SQuINTing at VQA Models: Introspecting VQA Models with Sub-Questions

Ramprasaath R. Selvaraju[1]* Purva Tendulkar[1] Devi Parikh[1]
Eric Horvitz[2] Marco Tulio Ribeiro[2] Besmira Nushi[2] Ece Kamar[2]
[1]Georgia Institute of Technology, [2]Microsoft Research
{ramprs, purva, parikh}@gatech.edu
{horvitz, marcotcr, benushi, eckamar}@microsoft.com

Abstract

Existing VQA datasets contain questions with varying levels of complexity. While the majority of questions in these datasets require perception *for recognizing existence, properties, and spatial relationships of entities, a significant portion of questions pose challenges that correspond to* reasoning tasks *– tasks that can only be answered through a synthesis of perception and knowledge about the world, logic and / or reasoning. Analyzing performance across this distinction allows us to notice when existing VQA models have consistency issues; they answer the reasoning questions correctly but fail on associated low-level perception questions. For example, in Figure 1, models answer the complex reasoning question "Is the banana ripe enough to eat?" correctly, but fail on the associated perception question "Are the bananas mostly green or yellow?" indicating that the model likely answered the reasoning question correctly but for the wrong reason. We quantify the extent to which this phenomenon occurs by creating a new Reasoning split of the VQA dataset and collecting VQA-introspect, a new dataset[1] which consists of 238K new perception questions which serve as sub questions corresponding to the set of perceptual tasks needed to effectively answer the complex reasoning questions in the Reasoning split. Our evaluation shows that state-of-the-art VQA models have comparable performance in answering perception and reasoning questions, but suffer from consistency problems. To address this shortcoming, we propose an approach called Sub-Question Importance-aware Network Tuning (SQuINT), which encourages the model to attend to the same parts of the image when answering the reasoning question and the perception sub question.We show that SQuINT improves model consistency by ∼5%, also marginally improving performance on the Reasoning questions in VQA, while also displaying better attention maps.*

*Research performed in part during an internship at Microsoft Research
[1]Our dataset can be found at aka.ms/vqa-introspect.

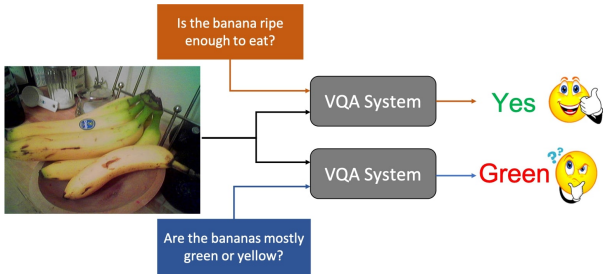


Figure 1: A potential reasoning failure: Current models answer the Reasoning question "Is the banana ripe enough to eat?" correctly with the answer "Yes". We might assume that doing so stems from perceiving relevant concepts correctly – perceiving yellow bananas in this example. But when asked "Are the bananas mostly green or yellow?", the model answers the question incorrectly with "Green" – indicating that the model possibly answered the original Reasoning question for the wrong reasons even if the answer was right. We quantify the extent to which this phenomenon occurs in VQA and introduce a new dataset aimed at stimulating research on well-grounded reasoning.

1. Introduction

Human cognition is thought to be compositional in nature: the visual system recognizes multiple aspects of a scene which are combined into shapes [7] and understandings. Likewise, complex linguistic expressions are built from simpler ones [5]. Similarly, tasks like Visual Question Answering (VQA) require models to perform inference at multiple levels of abstraction. For example, to answer the question, "Is the banana ripe enough to eat?" (Figure 1), a VQA model has to be able to detect the bananas and extract associated properties such as size and color (perception), understand what the question is asking, and reason about how these properties relate to known properties of edible bananas (ripeness) and how they manifest (yellow versus green in color). While "abstraction" is complex and spans distinctions at multiple levels of detail, we focus on separating questions into Perception and Reasoning questions. Perception questions only require visual perception to recognize existence, physical properties or spatial relationships among entities, such as "What color is the banana?" or

"What is to the left of the man?", while Reasoning questions require the composition of multiple perceptual tasks and knowledge that harnesses logic and prior knowledge about the world, such as "Is the banana ripe enough to eat?".

Current VQA datasets [3, 6, 15] contain a mixture of Perception and Reasoning questions, which are considered equivalent for the purposes of evaluation and learning. Categorizing questions into Perception and Reasoning promises to promote a better assessment of visual perception and higher-level reasoning capabilities of models, rather than conflating these capabilities. Furthermore, we believe it is useful to identify the Perception questions that serve as subtasks in the compositional processes required to answer the Reasoning question. By elucidating such "sub-questions," we can check whether the model is reasoning appropriately or if it is relying on spurious shortcuts and biases in datasets [1]. For example, we should be cautious about the model's inferential ability if it simultaneously answers "no" to "Are the bananas edible?" and "yellow" to "What color are the bananas?", even if the answer to the former question is correct. The inconsistency between the higher-level reasoning task and the lower-level perception task that it builds upon suggests that the system has not learned effectively how to answer the Reasoning question and will not be able to generalize to same or closely related Reasoning question with another image. The fact that these sub-questions are in the same modality (i.e. questions with associated answers) allows for the evaluation of any VQA model, rather than only models that are trained to provide justifications. It is this key observation that we use to develop an evaluation methodology for Reasoning questions.

