Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing

Vedika Agarwal[1,3*] Rakshith Shetty[1] Mario Fritz[2]

[1]Max Planck Institute for Informatics
Saarland Informatics Campus
[2] CISPA Helmholtz Center for Information Security
Saarland Informatics Campus
[3]TomTom

vedika.agarwal@tomtom.com rakshith.shetty@mpi-inf.mpg.de fritz@cispa.saarland

Abstract

Despite significant success in Visual Question Answering (VQA), VQA models have been shown to be notoriously brittle to linguistic variations in the questions. Due to deficiencies in models and datasets, today's models often rely on correlations rather than predictions that are causal w.r.t. relevant evidence. In this paper, we propose a novel way to analyze and measure the robustness of the state of the art models w.r.t semantic visual variations as well as propose ways to make models more robust against spurious correlations. Our method performs automated semantic image manipulations and tests for consistency in model predictions to quantify the model robustness as well as generate synthetic data to counter these problems. We perform our analysis on three diverse, state of the art VQA models and diverse question types with a particular focus on challenging counting questions. In addition, we show that models can be made significantly more robust against inconsistent predictions using our edited data. Finally, we show that results also translate to real-world error cases of state of the art models, which results in improved overall performance.

1. Introduction

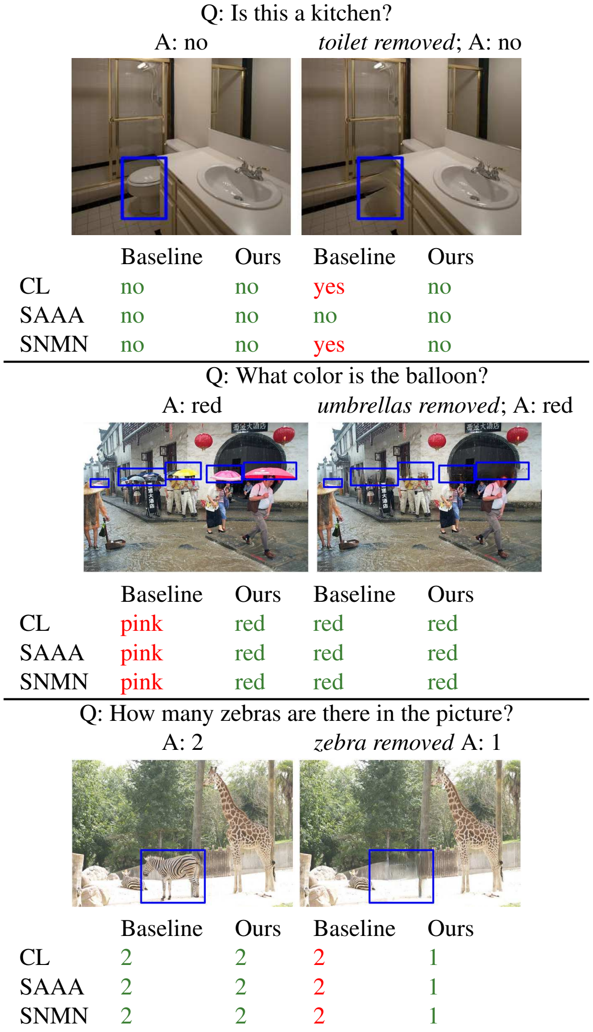
VQA allows interaction between images and language, with diverse applications such as interacting with chat bots to assisting visually impaired people. In these applications we expect a model to answer truthfully and based on the evidence in the image and the actual intention of the question. Unfortunately, this is not always the case even for state of the art methods. Instead of "sticking to the facts", models frequently rely on spurious correlations and follow biases induced by data and/or model. For instance, recent works [27, 26] have shown that the VQA models are brittle to linguistic variations in questions/answers. Shah *et al.* in [27] introduced VQA-Rephrasings dataset to expose the brittleness of the VQA models to linguistic variations and proposed cyclic consistency to improve their robustness. They show that if a model answers 'Yes' to the question: 'Is it safe to turn left?', it answers 'No' when the question is rephrased to 'Can one safely turn left?'. Similarly Ray *et al.* in [26] introduced ConVQA to quantitatively evaluate the consistency for VQA towards different generated entailed questions and proposed data augmentation module to make the models more consistent.

While previous works have studied linguistic modifications, our contribution is the first systematic study of automatic visual content manipulations at scale. Analogous to rephrasing questions for VQA, images can also be semantically edited to create different variants where the same question-answer (QA) pair holds. One sub-task of this broader semantic editing goal is object removal. One can remove objects in such a way that the answer remains invariant (wherein only objects irrelevant to the QA are removed) as shown in Figure 1 (top/middle). Alternately one could also make covariant edits where we remove the object mentioned in the QA and hence expect the answer to change in a predictable manner as shown in Figure 1 (bottom). We explore both invariant and covariant forms of editing and quantify how consistent models are under these edits.

