Interpreting the Rhetoric of Visual Advertisements

Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, Adriana Kovashka

Abstract—Visual media have important persuasive power, but prior computer vision approaches have predominantly ignored the persuasive aspects of images. In this work, we propose a suite of data and techniques that enable progress on understanding the messages that visual advertisements convey. We make available a dataset of 64,832 image ads and 3,477 video ads, annotated with ten types of information: the topic and sentiment of the ad; whether it is funny, exciting, or effective; what action it prompts the viewer to do, and what arguments it provides for why this action should be taken; symbolic associations that the ad relies on; the metaphorical object transformations on which especially creative ads rely; and the climax in video ads. We develop methods that use multimodal cues, i.e. both visuals and slogans, for both the image and video domains. Our methods rely on finding poignant content spatially and temporally. We also examine the creative story construction in ads: for videos, we learn to predict when the climax occurs (if any), and how effective the story is; for images, we analyze how object transformations in ads metaphorically depict product properties.

Index Terms—visual reasoning, vision and language, video understanding, representation learning, visual rhetoric, atypicality

1 Introduction

Visual media are informative, but they are also manipulative, intentionally or unintentionally [1], [2], [3], [4]. Targeted campaigns to change public opinions on matters with economic and social impact have been effective [5], [6]. Well-created ads gain great popularity and are seen by many, thus entering our common consciousness [7]. The public response to political images has caused policy changes as well as major governmental decisions on issues such as war involvement and admitting refugees [8], [9].

Despite the importance of the persuasive nature of visual media, there is a scarcity of computer vision approaches to understand visual rhetoric. While we have made impressive progress on inferring the explicit content in the media (e.g. objects, scenes, actions), the implicit nuances of the media have been overlooked, partly due to the significant challenges that this task poses. Sometimes the message of an advertisement image is simple, and can be inferred from body language, as in the "We can do it" ad (A) in Fig. 1. Other images convey more complex or clever messages, whose decoding relies on human visual recognition (including generalization), association, and reasoning capabilities. For example, in Fig. 1 (B), one might infer that because the eggplant and pencil form the same object, the pencil gives a very real, *natural* eggplant color. In Fig. 1 (C), one might conclude that Burger King burgers are delicious, since even employees from competitor restaurants (McDonalds) secretly buy them. In Fig. 1 (D), lungs symbolize breathing and by extension, life. However, a human first has to recognize the groups of trees as lungs, which might be difficult for a computer vision system to do, due to the atypical texture. In Fig. 1 (E), the viewer has to infer that the woman went on vacation from the fact that she is carrying a suitcase,

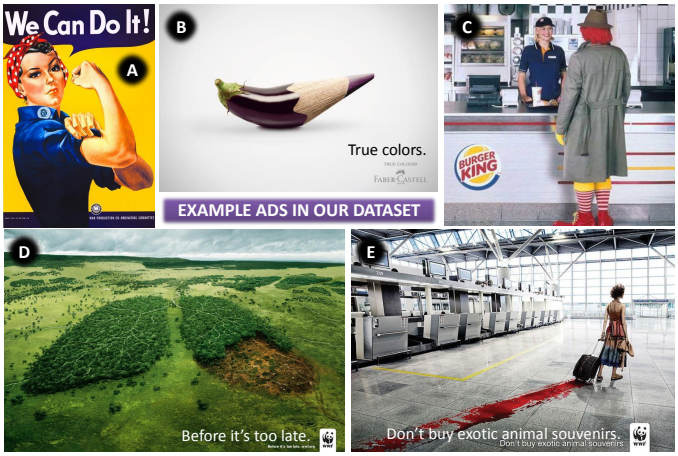


Fig. 1: Example advertisements from our dataset that require challenging visual recognition and reasoning. Despite the potential applications of understanding the messages of ads, this problem has not been tackled in computer vision.

and then surmise that she is bringing dead animals from the blood trailing behind her suitcase. A human knows this because she associates blood with injury or death. These are just a few examples of how ads use different types of *visual rhetoric* to convey their message, namely: association and symbolism, common-sense reasoning, and recognition of non-photorealistic objects. Understanding advertisements automatically requires decoding this rhetoric.

We argue that the ability to automatically understand the explicit or implicit messages of persuasive images is important. We focus on visual advertisements, and propose the following advancements on this task. First, because ads rely on human perception and reasoning, we collect and learn from human annotations on ads. Second, we explicitly model symbolic associations that ads exploit, and use these

- *All authors performed the work at University of Pittsburgh. James Hahn is now at Georgia Tech.*
  E-mail: kovashka@cs.pitt.edu

Manuscript received ???.

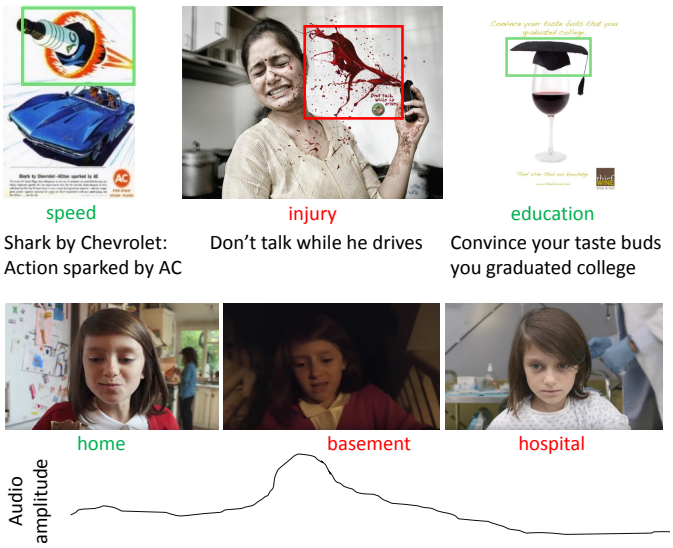


Fig. 2: The core idea of our method: We rely on spatial or temporal attention (symbols, climax), and two modaliities (visual and text/audio), to understand the messages of image and video ads. At the top, we show symbols (rocket-speed, blood-injury, cap-education) and slogans embedded in the ads. At the bottom, we show frames from a video ad, along with predicted scene types, and the audio amplitude; we use these to infer the most important temporal region, or the climax, of a video.

to predict how a human would explain the message of a particular ad. Third, we model the structure of ads through attention mechanisms that mimic human associations and sentiments. Finally, since ads are multimodal, we exploit cues from the multiple channels to facilitate ad understanding. We show that our techniques improve the performance of off-the-shelf vision methods by 46%, for the more challenging type of ads in our dataset, namely public service announcements, and 31% for product ads.

In more detail, we *propose a large, richly annotated dataset*, that enables initial progress on the challenging task of automatically understanding the messages that visual advertisements convey. We released a dataset of 64,832 image ads, and 3,477 video ads. We provide ten key types of annotations on these: topics, sentiments, actions that the ad calls for, arguments it provides for taking the suggested actions, humor, excitement, effectiveness, symbolism, atypical or metaphorical object portrayals, and climax.

We propose a set of methods to predict these annotations. Our methods have *two key characteristics:* (1) they look for poignant regions in space and time, to make their decisions, and (2) they use multiple modalities, i.e. image/video and text/speech/sound. We focus on the task of *inferring the suggested action* (what the viewer should do) and *provided arguments* (why they should do it, according to the ad). We propose a novel method that embeds images and action-reason (what-why) statements, to allow retrieval of statements given an image. In particular, we train an attention model to select regions likely to contain symbolism, and use them to represent the image. The representation is affected by a mapping of visuals to symbolic meaning (e.g. gun-danger or speed-bullet, see top of Fig. 2), and also depends on slogans automatically extracted from the image. For video ads, we perform the same task in similar fashion, but use the speech in the audio track as a cue.

To better understand *how stories are constructed in video ads*, we develop two additional, related methods. One learns to predict the climax of a video, based on cues extracted in unsupervised fashion, e.g. audio amplitude and optical flow, and ones trained on a disjoint dataset, e.g. a distribution of the objects, places, and facial expressions in the video. The likelihood of an image frame to belong to the climax of a video can be used as an attention mechanism– more important content is near the climax. We learn to predict the effectiveness of a video using similar features.

Finally, to understand the *story in image ads* in greater depth, we examine a subset of ads that contain intentionally non-photorealistic, manipulated, metaphorical object portrayals (see Fig. 5). We annotate over 4,000 ads deemed atypical, with the type of object transformation they show (e.g. an object with the texture of another, or parts from two objects combined). We show analysis of the type of transformations ads use, that in the long run, can be used to semi-automatically generate ads.

This paper extends our prior work [10], [11], [12]. The major differences are as follows. First, we include new methods, results and analysis on two tasks, predicting action-reason statements in image and video ads. Unlike our previous work, here we incorporate the slogans embedded in image ads, and the audio channel in video ads, to infer the statement. This leads to a large boost in performance. We show new ablation results, and a new result comparing the difficulty of predicting *what* the viewer should do as opposed to *why* they should do it. Second, we develop methods for a new task that we did not previously tackle, namely predicting the effectiveness of a video ad. We test a variety of features for this task, both new and adapted from previous methods, and show that their combination works best. Third, we propose a new addition to our dataset which analyses the creative and metaphorical transformations of objects that ads use.

The potential impact of our work is as follows. Most directly, our work can be used to train models that can group ads according to different topics, and has recently been used to explore adversarial techniques to ad blocking [13]. Decoding ads would allow us to generate descriptions of these ads for the visually impaired, and thus give them richer access to the content shown in newspapers or on TV. Thus, our work is the rhetoric-aware analogue to recent captioning work and visual question answering for blind users [14]. It can also be employed in parallel with focus groups [4] to gauge quality and effectiveness before an ad is finalized and released. Using automatic understanding of the strategies that ads use, we can help viewers become more aware of how ads are tricking them into buying certain products. It can be used to gauge media biases, which could either strengthen or burst information bubbles [15], [16]. Finally, ads are both affected by, and affect, how we think about objects and people, thus our work broadly fits into recent efforts to cope with social biases creeping into machine models [17], [18].

2 Related Work

Our work aims to automatically analyze one type of visual media, using novel techniques in the space of image-text embeddings, attention, atypicality, and video analysis. We discuss the most relevant work below.

Automated media analysis. There is a small body of work in analyzing persuasion and social phenomena as portrayed in the media domain. [19] analyze in what light a photograph portrays a politician, and [20] examine how the facial features of a candidate determine the outcome of an election. [21] examine the facial attributes of faces in politics, and [22] examine the variance of faces in ads. Some work also analyzes events (e.g. protests) as reported in social media [23]. This work primarily applies to images of people. Also related is work in parsing infographics, charts and comics [24], [25], [26]. In particular, these focus on modeling attention, or extracting information and answering questions about comic books. In contrast to these, our interest is analyzing the *implicit* arguments ads were created to make.

