

## ARTICLE TYPE

# Generating and Evaluating Explanations of Attended and Error-Inducing Input Regions for VQA Models

Arijit Ray<sup>\*1</sup> | Michael Cogswell<sup>1</sup> | Xiao Lin<sup>1</sup> | Kamran Alipour<sup>2</sup> | Ajay Divakaran<sup>1</sup> | Yi Yao<sup>1</sup> | Giedrius Burachas<sup>1</sup>

<sup>1</sup>Center for Vision Technologies, SRI International, Princeton, NJ, USA

<sup>2</sup>Department of Computer Science, University of California, San Diego, La Jolla, CA, USA

## Correspondence

\*Arijit Ray, Email: [arijit.ray93@gmail.com](mailto:arijit.ray93@gmail.com)

## Summary

Attention maps, a popular heatmap-based explanation method for Visual Question Answering (VQA), are supposed to help users understand the model by highlighting portions of the image/question used by the model to infer answers. However, we see that users are often misled by current attention map visualizations that point to relevant regions despite the model producing an incorrect answer. Hence, we propose Error Maps that clarify the error by highlighting image regions where the model is prone to err. Error maps can indicate when a correctly attended region may be processed incorrectly leading to an incorrect answer, and hence, improve users' understanding of those cases. To evaluate our new explanations, we further introduce a metric that simulates users' interpretation of explanations to evaluate their potential helpfulness to understand model correctness. We finally conduct user studies to see that our new explanations help users understand model correctness better than baselines by an expected 30% and that our proxy helpfulness metrics correlate strongly ( $\rho > 0.97$ ) with how well users can predict model correctness.

## KEYWORDS:

VQA, Error Maps, Attention Maps, Explanations, Evaluations

## 1 | INTRODUCTION

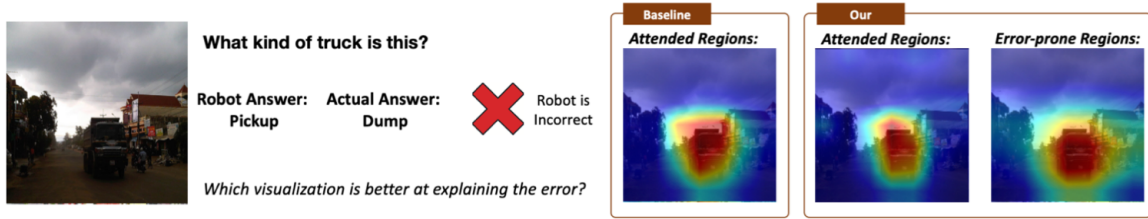
Similar to human-human collaboration, users should have good mental models of machines' capabilities for effective human-machine collaboration<sup>1</sup>. This can help users calibrate their expectations on models' robustness to establish a proper level of trust of machine decisions<sup>2</sup>. To improve users' mental models of machines, there has been a recent focus on generating "explanations" that shed light on machines' inner workings<sup>3</sup>. In the context of VQA, visualizing an attention map over an image is a popular form of explanation that highlights relevant regions used by the system for answering the question<sup>4,5,6,7</sup>.

For example, as shown in Figure 1, to explain the answer "pickup" produced by a VQA system to the question "What kind of truck is this?", an attention map is generated to highlight the vehicle region.

Users' ability to understand VQA model's performance while using such explanations is usually tested by asking users to predict model correctness while viewing the explanations along with the input question and image<sup>8,9</sup>. However, in these studies, attention maps haven't been shown to be very helpful<sup>8,10,9,11</sup> in assisting users to predict VQA system correctness.

To analyze how humans interpret explanations and why current attention maps don't help, we first observe that users generally think a VQA model will produce a correct answer when the attention maps point to the relevant evidence in the image (measured

<sup>0</sup>Project page: <https://bit.ly/3cf9Tc>



**FIGURE 1** We model how humans interpret heatmap explanations and propose novel explanations that are helpful to users for understanding system performance. We introduce error maps that point to regions that potentially cause system error. If error maps align with attended regions (as pointed to by attention maps), it indicates a possibility of system failure, and vice versa.

using human attention as described in Section 3 and illustrated in Figure 3). However, as shown in Figure 1, we frequently observe that attention maps point to relevant regions (the truck) even when VQA answers incorrectly (“pickup” instead of “dump”). This is misleading to a user who would intuitively guess that the system will be correct because it looked at the right locations.

To clarify cases where an incorrect answer is produced in spite of attention pointing to the relevant regions, we consider an additional mode of explanation to highlight image areas that may be processed incorrectly - we refer to this as an Error Map. Taking Figure 1 for example, the Error Map points to the relevant vehicle region indicating that the vehicle may not be processed correctly (maybe because of low illumination, complexity of the entity, similarity to confusing entities, etc.). We generate error maps by training a Justifying module to learn model correctness and then using GradCAM<sup>12</sup> to highlight visual evidence that contribute to a prediction of model error.

In order to measure the potential of our explanations to help users predict model correctness, we introduce a proxy helpfulness metric by simulating the way humans interpret them. The metric is based on 1) judging whether the explanation highlights information the human thinks is relevant and 2) using the relevance as an indicator to decide if the system can produce the correct answer or not. Attention maps that point to relevant regions for correct answers would be helpful according to such a metric. We also outline a few ways to optimize the visualization of attention maps to maximize these helpfulness characteristics.

Finally, we demonstrate that our new error maps, combined with attention maps, can improve users’ ability to predict model performance compared to using baseline attention maps or using no explanations.

