Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions

Arijit Ray[1], **Gordon Christie**[1], **Mohit Bansal**[2], **Dhruv Batra**[3,1], **Devi Parikh**[3,1]
[1]Virginia Tech [2]UNC Chapel Hill [3]Georgia Institute of Technology
{ray93,gordonac,dbatra,parikh}@vt.edu
mbansal@cs.unc.edu

Abstract

Visual Question Answering (VQA) is the task of answering natural-language questions about images. We introduce the novel problem of determining the *relevance of questions to images* in VQA. Current VQA models do not reason about whether a question is even related to the given image (*e.g., What is the capital of Argentina?*) or if it requires information from external resources to answer correctly. This can break the continuity of a dialogue in human-machine interaction. Our approaches for determining relevance are composed of two stages. Given an image and a question, (1) we first determine whether the question is visual or not, (2) if visual, we determine whether the question is relevant to the given image or not. Our approaches, based on LSTM-RNNs, VQA model uncertainty, and caption-question similarity, are able to outperform strong baselines on both relevance tasks. We also present human studies showing that VQA models augmented with such question relevance reasoning are perceived as more intelligent, reasonable, and human-like.

1 Introduction

Visual Question Answering (VQA) is the task of predicting a suitable answer given an image and a question about it. VQA models (*e.g.,* (Antol et al., 2015; Ren et al., 2015)) are typically discriminative models that take in image and question representations and output one of a set of possible answers.

Our work is motivated by the following key observation – all current VQA systems always output an answer *regardless of whether the input question makes any sense for the given image or not*. Fig. 1 shows examples of relevant and irrelevant questions. When VQA systems are fed irrelevant questions as input, they understandably produce nonsensical answers (Q: *"What is the capital of Argentina?"* A: *"fire hydrant"*). Humans, on the other hand, are unlikely to provide such nonsensical answers and will instead answer that this is irrelevant or use another knowledge source to answer correctly, when possible. We argue that this implicit assumption by all VQA systems – that an input question is always relevant for the input image – is simply untenable as VQA systems move beyond standard academic datasets to interacting with real users, who may be unfamiliar, or malicious. The goal of this work is to make VQA systems more human-like by providing them the capability to identify relevant questions.

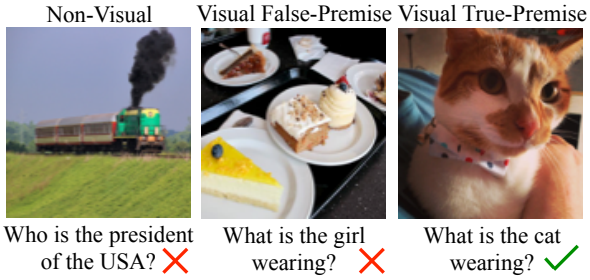

Figure 1: Example irrelevant (non-visual, false-premise) and relevant (visual true-premise) questions in VQA.

While existing work has reasoned about cross-modal similarity, being able to identify whether a question is relevant to a given image is a novel problem with real-world applications. In human-robot interaction, being able to identify questions that are dissociated from the perception data available is important. The robot must decide whether to process the scene it perceives or query external world knowledge resources to provide a response.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 919–924,
Austin, Texas, November 1-5, 2016.

As shown in Fig. 1, we study three types of question-image pairs: **Non-Visual**. These questions are not questions about images at all – they do not require information from *any* image to be answered (*e.g., "What is the capital of Argentina?"*). **Visual False-Premise**. While visual, these questions do not apply to the given image. For instance, the question *"What is the girl wearing?"* makes sense only for images that contain a girl in them. **Visual True-Premise**. These questions are relevant to (*i.e.,* have a premise which is true) the image at hand.

We introduce datasets and train models to recognize both non-visual and false-premise question-image (QI) pairs in the context of VQA. First, we identify whether a question is visual or non-visual; if visual, we identify whether the question has a true-premise for the given image. For visual vs. non-visual question detection, we use a Long Short-Term Memory (LSTM) recurrent neural network (RNN) trained on part of speech (POS) tags to capture visual-specific linguistic structure. For true vs. false-premise question detection, we present one set of approaches that use the uncertainty of a VQA model, and another set that use pre-trained captioning models to generate relevant captions (or questions) for the given image and then compare them to the given question to determine relevance.

