Guiding Visual Question Answering with Attention Priors

Thao Minh Le, Vuong Le, Sunil Gupta, Svetha Venkatesh, Truyen Tran
Applied Artificial Intelligence Institute, Deakin University, Australia
{thao.le,vuong.le,sunil.gupta,svetha.venkatesh,truyen.tran}@deakin.edu.au

Abstract

The current success of modern visual reasoning systems is arguably attributed to cross-modality attention mechanisms. However, in deliberative reasoning such as in VQA, attention is unconstrained at each step, and thus may serve as a statistical pooling mechanism rather than a semantic operation intended to select information relevant to inference. This is because at training time, attention is only guided by a very sparse signal (i.e. the answer label) at the end of the inference chain. This causes the cross-modality attention weights to deviate from the desired visual-language bindings. To rectify this deviation, we propose to guide the attention mechanism using explicit linguistic-visual grounding. This grounding is derived by connecting structured linguistic concepts in the query to their referents among the visual objects. Here we learn the grounding from the pairing of questions and images alone, without the need for answer annotation or external grounding supervision. This grounding guides the attention mechanism inside VQA models through a duality of mechanisms: pre-training attention weight calculation and directly guiding the weights at inference time on a case-by-case basis. The resultant algorithm is capable of probing attention-based reasoning models, injecting relevant associative knowledge, and regulating the core reasoning process. This scalable enhancement improves the performance of VQA models, fortifies their robustness to limited access to supervised data, and increases interpretability.

1. Introduction

Visual reasoning is the new frontier of AI wherein facts extracted from visual data are gathered and distilled into higher-level knowledge in response to a query. Successful visual reasoning methodology estimates the cross-domain association between the symbolic concepts and visual entities in the form of attention weights. Such associations shape the knowledge distillation process, resulting in a unified representation that can be decoded into an answer. In the exemplar reasoning setting known as Visual Question Answering (VQA), attention plays a pivotal role in modern

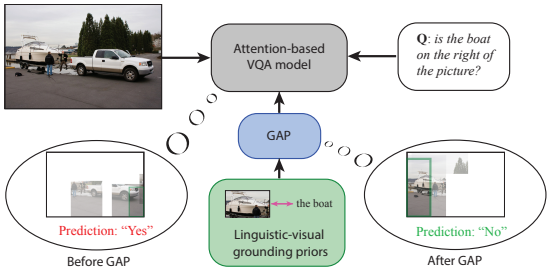

Figure 1. We introduce Grounding-based Attention Prior mechanism (blue box) which considers the linguistic-visual associations between a query-image pair and refines the attentions inside VQA models (gray box). This boosts the models' performance, reduces their reliance on supervised data and increases their interpretability.

systems [3, 15, 20, 23, 31]. Ideal attention scores must be both *relevant* and *effective*: Relevance implies that attention is high when the visual entity and linguistic entity refer to the same concept; Effectiveness implies that the attention derived leads to good VQA performance.

However, in typical systems, the attention scores are computed on-the-fly: unregulated at inference time and guided at training time by the gradient from the groundtruth answers. Analysis of several VQA attention models shows that these attention scores are usually neither relevant nor guaranteed to be effective [6]. The problem is even more severe when we cannot afford to have enough labeled answers due to the cost of the human annotation process. A promising solution is providing pre-computed guidance to direct and hint the attention mechanisms inside the VQA models towards more appropriate scores. Early works use human attention as the label for supervising machine attention [36, 40]. This simple and direct attention perceived by humans is not guaranteed to be optimal for machine reasoning [7, 8]. Furthermore, because annotating attention is a complex labeling task, this process is inherently costly, inconsistent and unreliable [40]. Finally, these methods only regulate the attention scores in training stage without directly adjust them in inference. Different from these approaches, we leverage the fact that such external guidance is pre-existing in the query-image pairs and can be extracted without any additional labels. *Using*

pre-computed language-visual associations as an inductive bias for attention-based reasoning without further extra labeling remains a desired but missing capability.

Exploring this underlying linguistic-visual association for VQA, we aim to distill the compatibility between entities across input modalities in an unsupervised manner from the query-image pairs without explicit alignment grouthtruths, and use this knowledge as an inductive bias for the attention mechanism thus boosting reasoning capability. To this end, we design a framework called *Grounding-based Attention Prior* (GAP) to (1) extract the alignments between linguistic-visual region pairs and (2) use these pair-wise associations as an inductive bias to guide VQA's attention mechanisms.

For the first task, we exploit the pairing between the questions and the images as a weakly supervised signal to learn the mapping between words and image regions. By exploiting the implicit supervising signals from the pairing, this requires no further annotation. To overcome the challenge of disparity in the co-inferred semantics between query words and image regions, we construct a parse tree of the query, extract the nested phrasal expressions and ground them to image regions. These expressions semantically match image regions better than single words and thus create a set of more reliable linguistic-visual alignments.

The second task aims at using these newly discovered alignments to guide reasoning attention. This guidance process is provided through two complementary pathways. First, we pre-train attention weights to align with the pre-computed grounding. This step is done in an unsupervised manner without access to the answer groundtruths. Second, we use the attention prior to directly regulate and refine the attention weights guided by the groundtruth answer through back-propagation to not deviate too far away from it. This is modulated by a learnable gate. These dual guidance pathways are a major advancement from previous attention regularization methods [40, 49] as the linguistic-visual compatibility is leveraged directly and flexibly in both training and inference rather than simply as just regularization.