In summary, our major contributions include: 1) We reveal **characteristics that make heatmap-based explanations helpful** for predicting model’s correctness. We introduce a proxy metric based on these characteristics to check how helpful an explanation may be to understand model correctness. **This bypasses the need for multiple expensive and laborious user studies during development.** 2) We propose a novel mode of explanation, **Error Maps, that infers the regions that induce error in the output.** We demonstrate the effectiveness of error maps, combined with attention maps, in improving helpfulness of explanations from the perspective of assisting users to predict model correctness. 3) We employ our metrics to optimize the visualization of attention maps from multiple heads of transformers<sup>13</sup>, which otherwise is done via manual inspection in literature. 4) We conduct **user studies to verify the effectiveness of our new attention and error maps.**

In the sections below, we first describe our VQA model and how attention maps are commonly generated in literature. We then describe how users interpret these attention maps, the problem with current attention maps, how error maps can help improve helpfulness. By simulating user interpretation, we describe our proxy metric for evaluating helpfulness of the attention/error maps. We then describe how we generate error maps and their evaluation.

## 2 | VQA-Net

We first describe our AI system of choice - a Visual Question Answering (VQA) model, which we use to study what makes explanations helpful to make better explanations. We build a transformer-based<sup>13</sup> VQA system that answers questions about images.

As shown in Figure 2, our VQA-Net takes an image,  $I$  and a question,  $Q$  as input and outputs a probability distribution over 3129 answer choices. The image  $I$  is encoded into a  $7 \times 7 \times 2048$  feature map using a Resnet152<sup>14</sup> and further encoded into a  $49 \times 512$  input feature matrix,  $H_m$ . We also extract 36 object-based features from the image using a Region Proposal Network<sup>15</sup>,

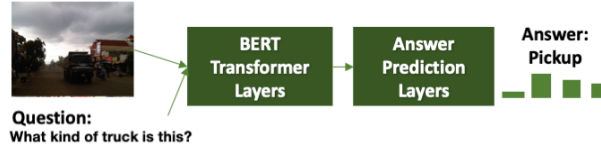


FIGURE 2 High Level diagram of our VQA model

which is encoded into a  $36 \times 512$  object feature array,  $H_o$ . Finally, the question is encoded into a  $30 \times 512$  feature array,  $H_q$  with a max word length of 30<sup>1</sup>.

We employ transformer-based attention layers (referred to as **AT-BERT**) that jointly take in the image and question features. The input to the AT-BERT layers are concatenated image and question features ( $H_m$ ,  $H_o$  and  $H_q$ ). Hence, the input dimension to the AT-BERT layers consists of 115 ( $30 + 36 + 49$ ) input tokens of dimension 512 each. Our AT-BERT has 4 layers with 12 heads each. In each head and layer of AT-BERT, 115 input tokens generate an attention over all other 115 input tokens. Hence, a AT-BERT module with  $l$  layers and  $h$  heads produces weights,  $W$ , of dimension  $l \times h \times d \times d$  attention weights.  $d$  is the number of input tokens, which is 115 in our case. Finally, the VQA-Net predicts a softmax probability distribution over 3129 answer choices (i.e., the top VQA2.0<sup>16</sup> answers) from the attention weighted feature values. Our VQA-Net is trained on the VQA 2.0 dataset<sup>17</sup>.

**Attention Map Generation** Following common practice in literature<sup>18</sup>, we display attention by choosing the layers and heads according to a visual inspection on a held-out val-train set. Specifically, the attention is selected by averaging the attention weights on image features by all input tokens in all heads of the last layer of AT-BERT.

### 3 | HUMAN INTERPRETATION OF EXPLANATIONS

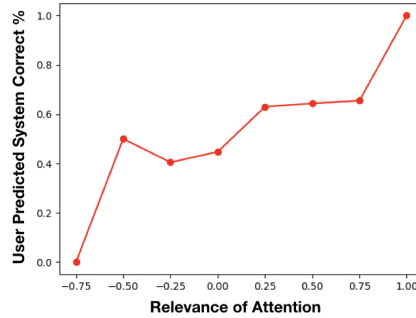
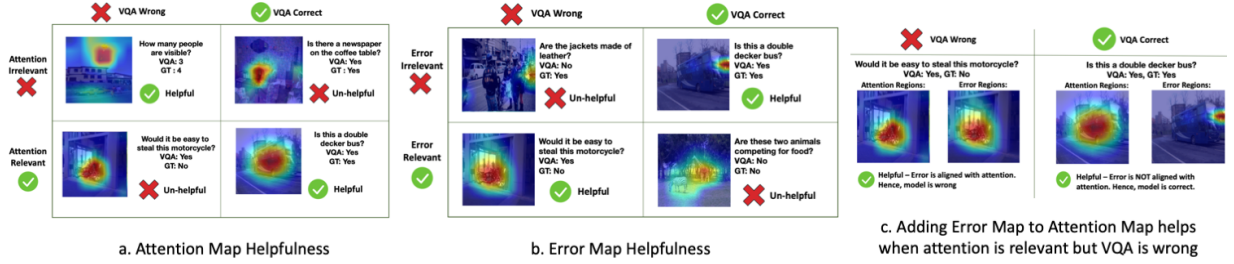


FIGURE 3 Users believe a model will be correct when the attention points to relevant areas of the image.

Our goal is to improve human's understanding of the model when seeing explanations such as attention maps as described in the previous section. In order to test users' understanding of the model, we conduct a user-study to see how well users can predict the performance of the network when seeing these attention maps. Given an Image-Question (IQ) pair and an attention map, users needed to judge if a VQA model can produce the correct answer or not. We compare this to a group of users seeing no explanations (only image and question) while trying to predict model correctness. We provided a bonus incentive that was scaled to their accuracy in predicting the VQA performance.

