# Multimodal Representation of Advertisements Using Segment-level Autoencoders

Krishna Somandepalli
Signal Analysis and Interpretation Laboratory
Los Angeles, CA, USA
somandep@usc.edu

Victor Martinez
Signal Analysis and Interpretation Laboratory
Los Angeles, CA, USA
victorrm@usc.edu

Naveen Kumar
Signal Analysis and Interpretation Laboratory
Los Angeles, CA, USA
knaveen87@gmail.com

Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory
Los Angeles, CA, USA
shri@ee.usc.edu

## ABSTRACT

Automatic analysis of advertisements (ads) poses an interesting problem for learning multimodal representations. A promising direction of research is the development of deep neural network autoencoders to obtain inter-modal and intra-modal representations. In this work, we propose a system to obtain segment-level unimodal and joint representations. These features are concatenated, and then averaged across the duration of an ad to obtain a single multimodal representation. The autoencoders are trained using segments generated by time-aligning frames between the audio and video modalities with forward and backward context. In order to assess the multimodal representations, we consider the tasks of classifying an ad as funny or exciting in a publicly available dataset of 2,720 ads. For this purpose we train the segment-level autoencoders on a larger, unlabeled dataset of 9,740 ads, agnostic of the test set. Our experiments show that: 1) the multimodal representations outperform joint and unimodal representations, 2) the different representations we learn are complementary to each other, and 3) the segment-level multimodal representations perform better than classical autoencoders and cross-modal representations – within the context of the two classification tasks. We obtain an improvement of about 5% in classification accuracy compared to a competitive baseline.

## KEYWORDS

multimodal joint representation; autoencoders; advertisements

**ACM Reference Format:**
Krishna Somandepalli, Victor Martinez, Naveen Kumar, and Shrikanth Narayanan. 2018. Multimodal Representation of Advertisements Using Segment-level Autoencoders. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3243026

*ICMI '18, October 16–20, 2018, Boulder, CO, USA*
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243026

## 1 INTRODUCTION

Video advertisements (ads) or TV commercials have become an indispensable tool for marketing. Media advertising spending in the United States for the year 2017 was about 206 Billion USD[1]. Companies not only invest heavily in advertising, but several companies generate revenue from ads. For example, the ad revenue for Alphabet, Inc., rose from about 43 to 95 Billon USD from 2012–2017. Considering the sheer number of ads being produced, it has become crucial to develop tools for a scalable and automatic analysis of ads.

In the seminal work of decoding advertisements, Williamson [28] states, "we can only understand advertisements ... by analysing the way in which they work". Previous research in advertising has shown a link between humorous or exciting content and ad effectiveness [13]. These studies have been limited in their ability to generalize their results, due to sampling methods or the sample sizes. An automatic and scalable analysis of ads might provide an understanding of the most valuable design choices that are needed to produce an effective ad.

In this context, the multimodal nature of ads can be leveraged to learn robust representations that can relate them to impact. Ads often contain video, audio and text (spoken language) modalities. With the advent of deep learning and big data, it is possible to obtain semantic attributes for these modalities beyond tasks such as labeling objects in images, or detecting music in audio. For example, features from a network pre-trained to detect high-level attributes like smiling [12], actions [25] or quality of an athletic action [18] may form better input features for learning representations.

In this work, we demonstrate that the audio-visual representations from multimodal autoencoders can improve the performance of classification tasks of ads compared to the unimodal features. Our experiments suggest that the representations from autoencoders trained on larger, unlabeled data not only improve classification accuracy on testing with unseen data, but also capture complementary information across the modalities. All models were trained on TensorFlow 1.4.0 and Keras 2.1.5 [6]. We have made the trained models, features and related code publicly available[2].

## 2 RELATED WORK

Hussain et. al. [11] analyzed the "visual rhetoric of ads" using image representations. This study has compiled and examined a dataset of

[1] www.statista.com
[2] https://github.com/usc-sail/mica-multimodal-ads

over 64,000 image ads, and 3,477 video ads. This data also includes human annotations for whether an ad is funny or exciting, among other labels such as topics, sentiment etc. In this work, we focus on the two binary classification tasks of whether an ad is 'funny' or 'exciting'. These tasks were chosen because they have the most number of human-labeled samples.

