Breaking Shortcuts by Masking for Robust Visual Reasoning

Keren Ye, Mingda Zhang and Adriana Kovashka
Department of Computer Science, University of Pittsburgh, Pittsburgh PA, USA
{yekeren, mzhang, kovashka}@cs.pitt.edu

Abstract

Visual reasoning is a challenging but important task that is gaining momentum. Examples include reasoning about what will happen next in film, or interpreting what actions an image advertisement prompts. Both tasks are "puzzles" which invite the viewer to combine knowledge from prior experience, to find the answer. Intuitively, providing external knowledge to a model should be helpful, but it does not necessarily result in improved reasoning ability. An algorithm can learn to find answers *to the prediction task yet not perform generalizable* reasoning. *In other words, models can leverage "shortcuts" between inputs and desired outputs, to bypass the need for reasoning. We develop a technique to effectively incorporate external knowledge, in a way that is both interpretable, and boosts the contribution of external knowledge for multiple complementary metrics. In particular, we mask evidence in the image and in retrieved external knowledge. We show this masking successfully focuses the method's attention on patterns that generalize. To properly understand how our method utilizes external knowledge, we propose a novel side evaluation task. We find that with our masking technique, the model can learn to select useful knowledge pieces to rely on.*[1]

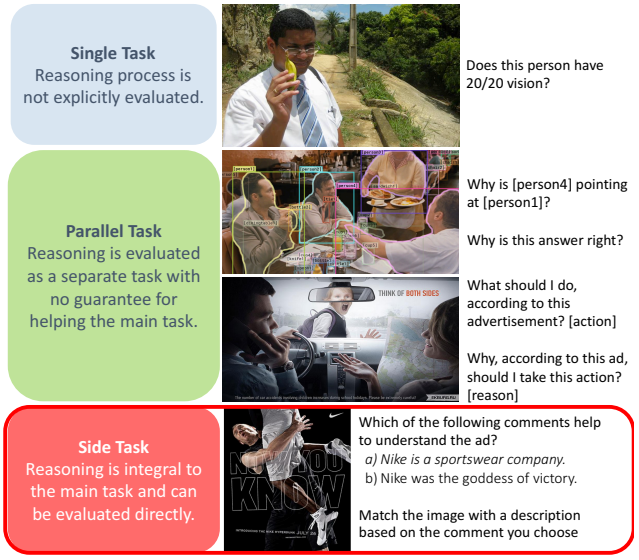

Figure 1: **Visual reasoning tasks**. Previous definitions either oversimplify reasoning (as answering, top) or treat it as a standalone task parallel to answering (middle). One of our contributions is a new evaluation side task (bottom) that checks the decisions made by our model, i.e. which knowledge pieces it selected to complete the answering task.

1. Introduction

Visual reasoning is an important family of problems including visual question answering (VQA) [5, 9, 12, 41] and visual commonsense reasoning (VCR) [56]. The name "reasoning" bears a flavor of classic AI and structured logic-inspired inference steps; one might argue that a human accumulates knowledge as they mature, and they store this knowledge in a metaphorical "knowledge base", then retrieve information from it as needed. Indeed, some approaches to VQA/VCR do rely on structured, symbolic reasoning [4, 18, 44, 47]. However, in many domains state of the art performance is achieved by end-to-end transformer models [6, 27, 43] or other attention models [3, 16] which do not perform structured reasoning. These models excel when sufficient labeled data is available, and potentially a large pool of image-text data in a disjoint domain, because they can effectively learn to mimic patterns in the data.

However, we highlight two *limitations* of existing methods for reasoning tasks. First, even though human reasoning is grounded in knowledge accumulated over the years from multiple sources, most methods just leverage data from the human-curated target dataset. Second, these models often learn shortcuts which do not generalize well; for example, they might learn to perform string or object matching between question/image and answers, rather than reasoning about properties and causality.

We propose a mechanism to effectively incorporate external knowledge for a task that especially requires it. To properly leverage the benefit that external knowledge can provide, we enable the model to filter irrelevant knowledge,

[1] Our code is available at https://github.com/yekeren/Ads-KB.

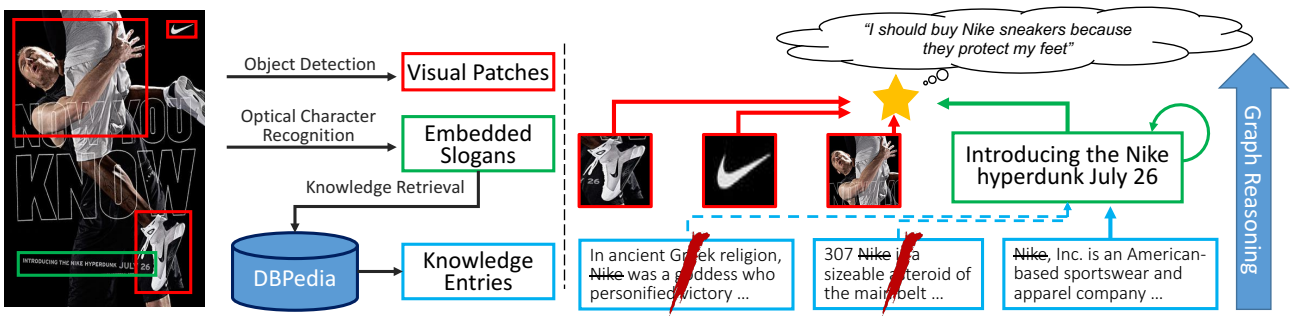

Figure 2: **Overview of the proposed model.** Given a single image ad, we first expand the representation using object detection and OCR, and also retrieve relevant knowledge based on slogan snippets (left). We build a graph-based model to infer the ad's message (top-right) using all available information (right), but allow the model to filter irrelevant knowledge (shown with red slashes). For more effective training, we randomly mask query keywords and other tokens (crossed-out).

and make training more robust, by partly blocking the effect of shortcuts through masking.

