Automatic Understanding of Image and Video Advertisements

Zaeem Hussain Mingda Zhang Xiaozhong Zhang Keren Ye
Christopher Thomas Zuha Agha Nathan Ong Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{zaeem, mzhang, xiaozhong, yekeren, chris, zua2, nro5, kovashka}@cs.pitt.edu

Abstract

There is more to images than their objective physical content: for example, advertisements are created to persuade a viewer to take a certain action. We propose the novel problem of automatic advertisement understanding. To enable research on this problem, we create two datasets: an image dataset of 64,832 image ads, and a video dataset of 3,477 ads. Our data contains rich annotations encompassing the topic and sentiment of the ads, questions and answers describing what actions the viewer is prompted to take and the reasoning that the ad presents to persuade the viewer ("What should I do according to this ad, and why should I do it?"), and symbolic references ads make (e.g. a dove symbolizes peace). We also analyze the most common persuasive strategies ads use, and the capabilities that computer vision systems should have to understand these strategies. We present baseline classification results for several prediction tasks, including automatically answering questions about the messages of the ads.

1. Introduction

Image advertisements are quite powerful, and web companies monetize this power. In 2014, one fifth of Google's revenue came from their AdSense product, which serves ads automatically to targeted users [1]. Further, ads are an integral part of our culture. For example, the two top-left ads in Fig. 1 have likely been seen by every American, and have been adapted and reused in countless ways. In terms of video ads, Volkswagen's 2011 commercial "The Force" had received 8 million views before it aired on TV [25].

Ads are persuasive because they convey a certain message that appeals to the viewer. Sometimes the message is simple, and can be inferred from body language, as in the "We can do it" ad in Fig. 1. Other ads use more complex messages, such as the inference that because the eggplant and pencil form the same object, the pencil gives a very real, natural eggplant color, as in the top-right ad in Fig. 1.

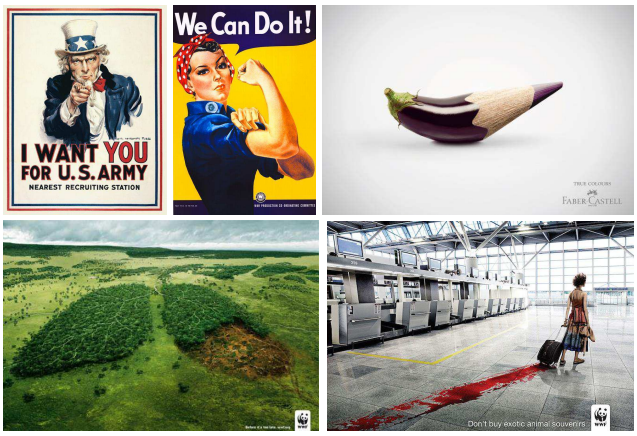

Figure 1: Two iconic American ads, and three that require robust visual reasoning to decode. Despite the potential applications of ad-understanding, this problem has not been tackled in computer vision before.

Decoding the message in the bottom-right ad involves even more steps, and reading the text ("Don't buy exotic animal souvenirs") might be helpful. The viewer has to infer that the woman went on vacation from the fact that she is carrying a suitcase, and then surmise that she is carrying dead animals from the blood trailing behind her suitcase. A human knows this because she associates blood with injury or death. In the case of the "forest lungs" image at the bottom-left, 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 to do. These are just a few examples of how ads use different types of *visual rhetoric* to convey their message, namely: common-sense reasoning, symbolism, and recognition of non-photorealistic objects. Understanding advertisements *automatically* requires decoding this rhetoric. This is a challenging problem that goes beyond listing objects and their locations [72, 21, 61], or even producing a sentence about the image [76, 14, 33], because ads are as much about

how objects are portrayed and *why* they are portrayed so, as about *what* objects are portrayed.

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 ad, and public service announcements, such as the anti-animal-souvenirs ad). Our ads cover a diverse range of subjects. We ask Amazon Mechanical Turk workers to tag each ad with its topic (e.g. what product it advertises or what the subject of the public service announcement is), what sentiment it attempts to inspire in the viewer (e.g. disturbance in the environment conservation ad), and what strategy it uses to convey its message (e.g. it requires understanding of physical processes). We also include crowdsourced answers to two questions: "What should the viewer do according to this ad?" and "Why should he/she do it?" Finally, we include any symbolism that ads use (e.g. the fact that a dove in an image might symbolically refer to the concept of "peace"). We also develop a dataset of over 3,000 *video* ads with similar annotations (except symbolism), and a few extra annotations (e.g. "Is the ad funny?" and "Is it exciting?") Our data collection and annotation procedures were informed by the literature in Media Studies, a discipline which studies the messages in the mass media, in which one of the authors has formal training. 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. We hope it will spur progress on the novel and important problem of decoding ads.