The dominant learning paradigm for teaching models to answer VQA tasks assumes that models are given <image, question, answer> triplets, with no additional annotation on the relationship between the question and the compositional steps required to arrive at the answer. As reasoning questions become more complex, achieving good coverage and generalization with methods used to date will likely require a prohibitive amount of data. Alternatively, we employ a hierarchical decomposition strategy, where we identify and link Reasoning questions with sets of appropriate Perception sub-questions. Such an approach promises to enable new efficiencies via compositional modeling, as well as lead to improvements in the consistency of models for answering Reasoning questions. Explicitly representing dependencies between Reasoning tasks and the corresponding Perception tasks also provides language-based grounding for reasoning questions where visual grounding [14, 18] may be insufficient, e.g., highlighting that the banana is important for the question in Figure 1 does not tell the model how it is important (i.e. that color is an important property rather than size or shape). Again, the fact that such grounding is in question-answer form (which models already have to deal with) is an added benefit. Such annotations allow for attempts to enforce reasoning devoid of shortcuts that do not generalize, or are not in line with human values and business rules, even if accurate (e.g. racist behavior).

We propose a new split of the VQA dataset, containing only Reasoning questions (defined previously). Furthermore, for questions in the split, we introduce VQA-introspect, a new dataset of 238K associated Perception sub-questions which humans perceive as containing the sub-questions needed to answer the original questions. After validating the quality of the new dataset, we use it to perform fine-grained evaluation of state-of-the-art models, checking whether their reasoning is in line with their perception. We show that state-of-the-art VQA models have similar accuracy in answering perception and reasoning tasks but have problems with consistency; in 28.14% of the cases where models answer the reasoning question correctly, they fail to answer the corresponding perception sub-question, highlighting problems with consistency and the risk that models may be learning to answer reasoning questions through learning common answers and biases.

Finally, we introduce SQuINT – a generic modeling approach that is inspired by the compositional learning paradigm observed in humans. SQuINT incorporates VQA-introspect annotations into learning with a new loss function that encourages image regions important for the sub-questions to play a role in answering the main Reasoning questions. Empirical evaluations demonstrate that the approach results in models that are more consistent across Reasoning and associated Perception tasks with no major loss of accuracy. We also find that SQuINT improves model attention maps for Reasoning questions, thus making models more trustworthy.

2. Related Work

Visual Question Answering [3], one of the most widely studied vision-and-language problems, requires associating image content with natural language questions and answers (thus combining perception, language understanding, background knowledge and reasoning). However, it is possible for models to do well on the task by exploiting language and dataset biases, e.g. answering "yellow" to "What color is the banana?" without regard for the image or by answering "yes" to most yes-no questions [1, 12, 18, 20, 2]. This motivates additional forms of evaluation, e.g. checking if the model can understand question rephrasings [19] or whether it exhibits logical consistency [16]. In this work, we present a novel evaluation of questions that require reasoning capabilities, where we check for consistency between how models answer higher level Reasoning questions and how they answer corresponding Perception sub-questions.

A variety of datasets have been released with attention annotations on the image pointing to regions that are important to answer questions ([4, 10]), with corresponding work

on enforcing such grounding [17, 14, 18]. Our work is complementary to these approaches, as we provide language-based grounding (rather than visual), and further evaluate the link between perception capabilities and how they are composed by models for answering Reasoning tasks. Closer to our work is the dataset of Lisa et al. [10], where natural language justifications are associated with (question, answer) pairs. However, most of the questions contemplated (like much of the VQA dataset) pertain to perception questions (e.g. for the question-answer "What is the person doing? Snowboarding", the justification is "...they are on a snowboard ..."). Furthermore, it is hard to use natural language justifications to evaluate models that do not generate similar rationales (i.e. most SOTA models), or even coming up with metrics for models that do. In contrast, our dataset and evaluation is in the same modality (QA) that models are already trained to handle.

3. Reasoning-VQA and VQA-introspect

In the first part of this section, we present an analysis of the common type of questions in the VQA dataset and highlight the need for classifying them into Perception and Reasoning questions. We then define Perception and Reasoning questions and describe our method for constructing the Reasoning split. In the second part, we describe how we create the new VQA-introspect dataset through collecting sub-questions and answers for questions in our Reasoning split. Finally, we describe experiments conducted in order to validate the quality of our collected data.

3.1. Perception vs. Reasoning

A common technique for finer-grained evaluation of VQA models is to group instances by answer type (yes/no, number, other) or by the first words of the question (what color, how many, etc) [3]. While useful, such slices are coarse and do not evaluate the model's capabilities at different points in the abstraction scale. For example, questions like "Is this a banana?" and "Is this a healthy food?" start with the same words and expect yes/no answers. While both test if the model can do object recognition, the latter requires additional capabilities in connecting recognition with prior knowledge about which food items are healthy and which are not. This is not to say that Reasoning questions are inherently harder, but that they require both visual understanding and an additional set of skills (logic, prior knowledge, etc) while Perception questions deal mostly with visual understanding. For example, the question "How many round yellow objects are to the right of the smallest square object in the image?" requires very complicated visual understanding, and is arguably harder than "Is the banana ripe enough to eat?", which requires relatively simple visual understanding (color of the bananas) and knowledge about properties of ripe bananas. Regardless of difficulty, categorizing questions as Perception or Reasoning is useful for both detailed model evaluation based on capabilities and also improving learning, as we demonstrate in later sections. We now proceed to define these categories more formally.

Perception : We define Perception questions as those which can be answered by detecting and recognizing the existence, physical properties and / or spatial relationships between entities, recognizing text / symbols, simple activities and / or counting, and that do not require more than one hop of reasoning or general commonsense knowledge beyond what is visually present in the image. Some examples are: "Is that a cat? " (existence), "Is the ball shiny?" (physical property), "What is next to the table?" (spatial relationship), "What does the sign say?" (text / symbol recognition), "Are the people looking at the camera?" (simple activity), etc. We note that spatial relationship questions have been considered reasoning tasks in previous work [9] as they require lower-level perception tasks in composition to be answered. For our purposes it is useful to separate visual understanding from other types of reasoning and knowledge, and thus we classify such spatial relationships as Perception.

Reasoning : We define Reasoning questions as non-Perception questions which require the synthesis of perception with prior knowledge and / or reasoning in order to be answered. For instance, "Is this room finished or being built?", "At what time of the day would this meal be served?", "Does this water look fresh enough to drink?", "Is this a home or a hotel?", "Are the giraffes in their natural habitat?" are all Reasoning questions.