In particular, we first present a graph-based model that represents the image meaning using visual information, embedded textual information, and information from external knowledge bases. Unlike most approaches that treat knowledge as always-correct-and-useful, we require the model to learn to filter out irrelevant information provided in the uncontrolled environment (e.g. paragraphs from DBPedia), by learning to dynamically downweight some graph edges.

Then, we use the presented model to study how the model uses external knowledge to reason. We find that models exploit "shortcuts", namely they can (1) select the true label, *without* (2) finding the correct and helpful knowledge for prediction. We call the former the *main task* and the latter a *side task* in that most benchmarks only evaluate (1) but not (2). We study the phenomenon in detail and use a stochastic masking technique which prevents the model from leveraging shortcuts. The masking forces the model to "work harder" and learn more generalizable relationships. With our proposed masking, our model better utilizes knowledge, resulting in gain in performance on the main task, but an even larger boost on the side task. Note that we do not collect annotations for the side task for training, but only for evaluation.

We test our framework on a reasoning task where external knowledge is especially necessary due to the creative nature of the images. We use the Ads dataset of [14] since advertisements naturally anchor the message of the image in information from external world (e.g., brand names, celebrities, etc.). Given a visual ad, the method should retrieve the correct "action-reason" statement which captures the *action* that the ad implies the viewer should take and *reasons* it provides for taking the suggested action. The word "reason" in this context is akin to "rationale". In contrast, by "reason*ing*" we mean the ability to use the right evidence to select a statement. Fig. 2 shows an example of expected correct reasoning on an ad: knowing Nike is a sportswear company and observing the shoe is the key to match to "*I should buy Nike sneakers because they protect my feet*".

The knowledge required to understand the ads is usually domain-specific and focused (e.g., details regarding brand names and persons). In contrast to using an image-text disjoint dataset to pre-train, as done for other VQA/VCR benchmarks, we use a sparse number of knowledge pieces (i.e. from DBPedia). Thus, it is more feasible to verify the correct use of knowledge in our setting (with our side task).

As for the benefits of using masking to learn generalizable features, we show that masking allows our model to improve the standard metrics used for evaluating advertisement understanding (main task). Besides, we verify that the external knowledge our method chose to use, is actually supporting the reasoning (side task). We show that the simple masking strategy more than *doubles* the accuracy of the knowledge selection (evaluated in the side task).

To summarize, our contributions are as follows:

- a bottom-up graph model that utilizes external knowledge and filters irrelevant knowledge,
- a method to effectively utilize external knowledge, by masking retrieved knowledge and image evidence, to prevent the model from learning shortcuts, and
- a new side task with annotations to evaluate reasoning.

2. Related work

Use of external knowledge in VQA. Early benchmarks [5, 8, 58] provide only the image and paired question/answers, but more recent approaches incorporate diverse resources external to the target corpus. In this work, we focus on discrete external knowledge (e.g. facts in a knowledge base), rather than pretraining on multimodal data in unsupervised fashion (i.e. learning better image and text representations). Some prior knowledge-based visual reasoning methods assume applicable facts or background knowledge are present in the VQA dataset itself [29, 39, 42, 47, 46]. However, in the NLP domain,

[15, 17, 48] showed on the SQuAD [35] benchmark that providing *always-relevant* knowledge is not good practice since the learned models do not necessarily properly use the facts to reason. Others methods [31, 32, 1] use a separate, general knowledge base (e.g. ConceptNet), predict whether the answer is in the knowledge base and choose the most suitable answer candidate. Our method assumes this more challenging setting (only general knowledge base is available). In contrast to [31, 32], we explicitly evaluate the ability of the method to choose relevant knowledge, without requiring additional annotations at training time (through our side task). Importantly, we propose a new mechanism to make the incorporation of knowledge more effective than simply using additional graph nodes [31, 32], and without requiring the use of a special relational engine [1].

Metrics for reasoning ability. VQA datasets typically ask a *single* question, i.e. while answering is explicitly evaluated, reasoning evaluation is only implicit. This setting is not suitable for verifying the effectiveness of external knowledge usage. In addition to the main metric (which measures the accuracy of answer prediction), we explicitly evaluate a method's reasoning capability, i.e., whether the model could find the correct knowledge piece to use. Other methods e.g. [56, 14] also incorporate additional metrics, e.g. the model needs to provide a rationale for its answer. However, they treat answering and reasoning as parallel tasks, and do not enforce the answer prediction to be based on the rationale. In our setting, answering directly depends on reasoning, thus evaluating answering verifies the output, while evaluating reasoning verifies the *inner workings* of the algorithm. See Fig. 1 for a comparison of tasks/metrics.

Dataset bias. Many works studied the VQA benchmark validity, e.g. [8, 58] retrospected on organizing the VQA challenge and proposed methods to improve the datasets. [36] studied the language priors in the VQA dataset, and forced the method to look at the image; we instead (implicitly) force it to look at external knowledge.

Ads understanding. We focus on advertisement understanding as our testbed to study the incorporation of external knowledge. This is because ads often appeal to human associations (guns are dangerous, vegetables are healthy) that are not explicitly stated in images and cannot be easily learned from the dataset itself. [14] provided action-reason statement annotated by multiple humans ("What action should the viewer take based on the ad? What reason does the ad provide for taking the suggested action?"). [52] proposed a cross-modal retrieval task to match the human-annotated statements with help from captioning and symbol prediction models, [2] used a symbolism-based attention model, and [33, 53] additionally used textual slogans in the image extracted with OCR techniques. Instead of using an embedding from a single modality or fusing the multi-modal features, we use a graph and allow message passing between modalities. The learned weights in the graph structure capture the model's reasoning and can be used to gauge "How does the model incorporate external knowledge to reason about an ad?".