Our proposed models achieve accuracies of 92% for detecting non-visual, and 74% for detecting false-premise questions, which significantly outperform strong baselines. We also show through human studies that a VQA system that reasons about question relevance is picked *significantly more often* as being more intelligent, human-like and reasonable than a baseline VQA system which does not. Our code and datasets are publicly available on the authors' webpages.

2 Related Work

There is a large body of existing work that reasons about cross-modal similarity: how well an image matches a query tag (Liu et al., 2009) in text-based image retrieval, how well an image matches a caption (Feng and Lapata, 2013; Xu et al., 2015; Ordonez et al., 2011; Karpathy and Fei-Fei, 2015; Fang et al., 2015), and how well a video matches a description (Donahue et al., 2015; Lin et al., 2014a).

In our work, if a question is deemed irrelevant, the VQA model says so, as opposed to answering the question anyway. This is related to perception systems that do not respond to an input where the system is likely to fail. Such failure prediction systems have been explored in vision (Zhang et al., 2014; Devarakota et al., 2007) and speech (Zhao et al., 2012; Sarma and Palmer, 2004; Choularton, 2009; Voll et al., 2008). Others attempt to provide the most meaningful answer instead of suppressing the output of a model that is expected to fail for a given input. One idea is to avoid a highly specific prediction if there is a chance of being wrong, and instead make a more generic prediction that is more likely to be right (Deng et al., 2012). Malinowski and Fritz (2014) use semantic segmentations in their approach to question answering, where they reason that objects not present in the segmentations should not be part of the answer.

To the best of our knowledge, our work is the first to study the relevance of questions in VQA. Chen et al. (2012) classify users' intention of questions for community question answering services. Most related to our work is Dodge et al. (2012). They extract visual text from within Flickr photo captions to be used as supervisory signals for training image captioning systems. Our motivation is to endow VQA systems the ability to detect non-visual questions to respond in a human-like fashion. Moreover, we also detect a more fine-grained notion of question relevance via true- and false-premise.

3 Datasets

For the task of detecting visual vs. non-visual questions, we assume all questions in the VQA dataset (Antol et al., 2015) are visual, since the Amazon Mechanical Turk (AMT) workers were specifically instructed to ask questions about a displayed image while creating it. We also collected non-visual philosophical and general knowledge questions from the internet (see supplementary material). Combining the two, we have 121,512 visual questions from the validation set of VQA and 9,952[1] generic non-visual questions collected from the internet. We call this dataset Visual vs. Non-

[1]High accuracies on this task in our experiments indicate that this suffices to learn the corresponding linguistic structure.

Visual Questions (VNQ).

We also collect a dataset of true- vs. false-premise questions by showing AMT workers images paired with random questions from the VQA dataset and asking them to annotate whether they are applicable or not. We had three workers annotate each QI pair. We take the majority vote as the final ground truth label.[2] We have 10,793 QI pairs on 1,500 unique images out of which 79% are non-applicable (false-premise). We refer to this visual true- vs. false-premise questions dataset as VTFQ.

Since there is a class imbalance in both of these datasets, we report the average per-class (*i.e.,* normalized) accuracy for all approaches. All datasets are publicly available.

4 Approach

Here we present our approaches for detecting (1) visual vs. non-visual QI pairs, and (2) true- vs. false-premise QI pairs.

4.1 Visual vs. Non-Visual Detection

Recall that the task here is to detect visual questions from non-visual ones. Non-visual questions, such as *"Do dogs fly?"* or *"Who is the president of the USA?"*, often tend to have a difference in the linguistic structure from that of visual questions, such as *"Does this bird fly?"* or *"What is this man doing?"*. We compare our approach (**LSTM**) with a baseline (**RULE-BASED**):

1. **RULE-BASED**. A rule-based approach to detect non-visual questions based on the part of speech (POS)[3] tags and dependencies of the words in the question. *E.g.,* if a question has a plural noun with no determiner before it and is followed by a singular verb (*"Do dogs fly?"*), it is a non-visual question.[4]

2. **LSTM**. We train an LSTM with 100-dim hidden vectors to embed the question into a vector and predict visual vs. not. Instead of feeding question words (['what', 'is', 'the', 'man', 'doing', '?']), the input to our LSTM is embeddings of POS tags of the words (['pronoun', 'verb', 'determiner', 'noun', 'verb']). Embeddings of the POS tags are learnt end-to-end. This captures the structure of image-grounded questions, rather than visual vs. non-visual topics. The latter are less likely to generalize across domains.

