A Study on Multimodal and Interactive Explanations for Visual Question Answering

Kamran Alipour,[1]* Jurgen P. Schulze,[1] Yi Yao,[2] Avi Ziskind,[2] and Giedrius Burachas[2]

[1]UC San Diego, La Jolla, CA
[2]SRI International, Princeton, NJ

Abstract

Explainability and interpretability of AI models is an essential factor affecting the safety of AI. While various explainable AI (XAI) approaches aim at mitigating the lack of transparency in deep networks, the evidence of the effectiveness of these approaches in improving usability, trust, and understanding of AI systems are still missing. We evaluate multimodal explanations in the setting of a Visual Question Answering (VQA) task, by asking users to predict the response accuracy of a VQA agent with and without explanations. We use between-subjects and within-subjects experiments to probe explanation effectiveness in terms of improving user prediction accuracy, confidence, and reliance, among other factors. The results indicate that the explanations help improve human prediction accuracy, especially in trials when the VQA system's answer is inaccurate. Furthermore, we introduce active attention, a novel method for evaluating causal attentional effects through intervention by editing attention maps. User explanation ratings are strongly correlated with human prediction accuracy and suggest the efficacy of these explanations in human-machine AI collaboration tasks.

1   Introduction

With recent developments in deep learning models and access to ever increasing data in all fields, we have witnessed a growing interest in using neural networks in a variety of applications over the past several years. Many complex tasks which required manual human effort are now assigned to these AI systems. To utilize an AI system effectively, users need a basic understanding of the system, i.e., they need to build a mental model of the system's operation for anticipating success and failure modes, and to develop a certain level of trust in that system. However, deep learning models are notoriously opaque and difficult to interpret and often have unexpected failure modes, making it hard to build trust. AI systems which users do not understand and trust are impractical for most applications, especially where vital decisions are made based on AI results. Previous efforts to address this issue and explain the inner workings of deep learning models include visualizing intermediate features of importance (Zeiler and Fergus 2014; Zhou et al. 2014;

---
*Email: kalipour@eng.ucsd.edu

Selvaraju et al. 2017) and providing textual justifications (Huk Park et al. 2018), but these studies did not evaluate whether these explanations aided human users in better understanding the system inferences or if they helped build trust. Prior work has quantified the effectiveness of their explanations by collecting user ratings (Lu et al. 2016a; Chandrasekaran et al. 2017) or checking their alignment with human attention (Das et al. 2017), but found no substantial benefit for the explanation types used in that study.

To promote understanding of and trust in the system, we propose an approach that provides transparency about the intermediate stages of the model operation, such as attentional masks and detected/attended objects in the scene. Also, we generate textual explanations that are aimed to explain *why* a particular answer was generated. Our explanations fall under the category of *local explanations* as they are intended to address inference on a specific run of the VQA system and are valid for that run. We offer extensive evaluations of these explanations in the setting of a VQA system. These evaluations are made by human subjects while performing a correctness prediction task. After seeing an image, a question, and some explanations, subjects are asked to predict whether the explainable VQA (XVQA) system will be accurate or not. We collect both the data on subject prediction performance and their explanation ratings during and after each prediction run.

We also introduce active attention - an interactive approach to explaining answers from a VQA system. We provide an interactive framework to deploy this new explanation mode. The interface is used to conduct a user study on the *effectiveness* and *helpfulness* of explanations in terms of improving user's performance in user-machine tasks and also their mental model of the system. The efficacy of explanations is measured using several metrics described below. We show that explanations improve VQA correctness prediction performance on runs with incorrect answers, thus indicating that explanations are very effective in anticipating VQA failure. Explanations rated as more helpful are more likely to help predict VQA outcome correctly. Interestingly, the user confidence in their prediction exhibits substantial correlation with the VQA system confidence (top answer probability). This finding further supports the notion that the

subjects develop a mental model of the XQA system that helps them judge when to trust the system and when not.

2   Related Work

Visual Question Answering. In the VQA task, the system provides a question and an image, and the task is to answer the question using the image correctly. The multimodal aspect of the problem, combining both natural language and visual features makes this a challenging task. The VQA problem was originally introduced in (Antol et al. 2015) and since then, multiple variations have been proposed and tested. A common approach is to use attentional masks that highlight specific regions of the image, conditioned on the question (Kazemi and Elqursh 2017; Lu et al. 2016b; Teney et al. 2017; Xu and Saenko 2015; Jiang et al. 2018b; Fukui et al. 2016; Xu and Saenko 2016; Teney et al. 2018).