In addition to creating the first pair of datasets for understanding ad rhetoric, we propose several baselines that will help judge progress on this problem. First, we formulate decoding ads as a question-answering problem. If a computer vision system understood the rhetoric of an ad, it should be able to answer questions such as "According to this ad, why should I not bully children?" This is a very challenging task, and accuracy on it is low. Second, we formulate and provide baselines for other tasks such as topic and sentiment recognition. These tasks are more approachable and have higher baseline accuracy. Third, we show initial experiments on how symbolism can be used for question-answering.

The ability to automatically understand ads has many applications. For example, we can develop methods that predict how effective a certain ad will be. 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. Further, if we can decode the messages of ads, we can perform better ad-targeting according to user interests. Finally, 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.

2. Related work

In this work, we demonstrate that there is an aspect of visual data that has not been tackled before, namely analyzing the *visual rhetoric* of images. This problem has been studied in Media Studies [79, 71, 51, 55, 54, 5, 44, 12]. Further, marketing research [83] examines how viewers react to ads and whether an ad causes them to buy a product. While decoding ads has not been studied in computer vision, the problem is related to several areas of prior work.

Beyond objects. Work on semantic visual attributes *describes* images beyond *labeling* the objects in them, e.g. with adjective-like properties such as "furry", "smiling", or "metallic" [39, 15, 56, 38, 68, 35, 36, 17, 77, 2, 27]. The community has also made first attempts in tackling content which requires subjective judgement or abstract analysis. For example, [59] learn to detect *how well* a person is performing an athletic action. [41] use the machine's "imagination" to answer questions about images. [32, 73] study the style of artistic photographs, and [13, 40] study style in architecture and vehicles. While these works analyze potentially subjective content, none of them analyze *what the image is trying to tell us*. Ads constitute a new type of images, and understanding them requires new techniques.

Visual persuasion. Most related to our work is the visual persuasion work of [29] which analyzes whether images of politicians portray them in a positive or negative light. The authors use features that capture facial expressions, gestures, and image backgrounds to detect positive or negative portrayal. However, many ads do not show people, and even if they do, usually there is not an implication about the qualities of the person. Instead, ads use a number of other techniques, which we discuss in Sec. 4.

Sentiment. One of our tasks is predicting the sentiment an ad aims to evoke in the viewer. [57, 58, 6, 42, 30, 45] study the emotions shown or perceived in images, but for generic images, rather than ones purposefully created to convey an emotion. We compare to [6] and show the success of their method does not carry over to predicting emotion in ads. This again shows that ads represent a new domain of images whose decoding requires novel techniques.

Prior work on ads. We are not aware of any work in decoding the meaning of advertisements as we propose. [4, 10] predict click-through rates in ads using low-level vision features, whereas we predict what the ad is about and what message it carries. [47] predict how much human viewers will like an ad by capturing their facial expressions. [82, 48] 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. [64, 19] detect whether the current video shown on TV is a commercial or not, and [65] detect human trafficking advertisements. [85] extract the object being advertised from commercials (videos), by looking for recurring patterns (e.g. logos). Human facial reactions, ad

placement and recognition, and detecting logos, are quite distinct from our goal of decoding the messages of ads.

Visual question-answering. One of the tasks we propose for advertisements is *decoding their rhetoric*, i.e. figuring out what they are trying to say. We formulate this problem in the context of visual question-answering. The latter is a recent vision-and-language joint problem [3, 62, 46, 80, 67, 81] also related to image captioning [76, 33, 14, 37, 16].

3. Image dataset

The first dataset we develop and make available is a large annotated dataset of image advertisements, such as the ones shown in Fig. 2 (more examples are shown in the supplementary file). Our dataset includes both advertisements for products, and ads that campaign for/against something, e.g. *for* preserving the environment and *against* bullying. We call the former "product ads," and the latter "public service announcements," or "PSAs". We refer to the product or subject of the ad as its "topic". We describe the image collection and annotation process below.

