measuring the deviations of these feature variables at different segments of the screen, the mean values of the screen's sub-segments and the differences between adjacent screen segments are calculated. We also segment the entire time duration of each video into 50/20/5 time segments as a hierarchical signal feature integration process and calculate the temporal statistics inside each segment including temporal mean and standard deviation, as well as the aggregated differences between adjacent frames.

Auditory/audio Features: Audio signal features include auditory loudness, onset density, and timbre centroid. We first segment the audio signal into 100 ms short segments and calculate the fast Fourier transform for each segment for analyzing its time-frequency energy distribution. This short segment length setting ensures appropriate analytical resolution in both the time domain and the frequency domain. Because the human auditory sensitivity varies with frequency, a computational auditory model [16] is employed to weight the response level to the energy distribution of audio segments. The loudness L_a is calculated as:

$$L_a = log_{10} \sum_{k=1}^{K} S(k)\eta(k)$$

where $S(k)$ and $\eta(k)$ denote the spectral magnitude and the frequency response strength respectively at frequency index k. K is the range of the frequency component. The loudness feature sequence is then segmented and temporal

characteristics like the mean and standard deviation in each segment are used as feature variables.

The audio onset density measures the time density of sonic events in each segment of $1/50^{th}$ of the entire video duration (typical segment length around 2 seconds). The onset detection algorithm [17] records onsets as time locations of large spectral content changes, and the amount of change as the respective onset significance. For each segment, we count onsets with significance value higher than a preset threshold and normalize the count by the segment length as the onset density. The timbre dimensions are measured from short time segments, similar to the loudness measurement above. The timbre centroid T_c for a short segment is calculated as:

$$T_c = \frac{\sum_{k=1}^{K} kS(k)}{\sum_{k=1}^{K} S(k)}$$

Textual Features: Our proposed framework employs word2vec [18] for semantic textual analysis. Word2vec provides a robust approach to word vector embedding, using a two-layer neural network with raw text as an input, to generate a vector embedding for each word in its vocabulary. The textual feature analysis module in our implementation first extracts and preprocesses the text transcription of each advertisement to obtain word tokens. We then use the 300-dimensional vectors pre-trained on the Google News Corpus [19] to obtain word2vec token embeddings.

B. Learning Multimodal Feature Representations

LSTMs for Sequential Feature Modeling: We employ an LSTM network to model the hierarchical temporal structures of the content features. At the core of the LSTM unit is a memory cell controlled by three sigmoidal gates: the input gate i decides whether the LSTM retains its current input x_t, the forget gate f enables the LSTM to forget its previous memory context c_{t-1}, and the output gate o controls the amount of memory context transferred to the hidden state h_t. The recurrences for the LSTM are defined as:

$$\begin{gathered} i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}) \\ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1}) \\ o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}) \\ c_t = f_t \odot c_{t-1} + i_t\phi(W_{xc}x_t + W_{hc}h_{t-1}) \\ h_t = o_t \odot \phi(c_t) \end{gathered}$$

where σ and ϕ are the sigmoid and hyperbolic tangent functions, $\odot$ denotes the product with the gate value and W_{ij} are the weight matrices containing the trained parameters.

We use an LSTM model with two layers to encode sequential multimedia features from Section III-A, employing a model of similar architecture for all the three input modalities. We generate a visual feature vector for temporal video frames of each advertisement, which forms the input to the first LSTM layer of the video model. We stack another LSTM hidden layer on top of this, as shown in Fig. 2,

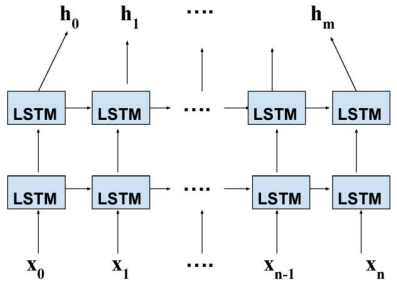

Figure 2. LSTM model with two hidden layers, each layer having 100 hidden units each, used for training individual input modalities.

which takes as input the hidden state encoding output from the first LSTM layer. Thus, the first hidden layer would create an aggregated encoding of the sequence of frames for each video, and the second hidden layer encodes the frame information to generate an aggregated embedding of the entire video.