Explainable models. Our focus is on ensuring and evaluating a model's ability to select reliable evidence (i.e. external knowledge), *not* on the explainability/interpretability of models to a human. We care about the correctness of knowledge pieces used, rather than how interpretable the model's selections are. Prior work [10, 13] collects explanation annotations and requires a model to point to the human-annotated reasons for an effect—for example, finding the spatial location in an image that directly affects a model's prediction. Unlike our work, these require annotation effort, i.e. humans provide explanations for training. Attention mechanisms [28, 34, 50, 55, 30, 40, 49, 51] and graph convolutional methods [21, 31, 38] are another way to achieve explainability. They optimize a primary goal while also learning the reliability of different evidence. Our approach is similar in that we do not require additional supervision, but we explicitly study the relation between *choosing correct supportive evidence* and *predicting the correct answer*.

3. Approach

We focus on one specific reasoning task, namely advertisement understanding. We incorporate image regions, text in the image, and external DBpedia knowledge [26], in a graph model. Because we retrieve knowledge from an open, general, real-world knowledge base, retrieved *irrelevant* pieces of knowledge dominate in count. We thus allow our model to select which pieces of knowledge and information to leverage, using learnable scalar edge weights.

One interesting but easy to neglect problem is that when the answer options can easily be matched to the image evidence, additional information (external knowledge) may not be necessary and hence may not help performance on the main task. Fig. 2 shows an example in the Ads dataset: given a Nike ad with an embedded slogan containing the word "Nike", the model must retrieve external knowledge to infer the particular properties that this ad demonstrates, so it can select the correct action-reason statement. However, the model can also find a shortcut and *not* perform reasoning, by merely looking for potential choices containing the brand name (e.g. simply matching "Nike" between the slogan, which is part of the input, and the word "Nike" in one of the answer options). Another example is the famous PepsiCo celebrity branding, where a naive model can simply *remember* popular celebrities and directly match them to "Pepsi" rather than understanding their shared characteristics (e.g. athleticism), thus it may generalize poorly if a new spokesperson is introduced in the ads. This means a model can correctly **answer** without **reasoning** correctly (i.e. without squeezing more useful information out of the

retrieved knowledge and without using the right knowledge). We refer to this phenomenon as a **shortcut effect**. Quantitatively, the gap between answering and reasoning is demonstrated by the difference in performance we obtain on the main answering task and the side knowledge selection task (Sec. 4). While we study shortcut effects in the Ads dataset, we want to point out that similar issues existing in other datasets. We show a small example in VCR [56], where the subject repetition seems to be the trick to answer the question without knowing the visual cues:

How is Jackie feeling? Avery is very excited.
How is Jackie feeling? Jackie is focused and active.

Below, we first describe the advertisement understanding task (Sec. 3.1). We introduce our overall framework and how we train (Sec. 3.2). We describe our image representation (Sec. 3.3-3.4) and knowledge selection mechanisms (Sec. 3.5). Finally, we describe our strategy for breaking shortcuts and forcing the model to "study harder" and learn more generalizable patterns (Sec. 3.6).

3.1. Task: Advertisement understanding

We focus on the advertisement understanding task [14] because it considers an interesting and practical scenario. First, ads exploit symbols that refer to content outside the image; thus, retrieving external knowledge is required. Second, unlike [39, 56], neither external knowledge nor reasoning rationales are available in clean form. Third, multiple modalities (image and slogan text) must be considered.

For each image, [14] provide three statements in which each is an action-reason pair (e.g., "I should buy Nike because it protects my feet."). There may be multiple plausible reasons per action, e.g. to buy "sportswear", the image may argue "it protects", "is cheap", or "celebrity wears it". Models are required to match an advertisement with the correct *action-reason* descriptive statement.

Given an ad image A, we assume it is composed of two parallel entity sets $A = \{V, T\}$, where V stands for visual signals and T represents the embedded slogans (i.e. textual signals). For each image, we generate a group of object proposals as the salient visual signals from the ad, noted as $V = \{v_1, v_2, \ldots, v_{|V|}\}$. We also use existing optical character recognition (OCR) engines to extract embedded text slogans as $T = \{t_1, t_2, \ldots, t_{|T|}\}$.

3.2. Training: Matching to the statements

We follow the approach in [52] and use triplet loss (Eq. 1) to optimize the cosine similarity $\text{cosine}(\mathbf{h}, \mathbf{s}) = \frac{\mathbf{h}\cdot\mathbf{s}}{\|\mathbf{h}\|\|\mathbf{s}\|}$ between advertisement representation $\mathbf{h}$ and answer choice statement embedding $\mathbf{s}$. Eq. 1 ensures that paired image and answer choices should be more similar than unpaired ones (i.e., $\text{cosine}(\mathbf{h}, \mathbf{s}_+) > \text{cosine}(\mathbf{h}, \mathbf{s}_-)$). $\mathbf{s}_+$ denotes the embedding of a paired annotation, $\mathbf{s}_-$ is a sampled statement embedding in the mini-batch, using semi-hard mining [37], and η is the margin in the triplet loss.

$$L(\mathbf{h}, \mathbf{s}) = \max(0, \text{cosine}(\mathbf{h}, \mathbf{s}_-) - \text{cosine}(\mathbf{h}, \mathbf{s}_+) + \eta) \quad (1)$$

We encode statements $\mathbf{s} = \mathbf{W}_s \text{BiLSTM}(\psi(s); \boldsymbol{\theta}_s) \in \mathbb{R}^{D\times 1}$, where ψ is the word embedding process, $\boldsymbol{\theta}_s$ denotes the parameters of the statement encoder, and $\mathbf{W}_s$ is for the linear layer. Below we describe how we represent the ad image $\mathbf{h}$ using a graph. During inference, models pick the most probable statement from candidates according to cosine similarity: $\underset{\mathbf{s}\in\text{candidates}}{\text{argmax}}\ \text{cosine}(\mathbf{h}, \mathbf{s})$.