3.1. Collecting ad images

We first assembled a list of keywords (shown in supp) related to advertisements, focusing on possible ad topics. We developed a hierarchy of keywords that describe topics at different levels of granularity. This hierarchy included both coarse topics, e.g. "fast food", "cosmetics", "electronics", etc., as well as fine topics, such as the brand names of products (e.g. "Sprite", "Maybeline", "Samsung"). Similarly, for PSAs we used keywords such as: "smoking", "animal abuse", "bullying", etc. 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 deduplication, we ended up with about 190,000 noisy images.

Finally, 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" (examples in supp). We marked as ads those images that at least 3/4 annotators labeled as an ad, obtaining 8,348 ads and 13,597 not-ads.

Second, we used these to train a ResNet [24] 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 evaluated on a held-out set from the human-annotated pool of 21,945 images. 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. Using the automatic classification step, we saved $1,300 in annotation costs. In total, we obtained **64,832** cleaned-up ads.

Type	Count	Example
Topic	204,340	Electronics
Sentiment	102,340	Cheerful
Action/Reason	202,090	I should bike because it's healthy
Symbol	64,131	Danger (+ bounding box)
Strategy	20,000	Contrast
Slogan	11,130	Save the planet... save you

Table 1: The annotations collected for our image dataset. The counts are before any majority-vote cleanup.

3.2. Collecting image ad annotations

We collected the annotations in Tab. 1, explained below. Note that we describe the strategies annotations in Sec. 4.

3.2.1 Topics and sentiments

The keyword query process used for image download does not guarantee that the images returned for 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 [47]. 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 [60]. We then semi-automatically clustered them 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 3 to 5 different workers. Inter-annotator agreement on topic labels was 85% (more details in supp). Examples are shown in Tab. 2. The distribution of topics and sentiments is illustrated in Fig. 3 (left); we see that sports ads and human rights ads inspire activity, while domestic abuse and human and animal rights ads inspire disturbance and empathy. Interestingly, we observe that domestic abuse ads inspire disturbance more frequently than animal rights ads do.

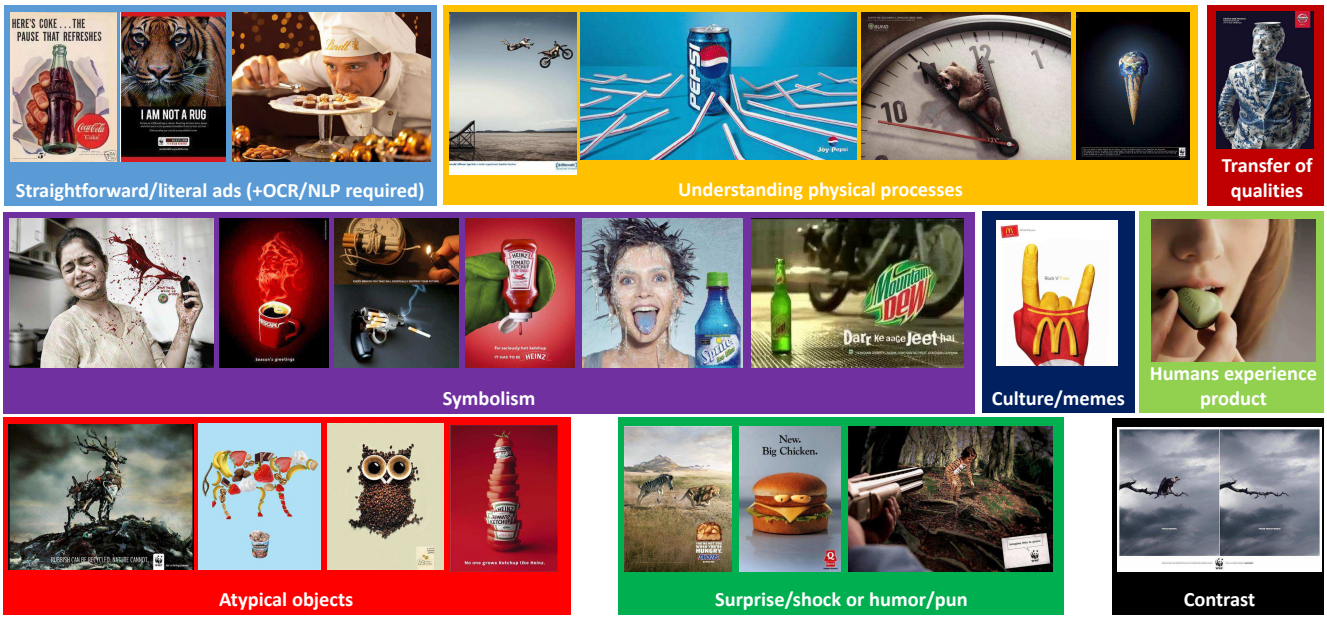

Figure 2: Examples of ads grouped by strategy or visual understanding required for decoding the ad.

