

Story Understanding in Video Advertisements

Keren Ye
yekeren@cs.pitt.edu

Kyle Buettner
buettnerk@pitt.edu

Adriana Kovashka
kovashka@cs.pitt.edu

Department of Computer Science
University of Pittsburgh
Pennsylvania, United States

Abstract

In order to resonate with the viewers, many video advertisements explore creative narrative techniques such as “Freytag’s pyramid” where a story begins with exposition, followed by rising action, then *climax*, concluding with denouement. In the dramatic structure of ads in particular, climax depends on changes in sentiment. We dedicate our study to understand the dynamic structure of video ads automatically. To achieve this, we first crowdsource climax annotations on 1,149 videos from the Video Ads Dataset, which already provides sentiment annotations. We then use both unsupervised and supervised methods to predict the climax. Based on the predicted peak, the low-level visual and audio cues, and semantically meaningful context features, we build a sentiment prediction model that outperforms the current state-of-the-art model of sentiment prediction in video ads by 25%. In our ablation study, we show that using our context features, and modeling dynamics with an LSTM, are both crucial factors for improved performance.

1 Introduction

Video advertisements are powerful tools for affecting the public opinion, by appealing to the viewers’ emotions [15]. To achieve persuasive power, many ads explore creative narrative techniques. One classic technique is “Freytag’s pyramid” where a story begins with exposition (setup), followed by rising action, then climax (action and sentiment peak), concluding with denouement or resolution (declining action) [12].

In this work, we model the dynamic structure of a video ad. We track the pacing and intensity of the video, using both the visual and audio domains. We model how emotions change over the course of the ad. We also model correlations between specific settings (e.g., child’s bedroom), objects (e.g., teddy bear) and sentiments (e.g., happy). We propose two methods to predict *climax*, “the highest dramatic tension or a major turning point in the action” [10], of a video. Then we use them along with rich context features to predict the *sentiment* that the video provokes in the viewer. Our framework is illustrated in Fig. 1. Our techniques are based on the following two hypotheses which we verify in our experiments.

First, we hypothesize that the climax of a video correlates with dramatic visual changes or intense content. Thus, we compute optical flow per frame and detect shot boundaries, then predict that climax occurs at those moments in the video where peaks in optical flow