Explainable AI. The effort to produce automated reasoning and explanations dates back to very early work in the AI field with direct applications in medicine (Shortliffe and Buchanan 1984), education (Lane et al. 2005; Van Lent, Fisher, and Mancuso 2004), and robotics (Lomas et al. 2012). For vision-based AI applications, several explanation systems draw the focus on discovering visual features important in the decision-making process (Zeiler and Fergus 2014; Hendricks et al. 2016; Jiang et al. 2017; Selvaraju et al. 2017; Jiang et al. 2018a). For visual question answering tasks, explanations usually involve image or language attention (Lu et al. 2016a; Kazemi and Elqursh 2017). Besides saliency/attention maps, other work has studied different explanation modes including layered attentions (Yang et al. 2016), bounding boxes around important regions (Anne Hendricks et al. 2018) or textual justifications (Shortliffe and Buchanan 1984; Huk Park et al. 2018).

In this paper, we propose a multi-modal explanation system which includes justifications for system behavior in visual, textual, and semantic formats. Unlike previous work that suggest explanations mostly relied on information produced by the AI machine, our approach benefits from combining AI-generated explanations and human-annotations for better interpretability.

Human studies. As an attempt to assess the role of an explanation system in building a better mental model of AI systems for their human users, several previous efforts focused on quantifying the efficacy of explanations through user studies. Some of these studies were developed around measuring trust with users (Cosley et al. 2003; Ribeiro, Singh, and Guestrin 2016), or the role of explanations to achieve a goal (Kulesza et al. 2012; Narayanan et al. 2018; Ray et al. 2019). Other works measured the quality of explanations based on improving the predictability of a VQA model (Chandrasekaran et al. 2018).

Despite their great insights into the efficacy of various explanation modes, previous studies do not interactively involve the human subjects in producing these explanations. In our study, we design an interactive environment for users to evaluate our multi-modal explanation system in helping users predict the correctness of a VQA model. Moreover,

The users also take part in generating explanations and receive online feedback from the AI machine.

3   The VQA Model

VQA deep learning models are trained to take an image and a question about its content and produce the answer to the question. The core model extracts features from natural language questions as well as images, combines them, and generates a natural language answer. Among various methods to train VQA systems to accomplish this task, the attention-based approach is specifically of our interest.

We use a 2017 SOTA VQA model with a ResNet (Szegedy et al. 2017) image encoder (figure 1) as our VQA agent. The model is trained on VQA2 dataset and uses an attention mechanism to select visual features generated by an image encoder and an answer classifier that predicts an answer from 3000 candidates. Moreover, we replaced Resnet with a Mask-RCNN (He et al. 2017) encoder to produce object attention explanations (similar to the approach used by (Ray et al. 2019)).

As illustrated in figure 1, our VQA model takes as input a 224×224 RGB image and question with at most 15 words. Using a ResNet, the model encodes the image to reach a $14 \times 14 \times 2048$ feature representation. The model encodes the input question to a feature vector of 512 dimensions using an LSTM model based on the GloVe (Pennington, Socher, and Manning 2014) embedding of the words. The attention layer takes in the question and image feature representations and outputs a set of weights to attend on the image features. The weighted image features, concatenated with the question representation, is used to predict the final answer from a set of 3000 answer choices.
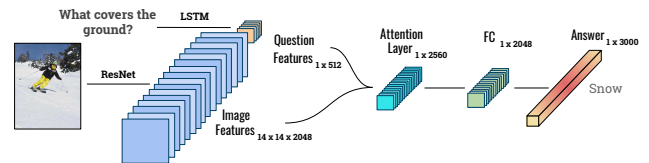

Figure 1: 2017 SOTA VQA Architecture.

4   Explanation Modes

Our XVQA system aims at explaining the VQA agent's behavior by combining the attention features generated in the VQA model with meaningful annotations from the input data. These annotations include labels, descriptions, and bounding boxes of entities in the scene and their connections with each other.

Our XVQA model either visualizes information from the inner layers of the VQA model or incorporates that information with annotations to explain model's inner work. The explanations are provided in different combinations to the subgroups of study participants to assess their effectiveness for accurate prediction.