Topic	Sentiment
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, shocked)
Cars, automobiles	Educated (enlightened, etc.)
Political candidates	Feminine (womanly, girlish)
Animal rights, animal abuse	Persuaded (impressed, etc.)
Smoking, alcohol abuse	Sad (depressed, etc.)

Table 2: A sample from our list of topics and sentiments. See supp for the full list of 38 topics and 30 sentiments.

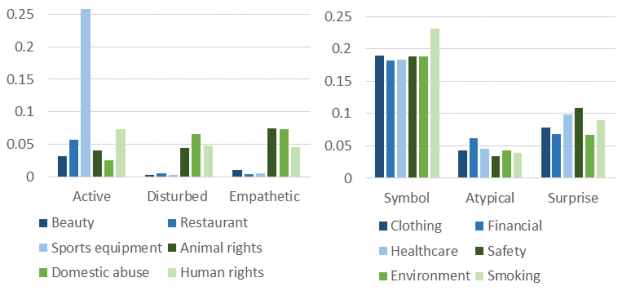

Figure 3: Statistics about topics and sentiments (left), and topics and strategies (right).

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

Table 3: Examples of collected question-answer pairs.

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

Table 4: Common words in responses to action and reason questions for selected topics, from the image dataset.

3.2.2 Questions and answers

We collected 202,090 questions and corresponding answers, with three question-answer pairs per image. Tab. 3 shows a few examples. We asked MTurk workers "What should you do, according to this ad, and why?" The answer then describes the message of the ad, e.g. "I should buy this dress because it will make me attractive." We re-

quired workers to provide answers in the form "I should [Action] because [Reason]." Since the question is always the same, we automatically reformatted the annotator's answer into a question-answer pair, as follows. The question became "Why should you [Action]?" and the answer became "Because [Reason]." For later tasks, we split this into *two* questions, i.e. we separately asked about the "What?" and the "Why?" However, Tab. 1 counts these as a single annotation. Examples of the most commonly used words in the questions and answers are shown in Tab. 4.

3.2.3 Symbols

In the second row of Fig. 2, 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 object or content that stands for some conceptual symbol is called "signifier", and the symbol is the "signified" [79].

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 straightforward), 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." More statistics are in supp.

3.2.4 Slogans

Additionally, for a small number of ads, we also asked MTurkers to write creative slogans that capture the message of the ad. While we did not use this data in our work, we obtained some intriguing answers, which we think can inspire interesting work on slogan generation.

3.3. 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 significantly, some of the more difficult tasks, such as identifying symbolism and question-answering, 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. In total, collecting annotations for both image and video ads cost **$13,030**.

For the tasks where MTurkers just had to select options, such as topics and sentiments, we relied on a majority vote to disregard low-quality work. For question-answering, we used heuristics, the number of short or repetitive responses and number of non-dictionary words in the answers, to shortlist suspicious responses. For symbolism, we manually reviewed a random subset of responses from each MTurker who did more than a prespecified number of tasks in a day.

4. How can we decode ads?

What capabilities should our computer vision systems have, in order to automatically understand the messages that ads convey, and their persuasive techniques? For example, would understanding the ad be straightforward if we had perfect object recognition? We develop a taxonomy that captures the key strategies that ads use. While ad strategy and type of visual understanding are not the same, they influence each other, so our analysis captures both.