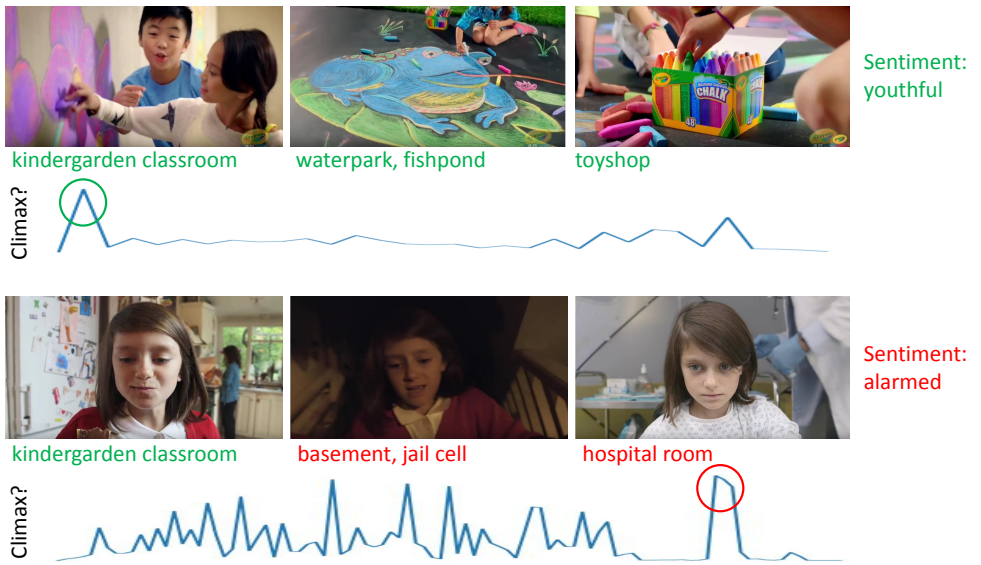


Figure 1: The key idea behind our approach. We want to understand the story being told in the ad video, and the sentiment it provokes. We hypothesize that the semantic content of each frame is quite informative and that we need to model the rising action to understand which temporal parts most contribute to the sentiment. We show the places recognized in the frames of two videos, as well as soft predictions about whether a certain frame corresponds to the climax of the video or not. While both videos start with images of children, which might indicate positive sentiment denoted in green (e.g. “youthful”), this positive trend only remains in the first video (indicated by places correlated with youthfulness, such as “toy shop”). In contrast, the second video changes course and shows unpleasant places (denoted in red) e.g. “basement” and “hospital room”. Because the climax in the second video occurs near the end, our method understands that it is these later frames that determine the sentiment (“alarmed”).

vectors or shot boundary changes occur. To measure dynamics in the audio domain, we extract the amplitude of the sound channel and predict climax when we encounter peaks in the amplitude. In addition to this unsupervised approach, we also show how to use the cues we develop as features, to predict climax in a supervised way. Both the unsupervised and supervised approaches greatly outperform the baseline tested.

Second, we hypothesize that video ads exploit associations that humans make, to create an emotional effect. We aim to predict the sentiment that an ad provokes in the viewer, and hypothesize that the setting and objects in the ad are greatly responsible for the sentiment evoked. We first extract predictions about the type of scene and type of objects in the ad, for each frame. We also hypothesize that the facial expressions of the subjects of the ad (i.e., the people in the ad) correlate with the sentiment provoked in the people watching it, so we also extract per-frame facial expression predictions. We treat sentiment prediction as a recurrent prediction task based on the scene, object, and emotion features, as well as features related to climax and standard ResNet [14] visual features.

To train our methods and test our hypotheses, we crowdsource climax annotations on 1,149 videos from the Video Ads Dataset of [17], and use the sentiment annotations provided.

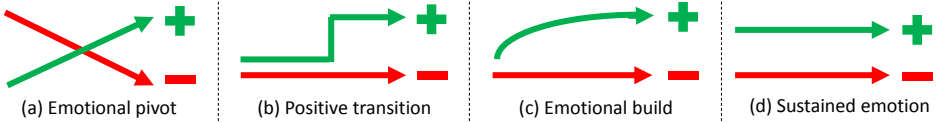


Figure 2: The “four archetypes of dramatic structure” in product ads [45] which motivate our approach. For PSAs, the roles of positive and negative sentiments might be reversed.

2 Related Work

Video dynamics and actions. Optical flow [8, 11, 12, 25, 29, 33] is a basic building block of video understanding. We use [29] due to its simplicity and reliable accuracy. Higher-level analysis of video includes human pose estimation [28, 32, 37] and action detection and recognition [5, 14, 39, 44]. Unlike these, optical flow does not capture semantics (such as the name of the action performed in a video). This is desirable in our case since a wide variety of activities can be exciting and climactic, so categorization is less useful. Anomaly detection [24] is also related, but rather than predicting what does not fit, we wish to predict how a video builds up and increases its dramatic content to create the climax.

Emotions. Researchers have been interested in predicting facial expressions and emotions for a long time [9, 10, 20]. Large datasets exist [4, 21, 26]. We train a facial expression model on [26] and apply it on faces detected in the video, as a cue for the viewers’ sentiment.

Movie and story understanding. We attempt to understand the stories told by video ads. Others have previously developed techniques for understanding various aspects of movies, such as their plot [27, 35] and the principal characters and their relations [41]. While there is no prior work on detecting climax in ads, some previous approaches model the tempo of other videos. For example, [23] use cues like “motion intensity” and “audio pace” to detect action scenes. [30] use the pacing of a movie to recognize its genre (action movies are faster-paced than dramas). [8] create video stories out of consumer videos, using story composition, dynamics and coherency, as cues. However, these works do not take emotions nor context such as scene and surrounding objects into account. We show semantic context features improve the performance of the unsupervised cues (e.g. “motion intensity”).

Advertisement and media understanding. There is a recent trend to attempt to understand the visual media with computer vision techniques. [16, 18, 22, 43] analyze the hidden messages of images, in news articles [18, 22] and advertisements [16]. [19, 40] examine the visual distinctions between people either running or voting in elections. We use the dataset of [16] for our study, and show that we greatly outperform their sentiment prediction model.