Research studies on ads prior to [11] have mostly focused on understanding the impact on the consumers. For example, [2, 5] used low-level image features such as intensity and color to predict click-through rates in ads. Similar work has been done in the audio domain. For example, the audio characteristics of an ad can be related to high-level concepts such clarity of the advertised message and its persuasiveness [9].

Audio-visual analysis of ads has historically focused on applications such as context-based video indexing [26], high-speed detection of commercials in MPEG streams [19], and retrieval of video segments for editing [29]. Although there has been some research in the field of Media Studies analyzing the audio-visual components of commercials [16], automatic understanding of high-level attributes (e.g., humor) has not been explored.

Multimodal deep learning [17] has shown promising results in obtaining robust representations across different modalities. A few prominent applications are audio-visual speech recognition in the wild [7], video hyperlinking [27], content indexing [22] and multimodal interaction analysis in the fields of emotion recognition and affect tracking [20].

The objective of multimodal representations is to obtain a feature vector that projects different unimodal representations onto a common space. See [4] for a survey on multimodal machine learning and its taxonomy. While there are several neural network architectures to learn these representations (e.g., sequence-to-sequence learning [3]), a popular approach is an autoencoder setup. An autoencoder with a linear decoder is akin to performing principal component analysis. Here, the model is trained to minimize the reconstruction error, and the features from the intermediate (middle) layer are used as common *representations* of the input modalities.

Autoencoders can be pre-trained on large domain-matched data, to obtain robust representations in an unsupervised fashion [17]. Motivated by this, we trained unimodal, and multimodal autoencoders on a large, unlabeled dataset to obtain joint representations for ads. Our experiments demonstrate the benefit of using such representations over audio-visual features.

## 3 DATASET

We use two datasets in this work. The primary dataset we consider is the video ads dataset released in [11]. This dataset originally had 3,477 ads from YouTube. However at the time of this work, only 2,720 ads were available. We consider two binary classification tasks: whether an ad is 'funny' or 'exciting'. Of the annotations provided in [11], 1,923 ads had the labels for 'funny', and 1,326 had labels for 'exciting'. Videos of duration less than 10s were excluded for this work. We refer to this dataset as *Ads-cvpr17*. The average duration of the ads in this dataset is 50.8 ± 27.9 seconds.

One of our goals in this paper is to show that the unsupervised autoencoder representations learned from a similar, larger, and unlabeled database of ads can provide better representations for the classification tasks. For this purpose we obtained ads shortlisted for the Cannes Lions Film Festival [1] through the years 2007–2017. This resulted in a dataset of 9,740 ads. The details of this dataset have been made publicly available[3]. Henceforth, we refer to this dataset as *Ads-Cannes*. The average duration of ads in this dataset is 90.3 ± 80.1 seconds.

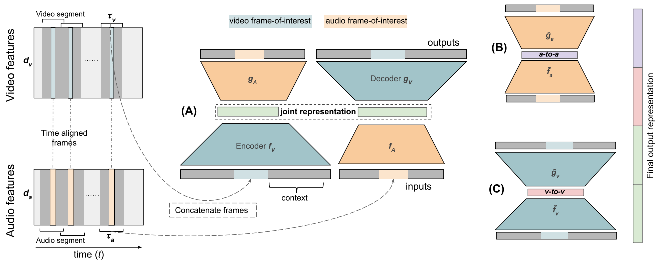

**Figure 1: Schematic diagram of segment-level autoencoders for (A) joint representation (B) audio: *a-to-a*, and (C) video: *v-to-v*. Input video and audio features of dimensions $d_v, d_a$ for an ad of duration $t$. Video and audio segments of length $\tau_v, \tau_a$**

## 4 METHODS

We first describe the choice of audio and video descriptors that we used in this work. We then present the procedure for training segment-level autoencoders using input frames to obtain unimodal and joint representations, followed by classification methods for the binary tasks.

### 4.1 Unimodal representations

*4.1.1 Video representations.* Prior work [11] has shown that features extracted from action recognition neural networks are effective toward automatic understanding on ads. Following this direction, we used the features from C3D network [25] which capture spatio-temporal features for recognizing actions in videos. It was pre-trained on the Sports-1M data [14] and fine-tuned on UCF-101 [23]. Global average pooling (GAP) in CNNs has been shown to improve localization and discriminability for action recognition tasks [30]. Hence, we perform GAP after the final convolution block in C3D network to obtain video features. We also replicated the results in [11] using the *fc7* features. These results were comparable to the results from GAP features.