Our analysis of the perception questions in the VQA dataset revealed that most perception questions have distinct patterns that can be identified with high precision regex-based rules. By handcrafting such rules (details can be found in the Supplementary) and filtering out perception questions, we identify 18% of the VQA dataset as highly likely to be Reasoning. To check the accuracy of our rules and validate their coverage of Reasoning questions, we designed a crowdsourcing task on Mechanical Turk that instructed workers to identify a given VQA question as Perception or Reasoning, and to subsequently provide sub-questions for the Reasoning questions, as described next. 89.25% of the times, at least 2 out of 3 workers classified our resulting questions as reasoning questions demonstrating the high precision of the regex-based rules we created.

3.2. VQA-introspect data

Given the complexity of distinguishing between Perception / Reasoning and providing sub-questions for Reasoning questions, we first train and filter workers on Amazon Mechanical Turk (AMT) via qualification rounds before we rely on them to generate high-quality sub-questions.

Worker Training - We manually annotate 100 questions

from the VQA dataset as Perception and 100 as Reasoning questions, to serve as examples. We first teach crowdworkers the difference between Perception and Reasoning questions by presenting definitions and showing several examples of each, along with explanations. Then, crowdworkers are shown (question, answer) pairs and are asked to identify if the given question is a Perception question or a Reasoning question [2]. Finally, for Reasoning questions, we ask workers to add all Perception questions and corresponding answers (in short) that would be necessary to answer the main question (details and interface can be found in the Supplementary). In this qualification HIT, workers have to make 6 Perception and Reasoning judgments, and they qualify if they get 5 or more answers right.

We launched further pilot experiments for the crowdworkers who passed the first qualification round, where we manually evaluated the quality of their sub-questions based on whether they were Perception questions grounded in the image and sufficient to answer the main question. Among those 540 workers who passed the first qualification test, 144 were selected (via manual evaluation) as high-quality workers, who finally qualified for attempting our main task.

Main task - In the main data collection, all VQA questions identified as Reasoning by regex-rules and a random subset of the questions identified as Perception were further judged by workers (for validation purposes). We eliminated ambiguous questions by further filtering out questions where there is high worker disagreement about the answer. We required at least 8 out of 10 workers to agree with the majority answer for yes/no questions and 5 out of 10 for all other questions. This labeling step left us with a Reasoning split that corresponds to $\sim$13% of the VQA dataset.

At the next step. each <question, image> pair labeled as Reasoning had sub questions generated by 3 unique workers [3]. Removing duplicate question, answer pairs left on average 3.1 sub-questions per Reasoning question. Qualitative examples from the resulting dataset are presented in Fig. 7.

The resulting train split of VQA-introspect contains 166927 sub questions for 55799 Reasoning questions in VQAv2 train, and the val split of VQA-introspect contains 71714 sub questions for 21677 Reasoning questions in VQAv2 val. This Reasoning split is not exhaustive, but is high precision (as demonstrated below) and contains questions that are not ambiguous, and thus is useful for evaluation and learning.

3.3. Dataset Quality Validation

In order to confirm that the sub-questions in VQA-introspect are really Perception questions, we did a further round of evaluation with workers who passed the worker qualification task described in Section C but had not provided sub-questions for our main task. In this round, 87.8% of sub-questions in VQA-introspect were judged to be Perception questions by at least 2 out of 3 workers.

It is crucial for the semantics of VQA-introspect that the sub-questions are tied to the original Reasoning question. While verifying that the sub-questions are necessary to answer the original question requires workers to think of all possible ways the original question could be answered (and is thus too hard), we devised an experiment to check if the sub-questions provide at least sufficient visual understanding to answer the Reasoning question. In this experiment, workers are shown the sub-questions with answers, and then asked to answer the Reasoning question without seeing the image, thus having to rely only on the visual knowledge conveyed by the sub-questions. At least 2 out of 3 workers were able to answer 89.3% of the Reasoning questions correctly in this regime (95.4% of binary Reasoning questions). For comparison, when we asked workers to answer Reasoning questions with no visual knowledge at all (no image and no sub-questions), this accuracy was 52% (58% for binary questions). These experiments give us confidence that the sub-questions in VQA-introspect are indeed Perception questions that convey components of visual knowledge which can be composed to answer the original Reasoning questions.

4. Dataset Analysis

The distribution of questions in our VQA-introspect dataset is shown in Figure 3. It is interesting to note that comparing these plots with those for the VQA dataset [3] show that the VQA-introspect dataset questions are more specific. For example, there are only 12 "why" questions in the dataset which tend to be reasoning questions. Also, for "where" questions, a very common answer in VQA was "outside" but answers are more specific in our VQA-introspect dataset (e.g., "beach", "street"). Figure 4 shows the distribution of question lengths in the Perception and Reasoning splits of VQA and in our VQA-introspect dataset. We see that most questions range from 4 to 10 words. Lengths of questions in the Perception and Reasoning splits are quite similar, although questions in VQA-introspect are slightly longer (the curve is slightly shifted to the right), possibly on account of the increase in specificity/detail of the questions.

One interesting question is whether the main question and the sub-questions deal with the same concepts. In order to explore this, we used noun chunks surrogates for concepts [4], and measured how often there was any overlap in concepts between the main question and the associated sub-

[2] We also add an "Invalid" category to flag nonsensical questions or those which can be answered without looking at the image

[3] A small number of workers displayed degraded performance after the qualification round, and were manually filtered

[4] Concepts are extracted with the Python spaCy library.