3 Approach

In The Advertising Research Handbook [45], dramatic structure has four prototypical forms, shown in Fig. 2 (based on [45] p.212). These structures depend on how positive and negative sentiment rises or declines. [45] examines product ads, and the changes in positive/negative sentiment are correlated with appearances of the brand. In public service announcements (PSAs), the role of positive/negative might be reversed, as PSAs often aim to create negative sentiment in order to change a viewer’s behavior. However, understanding the story of PSAs still depends on understanding the climax of (negative) sentiment. Thus, we first collect data (Sec. 3.1) and develop features (Sec. 3.2) that help us predict when climax occurs. We then develop features informative for sentiment (Sec. 3.3). We finally describe how we use these features to predict the type of sentiment and occurrences of climax (Secs. 3.4 and 3.5).

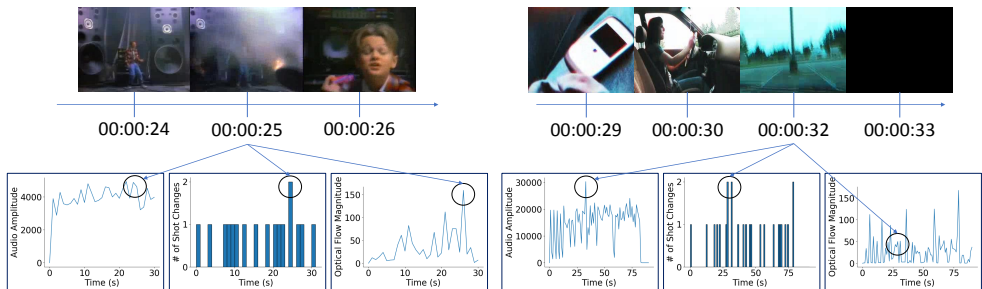


Figure 3: The audio, shot boundary frequency, and optical flow plots for two videos, along with frames from the videos corresponding to climactic points. The first video shows an “explosion” around the 25th second, and the second shows a car crash around the 32nd second. The circles correspond to the timestamp of the frames shown. In the first video, climax is detected well in each of the three plots. In the second, shot boundaries and audio are informative, but optical flow is not.

3.1 Climax and sentiment data

We use the Video Ads Dataset of [L6]. It contains 3,477 video advertisements with a variety of annotations, including the sentiment that the ad aims to provoke in the viewer. We collected climax annotations on a randomly chosen subset of 1,595 videos from this dataset, using the Amazon Mechanical Turk platform. We restricted participation on our tasks to annotators with at least 98% approval rate who submitted at least 1000 approved tasks in the past. We submitted each video for annotation to four workers. Each was asked to watch the video and could choose between two options, “the video has no climax” or “the video has climax.” If the latter, the worker was asked to provide the minute and second at which climax occurs (most videos are less than 1 min long). To ensure quality, annotators were also asked to describe what happens at the end of the video. Some of the videos in [L6]’s dataset were not available, so the annotators could also mark this option. We ended up with 1,149 videos that contain climax annotations. We manually inspected a subset of them and found the timestamps were quite reasonable. The descriptions of what happened at the end were often quite detailed. We will make this data publicly available upon publication.

3.2 Climax indicators

We first analyze the dynamics of the video, using both visual and audio channels. We plot time on the x-axis, and measurement of dynamics/activity on the y-axis (Fig. 3). We consider three indicators of rapid activity: the amplitude of audio signals, the occurrence of shot boundaries, and the magnitude of optical flow vectors between frames.

In particular, we extract these features and portray them as follows:

- **The audio amplitude a^k** , which is the max amplitude of audio for the k -th frame. We first extract the sound channel from the video, take a fixed number of samples from the sound wave per second, then compute the max across the samples for that frame.
- **The shot boundary indicator**, which is equal to 0 or 1 depending on whether a shot boundary occurs in the k -th frame. We use [2] for shot boundary extraction. In order to obtain more informative cues, we vary the parameters of [2] to get five 0/1 predictions per frame and use this 5D prediction b^k as the representation for the k -th frame. To generate the plot in Fig. 3, we aggregate information over all frames in a given second.