*4.1.2 Audio representations.* Audio is an integral part of TV commercials because it can convey abstract concepts in a short duration [9]. We use features from an audio event detection network, *AudioSet* [10]. The actions being considered in the C3D framework are different from the events classified in *AudioSet*. Hence, these features likely provide complementary information to the video features. *AudioSet* uses a modified VGG [21] architecture to classify audio events with log-mel spectrograms of audio as input.

### 4.2 Segment level autoencoders

*4.2.1 Segments from frames.* The ads considered in this work are of variable duration. Hence, the features obtained from the

[3] To be provided after blind review

different modalities do not have a fixed length. Additionally, the pretrained networks used for video and audio representations each use different number of input frames. In this context, sequence representation is a popular approach (See [4] for a survey).

Sequence modeling approaches commonly use a RNN or LSTM architecture to learn the mapping between sequences in the two modalities, often explicitly modeling alignment with the use of attention [24]. In this work, we approximate a sequence representation by training an autoencoder at the segment level instead of sequence-to-sequence learning. We refer to the window consisting of a *frame-of-interest* and a fixed context as a *segment*. This allows us to jointly model the multimodal representation at a shorter time-scale (about 7 seconds) thus generalizing easily to other ads. This is a common data augmentation approach used in sequence learning for datasets with longer sequences [10].

Segments are generated as follows:

(1) Identify a frame-of-interest in video $x_v$, and include $\delta$ seconds worth of frames for forward and backward context.
(2) Identify the audio frame-of-interest $x_a$ corresponding to $x_v$
(3) Concatenate the frames from the resulting video segment of length $\tau_v$ to form $\mathbf{X}_v$. Similarly, concatenate the frames of audio segment of length $\tau_a$ to form $\mathbf{X}_a$. (See Fig. 1)

Note that although $\tau_v \neq \tau_a$ because of different sampling rates for each modality, these windows correspond to the same time duration. This helps ensure that the video and audio segments used in the autoencoders are time-aligned.

*4.2.2 Multimodal autoencoders.* We train three different autoencoders using the segment-level samples. A schematic diagram of the three networks is shown in Fig. 1. The encoders and decoders in these networks are a sequence of fully connected (FC) layers. For brevity, we represent the parameters (weights and biases) of the encoder and decoder parts of the network as $f(\cdot)$ and $g(\cdot)$ respectively. The representation vectors are denoted as $\mathbf{Z}$. Using this notation, the three architectures can be described as follows:

**(A)** Joint autoencoder representation: $(\mathbf{Z}_{joint})$

$$\min_{f_a, f_v, g_a, g_v} \; w_v \sum_{i=1}^{n} (\mathbf{X}_v^{(i)} - \mathbf{Y}_v^{(i)})^2 + w_a \sum_{i=1}^{n} (\mathbf{X}_a^{(i)} - \mathbf{Y}_a^{(i)})^2 \qquad (1)$$

$$\text{where } \mathbf{H}_v = f_v(\mathbf{X}_v); \mathbf{H}_a = f_a(\mathbf{X}_a)$$

$$\mathbf{Z}_{joint} = [\mathbf{H}_v, \mathbf{H}_a]; \mathbf{Y}_a = g_a(\mathbf{Z}_{joint}); \mathbf{Y}_v = g_v(\mathbf{Z}_{joint})$$

where, $n$ is the batch size, and $\mathbf{Y}_v$ and $\mathbf{Y}_a$ are the decoded video and audio respectively. We minimize the weighted mean squared error (MSE). The weights $w_v$ and $w_a$ are set such that they sum to 1.

In this architecture, the middle layer (*joint representation*: $\mathbf{Z}_{joint}$) is formed by concatenating the encoded video and audio layers. The corresponding video and audio segments are decoded using this $\mathbf{Z}_{joint}$. This is different from the "classical" autoencoders used for joint representation [27], where a single common layer is used to encode the outputs of both modalities (*shared representation*). One benefit of our network architecture over classical design is that we can handle missing data from either of the modalities.