Main Reasoning Question:
- Is this a keepsake photo? "Yes"

Perception Sub-questions:
- Is this a black and white photo? "Yes"
- Is the woman wearing a white veil and holding flowers? "Yes"
- Is the woman wearing a veil? "Yes"
- What is the woman next to the man wearing? "Gown"

(a)

Main Reasoning Question:
- Is this giraffe at the zoo? "Yes"

Perception Sub-questions:
- Is the giraffe fenced in? "Yes"
- Is the grass shorter than 3 inches? "Yes"
- Is there a fence? "Yes"
- Is a fence around the giraffe? "Yes"

(b)

Main Reasoning Question:
- Does this appear to be an emergency? "Yes"

Perception Sub-questions:
- Are there a lot of ambulances? "Yes"
- Are people standing in the middle of the street? "Yes"
- Is there a firetruck? "Yes"
- Does the white vehicle say "ambulance"? "Yes"
- Does the red truck say "fire department"? "Yes"

(c)

Main Reasoning Question:
- Could this cause cavities? "Yes"

Perception Sub-questions:
- What food is on the plate? "Dessert"
- Is the boy eating whipped cream? "Yes"
- Is his mouth wide open? "Yes"

(d)

Main Reasoning Question:
- Is this a good idea for a rainy day? "No"

Perception Sub-questions:
- Is there a roof on the bus? "No"
- Does the vehicle have a roof? "No"

(e)

Main Reasoning Question:
- Are these dangerous for babies? "No"

Perception Sub-questions:
- What food is shown? "Peanut butter and banana"
- What food is on the plate? "Banana"
- Is there pudding on the plate? "Yes"

(f)

Main Reasoning Question:
- Is it evening? "Yes"

Perception Sub-questions:
- Are the lights on the train lit up? "Yes"
- Is there a lot of sunlight in the sky? "No"

(g)

Main Reasoning Question:
- How is this train powered? "Electricity"

Perception Sub-questions:
- Are there electric lines? "Yes"
- Is the train attached to a cable? "Yes"
- Are there wires above the tracks? "Yes"

(h)

Figure 2: Qualitative examples of Perception sub-questions in our VQA-introspect dataset for main questions in the Reasoning split of VQA. Main questions are in orange and sub questions are in blue. A single worker may have provided more than one sub questions for the same (image, main question) pair. More examples can be found in the Supplementary

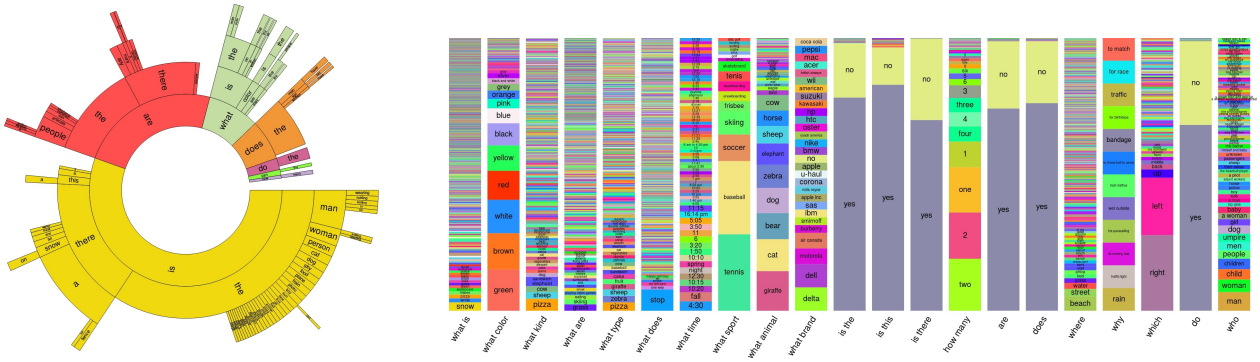


Figure 3: Left: Distribution of questions by their first four words. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show, Right: Distribution of answers per question type

question. Noun-chunks are only a surrogate and may miss semantic overlap otherwise present (e.g. through verb-noun connections like "fenced" and "a fence" in Figure 7 (b), sub-questions). With this caveat, we observe that there is overlap only 24.18% of the time, indicating that Reasoning questions in our split often require knowledge about concepts not explicitly mentioned in the corresponding Perception questions. The lack of overlap indicates that models cannot solely rely on visual perception in answering Reasoning tasks, but incorporating background knowledge and common sense understanding is necessary. For example, in the question "Is the airplane taking off or landing?", the concepts present are 'airplane' and 'landing', while for the associated sub-question "Are the wheels out?", the concept is 'wheels'. Though 'wheels' do not occur in the main question, the concept is important, in that providing this ground-

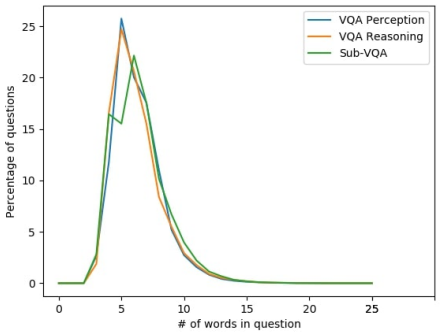


Figure 4: Percentage of questions with different word lengths for the train and val sub-questions of our VQA-introspect dataset.

ing might help the model explicitly associate the connection between airplane wheels and take-offs / landings.

5. Fine grained evaluation of VQA Reasoning

VQA-introspect enables a more detailed evaluation of the performance of current state-of-the-art models on Reasoning questions by checking whether correctness on these questions is consistent with correctness on the associated Perception sub-questions. It is important to notice that a Perception failure (an incorrect answer to a sub-question) may be due to a problem in the vision part of the model or a grounding problem – the model in Figure 5 may know that the banana is mostly yellow and use that information to answer the ripeness question, while, at the same time, fail to associate this knowledge with the word "yellow", or fail to understand what the sub-question is asking. While grounding problems are not strictly visual perception failures, we still consider them Perception failures because the goal of VQA is to answer natural language questions about an image, and the sub-question being considered pertain to Perception knowledge as defined previously. With this caveat, there are four possible outcomes when evaluating Reasoning questions with associated Perception sub-questions, which we divide into four quadrants:

Q1: Both main & sub-questions correct (M✓ S✓): While we cannot claim that the model predicts the main question correctly *because* of the sub-questions (e.g. the bananas are ripe *because* they are mostly yellow), the fact that it answers both correctly is consistent with good reasoning, and should give us more confidence in the original prediction.

