Greedy Gradient Ensemble for Robust Visual Question Answering

Xinzhe Han[2,1] Shuhui Wang[1]* Chi Su[3] Qingming Huang[1,2,4] Qi Tian[5]
[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China [3]Kingsoft Cloud, Beijing, China
[4]Peng Cheng Laboratory, Shenzhen, China [5]Cloud BU, Huawei Technologies, Shenzhen, China.
hanxinzhe17@mails.ucas.ac.cn, wangshuhui@ict.ac.cn, suchi@kingsoft.com
qmhuang@ucas.ac.cn, tian.qi1@huawei.com

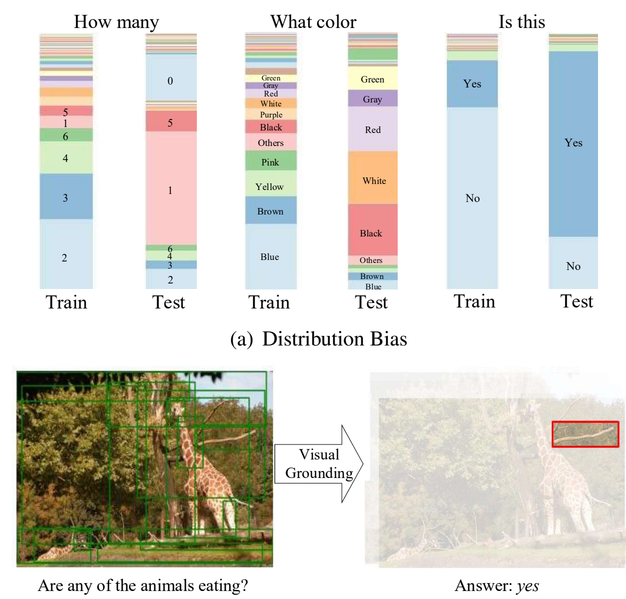
Abstract

Language bias is a critical issue in Visual Question Answering (VQA), where models often exploit dataset biases for the final decision without considering the image information. As a result, they suffer from performance drop on out-of-distribution data and inadequate visual explanation. Based on experimental analysis for existing robust VQA methods, we stress the language bias in VQA that comes from two aspects, i.e., distribution bias and shortcut bias. We further propose a new de-bias framework, Greedy Gradient Ensemble (GGE), which combines multiple biased models for unbiased base model learning. With the greedy strategy, GGE forces the biased models to over-fit the biased data distribution in priority, thus makes the base model pay more attention to examples that are hard to solve by biased models. The experiments demonstrate that our method makes better use of visual information and achieves state-of-the-art performance on diagnosing dataset VQA-CP without using extra annotations.

1. Introduction

Visual Question Answering (VQA) is a challenging task that requires both language-aware reasoning and image understanding. With advances in deep learning, neural networks [37, 34, 6, 13, 18, 17, 19, 29] that model the correlations between vision and language have shown remarkable results on large-scale benchmark datasets [3, 15, 23, 20].

However, recent studies have demonstrated that most VQA methods tend to rely on existing idiosyncratic biases in the datasets [15, 24, 43]. They often leverage superficial correlations between questions and answers to train the model without considering exact vision information. For example, a model may blindly answer "tennis" for the question "What sports ..." just based on the most common textual QA pairs in the train set. Unfortunately, models exploiting



(b) Shortcut Bias

Figure 1. Two aspects of language bias in VQA. **(a) Distribution Bias:** The answer distribution for certain question type is significantly long-tailed. **(b) Shortcut Bias:** The correct answers produced by the model may rely on the question-answer shortcut rather than proper visual grounding.

statistical shortcuts during training often show poor generalization ability to out-of-domain data, and hardly provide proper visual evidence for a certain answer.

Currently, the prevailing solutions for this problem can be categorized into ensemble-based [36, 7, 10], grounding-based [39, 42, 22] and counterfactual-based [8]. Similar to re-weighting and re-sampling strategies in traditional long-tailed classification [45, 25, 16, 31], ensemble-based methods re-weight the samples by the question-only branch. Grounding-based models stress a better use of image information according to human-annotated visual explanation [11, 21]. Newly proposed counterfactual-based meth-

*Corresponding author.

ods [8, 33] further combine these two lines of work and achieve better performance.

Nevertheless, it has been shown that existing methods have not fully leveraged both vision and language information. For example, Shrestha *et al*. [40] argue that improved accuracy in grounding-based methods [39, 42] does not actually emerge from proper visual grounding but some unknown regularization effects. Similar to [40], we further analyse all the three categories of existing work by control experiments in Section 3.2. We found that language bias in VQA is actually two-fold: (a) the statistical distribution gap between train and test, *i.e*., *distribution bias* shown in Figure 1(a), and (b) the semantic correlation between specific QA pairs, *i.e*., *shortcut bias* shown in Figure 1(b). Although long-tailed distribution in train set is usually considered to be one of the factors that increase shortcut bias, we experimentally demonstrate that they are actually two aspects of the language bias. Grounding supervision in [39] or ensemble regularization in [7, 10] does not necessarily force the model to focus on visual information as expected. To encourage the model to pay attention to the images, we need to explicitly model bothbiases and reduce them step by step.

Inspired by our empirical findings, we propose Greedy Gradient Ensemble (GGE), a new model-agnostic de-bias framework that ensembles biased models and the base model like gradient descent in functional space. The key idea of our method is to make use of the over-fitting phenomenon in deep learning. The biased part of data is greedily over-fitted by biased features, as a result, the expected base model can be learned with more ideal data distribution and focus on examples that are hard to solve with biased models.