**Users get misled by current attention maps** We observe that the users' accuracy in predicting correctness does not improve over not seeing any attention maps. In fact, we see a slight decrease (57.18% without explanations, 56.87% when seeing attention maps), suggesting that the attention maps may be misleading users. To understand why we see a decrease in correctness prediction accuracy, we first analyze how humans interpret these explanations and the fault with the current visualization of the attention maps.

<sup>1</sup>Questions below that length are padded with 0's



**FIGURE 4** **a.** Attention is perceived as helpful when it points to the relevant regions while the VQA model produces the correct answer (bottom-right quadrant) and points to the irrelevant regions while the VQA model produces incorrect answers (top-left quadrant). **b.** Error map is perceived as helpful when it points to the irrelevant regions when the VQA model answers correctly (top-right quadrant) and points to the relevant regions when the VQA model answers incorrectly (bottom-left quadrant). **c.** Combining attention and error maps improves the helpfulness of explanation especially for cases where attention points to the relevant regions while the VQA answers incorrectly (bottom-left quadrant of Figure 4). With error map, the incorrect answer is better explained (left panel, this figure).

### Users associate relevance with correctness

We sample subsets of IQ pairs from our user study set and plot the average relevance of attention in those sets to the percentage of cases users think the model will be correct at to compute Figure 3. We see that users predict the model will be correct more on average when attention maps pay attention to relevant information. Our definition of **Relevance** makes use of annotated human attentions for IQ pairs from the VQA<sup>16</sup> dataset collected by<sup>19</sup>. This dataset of attention maps show where humans looked to answer VQA questions. Using these human attention maps, we define the *Relevance*  $REL_A$  of an explanation  $A$  as the Spearman rank correlation  $\phi(\cdot)$  of the explanation with the human attention  $A_h$  for a given IQ pair<sup>2</sup>:

$$REL^A = \phi(A, A_h) \quad (1)$$

We choose a rank correlation over the commonly used Pearson’s correlation because we want to measure whether the directions of change of attention values are similar to that of the human attention values rather than the absolute magnitudes. Absolute magnitudes depend on the temperature of the attention distribution, which is irrelevant to the similarity we intend to measure, yet can still change the Pearson correlation significantly.

Using the fact that users associate relevance with correctness as seen in Figure 3, we can assume that users find it intuitive to understand model behavior when attention maps point to relevant regions when the system is correct and vice versa. Figure 4a illustrates this point- it is intuitive to understand machine’s correctness for the upper left quadrant (the attention is relevant and the machine is correct) and for the bottom right quadrant (the attention is irrelevant and the machine is incorrect). However if the attention is relevant and the answer is incorrect or vice versa (quadrants 2 and 3), it is unclear why the machine is correct/incorrect.

**Attention Maps seem to be relevant even for incorrect cases** We observe that the attention maps point to relevant image regions even when the model produces an incorrect answer. The average  $REL^A$  for the attention maps for correct answer cases is 0.32 and for wrong answer cases is 0.31. This indicates that there is very little visual difference between the correct and incorrect cases. Hence, we need another mode of explanation that explains why the model was incorrect despite attention pointing to similarly relevant areas as correct cases.

**How Error Maps can help** To clarify cases where the model produces an incorrect answer even when the attention is relevant, we introduce a novel method of explanation - Error Maps, that indicate whether the attended region will be processed correctly or not. If users indeed use the relevance of the attention as a signal for when the model is correct, we can assume that users would similarly use the relevance of the error map to judge if the model will answer incorrectly or not. If an error map points to relevant areas of the image or to the attended regions, a user would assume that those areas may not be processed correctly. As illustrated in Figure 4b, it is easy to understand why a machine failed when the error map points to the relevant region needed

<sup>2</sup> $7 \times 7$  attention values are flattened into a vector.

to answer the question (bottom left quadrant) and why a machine succeeded if it points away from the relevant regions (upper right quadrant).

#### 4 | A PROXY HELPFULNESS MEASURE

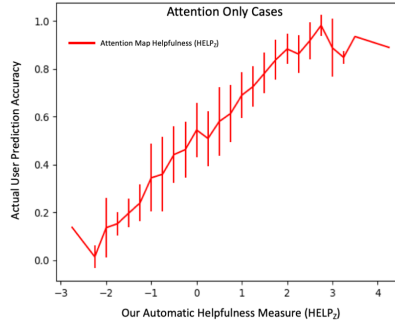
Using our knowledge that humans judge model correctness based on how well the explanations point to relevant areas, we design a proxy helpfulness metric based on relevance to evaluate the potential of an attention map to help users to understand machine correctness. We can use this metric to evaluate explanations quickly during development without having to conduct repeated user-studies. Note that, in this work, we focus on helpfulness of explanations rather than faithfulness of explanations to the model’s inner workings.

**Helpfulness of Attention Maps:** Our helpfulness metric for Attention Maps is higher if their relevance is higher for cases where the VQA model produces correct answers than it is for cases where the VQA model produces incorrect answers. The higher the difference, the better a user can visually tell apart when a model would be incorrect based on the attention relevance. As such, it is defined on a set  $S$  of IQ pairs where sometimes the model is wrong and sometimes the model is right.

