A Negative Case Analysis of Visual Grounding Methods for VQA

Robik Shrestha[1] **Kushal Kafle**[1,2] **Christopher Kanan**[1,3,4]

Rochester Institute of Technology[1] Adobe Research[2] Paige[3] Cornell Tech[4]

{rss9369, kk6055, kanan}@rit.edu

Abstract

Existing Visual Question Answering (VQA) methods tend to exploit dataset biases and spurious statistical correlations, instead of producing right answers for the right reasons. To address this issue, recent bias mitigation methods for VQA propose to incorporate visual cues (e.g., human attention maps) to better ground the VQA models, showcasing impressive gains. However, we show that the performance improvements are not a result of improved visual grounding, but a regularization effect which prevents over-fitting to linguistic priors. For instance, we find that it is not actually necessary to provide proper, human-based cues; random, insensible cues also result in similar improvements. Based on this observation, we propose a simpler regularization scheme that does not require any external annotations and yet achieves near state-of-the-art performance on VQA-CPv2[1].

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015), the task of answering questions about visual content, was proposed to facilitate the development of models with human-like visual and linguistic understanding. However, existing VQA models often exploit superficial statistical biases to produce responses, instead of producing the right answers for the right reasons (Kafle et al., 2019).

The VQA-CP dataset (Agrawal et al., 2018) showcases this phenomenon by incorporating different question type/answer distributions in the train and test sets. Since the linguistic priors in the train and test sets differ, models that exploit these priors fail on the test set. To tackle this issue, recent works have endeavored to enforce proper visual grounding, where the goal is to make models produce answers by looking at relevant visual regions (Gan et al., 2017; Selvaraju et al.,
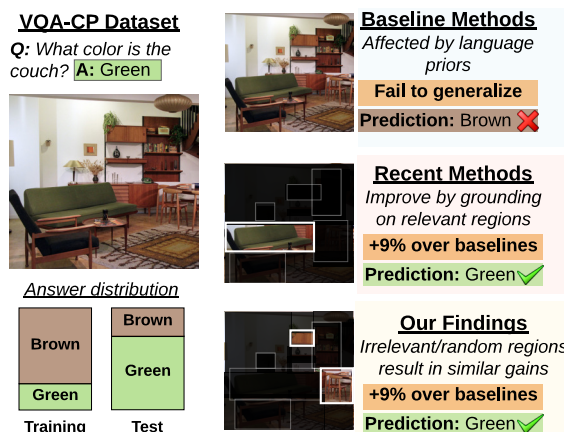

Figure 1: We find that existing visual sensitivity enhancement methods improve performance on VQA-CPv2 through regularization as opposed to proper visual grounding.

2019; Wu and Mooney, 2019), instead of exploiting linguistic priors. These approaches rely on additional annotations/cues such as human-based attention maps (Das et al., 2017), textual explanations (Huk Park et al., 2018) and object label predictions (Ren et al., 2015) to identify relevant regions, and train the model to base its predictions on those regions, showing large improvements (8-10% accuracy) on the VQA-CPv2 dataset.

Here, we study these methods. We find that their improved accuracy does not actually emerge from proper visual grounding, but from regularization effects, where the model *forgets* the linguistic priors in the train set, thereby performing better on the test set. To support these claims, we first show that it is possible to achieve such gains even when the model is trained to look at: a) irrelevant visual regions, and b) random visual regions. Second, we show that differences in the predictions from the

---

[1]https://github.com/erobic/negative_analysis_of_grounding

variants trained with relevant, irrelevant and random visual regions are not statistically significant. Third, we show that these methods degrade performance when the priors remain intact and instead work on VQA-CPv2 by hurting its train accuracy.

Based on these observations, we hypothesize that controlled degradation on the train set allows models to forget the training priors to improve test accuracy. To test this hypothesis, we introduce a simple regularization scheme that zeros out the ground truth answers, thereby always penalizing the model, whether the predictions are correct or incorrect. We find that this approach also achieves near state-of-the-art performance (48.9% on VQA-CPv2), providing further support for our claims.

While we agree that visual grounding is a useful direction to pursue, our experiments show that the community requires better ways to test if systems are actually visually grounded. We make some recommendations in the discussion section.