In the experiments, variants of GGE models are provided in ablation study, which demonstrates the generalization ability of our method and further supports our claim that distribution bias and question shortcut bias are complementary in VQA. To verify if a model can really use visual information for the answer decision, we further study the language bias in VQA from a visual modelling perspective. Quantitative and qualitative evaluations show that GGE can provide better visual evidence accompanied with predictions.

The major contributions are:

- We provide analysis for the language bias in VQA task and decompose the language bias into distribution bias and shortcut bias.
- We propose a new model-agnostic de-bias framework Greedy Gradient Ensemble (GGE), which sequentially ensembles biased models for robust VQA.
- On VQA-CP, our method makes better use of visual information and achieves state-of-the-art performance, with 17.34% gain against simple UpDn baseline without extra annotations. Code is available at https://github.com/GeraldHan/GGE.

2. Related work

2.1. De-bias with dataset construction

The most straightforward way to remove the dataset bias is to construct a balanced dataset. For instance, Zhang *et al*. [43] collect complementary abstract scenes with opposite answers for all binary questions. Similarly, VQA v2 [15] is introduced to weaken language priors in the VQA v1 dataset [3] by adding similar images with different answers for each question. Agrawal *et al*. [1] introduce a diagnosing VQA dataset under Changing Prior (VQA-CP) constructed with different answer distributions between the train and test splits. Most of the models that perform well on VQA v2 significantly drop on VQA-CP in Accuracy.

2.2. De-bias with model design

Collecting new large-scale datasets is costly. It is crucial to develop models that are robust to biases. Along with VQA-CP dataset [1], Agrawal *et al*. propose GVQA model that disentangles the visual concept recognition from the answer space prediction. LDP [22] and GVQE [27] exploit different information in questions for better question representation. These models require a pre-defined question parser, making them hard to implement.

Another line of work starts from visual grounding. Early works [35, 44] directly apply human grounding [11, 21] as supervision to attention maps, but the improvement is limited. HINT [39] and SCR [42] change supervised attention maps to Grad-CAM, which directly encourages the contribution of each object to be consistent with human annotations. Recent work [40] experimentally challenges the effectiveness of visual grounding in [39, 42], but does not find a good way to test if systems are actually visually grounded.

The most effective solution so far is ensemble-based, which formulates a question-only branch as explicit modelling for language bias. Ramakrishnan *et al*. [36] introduce an adversarial regularization to remove answer discriminative feature from the questions. RUBi [7], LMH [10] and PoE [28] re-weight samples based on the question-only prediction. Niu *et al*. [33] further improve ensemble strategies from a causal-effect perspective. CSS [8] combines grounding-based and ensemble-based methods with counterfactual samples synthesizing. Gat *et al*. [14] introduce a regularization by maximizing functional entropies (MFE), which forces the model to use multiple sources of information in multi-modal tasks. Nam *et al*. [32] propose a general framework LfF, which trains the de-biased classifier from a biased classifier. Compared to our work, they mainly focus on single-modality classification problems and their General Cross-Entropy (GCE) re-weighting strategy is less flexible, which relies on hyper-parameter in GCE and can only handle one pair of attributes in de-bias learning.

3. Revisiting Language Bias in VQA

3.1. Problem Definition

For base model, we consider the common formulation of VQA task as a multi-class classification problem. Given a dataset $\mathcal{D} = \{v_i, q_i, a_i\}_{i=1}^N$ consisting of an image $v_i \in \mathcal{V}$, a question $q_i \in \mathcal{Q}$ and a labelled answer $a_i \in \mathcal{A}$, we need to optimize a mapping $f_{VQ} : V \times Q \to \mathbb{R}^C$ which produces a distribution over the C answer candidates. Without loss of generality, the function is composed as following:

$$\tilde{a}_i = f_\theta(v_i, q_i) = c\left(m\left(e_v(v_i), e_q(q_i)\right)\right), \tag{1}$$

where $e_v : V \to \mathbb{R}^{n_v \times d_v}$ is an image encoder, $e_q : Q \to \mathbb{R}^{n_q \times d_q}$ is a question encoder, $m(.)$ denotes the multi-modal fusion or reasoning module, and $c(.)$ is the multi-layer perception classifier. The output is a vector $\tilde{a} \in \mathbb{R}^C$ indicating the probability belonging to each answer candidate.

3.2. Experimental Analysis for Language Bias

In recent work, Shrestha *et al*. [40] experimentally challenge the way grounding-based methods [42, 39] work on VQA-CP [1]. However, they did not provide insights into the language bias itself. In this section, from a new de-bias method perspective, we provide three control experiments for baseline model UpDn [2], grounding-based method HINT [39], ensemble-based method RUBi [7] LMH [10] and counterfactual-based method CSS [8] on VQA-CP and VQA v2 to discuss the language bias in VQA.

Inverse Grounding Annotation. To analyse the contribution of visual-grounding, we first experiment with HINT and CSS-V that use human attention as extra information. Following [40], we change human-annotated region importance scores [11] S_h to irrelevant grounding $S'_h = 1 - S_h$. As shown in Table 1, the performance of HINT$_{inv}$ and CSS-V$_{inv}$ is almost the same as the original models. This indicates that the Accuracy gains are not necessarily from looking at relevant regions [4]. Although the models correctly answer some hard questions, they still make predictions based on language information regardless of images. We refer to this unexpected solution as "inverse language bias".