[2] 78% of the time all three votes agree.

[3] We use spaCy POS tagger (Honnibal and Johnson, 2015).

[4] See supplement for examples of such hand-crafted rules.

4.2 True- vs. False-Premise Detection

Our second task is to detect whether a question Q entails a false-premise for an image I. We present two families of approaches to measure this QI 'compatibility': (i) using uncertainty in VQA models, and (ii) using pre-trained captioning models.

Using VQA Uncertainty. Here we work with the hypothesis that if a VQA model is uncertain about the answer to a QI pair, the question may be irrelevant for the given image since the uncertainty may mean it has not seen similar QI pairs in the training data. We test two approaches:

1. **ENTROPY**. We compute the entropy of the softmax output from a state-of-the art VQA model (Antol et al., 2015; Lu et al., 2015) for a given QI pair and train a three-layer multilayer perceptron (MLP) on top with 3 nodes in the hidden layer.

2. **VQA-MLP**. We feed in the softmax output to a three-layer MLP with 100 nodes in the hidden layer, and train it as a binary classifier to predict whether a question has a true- or false-premise for the given image.

Using Pre-trained Captioning Models. Here we utilize (a) an image captioning model, and (b) an image question-generation model – to measure QI compatibility. Note that both these models generate natural language capturing the semantics of an image – one in the form of statement, the other in the form of a question. Our hypothesis is that a given question is relevant to the given image if it is similar to the language generated by these models for that image. Specifically:

1. Question-Caption Similarity (**Q-C SIM**). We use NeuralTalk2 (Karpathy and Fei-Fei, 2015) pre-trained on the MSCOCO dataset (Lin et al., 2014b) (images and associated captions) to generate a caption C for the given image, and then compute a learned similarity between Q and C (details below).

2. Question-Question Similarity (**Q-Q' SIM**). We use NeuralTalk2 re-trained (from scratch) on the questions in the VQA dataset to generate a question Q' for the image. Then, we compute a learned similarity between Q and Q'.

Visual vs. Non-Visual		True- vs. False-Premise				
RULE-BASED	LSTM	ENTROPY	VQA-MLP	Q-GEN SCORE	Q-C SIM	Q-Q' SIM
75.68	**92.27**	59.66	64.19	57.41	74.48	**74.58**

Table 1: Normalized accuracy results (averaged over 40 random train/test splits) for visual vs. non-visual detection and true- vs. false-premise detection. **RULE-BASED** and **Q-GEN SCORE** were not averaged because they are deterministic.

We now describe our learned Q-C similarity function (the Q-Q' similarity is analogous). Our Q-C similarity model is a 2-channel LSTM+MLP (one channel for Q, another for C). Each channel sequentially reads word2vec embeddings of the corresponding language via an LSTM. The last hidden state vectors (40-dim) from the 2 LSTMs are concatenated and fed as inputs to the MLP, which outputs a 2-class (relevant vs. not) softmax. The entire model is learned end-to-end on the VTFQ dataset. We also experimented with other representations (*e.g.,* bag of words) for Q, Q', C, which are included in the supplement for completeness.

Finally, we also compare our proposed models above to a simpler baseline (**Q-GEN SCORE**), where we compute the probability of the input question Q under the learned question-generation model. The intuition here is that since the question generation model has been trained only on relevant questions (from the VQA dataset), it will assign a high probability to Q if it is relevant.

5 Experiments and Results

The results for both experiments are presented in Table 1. We present results averaged over 40 random train/test splits. **RULE-BASED** and **Q-GEN SCORE** were not averaged because they are deterministic.