2 Related Work

2.1 Biases in VQA

As expected of any real world dataset, VQA datasets also contain dataset biases (Goyal et al., 2017). The VQA-CP dataset (Agrawal et al., 2018) was introduced to study the robustness of VQA methods against linguistic biases. Since it contains different answer distributions in the train and test sets, VQA-CP makes it nearly impossible for the models that rely upon linguistic correlations to perform well on the test set (Agrawal et al., 2018; Shrestha et al., 2019).

2.2 Bias Mitigation for VQA

VQA algorithms without explicit bias mitigation mechanisms fail on VQA-CP, so recent works have focused on the following solutions:

2.2.1 Reducing Reliance on Questions

Some recent approaches employ a question-only branch as a control model to discover the questions most affected by linguistic correlations. The question-only model is either used to perform adversarial regularization (Grand and Belinkov, 2019; Ramakrishnan et al., 2018) or to re-scale the loss based on the difficulty of the question (Cadene et al., 2019). However, when these ideas are applied to the UpDn model (Anderson et al., 2018), which attempts to learn correct visual grounding,

these approaches achieve 4-7% lower accuracy compared to the state-of-the-art methods.

2.2.2 Enhancing Visual Sensitivities

Both Human Importance Aware Network Tuning (HINT) (Selvaraju et al., 2019) and Self Critical Reasoning (SCR) (Wu and Mooney, 2019), train the network to be more sensitive towards salient image regions by improving the alignment between visual cues and gradient-based sensitivity scores. HINT proposes a ranking loss between human-based importance scores (Das et al., 2016) and the gradient-based sensitivities. In contrast, SCR does not require exact saliency ranks. Instead, it penalizes the model if correct answers are more sensitive towards non-important regions as compared to important regions, and if incorrect answers are more sensitive to important regions than correct answers.

3 Existing VQA Methods

Given a question $\mathcal{Q}$ and an image $\mathcal{I}$, *e.g.,* represented by bottom-up region proposals: v (Anderson et al., 2018), a VQA model is tasked with predicting the answer a:

$$P(a|\mathcal{Q}, \mathcal{I}) = f_{VQA}(v, \mathcal{Q}). \tag{1}$$

3.1 Baseline VQA Methods

Without additional regularization, existing VQA models such as the baseline model used in this work: UpDn (Anderson et al., 2018), tend to rely on the linguistic priors: $P(a|\mathcal{Q})$ to answer questions. Such models fail on VQA-CP, because the priors in the test set differ from the train set.

3.2 Visual Sensitivity Enhancement Methods

To reduce the reliance on linguistic priors, visual sensitivity enhancement methods attempt to train the model to be more sensitive to relevant visual regions when answering questions. Following (Wu and Mooney, 2019), we define the sensitivity of an answer a with respect to a visual region v_i as:

$$\mathcal{S}(a, v_i) := (\nabla_{v_i} P(a|\mathcal{I}, \mathcal{Q}))^T \mathbf{1}. \tag{2}$$

Existing methods propose the following training objectives to improve grounding using $\mathcal{S}$:

- **HINT** uses a ranking loss, which penalizes the model if the pair-wise rankings of the sensitivities of visual regions towards ground truth answers a_{gt} are different from the ranks computed from the human-based attention maps.

- **SCR** divides the region proposals into influential and non-influential regions and penalizes the model if: 1) $\mathcal{S}(a_{gt})$ of a non-influential region is higher than an influential region, and 2) the region most influential for the correct answer has even higher sensitivity for incorrect answers.

Both methods improve baseline accuracy by 8-10%. Is this actually due to better visual grounding?

4 Why Did the Performance Improve?

We probe the reasons behind the performance improvements of HINT and SCR. We first analyze if the results improve even when the visual cues are irrelevant (Sec. 4.2) or random (Sec. 4.3) and examine if their differences are statistically significant (Sec. 4.4). Then, we analyze the regularization effects by evaluating the performance on VQA-CPv2's train split (Sec. 4.5) and the behavior on a dataset without changing priors (Sec. 4.6). We present a new metric to assess visual grounding in Sec. 4.7 and describe our regularization method in Sec. 5.

4.1 Experimental Setup

We compare the baseline UpDn model with HINT and SCR-variants trained on VQAv2 or VQA-CPv2 to study the causes behind the improvements. We report mean accuracies across 5 runs, where a pre-trained UpDn model is fine-tuned on subsets with human attention maps and textual explanations for HINT and SCR respectively. Further training details are provided in the Appendix.

