Cycle-Consistency for Robust Visual Question Answering

Meet Shah[1] , Xinlei Chen[1], Marcus Rohrbach[1], Devi Parikh[1,2]

[1]Facebook AI Research, [2]Georgia Institute of Technology

{meetshah, xinleic, mrf}@fb.com, dparikh@gatech.edu

Abstract

Despite significant progress in Visual Question Answering over the years, robustness of today's VQA models leave much to be desired. We introduce a new evaluation protocol and associated dataset (VQA-Rephrasings) and show that state-of-the-art VQA models are notoriously brittle to linguistic variations in questions. VQA-Rephrasings contains 3 human-provided rephrasings for 40k questions spanning 40k images from the VQA v2.0 validation dataset. As a step towards improving robustness of VQA models, we propose a model-agnostic framework that exploits cycle consistency. Specifically, we train a model to not only answer a question, but also generate a question conditioned on the answer, such that the answer predicted for the generated question is the same as the ground truth answer to the original question. Without the use of additional annotations, we show that our approach is significantly more robust to linguistic variations than state-of-the-art VQA models, when evaluated on the VQA-Rephrasings dataset. In addition, our approach outperforms state-of-the-art approaches on the standard VQA and Visual Question Generation tasks on the challenging VQA v2.0 dataset.

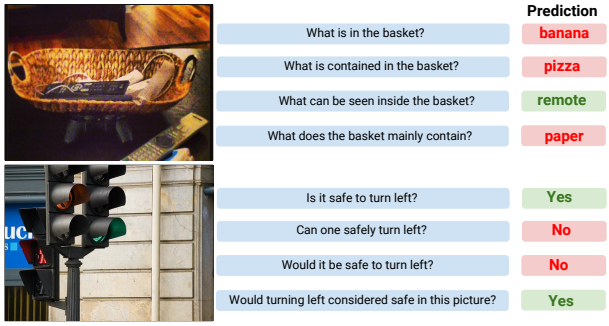


Figure 1. **Existing VQA models are brittle**. Shown above are examples from our new large-scale **VQA-Rephrasings** dataset that enables systematic evaluation of robustness of VQA models to linguistic variations in the input question. Also shown are answers predicted by a state-of-the-art VQA model [41]. We see that the model predicts different answers for different reasonable rephrasings of the same question. We propose a novel model-agnostic framework that exploits cycle consistency in question answering and question generation to make VQA models more robust, without using additional annotation. Moreover, it outperforms state-of-the-art models on the standard VQA and Visual Question Generation tasks on the VQA v2.0 dataset.

1. Introduction

Visual Question Answering (VQA) applications allow a human user to ask a machine questions about images – be it a user interacting with a visual chat-bot or a visually impaired user relying on an assistive device. As this technology steps out of the realm of curated datasets towards real-world settings, it is desirable that VQA models be robust to and consistent across reasonable variations in the input modalities. While there has been significant progress in VQA over the years [1, 17, 2, 9, 19, 41, 3, 4], today's VQA models are, however, far from being robust.

VQA is a task that lies at the intersection of language and vision. Existing works have studied the robustness and sensitiveness of VQA models to meaningful semantic variations in images [9], changing answer distributions [2] and adversarial attacks [39] to images. However, to the best of our knowledge, no work has studied the robustness of VQA models to linguistic variations in the input question. This is important both from the perspective of VQA being a benchmark to test multi-modal AI capabilities (do our VQA models really "understand" the question when answering it?) and for applications (human users are likely to phrase the same query in a variety of different linguistic forms). However, today's state-of-the-art VQA models are brittle to such linguistic variations as can be seen in Fig. 1.

One approach to make VQA models more robust is to collect a dataset with diverse rephrasings of questions to train VQA models. Alternatively, an automatic approach that does not require additional human intervention but results in a VQA model that is robust to linguistic variations observed in the natural language questions is desirable.

We propose a novel model-agnostic framework that relies on cycle consistency to learn robust VQA models without requiring additional annotation. Specifically, we train

the model to not just answer a question, but also to generate diverse, semantically similar variations of questions conditioned on the answer. We enforce that the answer predicted for a generated question matches the ground truth answer to the original question. In other words, the model is being trained to predict the same (correct) answer for a question and its (generated) rephrasing.

Advantages of our proposed approach are two-fold. First, enforcing consistent correctness across diverse rephrasings allows models to generalize to unseen semantically equivalent variations of questions at test time. The model achieves this by generating linguistically diverse rephrasings of questions on-the-fly and training with these variations. Second, a model trained generatively to generate a valid question given a candidate answer and image has a stronger multi-modal understanding of vision and language. Questions tend to have less learnable biases [26]. As a result, models that can jointly perform the task of question generation and question answering are less prone to taking "shortcuts" and exploiting linguistic priors in questions. Indeed, we find that models trained with our approach outperform existing state-of-the-art models on both VQA and Visual Question Generation (VQG) tasks on VQA v2.0 [9].