We employ a GAN-based [28] re-synthesis model to automatically remove objects. Our data generation technique helps us create exact complementary pairs of the image as shown in Figures 1, 2. We pick three recent models which represent different approaches to VQA to analyze robustness: a simple CNN+LSTM (CL) model, an attention-based model (SAAA [17]) and a compositional model (SNMN [11]). We show that all the three models are brittle to semantic variations in the image, revealing the false correlation that the models exploit to predict the answer. Furthermore, we show that training data augmentation with our synthetic set can improve models robustness.

Our motivation to create this complementary dataset

[*] Work done at Max Planck Institute for Informatics.



Q: Is this a kitchen?

A: no | *toilet removed*; A: no

	Baseline	Ours	Baseline	Ours
CL	no	no	yes	no
SAAA	no	no	no	no
SNMN	no	no	yes	no

Q: What color is the balloon?

A: red | *umbrellas removed*; A: red

	Baseline	Ours	Baseline	Ours
CL	pink	red	red	red
SAAA	pink	red	red	red
SNMN	pink	red	red	red

Q: How many zebras are there in the picture?

A: 2 | *zebra removed* A: 1

	Baseline	Ours	Baseline	Ours
CL	2	2	2	1
SAAA	2	2	2	1
SNMN	2	2	2	1

Figure 1: VQA models change their predictions as they exploit spurious correlations rather than causal relations based on the evidence. Shown above are predictions of 3 VQA models on original and synthetic images from our proposed IV-VQA and CV-VQA datasets. 'Ours' denote the models robustified with our proposed data augmentation strategy.

stems from the desire to study how accurate and consistent different VQA models are and to improve the models by the generated 'complementary' data (otherwise not available in the dataset). While data augmentation and cyclic consistency are making the VQA models more robust [16, 26, 27] towards the natural language part, we take a step forward to make the models consistent to semantic variations in the images. We summarize our main contributions as follows:

- We propose a novel approach to analyze and quantify issues of VQA models due to spurious correlation and biases of data and models. We use synthetic data to quantify these problems with a new metric that measures erroneous inconsistent predictions of the model.
- We contribute methodology and a synthetic dataset [1] that complements VQA datasets by systematic variations that are generated by our semantic manipulations. We complement this dataset by a human study that validates our approach and provides additional human annotations.
- We show how the above-mentioned issues can be reduced by a data augmentation strategy - similar to adversarial training. We present consistent results across a range of questions and three state of the art VQA methods and show improvements on synthetic as well as real data.
- While we investigate diverse question types, we pay particular attention to counting by creating an covariant edited set and show that our data augmentation technique can also improve counting robustness in this setting.

2. Related Work

Visual Question Answering. There has been growing interest in VQA [15, 29] recently, which can be attributed to the availability of large-scale datasets [7, 14, 3, 23, 4] and deep learning driven advances in both vision and NLP. There has been immense progress in building VQA models [20, 24, 22, 6] using LSTMs [9] and convolutional networks [18, 8] to models that span different paradigms such as attention networks [21, 17, 30] and compositional module networks [2, 12, 11, 13]. In our work, we pick a representative model from each of these three design philosophies and study their robustness to semantic visual variations.

Robustness in VQA. Existing VQA models often exploit language and contextual priors to predict the answers [31, 25, 7, 1]. To understand how much do these models actually see and understand, various works have been proposed to study the robustness of models under different variations in the input modalities. [1] shows that changing the prior distributions for the answers across training and test sets significantly degrades models' performance. [26, 27] study the robustness of the VQA models towards linguistic variations in the questions. They show how different re-phrasings of the questions can cause the model to switch their answer predictions. In contrast, we study the robustness of VQA models to semantic manipulations in the image and propose a data augmentation technique to make the models robust.

Data Augmentation for VQA. Data Augmentation has been used in VQA to improve model's performance either in the context of accuracy [16] or robustness against linguistic variations in questions [26, 27]. [16] generated new questions by using existing semantic annotations and a generative approach via recurrent neural network. They showed that augmenting these questions gave a boost of around 1.5% points in accuracy. [27] propose a cyclic-consistent training scheme where they generate different rephrasings

[1] https://rakshithshetty.github.io/CausalVQA/

of question (based on answer predicted by the model) and train the model such that answer predictions across the generated and the original question remain consistent. [26] proposes a data augmentation module that automatically generates entailed (or similar-intent) questions for a source QA pair and fine-tunes the VQA model if the VQA's answer to the entailed question is consistent with the source QA pair.

3. Synthetic Dataset for Variances and Invariances in VQA

While robustness w.r.t linguistic variations [27, 26] and changes in answer distributions [1] have been studied, we explore how robust VQA models are to semantic changes in the images. For this, we create a synthetic dataset by removing objects irrelevant and relevant to the QA pairs and propose consistency metrics to study the robustness. Our dataset is built upon existing VQAv2 [7] and MS-COCO [19] datasets. We target the 80 object categories present in the COCO dataset [19] and utilize a GAN-based [28] resynthesis technique to remove them. The first key step in creating this dataset is to select a candidate object for removal for each Image-Question-Answer (IQA) pair. Next, since we use an in-painter-based GAN, we need to ensure the removal of the object does not affect the quality of the image or QA in any manner. We introduce vocabulary mapping to take care of the former and area-overlapping criteria for the latter. We discuss these steps in detail to generate the edited set in irrelevant removal setting and later extend these to relevant object removal.