Through extensive experiments, we prove that this methodology is effective in both discovering the grounding and using them to boost the performance of attention-based VQA models across representative methods and datasets. These improvements surpass other methods' performance and furthermore require no extra annotation. The proposed method also significantly improves the sample efficiency of VQA models, hence less annotated answers are required. Fig. 1 illustrates the intuition and design of the method with an example of the improved attention and answer.

Our key contributions are:

1. A novel framework to calculate linguistic-visual alignments, providing pre-computed attention priors to guide attention-based VQA models;

2. A generic technique to incorporate attention priors into most common visual reasoning methods, fortifying them in performance and significantly reducing their reliance on human supervision; and,

3. Rigorous experiments and analysis on the relevance of linguistic-visual alignments to reasoning attention.

2. Related Work

Attention-based models are the most prominent approaches in VQA. Simple methods [3] only used single-hop attention mechanism to help machine select relevant image features. More advanced methods [52, 15, 23] and those relying on memory networks [50, 51] used multi-hop attention mechanisms to repeatedly revise the selection of relevant visual information. BAN [20] learned a co-attention map using expensive bilinear networks to represent the interactions between pairs of word-region. One drawback of these attention models is that they are only supervised by the answer groundtruth without explicit attention supervision.

Attention supervision is recently studied for several problems such as machine translation [27] and image captioning [26, 32, 54]. In VQA, attentions can be self-regulated through internal constraints [37, 28]. More successful regularization methods use external knowledge such as human annotations on textual explanations [49] or visual attention [36, 40]. Unlike these, we propose to supervise VQA attentions using pre-computed language-visual grounding from image-query pairs without using external annotation.

Linguistic-visual alignment includes the tasks of text-image matching [24], grounding referring expressions [53] and cross-domain joint representation [30, 42]. These groundings can support tasks such as captioning [54, 18]. Although most tasks are supervised by human annotations, contrastive learning [11, 47] allows machines to learn the associations between words and image regions from weak supervision of phrase-image pairs. In this work, we propose to explore such associations between query and image in VQA. This is a new challenge because the query is complex and harder to be grounded, therefore new method using grammatical structure will be devised.

Our work also share the **Knowledge distillation** paradigm [13] with cross-task [2] and cross modality [10, 29, 48] adaptations. Particularly, we distill visual-linguistic grounding and use it as an input for VQA model's attention. This also distinguishes our work from the recent **self-supervised pretraining methods** [43, 25] where they focus on a unified representation for a wide variety of tasks thanks to the access to enormous amount of data. Our work is theoritically applicable to complement the multimodal matching inside these models.

3. Preliminaries

A VQA system aims to deduce an answer y about an image $\mathcal{I}$ in response to a linguistic question q, $e.g.$, via

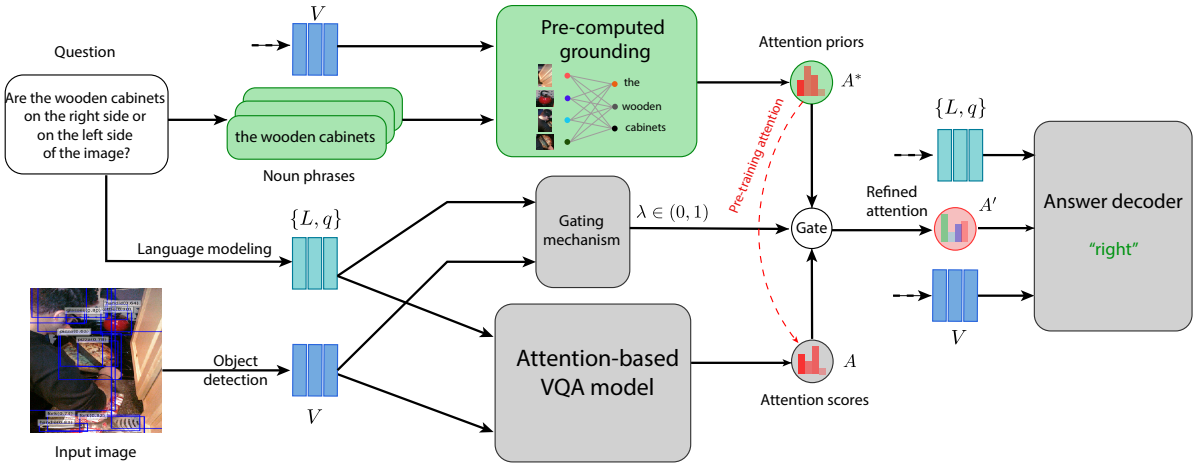

Figure 2. Overall architecture of a generic joint attention VQA model using Grounding-based Attention Prior (GAP) to guide the computation of attention weights. Vision-language compatibility pre-computed by an unsupervised framework (green boxes) serves as an extra source of information, providing inductive biases to guide attention weights inside attention-based VQA models towards more meaningful alignment.

$P(y \mid q, \mathcal{I})$. The query q is typically decomposed into a set of T linguistic entities $L = \{l_i\}_{i=1}^{T}$. These entities and the query q are then embedded into a feature vector space: $q \in \mathbb{R}^d$, $l_i \in \mathbb{R}^d$. In the case of sequential embedding popularly used for VQA, entities are query words; they are encoded with GloVe for word-level embedding [35] followed by RNNs such as BiLSTM for sentence-level embedding. Likewise the image $\mathcal{I}$ is often segmented into a set of N visual regions with features $V = \{v_j \mid v_j \in \mathbb{R}^d\}_{j=1}^{N}$ by an object detector, i.e., Faster R-CNN [39]. For ease of reading, we use the dimension d for both linguistic embedding vectors and visual representation vectors.