We conduct a z-test<sup>3</sup> for comparing the set of correct answer *Relevances* to the same for incorrect answers. Higher the t-statistic of the z-test, the higher the significance of the difference and hence, the more the helpfulness. Let  $REL_c^A$  be the set of correct answers relevances and  $REL_w^A$  for wrong answers, we compute

$$HELP_Z^A = ztest(REL_c^A, REL_w^A) \quad (2)$$

**This helpfulness metric indeed correlates to user accuracy on the task** As illustrated in Figure 5, we note that there is a



**FIGURE 5** Users’ accuracy in predicting model correctness gets better on subsets that have higher helpfulness scores as evaluated by our automated metric.

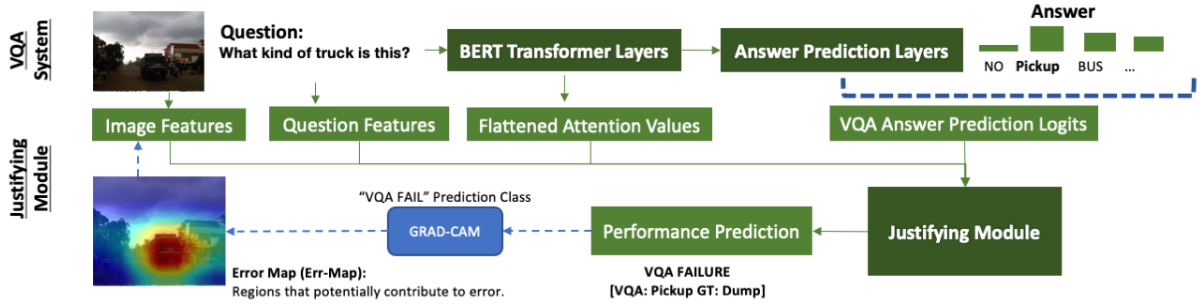
strong positive correlation between our  $HELP_Z^A$  score and actual users’ accuracy in predicting model correctness. To compute the graph in Figure 5, we sample subsets of our user study set of varying user accuracies and compute the helpfulness  $HELP_Z$  on those sets. We then plot the average users’ accuracy in predicting model performance for these binned  $HELP_Z$  values.

**Helpfulness of Error Maps:** Using the same intuition as for attention maps, we can assume that users would also use a relevance-based criteria to judge error maps. However, since error maps point to error contributing areas, the helpfulness metric should be higher if their relevance is lower for cases where the VQA model produces a correct answers than it is for cases where the VQA model produces incorrect answers.

To account for such an inverted behavior, we simply add a negative sign to indicate that a negative correlation is more helpful in the correlation based helpfulness:

$$HELP_Z^{ERR} = -ztest(REL_c^{ERR}, REL_w^{ERR}) \quad (3)$$

<sup>3</sup>z-test calculated using statsmodels.org Python package



**FIGURE 6** We train a Justifying Module that learns to predict VQA System’s performance. We compute the visual evidence for VQA Failure from the Justifying Module that highlights the error-contributing areas in the image for the given question (referred to as ErrMap). In the example shown, the question is “What kind of truck is this?”, the ErrMap regions point to the truck, indicating that it might confuse the system. The VQA understandably answers incorrectly.

With these definitions, a higher score in any *HELP* metric always means that the explanation is more helpful.

**Helpfulness of Joint Error & Attention Maps:** Up to this point, our helpfulness scores rely on human attention maps being available to determine the relevance of an explanation. We also propose a fully automatic helpfulness metric when both attention and error maps are shown together, by computing the relevance (Spearman correlation) between the Error Map and Attention Map  $REL^{EA}$ . As shown in Figure 4c, we hypothesize that a joint error and attention map explanation is helpful if the relevance of the error map to the attention map is higher for cases where the VQA model produces wrong answers than for cases where the VQA model produces correct answers. Intuitively, this means that the explanation is helpful when the error-prone regions point to the attention regions for cases when the VQA produces an incorrect answer, and vice versa. Consequently, we can helpfulness ( $HELP_Z^{EA}$ ):

$$HELP_Z^{EA} = -ztest(REL_c^{EA}, REL_w^{EA}) \quad (4)$$

## 5 | EXPLANATION GENERATION

In this section, we outline how we generate our error maps to improve helpfulness. We also outline a few ways we optimize our attention map visualization to maximize the proxy helpfulness characteristics.

### 5.1 | Error Map Generation

As shown in the right side of Figure 6, we augment our VQA Model by a Justifying Module that learns to predict whether the VQA model will fail or not.

The Justifying Module is conditioned on the input image features  $H_m$ , word2vec<sup>20</sup> question features<sup>4</sup>, AT-BERT weights  $W$ , and the model’s answer logits. We encode the image using a convolution layer and a fully connected layer into a 96-dim vector. The question features, AT-BERT weights and answer logits are each encoded to 96 dimension hidden vectors using two fully connected layers (96 neurons each). We concatenate the four 96-dim vectors to form a latent vector  $h_j$  that describes the context needed to predict the correctness.

We predict the VQA-Net’s performance from the latent embedding  $h_j$ . We use a fully connected layer on  $h_j$  followed by a sigmoid activation to output a binary probability for failure (output is 1 when VQA-System fails). Next, we use GradCAM<sup>12</sup>, which uses gradients to highlight input regions that are important for a failure output. Specifically, we apply it to the logit of the failure predictor output to highlight regions in the image that contribute to the Justifying Module predicting *Failure*. We use the gradient with respect to the image features  $H_m$  to compute the Error Map.

<sup>4</sup>word2vec<sup>20</sup> features are averaged across all question words



## 5.2 | Finding Helpful Attention Maps

To produce a 2-D attention map from AT-BERT layers, we previously described a hand-crafted way to generate attention maps in Section 2. Here, we outline a few ways to optimize the visualization by utilizing our HELPMetrics.