3.3. Image representation graph: Nodes and edges

Briefly, an image $\mathbf{h}$ is partially represented using slogan text found in the image; in turn, these slogans are represented using external information found using the slogans as queries. Our image representation graph contains four types of nodes (image, slogan, knowledge and a global node), and three types of edges connecting these nodes.

Image nodes. For each image proposal $v_i \in V$, we use a pre-trained model to extract its feature $\text{CNN}(v_i)$. The embedding of v_i, denoted as $\mathbf{v}_i \in \mathbb{R}^{D\times 1}$, is obtained as a linear projection $\mathbf{v}_i = \mathbf{W}_v \text{CNN}(v_i)$ where $\mathbf{W}_v$ is the parameter.

Slogan nodes. We represent each OCR-detected slogan $t_i \in T$ using a BiLSTM encoder, then project it into the same feature space as the image: $\mathbf{t}_i^{(0)} = \mathbf{W}_t \text{BiLSTM}(\psi(t_i); \boldsymbol{\theta}_t) \in \mathbb{R}^{D\times 1}$. As OCR may produce noisy detections, model weights $\boldsymbol{\beta}$ discussed below (Eq. 3) choose which OCR results to use.

Knowledge nodes. Since the embedded slogans in ads are usually succinct, abbreviated, or ambiguous [24, 57], an external database will be used as a source of knowledge to help enriching and clarifying the meaning of the slogans. Specifically, we send each word in slogan t_i to the DBpedia knowledge base [26] as a query. This retrieval process φ returns a set of related comments. For example, φ("WWF")[2] returns the explanations of "Windows Workflow Foundation", "Words with Friends", and "World Wide Fund for Nature". We take the union of the retrieved knowledge entries to enrich a slogan, denoted as $\phi(t_i) = \bigcup_{q\in t_i} \varphi(q)$. In Fig. 2, the blue boxes show these extended pieces of knowledge for a specific slogan. Our model will learn to select the relevant ones using the weights $\boldsymbol{\alpha}$ in Eq. 2.

For external knowledge $k_{i,j} \in \phi(t_i)$ (with j ranging over all retrieved comments for slogan t_i), we use a separate BiLSTM encoder $\mathbf{k}_{i,j} = \mathbf{W}_k \text{BiLSTM}(\psi(k_{i,j}); \boldsymbol{\theta}_k) \in \mathbb{R}^{D\times 1}$. Note that knowledge nodes share the word embedding process ψ with slogan nodes and human-annotated statements but not the BiLSTM encoder, because we suppose word meanings in different modalities (DBpedia comments, slogans, action-reason statements) are the same, but the grammar structures may differ.

[2] http://dbpedia.org/page/WWF

Edges. We build an inference graph (DAG) to capture the relationships for a better understanding of the image. We treat all the proposals, slogans, and knowledge pieces as nodes, with the knowledge nodes connected to the associated slogans by `IsADescriptionOf` edges. Next, we add a *global node* as an overall representation and connect all proposals and slogans to it using `ContributesTo` edges. The representation of the global node will be used to facilitate message passing and graph inference (described next). We also add extra `IsIdenticalTo` self-looping connections to all slogan nodes. Fig. 2 shows an example.

3.4. Image representation graph: Inference

Our method propagates information in a bottom-up manner and adjusts edge weights to optimize the final image representation $\mathbf{h}$ (Eq. 1). This inference procedure is similar to the Graph Convolutional Network (GCN) [21] in that we both use message passing to deduce the uncertain node embeddings. However, we fuse global context information to compute the edge weights, while GCN considers only the local information among neighbors.

Updating slogan embeddings. The slogan t_i chooses a meaning (soft selection using the $\boldsymbol{\alpha}$ weights) among its initial embedding $\mathbf{t}_i^{(0)}$ and representations of the retrieved DBpedia comments $\mathbf{k}_{i,j}$.

$$\mathbf{t}_i^{(1)} = \underbrace{\alpha_{i,0}\mathbf{t}_i^{(0)}}_{\text{original meaning}} + \underbrace{\sum_{j=1}^{|\phi(t_i)|} \alpha_{i,j}\mathbf{k}_{i,j}}_{\text{descriptions from extra knowledge}} \tag{2}$$

The weight vector $\boldsymbol{\alpha}_i \in \mathbb{R}^{1+|\phi(t_i)|}$ denotes the incoming edge scores for a slogan node t_i, where $\alpha_{i,0}$ is the weight of the self-loop edge `IsIdenticalTo`, and $\alpha_{i,j}$ ($j \in \{1, \ldots, |\phi(t_i)|\}$) are the weights of `IsADescriptionOf` edges. We require that $\sum_{j=0}^{|\phi(t_i)|} \alpha_{i,j} = 1$. We describe how we learn $\boldsymbol{\alpha}$ shortly.