4.1 Spatial attention

As introduced by previous work, the primary purpose of spatial attention is to show the parts of the image the model focuses on while preparing the answer. Attentions maps here are question-guided and more weighted in the areas of the image that make a higher contribution in the response generated by the model. The model computes the attentions based on image features in ResNet (Szegedy et al. 2017) layers and the question input. The final values in the attention map is a nonlinear function of image and question feature channels (figure 2)
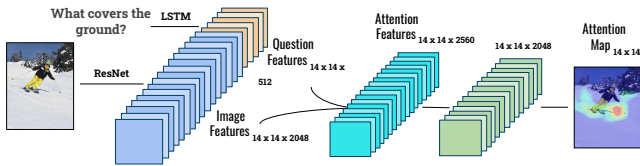

Figure 2: Attention map generated based on the input features in XVQA model.

Users try to develop an understanding of the way the model analyzes an image based on the question by looking at these attentions maps (example provided in figure 3).

Figure 3: Spatial attention explanation generated for the question: "What are the girls skating inside of?".

4.2 Active attention

Our model provides this explanation mode for the users within a feedback loop. Users can utilize this feature to *alter* a model's attention map to *steer* the model's attention and the way the answer is generated. In this feedback loop, users first see the model's answer based on the original attention map, and then they modify the attention to create a different response.

The active attention trial has a two-step task to complete. The first step is very similar to spatial attention trials where users make their prediction based on the attention map generated by the VQA model. The subject then observes the prediction results and realizes whether the system is accurate or not. At the second step, the subject is asked to draw a new attention map. Using the manually drawn attention map, the model processes the image and question one more time and produces a second answer.

In the feedback loop, the model directly multiplies the user-generated attention map into the image feature map (figure 4). This operation accentuates the image features in the highlighted areas and mitigates the features in irrelevant sections of the image.

The purpose of this operation is to allow the subject to get involved in the inference process and provide feedback to the model interactively. In cases where the model answers the questions incorrectly, subjects attempt to correct the model's response by drawing the attention. Otherwise, for those cases where the model is already accurate, subjects try to create a different answer by altering the attention map.
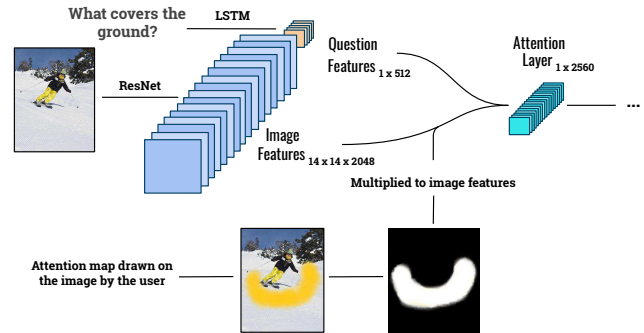

Figure 4: The architecture of active attention loop within the XVQA model.

4.3 Bounding boxes

The bounding boxes in this model are generated based on the annotations in the Visual Genome dataset and can carry important information about the scene. A combination of the attention maps created by the model and these annotations can produce explanations of the system behavior on a conceptual level. We calculate the average attention weight of the bounding boxes in the image based on the spatial attention maps and keep the top K ($K = 5$ in our studies) boxes as an indicator of most related objects in the scene contributing to the system's answer (figure 5)

4.4 Scene graph

The bounding box annotations are completed by the scene graph information which illustrates the relationships between different objects in the scene. The connections are in the form of *subject-predicate-object* phrases and can indicate object attributes or interactions. In the Visual Genome (VG) dataset, the object labels, their bounding boxes and the scene graph connecting them provide a structured, formalized representation of components in each image (Krishna et al. 2017). For each question, we filter objects in the scene graph based on the attention weights of their bounding boxes (figure 6). The users can interactively locate active objects of the scene graph and see their bounding boxes in the input image.
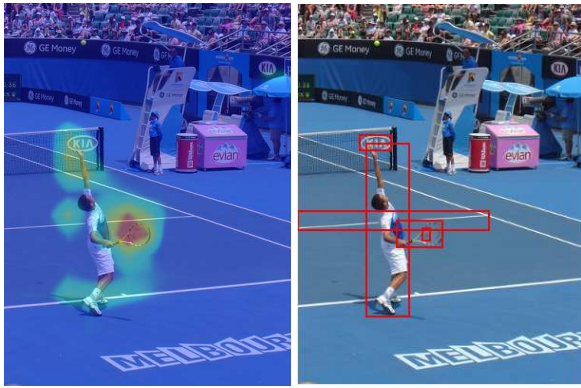
Figure 5: Bounding box explanations generated based on spatial attention weights for the question "What is the man doing?".
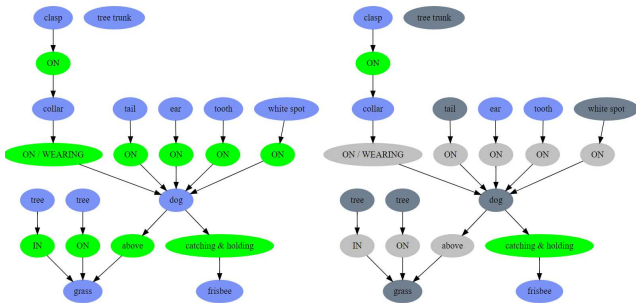