Vision-only Model. The second experiment aims to analyse the function of the ensemble branch in RUBi and LMH. For the base model, we only feed the vision feature without multi-modal fusion to the answer classifier:

$$\tilde{a}_i = c\left(e_v(v_i)\right). \tag{2}$$

There is no question information for classification in base model, and thus obviously no shortcut between QA pairs to reduce. As shown in Table 1, RUBi$_{vo}$ degrades a lot, but LMH$_{vo}$ still surpasses UpDn$_{vo}$ by a large margin in Accuracy. Apart from restraining shortcuts between question-answer pairs, we think the improved Accuracy in LMH mainly comes from penalizing the most common answers in the train set, which leads to a more balanced classifier according to inverse distribution. This means the distribution bias in LMH plays a different role compared with the question shortcut in RUBi.

Table 1. Experimental analysis for representative methods on VQA-CP and VQA v2. Footnote inv stands for Inverse Grounding Annotation, vo for Vision-only, and is for Inverse-Supervision.

Method	VQA-CP	VQA 2.0
UpDn [2]	39.89	63.79
HINT [39]	47.50	63.38
RUBi [7]	45.42	58.19
LMH [10]	52.73	56.35
CSS [8]	58.11	53.15
HINT$_{inv}$	47.20	60.33
CSS-V$_{inv}$	58.05	54.39
UpDn$_{vo}$	33.18	45.67
RUBi$_{vo}$	23.53	46.11
LMH$_{vo}$	43.68	27.18
UpDn$_{vo,is}$	39.44	40.03
UpDn$_{is}$	42.12	60.85
RUBi$_{is}$	48.42	59.10
LMH$_{is}$	58.12	43.29

Inverse Supervision for Balanced Classifier. To directly verify if such "inverse distribution bias" can improve Accuracy, inspired by the two-round training in CSS [8], we design a simple "inverse supervision" strategy. For each iteration, the parameters are updated two rounds with different supervisions. In the first round, we train the model supervised by ground-truth label $\mathcal{A}$ and get the prediction $P(a)$. The top-N answers with the highest predicted probabilities are selected as $\mathbf{a}^+$. In the second-round training, the label is defined as $\hat{\mathcal{A}} = \{a_i | a_i \in \mathcal{A}, a_i \notin \mathbf{a}^+\}$. This strategy is actually a simplified version of CSS [8] without object/question masks. In this way, the model continuously penalizes the most confident answers in the first round training, thus formulates a more balanced classifier according to inverse distribution bias. The Accuracy improvement in UpDn$_{vo,is}$ reveals the existence of distribution bias. The result of RUBi$_{is}$ further indicates that distribution bias and shortcut bias are complementary. LMH$_{is}$ is even comparable to CSS that uses extra annotations. However, this method leads to catastrophic degradation on the in-distribution dataset VQA v2 as shown in Table 1.

According to the above experiments, we obtain the following valuable insights: 1) Good Accuracy can not guarantee that the system is really visually grounded for answer classification. Grounding supervision or question-only regularization may encourage models to make use of inverse language bias rather than better visual information for higher Accuracy. 2) Distribution bias and shortcut bias are complementary aspects of language bias in VQA. A single ensemble branch is unable to model such two types of biases.

4. Method

Based on the above findings, we propose GGE, a new model-agnostic de-bias learning paradigm, which removes distribution bias and shortcut bias step by step, thus forces the model to focus on images.

4.1. Greedy Gradient Ensemble

Let (X, Y) denote the train set, where X is the space of observations, and Y is the space of labels. Following previous VQA methods, we mainly consider the classification problem with binary cross-entropy (BCE) loss

$$\mathcal{L}(P, Y) = -\sum_{i=1}^{C} y_i \log(p_i) + (1 - y_i) \log(1 - p_i), \quad (3)$$

where C denotes the number of classes. $p_i = \sigma(z_i)$ where z_i is the predicted logit for class i and $\sigma(.)$ is the sigmoid function. Baseline methods directly minimize the loss between the prediction $f(X; \theta)$ and label Y

$$\min_{\theta} \mathcal{L}\left(\sigma(f(X; \theta)), Y\right). \quad (4)$$

Since $f(.)$ is over-parametrized DNNs, the model is easy to over-fit the dataset biases and suffers from poor generalization ability.

For our method, we make use of this kind of over-fitting in deep learning. Assume $\mathcal{B} = \{B_1, B_2, \ldots, B_M\}$ to be a set of bias features that can be extracted based on prior knowledge. This time we fit the ensemble of bias models and base model to label Y

$$\min_{\phi, \theta} \mathcal{L}\left(\sigma\left(f(X; \theta) + \sum_{i=1}^{M} h_i(B_i; \phi_i)\right), Y\right), \quad (5)$$

where $h_i(.)$ is a biased model for certain biased feature. Ideally, we hope the biased part of data is only over-fitted by the bias models, thus the base model can be learned with unbiased data distribution. To achieve this goal, we propose GGE in which biased models have a higher priority to over-fit the dataset with greedy strategy.