We also observed that one reason for limited development of VQA models robust to linguistic variations in input questions is due to the lack of a benchmark to measure robustness. A lack of such a benchmark makes it hard to quantitatively realize the inflated capabilities and limited multi-modal understanding of modern VQA models . To enable quantitative evaluation of robustness and consistency of VQA models across linguistic variations in input questions, we collect a large-scale dataset – **VQA-Rephrasings** (Section 4) based of the VQA v2.0 dataset [9]. VQA-Rephrasings contains 3 human-provided rephrasings for ∼40k questions on ∼40k images from the validation split of the VQA v2.0 dataset. We also propose metrics to measure the robustness of VQA models across different question rephrasings. Further, we benchmark several state-of-the-art VQA models [3, 5, 19, 41] on our proposed VQA-Rephrasings dataset to highlight the fragility of VQA models to question rephrasings. We observe a significant drop when VQA models are required to be consistent in addition to being correct (Section 5), which reinforces our belief that existing VQA models do not understand language "enough". We show that VQA models trained with our approach are significantly more robust across question rephrasings than their existing counterparts on the proposed VQA-Rephrasings dataset.

In this paper, our contributions are the following:

- We propose a model-agnostic cycle-consistent training scheme that enables VQA models to be more robust to linguistic variations observed in natural language open-ended questions.
- To evaluate the robustness of VQA models to linguistic variations, we introduce a large-scale **VQA-Rephrasings** dataset and an associated consensus score. VQA-Rephrasings consists of 3 rephrasings for ∼40k questions on ∼40k images from the VQA v2.0 validation dataset, resulting in a total of ∼120k question rephrasings by humans.
- We show that models trained with our approach outperform state-of-the-art on the standard VQA and Visual Question Generation tasks on the VQA v2.0 dataset and are significantly more robust to linguistic variations on VQA-Rephrasings.

2. Related Work

Visual Question Answering. There has been tremendous progress in building models for VQA using LSTMs [13] and convolutional networks [22]. VQA models spanning paradigms like attention networks [40, 19], module networks [14, 4, 17], relational networks [32] and multi-modal fusion [5] have been proposed. Our method is model-agnostic and is applicable with any VQA architecture.

Robustness. Robustness of VQA models has been studied in several contexts [2, 39, 9]. For example, [2] studies the robustness of VQA models to changes in the answer distributions across training and test settings; [42] analyzes the extent of visual grounding in VQA models by studying robustness of VQA models to meaningful semantic changes in images; [39] shows that despite the use of an advanced attention mechanism, it is easy to fool a VQA model with very minor changes in the image. Our work, however, aims to complete the study in robustness by benchmarking and improving robustness of VQA models to linguistic and compositional variations in questions in the form of rephrasings. Robustness has also been studied in natural language processing (NLP) systems [7, 12] in contexts of bias [35, 34], domain-shift [23] and syntactic variations [15]. We study this in the context of visual question answering which is a multi-modal task which grounds language into the visual world.

(Visual) Question Generation. Generating questions conditioned on an image was introduced in [29] and a large-scale VQG dataset was collected by [30] to evaluate visually grounded question generation capabilities of models. More recently, there has been work on generating questions that are diverse [16, 40]. While these techniques generate questions about an image in an answer-agnostic manner, techniques like [26] propose a variational LSTM based model trained with reinforcement learning to generate answer-specific questions for an image. More recently, [24] generates answer-specific questions for specific question-types by modelling question generation as a dual task of question

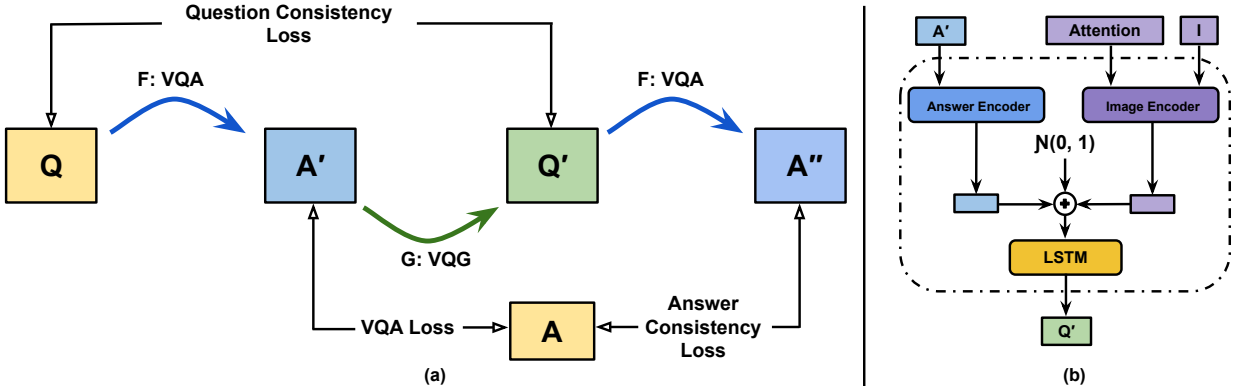


Figure 2. (a) **Abstract representation of the proposed cycle-consistent training scheme:** Given a triplet of image I, question Q, and ground truth answer A, a VQA model is a transformation $F : (Q, I) \mapsto A'$ used to predict the answer A'. Similarly, a VQG model $G : (A', I) \mapsto Q'$ is used to generate a rephrasing Q' of Q. The generated rephrasing Q' is passed through F to obtain A'' and consistency is enforced between Q and Q' and between A' and A''. Image I is not shown for clarity. (b) **Detailed architecture of our visual question generation module** G. The predicted answer A' and image I are embedded to a lower dimension using task-specific encoders and the resulting feature maps are summed up with additive noise and fed to an LSTM to generate questions rephrasings Q'.