The **global embedding** $\mathbf{h}$ is a weighted sum of image patches and updated slogan embeddings.

$$\mathbf{h} = \underbrace{\sum_{i=1}^{|V|} \beta_i \mathbf{v}_i}_{\text{messages from proposals}} + \underbrace{\sum_{i=|V|+1}^{|V|+|T|} \beta_i \mathbf{t}_i^{(1)}}_{\text{messages from slogans}} \tag{3}$$

Specifically, we define a vector $\boldsymbol{\beta} \in \mathbb{R}^{|V|+|T|}$ denoting the weights of different `ContributesTo` edges. The first $|V|$ values are the contributions of image proposals and the next $|T|$ denote slogans. We require $\sum_{i=1}^{|V|+|T|} \beta_i = 1$.

3.5. Image representation graph: Edge weights

The weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ allow our model to choose which knowledge pieces and slogans to use. We show the knowledge pieces chosen (with α larger than 0.05) in Fig. 3; thicker arrows correspond to larger values of α, β.

We use an image-guided attention mechanism to infer $\boldsymbol{\alpha}$ (Eq. 2) hence choose whether to incorporate the external information or maintain the original slogan feature. This choice depends (1) the relation between the node and the connected slogan target, and (2) the relation between the node and the image context. We use a group of three-layer perception models denoted as $\text{MLP}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ to model the relations between any two types of feature vectors ($\mathbf{x}, \mathbf{y} \in \mathbb{R}^{D\times 1}$). In Eq. 4, $[;]$ denotes concatenation, and $\cdot$ pointwise multiplication; $\boldsymbol{\theta} = (\mathbf{W}_1, \mathbf{W}_2)$ denotes parameters of a specific relation MLP, in which $\mathbf{W}_1$, $\mathbf{W}_2$ are parameters.

$$\text{MLP}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \mathbf{W}_2 \text{tanh}(\mathbf{W}_1 [\mathbf{x}; \mathbf{y}; \mathbf{x} \cdot \mathbf{y}]) \tag{4}$$

Eq. 5 defines the edge weights connecting to textual slogans t_i. We define the image context $\bar{\mathbf{v}} = \frac{1}{|V|}\sum_{i=1}^{|V|} \mathbf{v}_i$. $\boldsymbol{\theta}_\alpha^t$ and $\boldsymbol{\theta}_\alpha^c$ are the parameters of the node-slogan and node-context MLPs. These MLPs measure how strong is the relationship between a node and the target slogan, and between a node and the image context.

$$a_{i,j} = \begin{cases} \text{MLP}(\mathbf{t}_i^{(0)}, \mathbf{t}_i^{(0)}; \boldsymbol{\theta}_\alpha^t) + \text{MLP}(\mathbf{t}_i^{(0)}, \bar{\mathbf{v}}; \boldsymbol{\theta}_\alpha^c) \\ \qquad\qquad \text{when } j = 0 \\ \text{MLP}(\mathbf{k}_{i,j}, \mathbf{t}_i^{(0)}; \boldsymbol{\theta}_\alpha^t) + \text{MLP}(\mathbf{k}_{i,j}, \bar{\mathbf{v}}; \boldsymbol{\theta}_\alpha^c) \\ \qquad\qquad \text{when } 1 \le j \le |\phi(t_i)| \end{cases} \tag{5}$$

$$\boldsymbol{\alpha}_i = \text{softmax}(\boldsymbol{a}_i)$$

To compute weight vector $\boldsymbol{\beta}$, we update the slogan context $\bar{\mathbf{t}}^{(1)} = \frac{1}{|T|}\sum_{i=1}^{|T|} \mathbf{t}_i^{(1)}$, then use Eq. 6. This is a co-attention mechanism in that we use visual context to determine weights of slogan nodes, and use slogan context to decide contributions of image proposals. When there is no slogan detected, the image features will dominate.

$$b_i = \begin{cases} \text{MLP}(\mathbf{v}_i, \bar{\mathbf{t}}^{(1)}; \boldsymbol{\theta}_\beta^v) \\ \qquad\qquad \text{when } 1 \le i \le |V| \\ \text{MLP}(\mathbf{t}_i^{(1)}, \bar{\mathbf{v}}; \boldsymbol{\theta}_\beta^t) \\ \qquad\qquad \text{when } |V|+1 \le i \le |V|+|T| \end{cases} \tag{6}$$

$$\boldsymbol{\beta} = \text{softmax}(\boldsymbol{b})$$

3.6. Masking for effective knowledge utilization

As we show in our experiments, combining the knowledge directly with the image and text, despite the learned edge weights, achieves small gains over using image and text alone. As we show in Fig. 3, our model as described so far often ascribes small weights $\boldsymbol{\alpha}$ to external knowledge retrieved. We discussed this "shortcut learning" phenomenon in Sec. 1. Thus, we next focus the model's attention towards important cues and knowledge pieces for reasoning, using a set of automatic masking strategies. To cope with this problem, we propose a simple yet effective masking strategy to break shortcut learning. For example, we replace the query

from the retrieved paragraph with the out-of-vocabulary token. In this way, the two pieces of knowledge in Fig. 2 become "[oov] is a sportswear company" and "[oov] is the name of an asteroid". Then the model can figure out whether "sportswear" or "asteroid" helps more for understanding the ad. At test time, when the model sees a rare sportswear company, it can benefit from the retrieved knowledge and not fail due to failed word-matching.

Our masking is similar to dropout (which we **do use** for our **baseline**), but applied over pieces of evidence in the slogan, knowledge comments, or action-reason statements. It is also similar to masking in cross-modal transformer methods [27, 6] but (1) we do not train the method to recover the masked symbol, and (2) transformer methods do not employ external knowledge, which is the key focus of our work.

We experiment with the following masking strategies in which the first two are only applied during training while the last one is used for both training and inference.

- M_t randomly drops a detected textual (T) slogan, with a probability of 0.5.
- M_s randomly sets the query words (e.g. "WWF" or "Nike") in the human-annotated statements (S) to the out-of-vocabulary token, with probability 0.5.
- M_k replaces the DBpedia queries in the retrieved knowledge contents with the out-of-vocabulary token.