**Most Helpful AT-BERT Attention (BestBERT)** We first extract heat maps from the attention weights  $W$  for each layer and head. For layer  $l$  and head  $h$  the corresponding attention map over the  $7 \times 7 = 49$  image features is  $A^{l,h} = \frac{1}{115} \sum_i W_{l,h,i,:49}$ . We compute the  $HELP_Z^A$  scores of attentions from each AT-BERT layer and head,  $A^{l,h}$  on a held out val split. We choose the  $A^{l,h}$  with the highest  $HELP_Z^A$  (referred to as **BestBERT**). For completeness, we also generate attention by choosing the best head (and average the attention values over the layers for that head), referred to as **BestBERT-H**. Similarly, we also choose the best layer (and average attention over the heads), referred to as **BestBERT-L**.

**Justified Attention Generation (J-Att)** We also learn to predict human annotated attention<sup>19</sup> from the Justifying Module’s embedding,  $h_j$ , by tacking on another head to the module in addition to the failure prediction. This is similar to the work on generating justifying statements<sup>11</sup>, but done for attention maps. The embedding  $h_j$  is fed through a fully connected layer with a 49 dimensional sigmoid output and then reshaped into a  $7 \times 7$  attention map. The Justifying Attention and Performance Prediction heads are trained simultaneously with a joint loss. The first term of the loss minimizes the mean squared error between the predicted J-Att attention map and the corresponding human attention map<sup>19</sup>. The second term is binary cross entropy for failure prediction.

## 6 | EXPERIMENTAL SETTINGS

We first evaluate our new attention maps and error maps on our automated proxy helpfulness metrics. We then conduct user-studies to show improved user performance when using our attention + error map explanations. Using this user study, we also test whether our hypothesized helpfulness metrics for error maps and joint error+attention maps actually correlate to human performance at predicting model correctness. The test split of the Human Attention Dataset<sup>19</sup> is used for both automatic evaluation and user studies. It consists of 3120 IQ pairs. For choosing hyper-parameters and checkpoints, we use a held-out val set of 1000 IQ pairs separate from the 3120 test set.

**Metrics Shown** In the tables for automated metric-based evaluation, we show the following metrics as defined in Section 3.

- **Relevance ✓ / Relevance ×**: The average relevance of the explanation mode for cases when the system is correct (Relevance ✓) or incorrect (Relevance ×). Hence, it is the average value of  $REL^A$  for attention maps,  $REL^{ERR}$  for error maps, and  $REL^{EA}$  for joint attention-error map explanations.
- **HELP<sub>Z</sub>**: Helpfulness of the mode of explanation measured by the z-test approach as described in Section 4.

**Baselines** We compare our new attention maps to the current standard for generating them - AT-BERT as described in Section 2. To create a baseline error map based on AT-BERT, we use the error map version of helpfulness, producing BertErr by choosing the attention  $A^{l,h}$  that maximizes  $HELP_Z^{ERR}$  over all layers and heads.

**User Studies** Users were divided into four groups - no explanations group (None), a group that saw BaselineBERT attention maps (described in Section 3), a group that saw BestBERT Attention Maps + ErrMap Error Maps, and a group that saw J-Att + ErrMap. Workers in each group were kept separate and five independent workers predicted the performance on each IQ pair in each group to ensure the quality of our user studies. In total, we have 3300 ratings for the “None” group, 3000 for BaselineBERT, 2250 for BestBERT + ErrMap, and 3960 for J-Att + ErrMap.

## 7 | RESULTS AND ANALYSIS

We first establish the performance of the proposed explanations according to our automated metrics, showing that our attention maps (e.g., J-Att, BestBERT) and error map (ErrMap) score higher for helpfulness than baselines. We then conduct user studies to show that explanations with higher helpfulness scores indeed improve users’ ability to predict model performance. Below, we describe our key takeaways.

**TABLE 1** The *Relevance* and *Helpfulness* of attention and error maps. Higher *Helpfulness* is better. \* indicates a statistically significant Z-test t-statistic score (Z) for  $HELP_Z$  at 99% confidence level  $p < 0.01$ .

Attention/Error Maps	Relevance		$HELP_Z \uparrow$
	✓	×	
a. BaselineBERT	0.32	0.31	0.4
b. BestBERT-H	0.33	0.32	1.07
c. BestBERT-L	0.41	0.40	0.98
d. J-Att	0.57	0.54	<b>2.56*</b>
e. BestBERT	0.35	0.31	<b>3.87*</b>
f. BERTErr	0.208	0.227	1.52
g. ErrMap	0.209	0.28	<b>4.93*</b>

**TABLE 2** The *Relevance* and *Helpfulness* of explanations when various attentions maps are used in conjunction with ErrMap. Higher *Helpfulness* is better. \* indicates a statistically significant Z-test t-statistic score at 99% confidence level  $p < 0.01$

Attn Choice	Helpfulness	
	$HELP^A$	$HELP^{EA}$
a. BaselineBERT	0.4	<b>5.94</b>
b. BestBERT-H	1.07	<b>4.59</b>
c. BestBERT-L	0.98	<b>7.15</b>
d. BestBERT	2.56	<b>4.99</b>
e. J-Att	3.87	<b>7.71</b>