Viewing in the functional space, suppose we have $\mathcal{H}_m = \sum_{i=1}^{m} h_i(B_i)$ and we wish to find $h_{m+1}(B_{m+1})$ added to $\mathcal{H}_m$ so that the loss $\mathcal{L}\left(\sigma(\mathcal{H}_m + h_{m+1}(B_{m+1})), Y\right)$ decreases. In theory, the desired direction of h_{m+1} is the negative derivative of $\mathcal{L}$ at $\mathcal{H}_m$, where

$$-\nabla \mathcal{L}(\mathcal{H}_{m,i}) := \frac{\partial \mathcal{L}\left(\sigma(\mathcal{H}_m), Y\right)}{\partial \mathcal{H}_{m,i}} = 2y_{m,i}\sigma\left(-2y_{m,i}\mathcal{H}_{m,i}\right). \quad (6)$$

For a classification problem, we only care about the probability for class i: $\sigma(f_i(x)) \in \{0, 1\}$. Therefore, we treat the negative gradients as pseudo labels for classification and optimize the new model $h_{m+1}(B_{m+1})$ with BCE loss:

$$L_{m+1} = \mathcal{L}\left(\sigma(h_{m+1}(B_{m+1}; \phi_{m+1})), -\nabla \mathcal{L}(\mathcal{H}_m)\right). \quad (7)$$

After integrating all biased models, the expected base model f is optimized with

$$L_b(\theta) = \mathcal{L}\left(\sigma(f(X; \theta)), -\nabla \mathcal{L}(\mathcal{H}_M)\right). \quad (8)$$

In the test stage, we only use the base model for predictions.

More intuitively, for a sample that is easy to fit by biased models, the negative gradient of its loss $-\nabla \mathcal{L}(\mathcal{H}_M)$ (*i.e.*, the pseudo label for the base model) will become relatively small. $f(X; \theta)$ will pay more attention to samples that are hard to solve by previous ensemble biased classifiers $\mathcal{H}_M$.

In order to make the above paradigm adaptive to Batch Stochastic Gradient Decent (Batch SGD), we implement two optimization schedules GGE-iteration and GGE-together, as shown in Algorithm 1 and Algorithm 2 in Supplementary. GGE-tog jointly optimizes biased models and the base model with

$$L(\Theta) = L_b(\theta) + \sum_{m=1}^{M} L_m(\phi_m). \quad (9)$$

For GGE-iter, each model is iteratively updated within a certain data-batch iteration. More details for GGE are provided in Section A in Supplementary.

4.2. GGE Implementation for Robust VQA

Following analysis in Section 3, we define two biased features for VQA, *i.e.*, distribution bias and shortcut bias.

Distribution Bias. We define the distribution bias as answer distribution in the train set conditioned on question types

$$B_d^i = p(a_i | t_i), \quad (10)$$

where t_i denotes the type of question q_i. The reason for counting samples conditioned on question types is to maintain type information when reducing distribution bias. Question type information can only be obtained from the questions rather than the images, which does not belong to the language bias to be reduced.

Shortcut Bias. Shortcut bias is the semantic correlation between specific QA pairs. Similar to [7], we compose the question shortcut bias as a question-only branch

$$B_q^i = c_q\left(e_q(q_i)\right), \quad (11)$$

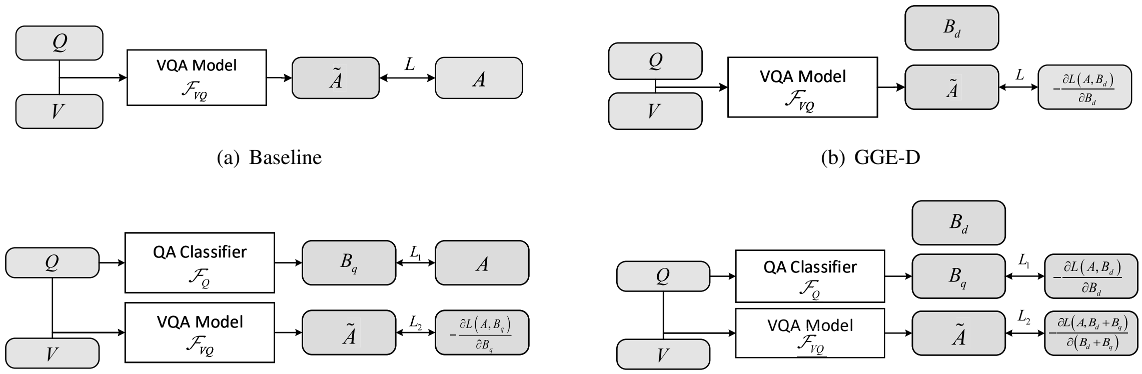
where $c_q : Q \to \mathbb{R}^C$.

To verify our claim that distribution bias and shortcut bias are complementary, we design three versions of GGE for ensembles of different language biases.

GGE-D only models distribution bias for ensemble, shown in Figure 2(b). The loss for the base model is

$$L = \mathcal{L}(\sigma(\tilde{A}), -\nabla \mathcal{L}(B_d, A)), \quad (12)$$

where $\tilde{A}$ is the predictions, and A is the labelled answers.



(a) Baseline (b) GGE-D

(c) GGE-Q (d) GGE-DQ

Figure 2. **Different versions of GGE**. V, Q and $\tilde{A}$ denote image, question, and answer prediction respectively. A is the human-annotated labels. B_d and B_q indicate the prediction from distribution bias and question shortcut bias respectively.