3.1. InVariant VQA (IV-VQA)

For the creation of this dataset, we select and remove the objects irrelevant to answering the question. Hence the model is expected to make the same predictions on the edited image. A change in the prediction would expose the spurious correlations that the model is relying on to answer the question. Some examples of the semantically edited images along with the original images can be seen in Figures 1, 2. For instance, in Figure 2 (top-right), for the question about the color of the surfboard, removing the person should not influence the model's prediction. In order to generate the edited image, we first need to identify person as a potential candidate which in turn requires studying the objects present in the image and the ones mentioned in the QA. Since we use VQA v2 dataset [7], where all the images overlap with MS-COCO [19], we can access the ground-truth bounding box and segmentation annotations for each image. In total, there are 80 different object classes in MS-COCO which become our target categories for removal.

Vocabulary mapping. To decide if we can remove an object, we need to first map all the object referrals in question and answer onto the 80 COCO categories. These categories are often addressed in the QA space by many synonyms or a subset representative of that class. For example- people, person, woman, man, child, he, she, biker all refer to the category: 'person'; bike, cycle are commonly used for the class 'bicycle'. To avoid erroneous removals, we create an extensive list mapping nouns/pronouns/synonyms used in the QA vocabulary to the 80 COCO categories. Table 1 shows a part of the object mapping list. The full list can be found in code-release for the project [2].

COCO categories	Additional words mapped
person	man, woman, player, child, girl, boy people, lady, guy, kid, he etc
fire hydrant	hydrant, hydrate, hydra
wine glass	wine, glass, bottle, beverage, drink
donut	doughnut, dough, eating, food, fruit
chair	furniture, seat
...	...

Table 1: Example of vocabulary mapping from QA space to COCO categories. If any of these words (in the right column) occur in the QA, these words are mapped to the corresponding COCO category (in the left column).

Let O_I represent the objects in the images (known via COCO segmentations), O_{QA} represent the objects in the question-answer (known after vocabulary mapping). Then our target object for removal, O_{target}, is given by $O_I - \{O_I \cap O_{QA}\}$. We assume that if the object is not mentioned in the QA, it is not relevant and hence can be safely removed.

Area-Overlap threshold. The next step is to make sure that the removal of O_{target} does not degrade the quality of the image or affect the other objects mentioned in the QA. Since we use an in-painter based GAN [28], we find that larger object removal is harder to in-paint leaving the images heavily distorted. In order to avoid such distorted images, we only remove the object if the area occupied by its largest instance is less than 10% of the image area. Furthermore, we also consider if the object being removed overlaps in any manner with the object that is mentioned in the QA. We quantitatively measure the overlap score as shown in Equation 1 where M_O denotes the dilated ground truth segmentation mask of all the instances of the object. We only remove the object if the overlap score is less than 10%.

$$\text{Overlap score}(O_{\text{target}}, O_{\text{QA}}) = \frac{(M_O)^{\text{target}} \cap (M_O)^{\text{QA}}}{(M_O)^{\text{QA}}} \tag{1}$$

Uniform Ground-Truth. Finally, we only aim to target those IQAs which have uniform ground-truth answers. In VQA v2 [7], all the questions have 10 answers, while it is good to capture diversity in open-ended question-answering, it also introduces ambiguity, especially in case

[2] https://github.com/rakshithShetty/CausalVQA

	IV-VQA			CV-VQA		
#IQA	train	val	test	train	val	test
real	148013	7009	63219	18437	911	8042
realNE	42043	2152	18143	13035	648	5664
edit	256604	11668	108239	8555	398	3743

Table 2: IV-VQA and CV-VQA distribution. Real refers to VQA [7] IQAs with uniform answers, realNE refers to IQAs for which no edits are possible (after vocabulary mapping and area-overlap threshold), edit refers to the edited IQA. We split the VQA val into 90:10 ratio, where the former is used for testing purpose and latter for validation.

of counting and binary question types. To avoid this ambiguity in our robustness evaluation, we build our edited set by only selecting to semantically manipulate those IQs which have a uniform ground truth answer.

Finally, we remove all the instances of the target object from the image for those IQAs which satisfy the above criteria using the inpainter GAN [28]. We call our edited set as IV-VQA as removal of objects does not lead to any change in answer, the answer is invariant to the semantic editing. Table 2 shows the number of edited IQAs in IV-VQA. While our algorithm involves both manually curated heuristics to select the objects to remove, and a learned in-painter-based GAN model to perform the removal, the whole pipeline is fully automatic. This allows us to apply it to the large-scale VQA dataset with 658k IQA triplets.