Next, we similarly encode the temporal audio feature sequence as a two hidden layer LSTM model. For the textual features, we first encode the 300-dimensional vector embedding of each word in the advertisement text transcription through the first hidden layer of an LSTM model. A second hidden layer is applied to this encoding to generate a summarized textual embedding for each advertisement.

Multimodal Deep Boltzmann Machine (MDBM): We employ the Gaussian-Bernoulli variant of the classical Restricted Boltzmann Machine [20], vertically stacking the RBMs to form a DBM [11]. Our implementation uses three DBMs to individually model the video, audio and text features. Each DBM has one visible layer $\mathbf{v} \in R^n$ with n units, and two hidden layers $h_i \in \{0,1\}^m$ with m units for $i = 1, 2$. Our architecture forms an MDBM configuration as in Fig. 3 by combining the three DBMs and adding an additional layer on top. The joint distribution over the three input modalities is defined as:

$$P(\mathbf{v^c}, \mathbf{v^a}, \mathbf{v^t}; \theta) = \tfrac{1}{Z(\theta)} \textstyle\sum_{\mathbf{h}} exp(-V - A - T + J)$$

where $\mathbf{v^c}$, $\mathbf{v^a}$ and $\mathbf{v^t}$ denote the visual, auditory and textual feature inputs over their respective pathways of V, A and T, J represents the joint layer at the top, and $Z(\theta)$ denotes the partition function. Here,

$$V = \sum_i \frac{(v_i^c - b_i^c)^2}{2\sigma_i^2} - \sum_{ij} \frac{v_i^c}{\sigma_i} W_{ij}^{(1c)} h_j^{(1c)} - \sum_{jl} W_{jl}^{(2c)} h_j^{(1c)} h_l^{(2c)}$$
$$- \sum_j b_j^{(1c)} h_j^{(1c)} - \sum_l b_l^{(2c)} h_l^{(2c)}; \quad J = \sum_{lp} W^{(3c)} h_l^{(2c)} h_p^{(3)}$$
$$+ \sum_{lp} W^{(3a)} h_l^{(2a)} h_p^{(3)} + \sum_{lp} W^{(3t)} h_l^{(2t)} h_p^{(3)}, \text{and}$$

A and T have similar expressions as V.

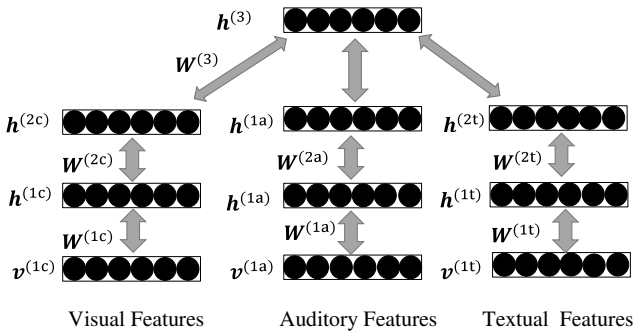

Figure 3. MDBM that models the joint distribution over the visual features, auditory features and textual features. All layers in this model are binary layers except for the bottom real valued layer.

$\mathbf{h} = \left\{h^{(1c)}, h^{(2c)}, h^{(1a)}, h^{(2a)}, h^{(1t)}, h^{(2t)}, h^{(3)}\right\}$ denotes the hidden variables, $\mathbf{W}$ denotes the weight parameters, and $\mathbf{b}$ denotes the biases. We first pre-train each modality-specific DBM individually with greedy layer-wise pretraining [21]. Then we combine them together and handle it as a multi-layer perceptron [22] for parameter tuning.