Q2: Main correct & sub-question incorrect (M✓ S✗): The Perception failure indicates that there might be a reasoning failure. While it is possible that the model is composing other perception knowledge that was not captured by the identified sub-questions (e.g. the bananas are ripe because they have black spots on them), it is also possible (and more likely) that the model is using a spurious shortcut or was correct by random chance.

Q3: Main incorrect & sub-question correct (M✗ S✓): The Perception failure here indicates a clear reasoning failure, as we validated that the sub-questions are sufficient to answer the main question. In this case, the model knows that the bananas are mostly yellow and still thinks they are not ripe enough, and thus it failed to make the "yellow bananas are ripe" connection.

Q4: Both main & sub-question incorrect (M✗ S✗): While the model may not have the reasoning capabilities to answer questions in this quadrant, the Perception failure could explain the incorrect prediction.

In sum, Q2 and Q4 are definitely Perception failures, Q2 likely contains Reasoning failures, Q3 contains Reasoning failures, and we cannot judge Reasoning in Q4.

As an example, we evaluate the Pythia model [11] (SOTA as of 2018)[5] along these quadrants (Table 1) for the Reasoning split of VQA. The overall accuracy of the model is 64.95%, while accuracy on Reasoning questions is 69.61%. We note that for 28.27% of the cases, the model is inconsistent, i.e., it answered the main question correctly, but got the sub question wrong. Further, we observe that 9.02% of the times the Pythia model gets *all* the sub questions wrong when the main question is right – *i.e.*, it seems to be severely wrong on its perception and using other paths (shortcuts or biases) to get the Reasoning question right .

6. Improving learned models with VQA-introspect

In this section, we consider how VQA-introspect can be used to improve models that were trained on VQA datasets. Our goal is to reduce the number of possible reasoning or perception failures (M✓ S✗ and M✗ S✓) without diminishing the original accuracy of the model.

6.1. Finetuning

The simplest way to incorporate VQA-introspect into a learned model is to fine-tune the model on it. We use the averaged binary cross entropy loss for the main question and the sub question as a loss function. Furthermore, to avoid catastrophic forgetting [13] of the original VQA data during finetuning, we augment every batch with randomly sampled data from the original VQA dataset.

6.2. Sub-Question Importance-aware Network Tuning (SQuINT)

The intuition behind Sub-Question Importance-aware Network Tuning (SQuINT) is that a model should attend to the same regions in the image when answering the Reasoning questions as it attends to when answering the associated Perception sub-questions, since they capture the visual components required to answer the main question. SQuINT does this by learning how to attend to sub-question

[5] source: https://visualqa.org/roe_2018.html

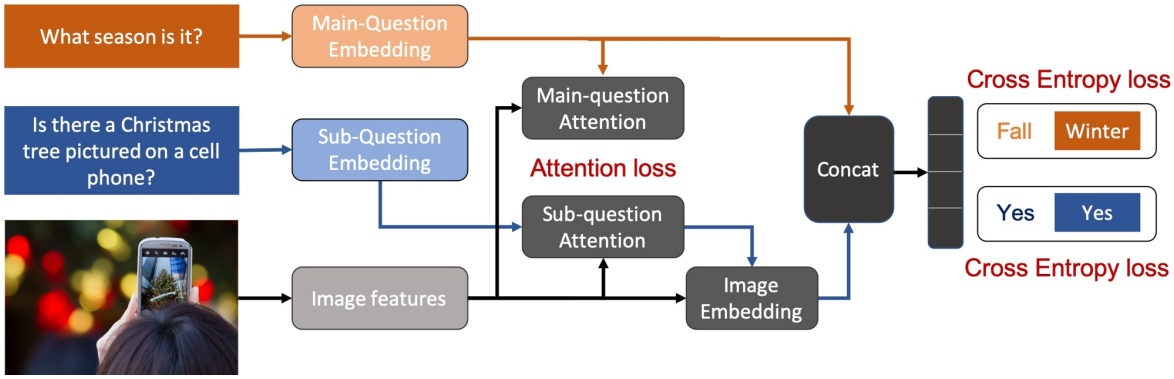


Figure 5: Sub-Question Importance-aware Network Tuning (SQuINT) approach: Given an image, a Reasoning question like "What season is it?" and an associated Perception sub-question like "Is there a Christmas tree pictured on a cell phone?", we pass them through the Pythia architecture [11]. The loss function customized for SQuINT is composed of three components: an attention loss that penalizes for the mismatch between attention for the main-question and the attention for the sub-question based on an image embedding conditioned on sub-question and image features, a cross entropy loss for answer of the main-question and a cross entropy loss for the answer of the sub-question. The loss function encourages the model to get the answers of both the main-question and sub-question right simultaneously, while also encouraging the model to use the right attention regions for the reasoning task.

regions of interest and reasoning over them to answer the main question. We now describe how to construct a loss function that captures this intuition.

Attention loss - As described in Section 3, the sub-questions in the dataset are simple perception questions asking about well-grounded objects/entities in the image. Current well-performing models based on attention are generally good at visually grounding regions in the image when asked about simple Perception questions, given that they are trained on VQA datasets which contain large amounts of Perception questions. In order to make the model look at the associated sub-question regions while answering the main question, we apply a Mean Squared Error (MSE) loss over the the spatial and bounding box attention weights.