- **The optical flow magnitude o^k** , which is computed as $\frac{1}{W*H} \sum_{i=1}^W \sum_{j=1}^H \sqrt{u_{i,j}^k{}^2 + v_{i,j}^k{}^2}$ where $u_{i,j}^k$ and $v_{i,j}^k$ are the horizontal and vertical optical flow components for each pixel (i, j) in the k -th frame. We use [49] to extract optical flow vectors.

3.3 Sentiment indicators

The Advertising Research Handbook [45] describes the dramatic structure of ads as closely depending on the emotion of the video. One type of structure (Fig. 2) is the “emotional pivot” where an ad starts with negative sentiment, which declines over time, to make room for increasing positive sentiment. The “emotional build” involves a gradual increase and climax in positive sentiment. Thus, the sentiment is equally crucial to understanding the story of the ad video as the climax. Since an ad targets an audience and wants to convince the audience to do something, it is the viewer’s sentiment that matters the most.

[46] contains annotations about what sentiment each ad video provokes in the viewer, collected from five annotators. These annotations involve 30 sentiments, both positive (e.g., cheerful, inspired, educated), negative (e.g., alarmed, angry) and neutral (e.g., empathetic). [46] also includes a baseline for predicting sentiment, using a multi-class SVM and C3D features [48]. The authors extract features from 16-frame video clips, then average the features. Thus, their model does not capture the dynamics and sequential nature of the video. We hypothesize that if we model how the content of the video changes *over time*, and consider the *context* in which the sentiment in the video is conveyed, we would be able to model sentiment more accurately. We model sentiment with the following intuitive context features:

- **The setting in each frame of the video**, i.e. the type of place/scene. Let $\mathbf{vp} = \{p_1, \dots, p_{365}\}$ be the vocabulary of places in the Places365 dataset [46]. We use a pre-trained prediction model from [46] to obtain a 365D vector $\mathbf{pl}^k = [l_1^k, \dots, l_{365}^k]$, where l_i^k is the probability that the k -th frame exemplifies the i -th place.
- **The objects found in the video**. Let $\mathbf{vo} = \{c_1, c_2, \dots, c_{80}\}$ be the vocabulary of the COCO object detection dataset [22]. We use the model of [45] trained on COCO to get the objects in a frame. We then use max-pooling to turn the detection results into an 80D fixed-length feature vector $\mathbf{ob}^k = [s_1^k, \dots, s_{80}^k]$, where s_i^k is the maximum confidence score among multiple instances of the same object class c_i in frame k .
- **The facial expressions in the video**. We observed that the overall sentiment that the video provokes *in the viewer* often depends on the emotions that the *subjects* of the video go through. For example, if a child in an ad video is initially “happy” but later becomes “sad,” the sentiment provoked in the adult viewer might be “alarmed” because something disturbing must have happened. Thus, we also model emotions predicted on faces extracted per frame. We first detect the faces using OpenFace [3]. We then extract the expression of each face using an Inception model [44] trained on the AffectNet dataset [26]. Two types of results are predicted: (1) the probability distribution among the eight expressions defined in AffectNet, and (2) the valence-arousal values for the face, saying how pleased and how active the person is (in range -1 to +1). We average the face expressions (10 values) for all faces detected in the k -th frame, to get the 10D final representation \mathbf{fa}^k .
- **The topic of the ads**. [46] defines a vocabulary of 38 topics in the ads domain and also provides annotations for these topics. We hypothesize the overall sentiment that the video provokes is related to the topic the ad belongs to. For example, “sports” ads usually convey “active” and “manly” sentiments, while “domestic violence” ads often make people feel “sad”. Thus, we designed a multi-task learning framework with two objectives: one for the topic and the other for the sentiment prediction, hoping the

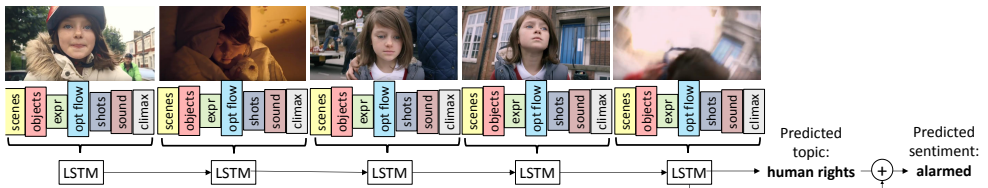


Figure 4: Our dynamic context-based approach. The last frame shows an explosion.

topic prediction can help the prediction of sentiments. We first use the video-level feature (the last hidden state of the LSTM) to predict the 38D topic distribution, then concatenate this 38D vector with the video-level feature to predict the sentiment. The idea is described in Fig. 4.

