Do explanations make VQA models more predictable to a human?

Arjun Chandrasekaran[*,1] **Viraj Prabhu**[*,1] **Deshraj Yadav** [*,1]
Prithvijit Chattopadhyay[*,1] **Devi Parikh**[1,2]
[1]Georgia Institute of Technology [2]Facebook AI Research
{carjun, virajp, deshraj, prithvijit3, parikh}@gatech.edu

Abstract

A rich line of research attempts to make deep neural networks more transparent by generating human-interpretable 'explanations' of their decision process, especially for interactive tasks like Visual Question Answering (VQA). In this work, we analyze if existing explanations indeed make a VQA model – its responses as well as failures – more predictable to a human. Surprisingly, we find that they do not. On the other hand, we find that human-in-the-loop approaches that treat the model as a black-box do.

1 Introduction

As technology progresses, we are increasingly collaborating with AI agents in *interactive* scenarios where humans and AI work together as a team, e.g., in AI-assisted diagnosis, autonomous driving, etc. Thus far, AI research has typically only focused on the AI in such an interaction – for it to be more accurate, be more human-like, understand our intentions, beliefs, contexts, and mental states.

In this work, we argue that for human-AI interactions to be more effective, humans must also understand the AI's beliefs, knowledge, and quirks.

Many recent works generate human-interpretable 'explanations' regarding a model's decisions. These are usually evaluated offline based on whether human judges found them to be 'good' or to improve trust in the model. However, their contribution in an interactive setting remains unclear. In this work, we evaluate the role of explanations towards making a model predictable to a human.

We consider an AI trained to perform the multi-modal task of Visual Question Answering (VQA) (Malinowski and Fritz, 2014; Antol et al., 2015), i.e., answering free-form natural language questions about images. VQA is applicable to scenarios where humans actively elicit information from visual data, and naturally lends itself to human-AI interactions. We consider two tasks that demonstrate the degree to which a human understands their AI teammate (we call Vicki) – Failure Prediction (FP) and Knowledge Prediction (KP). In FP, we ask subjects on Amazon Mechanical Turk to predict if Vicki will correctly answer a given question about an image. In KP, subjects predict Vicki's exact response.

We aid humans in forming a mental model of Vicki by (1) familiarizing them with its behavior in a 'training' phase and (2) exposing them to its internal states via various explanation modalities. We then measure their FP and KP performance.

Our key findings are that (1) humans are indeed capable of predicting successes, failures, and outputs of the VQA model better than chance, (2) explicitly training humans to familiarize themselves with the model improves their performance, and (3) existing explanation modalities do not enhance human performance.

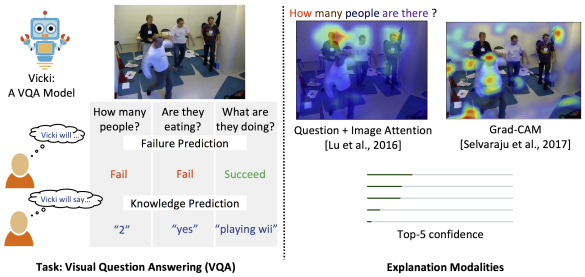


Figure 1: We evaluate the extent to which explanation modalities (right) and familiarization with a VQA model help humans predict its behavior – it's responses, successes, and failures (left).

2 Related Work

Explanations in deep neural networks. Several works generate explanations based on inter-

*Denotes equal contribution.

nal states of a decision process (Zeiler and Fergus, 2014; Goyal et al., 2016b), while others generate justifications that are consistent with model outputs (Ribeiro et al., 2016; Hendricks et al., 2016). Another popular form of providing explanations is to visualize regions in the input that contribute to a decision – either by explicitly attending to relevant input regions (Bahdanau et al., 2014; Xu et al., 2015), or exposing implicit attention for predictions (Selvaraju et al., 2017; Zhou et al., 2016).

Evaluating explanations. Several works evaluate the role of explanations in developing trust with users (Cosley et al., 2003; Ribeiro et al., 2016) or helping them achieve an end goal (Narayanan et al., 2018; Kulesza et al., 2012). Our work, however, investigates the role of machine-generated explanations in improving the *predictability* of a VQA model.

Failure prediction. While Bansal et al. (2014) and Zhang et al. (2014) predict failures of a model using simpler statistical models, we explicitly train a person to do this.

