ADVISE: Symbolism and External Knowledge for Decoding Advertisements

Keren Ye[0000−0002−7349−7762] and Adriana Kovashka[0000−0003−1901−9660]

University of Pittsburgh, Pittsburgh PA 15260, USA
{yekeren,kovashka}@cs.pitt.edu

Abstract. In order to convey the most content in their limited space, advertisements embed references to outside knowledge via symbolism. For example, a motorcycle stands for adventure (a positive property the ad wants associated with the product being sold), and a gun stands for danger (a negative property to dissuade viewers from undesirable behaviors). We show how to use symbolic references to better understand the meaning of an ad. We further show how anchoring ad understanding in general-purpose object recognition and image captioning improves results. We formulate the ad understanding task as matching the ad image to human-generated statements that describe the action that the ad prompts, and the rationale it provides for taking this action. Our proposed method outperforms the state of the art on this task, and on an alternative formulation of question-answering on ads. We show additional applications of our learned representations for matching ads to slogans, and clustering ads according to their topic, without extra training.

Keywords: advertisements · symbolism · question answering · external knowledge · vision and language · representation learning

1 Introduction

Advertisements are a powerful tool for affecting human behavior. Product ads convince us to make large purchases, e.g. for cars and home appliances, or small but recurrent purchases, e.g. for laundry detergent. Public service announcements (PSAs) encourage socially beneficial behaviors, e.g. combating domestic violence or driving safely. To stand out from the rest, ads have to be both eye-catching and memorable [71], while also conveying the information that the ad designer wants to impart. All this must be done in a limited space (one image) and time (however many seconds the viewer spends looking at the ad).

How can ads get the most "bang for their buck"? One technique is to make references to knowledge viewers already have, e.g. cultural knowledge, associations, and *symbolic mappings* humans have learned [54, 35, 57, 34]. These symbolic references might come from literature (e.g. a snake symbolizes evil or danger), movies (a motorcycle symbolizes adventure or coolness), common sense (a flexed arm symbolizes strength), or pop culture (Usain Bolt symbolizes speed).

In this paper, we describe how to use symbolic mappings to predict the messages of advertisements. On one hand, we model how components of the ad

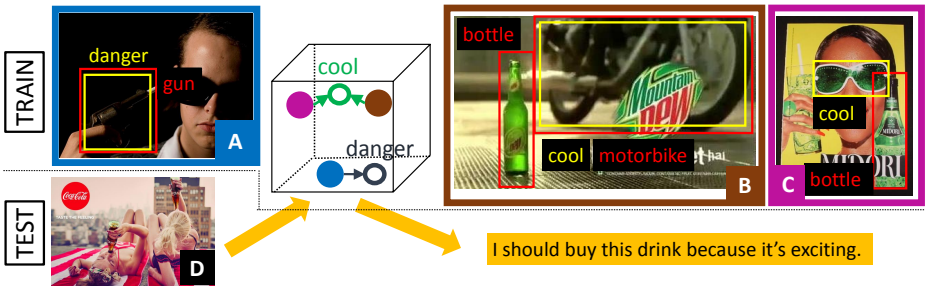


Fig. 1. Our key idea: Use symbolic associations shown in yellow (a gun symbolizes danger; a motorcycle symbolizes coolness) and recognized objects shown in red, to learn an image-text space where each ad maps to the correct statement that describes the message of the ad. The symbol "cool" brings images B and C closer together in the learned space, and further from image A and its associated symbol "danger." At test time (shown in orange), we use the learned image-text space to retrieve a matching statement for test image D. At test time, the symbol labels are *not* provided.

image serve as visual anchors to concepts outside the image, using annotations in the Ads Dataset of [22]. On the other hand, we use knowledge sources external to the main task, such as object detection models, to better relate ad images to their corresponding messages. Both of these are forms of using outside knowledge, and they both boil down to learning links between objects and symbolic concepts. We use each type of knowledge in two ways, as a constraint or as an additive component for the learned image representation.

We focus on the following multiple-choice task, implemented via ranking: Given an image and several statements, the system must identify the correct statement to pair with the ad. For example, for test image D in Fig. 1, the system might predict the right statement is "Buy this drink because it's exciting." Our method learns a joint image-text embedding that associates ads with their corresponding messages. The method has three components: (1) an image embedding which takes into account individual regions in the image, (2) constraints on the learned space from symbol labels and object predictions, and (3) an additive expansion of the image representation using a symbol distribution. These three components are shown in Fig. 1, and all of them rely on external knowledge in the form of symbols and object predictions. Note that we can recognize the symbolic association to danger in Fig. 1 via two channels: either a direct classifier that learns to link certain visuals to the "danger" concept, or learning associations between actual *objects* in the image which can be recognized by object detection methods (e.g. "gun"), and symbolic concepts. We call our method ADVISE: **AD**s **VI**sual **S**emantic **E**mbedding.

We primarily focus on public service announcements, rather than product (commercial) ads. PSAs tend to be more conceptual and challenging, often involving multiple steps of reasoning. Quantitatively, 59% of the product ads in the dataset of [22] are straightforward, i.e. would be nearly solved with tradi-

tional recognition advancements. In contrast, only 33% of PSAs use straightforward strategies, while the remaining 67% use challenging non-literal rhetoric. Our method outperforms several recent baselines, including prior visual-semantic embeddings [11, 10] and methods for understanding ads [22].