## 7.1 | Automatic Measurement of Explanations

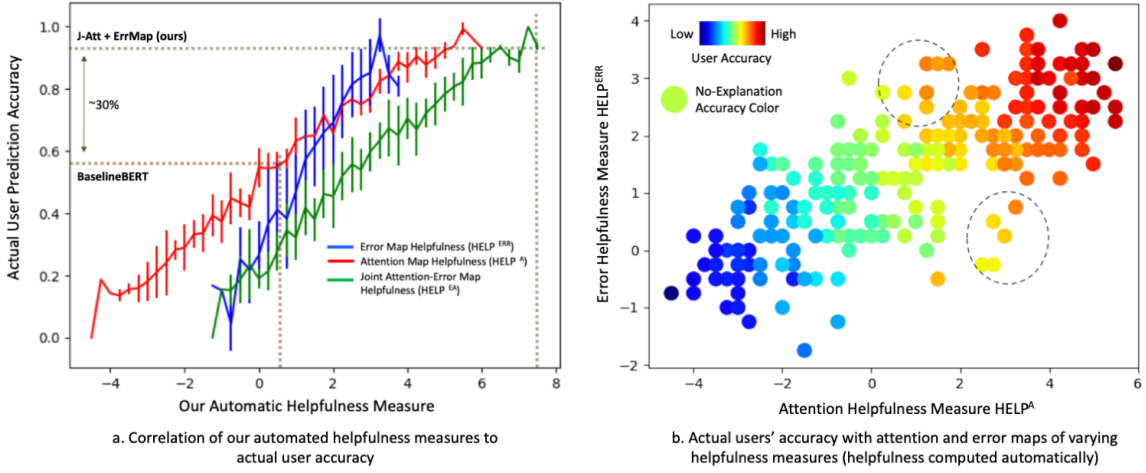
**Our Error Map has high Helpfulness** Helpfulness scores for attention and error maps are listed in Table 1. Our error map, ErrMap (row g), has the highest  $HELP_Z$  score when compared to all attention maps (rows a to e) and baseline error map (row f). This is reflected in the difference in relevance for system correct vs wrong cases, which is the highest in ErrMap.

**Our new attention maps are also more helpful than baselines, with BestBERT being the most helpful.** Our relevance and helpfulness metrics for attention maps are shown in Table 1. The proposed attention maps, J-Att (row d), BestBERT (row e), BestBERTHead (row b), and BestBERTLayer (row c) all show higher *Helpfulness* compared to the baseline (row a). This is reflected in the relevance values, where a larger gap between correct and incorrect relevance leads to higher *Helpfulness*. BestBERT has the highest difference between the correct and wrong cases' relevance, and the highest *Helpfulness*.

**Justifying Attention has high relevance, but relatively low helpfulness.** BestBERT outperforms J-Att, which is supervisedly trained to be human like. This shows that when we train attentions to be human-like, they seem more relevant (points to areas a human would point to for the question), but they may not be helpful for understanding model performance because they are more likely to be relevant even when the model is incorrect.

**Error Maps used in tandem with attention maps improve helpfulness.** In Table 2, we observe that the  $HELP^{EA}$  scores of attention maps with our ErrMap are higher the attention map's  $HELP^A$  scores. This is especially true for high relevance, yet low helpfulness attention maps (column  $HELP^A$  vs  $HELP^{EA}$ , Table 2). For example, BaselineBERT and J-Att which have high relevance even for wrong answers are not very helpful (row a, e, same as Table 1 numbers). When combined with ErrMap the joint helpfulness increases significantly. This is because our ErrMap points to relevant areas when the model errs, and hence, the alignment with an always-relevant attention map can easily indicate failure.





**FIGURE 7 a.** Our automatic metrics for evaluating the *Helpfulness* of attention and error maps correlates to and can reflect users’ ability to predict model performance. **b.** The distribution of user prediction accuracy across varying  $HELP_Z^A$  of attention maps and  $HELP_Z^{ERR}$  of error maps. Having either helpful Attention or Error maps can improve user accuracy as highlighted by the off diagonal dashed circles.

**TABLE 3** User accuracy on our user-study set (randomly sampled subset of val data) for predicting VQA-Net performance. Subset HELP indicates the HELP score on the user study set used.

Expl. Mode	Subset HELP	User Pred Acc		
		AVG	✓	×
a. None	—	57.18	64.02	49.31
b. BaselineBERT	0.57	56.87	63.16	49.56
c. BestBERT+ErrMap	2.54	60.00	59.53	<b>60.42</b>
d. J-Att + ErrMap	<b>4.37</b>	<b>60.13</b>	<b>65.21</b>	55.06

## 7.2 | User-based Evaluation

We first evaluate our proxy helpfulness metrics for error maps and joint error+attention maps and check whether they correlate to actual human accuracy on the task. Similar to how we tested whether our attention helpfulness metric in Figure 5, we plot the average user accuracy for sets of IQ pairs with varying  $HELP_Z^{ERR}$  and  $HELP_Z^{EA}$  values in Fig. 7.

**Our automated *Helpfulness* metrics for error maps and joint error+attention maps correlate with users’ accuracy to predict model performance** As visualized in Figure 7a, we see a strong positive correlation (Pearson Correlation  $> 0.9$ ) with  $HELP_Z$  scores for attention and error maps to the user prediction accuracy when using those explanations. The red line shows the attention only  $HELP_Z^A$ , the blue line shows the  $HELP_Z^{ERR}$ , and the green line shows the  $HELP_Z^{EA}$ .

**Our explanations improve user prediction accuracy** Table 3 shows the users’ overall prediction accuracy (AVG), accuracy for cases with the VQA system producing correct answers (✓), and accuracy for cases with the VQA system producing wrong answers (×). We observe an increase in overall accuracy using ErrMap along with BestBERT (row c) or J-Att (row d) explanations than when using no explanations (row a) or when using Baseline BERT explanations (row b). Note that the HELP score on the set used in the user study is lower than that on overall val data. Hence, we also mark the user accuracy corresponding to the set that has the helpfulness scores closest to our overall test data using dotted lines for baselineBERT (attention-only) and our best performing mode, J-Att + ErrMap (joint attention-error). We note a 30% improvement in user accuracy (Fig. 7a).