We also use features from the last layer of a ResNet trained on ImageNet [14, 51].

3.4 Unsupervised climax prediction

We can directly predict that climax occurs at times which are peaks in terms of shot boundary frequency, optical flow magnitude, or audio amplitude. Since the shot boundary frequency can be the same for many timeslots, we look for the longest sequence of timeslots which contain at least one shot boundary and predict the center of this “run” as a peak. Optical flow magnitudes and audio amplitudes are compared on a second-by-second basis. We extract the top- k maximal responses from each plot, predict these as climax, and evaluate the performance in Sec. 4.3.

3.5 Supervised prediction

We predict climax using an LSTM (with 64 hidden units) that outputs 0/1 for each frame, where 1 denotes that the frame is predicted to contain climax. The frame-level features used are ResNet features (2048D), optical flow magnitude o^k (1D), the shot boundary indicator b^k (5D), the sound amplitude a^k (1D), the place representation pl^k (365D), the object representation ob^k (80D), and the facial expression feature fa^k (10D).

For the sentiment prediction task, we also use an LSTM with 64 hidden units. We use the same frame-level features as the climax prediction. Moreover, we also add the predicted climax (1D) as extra information. Ads topics are used as both an additional loss/constraint and an extra feature for the sentiment prediction (see Fig. 4).

3.6 Discussion

The advantages of our approach are as follows. First, the distribution of object, place, and facial expression probability vectors is much lower-dimensional than ResNet features, so given the limited size of the Video Ads Dataset (3,477), formulating the problem as learning a mapping from objects/scenes/facial expressions to sentiments/climax is much more feasible. The optical flow, shot boundary, and sound features are also very low-dimensional, and have clear correlation with the presence of climax. Further, understanding the sentiment of a video and its climax are related tasks. Thus, it is intuitive that climax predictions should be allowed to affect sentiment prediction; this is the idea shown in Fig. 1 where we use climax to select the part of the video which affects the elicited sentiment the most. We show in Sec. 4.4 (Table 4) that our semantic/climax features outperform the ResNet features, and the combination of the two achieves the strongest performance.

4 Experimental Validation

We first describe our experimental setup and training procedure, then present quantitative and qualitative results on the climax and sentiment prediction tasks.

4.1 Evaluation metrics

For the climax prediction task, we use the recall of the top- k prediction ($k = 1, 3$) to measure performance. Since exactly matching the ground-truth climax timestamp is challenging, we apply an error window saying that the prediction is treated as correct if the ground-truth climax is close (within 0, 1, 2 sec). We treat the prediction as correct if it recalls any of the ground-truth annotations for that video, except rejected work. Table 1 shows the results.

To measure how well the model’s prediction agrees with the sentiment annotations, we compute mean average precision (mAP) and top-1 accuracy (acc@1) based on three forms of agreement (agree with k , where $k = 1, 2, 3$). “Agree with k ” means that we assign a ground-truth label to a video only if at least k annotators agree on the existence of the sentiment. The acc@1 is the fraction of correct top-1 predictions across all videos, and the mAP is the mean of the average precision over evenly spaced recall levels. Tables 2, 3 and 4 show the results.

4.2 Training and implementation details

For training both the climax and sentiment prediction models, we use the TensorFlow [2] deep learning framework. We split the Video Ads Dataset [16] (3,477 videos) into train/val/test (60%/20%/20%), resulting in around 2,000 training examples for the sentiment prediction task and about 700 training examples for the climax prediction task (since only 1,149 of the 3,477 videos have climax annotation). We report our results using five-fold cross-validation.