Figure 6: Left: input scene graph. Right: scene graph filtered based on the attention map weights generated by the model in response to a question.

4.5 Object attention

Inspired by previous work (Ray et al. 2019), we added a MASK-RCNN image encoder to our model to produce explanations on object-level. This encoder is specifically used by the XVQA model, as the VQA model still uses Resnet encoder to produce answers.

Model creates object attention masks based on attention modules to highlight objects with greater contributions to the inference process. As opposed to spatial attention explanations, object attentions have the capability to segment certain entities in the scene to illustrate a more meaningful explanation for system answers(Figure 7). For more details on the implementation of this technique, please refer to (Ray et al. 2019).

4.6 Textual explanation

Along with visual explanations, we also integrate natural language (NL) explanations in our XVQA system. Our technique is derived from the work done by (Ghosh et al. 2019) which uses the annotations of entities in an image (extracted from the scene graph), and the attention map generated by a VQA model while answering the question.

For a given question-image pair, our textual explanation module uses the visual attention map to identify the most

(a) Spatial attention   (b) Object attention

Figure 7: (b) Object-level attention compared with (a) spatial attention.

relevant parts of the image. The model then retrieves the bounding boxes of entities that highly overlap with these regions.

The model eventually identifies those entities most relevant to the answer based on their spatial relevance on the image and their NL representation. The region descriptions for the most relevant entities form the textual explanations. A sample output generated by this technique is illustrated in figure 8.
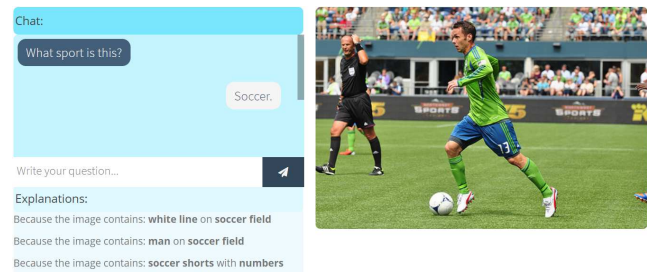

Figure 8: Sample results from the NL module producing a textual explanation for the model's answer.

5 Experimental design

For a careful evaluation of all mentioned explanation modes, we implement an interactive interface where users can take part in a user-machine prediction task. The test starts with an introduction section and continues in the form of a series of trials where the task in each trial is to estimate the VQA system's answer correctness.

Within the introduction section, the subjects are also informed of their interaction with an AI system without any implications of its accuracy to avoid any prior bias in their mental model of the system. The subjects are also provided with a set of instructions to perform the tasks and work with the interface effectively.

5.1 User task

On each trial, users enter their prediction of whether they think the system would answer the system correctly or not and then declare their level of confidence in their answer on a Likert scale. Afterward, the subjects view the ground-truth, system's top-five answers, and their probabilities in order. The system also provides its overall confidence/certainty based on normalized Shannon entropy of the answer probability distribution.

To prevent the effect of fatigue on performance in groups with longer trials, the test for each subject is limited to a one-hour session. Participants are asked to go through as many trials as possible within that period.

5.2 Trials

There are two types of trials in the experiment: no-explanation trials, and explanation trials. In no-explanation trials, subjects estimates system's accuracy only based on the input image and question.

In explanation trials, the subjects first see the inputs and system's explanations. Before estimating the correctness of system's answer, subjects are asked to rate each explanation's helpfulness towards better predicting system's accuracy. At the end of each explanation trial, subjects rate their reliance on the explanations to predict system's accuracy. Figure 9 depicts the order of actions in a trial in our evaluation system.

Each test session starts with a practice block consisting of two trials. The practice trials are only purposed to familiarize the subjects with the flow of the test and are not considered in any of the final results. The rest of the test is carried out in blocks where each block includes five trials.