4.2 Training on Irrelevant Visual Cues

In our first experiment we studied how irrelevant visual cues performed compared to relevant ones. We fine-tune the model with irrelevant cues defined as: $\mathcal{S}_{irrelevant} := (1 - \mathcal{S}_h)$, where, $\mathcal{S}_h$ represents the human-based importance scores. As shown in the 'Grounding using irrelevant cues' section of Table 1, both HINT and SCR are within 0.3% of the results obtained from looking at relevant regions, which indicates the gains for HINT and SCR are not necessarily from looking at relevant regions.

4.3 Training on Random Visual Cues

In our next experiment we studied how random visual cues performed with HINT and SCR. We assign random importance scores to the visual regions: $\mathcal{S}_{rand} \sim uniform(0, 1)$. We test two variants of randomness: **Fixed random regions**, where

Table 1: Results on VQA-CPv2 and VQAv2 datasets for the baseline UpDn, visual sensitivity enhancement methods (HINT and SCR) and our own regularization method, including the published (pub.) numbers.

	VQA-CPv2		VQAv2	
	Train	Test	Train	Val
Baseline - Without visual grounding				
UpDn	84.0	40.1	83.4	64.4
Grounding using human-based cues				
HINT$_{pub.}$	N/A	46.7	N/A	63.4[1]
SCR$_{pub.}$	N/A	**49.5**	N/A	62.2
HINT	73.9	48.2	75.7	61.3
SCR	75.9	49.1	77.9	61.3
Grounding using irrelevant cues				
HINT	71.2	48.0	73.5	60.3
SCR	75.7	49.2	74.1	59.1
Grounding using fixed random cues				
HINT	72.0	48.1	73.0	59.5
SCR	70.0	49.1	78.0	61.4
Grounding using variable random cues				
HINT	71.9	48.1	72.9	59.4
SCR	69.6	49.2	78.1	61.5
Regularization by zeroing out answers				
Ours$_{1\% \ fixed}$	**78.0**	48.9	**80.1**	**62.6**
Ours$_{1\% \ var.}$	77.6	48.5	80.0	**62.6**
Ours$_{100\%}$	75.7	48.2	79.9	62.4

[1] The published number is a result of fine-tuning HINT on the entire training set, but as described in Sec. 4.6, other published numbers and our experiments fine-tune only on the instances with cues.

$\mathcal{S}_{rand}$ are fixed once chosen, and **Variable random regions**, where $\mathcal{S}_{rand}$ are regenerated every epoch. As shown in Table 1, both of these variants obtain similar results as the model trained with human-based importance scores. The performance improves even when the importance scores are changed every epoch, indicating that it is not even necessary to look at the *same* visual regions.

4.4 Significance of Statistical Differences

To test if the changes in results were statistically significant, we performed Welch's t-tests (Welch, 1938) on the predictions of the variants trained on relevant, irrelevant and random cues. We pick Welch's t-test over the Student's t-test, because the latter assumes equal variances for predictions from different variants. To perform the tests, we first randomly sample 5000 subsets of non-overlapping test instances. We then average the accuracy of each subset across 5 runs, obtaining 5000 values. Next, we run the t-tests for HINT and SCR separately on the subset accuracies. As shown in Table 2, the *p*-values across the variants of HINT and SCR are

Table 2: p-values from the Welch's t-tests and the percentage of overlap between the predictions (Ovp.) of different variants of HINT and SCR.

Methods	p	Ovp.(%)
HINT variants against Baseline		
Default vs. Baseline	0.0	83.6
Irrelevant vs. Baseline	0.0	82.4
Fixed Random vs. Baseline	0.0	82.0
Variable Random vs. Baseline	0.0	81.5
Among HINT variants		
Default vs Irrelevant	0.3	89.7
Default vs Fixed random	0.7	90.9
Default vs Variable random	0.6	91.9
Irrelevant vs Fixed random	0.5	95.6
Irrelevant vs Variable random	0.7	93.9
Fixed random vs Variable random	0.9	96.9
SCR variants against Baseline		
Default vs. Baseline	0.0	85.6
Irrelevant vs. Baseline	0.0	84.2
Fixed Random vs. Baseline	0.0	80.7
Variable Random vs. Baseline	0.0	80.6
Among SCR variants		
Default vs Irrelevant	0.6	92.0
Default vs Fixed random	0.8	89.3
Default vs Variable random	0.6	89.5
Irrelevant vs Fixed random	0.4	91.7
Irrelevant vs Variable random	1.0	91.6
Fixed random vs Variable random	0.4	96.7

greater than or equal to 0.3. Using a confidence level of 95% ($\alpha = 0.05$), we fail to reject the null hypothesis that the mean difference between the paired values is 0, showing that the variants are not statistically significantly different from each other. We also compare the predictions of HINT/SCR against baseline, and find that p-values are all zeros, showing that the differences have statistical significance.

