Reframing Explanation as an Interactive Medium: The EQUAS (Explainable QUestion Answering System) Project

Authors: Dhruv Batra, William Ferguson, Raymond Mooney, Devi Parikh, Antonio Torralba, David Bau, David Diller, Josh Fasching, Jaden Fiotto-Kaufman, Yash Goyal, Jeff Miller, Kerry Moffitt, Alex Montes de Oca, Ramprasaath R. Selvaraju, Ayush Shrivastava, Jialin Wu

Institutional Affiliations: Raytheon BBN Technologies, Cambridge Massachusetts; Georgia Institute of Technology, Atlanta Georgia; University of Texas at Austin, Austin Texas; Massachusetts Institute of Technology, Cambridge Massachusetts

Acknowledgement: The information provided in this document is derived from an effort sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), and awarded to Raytheon BBN Technologies under Contract Number FA8750-18-C-0004.

Keywords: Explainable artificial intelligence (XAI), Visual Question Answering (VQA), Human/Computer interaction (HCI), tasking and adapting agents

Main Text:

1 Introduction

In this paper, we share the insights and lessons that the EQUAS (Explainable QUestion Answering System) team discovered and explored during our three and a half year effort in DARPA's Explainable Artificial Intelligence (XAI) Program. In particular, we learned that simply presenting explanations in the form of static chunks of information (heatmaps, feature lists, diagrams, etc.) has a limited (but mostly positive) effect on human/machine interaction, building appropriate trust and enabling people to build useful and accurate mental models of the AI system. If instead, these explanations can support interactions, either by being editable or through iterating with a human user, then the power of explanation rises considerably. Indeed explanation-based interaction may be central to support interacting with an AI system while understating its virtual, intentions and choices.

In this paper, we will sample the journey that we took to explicate explanatory competence. Our successes are preliminary, our conclusions are still tentative but the core of what we (and other teams in the explainable artificial intelligence program) learned should be important as explainable AI research moves forward.

2 Leveraging explanations to make vision and language models more grounded:

Today's state-of-the-art deep models – especially for vision and language tasks – are known to rely heavily on superficial correlations in training data. As a result, these models are often biased by language priors, and do not make predictions that are sufficiently grounded in the image content. This section describes work that effectively inverts the popular process of generating salience maps to explain to humans where in images a machine pays attention during task performance., It does this by collecting ground truth data from humans about which regions of images are most important during task performance, and using that data during training as a kind of explanation to guide the model's attention within the image. This enables it to improve its performance via more robust visual grounding.

Extending insights gained from earlier work on Gradient-weighted Class Activation Mapping (Grad-CAM1)) (that uses the gradient information flowing into the last convolutional layer of a Convolutional Neural Network (CNN) to assign importance values to each neuron and thereby generate a salience map), this work establishes a generic approach called Human Importance-aware Network Tuning (HINT). HINT encourages deep networks to be sensitive to the same input regions as humans by optimizing the alignment between human attention maps and gradient-based network importance – ensuring that models learn not just to look at but rather rely on visual concepts that humans found relevant for a task when making predictions. We apply HINT to Visual Question Answering (VQA) and Image Captioning tasks, outperforming top approaches on splits that penalize over-reliance on language priors using human attention demonstrations for just 6% of the training data.

HINT estimates the importance of input regions through gradient-based explanations and tunes the network parameters to align with the regions deemed important by humans.

We evaluate HINT using VQA-CP[3] – a restructuring of VQAv2 that is designed such that the answer distribution in the training set differs significantly from that of the test set. Without proper visual grounding, models trained on this dataset will generalize poorly to the test distribution. The HINTed model significantly improves over its base architecture alone by a sever percentage point gain in overall accuracy. Further, it outperforms existing approaches based on the same base architecture (41.17 vs 46.73), setting a new state-of-the-art for this problem, at the time of publication. We do note that our approach uses additional supervision in the form of human attention maps for 6% of training images.

Please refer to Selvaraju, et al[4] for more details.