Five of the authors each labeled 100 ads with the strategy the ad uses. We did so using a shared spreadsheet where we progressively added new strategies in free-form text, as we encountered them, or selected a previously listed strategy. After all 5x100 images were annotated, one author checked for consistency and iteratively merged similar strategies, resulting in a final list of nine strategies shown in Fig. 2:

- *Straightforward/literal* ads that only require object recognition and text recognition and understanding;
- Ads that imply some dynamic *physical process* is taking place, and this process is the reason why the product is valuable (e.g. the straws are striving towards the can) or why action must be taken (e.g. the arms of the clock are crushing the bear, so time is running out);
- Ads where *qualities* of one object (e.g. the fragility of a teacup) *transfer* to another (the person);
- Ads where an object *symbolizes* an external concept;
- Ads that make references to *cultural* knowledge;
- Ads that illustrate the qualities of a product with a person *experiencing* them;
- Ads that show *atypical* non-photorealistic objects;
- Ads that convey their message by *surprising, shocking* or entertaining the viewer through *humor*;
- Ads that demonstrate the qualities of products, or dangers of environmental processes, through *contrast*.

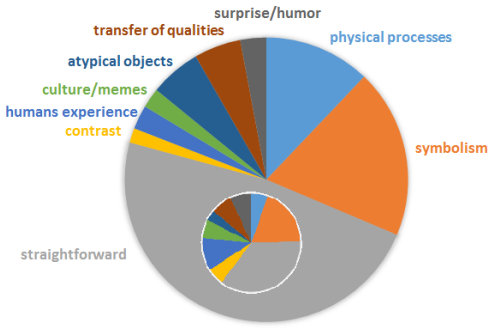

Figure 4: Strategies and visual understanding statistics. The main figure shows annotations from the authors; the inset shows annotations from MTurk workers. Best in color.

Each image could be labeled with multiple strategies. We computed what fraction of the total number of strategy instances across ads belong to each strategy. We illustrate the result in the main chart in Fig. 4. In order to compute statistics over more ads, we also asked MTurk workers to label the strategy for a set of 4000 ads. We obtained a similar chart (shown as the inset in Fig. 4) where the more rare strategies appeared slightly more commonly, likely because different viewers have a different internal "clustering" of strategies, and our MTurk annotators are not vision experts. In both the authors' pie chart and the crowdsourced one, straightforward ads and symbolic ads are most common.

Based on the statistics in Fig. 4, the straightforward strategy which can be decoded with perfect object recognition accounts for less than 50% of all strategy instances. Thus, as a community, we also must tackle a number of other challenges summarized below, to enable ad-decoding. Note that decoding ads involves a somewhat unique problem: the number of ads for each strategy is not very large, so applying standard deep learning techniques may be infeasible.

- We need to develop a method to decode symbolism in ads. We make an initial attempt at this task, in Sec. 6.2.
- We need to develop techniques to understand physical processes in ads (e.g. the straws *striving towards* the can, or the bear *being crushed*). There is initial work in understanding physical forces [53, 52, 86] and object transformations [26], but this work is still in its infancy and is not sufficient for full physical process understanding, as we need for many ads.
- We need robust algorithms that can recognize objects in highly non-photorealistic modalities. For example, vision algorithms we experimented with were unable to recognize the deer, cow, owl and bottle under "Atypical objects" in Fig. 2. This may be because here these objects appear with very distinct texture from that seen in training images. There is work in learning domain-invariant representations [22, 75, 7, 8, 18, 20, 31, 43, 11], but a challenge in the case of ads is that data from each "domain" (the particular way e.g. the deer is portrayed) may be limited to single examples.
- We need techniques to understand what is surprising or funny in an ad. There is initial work on abnormality [63, 78] restricted to modeling co-occurrences of objects and attributes, and on humor [9] in cartoons, but surprise/humor detection remains largely unsolved.

Finally, we also analyze correlations between ad topics and ad strategies, in Fig. 3 (right). We see that symbols are used in a variety of ads, but most commonly in smoking ads. Financial ads use atypical portrayals of objects most frequently, and healthcare and safety ads use surprise.

5. Video dataset

Video advertisements are sometimes even more entertaining and popular than image ads. For example, an Old Spice commercial[1] has over 54 million views. However, commercials are expensive to make, and might cost several million USD to air [23]. Thus, there are fewer commercials available on the web, hence our video dataset is smaller.

5.1. Collecting ad videos

We obtained a list of 949 videos from an Internet service provider. However, we wanted to increase the size of the dataset, so we additionally crawled YouTube for videos, using the keywords we used to crawl Google for images. We picked videos that have been played at least 200,000 times and have 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 presumably less sensitive to slight variations. If both the start and end frames of two videos matched according to a hashing algorithm [84], 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.