Validation by Humans. We get a subset (4.96k IQAs) of our dataset validated by three humans. The subset is selected based on an inconsistency analysis of 3 models covered in the next Section 4. Every annotator is shown the edited IQA and is asked to say if the answer shown is correct for the given image and question (yes/no/ambiguous). According to the study, 91% of the time all the users agree that our edited IQA holds. More details about the study are in the supplementary material (section A.2).

3.2. CoVariant VQA (CV-VQA)

An alternate way of editing images is to target the object in the question. Object-specific questions like counting, color, whether the object is present or not in the image are suitable for this type of editing. We choose counting questions where we generate complementary images with one instance of the object removed. If the model can count n instances of an object in the original image, it should also be able to count $n-1$ instances of the same object in the edited image. Next, we will describe how to generate this covariant data for counting.

First, we collect all the counting questions in the VQA set: selecting questions which contained words ‘many’ and ‘number of’ and which had numeric answers. Next, we focus on removing instances of the object which is to be counted in the question. Vocabulary mapping is used to identify the object mentioned in the question O_Q. Then only those images are retained where the number of the target object instances according to COCO segmentations match the IQA ground-truth answer A given by 10 human annotators.

For the generation of this set, we use the area threshold as 0.1, we only intend to remove the instance if it occupies less than 10% of the image. Furthermore for overlap, since we do not want the removed instance to interfere with the other instances of the object, two masks considered to measure the score are: (1). dilated mask of instance to be removed (2). dilated mask of all the other instances of the object. The object is only removed if the overlap score is zero.

We call our edited set as CV-VQA since removal of the object leads to a covariant change in answer. Table 2 shows the number of edited IQAs in VQA-CV. Figure 2 (bottom row) shows a few examples from our edited set. We only target one instance at a time. More such visual examples can be found in supplementary (section B.2) .

4. Consistency analysis

The goal of creating edited datasets is to gauge how consistent are the models to semantic variations in the images. In IV-VQA, where we remove objects irrelevant to the QA from the image, we expect the models predictions to remain unchanged. In CV-VQA, where one of the instances to be counted is removed, we expect the predicted answer to reduce by one as well. Next, we briefly cover the models' training and then study their performances both in terms of accuracy and consistency. We propose consistency metrics based on how often the models flip their answers and study the different type of flips qualitatively and quantitatively.

VQA models and training. For comparison and analysis, we select three models from the literature, each representing a different design paradigm: a simple CNN+LSTM (CL) model, an attention-based model (SAAA [17]) and a compositional model (SNMN [11]). We use the official code for training the SNMN [11] model, [10]. SAAA [17] is trained using the code available online [32]. We modified this SAAA code in order to get CL model by removing the attention layers from the network. As we use the VQA v2 val split for consistency evaluation and testing, the models are trained using only the train split. Further details of these models and hyper-parameters used can be found in the supplementary (section B.1). Table 3 shows the accuracy scores on VQA v2 val set for models trained by us along with similar design philosophy models benchmarked in [1] and [7]. The models chosen by us exceed the performance of other models within the respective categories.

Consistency. The edited data is created to study the robustness of the models. Since we modify the images in con-