3 Contrastive explanation for teaching

In this work, we study discriminative explanations of the form 'For input *X*, why did the model predict *Y* instead of *Z*?' One way to answer this question is through counterfactual reasoning, i.e., how should I change the input minimally such that the outcome changes from *Y* to *Z*. In the context of the task of image classification, given a 'query' image I_1 for which a classification model predicts class c_1, a counterfactual visual explanation identifies how I_1 could change such that the system would output a different specified class c_2.

To generate these counterfactual visual explanations, we develop a technique where we first select a 'distractor' image I_2 that the model predicts as class c_2. Then we identify spatial regions in I_1 and I_2 such that replacing the identified region in I_1 with the identified region in I_2 would push the model towards classifying I_1 as c_2 (refer to Figure 1). We apply our approach to multiple image classification datasets generating qualitative results displaying the interpretability and discriminative nature of our counterfactual explanations. More details about the approach and results can be found in Goyal, et al[5].

We investigate if our counterfactual explanations can help in teaching a fine-grained bird classification task to lay people. To evaluate this, we design a machine teaching interface where we first train the subjects for the fine-grained task using our counterfactual explanations and then test them on new instances. Our interface is shown in Figure 2. We compare our human study with two baselines, which only differ in terms of the feedback shown to human subjects -- no explanation or a non-counterfactual, feature-attribution explanation generated via GradCAM[1] in place of our counterfactual explanations.

The mean test accuracy with counterfactual explanations is 78.77%, with GradCAM explanations; it is 74.29% while the mean test accuracy without any explanations is 71.09%.

Therefore, our studies indicate that our counterfactual explanations are more effective in this machine teaching task as compared to non-counterfactual feature attribution explanations or no explanations at all.

4 Rewriting Rules in a Generative Model

While many explanation methods focus on explaining a single prediction at a time, we can also investigate how a user can understand the general rules in a model. We ask, can a user directly manipulate and change the internals of a generative model by understanding its structure directly, and without training the model on any new images? The goal is not to alter just a single image, but to edit the generalized computational rules encoded within the model, even when training data is unavailable. For example, can a person directly manipulate a trained model's rule for the appearance of the top of a tower so that *all* the towers have trees growing from them? See Figure 3.

To enable such direct model editing, we develop a method[6] for rewriting a single layer of a deep generative model as a linear associative memory. Our method views the weights of a layer as a matrix storage (an optimal linear associative memory[7] that associates vector input keys with vector output values, and allows insertion of a new memory by performing an error-minimizing rank-one change in the weights of the matrix. To enable a user to perform such an update in an interpretable way, we create a user interface (See Figure 4) that allows a user to write into a specific memory selecting a small handful of examples that are used to infer the location and the new content for the memory to replace within the layer.

We demonstrate our method on Progressive Generative Adversarial Network (GAN[8]) and StyleGAN[9] models trained to model a variety of image data sets, and we benchmark our method against prior methods for propagating edits from one image to other images by user's evaluation of image realism of edited human faces. We find that our method provides more realistic changes than the previous leading image-propagation editing method: "neural best buddies"[10] (90.2% of our edits of human face outputs are considered more realistic). We find that our method reveals the structure of the model, for example, that a human face model separately memorizes rules for parts of child faces from adult faces, so that editing a single memorized rule can edit child faces without modifying the faces of adults at all.

5 Improving VQA and its Explanations by Comparing Competing Explanations

We have also developed a novel framework that uses explanations for competing answers to help VQA systems select the correct answer[11]. Instead of using explanations to elucidate reasons for the system's answer to a human user, this approach uses them to allow the system itself to more deeply examine the rationales for competing potential answers, and reweight them based on this additional information. We have shown that this improves *both* system accuracy as well as the quality of its explanations as evaluated by humans. An example of the system comparing textual explanations and using them to reweight competing answers to a VQA problem is shown in Figure 5.

Our framework is end-to-end trainable and therefore applicable to any differentiable VQA system. It learns better representations for the questions and visual content by training to retrieve explanations, and achieves state-of-the-art results by further jointly considering competing explanations. Using human textual explanations, our framework builds better representations for the questions and visual content, and then reweights confidences in the answer candidates.