answering. Unlike [24], our method is not restricted to generating questions only for specific question types. Different from previous works, the goal of our VQG component is to automatically generate question rephrasings that make the VQA models more robust to linguistic variations. To the best of our knowledge, we are the first to demonstrate that the VQG module can be used to improve VQA accuracy in a cycle-consistent setting.

Cycle-Consistent Learning. Using cycle-consistency to regularize the training of models has been used extensively in object tracking [36], machine translation [10], unpaired image-to-image translation [43] and text-based question answering [37]. Consistency enables learning of robust models by regularizing transformations that map one interconnected modality or domain to the other. While cycle consistency has been used vastly in the domains involving a single modality (text-only or image-only), it hasn't been explored in the context of multi-modal tasks like VQA. Cycle-consistency in VQA can be also thought of as an online data-augmentation technique where the model is trained on several generated rephrasings of the same question.

3. Approach

We now introduce our cycle-consistent scheme to train robust VQA models. Given a triplet of image I, question Q, and ground truth answer A, a generic VQA model can be formulated as a transformation $F : (Q, I) \mapsto A'$, where A' is the answer predicted by the model as in Fig. 2(a). Similarly, a generic VQG model can be formulated as a transformation $G : (A, I) \mapsto Q'$ as in Fig. 2(b). For a given (I, Q, A) triplet, we first obtain an answer prediction A' using the VQA model F for the original question Q. We then use the predicted answer A' and the image I to generate a question Q' which is semantically similar to Q using the VQG model G. Lastly, we obtain a answer prediction A'' for the generated question Q'.

Our design of consistency components is inspired by two beliefs. Firstly, a model which can generate a semantically and syntactically correct question given a answer and an image, has a better understanding of the cross-modal connections among the image, the question and the answer, which make them a valid (I, Q, A) triplet. Secondly, assuming the generated question Q' is a valid rephrasing of the original question, a robust VQA model should answer this rephrasing with the same answer as the original question Q. In practice, however, there are several challenges that inhibit enforcement of cycle-consistency in VQA. We discuss these challenges and describe the key components of our framework geared to tackle them in the following sections.

3.1. Question Generation Module

Since VQA is a setting where there is high disparity in the information content of involved modalities (a question and answer pair is a very lossy compressed representation of the image), learning transformations that map one modality to another is non-trivial. In cycle-consistent models dealing with single-modalities, transformations need to be learned across different domains of the same modality (image or text) with roughly similar information contents. However in a multi-modality transformation like VQG, learning a transformation from a low information modality (such as answer) to high information modality (question) needs additional supervision. We provide this additional supervision to the VQG model in the form of attention. To generate a rephrasing Q', the VQG is guided to attend at regions of the image which were used by the VQA model to answer the original question Q. Unlike [24], this enables our models

to generate questions more similar to the original question from answers like "yes", which could possibly have a large space of plausible questions.

We model the question generation module G in a fashion similar to a conditional image captioning model. The question generation module consists of two linear encoders that transform attended image features obtained from VQA model and the distribution over answer space to lower dimensional feature vectors. We sum these feature vectors with additive noise and pass them through an LSTM which is trained to reconstruct the original question and optimized by minimizing the negative log likelihood with teacher-forcing. Note that unlike [26, 24] we do not pass the one-hot vector representing the answer obtained, or an embedding of the answer obtained to the question generation, but rather the predicted distribution over answers. This enables the question generation module to learn to map the model's confidence over answers to the generated question.

Throughout the paper, **Q-consistency** implies addition of a VQG module G on top of the base VQA model F to generate rephrasings Q' from the image I and the predicted answer A' with an associated Q-consistency loss $\mathcal{L}_G(Q, Q')$. Similarly, **A-consistency** implies passing all questions generated Q' by the VQG Model G to the VQA model F and an associated A-consistency loss $\mathcal{L}_{cycle}(A, A'')$. The overall loss can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_F(A, A') + \lambda_G \mathcal{L}_G(Q, Q') + \lambda_C \mathcal{L}_{cycle}(A, A'') \quad (1)$$

where $\mathcal{L}_F(A, A')$ and $\mathcal{L}_{cycle}(A, A'')$ (*i.e.* A-Consistency Loss) are cross-entropy losses, $\mathcal{L}_G(Q, Q')$ (*i.e.* Q-Consistency Loss) is sequence generation loss [28] and λ_G, λ_C are tunable hyperparameters.