A large family of VQA systems [31, 3, 15, 23, 20, 21] rely on attention mechanisms to distribute conditional computations on linguistic entities L and visual counterparts V. These models can be broadly classified into two groups: *joint-* and *marginalized- attention models*. [31, 3, 15] are among those who fall into the former, while [20, 21] and transformer-based models [43] are typical representative of the works in the latter category.

Joint attention models The most complete attention model includes a detailed pair-wise attention map indicating the contextualized correlation between word-region pairs used to estimate the interaction between visual and linguistic entities for the combined information. These attention weights are in the form of a 2D matrix $A \in \mathbb{R}^{T \times N}$. They often contain fine-grained relationships between each linguistic word to each visual region. The attention matrix A is derived by a sub-network $B_\theta(.)$ as $A_{ij} = B_\theta(e_{ij} \mid V, L)$, where each e_{ij} denotes the correlation between the linguistic entities l_i and the visual region v_j, and θ is network parameters of VQA models. Joint attention models contain the rich pairwise relation and often perform well. However, calculating and using this full matrix has a large overhead computation cost. A good approximation of this matrix is the marginalized vectors over rows and columns which is described next.

Marginalized attention models Conceptually, the matrix A is marginalized along columns into the *linguistic attention vector* $\alpha = \{\alpha_i\}_{i=1}^{T} \in \mathbb{R}^T$ and along rows into *visual attention vector* $\beta = \{\beta_j\}_{j=1}^{N}, \beta \in \mathbb{R}^N$. In practice, α and β are calculated directly from each pair of input image and query through dedicated attention modules. They can be implemented in different ways such as direct single-shot attention [3], co-attention [31] or multi-step attention [15]. In our experiment, we concentrate on two popular mechanisms: *single-shot attention* where the visual attention β is calculated directly from the inputs (V, q) and *alternating attention mechanism* where the visual attention β follows the linguistic attention α [31]. Concretely, α is estimated first, followed by the attended linguistic feature of the entire query $c = \sum_{i=1}^{T} \alpha_i * l_i$; then this attended linguistic feature is used to calculate the visual attention β. The alternating mechanism can be extended with multi-step reasoning [15, 23, 14]. In such case, a pair of attentions $\alpha_{i,k}$ and $\beta_{j,k}$ are estimated at each reasoning step k forming a series of them.

Answer decoder Attention scores drive the reasoning process producing a joint linguistic-visual representation on which the answer is decoded: $P\left(a \mid f(L, V, \text{att\_scores})\right)$ ("att_scores" refers to either visual attention vector β or attention matrix A). For marginalized attention models, the function $f(.)$ is a neural network taking as input the query representation q and the attended visual feature $\hat{v} = \sum_j \beta_j * v_j$ to return a joint representation. Joint attention models instead use the bilinear combination to calculate each component of the output vector of f [20]:

$$f_t \equiv (L^\top W_L)_t^\top A (V^\top W_V)_t, \quad (1)$$

where t is the index of output components and $W_L \in \mathbb{R}^{d\times d}$ and $W_V \in \mathbb{R}^{d\times d}$ are learnable weights.

4. Methods

We now present Grounding-based Attention Priors (GAP), an approach to extract the concept-level association between query and image and use this knowledge as attention priors to guide and refine the cross-modality attentions inside VQA systems. The approach consists of two main stages. First, we learn to estimate the linguistic-visual alignments directly from question-image pairs (Sec. 4.1, green boxes in Fig. 2). Second, we use such knowledge as inductive priors to assist the computation of attention in VQA (Sec. 4.2, Sec. 4.3, and lower parts in Fig. 2).

4.1. Structures for Linguistic-Visual Alignment

Grammatical structures for grounding. The task of *Linguistic-visual Alignment* aims to find the groundings between the linguistic entities (*e.g.*, query words $L = \{l_i\}_{i=1}^T$ in VQA) and vision entities (*e.g.*, visual regions $V = \{v_j\}_{j=1}^N$ in VQA) in a shared context. This requires the interpretation of individual words in the complex context of the query so that they can co-refer to the same concepts as image regions. However, compositional queries have complex structures that prevent state-of-the-art language representation methods from fully understanding the relations between semantic concepts in the queries [38]. We propose to better contextualize query words by *breaking a full query into phrases that refer to simpler structures,* making the computation of word-region grounding more effective. These phrases are called *referring expressions* (RE) [33] and were shown to co-refer well to image regions [19]. The VQA image-query pairing labels are passed to the REs of such query. We then ground words with contextualized embeddings within each RE to their corresponding visual regions. As the REs are nested phrases from the query, a word can appear in multiple REs. Thus, we obtain the query-wide *word-region grounding* by aggregating the grounding of REs containing the word. See Fig. 3 for an example on this process.

We extract query REs using a constituency parse tree $\mathcal{T}$[5].[1] In this structure, the query is represented as a set of nested phrases corresponding to subtrees of $\mathcal{T}$. The parser also provides the grammatical roles of the phrases. For example, the phrase "the white car" will be tagged as a *noun-phrase* while "standing next to the white car" is a *verb-phrase*. As visual objects and regions are naturally associated with noun-phrases, we select a set $E = \{E_r\}$ of all the noun phrases and wh-noun phrases[2] as the REs.

[1] Berkeley Neural Parser [22] in our implementation.

[2] noun phrases prefixed by a pronoun, *e.g.*, "which side", "whose bag".