In Tab. 2, we show that these strategies are more effective than masking over the image [25].

We found the masking strategy helps to significantly improve the main task of retrieving an answer. Moreover, when we evaluated the relevance of the knowledge pieces our model chose using weights α, we found an even more significant margin. While our masking strategy is specific to our target domain, masking in general merits exploration as a technique to aid in knowledge-based reasoning.

4. Experiments

Dataset. We use the data from the 2018 ad understanding challenge [23]. There are 51,223 *trainval* images paired with 161,557 annotated statements; and 12,805 *test* images, each with 3 correct statements and 12 incorrect distractions (15 in total). We use Google Cloud Vision OCR [7] to recognize the embedded textual slogans. We retrieve DBpedia comments based on detected slogans; an example SPARQL query is shown in our supplementary file. Eventually we obtain 443,747 detected textual slogans, and 30,747 unique knowledge descriptions, to be associated with the 64,028 images (*trainval+test*). Each image is annotated with, on average, 6.9 slogans and 27.5 DBpedia comments.

Main task metrics. Following the convention in the Ads challenge, we report accuracy (aka. precision@1) to compare against other methods from the challenge. However, we note that statement retrieval *accuracy* on the original task (3 correct with 12 incorrect statements) is not distinguishable enough, as many methods tie on this metric. To mitigate this issue, we additionally report min and avg *rank* (of the three correct statements) and *recall@K* scores, inspired by [20, 22, 45]. Further, we created two additional "harder" test sets named Sampled-100 and Sampled-500, where each image is accompanied by 3 correct statements and 97 (or 497) incorrect distracting options.

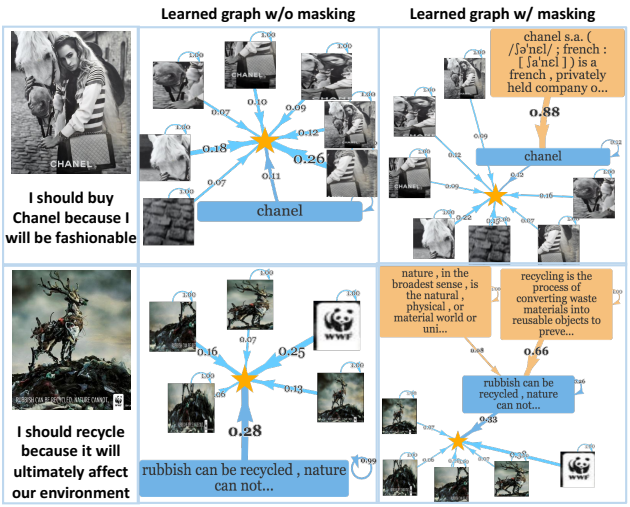

Figure 3: **Examples of the learned graphs (best with zoom).** We show the ad image and annotated action-reason statements on the left, the graph learned without masking in the middle, and that learned with masking (our approach) on the right. We show slogans in blue, DBpedia comments in orange, and the global node as a star. **Arrow thickness is correlated with learned weights α, β.** For visualization we removed all edges with small weights (threshold=0.05). Our method more **effectively leverages external information,** as it relies on appropriate knowledge (in orange) more than the baseline method w/o masking does.

Side task. We recruit human annotators to manually verify whether the retrieved knowledge is helpful for the ad understanding task. Specifically, for a given advertisement, we show all retrieved knowledge pieces and ask humans to annotate whether each piece is helpful or not in understanding the ad. These annotations serve as "gold standard" for knowledge selection evaluation (Sec. 4.3). We provide details in supp. Note these annotations are never used to train.

Training details. We use a pre-trained object detector [54] to generate 10 proposals per image and keep the 20 largest OCR detected regions. Note we only use the proposal regions, without any labels. Since we do not have ground-truth annotations for both objects and slogans, we manually verified proposal and OCR outputs are reasonable. Improving these models may increase performance, but we did not test alternatives since we care only about the relative contribution (with/without masking) as a mech-

Method	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10	Min Rank	Avg Rank
Results on the Challenge-15 task										
V,T	**87.3**	76.6	55.1	30.6	**28.4**	74.2	87.9	97.5	1.26	3.02
V,T+K	**87.3**	76.6	55.1	30.6	**28.4**	74.3	87.9	97.6	1.25	3.02
OURS: V,T+K(M)	**87.3**	**77.5**	**55.9**	**30.8**	**28.4**	**75.2**	**89.2**	**98.2**	**1.23**	**2.91**
Results on the Sampled-100 task										
V,T	79.8	66.5	46.9	26.2	26.0	64.4	74.9	83.5	2.38	7.52
V,T+K	80.0	67.0	47.0	26.1	26.0	64.9	75.1	83.4	2.29	7.49
OURS: V,T+K(M)	**80.2**	**67.9**	**47.9**	**26.8**	**26.1**	**65.8**	**76.6**	**85.4**	**2.14**	**6.56**
Results on the Sampled-500 task										
V,T	**65.5**	52.3	37.8	21.7	**21.3**	50.5	60.4	69.0	8.18	30.1
V,T+K	65.4	52.3	38.0	21.9	**21.3**	50.6	60.7	69.6	7.60	30.0
OURS: V,T+K(M)	64.8	**52.4**	**38.3**	**22.1**	21.1	**50.7**	**61.1**	**70.6**	**6.89**	**25.1**

Table 1: **Main result using three ranking task setups.** The best model in each group is shown in **bold**. High Precision and Recall scores, and low Rank scores, are better.

anism to make knowledge incorporation effective. To avoid missing undetected objects, we also add the entire image as a proposal. Our vocabulary for slogan, knowledge and statements consists of words that appeared more than 5 times in human-annotated statements or more than 20 times in OCR slogans or DBpedia comments. $\mathbf{v}_i$, $\mathbf{t}_i^{(0)}$, $\mathbf{t}_i^{(1)}$, $\mathbf{k}_{i,j}$, $\mathbf{h}$, $\mathbf{s}$ are all 200-D vectors. We use RMSprop with learning rate 0.001, batch size 128, and η (in triplet loss) of 0.2.