Our network is comparable in design to the bidirectional symmetrical deep neural networks, *BiDNN* [27]. This network was trained for cross-modal translation where the weights are shared for the layers adjacent to the joint representation. But, unlike [27], our objective is to obtain joint representation and not perform translation between the modalities. Hence we do not tie the weights for the layers adjacent to the middle layer. For example, a video feature that encodes information about the visual action (lunging) can be very different from the event captured by audio feature (burping: human sounds). But, the combined information from these two modalities may indicate that the ad is funny. We perform additional experiments to compare our proposed system with classical autoencoders, as well as the *BiDNN*.

**Table 1: Unimodal performance: autoencoder representation vs. features (Acc., Accuracy (%); $\mathbf{F}_1$, F1 score (%)**

| Task: | Funny | | Exciting | |
|---|---|---|---|---|
| No. Samples | 1923 | | 1326 | |
| Majority class baseline (Acc.) | 58.00 | | 60.80 | |
| **Performance measures** | **Acc.** | **F$_1$** | **Acc.** | **F$_1$** |
| C3D features | 71.95 | 64.92 | 70.43 | 65.51 |
| *v-to-v* representation ( $\mathbf{Z}_v$) | 74.32 | 68.21 | 75.42 | 81.02 |
| Audioset features | 76.10 | 74.40 | 74.98 | 69.01 |
| *a-to-a* representation ( $\mathbf{Z}_a$) | 78.32 | 75.82 | 79.79 | 84.17 |

**(B)** Segment-level audio representation $(\mathbf{Z}_a)$, and **(C)** , segment-level video representation $(\mathbf{Z}_v)$: We construct symmetric autoencoders to obtain segment-level audio and video representations. While (A) captures the shared *joint* representation of the two modalities, these networks capture the information from the individual modalities.

The three networks were trained independent of each other. The network layers were designed such that the dimension of middle layers was consistent; i.e., $|\mathbf{Z}_{joint}| = 2|\mathbf{Z}_v| = 2|\mathbf{Z}_a|$. Finally, we concatenate the three vectors to obtain a multimodal representation of dimension $2|\mathbf{Z}_{joint}|$. These are further averaged across time to obtain a single feature per ad.

## 4.3 Classification of ads as funny or exciting

We use a support vector machine (SVM) classifier with representations from the autoencoders as features. The parameters of the SVM are tuned on a development set, and the results are reported using a 10-fold cross validation. We used McNemar's chi-squared test with continuity correction [8] to test for differences in both the performance of the predicted labels with the ground-truth, as well as between different models. We conducted pairwise tests (with Bonferroni correction for multiple comparisons) to test the performance of the models trained on different representations. A significant difference in this pairwise test of these outcomes would indicate that the different representations provides complementary information for the classification task.

# 5 EXPERIMENTS AND RESULTS

In all our experiments, the multimodal autoencoders were trained on the *Ads-Cannes* dataset in an unsupervised fashion (with respect to the classification tasks). The middle layer representations for the *Ads-cvpr17* dataset were then used for the classification tasks. First, we trained the autoencoders using the segments from 9,740 ads. We then performed experiments to analyze two claims: 1) representations from segment-level autoencoders trained on a larger dataset are better than the frame-level features, and 2) multimodal feature representations provide more information than the individual modalities – for classification tasks on the *Ads-cvpr17* dataset. In all cases, segment-level or frame-level features were averaged across time to obtain a single feature vector per sample. The binary classification tasks were peformed as described in Sec. 4.3.

## 5.1 Multimodal autoencoders

As described in Sec. 4.1, we obtained video features for every 0.64s using the pre-trained C3D network. We obtained the output from the final convolution layer, and performed global-average-pooling, resulting in a 512-dimension vector. We obtained audio features (128-dimensional) for