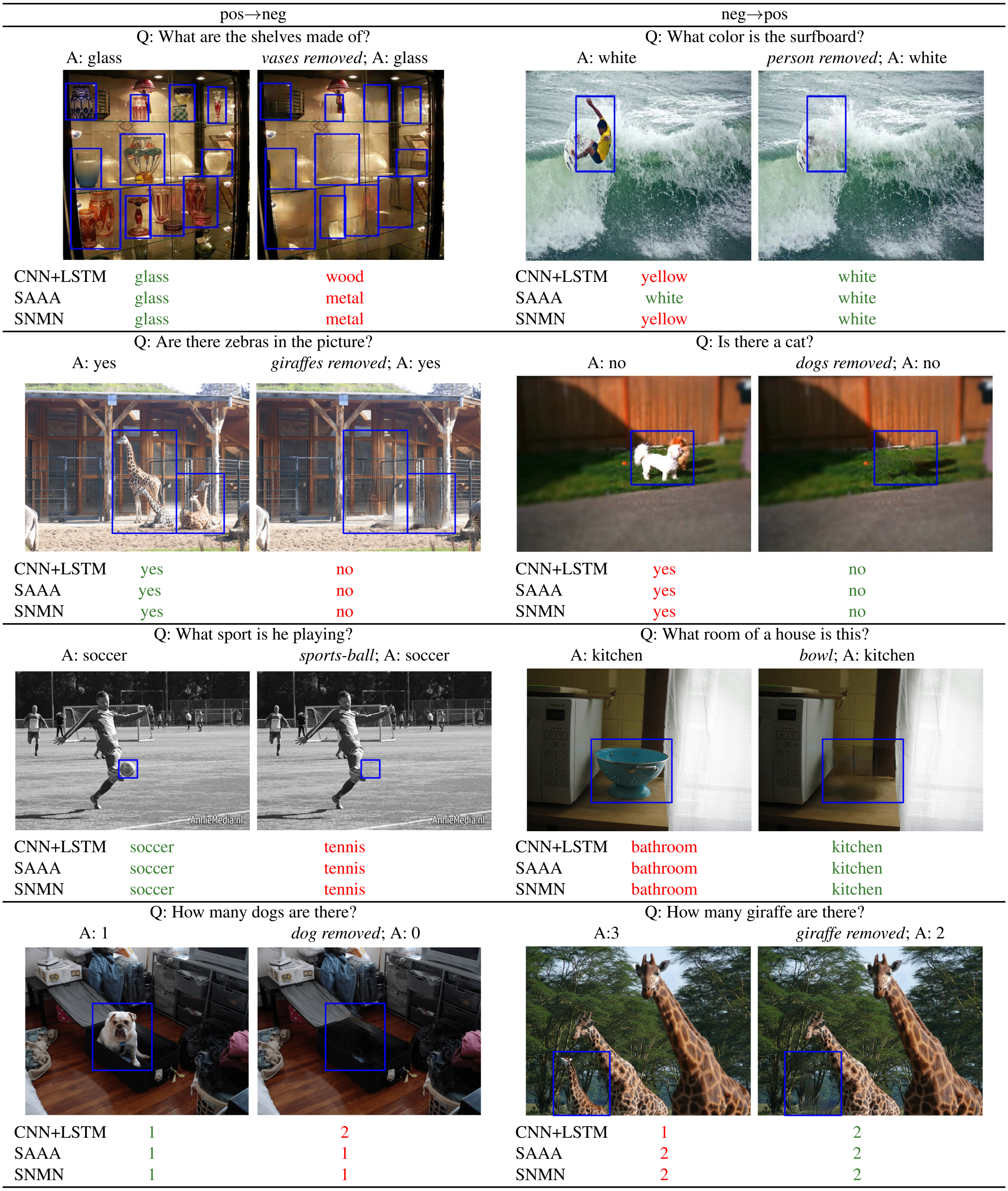


Figure 2: Existing VQA models exploit spurious correlations to predict the answer often looking at irrelevant objects. Shown above are the predictions for 3 different VQA models on original and edited images from our synthetic datasets IV-VQA and CV-VQA.

Trained by us		For comparison	
CNN+LSTM	53.32	d-LSTM Q + norm I [20]	51.61
SAAA [17]	61.14	SAN [30]	52.02
		HieCoAttn [21]	54.57
		MCB [5]	59.71
SNMN [11]	58.34	NMN [2]	51.62

Table 3: Accuracy (in %) of different models when trained on VQA v2 train and tested on VQA v2 val.

	CL (%)	SAAA (%)	SNMN (%)
Accuracy orig	60.21	70.26	66.04
Predictions flipped	17.89	7.85	6.52
pos→neg	7.44	3.47	2.85
neg→pos	6.93	2.79	2.55
neg→neg	3.53	1.58	1.12

Table 4: Accuracy-flipping on real data/IV-VQA test set.

	CL (%)	SAAA (%)	SNMN (%)
Accuracy orig	39.38	49.9	47.948
Predictions flipped	81.41	78.44	78.92
pos→neg	28.69	31.66	32.35
neg→pos	20.57	25.38	23.51
neg→neg	32.14	21.4	23.06

Table 5: Accuracy-flipping on real data/CV-VQA test set.

trolled manner, we expect the models predictions to stay consistent. Robustness is quantified by measuring how often models change their predictions on the edited IQA from the prediction on original IQ. On IV-VQA, a predicted label is considered "flipped" if it differs from the prediction on the corresponding unedited image. On CV-VQA, if the answer on the edited samples is not one less than the prediction on original image, it is considered to be "flipped".

We group the observed inconsistent behavior on edited data into three categories: 1. neg→pos 2. pos→neg 3. neg→neg. neg→pos flip means that answer predicted on the edit IQA was correct but the prediction on the corresponding real IQA was wrong. Other flips are defined analogously. In the neg→neg flip, answer predicted is wrong in both the cases. While all forms of label flipping show inconsistent behaviour, the pos→neg and neg→pos categories are particularly interesting. In these the answer predicted is correct before and afterward the edit, respectively. These metrics show that there is brittleness even while making correct predictions and indicate that models exploit spurious correlations while making their predictions.

Quantitative analysis. Table 4 shows the accuracy along with the consistency numbers for all the 3 models on the IV-VQA test split. Consistency is measured across edited IV-VQA IQAs and corresponding real IQAs from VQA v2. Accuracy is reported on real data from VQA v2 (original IQAs with uniform answers). We follow this convention throughout our paper. On the original data, we see that SAAA is the most accurate model (70.3%) as compared to SNMN (66%) and CL (60.2%). In terms of robustness towards the variations in the images, CL model is the least consistent- with a 17.9% flipping on the edit set compared to the predictions on the corresponding original IQA. For SAAA, 7.85% flips, making SNMN the most robust model with 6.522% flips. SAAA and SNMN are much more stable than CL. A point noteworthy here is that SNMN turns out to be the most robust despite its accuracy being lesser than SAAA. This shows that higher accuracy does not necessarily mean we have the best model, further highlighting the need to study and improve the robustness of the models. Of particular interest are the pos→neg and neg→pos scores, which are close to 7% each for the CL model. For a neg→pos flip, the answer to change from an incorrect answer to one correct answer of the 3000 possible answers (size of answer vector). If the removed object was not used by the model, as it should be, and editing caused uniform perturbations to the model prediction, this event would be extremely rare ($p(\text{neg} \rightarrow \text{pos}) = 1/3000 * 39.8 = 0.013\%$). However we see that this occurs much more frequently (6.9%), indicating that in these cases model was spuriously basing its predictions on the removed object and thus changed the answer when this object was removed.