For the climax prediction task, we use a one-layer LSTM model with 64 hidden units. At each timestamp, the model predicts a real value ranging from $[0, 1]$ (output of the sigmoid function) denoting whether the corresponding frame contains a climax. We then use the sigmoid cross entropy loss to constrain the model to mimic the human annotations. Considering the size of the dataset, we set both the input and output dropout keep probability of the LSTM cell to 0.5 to avoid over-fitting. We use the RMSprop optimizer with a decay factor of 0.95, momentum of $1e-8$, and learning rate of 0.0002. We train for 20,000 steps using a batch size of 32, and we use the recall of the top-1 prediction (the error window is set to “within 2 seconds”) to pick the best model on the validation set.

For the sentiment prediction task, we use the same procedure, but we pick the best model using mAP using “agreement with 2”. We use the last hidden state of the LSTM to represent the video feature and add a fully connected layer upon it to get the 38D topic representation. We then concatenate the 38D topic representation with the last hidden state of the LSTM and infer a 30D sentiment logits vector from the concatenated feature. The sigmoid cross entropy loss is also used here. Similar to [16], we found that using soft scores as ground-truth targets improves the performance and makes the training more stable. To deal with data imbalance for the rare classes, we sampled at most $5n$ negative samples if there were n positives.

4.3 Climax prediction

We show the results of unsupervised and supervised climax prediction in Table 1. We measure whether the predicted climax is within 0, 1, or 2 seconds of the ground-truth climax. We first show a heuristic-guess baseline which always predicts that climax occurs at 5 seconds for the top-1 prediction and at 5, 15 and 25 seconds for top-3. We then show the performance of the three unsupervised climax prediction methods described in Sec. 3.4. Next, we show the performance of 0/1 climax prediction (Sec. 3.5) using an LSTM with ResNet features

Method	top-1 prediction			top-3 prediction		
	w/in 0 s	w/in 1 s	w/in 2 s	w/in 0 s	w/in 1 s	w/in 2 s
baseline	0.031	0.083	0.121	0.122	0.299	0.430
shot boundary (unsup)	0.068	0.179	0.265	<i>0.221</i>	0.457	0.588
optical flow (unsup)	0.064	0.152	0.220	0.163	0.380	0.513
audio (unsup)	0.077	0.171	0.255	0.178	0.403	0.534
LSTM, ResNet only	0.071	<i>0.206</i>	0.290	0.190	0.400	0.523
LSTM, all feats (Ours)	0.077	0.209	<i>0.287</i>	0.226	<i>0.439</i>	<i>0.546</i>

Table 1: Climax prediction with best performer per setting in **bold** and second performer in *italics*. Unsupervised prediction performs quite well. Our supervised method achieves the best or second-best performance for all settings. For the “LSTM, ResNet only” approach, we guess the reason that it is competitive is that LSTM has the ability to capture the temporal dynamics to a certain degree.

Method	Agree with 1		Agree with 2		Agree with 3	
	mAP	acc@1	mAP	acc@1	mAP	acc@1
Hussain et al. [16]	0.283	0.664	0.135	0.435	0.075	0.243
Our model	0.313	0.712	0.160	0.449	0.094	0.241

Table 2: Our method outperforms prior art for sentiment prediction.

only, and finally our method using the features we proposed in both Sec. 3.2 and Sec. 3.3 (excluding the video-level topic feature).

We see that the unsupervised methods, and especially shot boundary and audio, greatly outperform the baseline. Interestingly, audio performs quite well in the hardest setting, only one shot at prediction and exact alignment between predicted and ground-truth climax. Shot boundary achieves the best performance in the two weakest settings (top-3 predictions, agreement within 1-2 seconds). In all settings, our method achieves the best or second-best performance.

4.4 Sentiment prediction

Table 2 shows our main result for sentiment prediction. We compare to Hussain et al. [16]’s method which is a multi-class SVM model using the C3D features [18]. This is the only prior method that attempts to predict sentiment on the Video Ads Dataset. We observe that our method improves upon [16]’s performance for most metrics. The improvement is more significant for mAP, which is more reliable because of the imbalance of the dataset. We improve the mAP compared to prior art by up to 25% in terms of agreement with 3 annotators. For reference, human annotators’ agreement with 1 (at least one other annotator) is 0.723.