Inferring a Joint Representation: Once we obtain high-level feature embeddings ($\mathbf{h^V}, \mathbf{h^A}, \mathbf{h^T}$) from the final hidden layer of the three respective models of audio, video and text, we concatenate the three hidden layer embeddings in a layer called the *fusion layer*, which enables us to explore the correlation between the three kinds of features. In order to minimize the impact of overfitting, we perform dropout regularization [23] on the fusion layer with a dropout probability of 0.5. The combined latent vector is passed through multiple dense layers with non-linear activation functions (rectified linear unit in our configuration), before being passed through a final softmax layer to predict the output class of the advertisement. We assume a binary classifier for the advertisements with two classes: effective or successful, and ineffective or unsuccessful. Thus, the probability of predicting a class label y is:

$$p(y|\mathbf{x_V}, \mathbf{x_A}, \mathbf{x_T}) \propto exp(W[\mathbf{h^V}; \mathbf{h^A}; \mathbf{h^T}] + \mathbf{b})$$

where y denotes the class, $\mathbf{x_V}, \mathbf{x_A}, \mathbf{x_T}$ are the video, audio and text features of advertisement x, W is the weight matrix, [;] denotes the concatenation operation and $\mathbf{b}$ the biases.

IV. Experiments

A. Dataset and Evaluation Metrics

We evaluated our proposed methodology on a dataset of 200 advertisements crawled from the *YouTube* website, spanning representative product categories such as food and beverages, clothing, consumer electronics, health and medicine, movie trailers, and service. The ground truth for whether an advertisement is successful/effective or not was based on three independent metrics. First, a user study was conducted on all advertisements comprising 900 responses to 96 questions on categories such as brand presence and reliability, emotion expressed in the advertisement, emotion induced in the user, attention paid while watching, its influence on the user etc. Most questions solicited ratings ranging from 'Strongly Disagree' (rated 1) to 'Strongly Agree' (rated 7). We considered advertisements with a mean rating less or equal to 3 (averaged over all questions) as ineffective, while the rest as effective. The rating results were anonymized, and the experiment content and procedures were approved by the Internal Review Board of the respective organizations involved (IRB number 2015B0249), and meet with standard Nielson Norman Group guidelines. Second, we scraped the users' *YouTube* comments on each advertisement, and calculated the strength of the sentiment expressed in them using a tool called SentiStrength [24]. The sentiment strength scores ranged from -5 to 5, and all advertisements having a mean score above a threshold of 2.5 were considered effective, and the rest as ineffective. Third, the number of 'likes' i.e. explicit positive feedback received by an advertisement video on *YouTube* is a clear indicator of its popularity. We calculate the measure of the likes an advertisement receives over its total number of views as a measure of its effectiveness. All advertisements with a likes to views ratio above the mean of the ratio values received by all advertisements were categorized as effective.

B. Results

We compare our method against the baseline classifiers of Linear SVM and Logistic Regression, which take as input a concatenation of the visual, auditory, and textual features (Table I). We trained our neural network models over 15 epochs, minimizing the binary cross entropy loss using the Adam [25] optimizer, with a learning rate of 0.001. We randomly selected 150 advertisements for training and 50 for testing our method, and averaged our results over 50 runs. Compared to our other models, *the multimodal LSTM model achieved the best accuracy and an F1-score greater than* 0.8*, and the difference in accuracy is significant*. It also has a false positive rate of 0. Using a multimodal joint feature representation gives a huge advantage over any of the individual models. The text-only LSTM model that classifies advertisements based only on textual features appears to perform better than the video-only and audio-only models, whose accuracies are below 50%.

In addition, we removed all occurrences of brand name from the advertisement text and found the accuracy of the Text-only model to reduce to nearly 46%, while the accuracy of the multimodal LSTM dropped to about 73%. This confirms that the presence of brand name is crucial in determining advertisement success. We also inspected the impact of the position of the brand name in the advertisement text i.e. its occurrence at the start, middle or end of the advertisement, but did not find any significant difference in performance.