In the CV-VQA setting, where we target counting and remove one instance of the object to be counted, we expect the models to maintain n/n-1 consistency on real/edited IQA. As we see from Table 5, the accuracy on orig set is quite low for all the models reflecting the fact that counting is a hard problem for VQA models. SAAA (49.9%) is the most accurate model with SNMN at 47.9% and CL at 39.4%. In terms of robustness, we see that for all 3 models are inconsistent more than 75%, meaning for >75% for the edited IQAs, if models could correctly count n objects in the original IQA, it wasn't able to count n-1 instances of the same object in the edited IQA. These numbers further reflect that counting is a difficult task for VQA models and enforcing consistency on it seems to break all 3 models. In the next section, we discuss these flips with some visual examples.

Qualitative analysis. We visualize the predictions of the models on a few original and edited IQAs for all the 3 models in Figure 2. The left half shows examples of pos→neg and the right half shows the neg→pos flips. Existing VQA models often exploit false correlations to predict the answer. We study the different kinds of flips in detail here and see how they help reveal these spurious correlations.

pos→neg. VQA models more often rely on the contextual information/ background cues/ linguistic priors to predict the answer rather than the actual object in the question. For instance, removal of the glass vases from the shelves in Figure 2 (Top-left) from the image causes all 3 models to flip their answers negatively, perhaps models were looking at the wrong object (glass vases) to predict the material of the shelves that also happened to be glass. In absence of giraffes, models cannot seem to spot the occluded zebras anymore- hinting that maybe they are confusing zebras with giraffes. Removing the sports-ball from the field make all 3 models falsely change their predictions to tennis without considering the soccer field or the players. In the bottom-left, we also see that if models were spotting the one dog rightly in the original image, on it's edited counterpart(with no dog anymore)- it fails to answer 0. Semantic edits impact the models negatively here exposing the spurious correlations being used by the models to predict the correct answer on the original image. These examples also show that accuracy should not be the only sole criterion to evaluate performance. A quick look at the Table 4 show that for IV-VQA, pos→neg flips comprise a major chunk (>40%) of all the total flips. For CV-VQA (refer Table 5) , these flips are 28-32% absolute- again reinforcing the fact that VQA models are far from learning to count properly.

neg→pos. Contrary to above, semantic editing here helps correct the predictions, meaning removal of the object causes the model to switch its wrong answer to one right answer by getting rid of the wrong correlations. For instance, removing the pink umbrella helps models predict correctly the color of the balloon Figure 1 (middle) . In Figure 2 (second-right), removing the dogs leave no animals behind and hence models now can correctly spot the absence of cat- hinting that they were previously confusing cats and dogs. In absence of the bowl, models can identify the room as kitchen- shows that too much importance is given to the bowl (which is falsely correlated to bathroom) and not to the objects in the background such as microwave. Towards the bottom-right, we see that removing a giraffe helps all the 3 models now- it's hard to say what is the exact reason for the behaviour but it indeed reflects upon the inconsistent behaviour of the models. From Table 4 we see that these flips also comprise a significant number of the total flips (>35%) for all the models. For CV-VQA (refer Table 5), these numbers are in range 20-25%, showing that counting is easier for these models when spurious correlations are removed.

neg→neg. These flips where answers change show the inconsistent behavior of models as well but since both the answers are wrong- they are harder to interpret. But in the end goal of building robust models, we expect consistent behavior even when making incorrect predictions.

All these flips show that existing VQA models are brittle to semantic variations in images. While VQA models are getting steadily better in terms of accuracy, we also want our models to be robust to visual variations. We want VQA models to not just be accurate but use the right cues to answer correctly. Accuracy combined with consistency can help us understand the shortcomings of the models.

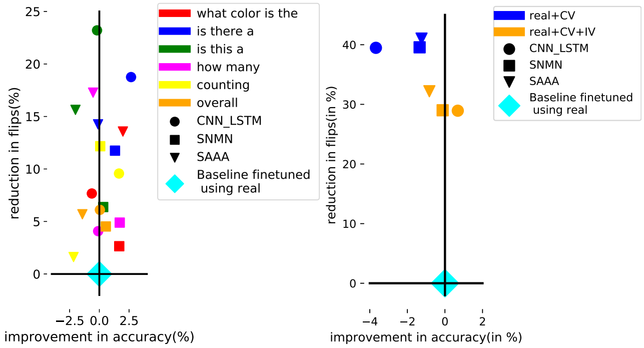


Figure 3: Accuracy-flipping results of finetuning experiments. Plots show relative performance of models finetuned using real+edit data w.r.t to using just real data.