GGE-Q only uses a question-only branch for shortcut bias. As shown in Figure 2(c), we first optimize the question-only branch with labelled answers

$$L_1 = \mathcal{L}(\sigma(B_q), A). \tag{13}$$

The loss for base model is

$$L_2 = \mathcal{L}(\sigma(\tilde{A}), -\nabla\mathcal{L}(\sigma(B_q), A)). \tag{14}$$

GGE-DQ uses both distribution bias and question shortcut bias. As shown in Figure 2(d), the loss for B_q is

$$L_1 = \mathcal{L}(\sigma(B_q), -\nabla\mathcal{L}(B_d, A)). \tag{15}$$

The loss for base model is

$$L_2 = \mathcal{L}(\sigma(\tilde{A}), -\nabla\mathcal{L}(\sigma(B_q) + B_d, A)). \tag{16}$$

We test both GGE-iter or GGE-tog for L_1 and L_2.

4.3. Connection to Boosting

Boosting [12, 38, 38, 9] is a widely used ensemble strategy for classification problems. The key idea of boosting is to combine multiple weak classifiers with high bias but low variance to produce a strong classifier with low bias and low variance. Each base learner has to be weak enough, otherwise, the first few classifiers will easily over-fit to the training data [5]. However, the neural networks' fitting ability is too strong to be "high bias" and "low variance" for boosting strategy, making it hard to use deep models as weak learners. In this paper, our method exploits this over-fitting phenomenon, making biased weak features to over-fit the bias distribution. In the test stage, we only use the base model trained with the gradient of biased models, thus removing language bias in VQA.

On the other hand, the idea of approximating negative gradients is very similar to Gradient Boost [30]. However, Gradient Boost has to greedily learn weak learners in turn. This will be costly for complicated neural networks via back-propagation. We design two strategies, GGE-iteration and GGE-together, in which the learners are updated along with Batch SGD.

5. Experiments

The experiments are conducted on both language-bias sensitive VQA-CP v2 [1] and standard VQA v2 [15]. Considering there is no validation set for VQA-CP, we simply choose the model in the last training epoch for comparison in consequent experiments. More implementation details can be found in Section C in the Supplementary.

5.1. Evaluation Metrics

For each model, we compare Accuracy, the standard VQA evaluation metric [3]. Moreover, a robust VQA model is expected to leverage both visual and language information, but good Accuracy is not enough to indicate the system is well visually grounded according to analysis in Sec. 3.

In [40], a new metric Correctly Predicted but Improperly Grounded (CPIG) is proposed to quantitatively assess visual grounding in VQA. An instance is regarded as correctly grounded if the ground-truth regions for the right answer (*e.g.*, HAT [21]) are within the model's top-N most sensitive visual regions. For convenience, we define $1 - CPIG$ as CGR (Correct Grounding for Right prediction):

$$\%CGR = \frac{N_{\text{rg,rp}}}{N_{\text{rp}}} \times 100\%, \tag{17}$$

where N_{rp} is the total number of right predictions, $N_{\text{rg,rp}}$ is the number of instances that are correctly answered with correct visual grounding. However, similar to results in [40], we find that CGR is not very discriminative across different methods as shown in Table 2 in Supplementary. The model with high CGR (*e.g.*, UpDn) may not actually use enough visual information for classification. If a model

Table 2. Experimental results on VQA-CP v2 test set and VQA v2 val set of state-of-the-art methods. **Best** and **second** performance are highlighted in each column. Methods with * use extra annotations (*e.g.*, human attention (HAT), explanations (VQA-X), or object label information). Methods with CGD are our reimplementation using released codes. Other results are reported in the original papers.

Method	Base	VQA-CP test					VQA v2 val			
		All	Y/N	Num.	Others	CGD	All	Y/N	Num.	Others
GVQA [1]	-	31.30	57.99	13.68	22.14	-	48.24	72.03	31.17	34.65
UpDn [2]	-	39.89	43.01	12.07	45.82	3.91	**63.79**	80.94	42.51	**55.78**
S-MRL [7]	-	38.46	42.85	12.81	43.20	-	63.10	-	-	-
HINT* [39]	UpDn	47.50	67.21	10.67	46.80	10.34	63.38	**81.18**	42.14	**55.66**
SCR* [42]	UpDn	49.45	72.36	10.93	48.02	-	62.2	78.8	41.6	54.4
AdvReg. [36]	UpDn	41.17	65.49	15.48	35.48	-	62.75	79.84	42.35	55.16
RUBi [7]	UpDn	45.42	63.03	11.91	44.33	6.27	58.19	63.04	41.00	54.43
LM [10]	UpDn	48.78	70.37	14.24	46.42	11.33	63.26	81.16	42.22	55.22
LMH [10]	UpDn	52.73	72.95	**31.90**	47.79	10.60	56.35	65.06	37.63	54.69
DLP [22]	UpDn	48.87	70.99	18.72	45.57	-	57.96	76.82	39.33	48.54
GVQE* [27]	UpDn	48.75	-	-	-	-	**64.04**	-	-	-
CSS* [8]	UpDn	41.16	43.96	12.78	47.48	8.23	59.21	72.97	40.00	55.13
CF-VQA(Sum) [33]	UpDn	53.69	**91.25**	12.80	45.23	-	63.65	**82.63**	**44.01**	54.38
GGE-DQ-iter (Ours)	UpDn	**57.12**	87.35	26.16	**49.77**	**16.44**	59.30	73.63	40.30	54.29
GGE-DQ-tog (Ours)	UpDn	**57.32**	87.04	27.75	**49.59**	**15.27**	59.11	73.27	39.99	54.39
RUBi [7]	S-MRL	47.11	68.65	20.28	43.18	-	61.16	-	-	-
GVQE* [27]	S-MRL	50.11	66.35	27.08	46.77	-	63.18	-	-	-
CF-VQA(Sum) [33]	S-MRL	54.95	**90.56**	21.88	45.36	-	60.76	81.11	**43.48**	49.58
MFE [14]	LMH	54.55	74.03	**49.16**	45.82	-	-	-	-	-
CSS* [8]	LMH	58.21	83.65	40.73	48.14	8.81	53.15	61.20	37.65	53.36