5.3 Study groups

The study involves six groups of participants. The control group (NE) does not see any explanation modes, so its task is reduced to predicting the system's correctness in trials. The explanation groups are exposed to either one or a combination of explanation modes before they make their prediction about the system's answer.

The control group (NE), only sees a block of no-explanation trials throughout the whole test. For the groups with explanation modes, the blocks toggle between explanation and no-explanation modes. The no-explanation blocks in explanation groups act as control tests to assess prediction quality and mental model progress as the users see more trials. (figure 10)

The explanation blocks view the explanations generated by the model before the users make their predictions and show the answer from the system along with the system's confidence afterwards. The no-explanation blocks only ask for user's prediction without exposing any explanations beforehand.

Group SA has an interactive workflow within which subjects first go through to the spatial attention explanation and then modify the attentions in a feedback loop. Each explanation group is dedicated to a specific explanation mode, except group SE which combines bounding box, scene graph
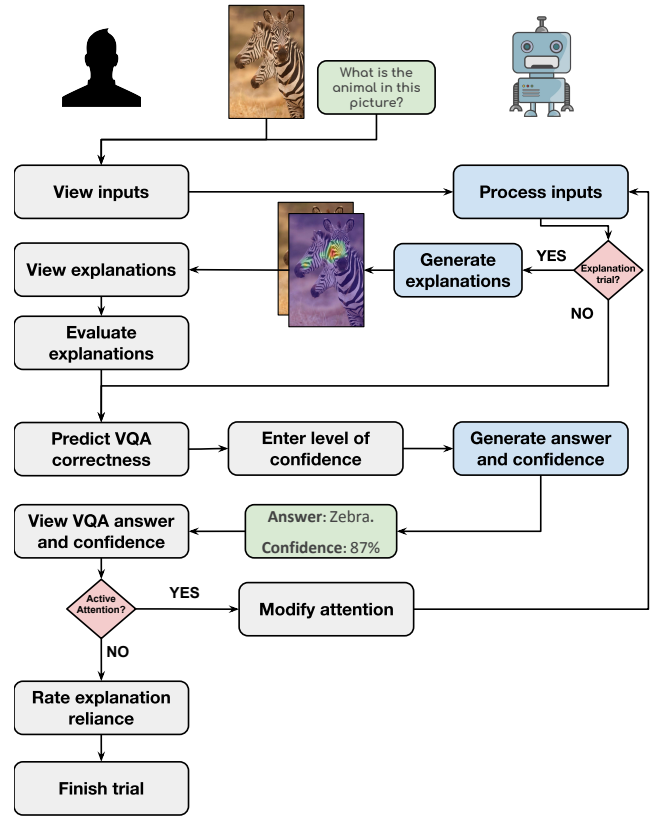

Figure 9: Flow-chart for a prediction evaluation task. The "Explanation trial?" conditional defines the type of trial as either explanation or control. The "Active attention?" conditional activates the feedback loop in case of active attention explanations.

and textual explanations. The study was conducted with 90 participants and a total number of more than 10,000 trials. Table 1 shows the number of participants in each group and the number of trials in each group.

A total number of 3969 image-question pairs were randomly selected from the overlap of VG dataset (Krishna et al. 2017) and VQA dataset (Goyal et al. 2017) to be used in the trials. The questions asked on each trial is selected from the VQA dataset and the annotations used in generating the explanations are extracted from the VG dataset. In the selection, all yes-no and counting questions were excluded to draw the focus of the test to non-trivial questions and less obvious answers with higher levels of detail in explanations.

6 Results

After assigning different groups of participants to specific combinations of explanations (including a control group that received no explanations) and having them perform the VQA prediction task, we evaluated different hypotheses about the explanations' impact on various aspects of human-machine task performance. The results are compared either based on the average of all trials within certain groups or based on the progress throughout the tests. Since the task in
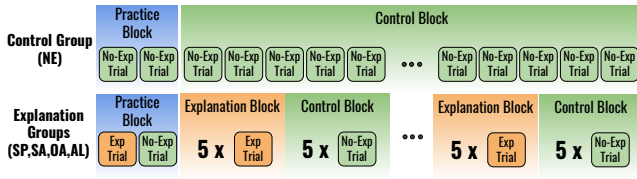
Figure 10: Structure of the test sessions in control group (NE) and explanation groups.

Table 1: User Study Design and Statistics.