Our experiments show improvements for our approach applied to Up-Down[12] (UpDn) and Learning Cross-Modality Encoder Representations from Transformers[13] (LXMERT) for the VQA-X dataset[14], which comes with human textual explanations for the answers.

After the base VQA system computes the top-*k* answers, our approach retrieves the most supportive explanations for each answer from the training set to construct the set of competing explanations. These explanations are used to help generate explanations for the current question. We learn to predict verification scores that indicate how well the retrieved or generated explanations support the predictions given the input question and visual content. The final answer is determined by jointly considering both the original answer probabilities and these verification scores. Details of the neural architecture and how it is trained are given in[11].

With respect to question-answering accuracy, we improve the original UpDn and LXMERT by 4.5 and 1.2 percentage points, respectively. UpDn benefits more from using competing explanations than LXMERT, but both improve. By using transformers, LXMERT already creates better, but less flexible, representations that are harder to improve upon by using explanations.

We evaluated the generated explanations using both automatic evaluation metrics comparing them to human explanations, and human evaluation employing crowdsourced judges. We compared to our previous state of the art VQA explanation system[15] as a baseline. Explanations from our new approach achieve better automatic evaluation scores, and more importantly, higher human ratings. In particular, human judges rate our explanations as good as or better than human explanations 55.6% of the time, whereas the baseline scores 49.5% by this metric. Full details of the evaluation are given in[11].

6 Instruction Following Experiment with Human in the loop

In this section, we study the performance of instruction following navigation agents with human in the loop. Vision and Language Navigation (VLN)[16] is an instantiation of the instruction following task where an agent is placed in a photo-realistic reconstruction of an indoor environment. The agent is given a natural language navigation instruction and asked to follow the trajectory. In this work, we study the following: if the model were to show a human the different ways in which it might execute the instruction, can a human identify which way is better, more accurately than the model's own beliefs? The visualization of how a model might execute the instruction can be thought of as a mode of explanation.

We use instructions from the unseen validation split of the VLN dataset[16], extract 30 trajectories using[17] and score the compatibility between each trajectory and the instruction using the VLN-Bidirectional Encoder Representations from Transformers (VLNBERT model[18].) Note that this compatibility score forms the model's own belief about which way of executing the instruction is best. For each instruction, we take the top five ranked trajectories and ask humans to select the best trajectory (out of five) based on how closely it follows the instruction mentioned. Using the human selected trajectory as the prediction, we evaluate the VLNBERT+Human performance on the nDTW metric[19,] which measures how closely the instruction was followed by the trajectory.

Results: VLNBERT achieves 60.2% on nDTW. The upper-bound Oracle performance that we get by selecting the best trajectory among five trajectories is 84.6% nDTW. When we evaluate the VLNBERT+Human model by considering all users individually, we see that it gets 69.0% nDTW (9.2% increase).

In this study, we observe that the performance of a state-of-the-art VLN agent can be increased by 13% on nDTW when paired with a human to select the best trajectory based on a visualization of the different ways in which the agent can execute the instructions.

7 One shot image detection and collaboration

One area where "explanation as a medium" can be useful is in the domain of one shot image detection. In this constrained setting the user only has one, or a few instances of a previously unseen target class (e.g. a new airplane) and wants to adapt the system to recognize this class in future data. We test these detectors by using them to find more instances of the target class in an unlabeled corpus of data, but such a corpus is by definition, not be available when the detector is defined – as it would be in a classic, image retrieval problem. User explanation is especially useful here as there are no more representatives of that target class (and therefore no false negatives to show) but the human user must still impart domain knowledge to the system.

We explored two modes of providing a one shot detection system with additional information from the user. The first mode explored augments a linear classifier trained on the one shot image and allowing the user to pick from five features chosen by a heuristic. We wanted to see if the user could select a feature such that if that features weight in the linear classifier were reassigned to be the most negative feature weight, the user would improve the classifier's F1 score for the new target class. We found it was possible (Turkers were (38%) correct vs. random selection (20%), however the impact that the user had on the F1 score was minimal compared to the amount of "damage" to the classifier. The overall average effect was -0.006 change in F1.