4.1. Qualitative examples

Fig. 3 shows learned edge weights. The weights (width of arrow) from visual objects, slogans and external knowledge towards the global node (star) reveal their relative contributions. The model without masking does not utilize the external knowledge effectively: all knowledge pieces have extremely small weights thus are omitted from the visualization. This indicates that even though the external knowledge is available, the model still tends to process superficial word pattern matching. Instead, when the entity information (potential shortcut) is masked from the retrieved comments, along with other info randomly sampled and masked, the model learns semantics from and thus **better exploits useful knowledge.** These results also suggest the **need to evaluate knowledge selection explicitly as a side task**, as we do in Tab. 3, as models may solve the main answering task but use irrelevant external knowledge or simply suppress all external knowledge.

4.2. Main result: Effectiveness of masking

In Tab. 1, let V denote the visual proposals, T the textual slogan information, and K the knowledge comments from DBPedia. M_t, M_s, and M_k, denote the different masking strategies described in Sec. 3.6. Simply "M" (for mask) means we use all three of them. By comparing V,T and V,T+K in each task, we see that simply adding knowledge achieves very marginal gains because the benefit of knowledge gets drowned-out due to shortcuts. However,

Method	Chall -15	Samp -100	Samp -500	Relative to V,T+K
V,T	3.02	7.52	30.11	
V,T+K	3.02	7.49	29.96	
V,T+K(M_t,M_s)	2.97	7.05	27.66	+7.68%
V,T+K(M_t,M_k)	2.93	6.74	26.04	+13.08%
V,T+K(M_s,M_k)	3.00	7.43	29.64	+1.07%
OURS: V,T+K(M_t,M_s,M_k)	**2.91**	**6.56**	**25.14**	+16.09%
V,T+K(M_v,M_t,M_s,M_k)	3.01	7.21	28.61	+4.51%

Table 2: **Average Rank on the ranking tasks.** Relative improvement is based on Sampled-500. Lower scores are better. The best method is shown in **bold**.

our masking strategy OURS: V,T+K(M) improves results over V,T+K on all tasks and almost all metrics. Accuracy (P@1) provides limited information because it only measures the easy-to-predict cases and all models are doing equally well. However, with the ranking metric and on the more challenging Sampled-100 and Sampled-500 test sets, we see our masking strategy brings significant and consistent performance gains. **Further, masking in conjunction with applying external knowledge (last row in each group) achieves better results compared to *not* using knowledge (first row).** Our method allows better reasoning (through external knowledge) by mitigating the effect of shallow matches (through masking).

Tab. 2 shows an ablation using the average rank metric. The table includes results for all three tasks, and we use the evaluation on the most difficult Sampled-500 to describe our improvement. First, directly adding knowledge (V,T+K v.s. V,T) does not help. The +K leads to only 0.5% improvement which is negligible (29.96 v.s. 30.11). However, if we apply masking to mitigate the effects of shortcut learning, the performance is improved by a large margin. As we compare OURS: V,T+K(M_t,M_s,M_k) to V,T+K, the average rank is reduced from 29.96 to 25.14 (-4.82 average rank or **+16.09% relative improvement when we use our proposed masking**). Further, we verify that removing any of the masking mechanisms, resulting in V,T+K(M_t,M_s), V,T+K(M_t,M_k), and V,T+K(M_s,M_k), leads to inferior performance (27.66, 26.04, 29.64 v.s. 25.14). We conclude the **useful information of external knowledge can be fully unleashed if and only if shortcut learning can be suppressed.**

Relation to dropout and Singh [25]. We highlight that avoiding shortcut differs from random dropout. First, though strategies M_t and M_s are similar to the dropout layer dropping information randomly during training, they are applied to textual tokens instead of neurons. Second, our M_k removes the query keywords from the retrieved knowledge paragraphs at *both* training and testing time. In Fig. 2, if "Nike" is *not* masked out ("~~Nike~~"), the model will consider the explanations regarding "goddess" and "asteroid" (blue boxes in the bottom-right) to be helpful because the keyword "Nike" overshadows the extra information. Finally,

Methods	Accuracy (%)
V,T+K	25.2
V,T+K(M_t,M_s)	**54.4**
V,T+K(M_t,M_k)	53.0
V,T+K(M_s,M_k)	25.9
OURS: V,T+K(M_t,M_s,M_k)	52.6

Table 3: **Accuracy(%) on the knowledge selection task.**

we provide comparison to the random masking of **visual regions** (similar to [25]). In Tab. 2, we denote M_v as randomly dropping an image region. We observe that it hurts the overall performance (+4.51% improvement over V,T+K for M_v v.s. +16.09% improvement for OURS). We argue that applying masking on regions did not focus on the key of knowledge utilization, unlike OURS: V,T+K(M_t,M_s,M_k).

