Abstract

While there have been many proposals on how to make AI algorithms more transparent, few have attempted to evaluate the impact of AI explanations on human performance on a task using AI. We propose a Twenty-Questions style collaborative image guessing game, Explanation-assisted Guess Which (ExAG) as a method of evaluating the efficacy of explanations in the context of Visual Question Answering (VQA) - the task of answering natural language questions on images. We study the effect of VQA agent explanations on the game performance as a function of explanation type and quality. We observe that “helpful” explanations are conducive to game performance (by almost 22% for “excellent” rated explanation games), and having at least one “correct” explanation is significantly helpful when VQA system answers are mostly noisy (by almost 30% compared to no explanation games). We also see that players develop a preference for explanations even when penalized and that the explanations are mostly rated as “helpful”.

1. Introduction

Deep networks are often deemed as un-interpretable and black-box models. This raises issues of trust and reliability for an end user. In the context of Visual Question Answering (VQA) [2], the task of answering questions asked to images, various methods for shedding light on the inner workings of these networks have been proposed — interpretable attention models [17, 13, 28], visual-semantic explanations [17, 19, 28, 22] and human-readable text-based justifying explanations [19] to mention a few. However, the empirical evidence that such a-priori devised explanations can actually be meaningful and useful for a human-machine collaborative task is lacking.

We propose a Twenty-Questions [15] style human-machine collaborative game, Explanation-Assisted Guess-Which (ExAG) using VQA [2] as the underlying task for evaluating the efficacy of explanations. An explanation, in this context, is defined as information that sheds light on the reasoning of the VQA agent for predicting a certain answer. In this game, the human and a VQA machine must collaborate to retrieve a secret image. The machine has access to the secret image while the human doesn’t. The job of the machine is to help the human identify the image by answering questions. Since the VQA machine is noisy in its answer predictions, our hypothesis is that humans will succeed more often in finding the correct image when the machine also explains its reasoning, compared to just answering. Practical applications of such a game are image retrieval using free-form queries. Finding the correct image in this game means a VQA system could successfully convey the right image in the player’s mind by simply answer-
2. Related Work

**Explainable AI** Early work on explainable models involves template-based systems that spanned from medical systems [23] to educational settings [14, 27]. Recent interest in explaining the inferences of a deep networks for computer vision applications includes introspective explanations that show the intermediate features of importance in making a decision [17, 19, 23, 10, 9, 22, 30], as well as post-hoc rationalization techniques such as justifying textual explanations [19] and generating visual explanations [12]. We focus on using attention-based visualization of important image regions [17, 23, 13, 25], object/scene detection [25, 11] and related question answers [21] for our game. [6] shows that humans and machines look differently at images when answering questions. Hence, it is not clear whether such explanations are indeed helpful to humans. In this paper, we try to quantify how much these explanation modes help in human-machine collaboration performance.

**Visual Question Answering** We use Visual Question Answering (VQA) [2] as the core underlying technology for our human-AI collaborative game. VQA was introduced by [2] as an AI complete task for image and text understanding. Most of the effective approaches to VQA consist of works with attention on image features [29, 17, 26, 28, 13] guided by the question in order to answer it. We implement a custom model that attends to both objects and free-form spatial regions in the image to answer the question similar to [29].

20 Questions Game Our choice of the image-guessing game is the visual version of the popular 20-questions game, which is more formally, a specific version of the classic Lewis Signaling Game [15]. There have also been efforts at training AI agents to play such an image-guessing game with humans/AI’s [8] using reinforcement learning - [7, 5] used this game to evaluate visual conversational agent performance. However, to our knowledge, we are the first ones to use such a game to evaluate the effect of VQA explanations on human-machine team performance.

**Mental Model of an AI System** Along the lines of quantifying explanation efficacy, [4, 3] quantifies whether attention-based explanation improves human prediction of VQA system answer correctness and accuracy. While they [4, 3] show no significant trend of being able to predict model outcome using attention-based explanations, we show that only good combined attention + related question-answer explanations are helpful in a game setting where multiple rounds of question-answering are involved. We also show that erroneous explanations are severely hurtful to game performance. [19] show that showing both attention-based and textual explanations is helpful for predicting model performance. We use related question-answers as a form of textual explanation and also see similar trends for such a collaborative question-answering task.