The second mode we explored creates a one shot detector from scratch by allowing the user to paint "aspects" on their one example of the target class to highlight discriminative qualities about that class. An aspect for the target class is defined as some distinguishing feature of that class such as a part (e.g. nose cone). Refinement of these aspects is assisted by the aid of a user interface (see Figure 6) where the user is presented with a list of images ranked according to how they are alike positively annotated image regions and different from negatively annotated image regions.

The user is able to refine their search in one of two ways by incorporating images from the returned query set, marking either positive or negative regions. These aspects are then used to build a two-layer network classifier. Work in this direction is still on going with further emphasis on how best to define negative features for annotation to better separate the decision space.

We learned several lessons by exploring "explanation as a medium" for human/machine learning (ML) system collaboration. "Expert" users who have some familiarity with the underlying system at work appear to do better than uninitiated users (Turkers). Giving users more freedom of control allows for using this collaboration in unexpected ways. They develop their own imperfect mental model of how the system works and try to use that knowledge for improvement gains. One unintended behavior in the second mode was that users found improvement in the overall classifier by only refining an aspect with negative images from the query set. By employing user studies, it might be possible for human/ML system designers to uncover more robust ways of having the pair work together.

8 Discussion

It has been our observation over the course of the XAI program that there are three functional or ontological characterizations of explanation that formal progression of sophistication and potential utility of explanation (Table 1).

The XAI program began focused on level one and that remained its official focus. Many performers moved on to level two in order to improve human/machine task performance. Some began to explore level three by considering how the user could adapt or correct the system's performance by changing its explanations or by adding rules.

We traveled a long way from our start with VQA, salience maps and named features. Early on presented explanations helped people with estimating competence and predicting success but the latter effect was slight. Many XAI performers were getting better results when users interacted with explanations. So we evolved the notion of explanation as an interactive medium – usually, between humans and AI systems during task performance (sections 3, 4). Explanations can also interact, via internal mechanisms to improve AI system performance – by training the system to align its attention with human explanations (section 2) or by using system generated explanations to decide among possible answers (section 5). In this case, the system's representation of its own explanation (as medium for metacognition) allowed a system to perform better by leveraging its descriptions of its own reasoning.

As we began to explore in our later work, interacting with explanations can enable profound capability for people to task and adapt ML agents. It works like this: Human interaction with any system requires an interpretable interface. An interpretable interface to an ML agent can be built from an explanation it provides by adding affordances for editing the explanation and designing the ML system to act in accordance with the modifications. In this way, editing an explanation can adapt a system's performance to some new, modified purpose (section 7). Additionally, there is still more value; ML systems can predict their future behavior (as a kind of planning) if this future behavior can be explained and the explanation made editable, then the system's "plan" *and the rational for that plan* (the intent, as embodied in the explanation) can be changed. This would allow for the agent to be effectively and robustly tasked. (Preliminary work on this technique appears in section 8). This deep tasking, wherein the agent knows its objective and the explanation for that objective will be critical to enable higher levels of autonomy.

Our work and the work of others in the XAI program has shown that people can learn something about ML system's representations and competence from statically presented explanations and use that knowledge. For example, salience explanations help people predict competence and contrastive explanations help people learn discriminative tasks. In the future, explanation as a medium for interaction shows promise for enabling human adapting, correcting and tasking of autonomous agents.

References:

1. Selvaraju R, R Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *ICCV*, 2017.
2. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6077-6086
3. Agrawal A, Batra D, Parikh D, Kembhavi A. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4971-4980
4. Selvaraju R, Lee S, Shen Y, Jin H, Ghosh S, Heck L, Batra D, Parikh D. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2591-2600
5. Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S. Counterfactual Visual Explanations. *ICML*, 2019.
6. Bau D Liu S, Wang T, Zhu J,Torralba A. Rewriting a deep generative model. In *European Conference on Computer Vision*, Springer, Cham 2020. 351-369 p.
7. Kohonen, T. Correlation matrix memories. IEEE transactions on computers 100, no. 4 1972.353-359 p.
8. Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017).
9. Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110-8119. 2020.
10. Aberman, Kfir, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. "Neural best-buddies: Sparse cross-domain correspondence." ACM Transactions on Graphics (TOG) 37, no. 4 (2018): 1-14.
11. Wu, J. and Mooney, R.J., "Improving VQA and its Explanations by Comparing Competing Explanations," Proceedings of the AAAI Workshop on Explainable Agency in AI, Feb. 2021.
12. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.;Gould, S.; and Zhang, L. "Bottom-Up and Top-Down Attention for Image Captioning and VQA." In CVPR, 2018.
13. Tan, H.; and Bansal, M. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers." In EMNLP, 2019.

14. Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.;Schiele, B.; Darrell, T.; and Rohrbach, M. "Multi-modal Explanations: Justifying Decisions and Pointing to the Evidence." In CVPR, 2018
15. Wu, J.; and Mooney, R. J., "Faithful Multimodal Explanation for Visual Question Answering." In ACL BlackboxNLPWorkshop, 2019.
16. Anderson, Peter, et al. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
17. Fried, Daniel, et al. "Speaker-follower models for vision-and-language navigation." arXiv preprint arXiv:1806.02724 (2018).
18. Majumdar, Arjun, et al. "Improving vision-and-language navigation with image-text pairs from the web." European Conference on Computer Vision. Springer, Cham, 2020.
19. Ilharco, Gabriel, et al. "General evaluation for instruction conditioned navigation using dynamic time warping." arXiv preprint arXiv:1907.05446 (2019).

Tables:

Table 1

Level	**Purpose**	**Challenge**	**Value**	**Focus**
1	Reveal or display the workings of the decision mechanism	Translate system internal representation and mechanism into a human meaningful, compact, unified form	Enable system debugging; enable trustworthiness assessment; enable mental modeling.	The explainer (the AI system)
2	Allow users of the system to make better choices	Adapt explanations to anticipate the information that he user needs	Build appropriate trust; enable behavior prediction; provide user satisfaction.	The "explainee" (typically, the human user)
3	Allow system and user to collaborate	System would need some theory-of-mind modeling to track common ground	Enable mutual reliance, team work, co-teaching; maintains common ground[†]	The interaction

---

[†] This may be one of the reasons that appropriate explanations engender trust;It is not simply that they assure the user that the system is working for sensible reasons. The overarching reason may be that users begin to believe that the system is "taking responsibility" to be understood. This is a profound implicit promise that all interlocutors make to each other, and which is necessary to facilitate communication.

Figure Legends:

Figure 1: Our approach explains why the query image (left) was classified as *1* (top row) or *Eared Grebe* (bottom row) rather than *4* or *Horned Grebe* by finding regions in a distractor image (middle) and the query image (red boxes) such that if the highlighted region in the query image looked like the highlighted region in the distractor image, the resulting composite (right) would be classified more confidently as *4* or *Horned Grebe*.

Figure 2: Our machine teaching interface. During the training phase (shown in (a)), if the participants choose an incorrect class, they are shown feedback (shown in (b)) highlighting the fine-grained differences between the two classes. At test time (shown in (c)), they must classify the birds from memory.

Figure 3: Our method enables a user to edit a generalized rule in a generative model, and can be used to remove watermarks; alter an existing pattern such as the density of crowds; or insert a new rule such as trees growing out of tops of towers.

Figure 4: Flow of user interface that allows a user to modify a single memorized association within a deep model. The user selects a region of an image (a) containing a new pattern they wish to insert in a new place in the model. The user selects several examples of contexts (b, c) where they wish the new pattern to appear. Our method inserts the new key-value pair in a layer of the model in order to change one rule.

Figure 5: An example of utilizing explanations to correct a VQA prediction. Although the original VQA confidence of the correct answer ``Yes'' is lower, the explanations for ``Yes'' support their answer better, resulting in a higher verification score and a final correct decision.

Figure 6: One shot image detection interface. Left column presents the images from the database determined most similar to the aspect annotated from the images on the right. The yellow lasso in the left column indicates the region of activation for the aspect.