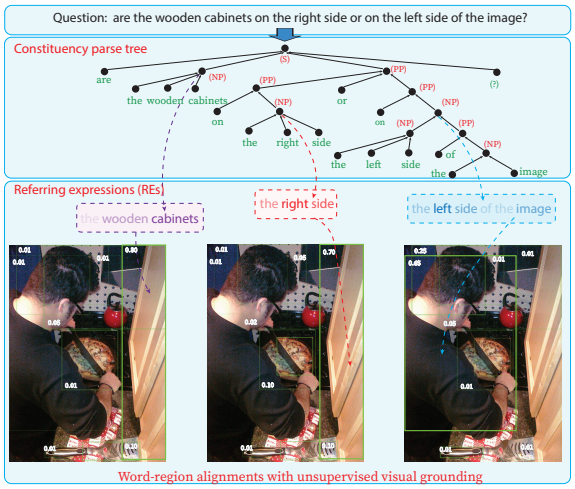

Figure 3. The query is parsed into a constituency parse tree to identify REs. Each RE serves as a local context for words. Words within each RE context are grounded to corresponding image regions. A word can appear in multiple REs, and thus its final grounding is averaged over containing REs, serving as inductive prior for VQA.

We denote the r^{th} RE as $E_r = \{w_i \,|\, w_i \in \mathbb{R}^d\}_{s_r \le i \le e_r}$ where s_r and e_r are the start and end word index of the RE within the query $L = \{l_i\}_{i=1}^T$. It has length $m_r = e_r - s_r + 1$. We now estimate the correlation between words in these REs and the visual regions $V = \{v_j \mid v_j \in \mathbb{R}^d\}_{j=1}^N$ by learning the neural association function $g_\delta(V, E_r)$ of parameter δ that generates a mapping $A_r^* \in \mathbb{R}^{m_r \times N}$ between words in the RE and the corresponding visual regions.

We implement $g_\delta(.)$ as the dot products of a contextualized embedding of word w_i in E_r with image regions in V, following the scaled dot-product attention [46].

Unsupervised training. To train the function $g_\delta(.)$, we adapt the recent contrastive learning framework [11] for phrase grounding to learn these word-region alignments from the RE-image pairs in an unsupervised manner, i.e. without explicit word-region annotations. In a mini batch $\mathcal{B}$ of size b, we calculate the positive mapping $A_r^* = \left(a_{r,i,j}^*\right) \in \mathbb{R}^{m_r \times N}$ on one positive sample (the RE E_r and the image regions V in the image that is paired with it) and $(b-1)$ negative mappings $\overline{A_{r,s}^*} = \left(\overline{a_{r,s,i,j}^*}\right) \in \mathbb{R}^{m_r \times N}$ where $1 \le s \le (b-1)$ from negative samples (the RE E_r and negative image regions $V_s' = \{v_{s,j}'\}$ from images that are not paired with it). We then compute linguistic-induced visual representations $v_i^* \in \mathbb{R}^d$ and $\overline{v_{s,i}^*} \in \mathbb{R}^d$ over regions for each word w_i:

$$v_i^* = \textstyle\sum_{v_j \in V} \text{norm}_j \left(a_{r,i,j}^*\right) W_v^\top v_j, \quad (2)$$

$$\overline{v_{s,i}^*} = \textstyle\sum_{v'_{s,j} \in V'_s} \text{norm}_j \left(\overline{a_{r,s,i,j}^*}\right) W_{v'}^\top v'_{s,j}, \quad (3)$$

where "norm$_j$" is a column normalization operator; $W_v \in \mathbb{R}^{d\times d}$ and $W_{v'} \in \mathbb{R}^{d\times d}$ are learnable parameters. We then push them away from each other by maximizing the linguistic-vision InfoNCE [34]:

$$\mathcal{L}_r(\delta)=\mathbb{E}_{\mathcal{B}}\left[\sum_{w_i\in E_r}\log\left(\frac{e^{\langle W_w^\top w_i, v_i^*\rangle}}{e^{\langle W_w^\top w_i, v_i^*\rangle}+\sum_{s=1}^{b-1}e^{\langle W_w^\top w_i, \overline{v_{s,i}^*}\rangle}}\right)\right]. \tag{4}$$

This loss maximizes the lower bound of mutual information $\text{MI}(V, w_i)$ between visual regions V and contextualized word embedding w_i [11].

Finally, we compute the word-region alignment $A^* \in \mathbb{R}^{T\times N}$ by aggregating the RE-image groundings:

$$A^* = \frac{1}{|E|}\sum_{r=1}^{|E|}\tilde{A}_r^*, \tag{5}$$

where $\tilde{A}_r^* \in \mathbb{R}^{T\times N}$ is the zero-padded matrix of A_r^*.

Besides making grounding more expressive, this divide-and-conquer strategy has extra benefits of augmenting the weak supervising labels from query-image to RE-image pairs, which provide more supervising signals (positive pairs) hence, better training of the contrastive learning framework.

The discovered grounding provides a valuable source of priors for VQA attention. Existing works [36, 40] use attention priors to regulate the gradient flow of VQA models during training, hence only constraining the attention weights indirectly. Unlike these methods, we directly guide the computation of attention weights via two pathways: through pre-training them without answers, and by refining in VQA inference on a case-by-case basis.

4.2. Pre-training VQA Attention

A typical VQA system seeks to ground linguistic concepts parsed from the question to the associated visual parts through cross-modal attention. However, this attention mechanism is guided only indirectly and distantly through sparse training signal of the answers. This training signal is too weak to assure that relevant associations can be discovered. To directly train the attention weights to reflect these natural associations, we pre-train VQA models by enforcing the attention weights to be close to the alignment maps A^* discovered through unsupervised grounding in Sec. 4.1.