3.2. Gating Mechanism

One of the assumptions of our proposed cycle-consistent training scheme is that the generated question is always semantically and syntactically correct. However, in practice this is not always true. Previous attempts [18] at naively generating questions conditioned on the answer and using them without filtering to augment the training data have been unsuccessful. Like the visual question answering module, the visual question generation module is also not perfect. Therefore not all questions generated by the question generator are coherent and consistent with the image, the answer and the original question. To overcome this issue, we propose a gating mechanism, which automatically filters undesirable questions generated by the VQG model before passing them to the VQA model for A-consistency. The gating mechanism is only relevant when used in conjunction with A-consistency. We retain only those questions which either the VQA model F can answer correctly or have a cosine similarity with the original question encoding greater than a threshold T_{sim}.

3.3. Late Activation

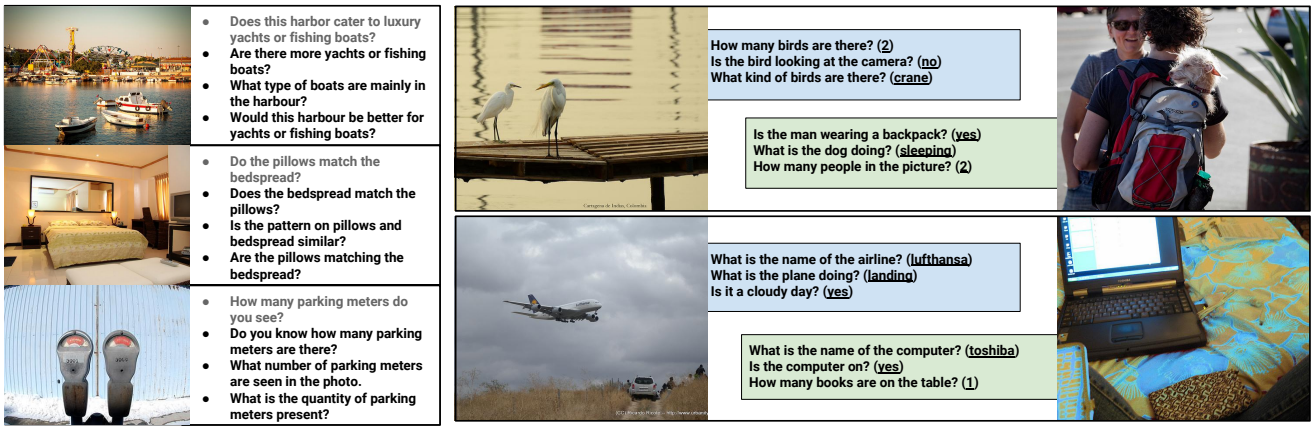
One key component of designing cycle consistent models is to prevent mode collapse. Learning cycle-consistent models in complex settings like VQA needs a carefully chosen training scheme. Since cycle-consistent models have several interconnected sub-networks learning different transformations, it is important to ensure that each of these sub-networks are working in harmony. For example, if the VQA model F and VQG model G are jointly trained and consistency is enforced in early stages of training, it is possible that both models can just "cheat" by both producing undesirable outputs. We overcome this by activating cycle-consistency at later stages of training, to make sure both VQA and VQG models have been sufficiently trained to produce reasonable outputs. Specifically, we enable the loss associated with cycle-consistency after a fixed A_{iter} iterations in the training process.

We find these design choices for question generation module, gating mechanism and late activation to be crucial for effectively training our model. We demonstrate this empirically via ablation studies in Table 2. As we want to increase the robustness of the VQA model to all generated variations, the weights between VQA models which answer the original question and the generated rephrasing are shared. Our formulation of cycle-consistency in VQA can be also thought of as an online data-augmentation technique where the model is trained on several generated rephrasings of the same question and hence is more robust to such anomalies during inference. We show that with clever training strategy, coupled with attention and carefully chosen architecture for question generation, incorporating cycle consistency for VQA is possible and not only leads to models that are better performing, but also more robust and consistent. In addition, we show that this robustness also imparts VQA models the ability to better predict their own failures.

4. VQA-Rephrasings Dataset

In this section, we introduce the VQA-Rephrasings dataset, which is the first dataset that enables evaluation of VQA models for robustness and consistency to different rephrasings of questions with the same meaning.

We use the validation split of VQA v2.0 [9] as our base dataset which contains a total of 214,354 questions spanning over 40,504 images. We randomly sample 40,504 questions (one question per image) from the base dataset to form a sampled subset. We collect 3 rephrasings of each question in the sampled subset using human annotators in two stages. In the first stage, humans were primed with the original question and the corresponding true answer and asked to rephrase the question such that answer



(a) (b)

Figure 3. **(a) Qualitative examples from our VQA-Rephrasings dataset**. The first question (shown in gray) in each block is the original question from VQA v2.0 validation set, the questions that follow (shown in black) are rephrasings collected in VQA-Rephrasings. **(b) Qualitative examples of answer conditioned question generation (input answer) by our VQG module**

to the rephrased question remains the same as the original answer. To ensure rephrasings from first stage are *syntactically* correct and *semantically* inline with the original question, we filter the collected responses in the next stage.