Table 2: Performance evaluation of unimodal vs. joint representations from the autoencoders

| Method/Task | Funny | | Exciting | |
|---|---|---|---|---|
| **Performance Measure** | **Acc.** | **$F_1$** | **Acc.** | **$F_1$** |
| Baseline (on complete data [11]) | 78.6 | – | 78.2 | – |
| BiDNN [27] | 78.01 | 75.71 | 79.02 | 82.17 |
| Classical autoencoder [27] | 71.54 | 61.92 | 77.10 | 81.02 |
| Best uni-modal ( $Z_a$) | 78.32 | 75.82 | 79.79 | 84.17 |
| Joint representation ( $Z_{joint}$) | 79.83 | 76.41 | 80.39 | 85.01 |
| **Multimodal representation** | **83.24** | 79.16 | **84.05** | 86.22 |

every 0.96s from the publicly available pre-trained *AudioSet* model[4]. Audio and video segments were constructed with these 'frame-level' features with a forward context and backward context of duration about 3.2 seconds. Each segment consisted of feature vectors corresponding to a duration of about 7s resulting in video and audio segments of size 11x512 (=5632) and 7x128(=896) respectively. We use a shift of 3 video-frames for obtaining consecutive video-segments. This resulted in a training set of 495,331 segments from the 9,740 ads.

The encoder and decoder parts of the autoencoders (see Sec. 4.2.2) were designed in a symmetric fashion. We use a shorthand notation to describe the network architecture. (N.B: due to the symmetric nature of the design, it is sufficient to describe the encoders)

**Video encoder** $f(X_v)$ :: **INP[(5632)] → FC[2816] → Dropout(0.2) → FC[1402] → Dropout(0.2) → FC[704] → FC[** $n_v$ **]**,
where **INP** = Input, **FC** = Fully connected layer with ReLu activation. $[\cdot]$ indicates the number of nodes in the layer.

**Audio encoder** $f(X_a)$ :: **INP[(896)] → FC[448] → FC[** $n_a$ **]**.

The number of nodes in the middle layer $n_a$ and $n_v$ was tuned over the set {16, 32, 64, 128, 256, 512, 1024} on the reconstruction loss of the training set. The networks were optimized with the Adam algorithm [15]. Training loss was monitored for parameter search because we train the models agnostic of the test dataset (i.e, *Ads-cvpr17*) Early stopping criterion (delta change of $10^{-6}$) on the training loss was used to terminate the network updates. The weights for the loss function, $w_v$ and $w_a$ were set to 0.75 and 0.25 respectively. Using these experiments we selected number of nodes for the middle layer, $n_a = n_v = 512$. The average reconstruction error (MSE) of the multimodal autoencoder was of the order, $10^{-3}$ for training set, and $10^{-2}$ for the test set. The MSE for the unimodal autoencoders (*a-to-a* and *v-to-v*) was of the order, $10^{-5}$ for training set, and $10^{-4}$ for the test set.

## 5.2 Performance evaluation

*5.2.1 Unimodal autoencoder representations.* Classification performance as measured by accuracy and F1-score shown in Table 1 supports our claim that the unimodal autoencoder representations perform better than the raw features – in the context of classifying whether an ad is funny or exciting. McNemar's chi-squared test[2] showed that the all the models outperform the majority class baseline. The audio-autoencoder representations (*a-to-a*) and the video-autoencoder (*v-to-v*) representations significantly outperformed the input audio and video features respectively ( $p << 0.01$). This suggests that (segment-level) autoencoders trained on larger, unlabeled data can provide 'better' representations for classification tasks.

*5.2.2 Multimodal autoencoder representations.* Performance evalutation results shown in Table 2 support our claim that multimodal information improves classification of whether an ad is funny or exciting. Note that the baseline results used in Table 2 are taken from [11]. These results are different when replicated on the data available (see acc. for C3D features in Table 1) . As such [11] is a competitive baseline since it had at least 700 more samples than the *Ads-cvpr17* dataset used in our paper. Our best uni-modal performance was comparable (if not better) than this baseline. We did not perform a significance test here due to lack of predicted labels from [11]. The concatenated multimodal representation (*final output representation* in Fig. 1) improved the classification accuracy by about 5% over the baseline.

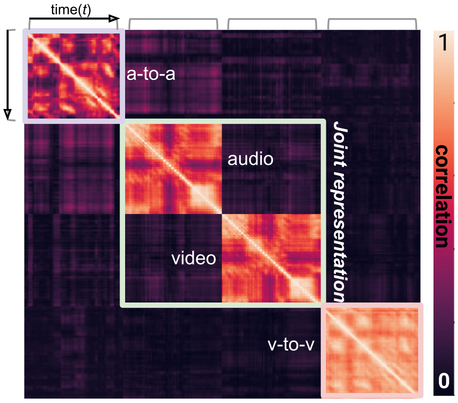

Figure 2: Similarity of representations with each other and across time in a sample

Pairwise McNemar's tests were conducted between the predicted outcomes for the two classification tasks from the four models (i.e., *v-to-v*, *a-to-a*, joint and multimodal representations). All pairwise tests except one (exciting: *a-to-a* vs. joint representation) survived multiple comparsion correction ( $p << 0.01$). The results from the chi-square test indicate that the models mislabel different samples during prediction. This suggests that the unimodal, joint and multimodal representations provide complementary information for the classification tasks considered.