Predicting placement and responses to ads. We are not aware of any work in decoding the meaning of advertisements as we propose. However, [27], [28] predict click-through rates in ads using low-level vision features, whereas we predict what the ad is about and what message it carries. [29], [30] determine the best placement of a commercial in a video stream, or of image ads in a part of an image using user affect and saliency. [31], [32] detect whether the current video shown on TV is a commercial or not, and [33] detect human trafficking advertisements. [13] modify ads to be indiscernible from regular images, in order to bypass ad-blockers. In terms of human responses to ads, [34] predict how much human viewers will like an ad by capturing their facial expressions. Human facial reactions, and ad placement and recognition, are quite distinct from our goal of decoding the messages of ads. There is also extensive research in the media studies, communications and advertising research community [35], [36] on how ads build rapport, but this research is not computational.

Image-text embeddings. We approach the task of inferring the meaning of an ad, as retrieving a statement that describes what the viewer should do and why, according to the ad. We do this through an image-text embedding that we learn. There has been great interest in joint vision-language tasks, e.g. captioning [37], [38], [39], [40], visual question answering [41], [42], [43], [44], and cross-domain retrieval [45], [46]. These often rely on learned image-text embeddings. [11], [46], [47] use triplet loss where an image and its corresponding human-provided caption should be closer in the space than pairs that do not match. [48] propose a bi-directional network to maximize correlation between matching images and text. None of these consider implicit persuasive intent, as we do.

Region proposals and attention. Our methods find the most relevant regions, in space (via region proposals) or time (via predicted climax). Region proposals [49], [50], [51] guide an object detector to regions likely to contain objects. Attention [52], [53] focuses prediction tasks on regions likely to be relevant. We show that for our task, the attended-to regions must be those likely to be visual anchors for symbolic references.

Atypical objects. As part of our analysis of the meaning of ads, we examine the creative, often atypical portrayal of objects (e.g. hybrid objects, atypically textured objects, etc.) In prior work, [54] build a generative model from typical objects, that learns typical relationships between objects, scenes and attributes in regular images, and uses it to predict whether an image is regular or abnormal. [55] rely on the distribution of object detection scores in different regions in the image. [56] detect objects placed in atypical contexts and environments. In contrast to these works, we consider finer nuances of transformations. [55] consider three types of transformations, while we consider eight. Further, we study *purposeful*, not accidental or purely artistic transformations, and examine the correlations between advertisement topics and transformation categories.

Effectiveness. Media arts papers have examined effectiveness as related to context [57], repetition [58], brand recognition [59], emotion and engagement [60], etc. Most of these papers require human input from surveys, and do not perform computational analysis or automatic prediction of effectiveness. [61] uses neural networks to predict TV ad effectiveness, but all 837 participants in the survey analyzed one of three ads, all of which were marketing toothpaste. Our dataset consists of close to forty topics and product types, thus making the task more challenging.

Video story and dynamics. One task in our work is to understand the structure of video ad stories. Others have developed techniques for understanding movie plots [43], [62] and the principal characters and their relations [63]. While there is no prior work on detecting climax in ads, some previous approaches model the tempo of other videos. [64] use cues like "motion intensity" and "audio pace" to detect action scenes. [65] use pacing to recognize movie genre since action movies are faster-paced than dramas. [66] create video stories out of consumer videos, using story composition and dynamics. We show semantic context features based on objects, scenes and emotions improve the performance of purely motion- or pace-based ones.

3 Dataset

We propose the problem of ad-understanding, and develop two datasets to enable progress on it. We collect a dataset of over 64,000 *image* ads (both product ads, such as the pencil or burger ads, and public service announcements, such as the anti-animal-souvenirs and environment-preservation ads from Fig. 1). Our ads cover a diverse range of thirty-eight subjects (e.g. food products, cosmetics, electronics, vehicles, travel, services, anti-violence public service announcements, etc.). We ask Amazon Mechanical Turk (MTurk) workers to tag each ad with its *topic* (e.g. what product it advertises or what the subject of the public service announcement is, e.g. "environment"), and what *sentiment* it attempts to inspire in the viewer (e.g. "disturbance" in the environment conservation ad). We also include crowdsourced answers to two questions: the *action* that the ad prompts ("What should the viewer do according to this ad?") and the *reason* it provides ("Why should he/she do it?"). We include any *symbolism* that the ad uses (e.g. trees/lungs symbolize "life", or a dove symbolizes "peace").

In addition to these core annotations, for some of our images, we also collect three additional types: strategy of the ad, annotator-created slogan, and presence/types of atypical objects. For 4,000 ads, we ask what *strategy* the ad uses to convey its message (e.g. it requires understanding of physical processes such as burning). For 2,000 ads, we ask annotators to invent a creative *slogan* for the ad. As an addition to our previous work, for 4,064 creative ads, we also examine how they compose object parts in metaphorical, attention-attracting ways; we collect 15,816 annotations about eight categories of atypicality.

We also annotate a dataset of almost 3,500 *video* ads with topic, sentiment, action, reason, humor, excitement, and effectiveness. First, we ask annotators "Is the ad *funny*?" and "Is it *exciting*?" We also ask "How effective is the ad?" on a scale of 1 (least) to 5 (most) effective. Finally, on 1,149 video ads, we ask annotators to mark the time at which the *climax* of the video occurs, if present.

Our data collection and annotation procedures were informed by the literature in Media Studies, a discipline which studies the messages in the mass media.[1] Our data is available at `http://www.cs.pitt.edu/~kovashka/ads/`. The dataset contains the ad images, video ad URLs, and annotations we collected, as well as a visualization tool to explore the dataset.

Below, we first describe how we collected the ad images and videos, then how we annotated them. For images and videos, we collected the annotations in Tables 1 and 2, respectively. We discuss most annotation types below. Please refer to [10] for a discussion of strategies and slogans.

TABLE 1: The annotations collected for our image dataset. The counts are before majority-vote cleanup. The *italicized* annotations are ones we do not use to train models currently.

Type	Count	Example
Topic	204,340	Clothing, Electronics
Sentiment	102,340	Cheerful, Disturbed
Action	202,090	I should bike more often...
Reason	202,090	... because it's healthy
Symbol	64,131	Danger (+ bounding box)
Atypical object	15,816	Tongue with strawberry texture
Strategy	*20,000*	*Contrast*
Slogan	*11,130*	*Save the planet... save you*

TABLE 2: The annotations for our video ads.

Type	Count	Example
Topic	17,345	Cars/automobiles, Safety
Sentiment	17,345	Cheerful, Amazed
Action	17,345	I should buy this car...
Reason	17,345	... because it is pet-friendly
Funny?	17,374	Yes/No
Exciting?	17,374	Yes/No
Effective?	16,721	Not/.../Extremely Effective
Climax	2,386	00:30 (timestamp)
English?	*15,380*	*Yes/No/Does not matter*

3.1 Collecting ad images

We first assembled a hierarchy of keywords that describe ad topics at different levels of granularity. This hierarchy included both coarse topics, e.g. "fast food", "cosmetics", "electronics", as well as fine topics, such as the brand names of products (e.g. "Sprite", "Maybeline", "Samsung"). For public service announcements (PSAs) we used keywords such as: "smoking", "animal abuse", "bullying". We used the entire hierarchy to query Google and retrieve all the images (usually between 600 to 800) returned for each query. We removed all images of size less than 256x256 pixels, and obtained an initial pool of about 220,000 noisy images.

Next, we removed duplicates from this noisy set. We computed a SIFT bag-of-words histogram per image, and used the chi-squared kernel to compute similarity between histograms. Any pair of images with a similarity greater than a threshold were marked as duplicates. After de-duplication, we ended up with about 190,000 noisy images.

We removed images that are not actually advertisements, using a two-stage approach. First, we selected 21,945 images, and submitted those for annotation on MTurk, asking "Is this image an advertisement? You should answer yes if you think this image could appear as an advertisement in a magazine." We showed plentiful examples to annotators to demonstrate what we consider to be an "ad" vs "not an ad". We marked as ads those images that at least three out of four annotators labeled as an ad, obtaining 8,348 ads and 13,597 not-ads. Second, we used these to train a ResNet [67] to distinguish between ads and not ads on the remaining images. We set the recall of our network to 80%, which corresponded to 85% precision on a held-out set. We ran that ResNet on our 168,000 unannotated images for clean-up, obtaining about 63,000 images labeled as ads. We allowed annotators to label ResNet-classified "ads" as "not an ad" in a subsequent stage; annotators only used this option in 10% of cases. Finally we obtained **64,832** ads.

1. Adriana Kovashka has formal training in Media Studies.

3.2 Collecting ad videos

Video advertisements reach broad audiences; e.g. an Old Spice commercial has over 56 million views. However, they are expensive to make [68], thus there are fewer commercials available on the web. We obtained a list of 949 ad videos from an Internet service provider. To increase the size of the dataset, we additionally crawled YouTube for videos, using the keywords we used to crawl Google for images. We picked videos that had been played at least 200,000 times and had more "likes" than "dislikes". We ran an automatic de-duplication step. For every video, we separately took (1) 30 frames from the beginning and (2) 30 from the end, lowered their resolution, then averaged over them to obtain a single image representation, which is less sensitive to slight variations. If both the start and end frames of two videos matched according to a hashing algorithm [69], they were declared duplicates. We thus obtained an additional set of 5,028 noisy videos, of which we submitted 3,000 for annotation on Mechanical Turk. We combined the ad/not ad cleanup with the remainder of the annotation process. We used intuitive metrics to ensure quality, e.g. we removed videos that were low-resolution, very old, spoofs, or simply not ads. We thus obtained **3,477** video ads in total.