locates the right object but still produces a wrong answer, it is a safe bet that it heavily relies on language bias instead of images for prediction. To quantitatively assess whether a model uses visual information for answer decision, we introduce CGW (Correct Grounding but Wrong prediction):

$$\%CGW = \frac{N_{\text{rg, wp}}}{N_{\text{wp}}} \times 100\%, \tag{18}$$

where N_{wp} is the number of wrong predictions, and $N_{\text{rg,wp}}$ is the number of instances for which the model provides the right visual evidences but wrong prediction. Bad cases like example 2 and 3 from UpDn in Fig. 4 are ignored by CGR but can be identified by high CGW.

For clearer comparison, we denote the difference of CGR and CGW as CGD (Correct Grounding Difference):

$$\%CGD = \%CGR - \%CGW. \tag{19}$$

CGD *only* evaluates whether the visual information is taken in answer decision, which is parallel with Accuracy. The key idea for CGD is that a model actually makes use of visual information should not only provide the right predictions based on the correct visual-groundings but also a wrong answer due to improper visual evidence as well. Detailed CGR and GCD for all experiments are provided in Table 2 in Supplementary. It shows that UpDn, HINT$_{inv}$ and CSS-V$_{inv}$ achieve comparable performance on Accuracy but significantly degrade on CGD. This meets our intuitive analysis that these methods do not fully exploit visual information for the answer decision. Although the visual-grounding annotations are not so reliable for some instances[1], CGD can offer statistically better distinction from the whole dataset level. More details for CGD are provided in Section B in the Supplementary.

5.2. Comparison with State-of-the-art Methods

We compare our best performed model GGE-DQ with existing state-of-the-art bias reduction techniques, including visual grounding-based methods HINT [39], SCR [42], ensemble-based methods AdvReg. [36], RUBi [7], LM (LMH) [10], MFE [14], new question encoding-based methods GVQE [27], DLP [33], counterfactual-based methods CF-VQA [33], CSS [8] , and recent proposed regularization method MFE [14].

Experiments on VQA-CP test set aim to evaluate whether VQA models effectively reduce language bias. As shown in Table 2, GGE-DQ achieves state-of-the-art performance without extra annotation. It outperforms the baseline model UpDn by 17% higher in Accuracy and 13% higher in CGD, which verifies the effectiveness of GGE on both answer classification and visual-grounding ability. Under the same base model UpDn, our method achieves the best performance in both Accuracy and CGD, with $\sim 5\%$ gain comparing to all other methods, even competitive with methods that use stronger base models.

For the comparison of question-type-wise results, incor-

[1]Not all examples in VQA v2 are annotated in VQAX [11]. Moreover, visual grounding for some instances are hard to evaluate (*e.g.*, questions that require global image information or without referring objects)

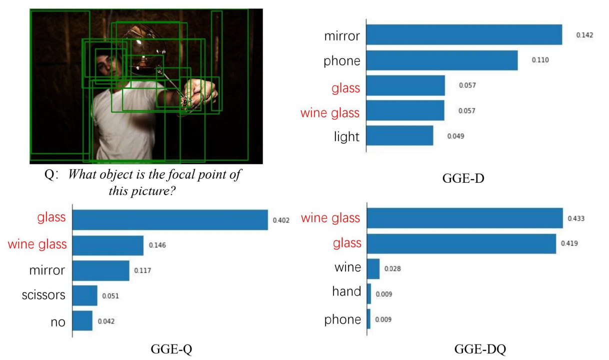


Figure 3. Predicted distribution for three variants of GGE.

porating GGE reduces the biases and improves the performance for all the question-types, especially the more challenging "other" question type [41]. CF-VQA [33] performs the best in Y/N, but worse than our methods in all other metrics. LMH [10], LMH-MFE [14] and LMH-CSS [8] surpass other methods in Num., and LMH-CSS even slightly outperforms GGE-DQ in overall Accuracy due to high performance in Num. (40.73%). Comparing LM and LMH, it is obvious that the performance gains in Num. are due to the additional regularization for entropy. However, methods with entropy regularization drop nearly 10% on VQA v2. This indicates that these models may over-correct the bias and largely use "inverse language bias".

5.3. Ablation Studies

In this section, we design various ablations to verify the effectiveness of greedy ensemble and our claim that distribution bias and question shortcut bias are two aspects of language bias. More results on VQA v2 are provided in Section D in the Supplementary.

The first group of ablations is to verify if the greedy ensemble can guarantee biased data is learned with biased models. We compare with other two ensemble strategies. **SUM-DQ** directly sums up the outputs of biased models and the base model. **LMH+RUBi** combines LMH [10] and RUBi [7]. It reduces distribution bias with LMH and shortcut bias with RUBi. The implementation details for these two ablations are provided in Section C in Supplementary.

As shown in Table 3, SUM-DQ performs even worse than baseline. Meanwhile, the Accuracy of LMH+RUBi is just similar to that of LMH, and about 6% worse than GGE-DQ. This shows that GGE can really force the biased data to be sequentially learned with biased models. Instances that are easy to predict based on distribution or shortcut bias will be well fitted by the corresponding biased model. As a result, the base model has to pay more attention to hard examples and consider more visual information for final decision.