Percentage of Overlaps: To further check if the variants trained on irrelevant or random regions gain performance in a manner similar to the models trained on relevant regions, we compute the overlap between their predictions on VQA-CPv2's test set. The percentage of overlap is defined as:

$$\% \, Overlap = \frac{n_{same}}{n_{total}} \times 100\%,$$

where, n_{same} denotes the number of instances where either both variants were correct or both were incorrect and n_{total} denotes the total number of test instances. As shown in Table 2, we compare $\%Overlap$ between different variants of HINT/SCR with baseline and against each other.
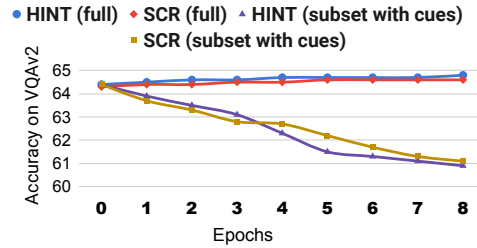

Figure 2: Accuracies for HINT and SCR on VQAv2's val set, when fine-tuned either on the full train set or on the subset containing visual cues.

We find $89.7 - 91.9\%$ and $89.5 - 92.0\%$ overlaps for different variants of HINT and SCR respectively. These high overlaps suggest that the variants are not working in fundamentally different manners.

4.5 Drops in Training Accuracy

We compare the training accuracies to analyze the regularization effects. As shown in Table 1, the baseline method has the highest training results, while the other methods cause $6.0 - 14.0\%$ and $3.3 - 10.5\%$ drops in the training accuracy on VQA-CPv2 and VQAv2, respectively. We hypothesize that degrading performance on the train set helps forget linguistic biases, which in turn helps accuracy on VQA-CPv2's test set but hurts accuracy on VQAv2's val set.

4.6 Drops in VQAv2 Accuracy

As observed by Selvaraju et al. (2019) and as shown in Fig. 2, we observe small improvements on VQAv2 when the models are fine-tuned on the entire train set. However, if we were to compare against the improvements in VQA-CPv2 in a fair manner, i.e., only use the instances with visual cues while fine-tuning, then, the performance on VQAv2 drops continuously during the course of the training. This indicates that HINT and SCR help forget linguistic priors, which is beneficial for VQA-CPv2 but not for VQAv2.

4.7 Assessment of Proper Grounding

In order to quantitatively assess visual grounding, we propose a new metric called: Correctly Predicted but Improperly Grounded (CPIG):

$$\%CPIG = \frac{N_{\text{correct ans, improper grounding}}}{N_{\text{correct ans}}} \times 100\%,$$

which is the number instances for which the most sensitive visual region used to correctly predict the

answer is not within top-3 most relevant ground truth regions, normalized by the total number of correct predictions. HINT and SCR trained on relevant regions obtained lower CPIG values that other variants (70.24% and 80.22% respectively), indicating they are better than other variants at finding relevant regions. However, these numbers are still high, and show that only 29.76% and 19.78% of the correct predictions for HINT and SCR were properly grounded. Further analysis is presented in the Appendix.

5 Embarrassingly Simple Regularizer

The usage of visual cues and sensitivities in existing methods is superfluous because the results indicate that performance improves through degradation of training accuracy. We hypothesize that simple regularization that does not rely on cues or sensitivities can also achieve large performance gains for VQA-CP. To test this hypothesis, we devise a simple loss function which continuously degrades the training accuracy by training the network to always predict a score of zero for all possible answers i.e. produce a zero vector ($\mathbf{0}$). The overall loss function can be written as:

$$L := BCE(P(\mathcal{A}), \mathcal{A}_{gt}) + \lambda BCE(P(\mathcal{A}), \mathbf{0}),$$

where, BCE refers to the binary cross entropy loss and $P(\mathcal{A})$ is a vector consisting of predicted scores for all possible answers. The first term is the binary cross entropy loss between model predictions and ground truth answer vector ($\mathcal{A}_{gt}$), and the second term is our regularizer with a coefficient of $\lambda = 1$. Note that this regularizer continually penalizes the model during the course of the training, whether its predictions are correct or incorrect.