3.3 Topics and sentiments

The keyword query process used for image download (Sec. 3.1) does not guarantee that the images returned for

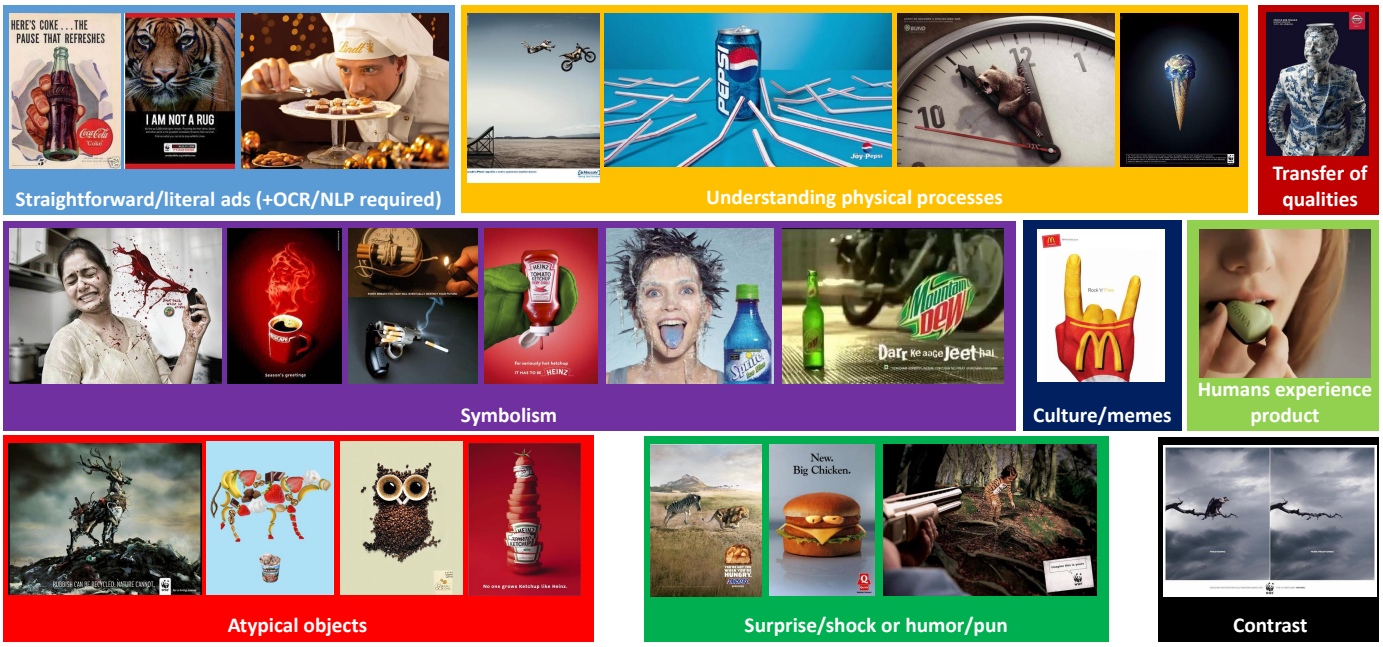


Fig. 3: Examples of ads grouped by strategy or visual understanding required for decoding the ad.

each keyword actually advertise that topic. Thus, we developed a taxonomy of products, and asked annotators to label the images with the topic that they advertise or campaign for. We also wanted to know how an advertisement makes the viewer feel, since the sentiment that the ad inspires is a powerful persuasion tool [34]. Thus, we also developed a taxonomy of sentiments. To get both taxonomies, we first asked annotators to write free-form topics and sentiments, on a small batch of images and videos. This is consistent with the "self report" approach used to measure emotional reactions to ads [70]. We then semi-automatically clustered the annotations and selected a representative set of words to describe each topic and sentiment type. We arrived at a list of 38 topics and 30 sentiments. In later tasks, we asked workers to select a single topic and one or more sentiments. We collected topic annotations on all ads, and sentiments on 30,340 ads. For each image, we collected annotations from three to five different workers. Inter-annotator agreement on topic labels was 85%. Some topic/sentiment classes are shown in Tab. 3. For videos, we showed workers six examples for how to annotate. The topic and sentiment options overlap with those used for images. The distribution for a subset of topics and sentiments for our video ads is illustrated in Fig. 4. We see cheerfulness is most common for beauty and soda ads, eagerness for soda ads, creativeness for electronics ads, and alertness for political ads.

TABLE 3: A sample from our list of topics and sentiments.

Topics	Sentiments
Restaurants, cafe, fast food	Active (energetic, etc.)
Coffee, tea	Alarmed (concerned, etc.)
Sports equipment, activities	Amazed (excited, etc.)
Phone, TV and web providers	Angry (annoyed, irritated)
Education	Cheerful (delighted, etc.)
Beauty products	Disturbed (disgusted, etc.)
Cars, automobiles	Educated (enlightened, etc.)
Political candidates	Feminine (womanly, girlish)
Animal rights, animal abuse	Persuaded (impressed, etc.)
Smoking, alcohol abuse	Sad (depressed, etc.)

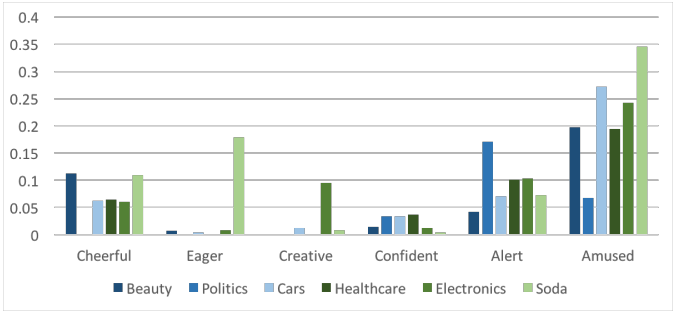


Fig. 4: Statistics about topics and sentiments.

3.4 Actions and reasons

We collected 202,090 actions and corresponding reasons, with three action-reason pairs per image. Tab. 4 shows a few examples. We required workers to provide answers in the form "I should [Action] because [Reason]." For later tasks, we split this into *two* questions, i.e. we separately asked about the "What?" (action) and the "Why?" (reason). Examples of the most commonly used words in the questions and answers are shown in Tab. 5. We followed the same procedure for images and videos.

3.5 Symbols

In the second row of Fig. 3, the first image uses blood to symbolize injury, the second symbolically refers to the holiday spirit via the steam, the third uses a gun to symbolize danger, the fourth uses an oven mitt to symbolize hotness, the fifth uses icicles to symbolize freshness, and the sixth uses a motorbike to symbolize adventure. Decoding symbolic references is difficult because it relies on human associations. In the Media Studies literature, the physical

TABLE 4: Examples of collected action-reason pairs.

Question	Answer
What should you do, acc. to the ad?	I should buy Nike sportswear.
Why, acc. to the ad, should you do it?	Because it will give me the determination of a star athlete.
What?	I should buy this video game.
Why?	Because it is a realistic soccer experience.
What?	I should drink Absolut Vodka.
Why?	Because they support LGBT rights.
What?	I should look out for domestic violence.
Why?	Because it can hide in plain sight.
What?	I should not liter in the ocean.
Why?	Because it damages the ocean ecosystem.

TABLE 5: Common words in responses to action and reason questions for selected topics, from the image dataset.

What should you do?			*Why* should you do it?		
Educat.	Travel	Smoking	Educat.	Travel	Smoking
go	go	smoke	help	fun	smoking
college	visit	cigarette	learn	beautiful	like
use	fly	buy	want	like	kill
attend	travel	stop	career	want	make
school	airline	quit	things	great	life

(a) Coffee (b) Shards (c) Crisps (d) Manfish

(e) Animal freedom (f) Delicious (g) Man without head

(h) Jar in onion (i) Beer in tree

Fig. 5: Atypical object transformations in ads: (a-c, e) atypical textures, (d, f) parts of objects combined, (g) parts missing, (h) one object in another, (i) object in new context.

object or content that stands for some conceptual symbol is called "signifier", and the symbol is the "signified" [35].

We develop a list of symbols (concepts, signifieds) and corresponding training data, using the help of MTurkers. We use a two-stage process. First, we ask annotators whether an ad can be interpreted literally (i.e. is straight-forward), or it requires some non-literal interpretation. For simplicity, we treat all non-literal strategies as symbolism. If the majority of MTurkers respond the ad is non-literal, it enters a second stage, in which we ask them to label the signifier and signified. In particular, we ask them to draw a bounding box (which denotes the signifier) and label it with the symbol it refers to (the signified). 13,938 of all images were found to contain symbolism. We prune extremely rare symbols and arrive at a list of 221 symbols, each with a set of bounding boxes. The most common symbols are: "danger," "fun," "nature," "beauty," "death," "sex," "health," and "adventure." We use these symbol annotations to infer the messages of ads. We show that finding what regions in an image are symbolically salient is crucial for good performance.

3.6 Atypical object transformations

Creative ads use symbolism to convey properties, or sometimes create entirely new, abstract, atypical objects to convey a property. Examples are shown in Fig. 5. To complement our prior work, we collect annotations about the manner in which ad designers transform and combine objects to convey properties, analogies, humor or alarm.

First, we defined our objects of interest as those that are either deformed beyond the range of typical deformations (e.g. viewpoint, pose) that one might see in realistic imagery, or that are in an unlikely context. We manually examined all ads in our dataset for whether they potentially matched this definition, or were unambiguously regular images (not of interest). We found a total of 11,795 images of potentially transformed objects.

Second, we devised a taxonomy of object transformations; one author grew and collapsed it iteratively as she examined a subset of 42 images potentially of interest. The final taxonomy is the following (with examples in Fig. 5):

1) One type of texture: In this category objects are not in their original texture, and the new texture is borrowed from another object, e.g. owl with texture of coffee beans (Fig. 5a), car with texture of glass shards (Fig. 5b), or man with texture of turtle (Fig. 5e).
2) Texture from separate objects: The texture has been created from combining small objects, and there is distance between the objects that create the texture; e.g. apple made of apple crisps (Fig. 5c).
3) One object inside another; e.g. jar in an onion (Fig. 5h).
4) Object with missing part; e.g. man wo/ head (Fig. 5g).
5) Hybrid object: New object from parts of other objects; e.g. man/fish or thumb/candy (Figs. 5d, 5f).
6) Bent object: Solid objects which have been deformed from their original shape by bending, twisting, etc.
7) Liquid deformed object: Like bent object but for liquids.
8) Context replacement: One object is placed in the context where another typically occurs; e.g. beer in tree (Fig. 5i).
9) The image contains an atypical object but it cannot be assigned to any of the aforementioned categories.
10) The image does not contain any atypical objects.

Third, we obtained annotations on 4,064 images. Annotators were shown definitions and four examples per category, and were asked to choose one or several of the categories for each image. One image per batch came from the 42 we manually labelled and was used for quality control.

Fourth, annotators also had to provide *fine-grained* details for each category. For category 1, we asked what object has the new texture, and what this new texture is (e.g. a [human] with [turtle] skin). For category 2, we asked what object was created from small objects, and what these are

TABLE 6: Annotation statistics for atypical objects.