Legibility. Dragan et al. (2013) describe the intent-expressiveness of a robot as its trajectory being expressive of its goal. Analogously, we evaluate if explanations of the intermediate states of a VQA model are expressive of its output.

Humans adapting to technology. Wang et al. (2016) and Pelikan and Broth (2016) observe humans' strategies while adapting to the limited capabilities of an AI in interactive language games. In our work we explicitly measure to what extent humans can form an accurate model of an AI, and the role of familiarization and explanations.

3 Setup

Agent. We use the VQA model by Lu et al. (2016) as our AI agent (that we call Vicki). The model processes the question at multiple levels of granularity (words, phrases, entire question) and at each level, has explicit attention mechanisms on both the image and the question[1]. It is trained on the train split of the VQA-1.0 dataset (Antol et al., 2015). Given an image and a question about the image, it outputs a probability distribution over 1000 answers. Importantly, the model's image and question attention maps provide access to its 'internal states' while making a prediction.

Vicky is *quirky* at times, i.e., has biases, albeit in a predictable way. Agrawal et al. (2016) outlines several such quirks. For instance, Vicki has a limited capability to understand the image – when asked the color of a small object in the scene, say a soda can, it may simply respond with the most dominant color in the scene. Indeed, it may answer similarly even if no soda can is present, i.e. if the question is irrelevant.

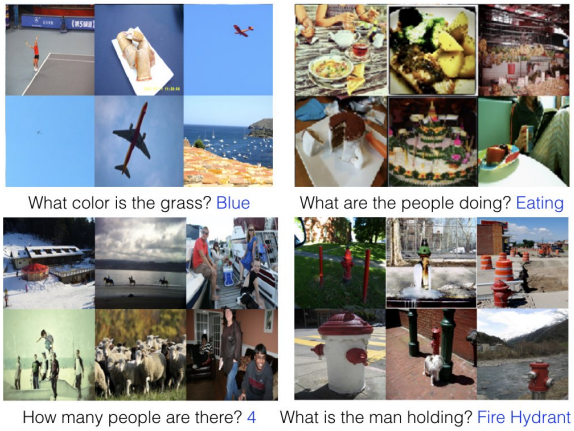


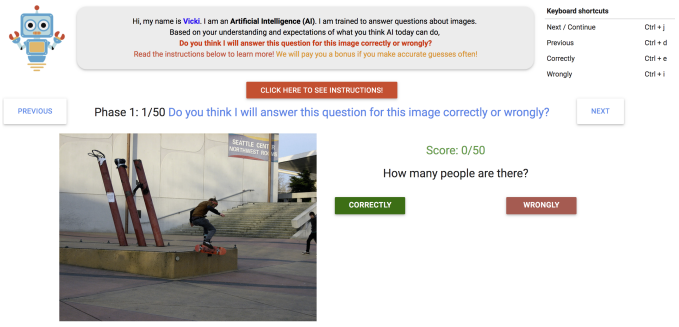
Figure 2: These montages highlight some of Vicki's quirks. For a given question, Vicki has the same response to each image in a montage. Common visual patterns (that Vicki presumably picks up on) within each montage are evident.

Further, Vicki has a limited capability to understand free-form natural language, and in many cases, answers questions based only on the first few words of the question. It is also generally poor at answering questions requiring "common sense" reasoning. Moreover, being a discriminative model, Vicki has a limited vocabulary (1k) of answers. Additionally, the VQA 1.0 dataset contains label biases; therefore, the model is very likely to answer "white' to a "what color" question (Goyal et al., 2016a).

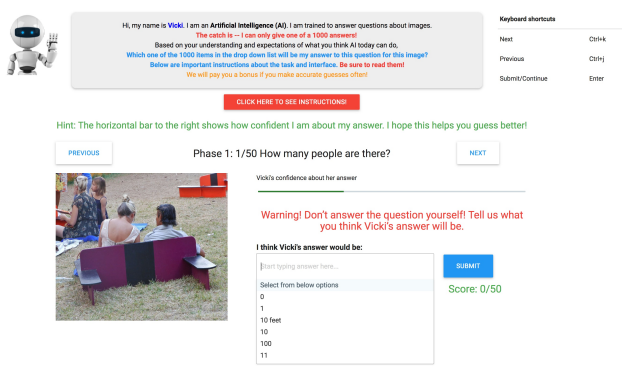
To get a sense for this, see Fig. 2 which depicts a clear pattern. In top-left, even when there is no grass, Vicki tends to latch on to one of the dominant colors in the image. For top-right, even when there are no people in the image, it seems to respond with what people could *plausibly* do in the scene if they were present. In this work, we measure to what extent lay people can pick up on these quirks by interacting with the agent, and whether existing explanation modalities help do so.