Cross Entropy loss - While the attention loss encourages the model to look at the right regions given a complex Reasoning question, we need a loss that helps the model learn to reason given the right regions. Hence we apply the regular Binary Cross Entropy loss on top of the answer predicted for the Reasoning question given the sub-question attention. In addition we also use the Binary Cross Entropy loss between the predicted and GT answer for the sub-question.

Total SQuINT loss - We jointly train with the attention and cross entropy losses. Let A_{reas} and A_{sub} be the model attention for the main reasoning question and the associated sub-question, and gt_{reas} and gt_{sub} be the ground-truth answers for the main and sub-question respectively. Let $o_{reas}|A_{sub}$ be the predicted answer for the reasoning question given the attention for the sub-question. The SQuINT loss is formally defined as:

$$\begin{aligned}\mathcal{L}_{\textbf{SQuINT}} &= \text{MSE}(A_{reas}, A_{sub}) \\ &+ \lambda_1 \text{BCE}(o_{reas}|A_{sub}, gt_{reas}) \\ &+ \lambda_2 \text{BCE}(o_{sub}, gt_{sub})\end{aligned}$$

The first term encourages the network to look at the same regions for reasoning and associated perception questions, while the second and third terms encourage the model to give the right answers to the questions given the attention regions. The loss is simple and can be applied as a modification to any model that uses attention.

7. Experiments

In this section, we perform fine grained evaluation of VQA reasoning as detailed in Section 5, using the SOTA model **Pythia** [11] as a base model (although any model that uses visual attention would suffice). We trained the base model on VQAv2, and evaluated the baseline and all variants on the Reasoning split and corresponding VQA-introspect val sub-questions[6]. As detailed in Section 6, **Pythia + VQA-introspect data** corresponds to finetuning the base model on train VQA-introspect subquestions, while **Pythia + SQuINT** finetunes Pythia model such that it now attends to the same regions for main questions and associated sub-questions. In Table 1, we report the reasoning breakdown detailed in Section 5. We also report a few additional metrics: **Consistency** refers to how often the model predicts the sub-question correctly given that it answered the main question correctly, while **Consistency (balanced)** reports the same metric on a balanced version of the sub-questions (to make sure models are not exploiting biases to gain consistency). **Attention Correlation** refers to the correlation between the attention embeddings of the main and sub-question. Finally, we report **Overall** accuracy (on the whole evaluation dataset), and accuracy on the Reasoning split (**Reasoning Accuracy**). Note that our approach does not require sub-questions at test time. We use $\lambda_1 = 0.1$ and $\lambda_2 = 1$ with a learning rate of 0.01 in our experiments.

The results in Table 1 indicate that fine-tuning on VQA-

[6]Note that this is different from our CVPR camera-ready version where the experiments were conducted on the previous version of the VQA-introspect dataset with a different base model that was trained on VQAv1.

Method	Consistency Metric							VQA Accuracy	
	M✓ S✓ ↑	M✓ S✗ ↓	M✗ S✓ ↓	M✗ S✗ ↓	Consistency% ↑	Consistency% (balanced) ↑	Attn Corr ↑	Overall ↑	Reasoning (M✓ S✓ + M✓ S✗) ↑
Pythia	50.05	19.73	17.40	12.83	71.73	75.67	0.71	**64.95**	69.61
Pythia + VQA-introspect data	53.21	16.62	19.40	10.77	76.20	74.58	0.71	64.60	69.64
Pythia + SQuINT	**53.90**	16.24	19.34	10.52	**76.84**	**75.76**	**0.75**	64.73	**69.88**

Table 1: Results on held out VQA-introspect val set for (1) Consistency metrics along the four quadrants described in Section 5 and Consistency and Attention Correlation metrics as described in Section 5 (metrics), and (2) Overall and Reasoning accuracy. The Reasoning accuracy is obtained by only looking at the number of times the main question is correct (M✓ S✓ + M✓ S✗) ignoring repetitions of the main question due to multiple sub-questions.

Figure 6: Qualitative examples showing the model attention before and after applying SQuINT. (a) shows an image along with the reasoning question, *'Is the man airborne?'*, for which the Pythia model looks at somewhat irrelevant regions and answers "No" incorrectly. Note how the same model correctly looks at the feet to answer the easier sub-question, *'Does the man have his feat on the ground?'*. After applying SQuINT, which encourages the model to use the perception based sub question attention while answering the reasoning question, it now looks at the feet and correctly answers the main question.

introspect (using data augmentation or SQuINT), increases consistency without hurting accuracy or Reasoning accuracy. Correspondingly, our confidence that it actually learned the necessary concepts when it answered Reasoning questions correctly should increase.

The **Attention Correlation** numbers indicate that SQuINT really is helping the model use the appropriate visual grounding (same for main-question as sub-questions) at test time. This effect does not seem to happen with naive finetuning on VQA-introspect. We present qualitative validation examples in Figure 6, where the base model attends to irrelevant regions when answering the main question (even though it answers correctly), while attending to relevant regions when asked the sub-question. The model finetuned on SQuINT, on the other hand, attends to regions that are actually informative in both main and sub-questions (notice that this is evaluation, and thus the model is not aware of the sub-question when answering the main question and vice versa). This is further indication that SQuINT is helping the model reason in ways that will generalize when it answers Reasoning questions correctly, rather than use shortcuts.

8. Discussion and Future Work

The VQA task requires multiple capabilities in different modalities and at different levels of abstraction. We introduced a hard distinction between Perception and Reasoning which we acknowledge is a simplification of a continuous and complex reality, albeit a useful one. In particular, linking the perception components that are needed (in addition to other forms of reasoning) to answer reasoning questions opens up an array of possibilities for future work, in addition to improving evaluation of current work. We proposed preliminary approaches that seem promising: finetuning on VQA-introspect and SQuINT both improve the consistency of the SOTA model with no discernible loss in accuracy, and SQuINT results in qualitatively better attention maps. We expect future work to use VQA-introspect even more explicitly in the modeling approach, similar to current work in explicitly composing visual knowledge to improve *visual* reasoning [8]. In addition, similar efforts to ours could be employed at different points in the abstraction scale, e.g. further dividing complex Perception questions into simpler components, or further dividing the Reasoning part into different forms of background knowledge, logic, etc. We consider such efforts crucial in the quest to evaluate and train models that truly generalize, and hope VQA-introspect spurs more research in that direction.