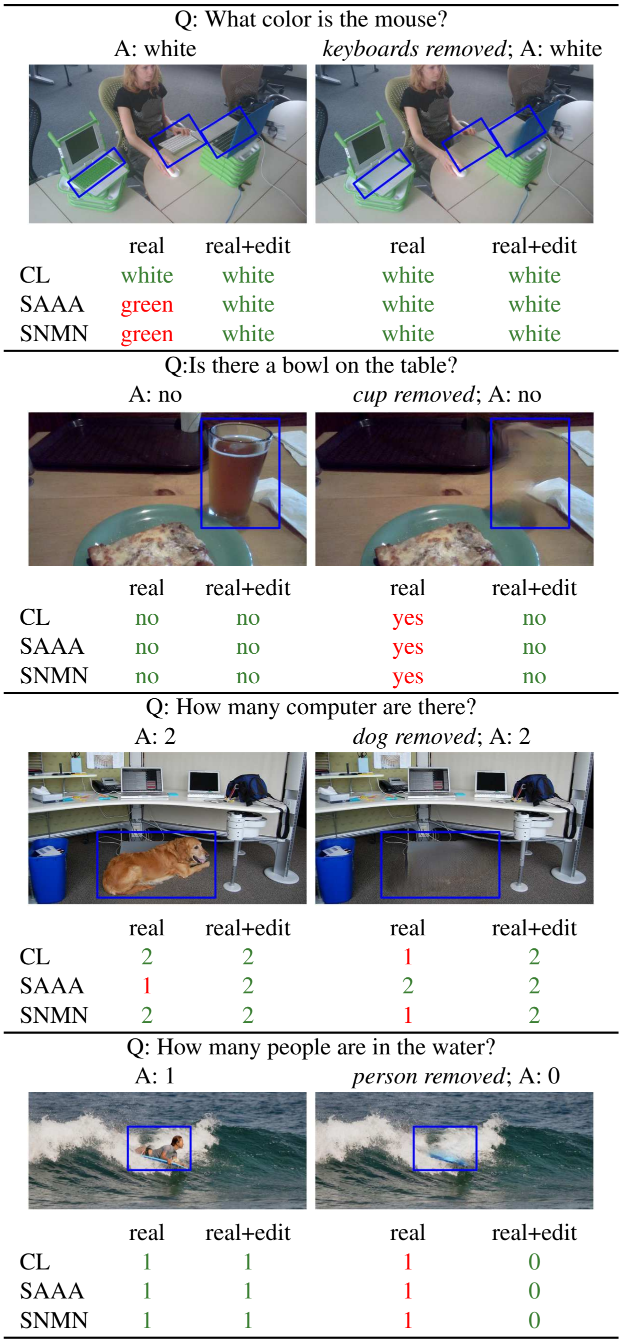
5. Robustification by Data Augmentation

In the previous section, we see that VQA models are brittle to semantic manipulations. While these flips expose the inconsistent behaviour, they also show the underlying scope of improvement for VQA models and can be used to make the models more robust. In order to leverage the variances brought in by the synthetic data, we finetune all the models using real and real+synthetic data. Our analysis shows that using synthetic data significantly reduces inconsistency across a variety of question types.

For fine-tuning experiments, we use a strict subset of IV-VQA with an overlap score of zero. The performance of all the baseline models on this strict subset remains similar to Table 4 (refer supplementary- section C.1). For SNMN, the model trained using a learning rate of $1e^{-3}$ is unstable while fine-tuning and hence we use a lower learning rate $2.5e^{-4}$ to train the model and further finetune this model.

InVariant VQA Augmentation. In order to train and test different models, we aim at specific question types and see if we are able to boost the model's performance on that question type. We select 4 question types based on how much they are affected from editing (i.e total number of flips/ total number of original IQA per question type) and if that question category has significant number of flipped labels in order to ensure we have enough edited IQAs for finetuning. Hence, we select the given 3 question categories and run our experiments on these splits: 1. 'what color is the' 2. 'is there a' 3. 'is this a' 4. 'how many'. Additionally we focus on all the counting questions. All these specialized splits have around 6.3k-12.5k IQAs in the real train split with 10.8k-15.2k in edit train split.

For each question-type, we finetune all the models with corresponding real + IV-VQA IQAs for the particular question type. For a fair baseline, we also finetune all the models using just real data. Figure 3 (left) shows how different



Q: What color is the mouse?

A: white | *keyboards removed*; A: white

	real	real+edit	real	real+edit
CL	white	white	white	white
SAAA	green	white	white	white
SNMN	green	white	white	white

Q:Is there a bowl on the table?

A: no | *cup removed*; A: no

	real	real+edit	real	real+edit
CL	no	no	yes	no
SAAA	no	no	yes	no
SNMN	no	no	yes	no

Q: How many computer are there?