	Group	Subjects	Trials
NE	Control group	15	4124
SP	Spatial attention	15	1826
SA	Active attention	15	1021
SE	Semantic	15	1261
OA	Object attention	15	1435
AL	All explanations	15	846
	Total	90	10,513

each group and trial can be different than other groups and trials, the number of trials finished by subjects vary between groups and even within groups.

6.1 Impact on user-machine task performance

The first metric we used to assess user-machine task performance is the user's accuracy for predicting the machine's correctness, and whether this is affected by explanations. We tested for any effect (positive or negative) between accuracy and presence of explanations, using a chi-squared test. The results from different groups show an overall accuracy increase in all explanation groups compared to the control group, however this is statistically significant only for cases where the system's answer is wrong (see figure 11).
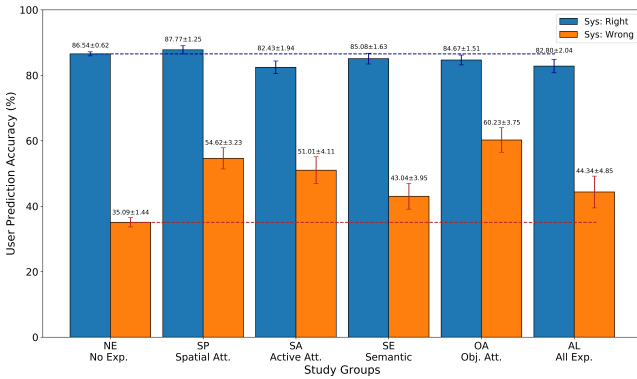

Figure 11: The average values of user's prediction accuracy (user performance) compared between different groups (sys:right p = 0.061, sys:wrong p< 0.0001, overall p=0.0001).

The progress of prediction accuracy is also another metric to quantify subject's progress in understanding and predict-
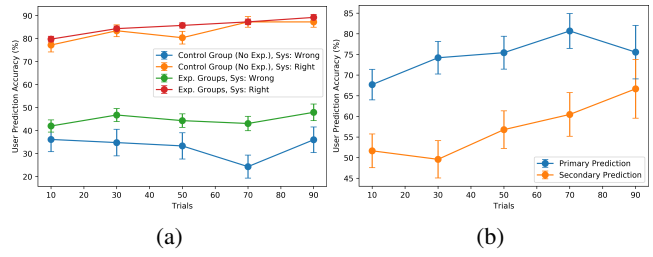

(a) (b)

Figure 12: (a) The progress of user's prediction accuracy compared between the control group and all explanation groups. The results are separated based on the accuracy of the system. (b) Prediction accuracy progress in Active Attention explanation group. primary prediction is made based on model's original attention map and secondary prediction is based on the modified attention map that the subject provides.

ing systems behaviour. As subjects go through trials in different groups, we compare the improvement of their mental model based on their prediction accuracy (figure 12a). As shown in figure 12a, in both cases whether the system is right or wrong, the subjects in explanation groups show a more steady improvement in their prediction accuracy.

6.2 User explanation helpfulness ratings

Before making a prediction about the VQA model's answer, users rate each explanation mode based on how much it helped them in the prediction task, a rating that we call "explanation helpfulness". Comparing these helpfulness ratings with the users' prediction accuracy reveals a positive correlation with accuracy improvement (accuracy after minus accuracy before) and helpfulness of explanations, but only in cases where the system is *right*. Figure 14b implies that when users find explanations helpful, they do better on the prediction task. On the other hand, a higher rating for explanations when the system is wrong (figure 14a) has lead to lower human prediction accuracy. This observation shows the effective role of explanations in the process of decision making for users.

6.3 Active Attention explanation

Within group SA, subjects view and interact with active attention explanations before making their prediction. Similar to spatial attention, users first make a prediction based on the attention map made by VQA model. On the second step, subjects draw a new attention map for the model in the purpose of changing networks answer. Subjects can compare their attention with model's attention and the answer created based on each of them. Figure 12b illustrates the trend of prediction accuracy progress as subjects interact with active attentions. While the explanation helps subjects improve their primary prediction of system's correctness, they also substantially improve in predicting system when working with their modified attention (secondary prediction).
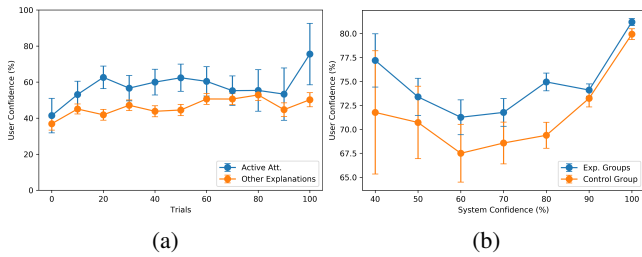
Figure 13: (a) User confidence progression comparison between the active attention group and other groups. (b) User confidence in prediction vs. system confidence in answer.