Acknowledgements. We are grateful to Dhruv Batra for providing very useful feedback on this work. The Georgia Tech effort was supported in part by NSF, AFRL, DARPA, ONR YIPs, ARO PECASE, Amazon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

Appendices

A. Introduction

This supplementary material is organized as follows. We first provide a sample of the kind of regex-based rules that we used to arrive at reasoning questions. We then provide the interface we designed for training and evaluating Mechanical turk workers and the interface for collecting the main dataset. We then show randomly sampled responses from workers.

B. Perception-VQA vs Reasoning-VQA

In the first part of this section, we revisit our definition of Perception and Reasoning questions and later we describe our rules for constructing the Reasoning split.

B.1. Perception vs. Reasoning

Perception : As mentioned in section 3.1 of the main paper, we define Perception questions as those which can be answered by detecting and recognizing the existence, physical properties and / or spatial relationships between entities, recognizing text / symbols, simple activities and / or counting, and that do not require more than one hop of reasoning or general commonsense knowledge beyond what is visually present in the image. Some examples are: "Is that a cat? " (existence), "Is the ball shiny?" (physical property), "What is next to the table?" (spatial relationship), "What does the sign say?" (text / symbol recognition), "Are the people looking at the camera?" (simple activity), etc.

Reasoning : We define Reasoning questions as non-Perception questions which require the synthesis of perception with prior knowledge and / or reasoning in order to be answered. For instance, "Is this room finished or being built?", "At what time of the day would this meal be served?", "Does this water look fresh enough to drink?", "Is this a home or a hotel?", "Are the giraffes in their natural habitat?" are all Reasoning questions.

B.2. Rules

As mentioned in section 3.1 of the main paper, our analysis of the perception questions in the VQA dataset revealed that most perception questions have distinct patterns that can be identified with high precision regex-based rules. In Table 2 we provide a list of top-40 regex rules based on the percentage of data the rule eliminated.

By hand-crafting such rules (as seen in Table 2) and filtering out perception questions, we identify 18% of the VQA dataset as highly likely to be Reasoning.

B.3. Validating rules

To check the accuracy of our rules, we designed a crowdsourcing task on Mechanical Turk that instructed workers to identify a given VQA question as Perception or Reasoning, and to subsequently provide sub-questions for the Reasoning questions.

Validating Precision. As mentioned in Section 3.1, 94.7% of the times, trained workers classified our resulting questions as reasoning questions demonstrating the high precision of the regex-based rules we created.

C. VQA-introspect

In this section, we describe how we collect sub-questions and answers for questions in our Reasoning split.

Given the complexity of distinguishing between Perception / Reasoning and providing sub-questions for Reasoning questions, we first train and filter workers on Amazon Mechanical Turk (AMT) via qualification rounds before we rely on them to generate high-quality sub-questions.

Worker Training - We manually annotate 100 questions from the VQA dataset as Perception and 100 as Reasoning questions, to serve as examples. We first teach workers the difference between Perception and Reasoning questions by defining them and showing several examples of each, along with explanations. Then, workers are shown (question, answer) pairs and are asked to identify if the given question is a Perception question or a Reasoning question [7]. Finally, for Reasoning questions, we ask workers to add all Perception questions and corresponding answers (in short) that would be necessary to answer the main question. In this qualification HIT, workers have to make 6 Perception and Reasoning judgments, and they qualify if they get 5 or more answers right. This interface can be found here.

We launched further pilot experiments for the workers who passed the first qualification round, where we manually evaluated the quality of their sub-questions based on 2 criteria : (1) The sub-questions should be Perception questions grounded in the image, and 2) The sub-questions should be sufficient to answer the main Reasoning question. Among those 540 workers who passed the first qualification test, 144 were selected (via manual evaluation) as high-quality workers, which finally qualified for attempting our main task.

Main task - In the main data collection, all VQA questions that got identified as Reasoning by regex-rules (section B) and a random subset of the questions identified as Perception were further judged by workers (for validation purposes). We eliminated ambiguous questions by further filtering out questions where there is high worker disagreement about the answer. We require at least 8 out of 10 workers to agree with the majority answer for yes/no questions and 5 out of 10 for all other questions, which leaves us with a split that corresponds to $\sim$13% of the VQA dataset. This interface can be found here.

C.1. VQA-introspect

Each <question, image> pair labeled as Reasoning had sub questions generated by by 3 unique workers [8]. On average we have 2.60 sub-questions per Reasoning question.

Randomly sampled qualitative examples from our collected dataset are shown in Fig. 7.

[7] We also add an "Invalid" category to flag nonsensical questions or those which can be answered without looking at the image

[8] A small number of workers displayed degraded performance after the qualification round, and were manually filtered