For joint attention VQA models, this is achieved through minimizing the Kullback-Leibler divergence between vectorized forms of the VQA visual attention weights A and the prior grounding scores A^*:

$$\mathcal{L}_{\text{pre-train}} = KL(\text{norm}\circ\text{vec}(A^*) \parallel \text{norm}\circ\text{vec}(A)), \tag{6}$$

where norm$\circ$vec flattens a matrix into a vector followed by a normalization operator ensuring such vector sums to one.

For marginalized attention models, we first marginalize $A^* = \left(a_{i,j}^*\right)$ into a vector of visual attention prior:

$$\beta^* = \frac{1}{T}\sum_{i=1}^{T}\text{norm}_j(a_{i,j}^*). \tag{7}$$

The pre-training loss is the KL divergence between the attention weights and their priors:

$$\mathcal{L}_{\text{pre-train}} = KL(\beta^* \parallel \beta). \tag{8}$$

4.3. Attention Refinement with Attention Priors

4.3.1 Marginalized attention refinement

Recall from Sec. 3 that a marginalized attention VQA model computes linguistic attention over T query words $\alpha \in \mathbb{R}^T$ and visual attention over N visual regions $\beta \in \mathbb{R}^N$. In this section, we propose to directly refine these attentions using attention priors $A^* = \left(a_{i,j}^*\right) \in \mathbb{R}^{T\times N}$ learned in Sec. 4.1. First, A^* is marginalized over rows and columns to obtain a pair of attention priors vectors $\alpha^* \in \mathbb{R}^T$ and $\beta^* \in \mathbb{R}^N$:

$$\alpha^* = \frac{1}{N}\sum_{j=1}^{N}\text{norm}_i(a_{i,j}^*);\ \beta^* = \frac{1}{T}\sum_{i=1}^{T}\text{norm}_j(a_{i,j}^*). \tag{9}$$

We then refine α and β inside the reasoning process through a gating mechanism to return refined attention weights α' and β' in two forms:

Additive form:

$$\alpha' = \lambda\alpha + (1-\lambda)\,\alpha^*, \quad \beta' = \gamma\beta + (1-\gamma)\,\beta^*, \tag{10}$$

Multiplicative form:

$$\alpha' = \text{norm}\Big((\alpha)^{\lambda}(\alpha^*)^{(1-\lambda)}\Big), \beta' = \text{norm}\Big((\beta)^{\gamma}(\beta^*)^{(1-\gamma)}\Big), \tag{11}$$

where "norm" is a normalization operator; $\lambda \in (0,1)$ and $\gamma \in (0,1)$ are outputs of learnable gating functions that decide how much attention priors contribute per words and regions. Intuitively, these gating mechanisms are a solution to maximizing the agreement between two sources of information: $\alpha' = \text{argmin}\left(\lambda * D(\alpha', \alpha) + (1-\lambda) * D(\alpha', \alpha^*)\right)$, where $D(P_1, P_2)$ measures the distance between two probability distributions P_1 and P_2. When $D \equiv$ Euclidean distance, it gives Eq. (10) and when $D \equiv$ KL divergence between the two distributions, it is Eq. (11) [12] (See Supp. for proofs). The same intuition applies for the calculation of β'.

The learnable gates for λ and γ are implemented as a neural function $h_\theta\left(.\right)$ of visual regions $\bar{v}$ and the question q:

$$\lambda = h_\theta\left(\bar{v}, q\right). \tag{12}$$

For simplicity, $\bar{v}$ is the arithmetic mean of regions in V.

For multi-step reasoning, we apply Eqs. (10, 11) step-by-step. As each reasoning step k is driven by an intermediate control c_k (Sec. 3), it affects the learning of the gate by:

$$\lambda_k = p_\theta\left(c_k, h_\theta\left(\bar{v}, q\right)\right). \tag{13}$$

4.3.2 Joint attention refinement

In joint attention VQA models, we can directly use matrix $A^* = \left(a_{ij}^*\right) \in \mathbb{R}^{T\times N}$ without marginalization. With slight abuse of notation, we denote the output the modulating gate for attention refinement as $\lambda \in (0,1)$ sharing similar role with the gating mechanism in Eq. (12):

$$A' = \begin{cases} \lambda A + (1-\lambda)\,B^* & \text{(add.)} \\ \text{norm}\left((A)^{\lambda}\left(B^*\right)^{(1-\lambda)}\right) & \text{(multi.)} \end{cases} \tag{14}$$

where $B^* = \text{norm}_{ij}\left(a_{ij}^*\right)$.

Method	VQA v2 standard val↑			
	All	Yes/No	Num	Other
UpDn+Attn. Align [40]	63.2	81.0	42.6	55.2
UpDn+AdvReg [37]	62.7	79.8	42.3	55.2
UpDn+SCR (w. ext.) [49]	62.2	78.8	41.6	54.5
UpDn+SCR (w/o ext.) [49]	62.3	77.4	40.9	56.5
UpDn+DLR [17]	58.0	76.8	39.3	48.5
UpDn+RUBi† [4]	62.7	79.2	42.8	55.5
UpDn+HINT [40]	63.4	**81.2**	43.0	55.5
UpDn+GAP	**64.3**	**81.2**	**44.1**	**56.9**

Table 1. Comparison between GAP and other attention regularization methods using UpDn on VQA v2. Results of other methods are taken from their respective papers. †Our reproduced results.