**Our explanations help especially in predicting failure** We note an increase of 11% (Table 3 comparing rows a and c in column ×) in user accuracy to predict failure of the VQA system.

**Gameability of the automated metric** We discuss a few contrived explanations that may not reflect the model’s logic for processing the QI pair, but can score high on our metric.

- We generate a centered gaussian attention for all images. We note that this has reasonably high helpfulness ( $HELP_Z^A = 2.35$ ). However, this is indeed a helpful explanation about VQA behavior since a high  $HELP_Z^A$  indicates that VQA models tend to answer correctly when the required objects are in the centered attention region.
- As a strong upperbound, we generate centered attentions when our Justifying Module predicts the system will be correct and randomized attention otherwise. Since the Justifying module is reasonably accurate at predicting VQA performance (70% accurate), attention is relevant for correct answers, and random (low relevance) for incorrect answers. Hence, this attention has high helpfulness ( $HELP_Z^A = 14.51$ ) since it gives a clear visual indicator of VQA correctness (i.e., the ground truth for the user prediction task).

Note that these explanations are gamed because they are not faithful to the internal workings of the model. However, the gamed explanations can indeed provide some helpful insights to users by revealing certain biases.

**Having either helpful Attention or Error Maps is beneficial for predicting VQA performance** In Figure 7b, we plot the user study results with respect to the  $HELP_Z^A$  of attention maps along the x-axis and the  $HELP_Z^{ERR}$  of error maps along the y-axis. The color of the dots indicate the user prediction accuracy (blue is lower and red is higher). We also show the color corresponding to the accuracy for users seeing no explanations for comparison. While it is expected that the user accuracy is higher when both attention and error maps are more helpful, we note that even having either one as helpful can have reasonably high user accuracy as highlighted by the off-diagonal circles.

## 8 | RELATED WORKS

Historically, explainable AI (XAI) spanned from medical systems<sup>21</sup> to using AI for educational purposes<sup>22,23</sup>. There has been recent thrust in developing XAI systems<sup>3,24</sup> for deep networks and a major goal has been to improve users’ mental model of systems for better human-AI collaboration<sup>8,25,11,10,9</sup> or discover biases in models<sup>26</sup>. However, most of these works show no significant improvement of users’ accuracy to predict performance with attention maps or show somewhat deleterious effects<sup>27</sup>. In this paper, we argue that to design helpful explanations, we need to have a modeling of users’ mental model of AI (perception of AI output). We use that to develop helpful attention and error maps to improve users’ ability to predict the model performance.

There have also been works on modeling AI to understand a human’s mental model - their actions, motivations or mental states<sup>28,29,30</sup>, human interactions<sup>31</sup>, or anticipating their beliefs<sup>32</sup>. Systems have also been trained to learn how to behave in society<sup>33</sup> for better human-AI collaboration<sup>34</sup>. While they mostly focus on how to model human behavior or interactions, we focus on modeling how humans interpret system outputs to aid their mental model of the AI system.

We use Visual Question Answering (VQA)<sup>16</sup> as our choice of AI system. Most effective approaches use attention to weigh image and text features<sup>35,7,36,37,38</sup>. Recently, transformer-based<sup>13</sup> attention is used<sup>39,40,41,18</sup>. However, most of these works do not specify how to effectively visualize the high dimensional transformer attentions. We utilize our automated helpfulness metrics to compute effective ways of displaying attention. There are also works on improving saliency-based explanations by making them more faithful<sup>42,43,44</sup>, using model introspection<sup>43</sup>, grounding to evidence<sup>45</sup>, or making them human-like<sup>46</sup>. However, we focus on how to design explanations to not just be relevant, but also be helpful to human users.

To improve helpfulness, we also introduce a error maps for pointing to error-prone regions for VQA. Predicting failures have also been proposed earlier on other domains for better human understanding<sup>47,48,49</sup> and to improve models<sup>50</sup>. This supports our motivation for introducing error maps in tandem with attention explanations to improve users’ ability to predict model performance for VQA.

## 9 | CONCLUSIONS

In this paper, we proposed Error Maps in VQA as a novel explanation method that improve helpfulness over just showing attention maps. We also revealed characteristics of heatmap-style explanations that make them helpful to users and designed an automated proxy helpfulness metric based on it. We anticipate that our automated metrics can aid rapid development of heatmap explanations without having to conduct repeated user-studies, and that our proposed Error Map explanations can be applied to other AI domains to improve users’ mental model of those systems for better human-AI collaboration.