Tasks: Failure Prediction (FP). Given an image and a question about the image, we measure how well a person can predict if Vicki will successfully answer the question. A person can presumably predict the failure modes of Vicki well if they have a good sense of its strengths and weaknesses.

[1] We use question-level attention maps in our experiments.



(a) The Failure Prediction (FP) interface.



(b) The Knowledge Prediction (KP) interface.

Figure 3: (a) A person guesses if a VQA model (Vicki) will answer this question for this image correctly or wrongly. (b) A person guesses what Vicki's exact answer will be for this QI–pair.

Knowledge Prediction (KP). In this task, we aim to obtain a fine-grained measure of a person's understanding of Vicki's behavior. Given a QI–pair, a subject guesses Vicki's exact response from a set of its output labels. Snapshots of our interfaces can be seen in Fig. 3.

4 Experimental Setup

In this section we investigate ways to make Vicki's behavior more predictable to a subject. We approach this by – providing instant feedback about Vicki's actual behavior on each QI pair once the subject responds, and exposing subjects to various explanation modalities that reveal Vicki's internal states before they respond.

Data. We identify a subset of questions in the VQA-1.0 (Antol et al., 2015) validation split that occur more than 100 times. We select 7 diverse questions[2] from this subset that are representative of the different types of questions (counting, yes/no, color, scene layout, activity, etc.) in the dataset. For each of the 7 questions, we sample a set of 100 images. For FP, the 100 images are random samples from the set of images on which the question was asked in VQA-1.0 val. For the KP task, these 100 images are random images from VQA-1.0 val. Ray et al. (2016) found that randomly pairing an image with a question in the VQA-1.0 dataset results in about 79% of pairs being irrelevant. This combination of relevant and irrelevant QI-pairs allows us to test subjects' ability to develop a robust understanding of Vicki's behavior across a wide variety of inputs.

[2]What kind of animal is this? What time is it? What are the people doing? Is it raining? What room is this? How many people are there? What color is the umbrella?

Study setup. We conduct our studies on Amazon Mechanical Turk. Each task (HIT) comprises of 100 QI-pairs where for simplicity (for the subject), a single question is asked across all 100 images. The annotation task is broken down into a train and test phase of 50 QI-pairs each. Over all settings, 280 workers took part in our study (1 unique worker per HIT), resulting in 28k human responses. Subjects were paid an average of \$3 base plus \$0.44 performance bonus, per HIT.

There are some challenges involved in scaling data-collection in this setting: (1) Due to the presence of separate train and test phases, our AMT tasks tend to be unusually long (mean HIT durations across the tasks of FP and KP $= 10.11 \pm 1.09$ and 24.49 ± 1.85 min. respectively). Crucially, this also reduces the subject pool to only those willing to participate in long tasks. (2) Once a subject participates in a task, they cannot do another because their familiarity with Vicki would leak over. This constraint causes our analyses to require as many subjects as tasks. Since work division in crowdsourcing tasks follows a Pareto principle (Little, 2009), this makes data collection very slow.

In light of these challenges, we focus on a small set of questions to systematically evaluate the role of training and exposure to Vicki's internal states.

4.1 Evaluating the role of familiarization

To familiarize subjects with Vicki, we provide them with instant feedback during the train phase. Immediately after a subject responds to a QI–pair, we show them whether Vicki actually answered the question correctly or not (in FP) or what Vicki's response was (in KP), along with a running score of how well they are doing. Once training is complete, no further feedback is pro-

vided and subjects are asked to make predictions for the test phase. At the end, they are shown their score and paid a bonus proportional to the score.

Failure Prediction. In FP, always guessing that Vicki answers 'correctly' results in 58.29% accuracy, while subjects do slightly better and achieve 62.66% accuracy, even without prior familiarity with Vicki (No Instant Feedback (IF)). Further, we find that subjects that receive training via instant feedback (IF) achieve 13.09% higher mean accuracies than those who do not (see Fig 4; IF vs No IF for FP (left)).