As shown in Table 1, we present results when this loss is used on: a) Fixed subset covering 1% of the dataset, b) Varying subset covering 1% of the dataset, where a new random subset is sampled every epoch and c) 100% of the dataset. Confirming our hypothesis, all variants of our model achieve near state-of-the-art results, solidifying our claim that the performance gains for recent methods come from regularization effects.

It is also interesting to note that the drop in training accuracy is lower with this regularization scheme as compared to the state-of-the-art methods. Of course, if any model was actually visually grounded, then we would expect it to improve performances on both train and test sets. We do not observe such behavior in any of the methods, indicating that they are not producing right answers for the right reasons.

6 Discussion on Proper Grounding

While our results indicate that current visual grounding based bias mitigation approaches do not suffice, we believe this is still a good research direction. However, future methods must seek to verify that performance gains are not stemming from spurious sources by using an experimental setup similar to that presented in this paper. We recommend that both train and test accuracy be reported, because a model truly capable of visual grounding would not cause drastic drops in training accuracy to do well on the test sets. Finally, we advocate for creating a dataset with ground truth grounding available for 100% of the instances using synthetically generated datasets (Kafle et al., 2017; Kafle and Kanan, 2017; Kafle et al., 2018; Acharya et al., 2019b; Hudson and Manning, 2019; Johnson et al., 2017), enabling the community to evaluate if their methods are able to focus on relevant information. Another alternative is to use tasks that explicitly test grounding, e.g., in visual query detection an agent must output boxes around any regions of a scene that match the natural language query (Acharya et al., 2019a).

7 Conclusion

Here, we showed that existing visual grounding based bias mitigation methods for VQA are not working as intended. We found that the accuracy improvements stem from a regularization effect rather than proper visual grounding. We proposed a simple regularization scheme which, despite not requiring additional annotations, rivals state-of-the-art accuracy. Future visual grounding methods should be tested with a more comprehensive experimental setup and datasets for proper evaluation.

References

Manoj Acharya, Karan Jariwala, and Christopher Kanan. 2019a. VQD: Visual query detection in natural scenes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1955–1961, Minneapolis, Minnesota. Association for Computational Linguistics.

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019b. Tallyqa: Answering complex counting questions. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980.

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

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

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 839–850.

Abhishek Das, Harsh Agrawal, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

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

Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1811–1820.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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 (CVPR)*, volume 1, page 3.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 1–13, Minneapolis, Minnesota. Association for Computational Linguistics (ACL).

Drew A Hudson and Christopher D Manning. 2019. 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 (CVPR)*, pages 6700–6709.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8779–8788.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997. IEEE.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1983–1991. IEEE.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, pages 198–202.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1541–1551.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bernard L Welch. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362.

Jialin Wu and Raymond Mooney. 2019. Self-critical reasoning for robust visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8601–8611.

A Appendix

A.1 Training Details

We compare four different variants of HINT and SCR to study the causes behind the improvements including the models that are fine-tuned on: 1) relevant regions (state-of-the-art methods) 2) irrelevant regions 3) fixed random regions and 4) variable random regions. For all variants, we fine-tune a pre-trained UpDn, which was trained on either VQA-CPv2 or VQAv2 for 40 epochs with a learning rate of 10^{-3}. When fine-tuning with HINT, SCR or our method, we also use the main binary cross entropy VQA loss, whose weight is set to 1. The batch size is set to 384 for all of the experiments.

HINT

Following (Selvaraju et al., 2019), we train HINT on the subset with human-based attention maps (Das et al., 2017), which are available for 9% of the VQA-CPv2 train and test sets. The same subset is used for VQAv2 too. The learning rate is set to 2×10^{-5} and the weight for the HINT loss is set to 2.

SCR

Since (Wu and Mooney, 2019) reported that human-based textual explanations (Huk Park et al., 2018) gave better results than human-based attention maps for SCR, we train all of the SCR variants on the subset containing textual explanation-based cues. SCR is trained in two phases. For the first phase, which strengthens the influential objects, we use a learning rate of 5×10^{-5}, loss weight of 3

Table A3: Results on VQA-CPv2 and VQAv2 datasets for the baseline UpDn, visual sensitivity enhancement methods (HINT and SCR) and our own regularization method, including the published (pub.) numbers.