In addition to showing how to use external knowledge to solve ad-understanding, we demonstrate how recent advances in object recognition help with this task. While [22] evaluates basic techniques, it does not employ recent advances like region proposals [16, 50, 38, 14] or attention [7, 70, 67, 56, 69, 49, 45, 39, 12, 75, 47].

To summarize, our contributions are as follows:

- We show how to effectively use symbolism to better understand ads.
- We show how to make use of noisy caption predictions to bridge the gap between the abstract task of predicting the message of an ad, and more accessible information such as the objects present in the image. Detected objects are mapped to symbols via a domain-specific knowledge base.
- We improve the state of the art in understanding ads by 21%.
- We show for "abstract" PSAs, conceptual knowledge helps more, while for product ads, general-purpose object recognition techniques are more helpful.

The remainder of the paper is organized as follows. We overview related work in Sec. 2. In Sec. 3.1, we describe our ranking task, and in Sec. 3.2, we describe standard triplet embedding on ads. In Sec. 3.3, we discuss the representation of an image as a combination of region representations, weighed by their importance via an attention model. In Sec. 3.4, we describe how we use external knowledge to constrain the learned space. In Sec. 3.5, we develop an optional additive refinement of the image representation. In Sec. 4, we compare our method to the state of the art, and conduct ablation studies. We conclude in Sec. 5.

2 Related Work

Advertisements and multimedia. The most related work to ours is [22] which proposes the problem of decoding ads, formulated as answering the question "*Why* should I [action]?" where [action] is what the ad suggests the viewer should do, e.g. buy a car or help prevent domestic violence. The dataset contains 64,832 image ads. Annotations include the topic (product or subject) of the ad, sentiments and actions the ad prompts, rationales provided for why the action should be done, symbolic mappings (referred to as signifier-signified, e.g. motorcycle-adventure), etc. Considering the media domain more broadly, [26] analyze in what light a photograph portrays a politician, and [27] examine how the facial features of a candidate determine the outcome of an election. This work only applies to images of people. Also related is work in parsing infographics, charts and comics [4, 29, 23]. In contrast to these, our interest is analyzing the *implicit* arguments ads were created to make.

Vision, language and image-text embeddings. Recently there is great interest in joint vision-language tasks, e.g. captioning [63, 28, 9, 25, 2, 70, 62, 61, 73, 68, 13,

47, 55, 8, 32], visual question answering [3, 72, 41, 69, 56, 66, 59, 76, 77, 19, 64, 24, 60], and cross-domain retrieval [6, 5, 74, 36]. These often rely on learned image-text embeddings. [11, 30] use triplet loss where an image and its corresponding human-provided caption should be closer in the space than pairs that do not match. [10] propose a bi-directional network to maximize correlation between matching images and text, akin to CCA [18]. None of these consider images with implicit persuasive intent, as we do. We compare against [11, 10] in Sec. 4.

External knowledge for vision-language tasks. [66, 64, 24, 77, 60] examine the use of knowledge bases and perform explicit reasoning for answering visual questions. [62] use external sources to diversify their image captioning model. [43] learn to compose object classifiers by relating semantic and visual similarity. [42, 15] use knowledge graphs or hierarchies to aid in object recognition. These works all use mappings that are objectively/scientifically grounded, i.e. lion is a type of cat. In contrast, we use cultural associations that arose in the media/literature and are internalized by humans, e.g. motorcycles are associated with adventure.

Region proposals and attention. Region proposals [16, 50, 38, 14] guide an object detector to regions likely to contain objects. Attention [7, 70, 67, 56, 69, 49, 45, 39, 12, 75, 47] 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.

3 Approach

We learn an embedding space where we can evaluate the similarity between ad images and ad messages. We use symbols and external knowledge in three ways: by representing the image as a weighted average of its regions that are likely to make symbolic references (Sec. 3.3), by enforcing that images with the same symbol labels or detected objects are close (Sec. 3.4), and by enhancing the image representation via an attention-masked symbol distribution (Sec. 3.5). In Sec. 4 we demonstrate the utility of each component.

3.1 Task and dataset

In [22], the authors tackled answering the question "Q: Why should I [action]?" with "A: [one-word reason]." An example question-answer pair is "Q: Why should I speak up about domestic violence? A: bad." In other words, question-answering is formulated as a classification task. The ground-truth one-word answers in [22]'s evaluation are picked from human-provided full-sentence answers, also available in the dataset. 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 [22] cannot convey both the meaning of "protect" and "feet" while the full-sentence answer "Because it will protect my feet" does capture both. On the other hand, 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 we show that we outperform prior art on the original question-answering task of [22], we focus on an alternative formulation.

We ask the system to pick which *action-reason statement* is most appropriate for the image. 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.*" For each image, we use three related statements (i.e. statements provided by humans for this image) and randomly sample 47 unrelated statements (written for *other* images). The system must rank these 50 statements based on their similarity to the image.