5.2. Collecting video ad annotations

We collected the types of annotations shown in Tab. 5. We showed workers examples for how to annotate, on six videos. The topic and sentiment multiple-choice options overlap with those used for images. We also obtained answers to the questions "What should you do according to this video?" and "Why should you do this, according to the

[1] https://www.youtube.com/watch?v=owGykVbfgUE

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

Table 5: The annotations collected for our video ad dataset.

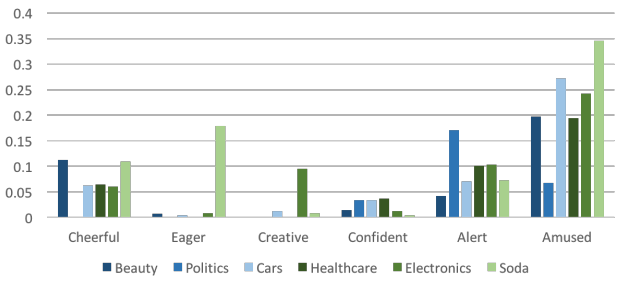

Figure 5: Statistics of the video dataset.

What should you do?			*Why* should you do it?		
Educat.	Travel	Charity	Educat.	Travel	Charity
univ.	visit	support	help	fun	help
enroll	go	donate	get	family	cancer
college	vacation	charity	degree	travel	need
online	travel	money	univ.	place	children
attend	use	foundat.	offer	vacation	people

Table 6: Common responses to action and reason questions.

video?" We show statistics in Fig. 5 and Tab. 6, and more in supp. For example, we see cheerfulness is most common for beauty and soda ads, eagerness for soda ads, creativeness for electronics ads, and alertness for political ads.

Our video dataset has two additional annotations, namely whether a video ad is "funny" or "exciting". Since a video has more space/time to convey its message, we thought humor and excitement are more typical for video ads. In contrast, we found symbolism less typical.

6. Experiments

Thus far, we described our collected data, and analysis on what is required in order to be able to decode ads. We now describe our evaluation of baselines for several prediction tasks on ads. For most prediction tasks we treat the possible labels as mutually exclusive, and report accuracy. For symbolism detection, we predict multiple labels per image, and report the overall F-score.

6.1. Question-answering for image ads

We first evaluate how well an existing question-answering method performs on our questions about ads. In order to answer questions about ads, computer vision systems need to understand the implicit visual rhetoric and persuasion techniques of ads. We show existing methods do *not* have this capability, via their low performance.

Because our amount of data is limited (64,832 images with 3 or 5 question-answer pairs per image), we opted for a simple question-answering approach [3], trained on our ads data. We use a two-layer LSTM to encode questions in 2048D, and the last hidden layer of VGGNet [69] to encode images in 4096D. We then project these to 1024D each, concatenate them, and add a 1000D softmax layer, which is used to generate a single-word answer. For each image, we have three reformatted questions about the persuasive strategy that the ad uses, of the type "Why should you [Action]?" where Action can be e.g. "buy a dress". We select one of the three questions for training/testing, namely the one whose words have the highest average TFIDF score. The TFIDF scores are calculated based on all questions and answers. We pair that question with modified versions of the three answers that annotators provided, of the form "Because [Reason]," where Reason can be e.g. "it will make me pretty." Since our data originally contains *sentence* (not single-word) answers, we trim each of the three answers to the single most "contentful" word, i.e. the word that has the highest TFIDF score.

We consider the predicted answer to be correct if it matches any of the three human answers. Frequently our human annotators provide different answers, due to synonymy or because they interpreted the ad differently. This is in contrast to more objective QA tasks [3] where answers are more likely to converge. Similarly, the QA method might predict a word that is related to the annotator answers but not an exact match. Thus, our QA task is quite challenging. Using the approach described, we obtain **11.48%** accuracy. This is lower than the accuracy of the "why" questions from the original VQA (using a different test setup).

To simplify the QA prediction task, we also conducted an experiment where we clustered the original full-sentence answers into 30 "prototype" answers, and trained a network to predict one of the 30 cluster IDs. The intuition is that while there are many different words annotators use to answer our "Why" questions, there are common patterns in the reasons provided. The baseline network (using a 128D question encoding and 512D image encoding) achieved **48.45%** accuracy on this task. We next attempt to improve these numbers via symbolism decoding.

6.2. Symbolism prediction

We use an attention model [66, 28] to detect symbols; we found that to work slightly better than direct classifica-

tion. Details of the model are provided in supp. Multiple symbols might be associated with different regions in the image. The network achieves F-score of **15.79%** in distinguishing between the 221 symbols.