Table I. Classification results using various classifiers and ground truth metrics (best performance in bold)

Model	Ground truth	Accuracy	F1
Linear SVM	Comment sentiment	0.58	0.565
Linear SVM	Likes/visits	0.586	0.568
Linear SVM	User study	0.565	0.541
Logistic Regression	Comment sentiment	0.468	0.44
Logistic Regression	Likes/visits	0.55	0.529
Logistic Regression	User study	0.542	0.52
Multimodal DBM	Comment sentiment	0.60	0.66
Multimodal DBM	Likes/visits	0.61	0.71
Multimodal DBM	User study	0.66	0.64
Multimodal LSTM	Comment Sentiment	0.786	0.765
Multimodal LSTM	Likes/visits	0.8	0.769
Multimodal LSTM	User study	**0.83**	**0.81**
Video-only LSTM	Comment Sentiment	0.34	0.334
Video-only LSTM	Likes/visits	0.39	0.378
Video-only LSTM	User study	0.44	0.408
Audio-only LSTM	Comment sentiment	0.365	0.341
Audio-only LSTM	Likes/visits	0.401	0.4
Audio-only LSTM	User study	0.416	0.37
Text-only LSTM	Comment sentiment	0.478	0.445
Text-only LSTM	Likes/visits	0.49	0.468
Text-only LSTM	User study	0.52	0.52

Regardless of the evaluation measure adopted, the multimodal LSTM model significantly outperforms the other models, followed by the multimodal DBM model. Hence, for the purpose of evaluation and model selection we hypothesize that one can employ metrics derived from easily available online information such as likes, views and comments, rather than opting for the more expensive method of performing a user study. Having said this, we note that using the user study as ground truth shows a statistically significant performance improvement. For instance, the difference in accuracy between the multimodal LSTM models evaluated on the user study and on the ratio of likes per visits has a p-value of 0.04123 whereas the difference in F1-score between the two has a p-value of 0.0133, using a McNemar's paired significance test.

In addition, we seek to study the mapping relationships between multimedia attributes and advertisement success. For this purpose we choose a random forest classifier (yielding a classification accuracy of 0.55). In case of the visual attributes, the average intensity and average chroma for the first and second video segments are found to be important. The average saturation span and average chroma span for the fifth spatial zone i.e. the central zone of the screen have also been recognized as essential. In case of audio features, we obtain as important the onset spectrum strength, onset spectrum variation, and onset density dynamic range for the second and penultimate audio partitions.

In order to validate the importance of the above features, we performed experiments using our proposed model after excluding these particular audio-visual features from the input and using the user study as a ground truth metric. The textual input remained the same as earlier. We found a significant reduction in classification accuracy for the LSTM model, down to about 67%, while the accuracy of the DBM model went down to about 61%. We then utilized just the top important audio-visual features and the entire textual feature set as the input data. The accuracy of both models was found to reduce to 70% and 63% respectively, with the reduction also possibly due to loss of information via feature elimination. However, using only the important features still yields a reasonable classification accuracy. Thus, the above identified video and audio features are indeed essential in identifying and characterizing effective advertisements.

V. Discussion

Our findings show that the video segments in the first few seconds of an advertisement significantly mark out effective advertisements from ineffective ones. This concurs with the classical interpretations from marketing literature [1], which argue that viewers pay very little attention to details while watching advertisements, and thus effective advertisements must first attract the target viewers attention before delivering its main message. Our results also show that audio features of the second audio partition predicts advertisement effectiveness; agreeing with findings in marketing literature that stating relevant music can grab viewers' attention [26]. However what is new in our results is that the order in which the visual and auditory elements occur in an advertisement attracts and holds viewers' attention. Our finding that the video features of the central part of an advertisement is important for its effectiveness suggests that this is where the core message of the advertisement is embedded. We also find that brand mentions in an advertisement makes it more effective. This result finds support in the branding literature [27]. However, our finding that the temporal location of brand-mention is irrelevant is not supported by marketing literature. Scholars report that it is better to convey brand names and logos after the introductory attention-grabbing phase. We offer two reasons for our results. First, we do not control for the product category effect. For example, the temporal location of brand mention may vary based on if the advertisement is about a drug or a vacation. Second, we do not control for the brand's stage in its product life cycle. That is, where the brand name and logo appears may matter differently if we are dealing with a new brand or a mature brand. Our small sample size did not permit this analysis.

One unexpected finding from our predictive analysis is that text features explain advertisement effectiveness more than either video or audio features on average (see Table I). Recall that our text features come from the transcription of the voice-over in the advertisement. This means that viewers prefer advertisements in which brand benefits are conveyed verbally and with which they can connect emotionally. This finding concurs with the argument in [28] stating that though visual and auditory information grab our attention quickly,

verbal information is cognitively demanding, forming strong associations in our brain because when we hear words we spontaneously extract meaning from them.