This ranking task is akin to multiple-choice question-answering, which was also used in prior VQA works [3, 59], but unlike these, we do not take the question as input. Similarly, in image captioning, [28, 11] look for the most suitable image description from a much larger candidates pool.

3.2 Basic image-text triplet embedding

We first directly learn an embedding that optimizes for the ranking task. We require that the distance between an image and its corresponding statement should be smaller than the distance between that image and any other statement, or between other images and that statement. In other words, we minimize:

$$L(\boldsymbol{v}, \boldsymbol{t}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \Big[\underbrace{\sum_{j \in N_{vt}(i)} \left[\|\boldsymbol{v}_i - \boldsymbol{t}_i\|_2^2 - \|\boldsymbol{v}_i - \boldsymbol{t}_j\|_2^2 + \beta \right]_+}_{\text{image as anchor, rank statements}} + \underbrace{\sum_{j \in N_{tv}(i)} \left[\|\boldsymbol{t}_i - \boldsymbol{v}_i\|_2^2 - \|\boldsymbol{t}_i - \boldsymbol{v}_j\|_2^2 + \beta \right]_+}_{\text{statement as anchor, rank images}} \Big] \tag{1}$$

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

$$N_{vt}(i) = \underset{\substack{A \subseteq \{1,...,K\}, \\ |A| = k'}}{\arg\min} \sum_{\substack{j \in A, \\ i \neq j}} \|\boldsymbol{v}_i - \boldsymbol{t}_j\|_2^2, \qquad N_{tv}(i) = \underset{\substack{A \subseteq \{1,...,K\}, \\ |A| = k'}}{\arg\min} \sum_{\substack{j \in A, \\ i \neq j}} \|\boldsymbol{t}_i - \boldsymbol{v}_j\|_2^2 \tag{2}$$

Image embedding. We extract the image's Inception-v4 CNN feature (1536-D) using [58], then use a fully-connected layer with parameter $\boldsymbol{w} \in \mathbb{R}^{200\times1536}$ to project it to the 200-D joint embedding space:

$$\boldsymbol{v} = \boldsymbol{w} \cdot CNN(\boldsymbol{x}) \tag{3}$$

Text embedding. We use mean-pooling to aggregate word embedding vectors into 200-D text embedding $\boldsymbol{t}$ and use GloVe [48] to initialize the embedding matrix. There are two reasons for us to choose mean-pooling: (1) comparable performance to the LSTM[1], and (2) better interpretability. By using mean-pooling, image and words are projected to the same feature space, allowing us to assign word-level semantics to an image, or even to image regions. In contrast, LSTMs encode meaning of nearby words which is undesirable for interpretability.

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" 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 [53, 17, 11, 65]. Our experiment (in supp) shows this trade-off is better than either extreme.

3.3 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. Thus, we represent an image as a collection of its constituent regions, using an attention module to aggregate all the representations from different regions.

Importantly, the chosen regions should be those likely to serve as visual anchors for symbolic references (such as the motorcycle or shades in Fig.1, rather than the bottles). Thus we consider all the 13,938 images, which are annotated as containing symbols, each with up to five bounding box annotations. 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. More specifically, we use the SSD object detection model [38] implemented by [20], pre-train it on the COCO [37] dataset, and fine-tune it with the symbol bounding box annotations [22]. We show in Sec. 4.3 that this fine-tuning is crucial, i.e. general-purpose regions such as COCO boxes produce inferior results.

We use bottom-up attention [1, 60, 31] to aggregate the information from symbolic regions (see Fig. 2). More specifically, we use the Inception-v4 model [58] to extract the 1536-D CNN features for all symbol proposals. Then, for each CNN feature $\boldsymbol{x}_i, i \in \{1, \ldots, M\}$ (we set $M = 10$, i.e., 10 proposals per image), a fully-connected layer is applied to project it to: 1) a 200-D embedding vector $\boldsymbol{v}_i$

[1] Non-weighted/weighted mean-pooling of word embeddings achieved 2.45/2.47 rank. The last hidden layer of an LSTM achieved 2.74 rank, while non-weighted/weighted averaging of the hidden layers achieved 2.43/2.46, respectively. Lower is better.

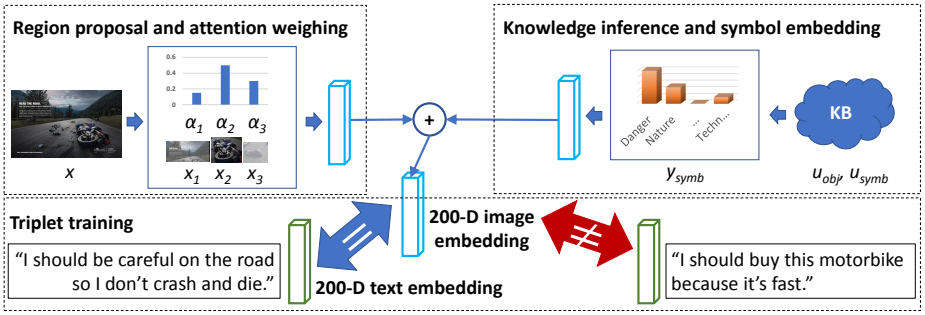


Fig. 2. Our image embedding model with knowledge branch. In the main branch (top left), multiple image symbolic anchors are proposed. Attention weighting is applied, and the image is represented as a weighted combination of the regions. The knowledge branch (top right) predicts the existence of symbols, maps these to 200-D, and adds them to the image embedding. We then perform triplet training to learn such an embedding space that keeps images close to their matching action-reason statements.