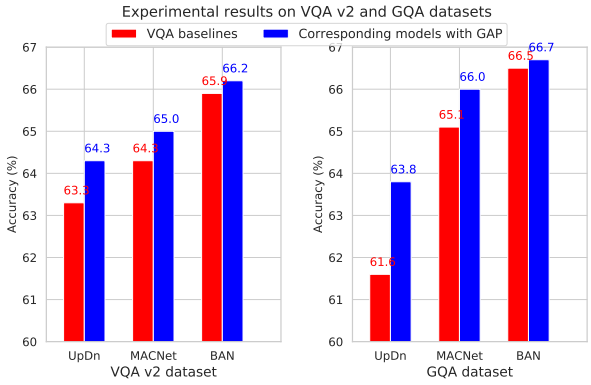

Figure 4. GAP's universality across different baselines and datasets.

4.4. Two-stage Model Training

We perform a two-step pre-training/fine-tuning procedure to train models using the attention priors: (1) unsupervised pre-training VQA without answer decoder with attention priors (Sec. 4.2), and (2) fine-tune full VQA models with attention refinement using answers, i.e. by minimizing the VQA loss $-\log P(y \mid q, \mathcal{I})$.

5. Experiments

We evaluate our approach GAP on two representative marginalized VQA models: Bottom-Up Top-Down Attention (UpDn) [3] for single-shot, MACNet [15] for multi-step compositional attention models; and a joint attention model of BAN [20]. Experiments are on two datasets: VQA v2 [9] and GQA [16]. Unless stated otherwise, we we additive gating (Eq. (10)) for experiments with UpDn and MACNet, and multiplicative forms (Eq. (11)) for BAN. Implementation details and extra results are available in the Supplement.

5.1. Experimental Results

Enhancing VQA performance We compare GAP against the VQA models based on the UpDn baseline that utilize external priors and human annotation on VQA v2. Some of these methods use internal regularization: *adversarial regularization* (AdvReg) [37], *attention alignment* (Attn. Align) [40]; and some use human attention as external supervision: *self-critical reasoning* (SCR) [49] and *HINT* [40]. While these methods mainly aim at designing regularization schemes to exploit the underlying data generation process of VQA-CP datasets [1] where it deliberately builds the train and test splits with different answer distributions. This potentially leads to overfitting to the particular test splits and accuracy gains do not correlate to the improvements of actual grounding [41]. On the contrary, GAP does not rely on those regularization schemes but aims at directly improving the learning of attention inside VQA models to facilitate reasoning. In other words, GAP complements the effects of the aforementioned methods on VQA-CP (See Supplement).

Table 1 shows that our approach *(UpDn+GAP)* clearly has advantages over others in improving the UpDn baseline. The favorable performance is consistent across all question types, especially on "Other" question type, which is the most important and challenging for open-ended answers [44, 45].

Compared to methods using external attention annotations (*UpDn+SCR*, *UpDn+HINT*), the results suggest that GAP is effective in using attention priors (both learning and inference), especially when our priors are extracted in an unsupervised manner without the need for human annotation.

Universality across VQA models GAP is theoretically applicable to any attention-based VQA models. We evaluate the universality of GAP by trialing it on a wider range of baseline models and datasets. Figure 4 summarizes the effects of GAP on UpDn, MACNet and BAN on the large-scale datasets VQA v2 and GQA.

It is clear that GAP consistently improves upon all baselines over all datasets. GAP is beneficial not only for the simple model UpDn, but also for the multi-step model (MACNet). We observe the best effects when applied at early reasoning steps where attention weights are yet to converge.

Between datasets, the improvement is stronger on GQA than on VQA v2, which is explained by the fact that GQA has a large portion of compositional questions which our unsupervised grounding learning can benefit from.

The improvements are less significant with BAN which already has large capacity model at the cost of data hunger and computational expensiveness. In the next section, we show that GAP significantly reduces the amount of supervision needed for these models compared to the baseline.

Sample efficient generalization We examine the generalization of the baselines and our proposed methods when analyzing sample efficiency with respect to the number of annotated answers required. Fig. 5 shows the performance of the chosen baselines on the validation sets of VQA v2 (left column) and GQA dataset (right column) when given

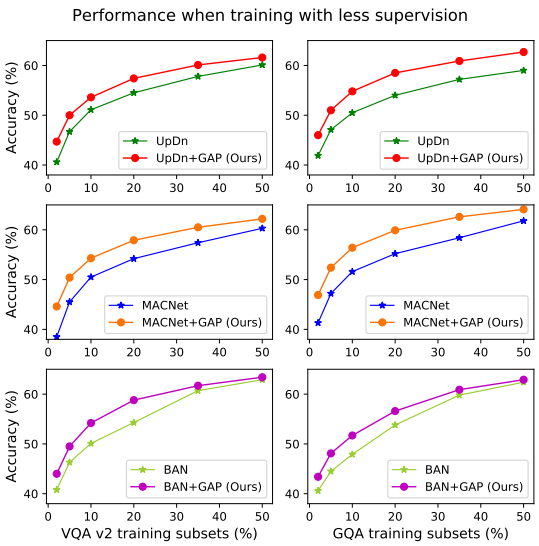

Figure 5. GAP improves generalization capability with limited access to grouthtruth answers.

different fractions of the training data. In particular, when reducing the number of training instances with groundtruth answers to under 50% of the training set, GAP considerably outperforms all the baseline models in accuracy across all datasets by large margins. For example, when given only 10% of the training data, GAP performs better than the strongest baseline BAN among the chosen ones by over 4.1 points on VQA v2 (54.2% vs. 50.1%) and nearly 4.0 points on GQA (51.7% vs. 47.9%). The benefits of GAP are even more significant for MACNet baseline which easily got off the track in the early steps without large data. The results strongly demonstrate the benefits of GAP in reducing the reliance on supervised data of VQA models.