In the second stage, humans were primed with the original question and it's rephrasing and were asked to label the rephrasing invalid if: (a) the plausible answer to the original question and it's rephrasing is different (*i.e.* if the question and it's rephrasing have different intents) or (b) if the rephrasing is grammatically incorrect. We collected 121,512 rephrasings from the original 40504 questions in the first stage. Of these, 1320 rephrasings were flagged as invalid in the second stage and were rephrased again in the first stage. The final dataset consists of 162,016 questions (including the original 40,504 questions) spanning 40,504 images with an average of ∼3 rephrasings per original question. A few qualitative examples from the collected dataset can be seen in Fig. 3(a). Additional details about the data collection, interfaces used and exhaustive dataset statistics can be found in supplementary materials.

Consensus Score. Intuitively, for a VQA model to be consistent across various rephrasings of the same question, the answer to all rephrasings should be the same. We measure this by a Consensus Score $CS(k)$. For every group Q consisting of n rephrasings, we sample all subsets of size k. The consensus score $CS(k)$ is defined as the ratio of the number of subsets where *all* the answers are correct and the total number of subsets of size k. The answer to a question is considered correct if it has a non-zero VQA Accuracy θ as defined in [1]. CS(k) is formally defined as:

$$CS(k) = \sum_{Q' \subset Q, |Q'|=k} \frac{\mathcal{S}(Q')}{{}^{n}C_{k}} \quad (2)$$

$$\mathcal{S}(Q') = \begin{cases} 1 & \text{if } \forall q \in Q' \ \theta(q) > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Where ${}^{n}C_{k}$ is number of subsets of size k sampled from a set of size n. As consensus score is a all-or-nothing score, to achieve a non-zero consensus score at k for a group of questions Q, the model has to answer at least k questions correctly in a group of questions Q. When $k = |Q|$ (*e.g.* when $k = 4$ in VQA-Rephrasings), the model needs to answer all rephrasings of a question and the original question correctly in order to get a non-zero consensus score. It is evident that a model with higher average consensus score at high values of k is quantitatively more robust to linguistic variations in questions than a model with a lower score.

5. Experiments

5.1. Consistency Performance

We start by benchmarking a variety of existing VQA models on our proposed VQA-Rephrasings dataset.

MUTAN [5] [1] parametrizes bilinear interactions between visual and textual representations using a multimodal low-rank decomposition. MUTAN uses skip-thought [20] sentence embeddings to encode the question and Resnet-152 [11] to encode images. MUTAN achieves 63.20% accuracy on VQA v2.0 test-dev. Among all models we analyze, MUTAN is the only model which uses sentence embeddings to encode questions.

Bottom-Up Top-Down Attention (BUTD) [3] [2] incorporates bottom-up attention in VQA by extracting features associated with image regions proposed by Faster-RCNN [33] pretrained on Visual Genome [21]. BUTD

[1] https://github.com/Cadene/vqa.pytorch
[2] https://github.com/hengyuan-hu/bottom-up-attention-vqa

Model	CS(k)				VQA Accuracy	
	k=1	k=2	k=3	k=4	ORI	REP
MUTAN [5]	56.68	43.63	38.94	32.76	59.08	46.87
BUTD [3]	60.55	46.96	40.54	34.47	61.51	51.22
BUTD + CC	**61.66**	**50.79**	**44.68**	**42.55**	**62.44**	**52.58**
Pythia [41]	63.43	52.03	45.94	39.49	64.08	54.20
Pythia + CC	**64.36**	**55.45**	**50.92**	**44.30**	**64.52**	**55.65**
BAN [19]	64.88	53.08	47.45	39.87	64.97	55.87
BAN + CC	**65.77**	**56.94**	**51.76**	**48.18**	**65.87**	**56.59**

Table 1. **Consensus performance on VQA-Rephrasings dataset**. CS(k) as defined in Eq. 2 is consensus score which is non-zero only if *at least* k rephrasings are answered correctly, zero otherwise; averaged across all group of questions. ORI represent a split of questions from VQA-Rephrasings which are original questions from VQA v2.0 and their corresponding rephrasings are represented by the split REP. Models trained with our cycle-consistent (CC) framework consistently outperform their baseline counterparts at all values of k.