In summary, our findings show that effective advertisements in our dataset exhibit strong creative design and production cues in visual, auditory and linguistic dimensions. Yet, they can be computationally analyzed and replicated in a rational pattern analysis process. Targeting audience acceptance and persuasive effectiveness, and creating effective advertisements demands strategic allocations of feature patterns from video, audio and textual script dimensions to form persuasive narrative patterns, e.g., first draw the viewer's attention in one content dimension and then deliver the brand message through another content dimension.

VI. Conclusion

We implemented a computational framework for modeling the temporal patterns of visual, auditory, and textual features extracted from multimedia content of online advertisement videos, and explored the correlations between these content features and various subjective metrics of advertisement effectiveness. The temporal patterns in multiple feature dimensions are modeled by three individual neural network models, which are fused together to yield a joint embedding for representing the pattern dependencies between different feature dimensions for predictive modeling and analysis. The predictive performance of our approach was validated using subjective ratings from a dedicated user study, the text sentiment strength of online comments and the likes/views ratio from *YouTube* web platform. The strong predictive performance obtained from the LSTM variants we proposed seem in line with its recent success in multimedia information retrieval and natural language processing [13], [14]. In the future, we plan to automate our feature engineering process utilizing CNNs trained on datasets such as ImageNet, and to develop models capable of generating more fine-grained predictive results rather than binary effectiveness indicators.

An expanded version of this paper is avilable at [29]. This extended version provides more detailed coverage of experimental configurations and predictive analysis results.

Acknowledgments

This work is supported by NSF awards NSF-EAR-1520870 and NSF-SMA-1747631.

References

[1] L. M. Lodish, M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens, "How tv advertising works: A meta-analysis of 389 real world split cable tv advertising experiments," *Journal of Marketing Research*, 1995.

[2] J. T. Cacioppo and R. E. Petty, "The elaboration likelihood model of persuasion," *NA-Advances in Consumer Research Volume 11*, 1984.

[3] J. Deighton, D. Romer, and J. McQueen, "Using drama to persuade," *Journal of Consumer research*, 1989.

[4] B. Wang, J. Wang, and H. Lu, "Exploiting content relevance and social relevance for personalized ad recommendation on internet tv," *ACM TOMM*, 2013.

[5] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, "Online video recommendation based on multimodal fusion and relevance feedback," in *ACM CIVR*, 2007.

[6] L. Elin and A. Lapides, *Designing and Producing the Television Commercial*. Pearson, 2004.

[7] J. Campbell, *The Hero With a Thousand Faces*. New World Library, 2008.

[8] B. Block, *The Visual Story: Creating the Visual Structure of Film, TV and Digital Media*. CRC Press, 2013.

[9] N. Graakjaer, *Analyzing Music in Advertising: Television Commercials and Consumer Choice*. Routledge, 2014.

[10] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[11] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial Intelligence and Statistics*, 2009.

[12] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *ICCV*, 2015.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014.

[14] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving lstm-based video description with linguistic knowledge mined from text," *arXiv preprint arXiv:1604.01729*, 2016.

[15] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *NIPS*, 2012.

[16] B. C. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[17] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[19] "Google news pretrained word2vec vectors," 2016. [Online]. Available: https://code.google.com/archive/p/word2vec/

[20] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," DTIC Document, Tech. Rep., 1986.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, 2006.

[22] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, 1990.

[23] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *JMLR*, 2014.

[24] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *JASIST*, 2012.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] G. C. Bruner, "Music, mood, and marketing," *Journal of marketing*, 1990.

[27] M. C. Campbell and K. L. Keller, "Brand familiarity and advertising repetition effects," *JCR*, 2003.

[28] K. L. Keller *et al.*, *Strategic Brand Management*. Upper Saddle River, NJ: Prentice Hall, 1998.

[29] N. Vedula *et al.*, "Multimodal Content Analysis for Effective Advertisements on YouTube," *ArXiv e-prints, https://arxiv.org/abs/1709.03946*, 2017.