Rules				Amount of Data	
Starts with	Contains	Not contains	Length	# questions	% data
How many	-	-	-	48656	10.96
-	color	-	-	47956	10.81
What is the	-	-	-	40988	9.24
What	on	-	-	29031	6.54
What	in	-	-	21876	4.93
Is there	-	-	-	16494	3.72
-	wear	['appropriate', 'acceptable', 'etiquitte']	-	15530	3.50
-	wearing	-	-	14940	3.37
Is this a	-	-	4	14814	3.34
Where	-	-	-	12409	2.80
-	old	-	-	11197	2.52
What kind of	-	-	-	11186	2.52
What are	-	-	-	10524	2.37
-	on?	-	-	9040	2.04
Are there	-	-	-	8665	1.95
What type of	-	-	-	7955	1.79
-	doing?	-	-	7288	1.64
-	holding	-	-	7137	1.61
-	low	-	-	6596	1.49
-	round?	-	-	6242	1.41
Do	have	-	-	6213	1.40
Is the	on the	-	-	5375	1.21
Are these	-	['homemade', 'healthy', 'domesticated', etc.]	3	5320	1.20
Is the	in the	['wild', 'mountain', 'desert', 'woods', etc.]	-	5108	1.15
Does	have	-	-	5078	1.14
-	number	-	-	4477	1.01
What is this	-	-	-	3970	0.89
Is	ed?	['overexposed?', 'doctored?', 'ventilated?', etc.]	3	3940	0.88
Is	ing?	['horrifying?', 'relaxing?', 'competing?', etc.]	3	3870	0.88
Is	on	-	3	3622	0.82
Who	on	-	-	3563	0.80
-	shown?	-	-	3501	0.79
What sport	-	-	-	3412	0.77
-	sun	-	-	3260	0.73
-	see	-	-	3238	0.73
-	visible	-	-	3076	0.69
What	say?	-	-	3238	0.69
What	playing?	-	-	3076	0.69
Are the	in the	['US', 'wild', 'team', 'or', etc.]	-	3010	0.68
What	playing?	-	-	3076	0.69
Are	on the	-	-	2932	0.66

Table 2: Our rules for eliminating perception questions. Length refers to the # words in the question. We show top-40 rules based on data eliminated.

Main Reasoning Question:
- Is this at a residence or restaurant? "Restaurant"

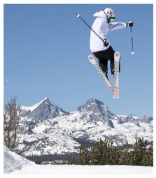
Perception Sub-questions:
- Does the napkin have a name of a cafe or a restaurant written on it? "Yes"

(a)

Main Reasoning Question:
- Has the pizza cutter been used yet? "No"

Perception Sub-questions:
- Is the pizza cut? "no"
- Is the pizza cut up yet, or one piece? "One piece"

(b)



Main Reasoning Question:
- Is he going to land safely? "Yes"

Perception Sub-questions:
- Are the skis pointed toward the ground? "Yes"
- Is the person facing the direction that they are falling? "Yes"
- Are the skis below the persons body and above the ground? "Yes"
- Is the person in the air? "Yes"

(c)

Main Reasoning Question:
- Is this a happy couple? "Yes"

Perception Sub-questions:
- Are the people smiling? "Yes"
- Is the bride smiling? "Yes"
- Did the couple just get married? "Yes"

(d)

Main Reasoning Question:
- Would it be safe to suggest most of the vegetation shown would not hide this animal? "Yes"

Perception Sub-questions:
- What kind of animal is this? "Elephant"
- Is the elephant larger than the tree trunks? "Yes"
- What kind of plants are around the elephant? "Trees"

(e)

Main Reasoning Question:
- Could you pick up this pizza to eat it? "No"

Perception Sub-questions:
- Is the sauce thick and running down the side? "Yes"
- Is the topping heavy, loose with lots of sauce? "Yes"
- Is there a very thick slice of pizza? "Yes"

(f)

Main Reasoning Question:
- Are the animals in their natural habitat? "Yes"

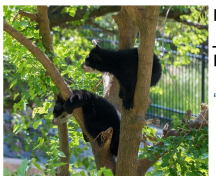
Perception Sub-questions:
- Is the bear touching a log in the water? "Yes"
- Does the bear have a wild water source to thrive in? "Yes"
- Is the bear in water? "Yes"
- Is the bear in a zoo? "No"
- Is the bear caged or fenced in? "No"

(g)

Main Reasoning Question:
- Is this a vegan dish? "Yes"

Perception Sub-questions:
- Is there any meat on the dish or dairy? "No"
- Are only vegetables shown? "Yes"
- Is there meat or dairy on the plate? "No"

(h)



Main Reasoning Question:
- Was this picture taken in Australia? "Yes"

Perception Sub-questions:
- What type of animals are climbing the tree? "small bears"
- Are there koala bears in the tree? "Yes"

(i)

Main Reasoning Question:
- Do you need practice to use one of these? "Yes"

Perception Sub-questions:
- Are the person's feet on a snowboard? "Yes"
- What is the object the person is standing on? "Snowboard"
- Has a person fallen? "Yes"

(j)

Main Reasoning Question:
- Is the cat a tabby? "No"

Perception Sub-questions:
- What is the color of the cat? "Black"
- Is the cat orange? "No"
- What color is the cat? "Yes"

(k)

Main Reasoning Question:
- Was this taken during the day? "No"

Perception Sub-questions:
- Is it dark outside? "Yes"

(l)

Main Reasoning Question:
- Could this be a foreign country? "Yes"

Perception Sub-questions:
- What type of vehicles are on the street? "Scooters"
- Are there people riding a motorbike? "Yes"
- Are the helmets the style worn in America? "No"

(m)

Main Reasoning Question:
- Are these dangerous for babies? "No"

Perception Sub-questions:
- What food is shown? "Peanut butter and banana"
- What food is on the plate? "Banana"
- Is there pudding on the plate? "Yes"

(n)

Figure 7: Randomly sampled qualitative examples of Perception sub-questions in our VQA-introspect dataset for main questions in the Reasoning split of VQA. Main questions are written in orange and sub questions are in blue. A single worker may have provided more than one sub questions for the same (image, main question) pair.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 2

[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. 2, 3, 4

[4] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *EMNLP*, 2016. 2

[5] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 1

[6] 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

[7] Donald D Hoffman and Whitman A Richards. Parts of recognition. *Cognition*, 18(1-3):65–96, 1984. 1

[8] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. 8

[9] 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. 3

[10] 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. 2, 3

[11] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 6, 7

[12] Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1338–1346, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 2

[13] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 6

[14] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *AAAI*, 2018. 2, 3

[15] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961, 2015. 2

[16] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3

[18] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3

[19] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019. 2

[20] 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