In the second group of experiments, we experimentally compare distribution bias and shortcut bias. The case analysis in Figure 3 shows that GGE-D only uniforms predictions, which mainly improves Y/N as shown in Table 3. B_q works like "hard example mining" but will also introduce some noise (*e.g.* "mirror" and "no" in this example) due to inverse distribution bias. Reducing B_d at the first stage can further encourage the discovery of the hard examples and force the base model to capture visual information. In Figure 3, the correct answer has higher confidence and the top predictions are all based on the image. As shown in Table 3, GGE-DQ surpasses single-bias versions by $\sim$10%. This well verifies our claim that distribution bias and shortcut bias are two complementary aspects of language bias.

Table 3. Ablation study for different versions of GGE on VQA-CP v2 test set. **Best** results are highlighted in the columns.

Method	All	Y/N	Others	Num.	CGD
Baseline	39.89	43.01	45.80	11.88	3.91
SUM-DQ	35.46	42.66	38.01	12.38	3.10
LMH+RUBi	51.54	74.55	47.41	22.65	6.12
GGE-D	48.27	70.75	47.53	13.42	14.31
GGE-Q-iter	43.72	48.17	48.78	14.24	6.70
GGE-Q-tog	44.62	47.64	48.89	14.34	6.63
GGE-DQ-iter	57.12	**87.35**	**49.77**	26.16	**16.44**
GGE-DQ-tog	**57.32**	87.04	49.59	**27.75**	15.27

5.4. Generalization of GGE

Self-Ensemble. The performance of GGE largely depends on the predefined biased features, which requires prior knowledge of the task or dataset.In order to further discuss the generalization of GGE, we test a more flexible Self-Ensemble fashion (GGE-SF) on VQA-CP. GGE-SF takes the joint representation $r_i = m\left(e_v(v_i), e_q(q_i)\right)$ itself as the biased feature instead of predefined question-only branch, the biased prediction is

$$B_{s_i} = c_s\left(r_i\right), \tag{20}$$

where $c_s : r \rightarrow \mathbb{R}^C$ is the classifier of the biased model. The training process is the same as GGE-Q.

As shown in Table 4, GGE-SF still surpasses the baseline even without predefined biased features. This means that the base model itself can also be regarded as a biased model, as long as the tasks or datasets are biased enough. Moreover, if we first remove distribution bias with GGE-D before Self-Ensemble, the performance of GGE-D-SF is also comparable to existing state-of-the-art methods.

Generalization for Loss Function. For a fair comparison with previous work, we adopt Sigmoid+BCE loss for the above experiments. Actually, GGE is agnostic for classification losses. We provide extra experiments for Softmax+CE loss in Table 4. The implementation for GGE$_{sxce}$ is provided in Section A in the Supplementary.

Generalization for Base Model. GGE is also agnostic for base model choices. We provide extra experiments with BAN [26] and S-MRL [7] as base model. The results are provided in Section D in the Supplementary.

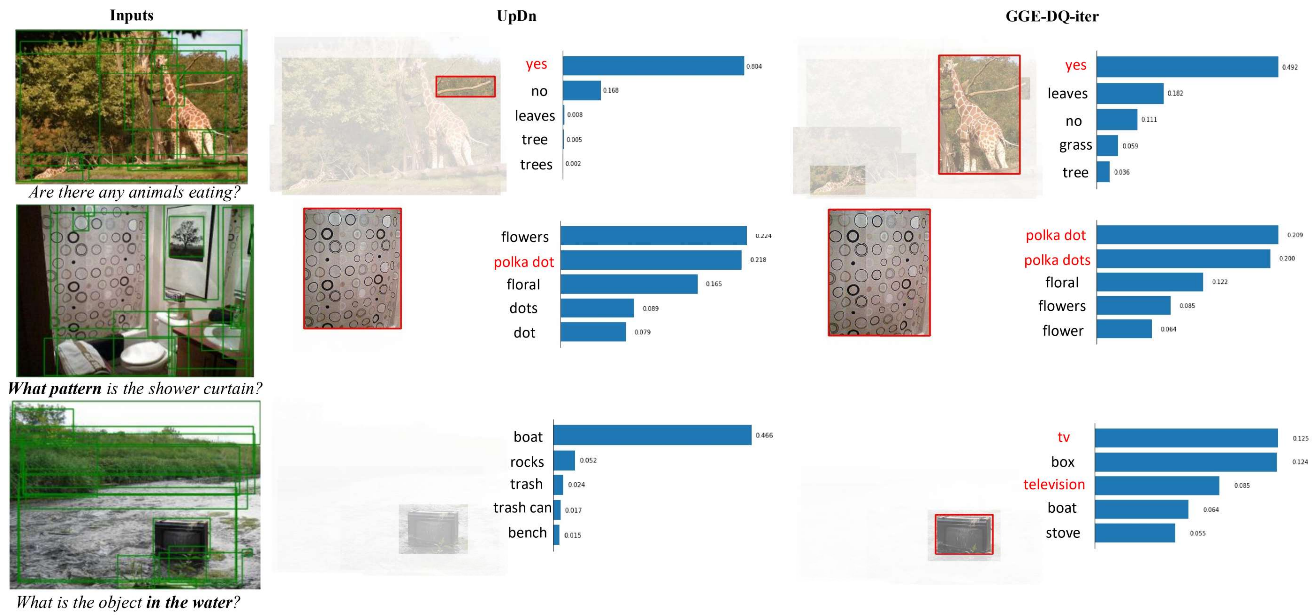


Figure 4. **Qualitative Evaluation for GGE-DQ**. We provide a comparison between UpDn and GGE-DQ on the visualization of the most sensitive regions and confidence of the top-5 answers. Red answers denote the ground-truth.