Knowledge Prediction. In KP, answering each question with Vicki's most popular answer overall ('no') would lead to an accuracy of 13.4%. Additionally, answering each question with its most popular answer *for that question* leads to an accuracy of 31.43%. Interestingly, subjects who are unfamiliar with Vicki (No IF) achieve 21.27% accuracy – better than the most popular answer overall, but worse than the question-specific prior over its answers. The latter is understandable as subjects unfamiliar with Vicki do not know which of its 1000 possible answers the model is most likely to predict for each question.

We find that mean performance in KP with IF is 51.11%, 29.84% higher than KP without IF (see Fig 4; IF vs No IF for KP (right)). It is apparent that just from a few (50) training examples, subjects succeed in building a mental model of Vicki's behavior that generalizes to new images. Additionally, the 29.84% improvement over No IF for KP is significantly larger than that for FP (13.09%). This is understandable because a priori (No IF), KP is a much harder task as compared to FP due to the increased space of possible subject responses given a QI-pair, and the combination of relevant and irrelevant QI-pairs in the test phase.

Questions such as 'Is it raining?' have strong language priors – to these Vicki often defaults to the most popular answer ('no'), irrespective of image. On such questions, subjects perform considerably better in KP once they develop a sense for Vicki's inherent biases via instant feedback. For open-ended questions like 'What time is it?', feedback helps subjects (1) narrow down the 1000 potential options to the subset that Vicki typically answers with – in this case time periods such as 'daytime' rather than actual clock times and (2) identify correlations between visual patterns and Vicki's answer. In other cases like 'How many people are in the image?' the space of possible answers is clear a priori, but after IF subjects realize that Vicki is bad at detailed counting and bases its predictions on coarse signals of the scene layout.

4.2 Evaluating the role of explanations

In this setting, we show subjects an image, a question, and one of the explanation modalities described below. We experiment with 3 qualitatively different modalities (see Fig.1, right):

Confidence of top-5 predictions. We show subjects Vicki's confidence in its top-5 answer predictions from its vocabulary as a bar plot (of course, we do not show the actual top-5 predictions). **Attention maps.** Along with the image we show subjects the spatial attention map over the image and words of the question which indicate the regions that Vicki is looking at and listening to, respectively. **Grad-CAM.** We use the CNN visualization technique by Selvaraju et al. (2017), using the (implicit) attention maps corresponding to Vicki's most confident answer.

Automatic approaches. We also evaluate automatic approaches to detect Vicki's failure from its internal states. We find that both, a decision stump on Vicki's confidence in its top answer, and on the entropy of its softmax output, result in an FP accuracy of 60% on our test set. A Multi-layer Perceptron (MLP) trained on Vicki's output 1000-way softmax to predict success vs failure, achieves an FP accuracy of 81%. Training it on just the top-5 softmax outputs achieves an FP accuracy of 61.43%.

Training an MLP which takes as input question features (average word2vec embeddings (Mikolov et al., 2013) of words in the question) concatenated with image features (fc7 from VGG-19) to predict success vs failure (which we call ALERT following (Zhang et al., 2014)) achieves an FP accuracy of 65%. Training an MLP on identical question features as above but concatenated with Grad-CAM saliency maps leads to FP accuracy of 73.14%. [3] Note that we only report machine results to put human accuracies in perspective. We do not draw any inferences about the relative capabilities of both.

Results. Average performance of subjects in the test phases of FP and KP, for different experimental settings are summarized in Fig. 4. In the first setting, we show subjects an explanation modality

[3]These methods are trained on 66% of VQA-1.0 val. The remaining data is used for validation.