Model	val	test-dev
MUTAN [5]	61.04	63.20
BUTD [3]	65.05	66.25
+ Q-consistency	65.38	66.83
+ A-consistency	60.84	62.18
+ Gating	**65.53**	**67.55**
Pythia [41]	65.78	68.43
+ Q-consistency	65.39	68.58
+ A-consistency	62.08	63.77
+ Gating	**66.03**	**68.88**
BAN [19]	66.04	69.64
+ Q-consistency	66.27	69.69
+ A-consistency	64.96	66.31
+ Gating	**66.77**	**69.87**

Table 2. **VQA Performance and ablation studies on VQA v2.0 validation and test-dev splits**. Each row in blocks represents a component of our cycle-consistent framework added to the previous row. First row in each block represents the baseline VQA model F. Q-consistency implies addition of a VQG module G to generate rephrasings Q' from the image I and the predicted answer A' with an associated VQG loss $\mathcal{L}_{vqg}(Q, Q')$. A-consistency implies passing all the generated questions Q' to the VQA model F and an associated loss $\mathcal{L}_{cycle}(A, A')$. Gating implies the use of gating mechanism to filter undesirable generated questions in Q' and passing the remaining to VQA model F. Last row in each block is equivalent to the base VQA model (first row in each block) + cycle-consistency (CC) as used in other tables. Models trained with our cycle-consistent (last row in each block) framework consistently outperform baselines.

model won the VQA Challenge in 2017 and achieves 66.25% accuracy on VQA v2.0 test-dev.

Pythia [41] [3] extends the BUTD model by incorporating co-attention [27] between question and image regions. Pythia uses features extracted from Detectron [8] pretrained on Visual Genome. An ensemble of Pythia models won the 2018 VQA Challenge using extra training data from Visual Genome [21] and using Resnet[11] features. In this study, we use Pythia models which do not use Resnet features.

Bilinear Attention Networks (BAN) [19] [4] combines the idea of bilinear models and co-attention [27] between image regions and words in questions in a residual setting. Similar to [3], it uses Faster-RCNN [33] pretrained on Visual Genome [21] to extract image features. In all our experiments, for a fair comparison, we use BAN models which do not use additional training data from Visual Genome. BAN achieves the current state-of-the-art single-model accuracy of 69.64 % on VQA v2.0 test-dev without using additional training data from Visual Genome.

Implementation Details For all models trained with our cycle-consistent framework, we use the values T_{sim}=0.9, λ_G=1.0, λ_C=0.5 and A_{iter}=5500. When reporting results on the validation split and VQA-Rephrasings we train on the training split and when reporting results on the test split we train on both training and validation splits of VQA v2.0. Note that we *never* explicitly train on the collected VQA-Rephrasings dataset and use it purely for evaluation purposes. We use publicly available implementations of each backbone VQA model.

We measure the robustness of each of these models on our proposed VQA-Rephrasings dataset using the consensus score (Eq. 2). Table 1 shows the consensus scores at different values of k for several VQA models. We see that all models suffer significantly when measured for consistency across rephrasings. For *e.g.*, the performance of Pythia (winner of 2018 VQA challenge) has a consensus score of 39.49% at $k = 4$. Similar trends are observed for MUTAN, BAN and BUTD. The drop increases with increasing k, the number of rephrasings used to measure consistency. Models like BUTD, BAN and Pythia which use word-level encodings of the question suffer significant drops. It is interesting to note that even MUTAN which uses skip-thought based sentence encoding [20] suffers a drop when checked for consistency across rephrasings (from $k = 1$ to $k = 4$). We observe that BAN + CC model trained with our proposed cycle-consistent training framework outperforms its counterpart BAN and all other models at all values of k.

Fig 4 qualitatively compares the textual and visual attention (over image regions) over 4 rephrasings of a question. The top row shows attention and predictions from a Pythia model, while the bottom row shows attention and predic-