Table 4. Variants of GGE on VQA-CP v2. SF stands for Self-Ensemble, $sxce$ denotes models trained with softmax+CE loss.

Method	All	Y/N	Others	Num.
UpDn	39.89	43.01	45.80	11.88
UpDn$_{sxce}$	41.37	45.96	46.90	12.46
GGE-SF-iter	44.53	50.98	48.90	18.24
GGE-SF-tog	43.10	49.90	47.33	17.74
GGE-D-SF-iter	56.33	86.43	49.32	24.37
GGE-D-SF-tog	52.86	76.25	49.46	20.56
GGE$_{sxce}$-D	53.98	86.06	47.85	15.09
GGE$_{sxce}$-Q-iter	52.98	82.27	48.06	14.97
GGE$_{sxce}$-Q-tog	52.99	81.86	47.97	16.11
GGE$_{sxce}$-DQ-iter	56.25	85.08	48.56	24.78
GGE$_{sxce}$-DQ-tog	55.84	84.47	48.76	26.96

5.5. Qualitative Evaluation

Examples in Figure 4 illustrate how GGE-DQ makes of visual information for inference. From top to bottom, we provide three representative failure cases from baseline UpDn. The first example is about shortcut bias. Despite offering the right answer "yes", the prediction from UpDn is not based on the right visual grounding. On the contrary, GGE correctly grounds the giraffe that is eating leaves. The second example is about distribution bias. UpDn correctly grounds the curtain but still answers the question based on distribution bias ("flowers" is the most common answer for "what pattern..." in the train set). The last example is a case for reducing language prior apart from Yes/No questions. UpDn answers "boat" just based on the language context "in the water", while GGE-DQ provides correct answers "tv" and "television" with more salient visual grounding. These examples qualitatively verify our improvement in both Accuracy and visual explanation for the predictions. More examples and failure cases can be found in Supplementary.

6. Conclusion

In this paper, we experimentally analyse several methods for robust VQA and propose a new framework to reduce the language bias in VQA. We demonstrate that the language bias in VQA can be decomposed into distribution bias and shortcut bias and then propose a Greedy Gradient Ensemble strategy to removes such two kinds of preferences step by step. Experimental results demonstrate the rationality of our bias decomposition and the effectiveness of GGE. We believe the idea behind GGE is valuable and has the potential to become a generic method for dataset bias problems. In the future, we will extend GGE to solve bias problems for other tasks, provide a more rigorous analysis to guarantee model convergence, and learn to automatically detect different kinds of bias features without prior knowledge.

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102003, in part by National Natural Science Foundation of China: 62022083, 61620106009, 61836002 and 61931008, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, and in part by the Beijing Nova Program under Grant Z201100006820023. Authors are grateful to Kingsoft Cloud for free GPU computing support.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 2, 3, 5, 6

[2] 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*, pages 6077–6086, 2018. 3, 6

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 5

[4] Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. Why is attention not so attentive? *arXiv preprint arXiv:2006.05656*, 2020. 3

[5] Peter J Bickel, Yaacov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7(May):705–732, 2006. 5

[6] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. 1

[7] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *Advances in neural information processing systems*, pages 841–852, 2019. 1, 2, 3, 4, 6, 7

[8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. 1, 2, 3, 6, 7

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 5

[10] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Dont take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4060–4073, 2019. 1, 2, 3, 6, 7

[11] 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, 2, 3, 6

[12] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995. 5

[13] Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. Multi-modality latent interaction network for visual question answering. *arXiv preprint arXiv:1908.04289*, 2019. 1

[14] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 6, 7

[15] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 1, 2, 5

[16] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, 2018. 1

[17] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision*, pages 53–69, 2018. 1

[18] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[19] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10294–10303, 2019. 1

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 1

[21] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018. 1, 2, 5

[22] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *AAAI*, pages 11181–11188, 2020. 1, 2, 6

[23] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 1

[24] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017. 1

[25] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decou-

pling representation and classifier for long-tailed recognition. *ICLR*, 2020. 1

[26] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 7

[27] Gouthaman KV and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. *arXiv preprint arXiv:2007.06198*, 2020. 2, 6

[28] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[29] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019. 1

[30] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000. 5

[31] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019. 1

[32] Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *34th Conference on Neural Information Processing Systems (NeurIPS) 2020*. Neural Information Processing Systems, 2020. 2

[33] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *arXiv preprint arXiv:2006.04315*, 2020. 2, 6, 7

[34] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[35] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. *arXiv preprint arXiv:1709.06308*, 2017. 2

[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*, pages 1541–1551, 2018. 1, 2, 6

[37] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 1

[38] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990. 5

[39] 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 International Conference on Computer Vision*, pages 2591–2600, 2019. 1, 2, 3, 6

[40] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for VQA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181, Online, July 2020. Association for Computational Linguistics. 2, 3, 5

[41] 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. *arXiv preprint arXiv:2005.09241*, 2020. 7

[42] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. In *Advances in Neural Information Processing Systems*, pages 8604–8614, 2019. 1, 2, 3, 6

[43] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016. 1, 2

[44] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 349–357. IEEE, 2019. 2

[45] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020. 1