## ACKNOWLEDGEMENTS

We would like to thank Anirban Roy, Briland Hitaj and Karan Sikka for their comments and feedback. This research was developed with funding from DARPA under the XAI program. The views, opinions and/or findings expressed are those of the authors' and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.<sup>5</sup>

## References

1. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. Evidence for a collective intelligence factor in the performance of human groups. *science* 2010; 330(6004): 686–688.
2. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* 2017.
3. Gunning D, Aha DW. DARPA's explainable artificial intelligence program. *AI Magazine* 2019; 40(2): 44–58.
4. Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: ; 2018: 6077–6086.
5. Huk Park D, Anne Hendricks L, Akata Z, et al. Multimodal explanations: Justifying decisions and pointing to the evidence. In: ; 2018: 8779–8788.
6. Yang Z, He X, Gao J, Deng L, Smola A. Stacked attention networks for image question answering. In: ; 2016: 21–29.
7. Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. In: ; 2016: 289–297.
8. Chandrasekaran A, Yadav D, Chattopadhyay P, Prabhu V, Parikh D. It Takes Two to Tango: Towards Theory of AI's Mind. *arXiv preprint arXiv:1704.00717* 2017.
9. Alipour K, Schulze JP, Yao Y, Ziskind A, Burachas G. A Study on Multimodal and Interactive Explanations for Visual Question Answering. *arXiv preprint arXiv:2003.00431* 2020.
10. Ray A, Yao Y, Kumar R, Divakaran A, Burachas G. Can you explain that? Lucid explanations help human-AI collaborative image retrieval. In: . 7. ; 2019: 153–161.
11. Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T. Generating visual explanations. In: Springer. ; 2016: 3–19.
12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ; 2017: 618–626.
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: ; 2017: 5998–6008.
14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: ; 2016: 770–778.
15. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: ; 2017: 2961–2969.
16. Antol S, Agrawal A, Lu J, et al. Vqa: Visual question answering. In: ; 2015: 2425–2433.
17. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. 2017.
18. Alipour K, Ray A, Lin X, Schulze JP, Yao Y, Burachas GT. The Impact of Explanations on AI Competency Prediction in VQA. 2020: 25–32.

---

<sup>5</sup> Authors have no conflict of interest

19. Das A, Agrawal H, Zitnick CL, Parikh D, Batra D. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?. In: ; 2016.
20. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: ; 2013: 3111–3119.
21. Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. *Rule-based expert systems* 1984: 233–262.
22. Lane HC, Core MG, Van Lent M, Solomon S, Gomboc D. Explainable artificial intelligence for training and tutoring. tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES; 2005.
23. Van Lent M, Fisher W, Mancuso M. An explainable artificial intelligence system for small-unit tactical behavior. In: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. ; 2004: 900–907.
24. Adadi A, Berrada M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 2018; 6: 52138–52160.
25. Chandrasekaran A, Prabhu V, Yadav D, Chattopadhyay P, Parikh D. Do explanations make VQA models more predictable to a human?. *arXiv preprint arXiv:1810.12366* 2018.
26. Manjunatha V, Saini N, Davis LS. Explicit bias discovery in visual question answering models. In: ; 2019: 9562–9571.
27. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* 2018.
28. Vondrick C, Pirsiavash H, Torralba A. Anticipating Visual Representations from Unlabeled Video. 2016.
29. Vondrick C, Oktay D, Pirsiavash H, Torralba A. Predicting Motivations of Actions by Leveraging Text. 2016.
30. El Kaliouby R, Robinson P. Real-time inference of complex mental states from facial expressions and head gestures. In: Springer. 2005 (pp. 181–200).
31. Oliver NM, Rosario B, Pentland AP. A Bayesian computer vision system for modeling human interactions. *IEEE transactions on pattern analysis and machine intelligence* 2000; 22(8): 831–843.
32. Eysenbach B, Vondrick C, Torralba A. Who is Mistaken?. 2017.
33. Chuang CY, Li J, Torralba A, Fidler S. Learning to act properly: Predicting and explaining affordances from images. In: ; 2018: 975–983.
34. Puig X, Shu T, Li S, et al. Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration. *arXiv preprint arXiv:2010.09890* 2020.
35. Yu Jiang\*, Vivek Natarajan\*, Xinlei Chen\*, Rohrbach M, Batra D, Parikh D. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *arXiv preprint arXiv:1807.09956* 2018.
36. Teney D, Anderson P, He X, Hengel Avd. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. *arXiv preprint arXiv:1708.02711* 2017.
37. Xu H, Saenko K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Springer. ; 2016: 451–466.
38. Kazemi V, Elqursh A. Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering. *arXiv preprint arXiv:1704.03162* 2017.
39. Lu J, Batra D, Parikh D, Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: ; 2019: 13–23.

40. Sun Q, Fu Y. Stacked Self-Attention Networks for Visual Question Answering. In: ICMR '19. Association for Computing Machinery; 2019; New York, NY, USA: 207–211
41. Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019.
42. Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* 2018.
43. Chang CH, Creager E, Goldenberg A, Duvenaud D. Explaining Image Classifiers by Counterfactual Generation. 2018.
44. Bargal SA, Zunino A, Petsiuk V, et al. Guided zoom: Questioning network evidence for fine-grained classification. *arXiv preprint arXiv:1812.02626* 2018.
45. Hendricks LA, Hu R, Darrell T, Akata Z. Grounding visual explanations. In: Springer. ; 2018: 269–286.
46. Qiao T, Dong J, Xu D. Exploring human-like attention supervision in visual question answering. *arXiv preprint arXiv:1709.06308*.
47. Zhang P, Wang J, Farhadi A, Hebert M, Parikh D. Predicting failures of vision systems. In: ; 2014: 3566–3573.
48. Bansal A, Farhadi A, Parikh D. Towards transparent systems: Semantic characterization of failure modes. In: Springer. ; 2014: 366–381.
49. McGrath S, Mehta P, Zytek A, Lage I, Lakkaraju H. When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. 2020.
50. Patro BN, Lunayach M, Patel S, Namboodiri VP. U-cam: Visual explanation using uncertainty based class activation maps. In: ; 2019: 7444–7453.

## AUTHOR BIOGRAPHY



**Arijit Ray** is a Computer Scientist in the Center for Vision Technologies at SRI International. He has a MS in Computer Engineering from Virginia Polytechnic Institute and State University. His research involves explainable AI in vision/language domains.