(Eq. 4, $\boldsymbol{w} \in \mathbb{R}^{200\times1536}$), and 2) a confidence score a_i saying how much the region should contribute to the final representation (Eq. 5, $\boldsymbol{w}_a \in \mathbb{R}^{1\times1536}$). The final image representation $\boldsymbol{z}$ is a weighted sum of these region-based vectors (Eq. 6).

$$\boldsymbol{v}_i = \boldsymbol{w} \cdot CNN(\boldsymbol{x}_i) \tag{4}$$

$$a_i = \boldsymbol{w}_a \cdot CNN(\boldsymbol{x}_i), \quad \boldsymbol{\alpha} = softmax(\boldsymbol{a}) \tag{5}$$

$$\boldsymbol{z} = \sum_{i=1}^{M} \alpha_i \boldsymbol{v}_i \tag{6}$$

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

3.4 Constraints via symbols and captions

We next exploit the symbol labels which are part of [22]. Symbols are abstract words such as "freedom" and "happiness" that provide additional information humans sense from the ads. We add additional constraints to the loss terms such that two images/statements that were annotated with the same symbol are closer in the learned space than images/statements annotated with different symbols. In the *extra* loss term (Eq. 7), $\boldsymbol{s}$ is the 200-D embedding of a symbol word; $\boldsymbol{z}$ is the 200-D region-based image representation defined in Eq. 6; and $N_{sz}(i)$ and $N_{st}(i)$ are the negative image/statement sets of the i-th symbol in

the batch, defined similar to Eq. 2.

$$L_{sym}(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{t}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \Big[\underbrace{\sum_{j \in N_{sz}(i)} \left[\|\boldsymbol{s}_i - \boldsymbol{z}_i\|_2^2 - \|\boldsymbol{s}_i - \boldsymbol{z}_j\|_2^2 + \beta \right]_+}_{\text{symbol as anchor, rank images}} + \underbrace{\sum_{j \in N_{st}(i)} \left[\|\boldsymbol{s}_i - \boldsymbol{t}_i\|_2^2 - \|\boldsymbol{s}_i - \boldsymbol{t}_j\|_2^2 + \beta \right]_+}_{\text{symbol as anchor, rank statements}} \Big] \tag{7}$$

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. However, since the Ads Dataset contains insufficient data to properly model object categories, we use DenseCap [25] to bridge the objects defined in Visual Genome [33] 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". We create additional constraints: If two images/statements have similar DenseCap predicted captions, they should be closer than images/statements with different captions. The *extra* loss term is defined similar to Eq. 7 using $\boldsymbol{c}$ for the caption representations.

In our setting, word embedding weights are not shared among the three vocabularies (ads statement, symbols, and DenseCap predictions). Our consideration is that the meaning of the same surface words may vary in these domains thus they need to have different embeddings. We weigh the symbol-based and object-based constraints by 0.1 since they in isolation do not tell the full story of the ad. We found that it is not sufficient to use *any* type of label as constraint in the domain of interest (see supp): using symbols as constraints gives greater benefit than the topic (product) labels in [22]'s dataset, and this point is not discussed in the general proxy learning literature [44].

3.5 Additive external knowledge

In this section, we describe how to make use of external knowledge that is adaptively added, to compensate for inadequacies of the image embedding. This external knowledge can take the form of a mapping between physical objects and implicit concepts, or a classifier mapping pixels to concepts. Given a challenging ad, a human might look for visual cues and check if they remind him/her of concepts (e.g. "danger", "beauty", "nature") seen in other ads. Our model interprets ads in the same way: based on an external knowledge base, it *infers* the abstract symbols. In contrast to Sec. 3.4 which uses the *annotated* symbols at training time, here we use a *predicted* symbol distribution at both training and test time as a secondary image representation. Fig. 2 (top right) shows the general idea of the external knowledge branch. Note our model only uses external knowledge

to compensate its own lack of knowledge (since we train the knowledge branch after the convergence of the visual semantic embedding branch), and it assigns small weights for uninformative knowledge.

We propose two ways to additively expand the image representation with external knowledge, and describe *two ways of setting* $\boldsymbol{y}_{symb}$ in Eq. 8. Both ways are a form of knowledge base (KB) mapping physical evidence to concepts.

KB Symbols. The first way is to directly train classifiers to link certain visuals to symbolic concepts. We learn a multilabel classifier $\boldsymbol{u}_{symb}$ to obtain a symbol distribution $\boldsymbol{y}_{symb} = \mathit{sigmoid}(\boldsymbol{u}_{symb} \cdot \boldsymbol{x})$. We learn a weight α_j^{symb} for each of $j \in \{1, \ldots, C = 53\}$ symbols from the Ads Dataset, denoting whether a particular symbol is helpful for the statement matching task.