	VQA-CPv2	VQAv2
Baseline - Without visual grounding		
UpDn	0.0110	0.0155
Grounding using human-based cues		
HINT	0.1020	0.1350
SCR	0.0340	-0.0670
Grounding using irrelevant cues		
HINT	-0.0048	-0.0200
SCR	0.0580	-0.0100
Grounding using fixed random cues		
HINT	0.0510	0.0620
SCR	-0.0250	-0.0350
Grounding using variable random cues		
HINT	0.0570	0.0623
SCR	-0.0380	0.0246
Regularization by zeroing out answers		
$\text{Ours}_{1\%\ fixed}$	-0.1050	-0.1200
$\text{Ours}_{100\%}$	-0.0750	-0.0100

and train the model to a maximum of 12 epochs. Then, following (Wu and Mooney, 2019), for the second phase, we use the best performing model from the first phase to train the second phase, which criticizes incorrect dominant answers. For the second phase, we use a learning rate of 10^{-4} and weight of 1000, which is applied alongside the loss term used in the first phase. The specified hyperparameters worked better for us than the values provided in the original paper.

Our Zero-Out Regularizer

Our regularization method, which is a binary cross entropy loss between the model predictions and a zero vector, does not use additional cues or sensitivities and yet achieves near state-of-the-art performance on VQA-CPv2. We set the learning rate to: $\frac{2 \times 10^{-6}}{r}$, where r is the ratio of the training instances used for fine-tuning. The weight for the loss is set to 2. We report the performance obtained at the 8th epoch.

A.2 Results

Correlation with Ground Truth Visual Cues

Following (Selvaraju et al., 2019), we report Spearman's rank correlation between network's sensitivity scores and human-based scores in Table A3. For HINT and our zero-out regularizer, we use human-based attention maps. For SCR, we use textual explanation-based scores. We find that HINT

trained on human attention maps has the highest correlation coefficients for both datasets. However, compared to baseline, HINT variants trained on random visual cues also show improved correlations. For SCR, we obtain surprising results, with the model trained on irrelevant cues obtaining higher correlation than that trained on relevant visual cues. As expected, applying our regularizer does not improve rank correlation. Since HINT trained on relevant cues obtains the highest correlation values, it does indicate improvement in visual grounding. However, as we have seen, the improvements in performance cannot necessarily be attributed to better overlap with ground truth localizations.

A Note on Qualitative Examples

Presentation of qualitative examples in visual grounding models for VQA suffers from confirmation bias i.e., while it is possible to find qualitative samples that look at relevant regions to answer questions properly, it is also possible to find samples that produce correct answers without looking at relevant regions. We present examples for such cases in Fig. A3. We next present a quantitative assessment of visual grounding, which does not suffer from the confirmation bias.

Quantitative Assessment of Grounding

In order to truly assess if existing methods are using relevant regions to produce correct answers, we use our proposed metric: Correctly Predicted but Improperly Grounded (CPIG). If the CPIG values are large, then it implies that large portion of correctly predicted samples were not properly grounded. Fig. A4 shows % CPIG for different variants of HINT trained on human attention-based cues, whereas Fig. A5 shows the metric for different variants of SCR trained on textual explanation-based cues. We observe that HINT and SCR trained on relevant regions have the lowest % CPIG values (70.24% and 80.22% respectively), indicating that they are better than other variants in finding relevant regions. However, only a small percentage of correctly predicted samples were properly grounded (29.76% and 19.78% for HINT and SCR respectively), even when trained on relevant cues.

Breakdown by Answer Types

Table A4 shows VQA accuracy for each answer type on VQACPv2's test set. HINT/SCR and our regularizer show large gains in 'Yes/No' questions.

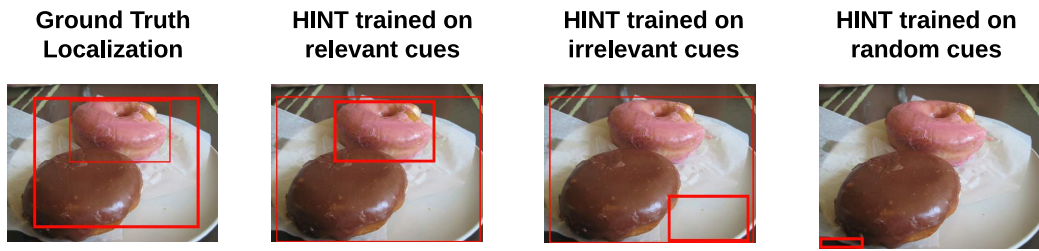Table A4: VQA accuracy per answer-type on VQACPv2 test set.