3. Game Outline

In our game setting, there are two agents: a state-of-the-art custom VQA deep learning model trained to answer questions about images (the “VQA machine”), and a human volunteer (the “player”) that has to guess a secret image picked by the machine while losing as few points as possible. A secret image is picked from a pool of 1500 images at random. We select \( N - 1 \) more images from the pool using a difficulty measure based on the VGG16 [23] FC7 distance that is adjusted so that the final \( N \) image set is challenging enough to require multiple questions and answers. The player starts with \( P_o \) points and is allowed to ask free-form questions to the VQA machine in order to guess the secret image and each question costs one point. The final score is \( P = P_o - Q \) if the correct image is guessed, where \( Q \) is the number of questions asked. Wrong image guess gets a...
score of $P = 0$. A success is defined as the player correctly selecting the secret image while keeping $P > 0$. Players are encouraged to keep $P$ as high as possible.

### 3.1. The VQA Model

![ResNet-based VQA model (Model A)](image)

**Figure 2.** ResNet-based VQA model (Model A). This network attends to ResNet 161 features. This model is based on [13].

We used two SOTA VQA models, one with a ResNet-based image encoder, referred to as Model A (see Figure 2) and the other uses both ResNet [25] and Mask-RCNN-based [11] image encoders, referred to as Model B (see Figure 3). Both models use an attention mechanism to select visual features generated by an image encoder and an answer classifier that predicts an answer from 3000 candidates. Our VQA models are very similar to [13, 26, 29]. More details about architecture and training are included in supplementary.

### 3.2. Modes of Explanations

We define explanations as information that provides insight into why the VQA predicted a certain answer. The insight can be visualizations of the evidence used to infer the answer, such as weights applied to visual encoder features (i.e., attention). It can also be rationalizations, such as stating the semantic beliefs about a fact that led to the answer. Below, we outline three modes of explanations and illustrate how they can be used in our ExAG game to aid players in succeeding in the game.

**Attention:**

We use spatial and object attention masks computed by the attention modules to highlight spatial locations/objects in the image that are weighted more heavily and thus contribute more as evidence in the inference process of answer prediction. A player can check if the attention masks correspond to the relevant part of the image given the question to determine if the machine generated answer is trustworthy.

Model A (Figure 2) employs spatial attention only on ResNet161 [25] features. Model B (Figure 3) employs two types of attentions - spatial attention that gates visual features in the image space and object attention derived from region proposals and centering at objects detected in the image. Accordingly, we have two types of attention-based explanations: 1) heatmaps based on spatial attention overlaid on the input images and 2) object segmentation masks weighted by object attention and overlaid on input images.

**Object and Scene Predictions:**

We display a list of the most relevant object and scene predictions observed in the image. The relevance of an object/scene $O$ is measured via an importance score $S(O)$ by $S(O) = \frac{\text{word2vec}(O, A)}{p(O|I)}$ where $I$ and $A$ denote the image and machine generated answer, respectively, and $\text{word2vec}(O, A)$ denotes the Word2Vec [18] similarity of the object and answer words. When using object-based attention, this importance score calculation is skipped and attention weights are used instead. The player needs to judge if the listed relevant objects make sense with the image and machine generated answer. This can help judging if the answer is correct or not.

**Answers to related questions (RelQAS or Related QAS):**

We list five questions from the VQA2.0 Validation Dataset that are most relevant to the asked question and answer them using the VQA. The relevance of a question is measured via a semantic similarity based on the averaged word2vec distance [18] over all the words in the question and answer for the image. The five answers predicted to these closely related questions are used as another mode of explanation [21]. Similar to explanation based on relevant object/scene, the player should judge the trustworthiness of machine decision based on the correctness as well as coherence of these answers. An example is shown in Figure 4.

### 3.3. Game Settings

**Setting A:**

In this setting, we show $N = 20$ images and the player has the option of receiving explanations for the answer or
Figure 4. Related question answer explanations shown in response to the question “what is the image?” for the given image. The answer predicted by the VQA for the image is “surfer” not. Each question asked costs one point and explanation, if requested, costs an additional two points. All explanation modes are shown once the player chooses to receive explanations. This includes heatmaps of spatial attention for all the images, a list of relevant objects/scenes for the secret image, and related question-answer pairs for the secret image. Note that the detected objects/scenes and the answers to the related questions are from actual model predictions and hence, noisy. Thus, they have to be correct/coherent with the original answer and not be misleading (listing objects of the similar distractor images) in order to be helpful.