5.2. Model Analysis

Performance of unsupervised phrase-image grounding To analyze the unsupervised grounding aspect of our model, (Sec. 4.1), we test the grounding model trained with VQA v2 on a mock test set from caption-image pairs on Flickr30K Entities. This out-of-distribution evaluation setting will show whether our unsupervised grounding framework can learn meaningful linguistic-visual alignments.

The performance of our new unsupervised linguistic-visual alignments using the query grammatical structure is shown in the top row of Table 2. This is compared against the alignment scores produced by the same framework but without breaking the query into REs (Middle row) and the random alignments (Bottom row). There is a 5 points gain from the random scores and over 1 point from the question-image pairs without phrases, indicating our linguistic-visual alignments is a reliable inductive prior for attention in VQA.

Model	R@1	R@5	R@10	Acc.
Unsup. RE-image grounding	14.1	35.6	45.5	45.4
Unsup. grounding w/o REs	12.0	33.0	42.9	44.3
Random alignment score (10 runs)	6.6	28.4	43.3	40.7

Table 2. Grounding performance of the unsupervised RE-image grounding when evaluated on out-of-distribution image-caption Flickr30K Entities test set. **R**ecall@k: fraction of phrases with bounding boxes that have IOU$\geq$0.5 with top-k predictions.

No.	Models	Acc.
1	UpDn baseline	63.3
2	+GAP w/ uniform-values vector	63.7
3	+GAP w/ random-values vector	63.6
4	+GAP w/ supervised grounding	64.0
5	+GAP w/ unsupervised visual grounding	64.3

Table 3. VQA performance on VQA v2 validation split with different sources of attention priors.

Models	Acc.
1. UpDn baseline, $\beta' \equiv \beta$ ($\gamma(\theta) \equiv 1.0$)	63.3
Attention as priors	
2. w/ $\beta' \equiv \beta^*(\gamma(\theta) \equiv 0.0)$	60.0
Effects of the direct use of attention priors	
3. +GAP w/o 1st stage fine-tuning	63.9
4. w/ 1st stage fine-tuning with attention priors	64.0
Effects of the gating mechanisms	
5. +GAP, fixed $\gamma(\theta) \equiv 0.5$	64.0
6. +GAP (multiplicative gating)	64.1
Effects of using visual-phrase associations	
7. +GAP (w/o extracted phrases from questions)	63.9
8. +GAP (full model)	64.3

Table 4. Ablation studies with UpDn on VQA v2.

Effectiveness of unsupervised linguistic-visual alignments for VQA We examine the effectiveness of our attention prior by comparing it with different ways of generating values for visual attention prior β^* on VQA performance. They include: (1) UpDn baseline (no use of attention prior) (2) uniform-values vector and (3) random-values vector (normalized normal distribution), (4) supervised grounding (pre-trained MAttNet [53] on RefCOCO [19]), and (5) GAP. Table 3 shows results on UpDn baseline. GAP is significantly better than the baseline and other attention priors (2-3-4). Especially our unsupervised grounding gives better VQA performance than the supervised one (Row 5). This surprising result suggests that pre-trained supervised model could not generalize out of distribution, and is worse than underlying grounding phrase-image pairs extracted unsupervisedly.

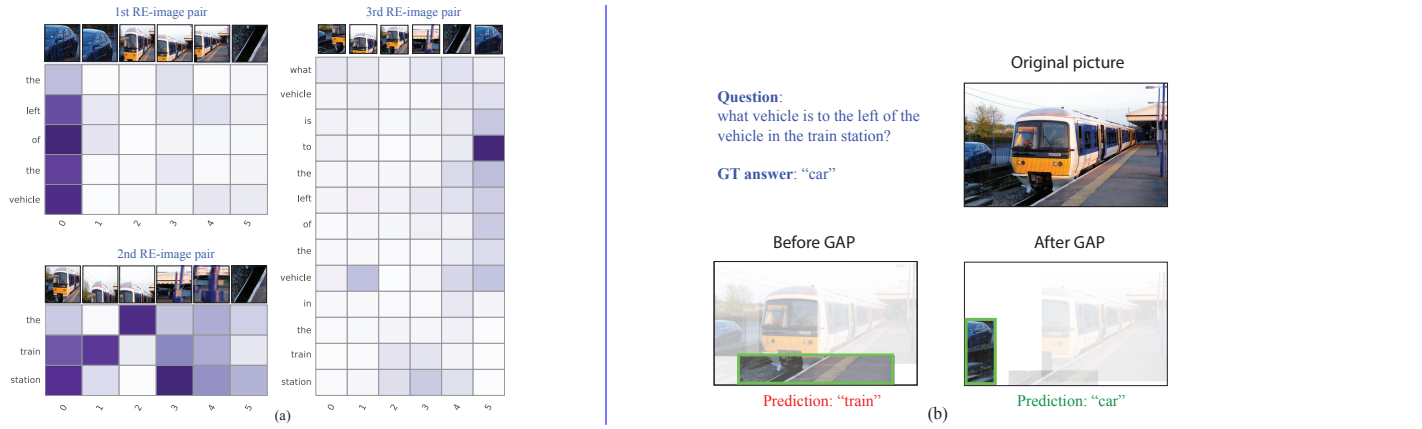

Figure 6. Qualitative analysis of GAP. (a) Region-word alignments of different RE-image pairs learned by our unsupervised grounding framework. (b) Visual attentions and prediction of UpDn model before (left) vs. after applying GAP (right). GAP shifts the model's highest visual attention (green rectangle) to more appropriate regions while the original puts attention on irrelevant parts.