Category	1	2	3	4	5	6	7	8
Count	252	213	142	75	260	141	125	487
% Agree	41	46	45	42	45	48	58	51

(e.g. a [heart] made out of small [fruits]). For category 3, we asked what object is inside vs outside (e.g. a [sauce jar] inside an [onion]). For category 4, we asked what object has missing parts. For category 5, annotators wrote down what objects have been combined to create a new object (e.g. [man, fish], or [baby bottle cap, ketchup cap]). For category 6, they stated in what way an object has been deformed (e.g. twisted). For category 7, we similarly asked what liquid has been deformed (e.g. juice). For category 8, we asked what object has been placed in the context of another object (e.g. a [beer] in a [tree], or a [hot dog] in place of a [baby], in a lady's arms). This further information was used to ensure quality, by forcing annotators to think harder about the task.

Table 6 shows statistics about the atypical object annotations. The first row shows the number of images with at least one annotation in this category. We see that category 8 (context replacement) is the most popular category, followed by category 5 (hybrid object). We also examine annotator agreement. On the task of binary typical/atypical categorization, on average 82% of annotators agreed with the final majority vote label (not shown in table), where chance is 67% and perfect agreement is 100%. For categories 1-8, we computed agreement as follows. If any annotators provided category C for image I, we checked what fraction of all annotators who labeled I chose C. Thus, agreement ranges between 33% and 100%. We see reasonable annotator agreement, especially for categories 7 and 8, where on average, 2 annotators out of 3 provide the same label. Note that for 658 of 1,122 images deemed atypical after majority voting, at least 2 annotators agree on the chosen transformation category.

We next analyzed the parts and objects that co-occur in our atypical ads, using the human-provided fine-grained annotations. For texture replacement, a common combination was body parts textured with foods, animals, and text. Body parts textured with animal textures were common for beauty ads. Restaurants and chocolate ads often used a human inside a piece of food. The hybrid object category often featured human-food and human-animal combinations, but those had different semantics. Human-food hybrids were often used by chocolate and ketchup ads, and frequently evoked the "eager" sentiment. In contrast, human-animal hybrids were featured in animal rights ads, and evoked "alertness" rather than "eagerness". Further, food-nature hybrids were used in seasoning ads, and evoked "eagerness". Human body parts combined with trees/nature were used in environment protection ads. In terms of correlations between action-reason words and object transformations, we observe that to illustrate being "natural" or "beauty", a fruit, tree or flower might be part of the hybrid. "Exotic" products feature human-animal hybrids. Being "fun", "easy", "strong" or "tough"' might be illustrated through a human-sport equipment hybrid. Phone-human hybrids imply being "safe" or "reliable".

In the future, this data can be used to model the relationship between particular part/texture combinations, to better infer the topic and meaning of an ad. Alternatively, we can develop an application that guides non-expert users through the creation of ads. Given an ad topic (and optionally, sentiment to be projected), the app can propose possible object parts and combinations, and show users segments they can combine to create a full object, then perform some post-processing to smooth the result. This app can help automate prior work on creating visual metaphors [71].

3.7 Climax

While symbolism (Sec. 3.5) and atypicality (Sec. 3.6) are common in image ads, we found them less typical in videos. Since a video has more time to convey its message, we chose to examine the structure of the story temporally. Thus, we next annotate the climax in a video, as a cue for which parts of the video are poignant, similar to how symbolism indicates poignant image regions.

We obtained timestamps showing the moment when the climax, or most dramatic point, in a video occurs. In The Advertising Research Handbook [4], dramatic structure has four prototypical forms, and these depend on how positive and negative sentiment rises or declines. [4] examines product ads, and the changes in positive/negative sentiment are correlated with appearances of the brand. In PSAs, understanding the story often depends on understanding the climax of negative sentiment.

We collected climax annotations on a randomly chosen subset of videos from our dataset, using MTurk. We submitted each video 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 had to provide the minute and second at which climax occurs. To ensure quality, annotators were also asked to describe what happens at the end of the video. We ended up with 1,149 videos that contain climax annotations. We manually inspected a subset and found the timestamps were often quite reasonable and descriptions were detailed.

3.8 Effectiveness

Effectiveness annotations were provided on a scale from "very ineffective" (rating 1) to "very effective" (rating 5), and contained self-reported estimates of effectiveness from viewers. The overall effectiveness for a video is computed as the mode of the five different annotators' ratings for that video. This resulted in 193 samples for rating 1, 261 for 2, 1319 for 3 (neutral), 426 for 4, and 1278 for 5 (very effective). In experiments, to ensure class balance, the class with lowest count determined the number of sampled videos from each class. Therefore, 965 samples in total were used. For the binary task of predicting "effective or not", we removed annotations with value 3.

3.9 Challenges of collection and quality control

A data collection task of this magnitude presented challenges on three fronts: speed of collection, cost, and quality. For each kind of annotation, we started with a price based on the estimated time it took to complete the task. As results would come in, we would adjust this price to account for the actual time taken on average and, often, also to increase

the speed with which the tasks were being completed. Even after increasing the pay, some of the more difficult tasks, such as identifying symbolism and annotating actions and reasons, would still take a long time to complete. For symbolism, we offered a bonus to MTurkers who would do a large number of tasks in one day. We found this incentive quite effective in speeding up the collection. In total, collecting all annotations cost about $15,000.

For the tasks where MTurkers had to select options, such as topics and sentiments, we relied on a majority vote to disregard low-quality work. For action-reason statements, we used heuristics, the number of short or repetitive responses and number of non-dictionary words in the answers, to shortlist suspicious responses for manual examination. For symbolism, we manually reviewed a random subset of responses from each MTurker who did more than a pre-specified number of tasks in a day. For climax and atypical objects, we looked at free-form text to gauge the effort that the annotator was putting in the annotation process, and reject careless submissions accordingly. For all tasks, we also restricted the tasks to workers who exceeded a threshold of approval ratings (95-98%), over 1000+ submissions.

4 Approach

We focus on two key tasks. First, we learn to predict the action and reason that an ad prompts. Second, we model the structure of the story that an ad conveys. In the image space, this takes the shape of analyzing the non-photorealistic object portrayals that ads create to metaphorically relate properties. In the video space, we look at the climax and effectiveness of the ad story. Our methods are unified through two characteristics. First, they look for visually and semantically poignant regions in space (in images) or time (in videos). Second, they consider two modalities, visual and language: text (in images) or speech (in videos).

4.1 Inferring actions and reasons

We first present an approach that allows us to retrieve an appropriate description for an ad image or video, i.e. one that describes *what the viewer should do and why*, according to the ad. We learn an embedding space where we can evaluate the similarity between ad images (or videos) and ad messages (Sec. 4.1.1). We represent an ad image as a weighted average of its regions that are likely to make symbolic references (Sec. 4.1.2), the slogans extracted from the image (Sec. 4.1.3), and knowledge from recognized symbols and objects (Sec. 4.1.4). For an ad video, we use bag of frames to represent its visual content (Sec. 4.1.5), and we extract its speech information to complement the visual feature (Sec. 4.1.6). Our full approach is defined in Sec. 4.1.7, which is a fused model. In Sec. 5.1 we demonstrate the utility of each component.

In Hussain *et al.* [10], we modeled a similar task as a classification problem, where given a question "Why should the viewer do [action]?", our system responds with a single-word [reason] answer. However, using a single word is insufficient to capture the rhetoric of complex ads. On one hand, summarizing the full sentence using only one word is too challenging, for example, for the question "Q: Why should I buy authentic Adidas shoes?", the ground-truth answer "feet" used in [10] cannot convey the meaning of the full-sentence answer "Because it will protect my feet". Further, picking one word as the answer may be misleading and imprecise, for example, for the "Q: Why should I buy the Triple Double Crunchwrap?", picking "short" from the sentence "Because it looks tasty and is only available for a short time" is problematic. Thus, while in [11] we show that we outperform prior art on the original question-answering task from [10], here we focus on the sentence retrieval task.

In particular, we ask the system to pick which *action-reason statement* is most appropriate for the image (or video). We retrieve statements in the format: "I should [action] because [*reason*]." e.g. "I should speak up about domestic violence because *being quiet is as bad as committing violence yourself.*" We gather the three or five annotated statements paired with each image and video respectively, and we use all these labeled data as paired ground-truth. For each image, we use three related statements (i.e. statements provided by humans for this image) and randomly sample 12 unrelated statements (written for *other* images). The system must rank these 15 statements based on their similarity to the image. Similarly, for each video, we use all five related statements, and randomly sample 20 unrelated ones (annotated for *other* videos). This sampling strategy ensures that the naive solution of random guess achieves roughly 20% accuracy on both the image and the video tasks.

4.1.1 Basic cross-modal embedding

We first directly learn an embedding that optimizes for the retrieval/ranking task. We require that the similarity between an image (video) and its corresponding statement should be higher than the similarity between that image (video) and any other statement, or between other images (videos) and that statement. Thus, we minimize Eq. 1:

$$L(\boldsymbol{v},\boldsymbol{t};\boldsymbol{\theta}) = \sum_{i=1}^{K}\Big[\underbrace{\sum_{j\in N_{vt}(i)}\max\left(0,\frac{\boldsymbol{v}_i^\intercal\boldsymbol{t}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{t}_j\|}-\frac{\boldsymbol{v}_i^\intercal\boldsymbol{t}_i}{\|\boldsymbol{v}_i\|\|\boldsymbol{t}_i\|}+\beta\right)}_{\text{image (video) as anchor, rank statements}} + \underbrace{\sum_{j\in N_{tv}(i)}\max\left(0,\frac{\boldsymbol{t}_i^\intercal\boldsymbol{v}_j}{\|\boldsymbol{t}_i\|\|\boldsymbol{v}_j\|}-\frac{\boldsymbol{t}_i^\intercal\boldsymbol{v}_i}{\|\boldsymbol{t}_i\|\|\boldsymbol{v}_i\|}+\beta\right)}_{\text{statement as anchor, rank images (videos)}}\Big] \tag{1}$$

where K is the batch size; β is the margin of triplet loss; $\boldsymbol{v}$ and $\boldsymbol{t}$ ($\boldsymbol{v},\boldsymbol{t}\in\mathbb{R}^{200\times1}$) are the visual and textual embeddings we are learning, respectively; $\boldsymbol{v}_i$, $\boldsymbol{t}_i$ correspond to the same ad and $\frac{\boldsymbol{v}_i^\intercal\boldsymbol{t}_i}{\|\boldsymbol{v}_i\|\|\boldsymbol{t}_i\|}$ measures the *cosine similarity* between the paired visual and textual embeddings; $N_{vt}(i)$ is the negative statement set for the i-th image (video), and $N_{tv}(i)$ is the negative visual set for the i-th statement, defined in Eq. 2. These two negative sets involve the most challenging k' examples within the size-K batch. A natural explanation is that Eq. 2 seeks to find a subset $A\subseteq\{1,...,K\}$ which involves the k' most confusing examples.