Setting B:

In Setting B, we show RelQAS and attention for all the images. This ensures that extra information is given for all images and only the coherence of the extra information can be used to aid in the game. In order to reduce the cognitive load on the players, we reduce image set size from $N = 20$ to $N = 5$ and make the images more similar to each other to maintain the same level of difficulty. Since we use object-based attention, we drop the “object/scene” explanations in this setting. We also drop the penalty for using explanations since we randomly assign each explanation type to a group of AMT workers. As the players play the game, they are also asked to rate how helpful the explanations are after each round of question answering. Note that at the time of rating, players do not yet know the outcome of the game or the secret image. So, their rating reflects how helpful the explanations are in narrowing down the secret image and identifying the proper question to ask next.

Figure 5. Attention visualization Version 1. The question is “What is in the image?”. The answer predicted is “banana”.

In order to examine the effect of different attention visualization, we have tested two variations:

• **Version 1**: Transparent spatial attention, transparent object attention with semantic labels (see Figure 5). The object labels come from a pool of 88 MSCOCO [16] classes detected by the Mask-RCNN [11].

• **Version 2**: Spatial attention visualized via the jet color map, transparent object attention without object labels (see Figure 6). Initial studies on AMT show this version to be more helpful in terms of game performance and ease of understanding (object labels are often confusing since the 88 class labels are different from VQA answer vocabulary, e.g., “dress” vs “shirt”). Hence, all future analyses are done with attention version 2.

4. Results

|                | # Plays | # Wins | Win Rate  |
|----------------|---------|--------|-----------|
| Total Plays    | 206     | 89     | 43.20%    |
| Expl. Used Atleast once | 157   | 75     | 47.77%    |
| No explanations used | 49    | 14     | 28.57%    |

4.1. Game Setting A

As a pilot study, the ExAG game setting A was played in a competitive setting (with cash rewards for group that won the most) by about 60 people grouped into 6 teams. The players were free to choose explanations or forgo them.

Of 206 total games played, the average win rate was 43%. We divided the game plays into games where explanations were never used (49) and those where explanations were used at least once (157; see Table 1). The win rate
Table 2. Each explanation type was randomly assigned to a group of AMT workers. “With Expl” are plays with explanations, “No Expl” are plays by the same worker group with no explanations. “Group Baseline” are the first 5 plays of that worker group without explanations to set an initial baseline for that group. “Overall” is the average across all three groups. “Overall Improv” refers to the improvement compared to the overall baseline.

| Expl Type | With Expl | No Expl | Group Baseline | Overall | Overall Improv |
|-----------|-----------|---------|----------------|---------|----------------|
|           | Score | Win Rate | Score | Win Rate | Score | Win Rate | Score | Win Rate |
| Attention | 6.23  | 66.67   | 6    | 64.92   | 5.66  | 62.1    | 0.42  | 2.82    |
| Rel QAS   | 6.8   | 71.48   | 6.03 | 64.54   | 6.02  | 65.45   | 0.99  | 7.63    |
| Both      | 6.44  | 69.03   | 5.83 | 63.25   | 5.68  | 63.75   | 0.63  | 5.18    |
| Overall   | 6.52  | 69.29   | 5.97 | 64.3    | 5.81  | 63.85   | 0.71  | 5.44    |

Figure 8. Histogram of ratings of how “helpful” explanations seemed while playing the game. These ratings were collected after each question was asked. Workers didn’t know the GT image or the outcome of the game while rating them.

with explanations was 47% (N=75) and 28% when explanations were never used (N=14). The z-test for proportions indicates that this is a statistically significant difference at 95% confidence level (p=0.019).

Moreover, as the players proceeded to play the ExAG game, they tended to opt for using explanations even though that resulted in additional 2 point penalty. Figure [7] shows this spontaneous adoption of explanation with increasing number of plays. A z-test comparing the proportion of games using the explanations during the first half of plays (61.2%, N=103) vs. during the second half (91.3%, N=103) indicates highly statistically significant increase in explanation use ($p < 10^{-5}$).

4.2. Game Setting B

Game Setting B shows related question-answer pairs and segmented objects (in attention) for all images and hence, players have to solely rely on the quality of explanations to the answer to aid their game. Hence, extra information about GT image cannot be misused. All analyses henceforth are performed on Game Setting B.