Model	Top-1 attn.	Top-5 attn.	Top-10 attn.
UpDn baseline	14.50	27.31	35.35
UpDn + GAP	16.76	29.32	36.53

Table 5. Grounding scores for top-1, top-5 and top-10 attention of UpDn before and applying GAP on GQA validation split.

Ablation studies To provide more insights into our method, we conduct extensive ablation studies on the VQA v2 dataset (see Table 4). Throughout these experiments, we examine the role of each component toward the optimal performance of the full model. Experiments (1, 2) in Table 4 show that UpDn model does not perform well with either only its own attention or with the attention prior itself. This supports our intuition that they complement each other toward optimal reasoning. Rows 5,6 show that a soft combination of the two terms is necessary.

Row 7 justifies the use of structured grounding. It shows that phrase-image grounding gives better performance than question-image pairs only. In particular, the extracted RE-image pairs improves performance from 63.9% to 64.3%. This clearly demonstrates the significance of the grammatical structure of questions as an inductive bias for inter-modality matching which eventually benefits VQA.

Quantitative results We quantify the visual attentions of the UpDn model before and after applying GAP on the GQA validation set. In particular, we use the grounding score proposed by [16] to measure the correctness of the model's attentions weights comparing to the groundtruth grounding provided. Results are shown in Table 5. Our method improves the grounding scores of UpDn by 2.26 points (16.76 vs. 14.50) for top-1 attention, 2.01 points (29.32 vs. 27.31) for top-5 attention and 1.18 points (36.53 vs. 35.35) for top-10 attention. It is to note that while the grounding scores reported by [16] summing over all object regions, we report the grounding scores attributed by top-k attentions to better emphasize how the attentions shift towards most relevant objects. This analysis complements the VQA performance in Table 3 in a more definitive confirmation of the role of GAP in improving both reasoning attention and VQA accuracy.

Qualitative results We analyze the internal operation of GAP by visualizing grounding results on a sample taken from the GQA validation set. The quality of grounding is demonstrated in Fig. 6(a) with the word-region alignments found for several RE-image pairs. With GAP, these good grounding eventually benefits VQA models by guiding their visual attentions. Fig. 6(b) shows visual attention of the UpDn model before and after applying GAP. The guided attentions were shifted towards more appropriate visual regions than attentions by UpDn baseline.

6. Conclusion

We have presented a generic methodology to semantically enhance cross-modal attention in VQA. We extracted the linguistic-vision associations from query-image pairs and used it to guide VQA models' attention with Grounding-based Attention Prior (GAP). Through extensive experiments across large VQA benchmarks, we demonstrated the effectiveness of our approach in boosting attention-based VQA models' performance and mitigating their reliance on supervised data. We also showed qualitative analysis to prove the benefits of leveraging grounding-based attention priors in improving the interpretability and trustworthiness of attention-based VQA models. Broadly, the capability to obtain the associations between words and vision entities in the form of common knowledge is key towards systematic generalization in joint visual and language reasoning.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pages 4971–4980, 2018. 5.1

[2] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301, 2018. 2

[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 *CVPR*, pages 6077–6086, 2018. 1, 2, 3, 3, 5

[4] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019. 5.1

[5] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 4.1

[6] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 1

[7] Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020. 1

[8] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011. 1

[9] 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 *CVPR*, pages 6904–6913, 2017. 5

[10] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2

[11] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *Computer Vision–ECCV 2020*, pages 752–768. Springer, 2020. 2, 4.1, 4.1

[12] Tom Heskes. Selecting weighting factors in logarithmic opinion pools. *Advances in neural information processing systems*, pages 266–272, 1998. 4.3.1

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[14] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. *ICCV*, 2019. 3

[15] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *ICLR*, 2018. 1, 2, 3, 3, 5

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 5, 5.2

[17] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11181–11188, 2020. 5.1

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 4.1, 5.2

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 1, 2, 3, 3, 5

[21] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *International Conference on Learning Representations*, 2017. 3

[22] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, 2018. 1

[23] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Dynamic language binding in relational visual reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 818–824, 2020. 1, 2, 3, 3

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2

[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[26] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[27] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, Dec. 2016. 2

[28] Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, and Liqiang Nie. Answer questions with right

image regions: A visual attention regularization approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2021. 2

[29] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 700–708, 2018. 2

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019. 2

[31] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *CVPR*, 2016. 1, 3, 3

[32] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. Learning to generate grounded visual captions without localization supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 2. Springer, 2020. 2

[33] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 4.1

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4.1

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 3

[36] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *AAAI*, volume 32, 2018. 1, 2, 4.1

[37] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *NeurIPS*, 2018. 2, 5.1, 5.1

[38] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP*, pages 3982–3992, 2019. 4.1

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 3

[40] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2591–2600, 2019. 1, 2, 4.1, 5.1, 5.1

[41] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. pages 8172–8181, 2020. 5.1

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *International Conference on Learning Representations*, 2020. 2

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 2, 3

[44] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *CVPR*, pages 1417–1427, 2021. 5.1

[45] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in neural information processing systems*, 2020. 5.1

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 4.1

[47] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021. 2

[48] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021. 2

[49] Jialin Wu and Raymond J Mooney. Self-critical reasoning for robust visual question answering. *NeurIPS*, 2019. 1, 2, 5.1, 5.1

[50] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, pages 2397–2406, 2016. 2

[51] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, pages 451–466. Springer, 2016. 2

[52] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. 2

[53] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension, 2018. 2, 5.2

[54] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4777–4786, 2020. 2