Note that learning a symbol detector is very challenging due to the variable data within each symbol. For example, the symbolic concept of "manliness" can be illustrated with an attractive man posing with a muscle car, a man surrounded by women, etc. We show some examples in supp.

To improve symbol predictions, we also experimented with grouping symbols into clusters based on synonymy and co-occurrence, obtaining 53 symbols (see supp for details). A model trained to distinguish between these 53 symbols achieved **26.84%** F-score.

We also did a preliminary experiment using symbolism for question-answering. For the 1000-way single-word prediction task, we used the class probability of each symbol as an extra feature to our QA network, and obtained slightly improved accuracy of **11.96%** (compared to 11.48% for the baseline). On the 30-way QA task, a method which replaced the baseline's image features with 3x512D ones obtained from a network fine-tuned to predict (1) symbols, (2) topics, and (3) sentiments, achieved **50%** accuracy (compared to 48.45%). Devising a better way to predict and use symbolism for question-answering is our future work.

6.3. Question-answering for video ads

We used the same process as above and the video features from Sec. 6.5. We achieved QA accuracy of **8.83%**.

6.4. Topic and sentiment on image ads

We chose the most frequent topic/sentiment as the ground-truth label. We trained 152-layer ResNets [24] to discriminate between our 38 topics and 30 sentiments. The network trained on topics achieved **60.34%** accuracy on a held-out set. The sentiment network achieved **27.92%** accuracy. Thus, predicting the topic of an ad is much more feasible with existing techniques, compared to predicting the message as in the QA experiments above.

For sentiments, we also trained a classifier on the data from the Visual Sentiment Ontology [6], to establish how sentiment recognition on our ads differs from recognizing sentiments on general images. We map [6]'s Adjective Noun Phrases (ANPs) to the sentiments in our data, by retrieving the top 10 ANPs closest to each of our sentiment words, measuring cosine similarity in word2vec space [49, 50]. We use images associated with all ANPs mapped to one of our sentiments, resulting in 21,523 training images (similar to our number of sentiment-annotated images). This achieves **6.64%** accuracy, lower than the sentiment accuracy on our ad data, indicating that sentiment on ads looks different than sentiment on other images.

6.5. Topic and sentiment on video ads

We believe the actions in the video ads may have significant impact on understanding the ads. Thus, we used the C3D network [74, 34] originally used for action recognition as a feature extractor. It is pre-trained on Sports-1M [34] and fine-tuned on UCF101 [70]. We converted videos into frames, and took consecutive 16 frames as a clip. We extracted fc6 and fc7 features for each clip and simply averaged the features for all clips within the same video.

We trained separate multi-class SVMs to distinguish between our 38 topics and 30 sentiments. We found fc7 shows better performance. With the optimal parameters from a validation set, we achieved **35.1%** accuracy for predicting video topics, and **32.8%** accuracy for sentiments. We had limited success with directly training a network for this task.

6.6. Funny/exciting for video ads

We used a similar strategy to predict "funny" and "exciting" binary labels on videos. We excluded videos that are ambiguous (i.e. obtained split positive/negative votes). We trained binary SVMs on the fc7 features, and obtained **78.6%** accuracy for predicting humor, and **78.2%** for predicting excitement. Note that a majority class baseline achieves only 58% and 60.8%, respectively. Thus, predicting humor and excitement is surprisingly feasible.

7. Conclusion

We have proposed a large annotated image advertisement dataset, as well as a companion annotated video ads dataset. We showed analysis describing what capabilities we need to build for vision systems so they can understand ads, and showed an initial solution for decoding symbolism in ads. We also showed baselines on several tasks, including question-answering capturing the subtle messages of ads.

We will pursue several opportunities for future work. We will further develop our symbolism detection framework, including additional weakly labeled web data for each symbol. We will also make use of knowledge bases for decoding ads. We will model video advertisements with an LSTM network and better features, and include audio processing in our analysis. We will use the topic, sentiment, humor and excitement predictions to improve the accuracy of question-answering. Finally, we will also pursue recognizing atypical objects and modeling physical processes.

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant Number 1566270. This research was also supported by a Google Faculty Research Award and an NVIDIA hardware grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] AdSense. https://en.wikipedia.org/wiki/AdSense.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] 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.

[4] 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.

[5] J. Bignell. *Media semiotics: An introduction*. Manchester University Press, 2002.