KB Objects. The second method is to learn associations between surface words for detected objects and abstract concepts. For example, what type of ad might I see a "car" in? What about a "rock" or "animal"? We first construct a knowledge base associating object words to symbol words. We compute the similarity in the learned image-text embedding space between symbol words and DenseCap words, then create a mapping rule ("[object] implies [symbol]") for each symbol and its five most similar DenseCap words. This results in a $53 \times V$ matrix $\boldsymbol{u}_{obj}$, where V is the size of DenseCap's vocabulary. Each row contains five entries of 1 denoting the mapping rule, and $V - 5$ entries of 0. Examples of learned mappings are shown in Table 3. For a given image, we use [25] to predict the three most probable words in the DenseCap vocabulary, and put the results in a multi-hot $\boldsymbol{y}_{obj} \in \mathbb{R}^{V \times 1}$ vector. We then matrix-multiply to accumulate evidence for the presence of all symbols using the detected objects: $\boldsymbol{y}_{symb} = \boldsymbol{u}_{obj} \cdot \boldsymbol{y}_{obj}$. We associate a weight α_{jl}^{symb} with each rule in the KB.

For both methods, we first use the attention weights α^{symb} as a mask, then project the 53-D symbol distribution $\boldsymbol{y}_{symb}$ into 200-D, and add it to the image embedding. This additive branch is most helpful when the information it contains is not already contained in the main image embedding branch. We found this happens when the discovered symbols are rare.

3.6 ADVISE: our final model

Our final **AD**s **VI**sual **S**emantic **E**mbedding loss combines the losses from Sec. 3.2, 3.3, 3.4, and 3.5:

$$\begin{aligned} L_{final}(\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{s}, \boldsymbol{c}; \boldsymbol{\theta}) = & L(\boldsymbol{z} + \boldsymbol{y}_{symb}, \boldsymbol{t}; \boldsymbol{\theta}) + \\ & 0.1\ L_{sym}(\boldsymbol{s}, \boldsymbol{z} + \boldsymbol{y}_{symb}, \boldsymbol{t}; \boldsymbol{\theta}) + 0.1\ L_{obj}(\boldsymbol{c}, \boldsymbol{z} + \boldsymbol{y}_{symb}, \boldsymbol{t}; \boldsymbol{\theta}) \end{aligned} \tag{8}$$

4 Experimental Validation

We evaluate to what extent our proposed method is able to match an ad to its intended message (see Sec. 3.1). We present the baselines against which we compare (Sec. 4.1), our metrics (Sec. 4.2), quantitative results on our main ranking

task (Sec. 4.3), results on QA as classification (Sec. 4.4) and on three additional tasks (Sec. 4.5). Please see the supplementary file for implementation details, in-depth quantitative results, and qualitative results.

4.1 Baselines

We compare our ADVISE method (Sec. 3.6) to the following approaches from recent literature. All methods are trained on the Ads Dataset [22], using a train/val/test split of 60%/20%/20%, resulting in around 39,000 images and more than 111,000 associated statements for training.

- Hussain-Ranking adapts [22], the only prior method for decoding the message of ads. This method also uses symbol information, but in a less effective manner. The original method combines image, symbol, and question features, and trains for the 1000-way classification task. To adapt it, we pointwise-add the image features (Inception-v4 as for our method) and symbol features (distribution over 53 predicted symbols), and embed them in 200-D using Eq. 1 (using hard negative mining), setting $\boldsymbol{v}$ to the image-symbol feature. We tried four other ways (described in supp) of adapting [22] to ranking, but they performed worse.
- VSE++ [11] (follow-up to [30]) uses the same method as Sec. 3.2. It is representative of one major group of recent image-text embeddings using triplet-like losses [46, 40, 28, 51].
- VSE, which is like VSE++ but without hard negative mining, for a more fair comparison to the next baseline.
- 2-way Nets uses our implementation of [10] (published code only demoed the network on MNIST) and is representative of a second type of image-text embeddings using reconstruction losses [10, 21].

4.2 Metrics

We compute two metrics: Rank, which is the averaged ranking value of the highest-ranked true matching statement (highest possible rank is 1, which means first place), and Recall@3, which denotes the number of correct statements ranked in the Top-3. We expect a good model to have low Rank and high Recall scores. We use five random splits of the dataset into train/val/test sets, and show mean results and standard error over a total of 62,468 test cases (removing statements that do not follow the template "I should [action] because [reason].").

4.3 Results on the main ranking task

We show the improvement that our method produces over state of the art methods, in Table 1. We show the better of the two alternative methods from Sec. 3.5, namely KB-Symbols. Since public service announcements (e.g. domestic violence or anti-bullying campaigns) typically use different strategies and sentiments than product ads (e.g. ads for cars or coffee), we separately show the result for

Table 1. Our main result. We show two methods that do not use hard negative mining, and three that do. Our method greatly outperforms three recent methods in retrieving matching statements for each ad. All methods are trained on the Ads Dataset of [22]. The best method is shown in **bold**, and the second-best in *italics*