Visual vs. Non-Visual Detection. We use a random set of 100,000 questions from the VNQ dataset for training, and the remaining 31,464 for testing. We see that **LSTM** performs 16.59% (21.92% relative) better than **RULE-BASED**.

True- vs. False-Premise Detection. We use a random set of 7,195 (67%) QI pairs from the VTFQ dataset to train and the remaining 3,597 (33%) to test. While the VQA model uncertainty based approaches (**ENTROPY**, **VQA-MLP**) perform reasonably well (with the MLP helping over raw entropy), the learned similarity approaches perform much better (10.39% gain in normalized accuracy). High uncertainty of the model may suggest that a similar QI pair was not seen during training; however, that does not seem to translate to detecting irrelevance. The language generation models (**Q-C SIM**, **Q-Q' SIM**) seem to work significantly better at modeling the semantic interaction between the question and the image. The generative approach (**Q-GEN SCORE**) is outperformed by the discriminative approaches (**VQA-MLP**, **Q-C SIM**, **Q-Q' SIM**) that are trained explicitly for the task at hand. We show qualitative examples of **Q-Q' SIM** for true- vs. false-premise detection in Fig. 2.

6 Human Qualitative Evaluation

We also perform human studies where we compare two agents: (1) **AGENT-BASELINE**– always answers every question. (2) **AGENT-OURS**– reasons about question relevance before responding. If question is classified as visual true-premise, **AGENT-OURS** answers the question using the same VQA model as **AGENT-BASELINE** (using (Lu et al., 2015)). Otherwise, it responds with a prompt indicating that the question does not seem meaningful for the image.

A total of 120 questions (18.33% relevant, 81.67% irrelevant, mimicking the distribution of the VTFQ dataset) were used. Of the relevant questions, 54% were answered correctly by the VQA model. Human subjects on AMT were shown the response of both agents and asked to pick the agent that sounded more intelligent, more reasonable, and more human-like after every observed QI pair. Each QI pair was assessed by 5 different subjects. Not all pairs were rated by the same 5 subjects. In total, 28 unique AMT workers participated in the study.

AGENT-OURS was picked 65.8% of the time as the winner, **AGENT-BASELINE** was picked only 1.6% of the time, and both considered equally (un)reasonable in the remaining cases. We also measure the percentage of times each robot gets picked

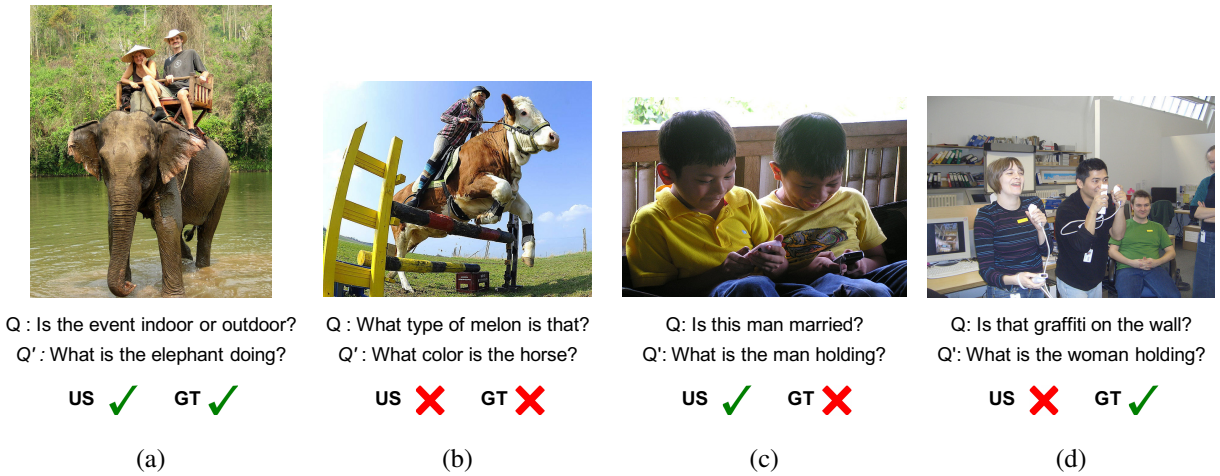

Figure 2: Qualitative examples for **Q-Q' SIM**. (a) and (b) show success cases, and (c) and (d) show failure cases. Our model predicts true-premise in (a) and (c), and false-premise in (b) and (d). In all examples we show the original question Q and the generated question Q'.

by the workers for true-premise, false-premise, and non-visual questions. These percentages are shown in Table 2.