[3] https://github.com/facebookresearch/pythia

[4] https://github.com/jnhwkim/ban-vqa

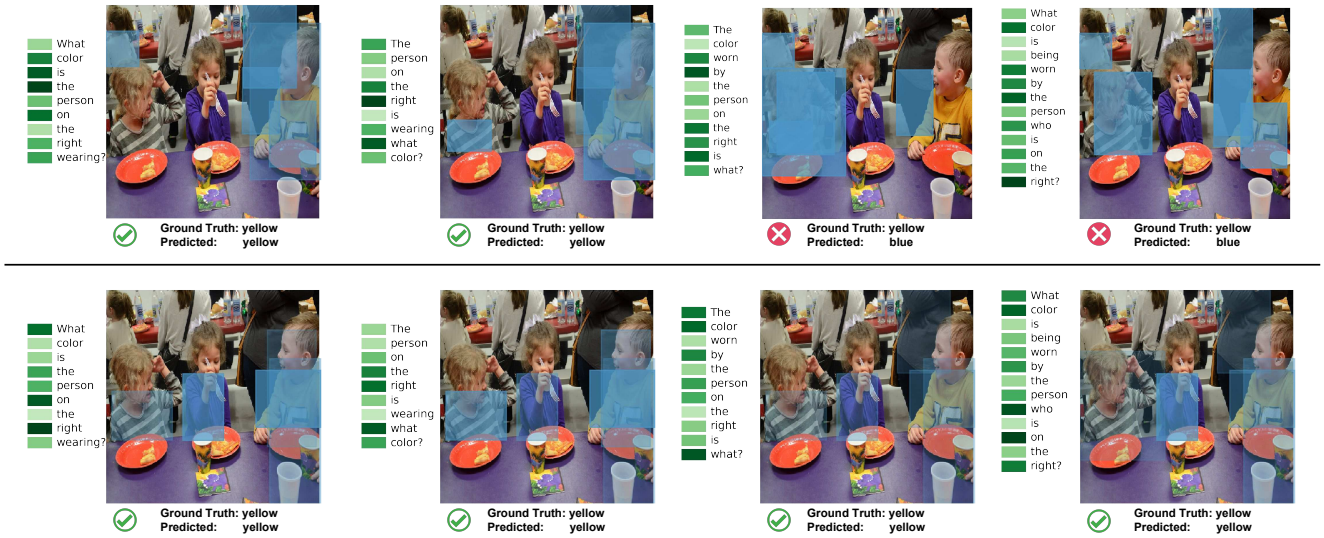


Figure 4. **Visualization of textual and image region attention across question variants:** The top row shows attention and predictions from a Pythia [41] model, the bottom row shows attention and predictions from the same Pythia model, but trained using our cycle-consistent approach. Our model attends to relevant image regions for all rephrasings and answers them correctly. The baseline Pythia counterpart, however, fails to attend over relevant image regions for some rephrasings.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDER
iQAN* [24]	0.582	0.467	0.385	0.320	0.617	0.276	2.222
Pythia + CC*	**0.708**	**0.561**	**0.438**	**0.339**	**0.627**	**0.284**	**2.301**
iVQA [26]	0.430	0.326	0.256	0.208	0.468	0.205	1.714
Pythia + CC	**0.486**	**0.368**	**0.287**	**0.226**	**0.556**	**0.225**	**1.843**

Table 3. **Question Generation Performance on VQA v2.0 validation set**, * signifies results on a constrained subset as done in [24]. CC represents models trained with our approach.

tions from the same Pythia model, but trained using our framework. Our model attends at relevant image regions for all rephrasings and answers all of them correctly. This qualitatively demonstrates the robustness of models trained with our framework.

5.2. Visual Question Answering Performance

We now evaluate our approach and various ablations on the standard task of question answering on VQA v2.0 dataset [9]. We compare the performance of several VQA models on the validation and test-dev splits of VQA v2.0. Table 2 shows the VQA scores of different models on validation and test-dev splits. We show that BUTD, Pythia and BAN models trained with our cycle-consistent framework outperform their corresponding baselines.

We show the impact of each component of our cycle-consistent framework by performing ablation studies on our models. We study the marginal effect of components like question consistency (Q-consistency), answer consistency (A-consistency) and gating mechanism by adding them step-by-step to the base VQA model F. Q-consistency implies addition of a VQG module G to generate rephrasings Q' from the image I and the predicted answer A' with an associated VQG loss $\mathcal{L}_{vqg}(Q, Q')$. As shown in Table 2, we see that addition of question consistency slightly improves performance of each VQA model. Inline with observations in [24], this shows that indeed models which can generate questions from the answer have better multi-modal understanding and in turn are better at visual question answering. A-consistency implies passing all the generated questions Q' to the VQA model F and an associated loss $\mathcal{L}_{cycle}(A, A')$. As seen in Table 2, we see that naively passing all the generated questions to the VQA model F leads to significant reduction in performance than the base model F. This goes in line with our earlier discussion that not all questions generated are *valid* rephrasings of the original question and hence enforcing consistency between the answers of two invalid pairs of questions naturally leads to degradation in performance. Finally, we show the effect of using our gating mechanism to filter undesirable generated questions in Q' and passing the remaining to VQA model F. We see that all VQA models perform consistently better when using a gating than just using Q-consistency.

We also experimented with Pythia model configurations

where the VQG model uses unattended image features (unlike the default setting which uses image features with attention from the VQA model). We found that with this configuration, our approach still shows improved performance over the baseline. However, the question generation quality is relatively poor, and the overall gain is smaller (3.58% in consistency $CS(k = 4)$ and 0.2% in VQA accuracy) compared to when using attention (8.08% and 0.5% respectively) – likely because attention helps in generating more-focused rephrasings.

5.3. Visual Question Generation Performance

Recall that our model also includes a VQG component which generates questions conditioned on an answer and image. Since the overall performance of our framework relies highly on the performance of question generation module, we evaluate our VQG component performance as well on commonly used image captioning metrics. We compare our VQG component to several answer-conditional VQG models on the VQA v2.0 dataset. We use standard image captioning metrics CIDEr [38], BLEU [31], METEOR [6] and ROUGE-L [25] as used in [26]. We compare our approach to two recently proposed visual question generation approaches. **iVQA** [26] uses a variational LSTM model trained with reinforcement learning to generate answer-specific questions for an image. Syntactic correctness, diversity and intent of the generated question are used to allocate rewards. **iQAN** [24] generates answer-specific questions by modelling question generation as a dual task of question answering and sharing parameters between question answering and question generation modules. Since iQAN can only generate a specific type of questions, for a fair comparison, we compare to iQAN only on a subset of the dataset containing questions from these specific types. As shown in Table 3, we observe that our question generation module trained with cycle-consistency consistently outperforms iVQA [26] and iQAN [24] on all metrics. A few qualitative examples of answer conditioned questions generated by our VQG model can be seen in Fig. 3(b).