A: 2 | *dog removed*; A: 2

	real	real+edit	real	real+edit
CL	2	2	1	2
SAAA	1	2	2	2
SNMN	2	2	1	2

Q: How many people are in the water?

A: 1 | *person removed*; A: 0

	real	real+edit	real	real+edit
CL	1	1	1	0
SAAA	1	1	1	0
SNMN	1	1	1	0

Figure 4: Visualizations from fine-tuning experiments using real/real+edit. Using real+edit makes models more consistent and in these examples- also accurate.

models, each specialized for a question type, behave when finetuned using real+synthetic data relative to finetuning using real data. The y axis denotes the reduction in flips and x axis represents the accuracy on the original set for. We observe that using synthetic data always reduces flipping as all the points lie above the $y = 0$ axis. The amount of reduction differs for each question type and varies from model to model. For instance, CL model has the highest reduction in flips for question 'is this a' with no change in accuracy and while question type 'how many' shows the least reduction. However for SAAA, 'how many' has the highest reduction with 2.5% drop in accuracy. For SNMN, counting has the highest reduction in flips. We also see that there are many points on the right side of $x = 0$ axis showing that synthetic data also help improve accuracy on the test set. Figure 4 shows some of the examples for these specialized models. As we can see, finetuning the model with IV-VQA dataset helps in improving consistency and leads to more accurate predictions both on real as well as synthetic data.

Additionally, we also finetune all the baseline models with all the real data in VQA-v2 + IV-VQA data. Overall, we find that there is 5-6% relative improvement in flips for all 3 models: CL (17.15→16.1), SAAA (7.53→7.09), SNMN (8.09→7.72) with marginal improvement in accuracy% in case of CL (60.21 →60.24), 1% reduction in accuracy in case of SAAA (70.25→69.25) and 0.6% improvement in accuracy for SNMN (67.65→68.02).

CoVariant VQA Augmentation. For counting, we create our CV-VQA edit set by removing one instance of the object being counted and evaluate the models on both accuracy and consistency. Following the procedure above, we finetune all the models using real data, real+CV and real+CV+IV IQAs. We evaluate the n/n-1 consistency for counting on CV-VQA for all the three models. The results are shown in Figure 3 (right). We see that using CV-VQA edit set reduces flipping by 40% for all 3 models with 1-4% drop in accuracy. Additionally we see that using CV-VQA + IV-VQA data reduce the flipping by 30%: CL (83.8→59.58), SAAA (77.74→52.71), SNMN (77.13→51.91)) with comparable accuracy: CL (43.65→43.94), SAAA (50.87→50.45) and SMNM (50.67→50.61). Figure 4 (Bottom) shows that models when trained using synthetic data can show a more accurate and consistent behaviour. Further consistency analysis with visualizations is in supplementary (section C.3).

6. Conclusion and Future Works

We propose a semantic editing based approach to study and quantify the robustness of VQA models to visual variations. Our analysis shows that the models are brittle to visual variations and reveals spurious correlation being exploited by the models to predict the correct answer. Next, we propose a data augmentation based technique to improve models' performance. Our trained models show significantly less flipping behaviour under invariant and covariant semantic edits, which we believe is an important step towards causal VQA models. By making our invariant and covariant VQA sets as well as evaluation and synthesis available to the community, we hope to support research in the direction towards causal VQA models.

References

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 2, 3, 4

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016. 2, 6

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[5] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468, Austin, Texas, Nov. 2016. Association for Computational Linguistics. 6

[6] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NeurIPS*, 2015. 2

[7] 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*, 2017. 2, 3, 4

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, volume 9, pages 1735–1780. MIT Press, Nov. 1997. 2

[10] Ronghang Hu. Official code release for explainable neural computation via stack neural module networks. https://github.com/ronghanghu/snmn, 2018. 4

[11] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, 2018. 1, 2, 4, 6

[12] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017. 2

[13] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018. 2

[14] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2

[15] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. In *CVIU*, 2017. 2

[16] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *INLG*, 2017. 2

[17] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. In *ArXiv*, volume abs/1704.03162, 2017. 1, 2, 4, 6, 9

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *IEEE*, volume 86, pages 2278–2324, Nov 1998. 2

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[20] Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. 2, 6

[21] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS, 2016*. 2, 6

[22] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, 2016. 2

[23] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NeurIPS*, 2014. 2

[24] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2

[25] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models. In *CVPR*, 2019. 2

[26] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation, 2019. 1, 2, 3

[27] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, June 2019. 1, 2, 3

[28] Rakshith Shetty, Mario Fritz, and Bernt Schiele. Adversarial scene editing: Automatic object removal from weak supervision. In *NeurIPS*, 2018. 1, 3, 4

[29] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Hengel. Visual question answering: A survey of methods and datasets. In *CVIU*, 07 2016. 2

[30] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2, 6

[31] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *CVPR*, 2016. 2

[32] Yan Zhang. Re-implementation of show, ask, attend, and answer: A strong baseline for visual question answering [17] in pytorch. https://github.com/Cyanogenoid/pytorch-vqa, 2017. 4