	Rank (Lower ↓ is better)		Recall@3 (Higher ↑ is better)	
Method	PSA	Product	PSA	Product
2-WAY NETS	4.836 (± 0.090)	4.170 (± 0.023)	0.923 (± 0.016)	1.212 (± 0.004)
VSE	4.155 (± 0.091)	3.202 (± 0.019)	1.146 (± 0.017)	1.447 (± 0.004)
VSE++	4.139 (± 0.094)	3.110 (± 0.019)	1.197 (± 0.017)	1.510 (± 0.004)
HUSSAIN-RANKING	*3.854* (± 0.088)	*3.093* (± 0.019)	*1.258* (± 0.017)	*1.515* (± 0.004)
ADVISE (ours)	**3.013** (± 0.075)	**2.469** (± 0.015)	**1.509** (± 0.017)	**1.725** (± 0.004)

Table 2. (Left) Ablation study on PSAs. All external knowledge components except attention improve over basic triplet embedding. (Right) Ablation on products. General-purpose recognition approaches, e.g. regions and attention, produce the main boost

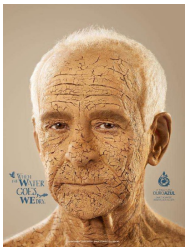
	PSA				Product			
			% improvement				% improvement	
Method	Rank ↓	Rec@3 ↑	Rank	Rec@3	Rank ↓	Rec@3 ↑	Rank	Rec@3
BASE TRIPLET	4.139	1.197			3.110	1.510		
GENERIC REGION	3.444	1.375	17	15	2.650	1.670	15	11
SYMBOL REGION	3.174	1.442	8	5	2.539	1.697	4	2
+ ATTENTION	3.258	1.428	-3	-1	2.488	1.726	2	2
+ SYMBOL/OBJECT	3.149	1.466	3	3	2.469	1.727	1	<1
+ KB OBJECTS	3.108	1.482	1	1	2.471	1.725	<1	<1
+ KB SYMBOLS	3.013	1.509	4	3	2.469	1.725	<1	<1

PSAs and products. We observe that our method greatly outperforms the prior relevant research. PSAs in general appear harder than product ads (see Sec. 1).

Compared to 2-WAY NETS [10], VSE which does *not* use hard negative mining is stronger by a large margin (14-23% for rank, and 19-24% for recall). VSE++ produces more accurate results than both 2-WAY NETS and VSE, but is outperformed by HUSSAIN-RANKING and our ADVISE. Our method is the strongest overall. It improves upon VSE++ [11] by 20-27% for rank, and 14-26% for recall. Compared to the strongest baseline, HUSSAIN-RANKING [22], our method is 20-21% stronger in terms of rank, and 13-19% stronger in recall. Fig. 3 shows a qualitative result contrasting the best methods.

We also conduct ablation studies to verify the benefit of each component of our method. We show the BASE TRIPLET embedding (Sec. 3.2) similar to VSE++; a GENERIC REGION embedding using image regions learned using [38] trained on the COCO [37] detection dataset; SYMBOL REGION embedding and ATTENTION (Sec. 3.3); adding SYMBOL/OBJECT constraints (Sec. 3.4); and including additive knowledge (Sec. 3.5) using either KB OBJECTS or KB SYMBOLS.

The results are shown in Table 2 (left for PSAs, right for products). We also show percent improvement of each new component, computed with respect to the previous row, except for KB OBJECTS and KB SYMBOLS, whose improvement is



VSE++: "I should try this makeup because its fun."

Hussain-ranking: "I should stop smoking because it destroys your looks."

ADVISE (ours): "I should be careful to how I treat Earth because when the water leaves we die."

VSE++: "I should wear Nivea because it leaves no traces."

Hussain-ranking: "I should be eating these because it has fresh ingredients."

ADVISE (ours): "I should buy GeoPack paper because their cutlery is eco-friendly."

Fig. 3. Our ADVISE method compared to the two stronger baselines. On the left, VSE++ incorrectly guessed this is a makeup ad, likely because often faces appear in makeup ads. HUSSAIN-RANKING correctly determined this is a PSA, but only our method was able to predict the topic, namely water/environment preservation. On the right, both HUSSAIN-RANKING and our method recognized the concepts of freshness/naturalness, but our method picked a more specific statement.

computed with respect to the third-to-last row, i.e. the method on which both KB methods are based. The largest increase in performance comes from focusing on individual regions within the image. This makes sense because ads are carefully designed and multiple elements work together to convey the message. We see that these regions must be learned as visual anchors to symbolic concepts (SYMBOL REGION vs GENERIC REGION) to further increase performance.

Beyond this, the story that the results tell differs between PSAs and products. Symbol/object constraints and additive branches are more helpful for the challenging, abstract PSAs that are the focus of our work. For PSAs, the additive inclusion of external information helps more when we directly predict the symbols (KB SYMBOLS), but also when we first extract objects and map these to symbols (KB OBJECTS). Note that KB SYMBOLS required 64,131 symbol labels. In contrast, KB OBJECTS relies on mappings between object and symbol words, which can be obtained more efficiently. While we obtain them as object-symbol similarities in our learned space, they could also be obtained from a purely textual, ad-specific resource. Thus, KB OBJECTS would generalize better to a new domain of ads (e.g. a different culture) where the data from [22] does not apply.