	True-Premise	False-Premise	Non-Visual
AGENT-OURS	22.7	78.2	65.0
AGENT-BASELINE	04.7	01.4	00.0
Both	27.2	03.8	10.0
None	45.4	16.6	25.0

Table 2: Percentage of times each robot gets picked by AMT workers as being more intelligent, more reasonable, and more human-like for true-premise, false-premise, and non-visual questions.

Interestingly, humans often prefer **AGENT-OURS** over **AGENT-BASELINE** even when *both models are wrong* – **AGENT-BASELINE** answers the question incorrectly and **AGENT-OURS** incorrectly predicts that the question is irrelevant and refuses to answer a legitimate question. Users seem more tolerant to mistakes in relevance prediction than VQA.

7 Conclusion

We introduced the novel problem of identifying irrelevant (*i.e.,* non-visual or visual false-premise) questions for VQA. Our proposed models significantly outperform strong baselines on both tasks. A VQA agent that utilizes our detector and refuses to answer certain questions significantly outperforms a baseline (that answers all questions) in human studies. Such an agent is perceived as more intelligent, reasonable, and human-like.

There are several directions for future work. One possibility includes identifying the premise entailed in a question, as opposed to just stating true- or false-premise. Another is determining what external knowledge is needed to answer non-visual questions.

Our system can be further augmented to communicate to users what the assumed premise of the question is that is not satisfied by the image, *e.g.,* respond to *"What is the woman wearing?"* for an image of a cat by saying *"There is no woman."*

Acknowledgements

We thank Lucy Vanderwende for helpful suggestions and discussions. We also thank the anonymous reviewers for their helpful comments. This work was supported in part by the following: National Science Foundation CAREER awards to DB and DP, Alfred P. Sloan Fellowship, Army Research Office YIP awards to DB and DP, ICTAS Junior Faculty awards to DB and DP, Army Research Lab grant W911NF-15-2-0080 to DP and DB, Office of Naval Research grant N00014-14-1- 0679 to DB, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donation to DB.

References

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

Long Chen, Dell Zhang, and Levene Mark. 2012. Understanding User Intent in Community Question Answering. In *WWW*.

Stephen Choularton. 2009. *Early Stage Detection of Speech Recognition Errors*. Ph.D. thesis, Macquarie University.

Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. 2012. Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition. In *CVPR*.

Pandu R Devarakota, Bruno Mirbach, and Björn Ottersten. 2007. Confidence estimation in classification decision: A method for detecting unseen patterns. In *ICAPR*.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé, III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting Visual Text. In *NAACL HLT*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *CVPR*.

Yansong Feng and Mirella Lapata. 2013. Automatic Caption Generation for News Images. *PAMI*, 35(4).

Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *EMNLP*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*.

Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014a. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *CVPR*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014b. Microsoft COCO: Common objects in context. In *ECCV*.

Dong Liu, Xian-Sheng Hua, Meng Wang, and HongJiang Zhang. 2009. Boost Search Relevance for Tag-based Social Image Retrieval. In *ICME*.

Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. 2015. Deeper LSTM and normalized CNN Visual Question Answering model. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`.

Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *NIPS*.

Arup Sarma and David D Palmer. 2004. Context-based Speech Recognition Error Detection and Correction. In *NAACL HLT*.

Kimberly Voll, Stella Atkins, and Bruce Forster. 2008. Improving the Utility of Speech Recognition Through Error Detection. *Journal of Digital Imaging*, 21(4).

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*.

Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. 2014. Predicting Failures of Vision Systems. In *CVPR*.

Tongmu Zhao, Akemi Hoshino, Masayuki Suzuki, Nobuaki Minematsu, and Keikichi Hirose. 2012. Automatic Chinese Pronunciation Error Detection Using SVM Trained with Structural Features. In *Spoken Language Technology Workshop*.