The games were played by 69 individual AMT workers. The instructions were to try to win the game by guessing the secret image correctly and were warned that an obvious lack of effort would lead to rejection. For AMT worker selection, worker qualification threshold was set to above 98% (number of HITs > 1000) and location set to US only for IRB regulations. We will anonymize and release all game play data publicly. The workers played 1469 games in total covering the three explanation modes - attention, related question-answer pairs and both, and also without any explanations.

Each individual worker always saw only one type of explanations while playing. Each worker was instructed to play at least 4 sessions with each session consisting of five games. Sessions with and without explanation alternate. For instance, the first session does not provide explanations, followed by a session with assigned choice of explanations. We also have a group that sees no explanations ever. The first block is used as the baseline no-explanation performance for the assigned explanation worker group. Workers who played with one explanation never play any other explanation type. Moreover, once each worker finished their HIT, they were not allowed to play another HIT so as to not mess up the first baseline block performance. We also kept track of worker ID to ensure game play progress was saved for each worker, and didn’t reset if they returned to the interface at a later time.
Overall impact of explanations

The overall impact of explanations on the performance of the ExAG game is summarized in Table 2. Both the average game score and the game win rate are modestly improved by explanations. Consistently with prior reports [4, 5], we don’t see a tangible effect of attention-based explanations when presented in isolation. When compared to the baseline no-explanation performance for the group (Baseline Group), related question-answers (relQAS) and showing combined (Attention + RelQAS) explanations improves game performance the most.

We observe that the helpfulness of explanations is not necessarily how correct they are, but a more elusive function of how telling they are to VQA success/failure. Hence, we collect two separate ratings for the explanations. While playing the game, workers are asked to self rate their “perceived helpfulness” of the explanation for zeroing down on an image after each question asked (Figure 8). Note that the workers do not know the GT image or the outcome of the game while rating. We also separately collect the correctness of the explanation from 3 independent AMT workers by showing them the explanation for the asked question and the GT image (Figure 9). We use these ratings to analyze how “correctness” and “perceived helpfulness” affect game performance.

Impact of explanations as a function of self-reported helpfulness ratings

We collect self-reported helpfulness ratings of explanations while workers are playing the game (Figure 8), while the play outcome is not yet known in order to avoid circularity in the ratings. We analyze game performance as a function of these ratings to see if explanations perceived as “helpful” do help game performance. Since workers didn’t know the correct image or the game outcome while rating, their decision was not affected by the (future) play outcome (win/lose) and likely reflected the helpfulness of explanations in narrowing down their image choices. The
Figure 11. Game win percentage as a function of independently rated explanation accuracy. Baseline Group Performance is the performance of the first 5 runs without explanations for the same group of workers to set an initial baseline for that group. Note that the baseline performance here may differ from Figure 10 because the correctness ratings were collected for a subset of the games and baseline was adjusted to reflect the baseline of only that subset of workers.

workers were asked to rate the explanations according to the following options: “Helping a lot” (Excellent), “Mostly helping” (Mostly Good), “Somewhat Helpful” (Somewhat), “Not helping much” (Not Much) and “Completely Confusing” (Misleading). For all plots, the ratings were averaged for a game and were binned into the original choices.

We see that overall (Figure 10a), explanations perceived as Excellent significantly increase game performance. When explanations are less helpful, performance is similar to playing without explanations supporting the notion that workers seem to ignore the explanations, or somewhat degraded as the workers were confused by them.

Next we break down explanation helpfulness-dependent performance by explanation type. Panel b of Figure 10 shows that combining attention and related QAs improves performance significantly for explanations rated as Excellent, but hurts performance slightly when rated below “somewhat”. Panel c of Figure 10 indicates that attention-based explanations don’t help much on their own, however, when they seem very helpful for making a choice, they do help game performance slightly. Excellent Related QAs were the most helpful for game performance when presented in isolation as in Figure 10d. We reason that this is due to the consistency of related question answers being a better indicator of VQA answer accuracy. We calculate the correlation coefficient of the explanation correctness to answer correctness and observe that it is 0.37 for related question-answers as compared to 0.33 for attention. This combined with the ease of parsing textual attention than heatmaps make related question answers more effective.