Figure 14: The average values of user's prediction accuracy for each explanation mode vs. user's ratings on explanations' helpfulness for cases where (a) the system is right, and (b) where the system is wrong. ($p < 0.0001$)

6.4 Impact of Active Attention on user confidence

Active attention explanation provides users with a feedback loop to modify the system's attention and see the result of attention changes in the model's answer. In trials with active attention explanation, users make two predictions: one based on the original spatial attention provided by the user, and a secondary prediction after they modify the attention map. We consider the accuracy of the primary prediction as an indicator of the user's mental model state. The secondary prediction is more specifically dependant on users general mental model of the attention map.

Comparing results from different explanation groups with the active attention group shows that users in the active attention group have higher average confidence in their primary predictions compared to other explanation groups (see figure 13a).

While the increase in user confidence points out the confidence and trust built by the active attention explanation, the average prediction accuracy in this group of participants is lower than other groups. These results suggest a higher potential for this technique to produce real insight into the model if used in multiple feedback loops instead of just one.

6.5 Impact on trust and reliance

Another important purpose of explanation systems is to create user trust in AI machines so that they can rely on the outcome of the system.

In our user study, we ask users about their level of reliance (in Likert scale) on the explanation section while predicting system's performance. Comparing users reliance with respect to their performance indicates a correlation between the reliance and users accuracy in those cases when the system is wrong (Figure 15a).

Moreover, users declare their level of confidence in their prediction on a Likert scale. Generally, we can assume the users' level of confidence in their prediction as a function of user confidence in the system and also the system's confidence in its answer. In the control group with no explanations, the level of confidence mainly stems from system performance in previous trials (mental model); while in other groups, the explanations have a direct effect on the level of confidence.

Figure 13b shows average user confidence compared with system confidence (provided to users after they make their
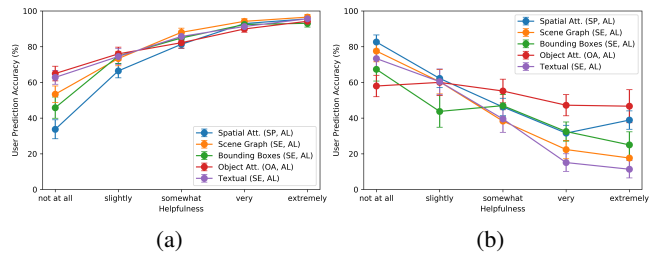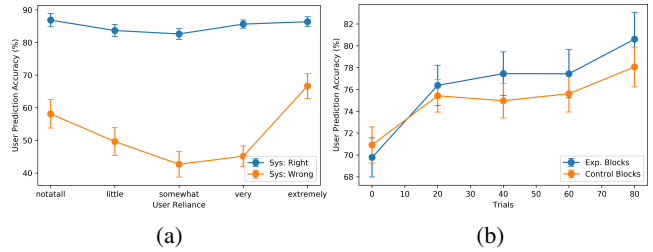

Figure 15: (a) Users prediction accuracy vs. their reliance on explanation divided by the accuracy of system. (b) Prediction accuracy growth in explanation groups compared between exp. blocks and no-exp. blocks.

predictions) in those cases when user's prediction is correct. The results indicate a consistent increase in user confidence when exposed to explanations against the control group with no explanations.

6.6 Impact of explanation goodness

As mentioned earlier, in explanation groups users go through blocks of trials. To assess the goodness of explanations in helping users predict systems answer, each block of trials with explanation is followed by a block without explanations. Comparing the user prediction accuracy between these blocks illustrates the progress of users mental model in presence of explanations (Figure 15b). Results indicate that within explanation blocks users have built a better mental model to predict system and made progress in understanding system answers.

7 Discussion

The overall assessment of user performance reveals a substantial improvement of prediction accuracy in the presence of explanations while the system is not correct. Users also provide higher ratings for the explanations when they perform better and vice versa. This direct correlation in all explanation modes strongly suggests the effectiveness of these explanations within the prediction task.

In group AL, although the subjects viewed all explanation modes, yet we do not see a higher level of accuracy compared to other groups. The feedback from the post-study

interviews pointed out two possible reasons for such observation: 1) the overwhelming amount of information in the group decreased the performance level for the subjects; 2) those cases where explanation modes conflicted with each other confused some of the subjects.