$$\begin{aligned} N_{vt}(i) &= \underset{A\subseteq\{1,...,K\},|A|=k'}{\arg\max}\sum_{j\in A,i\neq j}\frac{\boldsymbol{v}_i^\intercal\boldsymbol{t}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{t}_j\|} \\ N_{tv}(i) &= \underset{A\subseteq\{1,...,K\},|A|=k'}{\arg\max}\sum_{j\in A,i\neq j}\frac{\boldsymbol{t}_i^\intercal\boldsymbol{v}_j}{\|\boldsymbol{t}_i\|\|\boldsymbol{v}_j\|} \end{aligned} \tag{2}$$

Text embedding. For the *action-reason statement* of an image, we use LSTM [72] to encode it to an embedding $\boldsymbol{t}$. Assume an input sequence of words $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$ where T denotes sequence length, each word $\boldsymbol{s}_t \in \mathbb{R}^{V\times 1}$ ($t \in \{1 \ldots T\}$, V is the vocabulary size) is a one-hot representation, and the word embedding process $\varphi_{emb}(\boldsymbol{s}_t) = \boldsymbol{w}_{emb}^{\intercal}\boldsymbol{s}_t$ (we initialize $\boldsymbol{w}_{emb} \in \mathbb{R}^{V\times 200}$ from GloVe [73]) encodes each word $\boldsymbol{s}_t$ to a 200-D word embedding $\varphi_{emb}(\boldsymbol{s}_t)$. We use the last hidden state $h_T \in \mathbb{R}^{200\times 1}$ as the textual embedding $\boldsymbol{t}$.

For the action-reason statement of a video, we use mean-pooling (bag of words) of the word embedding vectors to represent $\boldsymbol{t}$, due to the limited training set size.

Hard negative mining. Different ads might convey similar arguments, so the sampled negative may be a viable positive. For example, for a car ad with associated statement "I should buy the car because it's fast", a hard negative "I should drive the car because of its speed" (provided on another image) may also be proper. Using the k' most challenging examples in the size-K batch (Eq. 2) is our trade-off between using all and using only the most challenging example, inspired by [47], [74], [75].

4.1.2 Image embedding using symbol regions

Since ads are carefully designed, they may involve complex narratives with several distinct components, i.e. several regions in the ad might need to be interpreted individually first to decode the full ad's meaning. As we show in [11], the chosen regions should be those likely to serve as visual anchors for symbolic references (such as the rocket, blood or graduation cap in Fig. 2). Our intuition is that ads draw the viewer's attention in a particular way, and the symbol bounding boxes, *without* symbol labels, can be used to approximate this. Thus we use all the 13,938 images annotated with symbolic references, each with up to five bounding box annotations. We use the SSD object detection model [51] implemented by [76], pre-train it on the COCO [77] dataset, and fine-tune it with the symbol bounding box annotations [10], to obtain region proposals.

We use bottom-up attention [40], [78], [79] to aggregate the information from symbolic regions (see Fig. 6). Specifically, we use the Inception-v4 model [80] to extract CNN features for all symbol proposals $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ (we set $M = 10$, i.e., 10 proposals per image), resulting in $\{\phi_{cnn}(\boldsymbol{x}_1), \ldots, \phi_{cnn}(\boldsymbol{x}_M)\}$ where $\phi_{cnn}(\boldsymbol{x}_m) \in \mathbb{R}^{1536\times 1}$. Then, for each CNN feature $\phi_{cnn}(\boldsymbol{x}_m)$, fully-connected layers are applied to project it to: 1) a 200-D embedding vector $\boldsymbol{v}_m$ (Eq. 3, $\boldsymbol{w}_{img} \in \mathbb{R}^{1536\times 200}$), and 2) an importance score α_m (Eq. 4, $\boldsymbol{w}_{attn} \in \mathbb{R}^{1536\times 1}$). The final image representation $\boldsymbol{z}_{img}$ is a weighted sum of these region-based vectors (Eq. 5).

$$\boldsymbol{v}_m = \boldsymbol{w}_{img}^{\intercal}\phi_{cnn}(\boldsymbol{x}_m) \tag{3}$$

$$\alpha_m = \frac{\exp(\boldsymbol{w}_{attn}^{\intercal}\phi_{cnn}(\boldsymbol{x}_m))}{\sum_{l=1}^{M}\exp(\boldsymbol{w}_{attn}^{\intercal}\phi_{cnn}(\boldsymbol{x}_l))} \tag{4}$$

$$\boldsymbol{z}_{img} = \sum_{m=1}^{M}\alpha_m\boldsymbol{v}_m \tag{5}$$

The loss to learn the image-text embedding is the same as Eq. 1, defined using the region-based image representation $\boldsymbol{z}_{img}$ as $\boldsymbol{v}$: $L(\boldsymbol{z}_{img}, \boldsymbol{t}; \boldsymbol{\theta})$.

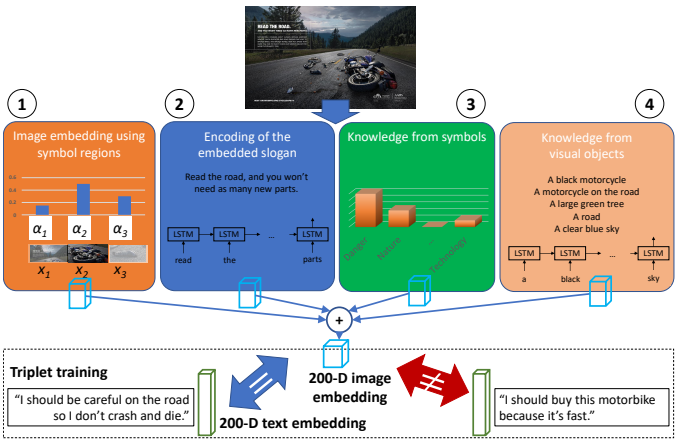


Fig. 6: Our image embedding model. In the image branch (1), multiple image symbolic anchors are proposed. Attention weighting is applied, and the image is represented as a weighted combination of the regions. The knowledge branch (3) predicts the existence of symbols and maps these to the 200-D embedding. For both the slogan (2) and visual objects captions (4) branches, we use LSTM to model the phrases. Pointwise addition is applied to fuse the features from four different modalities. We then perform triplet training to learn such an embedding space that keeps images close to their matching action-reason statements.

We demonstrate in [11] that (1) learning a region proposal network with attention, and (2) learning from symbol bounding boxes, greatly help the statement retrieval task. In particular, statement ranking results are worse if we use a generic pre-trained region proposal network. We argue general-purpose object detection models cannot capture nuance in ads since they ignore uncommon or abstract objects.

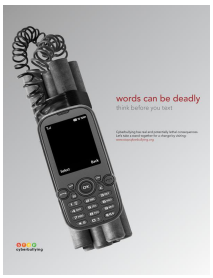
4.1.3 Encoding of the embedded slogan

In most cases, the image alone does not tell the full story of the ad, and may be (intentionally) ambiguous [81]. Thus, we also need to consider the *slogan* embedded in the ad to accurately retrieve statements about it. Two examples are shown in Fig. 7. It is clear that the ad understanding task becomes easier if we can read the slogans in both images, namely "words can be deadly." and "Human Trafficking. Don't ignore it". Inspired by these examples, we design a method to read the slogan information and thus improve the performance of inferring actions and reasons.

Given an image, we first use the Optical Character Recognition (OCR) functionality of the Google Cloud Vision API [82] to extract the text in the ad. We concatenate all the detected pieces into one. About half of the ads have up to 20 detected tokens (usually a word or part of a word). One-fifth has between 20 and 50 tokens, 14% has between 50 and 100 tokens, and the rest of the ads have over 100 tokens. We use a standard LSTM model to obtain a slogan embedding which results in $\boldsymbol{z}_{slg} \in \mathbb{R}^{200\times 1}$. The model using only the slogan modality can be trained using $L(\boldsymbol{z}_{slg}, \boldsymbol{t}; \boldsymbol{\theta})$ in Eq. 1.

4.1.4 Knowledge from symbols and visual objects

We next exploit the symbol labels which are part of [10]. Symbols are abstract words such as "freedom" and "happi-



(a) "words can be deadly. think before you text"
QA: I should be careful what I say because words can hurt like any weapons.

(b) "Human Trafficking. Don't ignore it."
QA: I should be aware of human trafficking because it is not always obvious.

Fig. 7: Example slogans from the image ads dataset. Both images require reasoning which makes the task challenging even for a human. However, given the slogan text information, understanding the message of the ads becomes easier.

ness" that provide additional information humans sense in the ads. Directly using the annotated symbols is unfeasible. On one hand, the symbol words labeled in the dataset are long-tailed because of the free-form annotations. On the other hand, the symbol annotations are available for only 13,938 images, and non-annotated images do not definitively lack the symbols. Therefore, our method of using symbol labels requires the training of a symbol classifier that generalizes beyond the annotated images. To make the learning more feasible, instead of training classifiers on all symbols, we base our work on the 53 symbol clusters in [10]. We learn a multilabel classifier $\boldsymbol{u}_{symb} \in \mathbb{R}^{1536\times 53}$ to obtain a symbol distribution $\sigma(\boldsymbol{u}_{symb}^{\intercal}\phi_{cnn}(\boldsymbol{x}))$ given the feature $\phi_{cnn}(\boldsymbol{x}) \in \mathbb{R}^{1536\times 1}$, where σ is *sigmoid*.

To use the additional knowledge (i.e., classifier $\boldsymbol{u}_{symb}$) regarding the symbols, we use a fully connected layer to project the symbol distribution to the joint embedding feature space, resulting in $\boldsymbol{z}_{symb} \in \mathbb{R}^{200\times 1}$:

$$\boldsymbol{z}_{symb} = \boldsymbol{w}_{symb}^{\intercal}\sigma(\boldsymbol{u}_{symb}^{\intercal}\phi_{cnn}(\boldsymbol{x})) \tag{6}$$

Much like symbols, the objects found in an image are quite telling of the message of the ad. For example, environment ads often feature animals, safe driving ads feature cars, beauty ads feature faces, drink ads feature bottles, etc. Since the Ads Dataset contains insufficient data to model object categories, we use DenseCap [39] to bridge the objects defined in Visual Genome [83] to the ads reasoning statements. More specifically, we use the DenseCap model to generate image captions and treat these as pre-fetched knowledge. For example, the caption "woman wearing a black dress" provides extra information about the objects in the image: "woman" and "black dress". Modeling the visual objects is similar to modeling the slogan (Sec. 4.1.3). We concatenate all the captions generated by DenseCap into a long textual description, then use an LSTM model to encode the sequence, resulting in objects-text embedding $\boldsymbol{z}_{obj}$.