We also check if the winning games had easier to distinguish images (since there is some randomness in the image set difficulty due to availability of certain types of images in the 1500 pool of images we select from) than non-winning games. We analyze the average VGG16 FC7 distance of images in the pool from GT image for winning and losing games for games with explanations and without explanations. We observe that for games with explanations, winning game image sets were just as difficult (or slightly more difficult) than non-winning games. For games without explanations, winning games have much lesser difficult image sets than non-winning games.
Impact of explanations as a function of independent explanation accuracy ratings

Since it is not obvious what makes an explanation helpful, we ran a separate AMT task to collect correctness ratings of the explanations given the image, the question that was asked and the answer given by the VQA model (Figure 9). The \{image, question, answer\} triplets were taken from the set of those used prior in the ExAG games. The workers were asked to rate the explanations according to the following options: “Exactly on-point” (Excellent), “Mostly on-point” (Mostly on-point), “Indifferent” (Indifferent), “Somewhat off” (Somewhat off) and “Completely Wrong” (Wrong). For all plots, the ratings were averaged for a game and were binned into the original choices.

Figure 11 shows game performance as a function of explanation accuracy ratings. We see that overall (panel 11a), combined (11b) and attention-based (11c) correct explanations are not sufficient for improving game performance (and hence being helpful), however, incorrect explanations can severely hurt performance. Related QAS answers, as long as they are mostly on-point, help game performance substantially (see Figure 11b).

Impact of explanations as a function of VQA answer accuracy

We also collect the answer accuracy ratings for the ExAG games using an independent AMT study. The workers were required to rate the answer for the question asked to the GT image as either “correct”, “somewhat correct” or “completely wrong”. We analyzed the game performance as a function of average VQA answer correctness in a game for game plays with explanations rated as mostly “correct” (average rating above indifferent) and without explanations.

Interestingly, we see that having at least one correct explanation significantly helps in games where the VQA answers are mostly incorrect as shown in Figure 12. This effect was observed only when the attention and related QA explanations were used together, and was not observed for attention-alone and related QA-alone explanations. We also didn’t see any significant improvement in performance when explanations were mostly “incorrect” (average rating below indifferent) for mostly incorrect answer games. Moreover, we observed a significant lack of single question incorrect answer games when played with explanations as compared to no explanation games. This seems to suggest that explanations help workers realize the answer may be wrong and encouraged them to ask more questions. However, it was the correct explanation that helped them fish out the correct answer among the mostly incorrect answers in the game. As expected, Without explanations (red line of Figure 12), game performance degraded as VQA answers got less accurate, suggesting that a player had no way of telling if an answer is wrong or correct without having the benefit of inspecting both types of explanations.

5. Conclusion

We propose the ExAG game as an evaluation framework for explanations and show that game performance correlates to the explanation helpfulness and correctness. Using ExaG, we analyze the effectiveness of various explanations. Our experiments provide empirical evidence that effective (rated very helpful and above) explanations help improve performance on a human-machine collaborative image guessing task whereas ineffective explanations hurt performance (Figure 8). Moreover, since the self-rated explanation helpfulness is not influenced by the (subsequent) outcome of the game, users can use their insight into explanation helpfulness to decide when to trust them and include them in decision making process vs. when not.

This is also supported by our observation that explanations help significantly when machine predictions are noisy (Figure 12). Without explanations, users blindly trust incorrect machine predictions, which hurts game performance. However, with explanations, users ask follow-up questions and are able to fish out the correct answer based on at least one good explanation.

Interestingly, we also observe that when image sets are difficult, players tend to win more when playing with explanations rather than without.

We believe our ExaG framework can serve towards effectively evaluating explanations. This will help designing more accurate and helpful explanations that improve human-AI collaboration outcomes in terms of performance, trust and satisfaction.
Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under the Explainable AI (XAI) program. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Wagner, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 2425–2433, 2015.

[3] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh. Do explanations make vqa models more predictable to a human? arXiv preprint arXiv:1810.12366, 2018.

[4] A. Chandrasekaran, D. Yadav, P. Chattopadhyay, V. Prabhu, and D. Parikh. It takes two to tango: Towards theory of ai’s mind. arXiv preprint arXiv:1704.00717, 2017.

[5] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh. Evaluating visual conversational agents via cooperative human-ai games. In Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP), 2017.

[6] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding, 163:90–100, 2017.