	Overall	Yes/No	Num	Other
Baseline - Without visual grounding				
UpDn	40.1	41.1	12.0	47.2
Grounding using human-based cues				
HINT	48.2	65.2	13.8	47.5
SCR	49.1	70.3	11.5	**48.0**
Grounding using irrelevant cues				
HINT	48.0	67.2	13.5	47.1
SCR	**49.2**	73.4	11.5	46.4
Grounding using fixed random cues				
HINT	48.1	66.9	13.8	46.9
SCR	49.1	**74.7**	12.2	45.1
Grounding using variable random cues				
HINT	48.1	67.1	**13.9**	46.9
SCR	49.2	**74.7**	12.2	45.1
Regularization by zeroing out answers				
Ours$_{1\% \ fixed}$	48.9	69.8	11.3	47.8
Ours$_{100\%}$	48.2	66.7	11.7	47.9

We hypothesize that the methods help forget linguistic priors, which improves test accuracy of such questions. In the train set of VQACPv2, the answer 'no' is more frequent than the answer 'yes', tempting the baseline model to answer 'yes/no' questions with 'no'. However, in the test set, answer 'yes' is more frequent. Regularization effects caused by HINT/SCR and our method cause the models to weaken this prior i.e., reduce the tendency to just predict 'no', which would increase accuracy at test because 'yes' is more frequent in the test set. Next, all of the methods perform poorly on 'Number (Num)' answer type, showing that methods find it difficult to answer questions that are most reliant on correct visual grounding such as: localizing and counting objects. Finally, we do not observe large improvements in 'Other' question type, most likely due to the large number of answers present under this answer type.

Accuracy versus Size of Train Set

We test our regularization method on random subsets of varying sizes. Fig. A6 shows the results when we apply our loss to $1 - 100\%$ of the training instances. Clearly, the ability to regularize the model does not vary much with respect to the size of the train subset, with the best performance occurring when our loss is applied to 1% of the training instances. These results support our claims that it is possible to improve performance without actually performing visual grounding.

| Ground Truth Localization | HINT trained on relevant cues | HINT trained on irrelevant cues | HINT trained on random cues |

Q: Is this food sweet? A: yes
Remarks: The most sensitive regions for irrelevant/random variants do not contain food, yet their answers are correct.

Q: What is the swimmer doing? A: surfing
Remarks: Models trained on irrelevant/random cues do not look at the swimmer at all, yet produce correct answer.
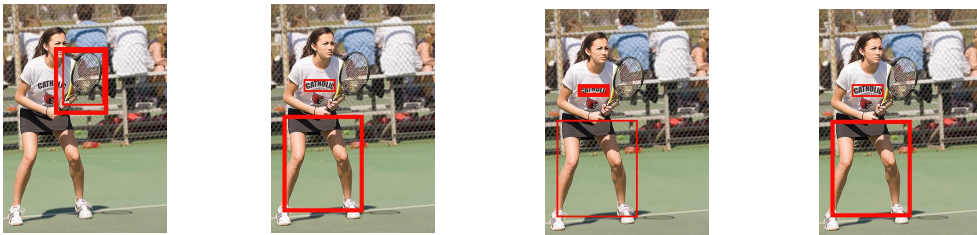
Q: Is the sport being played tennis or volleyball? A: tennis
Remarks: None of the variants look at relevant regions, and yet produce correct answer.

Q: Has the boy worn out his jeans? A: yes
Remarks: All of the variants look at both relevant and irrelevant regions to produce correct answer.