Table 3 examines the contribution of the features described in Sec. 3.2 and Sec. 3.3, and the use of an LSTM to model dynamics of the video. We compare against an LSTM that uses only ResNet features. We also compare to a bag-of-frames (BOF) method that rules out the effects of dynamics. It computes the final video-level representation by simply applying mean pooling among the frame-level features. We observe that our method (using the proposed features and LSTM) always outperforms the other methods in terms of mAP scores. Our method achieves significant improvement over the second-best method (10% for mAP and agreement with 2, and 21% for mAP and agreement with 3). In terms of accuracy, all methods perform similarly, and the best model (BOF, all features) also uses our proposed features.

Method	Agree with 1		Agree with 2		Agree with 3	
	mAP	acc@1	mAP	acc@1	mAP	acc@1
BOF, ResNet only	0.295	0.708	0.141	0.449	0.076	0.242
LSTM, ResNet only	0.302	0.716	0.145	0.451	0.074	0.242
BOF, all features (incl. ours)	0.302	0.719	0.146	0.462	0.078	0.248
LSTM, all features (Our model)	0.313	0.712	0.160	0.449	0.094	0.241

Table 3: In-depth evaluation of the components of our method for sentiment prediction.

	average	educated	alarmed	fashionable	angry
ResNet only (baseline)	0.074	0.036	0.117	0.047	0.007
objects	0.082	0.032	0.140	0.080	0.004
places	0.082	0.074	0.132	0.160	0.005
facial expressions	0.077	0.044	0.143	0.084	0.003
topic	0.086	0.032	0.143	0.136	0.009
optical flow	0.082	0.045	0.150	0.133	0.005
shot boundaries	0.080	0.037	0.151	0.110	0.003
audio	0.077	0.040	0.113	0.116	0.010
climax	0.079	0.025	0.119	0.082	0.011
all features except ResNet	0.080	0.038	0.104	0.036	0.007
all features (Our model)	0.094	0.026	0.099	0.202	0.005

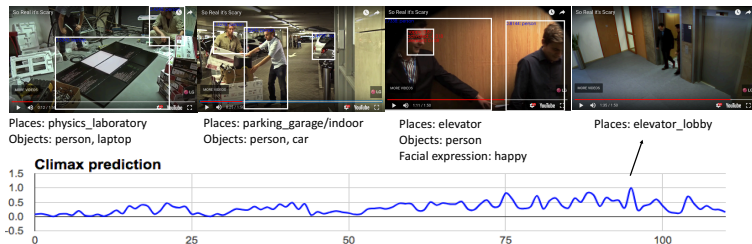
Table 4: Ablation study evaluating the benefit of each feature for sentiment prediction, using agreement with 3 mAP. In bold are all methods that outperform the baseline.

Table 4 verifies the benefit of each of our features. We show the LSTM-ResNet-only baseline from Table 3, then eight methods which add one of our features at a time, on top of this baseline. Next, we show an LSTM method which uses our features without the base ResNet feature, and finally, our full method. We use mAP for agreement with 3 in the table. We show the average result across all sentiment classes, then results for four individual ad sentiments. In bold are all methods which improve upon the ResNet baseline. We see that all of our features (the average column) contribute to the performance of our full method. Using all features except ResNet is stronger than using ResNet features alone. We note models based on individual features still show benefits on specific sentiment classes, and we believe the reason is that our fusion method is too simple to aggregate all the information.

We observe some intuitive results for the four chosen individual sentiments. We ranked sentiments by frequency in the dataset and picked the 6th, 7th, 9th and 13th most frequent. For “educated,” the places feature is most beneficial, which makes sense because “education” might occur in particular environments, e.g., classroom. As shown in our example ad in Fig. 4, the setting (e.g., places) and dramatic content changes (measured by optical flow and shot boundaries) are quite telling of the “alarmed” sentiment. Most features help greatly for the “fashionable” sentiment. For “angry”, audio is very helpful (43% improvement over ResNet), which makes sense since loud speaking might trigger or correlate with anger.

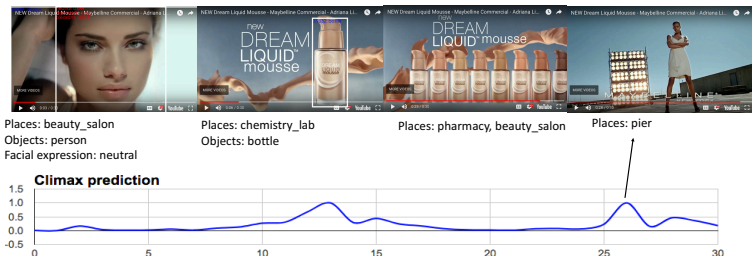
We show qualitative examples in Fig. 5. Our model’s features correctly predict “amazed” and “fashionable” while the baseline method does not. Our method relies on recognized places (e.g. laboratory, beauty salon), objects, facial expressions, and climax dynamics.