[7] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. arXiv preprint arXiv:1703.06585, 2017.

[8] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. CoRR, abs/1611.08481, 2016.

[9] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the International Conference on Computer Vision (ICCV), 2017.

[10] R. Fong and A. Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2980–2988. IEEE, 2017.

[12] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In European Conference on Computer Vision, pages 3–19. Springer, 2016.

[13] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162, 2017.

[14] H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, and D. Gomboc. Explainable artificial intelligence for training and tutoring. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES, 2005.

[15] D. Lewis. Convention: A philosophical study. John Wiley & Sons, 2008.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[17] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In Advances In Neural Information Processing Systems, pages 289–297, 2016.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[19] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. arXiv preprint arXiv:1802.08129, 2018.

[20] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[21] A. Ray, G. T. Buracha, K. Sikka, A. Roy, A. Ziskind, Y. Yao, and A. Divakaran. Make up your mind: Towards consistent answer predictions in vqa models.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. See https://arxiv.org/abs/1610.02391 v3, 7(8), 2016.

[23] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. Rule-based expert systems, pages 233–262, 1984.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, volume 4, page 12, 2017.

[26] D. Teney, P. Anderson, X. He, and A. v. d. Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. arXiv preprint arXiv:1708.02711, 2017.

[27] M. Van Lent, W. Fisher, and M. Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the National Conference on Artificial Intelligence, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.
Supplementary

VQA Architecture Details

We use Tensorflow [1] for all our implementations. The network we use for Game Setting 1 is shown in Figure 2. Details are as follows:

- **Input Image** - size 448×448. We center crop all images to the mentioned size during training and reshape during evaluation. We encode the image using a ResNet 161 [25] network.

- **Question Input** - Each question word is encoded using the Glove [20] 300 dimensional embeddings before feeding into an LSTM word by word. We take the final 512 dimensional LSTM features as the question feature. Embeddings are fine-tuned.

- **Attention** - For Resnet feature attention, we tile the question features (512) into $14 \times 14 \times 512$ and concatenate with $14 \times 14 \times 2048$ image features. Attention predicts a set of weights in the shape of $14 \times 14$ using a 2-D convolutional layer.

- **Answer classifier** - We concatenate weighted flattened ResNet features and question features and pass it through a fully connected layer to get 3000 answer logits. We compute a softmax to get probabilities.

For Game Setting 2, we use Network 2 as shown in Figure 3. Details are as follows:

- **Input Image** - For ResNet161 [25] image features, we center crop images to 448×448 similar to Network 1 to get $14 \times 14 \times 2048$ dimensional features. We also use the Region Proposal Network from Mask RCNN [11] to generate 100 object proposals per image. The input image size to Mask RCNN is $1024 \times 1024$ and the images are re-sized without cropping. We pool 1024 dimensional features from each of the 100 proposal boxes.

- **Question Input** - Each question word is encoded using the Glove [20] 300 dimensional embeddings before feeding into an LSTM similar to network 1. The embeddings are fine-tuned.

- **Attention** - We have two attention modules - one for attending to the ResNet features (same as Network 1) and one for attending to the 100 object proposals in the image (object attention). Resnet feature attention is same as setting 1. For object proposal attention, we concatenate the question features (512) to each 1024 image feature for the 100 proposals. Attention predicts a weight of shape $100 \times 1024$ using a 1-D convolutional layer.

- **Answer classifier** - We concatenate weighted flattened ResNet features, averaged weighted object features and question features and pass it through a fully connected layer to get 3000 answer logits. We compute a softmax to get probabilities.

Qualitative Results

Figure 13 and Figure 14 show qualitative examples of plays without and with explanations and how explanations may help in choosing the correct image more often when VQA answers are noisy.
Figure 13. This image shows two game plays without explanations (each row is a game-play example). When the VQA is fairly accurate, a user is easily able to pick out the correct image as shown in row 1. However, in the second row, the VQA answer leads to a reasonably wrong selection by the user since the fifth image is the image with the most prominent umbrella.

Figure 14. This image shows a game with both explanations. Even though the most prominent toilet is in the fourth image, note how the explanations make it clear that the fifth image could also be the GT image since it suggests that the machine saw a toilet there as well and that it probably mis-detected the bathtub or sink for a toilet. This hints the user to ask follow-up questions like “is there a sink?” to finally select the correct image.