The complementary nature of these representations was also examined by computing similarity of these vectors with each other and across time of an ad sample. Fig. 2 shows that the unimodal and joint representations are uncorrelated with each other (off-diagonal blocks). Additionally, the correlation of the individual representations across the time within an ad is similar (by visual inspection of the diagonal blocks). Finally, as shown in Table 2, the proposed segment-level autoencoders outperform the classification models trained on representations from *BiDNN*[27] and classical autoencoders.

## 6 CONCLUSIONS

In this paper, we propose segment-level autoencoders to obtain multimodal joint representations for analyzing advertisements. We show that representations obtained from training autoencoders on larger, unlabeled datasets are beneficial for classification tasks. Our experiments show that the unimodal and joint representations from the autoencoders provide complementary information, and improve classification performance over a competitive baseline, as well as compared to representations from classical and cross-modal autoencoders. In the future, we would like to allow our proposed approach to be more flexible in aligning segments across different modalities, which could be useful for representing asynchronous audio and visual events.

## 7 ACKNOWLEDGEMENTS

This work in part is supported by grants from NSF and Google. Victor Martinez is partially supported by the Mexican Council of Science and Technology (CONACyT).

[4] https://github.com/tensorflow/models/tree/master/research/audioset

[2] Reject $H_0$: marginal probabilities of each outcome are the same at $p < 0.01$

## REFERENCES

[1] [n. d.]. Cannes Lions. www.canneslions.com. [Online; accessed May-2017].

[2] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. 2012. Visual appearance of display ads and its effect on click through rate. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 495–504.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[5] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam. 2012. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 777–785.

[6] François Chollet et al. 2015. Keras. https://keras.io.

[7] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87–103.

[8] William D Dupont and Walton D Plummer. 1990. Power and sample size calculations: a review and computer program. *Controlled clinical trials* 11, 2 (1990), 116–128.

[9] Samaneh Ebrahimi, Hossein Vahabi, Matthew Prockup, and Oriol Nieto. 2018. Predicting Audio Advertisement Quality. (2018).

[10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 776–780.

[11] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1100–1110.

[12] Natasha Jaques, Weixuan'Vincent' Chen, and Rosalind W Picard. 2015. SmileTracker: automatically and unobtrusively recording smiles and their context. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1953–1958.

[13] Jean-Noel Kapferer, Gilles Laurent, et al. 1985. *Consumer involvement profiles: a new and practical approach to consumer involvement*. Technical Report.

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[15] D Kinga and J Ba Adam. 2015. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[16] Zizi Li. 2017. The âĂIJCelebâĂİ Series: A Close Analysis of Audio-Visual Elements in 2008 US Presidential Campaign Ads. (2017).

[17] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[18] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. 2014. Assessing the quality of actions. In *European Conference on Computer Vision*. Springer, 556–571.

[19] David A Sadlier. 2002. *Audio/visual analysis for high-speed TV advertisement detection from MPEG bitstream*. Ph.D. Dissertation. Dublin City University.

[20] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 2011. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*. Springer, 415–424.

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Cees GM Snoek and Marcel Worring. 2005. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications* 25, 1 (2005), 5–35.

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[25] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 4489–4497.

[26] S. Tsekeridou and I. Pitas. 1999. Audio-visual content analysis for content-based video indexing. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, Vol. 1. 667–672 vol.1. https://doi.org/10.1109/MMCS.1999.779279

[27] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. 2016. Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. In *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*. ACM, 37–44.

[28] Judith Williamson. 1978. *Decoding advertisements*. Vol. 4. Marion Boyars London.

[29] T. Zhang and C. C. J. Kuo. 2001. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing* 9, 4 (May 2001), 441–457. https://doi.org/10.1109/89.917689

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2921–2929.