In Table 3, we show the object-symbol knowledge base that KB OBJECTS (Sec. 3.5) uses. We show "synonyms" across three vocabularies: the 53 symbol words from [22], the 27,999 words from the action/reason statements, and the 823 words from captions predicted for ads. We compute the nearest neighbors for each word in the learned space. This can be used as a "dictionary": If I see a given object, what should I predict the message of the ad is, or if I want to make a point, what objects should I use? In triplet ID 1, we see to allude to "comfort," one might use a soft sofa. From ID 2, if the statement contains "driving," perhaps this is a safe driving ad, where visuals allude to safety and injury, and contain cars and windshields. We observe the different role of "ketchup" (ID 3) vs "tomato" (ID 4): the former symbolizes flavor, and the latter health.

Table 3. Discovered synonyms between symbol, action/reason, and DenseCap words

ID	Symbol	Statement	DenseCap
1	*comfort*	couch, sofa, soft	pillow, bed, blanket
2	safety, danger, injury	*driving*	car, windshield, van
3	delicious, hot, food	*ketchup*	beer, pepper, sauce
4	food, healthy, hunger	salads, food, salad	*tomato*

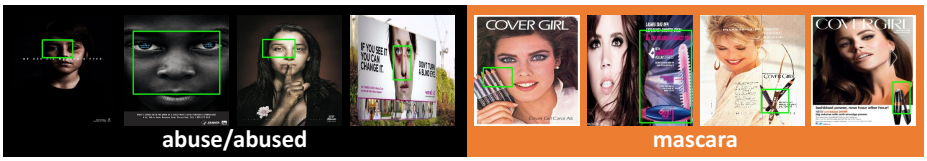


Fig. 4. Application for ads image retrieval (see details in supp). We extract the CNN feature of each image region (Eq. 4), then use the word embeddings of "abuse/abused" and "mascara" to retrieve the most similar image regions (denoted using green boxes).

In Fig. 4, we show the learned association between the individual words and symbolic regions. By learning from the ads image and statement pairs, our ADVISE model propagates words in the statement to the regions in the image thus associates each label-agnostic region proposal with semantically meaningful words. At training time, we have neither box-level nor word-level annotations.

4.4 Results on question-answering as classification

For additional comparison to [22], we evaluate our method on the question-answering task formulated as 1000-way single-word answer classification (Sec. 3.1). We now directly optimize for this classification task, but add our symbol-based region proposals, symbol/object constraints, and additive knowledge-based image representation. Our implementation of the method of Hussain et al. [22] pointwise-adds Inception-v4 image features and the symbol distribution, and obtains 10.03% top-1 accuracy on PSAs, and 11.89% accuracy on product ads (or 11.69% average across ads regardless of type, which is dominated by product ads, and is close to the 11.96% reported in [22]). Representing the image with a weighted summation of generic regions produced 10.42% accuracy for PSAs, and 12.45% for products (a 4% and 5% improvement, respectively). Using our method resulted in 10.94% accuracy for PSAs, and 12.64% for products (a 9% and 6% improvement over [22], respectively). Note that a method known to work well for many recognition tasks, i.e. region proposals, leads to very small improvement in the case of QA classification for ads, so it is unlikely that any particular method would lead to a large improvement on this task. This is why we believe the ranking task we evaluate in Sec. 4.3 is more meaningful.

Table 4. Other tasks our learned image-text embedding helps with. We show rank for the first two (lower is better) and homogeneity [52] for the third (higher is better)

Method	Hard statements (↓ better)	Slogans (↓ better)	Clustering (↑ better)
HUSSAIN-RANKING	5.595 (± 0.027)	4.082 (± 0.090)	0.291 (± 0.002)
VSE++	5.635 (± 0.027)	4.102 (± 0.091)	0.292 (± 0.002)
ADVISE (ours)	**4.827** (± 0.025)	**3.331** (± 0.077)	**0.355** (± 0.001)

4.5 Results on additional tasks

In Table 4, we demonstrate the versatility of our learned embedding, compared to the stronger two baselines from Table 1. None of the methods were retrained, i.e. we simply used the pre-trained embedding evaluated on statement ranking. First, we show a harder statement retrieval task: all statements that are to be ranked are from the same topic (e.g. all statements are about car safety or about beauty products). The second task uses creative captions that MTurk workers were asked to write for 2,000 ads in [22]. We rank these slogans, using an image as the query, and report the rank of the correct slogan. Finally, we check how well an embedding clusters ad images with respect to a ground-truth clustering defined by the topics of the ads.