4.3. Side task: Analyzing the knowledge utilization

We use a side task to measure how accurately the model could select the useful knowledge pieces from the noisy candidate pool. We use the edge weights methods learned, with and without our masking strategy. For each image, we take the learned weights for DBpedia comments (Eq. 2) as a knowledge importance score, and select the one with highest score using $\text{argmax}_{i,j}\ \alpha_{i,j}$. Then the model-selected knowledge is compared against human annotations, for an accuracy score. The procedure is **integral to the main task because the weights are learned automatically in it**. Note that methods did *not* receive supervision for this task at training time; instead, our masking strategy helps our method accomplish the task better than the baseline can. To the best of our knowledge, similar experiments have not been done in prior visual reasoning work. In knowledge-based VQA datasets, all provided knowledge pieces are relevant, but in our setting, the retrieved DBpedia knowledge pieces are usually noisy. Such noisy retrieval is more likely to happen in real-world applications.

The results are in Tab. 3. OURS: V,T+K(M_t,M_s,M_k) improves accuracy to 52.6% (+109% improvement!), and V,T+K(M_t,M_s) improves it to 54.4% compared to 25.2% for V,T+K (+115% improvement). **Masking doubles the ability of our method to retrieve appropriate knowledge,** by removing reliance on shortcuts. Further, this result **quantitatively shows the impact of shortcuts effects through the discrepancy of the main and side metrics** (16% gain in Tab. 2 compared to 109% in Tab. 3).

4.4. Comparison with the state-of-the-art

We compare our model to the approaches in the "Automatic Understanding of Visual Advertisements" challenge and other recent works. VSE trained by [52] uses only the image-level feature to represent the ad and triplet loss to optimize the model. ADNET [11] is similar but uses ResNet as the network backbone. ADVISE [52] aggregates proposal feature vectors to get the image representation, and incorporates knowledge from a pre-trained dense captioning model [19] and a symbol classifier. CYBERAGENT [33] is the first model that uses slogan texts embedded in the image. RHETORIC [53] is a hybrid model of both ADVISE and CYBERAGENT; it uses pointwise addition to integrate image and slogan, and is the current state-of-the-art.

Methods	Accuracy (%)
VSE [52]	62.0
ADNET [11]	65.0
ADVISE [52]	69.0
CYBERAGENT [33]	82.0
RHETORIC [53]	83.3
OURS	**87.3**

Table 4: **Accuracy(%) on the 2018 Ads challenge**. We compared our method to state-of-the-art models.

Tab. 4 shows the comparison to these approaches. Our model outperforms even the strongest baseline RHETORIC by 4.8% in terms of accuracy (87.3% v.s. 83.3%). While RHETORIC also incorporates both image and slogan information, our method represents this information in a more fine-grained manner using the graph. Besides, our method uses external knowledge from DBPedia.

We did not adapt and compare to more general VQA methods for two reasons: (1) our goal is to make external knowledge utilization effective for a simple method, not achieve state-of-the-art, and (2) many VQA methods are not applicable in our setting. The ads understanding requires in-depth understanding of the visual and the embedded textual features, as well as background information. Thus, general vision-language methods (e.g. VSE, ADNET) may perform poorly. Note however that our basic technique (Sec. 3.3) does incorporate ideas from well-known VQA work. Our method is an advanced version of bottom-up attention [3] in that the proposed graph also captures attention among knowledge, visual regions, and slogans. Like BERT-based [56, 27, 6] methods, we use attention to select relevant relationships. Diagnosing knowledge utilization in the BERT architecture is complex (e.g. due to many layers) so we use a simpler message passing structure and focus on effective knowledge usage.

5. Conclusion

Visual reasoning has attracted much attention, although the "reasoning" process is usually hidden behind a mixed or decoupled evaluation protocol. We proposed an effective method to incorporate external knowledge, and evaluated the gap between answering questions well and using the correct external knowledge and thus, the correct reasoning. Our masking strategy improved knowledge utilization on a challenging ads understanding task. Next we will *learn* how to mask and apply the strategy on additional datasets.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant No. 1718262. We thank the reviewers and AC for their insightful feedback.

References

[1] Somak Aditya, Yezhou Yang, and Chitta Baral. Explicit reasoning over end-to-end neural architectures for visual question answering. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 629–637, 2018.

[2] Karuna Ahuja, Karan Sikka, Anirban Roy, and Ajay Divakaran. Understanding visual ads by aligning symbols and objects using co-attention. In *CVPR Workshop towards Automatic Understanding of Visual Advertisements*, 2018.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[7] Google. Google cloud vision api. `https://cloud.google.com/vision/`.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, 2016.

[11] Murhaf Hossari, Soumyabrata Dev, Matthew Nicholson, Killian McCabe, Atul Nautiyal, Clare Conran, Jian Tang, Wei Xu, and François Pitié. Adnet: A deep network for detecting adverts. In *26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, 2018.

[12] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[16] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[17] Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[18] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[23] Adriana Kovashka and James Hahn. Automatic understanding of visual advertisements, June 2018. `https://evalai.cloudcv.org/web/challenges/challenge-page/86/overview`.

[24] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[25] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[26] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer,

et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2015.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[28] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[32] Medhini Narasimhan and Alexander G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *The European Conference on Computer Vision (ECCV)*, 2018.

[33] Mayu Otani, Yuki Iwazaki, and Kota Yamaguchi. Unreasonable effectiveness of ocr in visual advertisement understanding. In *CVPR Workshop towards Automatic Understanding of Visual Advertisements*, 2018.

[34] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[35] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[36] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[38] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G. Schwing. Factor graph attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[40] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[41] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[44] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *International Conference on Machine Learning (ICML)*, 2019.

[45] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[46] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[47] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 2018.

[48] Yicheng Wang and Mohit Bansal. Robust machine comprehension models via adversarial training. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018.

[49] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision (ECCV)*, 2016.

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.

[51] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[52] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *The European Conference on Computer Vision (ECCV)*, 2018.

[53] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Inferring

the messages of visual advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019.

[54] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[55] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[56] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[57] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *British Machine Vision Conference (BMVC)*, 2018.

[58] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.