Users show higher levels of confidence when exposed to active attention in explanation groups; although, the overall performance of the active attention group (SA) does not yet exceed the spatial attention group (SP). The reason behind this drawback can be active attention's limit to only visual features and not question features. Possibly, multiple feedback loops can also help users better understand the role of image features as only one of (and not all of) the contributors in the final answer.

In cases where the system is wrong, user's accuracy show an interesting correlation with user's reliance. The subjects seem to do well either when they are extremely relying on the explanations or when they are completely ignoring them. For those cases that the users ignore the explanations, post-study interviews suggest that the subjects made their decision based on their mental model of the system and previous similar trials.

8   Conclusion

We designed an interactive experiment to probe explanation effectiveness in terms of improving user prediction accuracy, confidence, and reliance in the context of a VQA task. The results of our study show that the explanations help to improve VQA accuracy, and explanation ratings approve the effectiveness of explanations in human-machine AI collaboration tasks.

To evaluate various modes of explanations, we conducted a user study with 90 participants. Users interactively rated different explanation modes and used them for predicting AI system behavior. The user-machine task performance results indicate improvements when users were exposed to the explanations. User confidence in predictions also improved when they viewed explanations which display the potential of our multi-modal explanation system in building user trust.

The strong correlation between the users' rating on explanation helpfulness and their performance in the prediction tasks shows the effectiveness of explanations in the user-machine task performance. Those explanations identified as more helpful helped users in cases where the system was accurate. On the other hand, in cases where the system was inaccurate, those explanations ranked as more helpful became more misleading.

We also introduced an interactive explanation mode (active attention) where users could directly alter the system's attention and receive feedback from it. Comparing the user confidence growth between active attention and other explanation groups shows a higher level of trust built in users, which shows the effectiveness of interactive explanations in building a better mental model of the AI system.

As a future direction, we may investigate other interactive explanation modes to maximize the performance in human-machine tasks. On the other hand, user feedback and ratings for the different modes explored in this study can guide us towards more effective explanation models in XAI systems.

References

Anne Hendricks, L.; Hu, R.; Darrell, T.; and Akata, Z. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 264–279.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

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

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

Cosley, D.; Lam, S. K.; Albert, I.; Konstan, J. A.; and Riedl, J. 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 585–592. ACM.

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

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Ghosh, S.; Burachas, G.; Ray, A.; and Ziskind, A. 2019. Generating natural language explanations for visual question answering using scene graphs and visual attention. *arXiv preprint arXiv:1902.05715*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*.

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

Huk Park, D.; Anne Hendricks, L.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2018. Multimodal explanations: Justifying decisions and pointing to the

evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8779–8788.

Jiang, Z.; Wang, Y.; Davis, L.; Andrews, W.; and Rozgic, V. 2017. Learning discriminative features via label consistent neural network. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 207–216. IEEE.

Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018a. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018b. Pythia v0.1: the winning entry to the VQA challenge 2018. *CoRR* abs/1807.09956.

Kazemi, V., and Elqursh, A. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *CoRR* abs/1704.03162.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Kulesza, T.; Stumpf, S.; Burnett, M.; and Kwan, I. 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*, 1–10. ACM.

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

Lomas, M.; Chevalier, R.; Cross II, E. V.; Garrett, R. C.; Hoare, J.; and Kopack, M. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 187–188. ACM.

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

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016b. Hierarchical question-image co-attention for visual question answering. *CoRR* abs/1606.00061.

Narayanan, M.; Chen, E.; He, J.; Kim, B.; Gershman, S.; and Doshi-Velez, F. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.

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

Ray, A.; Burachas, G.; Yao, Y.; and Divakaran, A. 2019. Lucid explanations help: Using a human-ai image-guessing game to evaluate machine explanation helpfulness. *arXiv preprint arXiv:1904.03285*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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*, 1135–1144. ACM.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

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

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Teney, D.; Anderson, P.; He, X.; and van den Hengel, A. 2017. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *CoRR* abs/1708.02711.

Teney, D.; Anderson, P.; He, X.; and van den Hengel, A. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4223–4232.

Van Lent, M.; Fisher, W.; and Mancuso, M. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Xu, H., and Saenko, K. 2015. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR* abs/1511.05234.

Xu, H., and Saenko, K. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 451–466. Cham: Springer International Publishing.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 21–29.

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

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2014. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.