5 Conclusion

We presented a method for matching image advertisements to statements which describe the idea of the ad. Our method uses external knowledge in the form of symbols and predicted objects in two ways, as constraints for a joint image-text embedding space, and as an additive component for the image representation. We also verify the effect of state-of-the-art object recognition techniques in the form of region proposals and attention. Our method outperforms existing image-text embedding techniques [10, 11] and a previous ad-understanding technique [22] by a large margin. Our region embedding relying on visual symbolic anchors greatly improves upon traditional embeddings. For PSAs, regularizing with external info provides further benefit. In the future, we will investigate other external resources for decoding ads, such as predictions about the memorability or human attention over ads, and textual resources for additional mappings between physical and abstract content. We will use our object-symbol mappings to analyze the visual variability the same object category exhibits when used for different ad topics.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number 1566270. This research was also supported by 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. We thank the anonymous reviewers for their feedback and encouragement.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
2. Anne Hendricks, L., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: Describing novel object categories without paired training data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
4. Bylinskii, Z., Alsheikh, S., Madan, S., Recasens, A., Zhong, K., Pfister, H., Durand, F., Oliva, A.: Understanding infographics through textual and visual tag prediction. arXiv preprint arXiv:1709.09215 (2017)
5. Cao, Y., Long, M., Wang, J., Liu, S.: Deep visual-semantic quantization for efficient image retrieval. In: CVPR (2017)
6. Chen, K., Bui, T., Fang, C., Wang, Z., Nevatia, R.: Amc: Attention guided multimodal correlation learning for image search. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
7. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: Computer Vision and Pattern Recognition (CVPR). IEEE (2016)
8. Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu, J., Sun, M.: Show, adapt and tell: Adversarial training of cross-domain image captioner. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
9. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
10. Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: CVPR (2017)
11. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612 (2017)
12. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
13. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: Stylenet: Generating attractive visual captions with styles. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
15. Goo, W., Kim, J., Kim, G., Hwang, S.J.: Taxonomy-regularized semantic deep convolutional neural networks. In: European Conference on Computer Vision. pp. 86–101. Springer (2016)
16. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
17. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)

18. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
19. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
20. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
21. Hubert Tsai, Y.H., Huang, L.K., Salakhutdinov, R.: Learning robust visual-semantic embeddings. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
22. Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Kovashka, A.: Automatic understanding of image and video advertisements. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
23. Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J., Daume, III, H., Davis, L.S.: The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
24. Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Inferring and executing programs for visual reasoning. In: ICCV (2017)
25. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
26. Joo, J., Li, W., Steen, F.F., Zhu, S.C.: Visual persuasion: Inferring communicative intents of images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
27. Joo, J., Steen, F.F., Zhu, S.C.: Automated facial trait judgment and election outcome prediction: Social dimensions of face. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3712–3720 (2015)
28. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
29. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European Conference on Computer Vision. pp. 235–251. Springer (2016)
30. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. In: TACL (2015)
31. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
32. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
33. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)

34. Leigh, J.H., Gabel, T.G.: Symbolic interactionism: its effects on consumer behaviour and implications for marketing strategy. Journal of Services Marketing **6**(3), 5–16 (1992)
35. Levy, S.J.: Symbols for sale. Harvard business review **37**(4), 117–124 (1959)
36. Li, X., Hu, D., Lu, X.: Image2song: Song retrieval via bridging image content and lyric words. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
38. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
39. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR (2017)
40. Mai, L., Jin, H., Lin, Z., Fang, C., Brandt, J., Liu, F.: Spatial-semantic image search by visual feature synthesis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
41. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
42. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
43. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
44. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: ICCV (2017)
45. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: CVPR (2017)
46. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Hierarchical multimodal lstm for dense visual-semantic embedding. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
47. Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
48. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
49. Ren, M., Zemel, R.S.: End-to-end instance segmentation with recurrent attention. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
50. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
51. Ren, Z., Jin, H., Lin, Z., Fang, C., Yuille, A.: Joint image-text representation by gaussian visual-semantic embedding. In: Proceedings of the 2016 ACM on Multimedia Conference. pp. 207–211. ACM (2016)
52. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: EMNLP-CoNLL. vol. 7, pp. 410–420 (2007)

53. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015)
54. Scott, L.M.: Images in advertising: The need for a theory of visual rhetoric. Journal of consumer research **21**(2), 252–273 (1994)
55. Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., Schiele, B.: Speaking the same language: Matching machine to human captions by adversarial training. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
56. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Computer Vision and Pattern Recognition (CVPR). IEEE (2016)
57. Spears, N.E., Mowen, J.C., Chakraborty, G.: Symbolic role of animals in print advertising: Content analysis and conceptual development. Journal of Business Research **37**(2), 87–95 (1996)
58. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017), https://arxiv.org/abs/1602.07261
59. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
60. Teney, D., Anderson, P., He, X., van den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
61. Vedantam, R., Bengio, S., Murphy, K., Parikh, D., Chechik, G.: Context-aware captions from context-agnostic supervision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
62. Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., Saenko, K.: Captioning images with diverse objects. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
63. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
64. Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A.: Fvqa: fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
65. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
66. Wu, Q., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Ask me anything: Free-form visual question answering based on knowledge from external sources. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
67. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision (ECCV). Springer (2016)
68. Yang, L., Tang, K., Yang, J., Li, L.J.: Dense captioning with joint inference and visual context. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
69. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Computer Vision and Pattern Recognition (CVPR). IEEE (2016)

70. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
71. Young, C.E.: The advertising research handbook. Ideas in Flight (2005)
72. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
73. Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.H., Kim, G.: Supervising neural attention models for video captioning by human gaze data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
74. Yu, Y., Ko, H., Choi, J., Kim, G.: End-to-end concept word detection for video captioning, retrieval, and question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
75. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
76. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
77. Zhu, Y., Lim, J.J., Fei-Fei, L.: Knowledge acquisition for visual question answering via iterative querying. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)