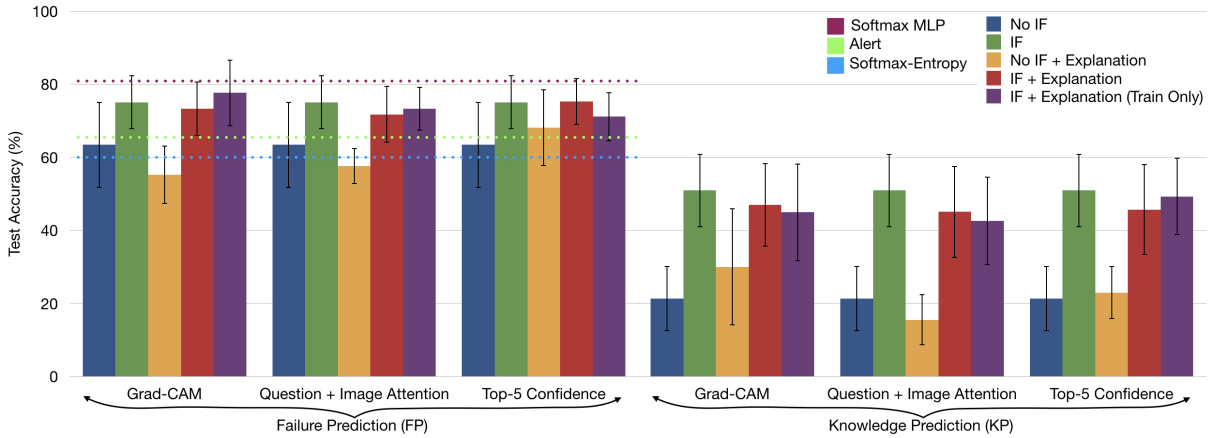


Figure 4: Average performance across subjects for Failure Prediction and Knowledge Prediction, across different settings: with or without (1) Instant feedback (IF) in the train phase, and (2) an explanation modality. Explanation modalities are shown in both train and test phases unless stated otherwise. Error bars are 95% confidence intervals from 1000 bootstrap samples. Note that the dotted lines are various machine approaches applied to FP.

with instant feedback (IF+Explanation). For reference, also see performance of subjects provided with IF and no explanation modality (IF).

We observe that on both FP and KP, subjects who received an explanation along with IF show no statistically significant difference in performance compared to those who did not. We see in Fig. 4, that both bootstrap based standard error (95% confidence intervals) overlap significantly.

Seeing that explanations in addition to IF does not outperform an IF baseline, we next measure whether explanations help a user not already familiar with Vicki via IF. That is, we evaluate if explanations help against a No IF baseline by providing an explanation only in the *test* phase, and no IF (see Fig 4; No IF + Explanation). Additionally, we also experiment with providing IF and an explanation *only* during the train phase (see Fig 4; IF + Explanation (Train Only)), to measure whether access to internal states during training can help subjects build better intuitions for model behavior without needing access to internal states at test time. In both settings however, we observe no statistically significant difference in performance over the No IF and IF baselines, respectively. [4]

5 Conclusion

As technology progresses, human-AI teams are inevitable. We argue that for these teams to be more effective, we should also be pursuing research directions to help humans understand the strengths, weaknesses, quirks, and tendencies of AI. We instantiate these ideas in the domain of Visual Question Answering (VQA), by proposing two tasks that help measure how well a human 'understands' a VQA model (we call Vicki) – Failure Prediction (FP) and Knowledge Prediction (KP). We find that lay people indeed get better at predicting Vicki's behavior using just a few 'training' examples, but surprisingly, existing popular explanation modalities do not help make its failures or responses more predictable. While previous works have typically assessed their interpretability or their role in improving human trust, our preliminary hypothesis is that these modalities may not yet help performance of human-AI teams in a goal-driven setting. Clearly, much work remains to be done in developing improved explanation modalities that can improve human-AI teams.

Future work involves closing the loop and evaluating the extent to which improved human performance at FP and KP translates to improved success of human-AI teams at accomplishing a shared goal. Co-operative human-AI games may be a natural fit for such an evaluation.

Acknowledgements. We thank Satwik Kottur for his help with data analysis, and for many fruitful discussions. We thank Aishwarya Agrawal and Akrit Mohapatra for their help with experiments. We would also like to acknowledge the workers on Amazon Mechanical Turk for their effort. This work was supported in part by NSF, AFRL, DARPA, Siemens, Google, Amazon, ONR YIP and ONR Grants N00014-16-1-2713.

[4] When piloting the tasks ourselves, we found it easy to 'overfit' to the explanations and hallucinate patterns.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Aayush Bansal, Ali Farhadi, and Devi Parikh. 2014. Towards transparent systems: Semantic characterization of failure modes. In *European Conference on Computer Vision*, pages 366–381. Springer.

Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592. ACM.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 301–308. IEEE.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016a. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*.

Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. 2016b. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10. ACM.

Greg Little. 2009. How many turkers are there.(dec 2009).

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

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

Hannah RM Pelikan and Mathias Broth. 2016. Why that nao?: How humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932. ACM.

Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in vqa: Identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*.

Sida I Wang, Percy Liang, and Christopher D Manning. 2016. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. 2014. Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3566–3573.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.