In our setting, word embedding weights are not shared among the three vocabularies (ads statement, slogan, and DenseCap words). Our intuition is that the meaning of the same surface words may vary in these domains thus they need to have different representations. The parameters of the LSTM models are also not shared, because the grammar and the sentence structures vary in the three domains.

4.1.5 Video embedding model

For the video domain, we treat the video as a *bag of frames* (BOF) which ignores the sequence order, due to the limited size of our video ads dataset. Consider a frame sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R$ of a video where R is the total number of frames. Given that we sample 1 frame per second, R also equals to the time duration measured in seconds. We use Inception-v4 [80] to extract frame features, resulting in $\phi_{cnn}(\boldsymbol{x}_1), \ldots, \phi_{cnn}(\boldsymbol{x}_R)$. The BOF model uses mean-pooling to represent the video contents. Eq. 7 shows the details where $\boldsymbol{w}_{vdo} \in \mathbb{R}^{1536\times 200}$ are the parameters.

$$\boldsymbol{z}_{vdo} = \frac{1}{R}\sum_{r=1}^{R}(\boldsymbol{w}_{vdo}^{\intercal}\phi_{cnn}(\boldsymbol{x}_r)) \tag{7}$$

Given the video representation $\boldsymbol{z}_{vdo} \in \mathbb{R}^{200\times 1}$, the basic video model that retrieves statements can be trained using $L(\boldsymbol{z}_{vdo}, \boldsymbol{t}; \boldsymbol{\theta})$ similar to Eq. 1. However, the image feature $\boldsymbol{v}$ is substituted with the video feature. Due to size of the dataset, we also change to using bag of words (mean-pooling of the word embeddings) to encode the *action-reason* statement to avoid overfitting.

4.1.6 Video embedding using speech-to-text

There is more information than the visual frames that can help distinguish the contents. For example, the audio may involve different styles of music and the speech may directly convey the ad messages. We focus on the speech information since we think they are better suited for the task of retrieving the action-reason statements.

Given a video, we use FFMPEG [84] to extract the audio track. We then invoke the Google Cloud Speech-to-text API [85] to extract text from the audio data. After getting the text tokens, we concatenate them to a single sentence and use mean-pooling during training to aggregate individual word embeddings, resulting in $\boldsymbol{z}_{spch} \in \mathbb{R}^{200\times 1}$. Training this individual model uses $L(\boldsymbol{z}_{spch}, \boldsymbol{t}; \boldsymbol{\theta})$.

4.1.7 ADVISE: our final model

Our final **AD**s **VI**sual **S**emantic **E**mbedding model uses late fusion (we use *pointwise-add*) to combine the components. For the full model on the image ads, components described in Sec 4.1.2, 4.1.3, 4.1.4 are used, and we optimize $L(\boldsymbol{z}_{img} + \boldsymbol{z}_{slg} + \boldsymbol{z}_{symb} + \boldsymbol{z}_{obj}, \boldsymbol{t}; \boldsymbol{\theta})$. For video ads, Sec 4.1.5 and 4.1.6 are used, and we optimize $L(\boldsymbol{z}_{vdo} + \boldsymbol{z}_{spch}, \boldsymbol{t}; \boldsymbol{\theta})$.

Note that a version of this method for images only appeared in [11]. However, that method did not include the slogan representation, and incorporated symbol/object knowledge in a different manner. Further, the method for retrieving action-reason for videos is completely new.

4.2 Predicting topics and sentiments

A simpler version of understanding the "message" of an ad is to simply predict what it is about. Thus, we also learn to predict the topic of an ad, and what sentiment it aims to evoke in the viewer. We treat these as categorization

tasks. We choose the most frequent topic/sentiment per image/video, among the multiple annotations provided, as the ground-truth label. We then train 152-layer ResNets [67] to discriminate between our 38 topics and 30 sentiments.

For video, we train similar predictors, but using action-based representations. We believe the actions in the video ads may have significant impact on understanding the ads. Due to the smaller size of our dataset, we only experiment with a simple approach to predict topic and sentiment. We use the C3D network [86], [87] originally used for action recognition as a feature extractor. It is pre-trained on Sports-1M [87] and fine-tuned on UCF101 [88]. We convert videos into frames, and take consecutive 16 frames as a clip. We extract fc6 and fc7 features for each clip and average the features for all clips within the same video, as our final video representation. We train separate multi-class SVMs to distinguish between our 38 topics and 30 sentiments.

4.3 Analyzing story structure and effectiveness

We next propose an approach for understanding the structure of the story told in a video ad. Specifically, we aim to detect when the climax of a video occurs. Climax can be used in downstream tasks, e.g. sentiment prediction, as we show in [12]. Next, we use features related to climax, to infer the effectiveness of the video ad.

To predict climax, we use the climax annotation dataset (Sec. 3.7). We model climax using a variety of cues, both non-semantic (e.g. dynamics, audio) and semantic ones (e.g. predicted places, objects, expressions):

- **Audio amplitude** $\boldsymbol{a}^k$, 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.
- **Shot boundaries**, equal to 0 or 1 depending on whether a shot boundary occurs in the k-th frame. We use [89] for shot boundary extraction. In order to obtain more informative cues, we vary the parameters of [89] to get five 0/1 predictions per frame and use this 5D prediction $\boldsymbol{b}^k$ as the representation for the k-th frame.
- **Optical flow magnitude** $\boldsymbol{o}^k$, 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, extracted using [90].
- **Facial expressions**: We observed that the response 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 change might correlate with climax because something disturbing must have happened. Thus, we first detect the faces using OpenFace [91]. We then extract the expression of each face using an Inception model [92] trained on the AffectNet dataset [93]. 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 $\boldsymbol{fa}^k$.
- **Setting**, i.e. the type of place/scene. Let $\boldsymbol{vp} = \{p_1,\ldots,p_{365}\}$ be the vocabulary of places in the Places365 dataset [94]. We use a pre-trained prediction model from [94] to obtain a 365D vector $\boldsymbol{pl}^k = [l_1^k,\ldots,l_{365}^k]$, where l_i^k is the probability that the k-th frame exemplifies the i-th place.
- **Objects**: Let $\boldsymbol{vo} = \{c_1,c_2,\ldots,c_{80}\}$ be the vocabulary of the COCO object detection dataset [77]. We use the model of [76] trained on COCO to get the objects in a frame. We then use max-pooling to obtain an 80D fixed-length feature vector $\boldsymbol{ob}^k = [s_1^k,\ldots,s_{80}^k]$, where s_i^k is the maximum confidence score among multiple instances of the same object class c_i, in frame k.

Finally, 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 for the k-th frame are ResNet features $\boldsymbol{x}_k \in \mathbb{R}^{2048\times 1}$, optical flow magnitude $\boldsymbol{o}_k \in \mathbb{R}^{1\times 1}$, shot boundary indicators $\boldsymbol{b}_k \in \mathbb{R}^{5\times 1}$, sound amplitude $\boldsymbol{a}_k \in \mathbb{R}^{1\times 1}$, place representation $\boldsymbol{pl}_k \in \mathbb{R}^{365\times 1}$, object representation $\boldsymbol{ob}_k \in \mathbb{R}^{80\times 1}$, and facial expressions $\boldsymbol{fa}_k \in \mathbb{R}^{10\times 1}$. We previously presented this method in [12].

Predicting effectiveness. From the point of view of the company that ordered an ad, the message of the ad might matter only in so much as the message makes the ad effective. To predict effectiveness, we use the features used for climax. We also add aggregate features that do not depend on the frame: the duration of the video, the average hue, the ground-truth topic and sentiment labels (encoded with one-hot representation), and whether the video is "funny" or "exciting" (binary labels). We found that of the climax features, only audio and a places distribution seem to be predictive of effectiveness. Due to the sparse labeled data (less than 1000 samples, see Sec. 3.8), we combine the features using AdaBoost [95].

5 Experiments

We show results on seven tasks. First, we examine our method's ability to rank action and reason statements (*what the viewer should do* and *why*). This is our core task that gauges how well the method understands the message of an ad. Second, we take an in-depth look at the story of the ad. We look for attractive visual content temporally (*climax*) and statically within image ads (*atypical objects*). We also show how well we can predict the *effectiveness* of video ads. We conclude with two simplified classification tasks, predicting the *topic* and *sentiment* of the ad. All experiments are conducted on the dataset described in Sec. 3 using a held-out set of the appropriate annotations.

5.1 Retrieving action/reason statements

For the task of inferring the actions and reasons of the image ads, we use the splits defined in [96]. We use the 51,223 *trainval* images which are paired with 161,557 annotated statements for training; and evaluate on the 12,805 *test* images paired with 40,178 statements. We use TensorFlow [97] to build our model. We use a learning rate of 0.001, and the RMSProp optimizer with 0.95 decay and 1e-8 momentum. We use a batch size of 128, and all models are trained for

TABLE 7: Action-reason statement ranking results; high Accuracy and low Min Rank is desired. **Bold** is best, *Italics* is second- and third-best.

	Accuracy		Min Rank	
Method	Product	PSA	Product	PSA
IMAGE ONLY	0.630	0.491	1.836	2.214
SLOGAN ONLY	0.791	0.677	1.599	1.788
IMAGE+SLOGAN	**0.847**	*0.712*	*1.320*	*1.635*
IMAGE+SYMBOLS	0.640	0.489	1.764	2.241
FULL METHOD	**0.847**	**0.733**	**1.282**	**1.554**
FULL METHOD (BOW)	*0.827*	*0.718*	*1.318*	*1.588*

roughly 60 epochs. To choose the best model, we use a held-out validation set with approximately 20% *trainval* data. For the similar action/reason ranking task on the video data, we split the 3,477 videos into trainval/test sets (80%/20%), resulting in 2,777 *trainval* and 700 *test* videos. We use a learning rate of 0.003 and roughly 170 epochs (3,000 steps). The remaining details are as for the image task.

We evaluate to what extent our proposed method (Sec. 4.1) is able to match an ad to its intended message; the message contains both the action and reason. We compare the following ablations of our method. All but the last one use an LSTM to encode the action-reason statements.