[6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the ACM International Conference on Multimedia*, 2013.

[7] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[8] L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] A. Chandrasekaran, A. Kalyan, S. Antol, M. Bansal, D. Batra, C. L. Zitnick, and D. Parikh. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] 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.

[11] C. M. Christoudias, R. Urtasun, M. Salzmann, and T. Darrell. Learning to recognize objects from unseen modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[12] M. Danesi. *Messages, signs, and meanings: A basic textbook in semiotics and communication*. Canadian Scholars Press, 2004.

[13] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes Paris look like Paris? *ACM Transactions on Graphics*, 31(4), 2012.

[14] 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.

[15] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[16] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[17] D. F. Fouhey, A. Gupta, and A. Zisserman. 3D shape attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015.

[19] J. M. Gauch and A. Shivadas. Finding and identifying unknown commercials using repeated video sequence detection. *Computer Vision and Image Understanding*, 103(1):80–88, 2006.

[20] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[21] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

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

[24] 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.

[25] M. Inbar. Little Darth Vader' reveals face behind the Force. http://www.nbcnews.com/id/41455377/41458412.

[26] P. Isola, J. J. Lim, and E. H. Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[28] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] 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.

[30] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *Proceedings of the ACM International Conference on Multimedia*, 2014.

[31] M. Kan, S. Shan, and X. Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[32] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.

[33] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] 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.

[35] A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision*, 114(1):56–73, 2015.

[36] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210, 2015.

[37] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[38] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[39] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[40] Y. J. Lee, A. Efros, M. Hebert, et al. Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[41] X. Lin and D. Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[42] H. Liu, B. Jou, T. Chen, M. Topkara, N. Pappas, M. Redi, and S.-F. Chang. Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology. In *ACM International Conference on Multimedia Retrieval*, 2016.

[43] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[44] S. Maasik and J. Solomon. *Signs of life in the USA: Readings on popular culture for writers*. Macmillan, 2011.

[45] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the ACM International Conference on Multimedia*, 2010.

[46] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[47] 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.

[48] T. Mei, L. Li, X.-S. Hua, and S. Li. Imagesense: towards contextual image advertising. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 8(1):6, 2012.

[49] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[50] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 2013.

[51] N. Mirzoeff. *The visual culture reader*. Psychology Press, 2002.

[52] R. Mottaghi, H. Bagherinezhad, M. Rastegari, and A. Farhadi. Newtonian image understanding: Unfolding the dynamics of objects in static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[53] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi. "What happens if..." Learning to predict the effect of forces in images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[54] W. M. O'Barr. *Culture and the ad: Exploring otherness in the world of advertising*. Westview Pr, 1994.

[55] L. C. Olson, C. A. Finnegan, and D. S. Hope. *Visual rhetoric: A reader in communication and American culture*. Sage, 2008.

[56] D. Parikh and K. Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[57] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[58] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? Predicting the Emotion Stimuli Map. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

[59] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014.

[60] K. Poels and S. Dewitte. How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, 46(1):18–37, 2006.

[61] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[62] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[63] 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.

[64] J. M. Sánchez, X. Binefa, and J. Vitrià. Shot partitioning based recognition of tv commercials. *Multimedia Tools and Applications*, 18(3):233–247, 2002.

[65] 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.

[66] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[67] 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.

[68] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[69] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[70] 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.

[71] M. Sturken, L. Cartwright, and M. Sturken. *Practices of looking: An introduction to visual culture*. Oxford University Press Oxford, 2001.

[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[73] C. Thomas and A. Kovashka. Seeing behind the camera: Identifying the authorship of a photograph. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[74] 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.

[75] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[76] 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.

[77] J. Wang, Y. Cheng, and R. Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[78] 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.

[79] J. Williamson. *Decoding advertisements*. 1978.

[80] Q. Wu, P. Wang, C. Shen, A. v. d. Hengel, and A. Dick. 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.

[81] 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.

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

[83] C. E. Young. *The advertising research handbook*. Ideas in Flight, 2005.

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

[85] G. Zhao, J. Yuan, J. Xu, and Y. Wu. Discovering the thematic object in commercial videos. *IEEE Transactions on Multimedia*, (3):56–65, 2011.

[86] B. Zheng, Y. Zhao, C. Y. Joey, K. Ikeuchi, and S.-C. Zhu. Detecting potential falling objects by inferring human action and natural disturbance. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.