5.4. Failure Prediction Performance

In previous results, we show that by training models to generate and answer questions while being consistent across both tasks leads to improvement in performance and robustness. Another way of testing robustness of these models is to see if models can predict their own failures. A robust model is less confident about an incorrect answer and vice versa. Motivated by this, we seek to verify if models trained with our cycle-consistent framework can identify their own failures *i.e.* correctly identify if they're wrong about a prediction. To this end, we use two failure predictions schemes. First, we naively threshold the confidence of the predicted answer. All answers above a particular threshold are marked as correctly answered and vice versa. Second, we design a failure prediction binary classification module (FP), which predicts for a given image I, question Q and answer A' (predicted by the base VQA model F), whether the predicted answer is correct for the given (I, Q) pair. The FP module is trained keeping the parameters of the base VQA model frozen. In Table 4, we show the failure prediction performance of the baseline VQA models and models trained with our proposed framework. It shows that the cycle consistency framework, even *without* an explicit failure predictor module, makes the models more calibrated – more capable of detecting their own failures. In both settings: (a) when using naive confidence thresholding (not marked as "+ FP" in the Table) and (b) using a specifically designed submodule to detect failures (marked as "+ FP"), models trained with our cycle-consistent training have higher F1 scores than their corresponding baselines. We see similar improvements in detecting failures for both BUTD and Pythia models, which shows that our cycle-consistency framework is model agnostic. This also shows that not only does cycle-consistent training make models robust to linguistic variations, but also allows them to be failure-aware.

Model	Precision	Recall	F1
BUTD [3]	0.71	0.78	0.74
+ FP	**0.74**	**0.85**	**0.79**
BUTD + CC	0.73	0.79	0.76
+ FP	**0.78**	**0.83**	**0.80**
Pythia [41]	0.74	0.79	0.76
+ FP	**0.76**	**0.88**	**0.82**
Pythia + CC	0.77	0.81	0.77
+ FP	**0.82**	**0.84**	**0.83**

Table 4. **Failure prediction performance on VQA v2.0 validation dataset**. Each row in blocks represents a component added to the previous row. CC represents models trained with our cycle-consistent framework and FP represents models with an additional binary classification Failure Prediction submodule to predict if the predicted answer A' is correct given a question and image pair (Q, I). For models trained without the FP module, scores are obtained by thresholding the answer confidences.

6. Conclusion

In this paper, we propose a novel model-agnostic training strategy to incorporate cycle consistency in VQA models to make them robust to linguistic variations and self-aware of their failures. We also collect a large-scale dataset, VQA-Rephrasings and propose a consensus metric to measure robustness of VQA models to linguistic variations of a question. We show that models trained with our training strategy are robust to linguistic variations, and achieve state-of-the-art performance in VQA and VQG on VQA v2.0 dataset.

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT*, 2016.

[5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[7] Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. Towards linguistically generalizable nlp systems: A workshop and shared task. *arXiv preprint arXiv:1711.01505*, 2017.

[8] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018.

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334. IEEE, 2017.

[10] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Jerry R Hobbs, Douglas E Appelt, John Bear, and Mabry Tyson. Robust processing of real-world natural-language texts. In *Proceedings of the third conference on Applied natural language processing*, pages 186–192. Association for Computational Linguistics, 1992.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 804–813. IEEE, 2017.

[15] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*, 2018.

[16] Unnat Jain, Ziyu Zhang, and Alexander Schwing. Creativity: Generating diverse questions using variational autoencoders. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5415–5424. IEEE, 2017.

[17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017.

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

[19] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*, 2018.

[20] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Yitong Li, Trevor Cohn, and Timothy Baldwin. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 21–27, 2017.

[24] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[26] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8611–8619, 2018.

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[28] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering, 2018.

[29] Issey Masuda Mora and Santiago Pascual de la Puente. Towards automatic generation of question answer pairs from images.

[30] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*, 2016.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

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

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] Michael Spranger, Jakob Suchan, and Mehul Bhatt. Robust natural language processing-combining reasoning, cognitive semantics and construction grammar for spatial language. *arXiv preprint arXiv:1607.05968*, 2016.

[35] Manfred Stede. The search for robustness in natural language understanding. *Artificial Intelligence Review*, 6(4):383–414, 1992.

[36] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010.

[37] Duyu Tang, Nan Duan, Zhao Yan, Zhirui Zhang, Yibo Sun, Shujie Liu, Yuanhua Lv, and Ming Zhou. Learning to collaborate for question answering and asking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1564–1574, 2018.

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[39] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4951–4961, 2018.

[40] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[41] 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.

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

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.