- IMAGE ONLY uses region proposals trained from our symbolism data, without symbol labels (Sec. 4.1.2).
- SLOGAN ONLY is the method that uses OCR to extract the slogan embedded in the image (Sec. 4.1.3).
- IMAGE+SLOGAN combines the image and slogan by optimizing $L(\boldsymbol{z}_{img} + \boldsymbol{z}_{slg}, \boldsymbol{t}; \boldsymbol{\theta})$.
- IMAGE+SYMBOLS uses additional knowledge from pre-trained multi-label symbol classifier $\boldsymbol{u}_{symb}$ (Sec. 4.1.4), and optimizes $L(\boldsymbol{z}_{img} + \boldsymbol{z}_{symb}, \boldsymbol{t}; \boldsymbol{\theta})$.
- FULL METHOD combines the region-based image representation, slogan, symbol and object (Sec. 4.1.7).
- FULL METHOD (BOW) is the same as the previous method but uses bag of words representation, i.e. we average the individual word embeddings to get the full-text embedding (for statement, slogan, and object).

In [11], we demonstrated that IMAGE ONLY greatly outperforms prior work such as [47], [48].

We evaluate the ablations in terms of two metrics: Accuracy which is the percentage of correct top-1 predictions; and Min Rank, which is the averaged ranking value of the best-ranked true matching statement (best possible rank is 1). We expect a good model to have high Accuracy and low Min Rank scores. We show results separately for product and public service announcement (PSA) ads, as in [11].

The results are shown in Tab. 7. The most important two modalities, as we expected, are the image (IMAGE ONLY) and slogan (SLOGAN ONLY). Adding symbols to the image representation helps for products, but less than adding the slogan. Interestingly, the slogan extracted from the image seems to be more helpful than the image itself. This is likely because the slogan is more straight-forward than the image. The image is designed to be attractive and may intentionally be ambiguous in isolation. Further, there is larger variance in how a particular message (e.g. "Don't smoke") might be visually portrayed (e.g. guns, body parts, burned textures) compared to the slogan text variance. Thus, our action-reason annotators rely on the slogan and may even borrow individual words in their annotations; still human-provided statements have very little overlap as the annotators need to *interpret* the slogan when writing the action-reason statements. For many ads, the slogan alone is ambiguous too (e.g. "winter collection" may be a fashion ad or a homelessness PSA). This explains why the fusion of both image and slogan modalities (IMAGE+SLOGAN) outperforms both IMAGE ONLY and SLOGAN ONLY. In particular, IMAGE+SLOGAN outperforms IMAGE ONLY by 34% on Product ads and 45% on PSAs, in terms of accuracy. This greater improvement on PSAs might be because these are less intuitive and more "clever" than product ads, thus hints from the slogan are more important. The inclusion of objects and symbols in our full method (FULL METHOD) improves the accuracy of IMAGE+SLOGAN on PSAs by 3%. Finally, aggregating text information using averaging (FULL METHOD (BOW)) provides slightly worse results compared to FULL METHOD, i.e. accuracy reduced by 2% on product ads and PSAs.

TABLE 8: Ranking action and reason statements separately, vs action-reason together. All methods shown use BOW. Numbers denote Min Rank (lower is better).

Method	Action-Reason	Action	Reason
IMAGE ONLY	1.755	2.007	2.157
SLOGAN ONLY	1.532	1.758	1.845
IMAGE+SLOGAN	**1.300**	**1.521**	**1.696**

We next break down the action-reason ranking task into two tasks, separately ranking action and reason. The results are shown in Tab. 8. In general, the task of ranking the combined action-reason is the easiest one since it only requires the model to be confident about either the action or the reason. The additional image information (IMAGE+SLOGAN VS SLOGAN ONLY) gives 18% reduction in rank while the extra slogan message (IMAGE+SLOGAN VS IMAGE ONLY) reduces rank by 35%. The action statement ranking is the second-easiest. Using the image gives 16% performance gain over slogan only. Predicting the reason statement is the most challenging, and offers the most limited room for improvement when using the image (9% over slogan only).

Finally, we show the performance on ranking statements for video ads (rather than image ads as before). We use the same metrics. We compare the following methods:

- FRAME ONLY (BOF) is the model that only uses the video representation (Sec. 4.1.5).
- SPEECH ONLY (BOF) only uses the text information extracted by speech recognition (Sec. 4.1.6).
- FRAME+SPEECH (BOF) combines the bag-of-frames encoded video and speech by optimizing $L(\boldsymbol{z}_{vdo} + \boldsymbol{z}_{spch}, \boldsymbol{t}; \boldsymbol{\theta})$ (Sec. 4.1.7). This is our final model.
- FRAME+SPEECH (LSTM) is similar but uses LSTM to encode all modalities (video, speech, and statement).

The results are shown in Tab. 10. Unlike the scenario in the image task, the directly detected spoken language (SPEECH ONLY (BOF)) is less useful than the pure visual cue (FRAME ONLY (BOF)); the visual feature is 10% better in terms of accuracy. This implies the ads designers did not put many unambiguous explanations in the conversation. However, we see that the conversation does help improve understanding, when used in combination with the frames: in terms of accuracy, FRAME+SPEECH (BOF) is 14% better

TABLE 9: Effectiveness prediction using the most promising frame-level climax features (last two) and video-level features.

Features/Method	Chance	Topic	Sentiment	Funny	Exciting	Hue	Duration	Audio	Places	Boosting
Accuracy (binary)	0.5000	0.5627	0.5386	0.5186	*0.5969*	0.5151	0.5562	0.5246	0.5224	**0.6131**
Accuracy (five-way)	0.2000	0.2094	0.2435	0.2416	*0.2461*	0.2300	0.2281	0.2195	0.2153	**0.2683**

TABLE 10: Ranking action-reason statements for video ads.

Method	Accuracy	Min Rank
FRAME ONLY (BOF)	0.560	*2.401*
SPEECH ONLY (BOF)	0.507	2.987
FRAME+SPEECH (BOF)	**0.639**	**2.053**
FRAME+SPEECH (LSTM)	*0.561*	2.547

TABLE 11: Climax prediction, using top-3 accuracy.

Method	w/in 0s	w/in 1s	w/in 2s
ResNet only	0.190	0.400	0.523
Audio (no training)	0.178	0.403	0.534
Ours	**0.226**	**0.439**	**0.546**

than FRAME ONLY (BOF). Finally, we compare the BOF approach to a more complex model with more learnable parameters (FRAME+SPEECH (LSTM)). We see that this latter models is 12% worse than the simpler BOF version. We surmise that the reason is the limited size of the video set. Note when we use an LSTM to represent the visual information, results are similar to the BOF version; the drop comes from the LSTM representation of the *text* information.

5.2 Understanding story structure and effectiveness

We next dive into the structure of video ads, by predicting when the climax occurs, and how effective the video may be, according to annotators. At each frame, the model in Sec. 4.3 predicts a real value ranging from $[0, 1]$ denoting whether the frame contains climax. We then use the sigmoid cross entropy loss to constrain the model to mimic the human annotations. Due to the size of the dataset, we set the input and output dropout keep probabilities of the LSTM cell to 0.5 to avoid over-fitting. We use RMSprop with 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 to pick the best model on the validation set. We sample frames at 1 sec intervals.

We measure the recall of the top-k predictions ($k = 1, 3$). 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.

We show in Table 11 that using our proposed features, climax can be predicted much more effectively. Further, if we simply extract the top-k maximal responses in the audio channel, and predict these as climax frames, without training, we achieve competitive results; in two of three columns, we actually outperform the trained visual-only ResNet using this approach. Note that climax is a simpler task than action-reason retrieval because at each frame, we need to predict a mapping from hidden layers to a 1D value (1 or 0), thus an LSTM works well. In contrast, for action-reason retrieval on video, we need to predict a mapping from hidden features to 200D.

TABLE 12: Topic and sentiment prediction.

Domain/Task	Topics	Sentiments
Image	0.603	0.279
Video	0.351	0.328

We next show the impact of the same features, along with a few additional features, to predict how effective a video story might be. We show a binary task and five-way task (of predicting the exact score, 1 through 5). We measure accuracy in both cases. We show the results in Table 9. We see that predicting effectiveness is generally a challenging task, and performance is low overall. The strongest features are excitement, topics, sentiment and duration. The boosted combination of features is the strongest, achieving 9% improvement over the single best method in the five-way task.

5.3 Recognizing objects with atypical texture

We next describe preliminary results that use our object transformation annotations. We train a VGG16 to distinguish between ads containing typical versus atypical objects. For training, we use the preliminarily labeled set of potentially atypical and certainly typical images; we use 22,568 images total, with equal number from each category. For validation and test, we use the data annotated by MTurk workers. To test, we use 1,106 images (553 for each category) on which all three annotators agree on the typical/atypical label. We use the Adam optimizer with a learning rate 10^{-5}. We thus achieve 66.91% accuracy on the test set. This is significantly higher than a random guess, indicating that atypicality can be visually distinguished. We do not pursue categorization of the eight transformation types (Sec. 3.6) because we find that these are not mutually exclusive. Using more sophisticated methods that explicitly model object context to recognize atypicality, or recognize semantic inconsistencies in the co-occurring objects, is the subject of our future work.

5.4 Predicting topics and sentiments

We conclude with results on two tasks that are a simplification of understanding the message of the ad. We predict the topic of the ad and the sentiment that it aims to provoke. We show the results in Table 12, for both images and video. We see that even though there are fewer sentiments in the vocabulary (30 as opposed to 38 as for topics), sentiment prediction is a more challenging task.

6 CONCLUSIONS

In this paper, we presented our dataset and approaches for understanding the messages that advertisements convey. We used approaches that rely on spatial and temporal attention, and combine cues from multiple modalities. We tested these methods on the tasks of predicting what the viewer

should do and why, measuring when the climax in a video occurs, how effective a video ad is deemed, and how objects are transformed to metaphorically convey properties.

Thus far we have approached ad understanding using supervised approaches and annotations primarily collected for the tasks of interest. However, ads reside in a broader media content, and on associations humans have acquired over the years. In particular, ads rely on human experiences about how life situations evolve, what properties different objects typically have, how other people behave, etc. Thus, in future work, our key focus will be to incorporate information and knowledge from external sources. On one hand, we will complement the cues elicited from the visual and textual channels, with context about the recognized objects in knowledge bases. Second, we will examine the context in which these objects and visuals appear in other media, such as newspapers, film, and non-ad YouTube videos. We believe this context will enable deeper and more accurate understanding of the messages of ads.

ACKNOWLEDGMENTS

The authors would like to thank Christopher Thomas, Xiaozhong Zhang, Zuha Agha, Nathan Ong, and Kyle Buettner for early work on this project, and Sanchayan Sarkar for help with the face data. This material is based upon work supported by the National Science Foundation under Grant Number 1566270. It was also supported by Google Faculty Research Awards and an NVIDIA hardware grant.