So real it's scary (<https://www.youtube.com/watch?v=NeXMxuNNIE8>)



Annotation:
amazed
Prediction (*without*
our features):
alarmed
Prediction (ours):
amazed

New dream liquid mousse (<https://www.youtube.com/watch?v=MTgeUVOxI8E>)



Annotation:
feminine, amazed,
fashionable
Prediction (*without*
our features):
alert
Prediction (ours):
fashionable

Figure 5: Qualitative results from our model.

5 Conclusion

We made encouraging progress in understanding the dynamic structure of a video ad. We hypothesized that climax correlates with dramatic visual and audio changes. We crowdsourced climax annotations on 1,149 videos from the Video Ads Dataset of [16] and used both unsupervised and supervised methods to predict the climax. By combining visual and audio cues with semantically meaningful context features, our sequential model (LSTM) outperforms the only prior work [16] by a large margin. To better understand the relations between the semantic visual cues and the sentiment each ad video provokes, we performed detailed ablations and found all the features we proposed help to understand the evoked sentiment. In the future, we will investigate other resources relevant to both climax and sentiment in video ads. We will also improve the interpretability of the model. Finally, our ablation studies show the limitation of the feature fusion method of our model thus we will investigate additional fusion strategies to further improve the performance.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number 1566270. This research was also supported by a Google Faculty Research Award and an NVIDIA hardware grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors also appreciate the help of Sanchayan Sarkar and Chris Thomas for preparing the data for and training the facial expression models.

References

- [1] Merriam-webster.com. <https://www.merriam-webster.com/dictionary/climax>.
- [2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [3] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 2016.
- [4] Carlos Fabian Benitez-Quiroz, Ramprakash Srinivasan, Aleix M Martinez, et al. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016.
- [5] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 25–36. Springer, 2004.
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.
- [7] Brandon Castellano. Pyscenedetect. <https://github.com/Breakthrough/PySceneDetect/>.
- [8] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Video-story composition via plot analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [9] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187, 2003.
- [10] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.
- [11] David Fleet and Yair Weiss. Optical flow estimation. In *Handbook of mathematical models in computer vision*, pages 237–257. Springer, 2006.
- [12] Gustav Freytag. *Freytag’s technique of the drama: an exposition of dramatic composition and art*. Scholarly Press, 1896.
- [13] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 759–768. IEEE, 2015.

- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [15] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [16] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110. IEEE, 2017.
- [17] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.
- [18] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–223, 2014.
- [19] Jungseock Joo, Francis F Steen, and Song-Chun Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720, 2015.
- [20] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.
- [21] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2309–2317. IEEE, 2017.
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.
- [23] Anan Liu, Jintao Li, Yongdong Zhang, Sheng Tang, Yan Song, and Zhaoxuan Yang. An innovative model of tempo and its application in action scene detection for movie analysis. In *WACV*, 2008.
- [24] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1975–1981. IEEE, 2010.
- [25] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016.

- [26] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.
- [27] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.
- [29] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.
- [30] Zeeshan Rasheed and Mubarak Shah. Movie genre classification by exploiting audio-visual features of previews. In *ICPR*, 2002.
- [31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [32] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, 2013.
- [33] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439. IEEE, 2010.
- [34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.
- [35] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelwagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2016.
- [36] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.
- [37] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014.

- [38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.
- [39] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4305–4314, 2015.
- [40] Yu Wang, Yang Feng, Zhe Hong, Ryan Berger, and Jiebo Luo. How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International Conference on Social Informatics*, pages 440–456. Springer, 2017.
- [41] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, 2009.
- [42] Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 786–794. ACM, 2017.
- [43] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *European Conference on Computer Vision (ECCV)*. Springer, 2018.
- [44] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2678–2687, 2016.
- [45] Charles E Young. *The advertising research handbook*. Ideas in Flight, 2008.
- [46] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.