Figure A3: Visualizations of most sensitive visual regions used by different variants of HINT to make predictions. We pick samples where all variants produce correct response to the question. The first column shows ground truth regions and columns 2-4 show visualizations from HINT trained on relevant, irrelevant and fixed random regions respectively.
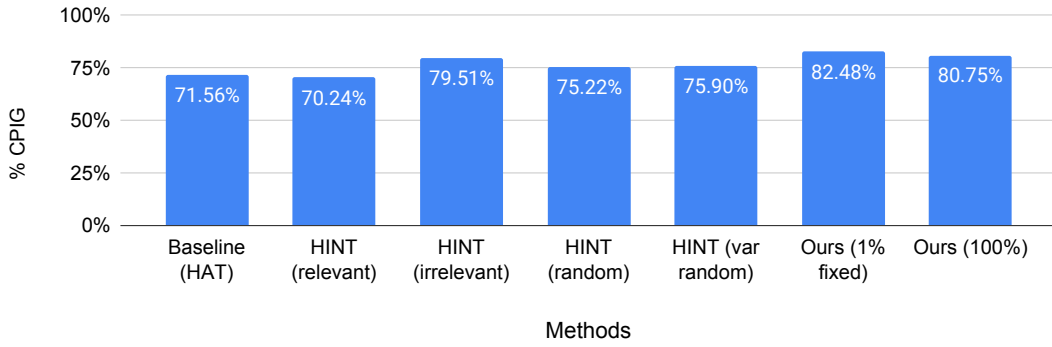
Figure A4: % CPIG for baseline and different variants of HINT and our method, computed using ground truth relevant regions taken from human attention maps (lower is better).
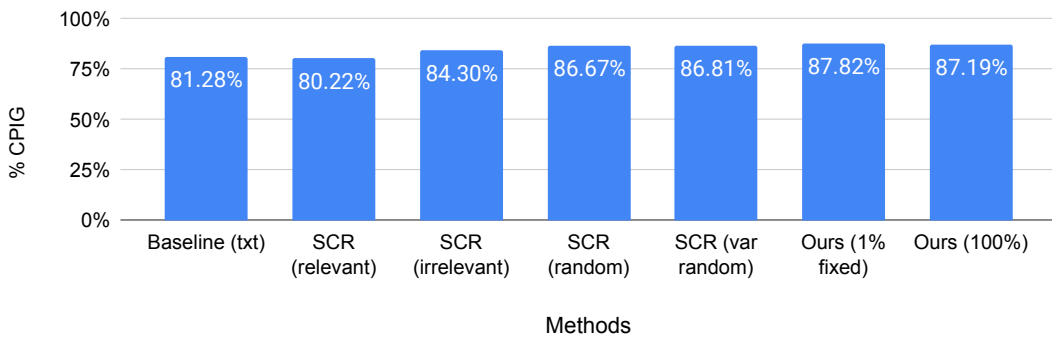

Figure A5: % CPIG for baseline and different variants of SCR and our method, computed using ground truth relevant regions taken from textual explanations (txt).
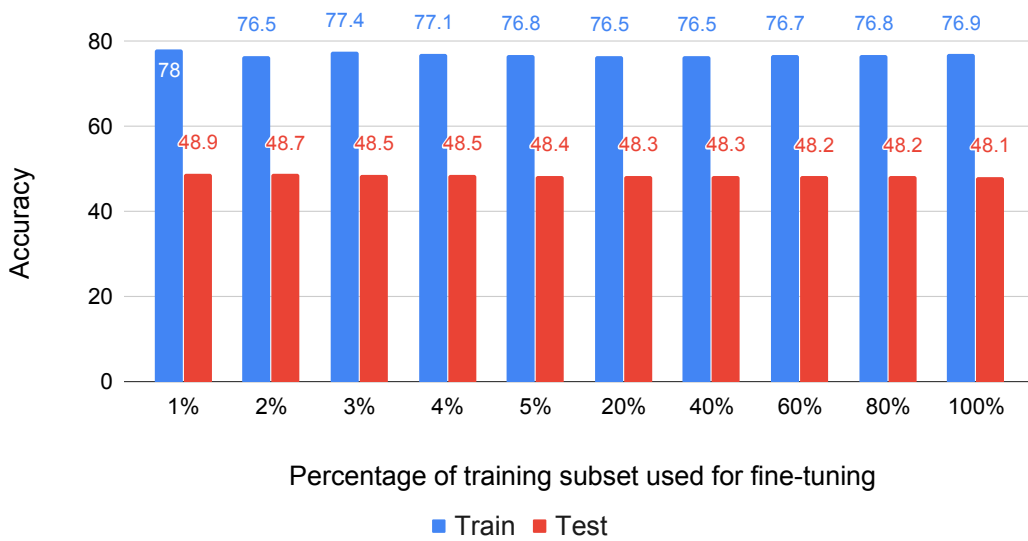

Figure A6: The regularization effect of our loss is invariant with respect to the dataset size.