REFERENCES

[1] N. Hollis, "Why good advertising works (even when you think it doesn't)," August 2011, https://www.theatlantic.com/business/archive/2011/08/why-good-advertising-works-even-when-you-think-it-doesnt/244252/.

[2] C. L. Muñoz and T. L. Towner, "The image is the message: Instagram marketing and the 2016 presidential primary season," *Journal of Political Marketing*, vol. 16, no. 3-4, pp. 290–318, 2017.

[3] K. Liebhart and P. Bernhardt, "Political storytelling on instagram: Key aspects of alexander van der bellens successful 2016 presidential election campaign," *Media and Communication*, vol. 5, no. 4, pp. 15–25, 2017.

[4] C. E. Young, *The advertising research handbook*. Ideas in Flight, 2008.

[5] X. Xu, R. L. Alexander, S. A. Simpson, S. Goates, J. M. Nonnemaker, K. C. Davis, and T. McAfee, "A cost-effectiveness analysis of the first federally funded antismoking campaign," *American journal of preventive medicine*, vol. 48, no. 3, pp. 318–325, 2015.

[6] B. Cosgrove, "The photo that changed the face of aids," November 2014, http://time.com/3503000/behind-the-picture-the-photo-that-changed-the-face-of-aids/ .

[7] M. O'Neill, "Old spice response campaign was more popular than obama," August 2015, https://www.adweek.com/digital/old-spice-response-campaign/.

[8] J. Battista, "Roger goodell, nfl rightly correct course with change in policy," August 2014, http://www.nfl.com/news/story/0ap3000000384987/article/roger-goodell-nfl-rightly-correct-course-with-change-in-policy.

[9] D. R. Winslow, "The pulitzer eddie adams didnt want," April 2011, https://lens.blogs.nytimes.com/2011/04/19/the-pulitzer-eddie-adams-didnt-want/.

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

[11] K. Ye and A. Kovashka, "Advise: Symbolism and external knowledge for decoding advertisements," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[12] K. Ye, K. Buettner, and A. Kovashka, "Story understanding in video advertisements," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[13] F. Tramèr, P. Dupré, G. Rusak, G. Pellegrino, and D. Boneh, "Ad-versarial: Defeating perceptual ad-blocking," *arXiv preprint arXiv:1811.03194*, 2018.

[14] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] P. Resnick, R. K. Garrett, T. Kriplean, S. A. Munson, and N. J. Stroud, "Bursting your (filter) bubble: strategies for promoting diverse exposure," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW) Companion*, 2013.

[16] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[17] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[19] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu, "Visual persuasion: Inferring communicative intents of images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] J. Joo, F. F. Steen, and S.-C. Zhu, "Automated facial trait judgment and election outcome prediction: Social dimensions of face," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[21] Y. Peng, "Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision," *Journal of Communication*, vol. 68, no. 5, pp. 920–941, 2018.

[22] C. Thomas and A. Kovashka, "Persuasive faces: Generating faces in advertisements," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[23] D. Won, Z. C. Steinert-Threlkeld, and J. Joo, "Protest activity detection and perceived violence estimation from social media images," in *Proceedings of the ACM Conference on Multimedia*, 2017.

[24] Z. Bylinskii, S. Alsheikh, S. Madan, A. Recasens, K. Zhong, H. Pfister, F. Durand, and A. Oliva, "Understanding infographics through textual and visual tag prediction," *arXiv preprint arXiv:1709.09215*, 2017.

[25] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[26] M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. Daume, III, and L. S. Davis, "The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern, "Visual appearance of display ads and its effect on click through rate," in *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2012.

[28] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam, "Multimedia features for click prediction of new ads in display advertising," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[29] K. Yadati, H. Katti, and M. Kankanhalli, "Cavva: Computational affective video-in-video advertising," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 15–23, 2014.

[30] T. Mei, L. Li, X.-S. Hua, and S. Li, "Imagesense: towards contextual image advertising," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 8, no. 1, p. 6, 2012.

[31] J. M. Sánchez, X. Binefa, and J. Vitrià, "Shot partitioning based recognition of tv commercials," *Multimedia Tools and Applications*, vol. 18, no. 3, pp. 233–247, 2002.

[32] J. M. Gauch and A. Shivadas, "Finding and identifying unknown commercials using repeated video sequence detection," *Computer Vision and Image Understanding*, vol. 103, no. 1, pp. 80–88, 2006.

[33] R. J. Sethi, Y. Gil, H. Jo, and A. Philpot, "Large-scale multimedia content analysis using scientific workflows," in *Proceedings of the ACM International Conference on Multimedia*, 2013.

[34] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. Picard, "Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads," *IEEE Transactions on Affective Computing*, 2014.

[35] J. Williamson, *Decoding advertisements*, 1978.

[36] P. Messaris, *Visual persuasion: The role of images in advertising*. Sage, 1997.

[37] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[38] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[42] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[45] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[47] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improved visual-semantic embeddings," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[48] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[49] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[51] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[52] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[53] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] B. Saleh, A. Elgammal, J. Feldman, and A. Farhadi, "Toward a taxonomy and computational models of abnormalities in images," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[55] P. Wang, L. Liu, C. Shen, Z. Huang, A. van den Hengel, and H. Tao Shen, "What's wrong with that object? identifying images of unusual objects by modelling the detection score distribution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[56] M. J. Choi, A. Torralba, and A. S. Willsky, "Context models and out-of-context objects," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 853–862, 2012.

[57] H. Stipp, "How context can make advertising more effective," vol. 58, no. 2, pp. 138–145, 2018.

[58] S. Singh and C. Cole, "The effects of length, content, and repetition on television commercial effectiveness," *Journal of Marketing Research*, vol. 30, pp. 91–104, 02 1993.

[59] H. Li and H.-Y. Lo, "Do you recognize its brand? the effectiveness of online in-stream video advertisements," *Journal of Advertising*, vol. 44, no. 3, pp. 208–218, 2015.

[60] T. Teixeira, M. Wedel, and R. Pieters, "Emotion-induced engagement in internet video advertisements," *Journal of Marketing Research*, vol. 49, no. 2, pp. 144–159, 2012.

[61] V. Ramalingam, B. Palaniappan, N. Panchanatham, and S. Palanivel, "Measuring advertisement effectivenessa neural network approach," *Expert Systems with Applications*, vol. 31, no. 1, pp. 159 – 163, 2006.

[62] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler, "Moviegraphs: Towards understanding human-centric situations from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[63] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "Rolenet: Movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[64] A. Liu, J. Li, Y. Zhang, S. Tang, Y. Song, and Z. Yang, "An innovative model of tempo and its application in action scene detection for movie analysis," in *Proceedings of Winter Applications of Computer Vision (WACV)*, 2008.

[65] Z. Rasheed and M. Shah, "Movie genre classification by exploiting audio-visual features of previews," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2002.

[66] J. Choi, T.-H. Oh, and I. So Kweon, "Video-story composition via plot analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[68] C. Groden, "This is how much a 2016 Super Bowl ad costs," http://fortune.com/2015/08/06/super-bowl-ad-cost/.

[69] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," Master's thesis, Upper Austria University of Applied Sciences, Austria, 2010.

[70] K. Poels and S. Dewitte, "How to capture the heart? Reviewing 20 years of emotion measurement in advertising," *Journal of Advertising Research*, vol. 46, no. 1, pp. 18–37, 2006.

[71] L. B. Chilton, S. Petridis, and M. Agrawala, "Visiblends: A flexible workflow for visual blends," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.

[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[73] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[74] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[75] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[76] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[77] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

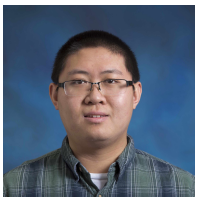
[78] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[79] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[80] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017. [Online]. Available: https://arxiv.org/abs/1602.07261
[81] M. Zhang, R. Hwa, and A. Kovashka, "Equal but not the same: Understanding the implicit relationship between persuasive images and text," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.
[82] "Google cloud vision api," https://cloud.google.com/vision/.
[83] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.
[84] "Ffmpeg," https://www.ffmpeg.org/.
[85] "Google cloud speech-to-text api," https://cloud.google.com/speech-to-text/.
[86] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
[87] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
[88] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[89] B. Castellano, "Pyscenedetect," https://github.com/Breakthrough/PySceneDetect/.
[90] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[91] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016.
[92] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[93] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, 2017.
[94] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
[95] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[96] "Automatic understanding of visual advertisements," https://evalai.cloudcv.org/web/challenges/challenge-page/86/overview.
[97] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

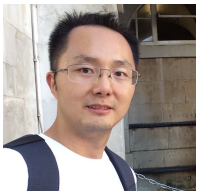
Narges Honarvar Nazari is a third-year Ph.D. student in Computer Science at the University of Pittsburgh. Her interest areas are computer vision, image processing and natural language processing. She received her Masters in Computer Sciences from the University of Southern California (2014-2015) and her Bachelors from Shahid Beheshti University (2008-2012).

James Hahn graduated with his BPhil in Computer Science at the University of Pittsburgh. He is now a Masters student at Georgia Tech. He is interested in computer vision, human-visual reasoning, knowledge representation, and neurocryptography. He has completed four software engineering internships, with the most most recent at Hulu in Santa Monica.

Zaeem Hussain is a fourth-year Ph.D. student in Computer Science at the University of Pittsburgh. Before Pitt, he was a research intern at the University of Queensland. He has a Masters degree in Computer Science from Lahore University of Management Sciences (2011-2014) and one in Applied Mathematics from University of Washington (2012-2013). He got his Bachelors in Electrical Engineering from University of Engineering and Technology, Lahore (2007-2011).



Mingda Zhang is a third-year Ph.D. student in Computer Science at University of Pittsburgh. His research focus is on the intersection of computer vision and natural language processing. He has completed research internships in Google AI, Seattle. Before coming to Pitt, he obtained his B.Sc. in Chemical Biology in 2013 from Peking University, China.



Keren Ye is a fourth-year Ph.D. student in Computer Science at the University of Pittsburgh (Pitt). His interests lie broadly in computer vision and natural language processing, including multi-modal learning, knowledge representation, and weakly supervised object detection. Before studying at Pitt, he worked as a software engineer at Baidu Inc. for 5 years. He got both of his Bachelors and Masters degrees (2004-2011) from Beihang University, China.

Adriana Kovashka is an Assistant Professor in Computer Science at the University of Pittsburgh. She got her PhD in August 2014 at the University of Texas at Austin. Her work has been published in CVPR, ICCV, ECCV, BMVC, ACL and AAAI, and has been funded by NSF, Google, Amazon and Adobe. She served as Area Chair for CVPR 2018-2020.