Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park[1], Lisa Anne Hendricks[1], Zeynep Akata[2,3], Anna Rohrbach[1,3],
Bernt Schiele[3], Trevor Darrell[1], and Marcus Rohrbach[4]

[1]EECS, UC Berkeley, [2]University of Amsterdam, [3]MPI for Informatics, [4]Facebook AI Research

Abstract

Deep models that are both effective and explainable are desirable in many settings; prior explainable models have been unimodal, offering either image-based visualization of attention weights or text-based generation of post-hoc justifications. We propose a multimodal approach to explanation, and argue that the two modalities provide complementary explanatory strengths. We collect two new datasets to define and evaluate this task, and propose a novel model which can provide joint textual rationale generation and attention visualization. Our datasets define visual and textual justifications of a classification decision for activity recognition tasks (ACT-X) and for visual question answering tasks (VQA-X). We quantitatively show that training with the textual explanations not only yields better textual justification models, but also better localizes the evidence that supports the decision. We also qualitatively show cases where visual explanation is more insightful than textual explanation, and vice versa, supporting our thesis that multimodal explanation models offer significant benefits over unimodal approaches.

1. Introduction

Explaining decisions is an integral part of human communication, understanding, and learning, and humans naturally provide both deictic (pointing) and textual modalities in a typical explanation. We aim to build deep learning models that also are able to explain their decisions with similar fluency in both visual and textual modalities. Previous machine learning methods for explanation were able to provide a text-only explanation conditioned on an image in context of a task, or were able to visualize active intermediate units in a deep network performing a task, but were unable to provide explanatory text grounded in an image.

We propose a new model which can jointly generate visual and textual explanations, using an attention mask to localize salient regions when generating textual rationales. We argue that to train effective models, measure the quality of the generated explanations, compare with other methods, and understand when methods will generalize, it is important to have access to ground truth human explanations. Unfortunately, there is a dearth of datasets which include examples of how humans justify specific decisions. Thus, we collect two new datasets, ACT-X and VQA-X, which allow us to train and evaluate our novel model, which we call the Pointing and Justification Explanation (PJ-X) model. PJ-X is explicitly multimodal: it incorporates an explanatory attention step, which allows our model to both visually point to the evidence and justify a model decision with text.

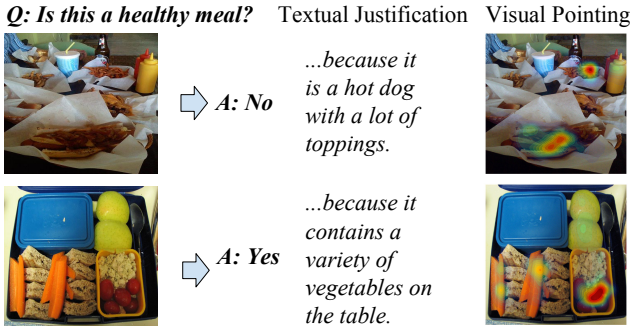

Figure 1: For a given question and an image, our Pointing and Justification Explanation (PJ-X) model predicts the answer and *multimodal* explanations which both point to the visual evidence for a decision and provide textual justifications. We show that considering multimodal explanations results in better explanations as visual and textual components complement each other.

To illustrate the utility of multimodal explanations, consider Figure 1. In both examples, the question "Is this a healthy meal?" is asked, and the PJ-X model correctly answers either "no" or "yes" depending on the visual input. To justify why the image is not healthy, the generated textual justification mentions the kinds of unhealthy food in the image ("hot dog" and "toppings"). In addition to mentioning the unhealthy food, our model is able to *point* to the hot dog in the image. Likewise, to justify why the image on the right is healthy, the textual explanation mentions "vegetables". The PJ-X model then points to the vegetables, which

are mentioned in the textual explanation, but not other items in the image, such as the bread.

We propose VQA and activity recognition as testbeds for studying explanations because they are challenging and important visual tasks which have interesting properties for explanation. VQA is a widely studied multimodal task that requires visual and textual understanding as well as commonsense knowledge. The newly collected VQA v2 dataset [16] includes complementary pairs of questions and answers. Complementary VQA pairs ask the same question of two semantically similar images which have different answers. As the two images are semantically similar, VQA models must employ finegrained reasoning to answer the question correctly. Not only is this an interesting and useful setting for measuring overall VQA performance, but it is also interesting when studying explanations. By comparing explanations from complementary pairs, we can more easily determine whether our explanations focus on the important factors for making a decision.

Additionally, we collect annotations for activity recognition using the MPII Human Pose (MHP) dataset [2]. Activity recognition in still images relies on a variety of cues, such as pose, global context, and the interaction between humans and objects. Though a recognition model can potentially classify an activity correctly, it is not capable of indicating which factors influence the decision process. Furthermore, classifying specific activities requires understanding finegrained differences (e.g., "road biking" and "mountain biking" include similar objects like "bike" and "helmet," but road biking occurs on a road whereas mountain biking occurs on a mountain path). Such finegrained differences are interesting yet difficult to capture when explaining neural network decisions.

In sum, we present VQA-X and ACT-X, two novel datasets of human annotated multimodal explanations for activity recognition and visual question answering. These datasets allow us to train the Pointing and Justification (PJ-X) model which goes beyond current visual explanation systems by producing *multimodal* explanations, justifying the predicted answer post-hoc through visual pointing and textual justification. Our datasets also allow us to effectively evaluate explanation models, and we show that the PJ-X model outperforms strong baselines. Importantly, by generating multimodal explanations, we outperform models which only produce visual or textual explanations.

2. Related Work

Explanations. Early textual explanation models span a variety of applications (e.g., medical [31] and feedback for teaching programs [19, 32, 9]). More recently, [17] developed a deep network to generate natural language justifications of a fine-grained classifier. Unlike our model, it does not provide multimodal explanations and is not trained on reference human explanations as no such dataset existed.

Many works have proposed methods to explain decisions visually. Some methods find discriminative visual patches [7, 11] whereas others aim to understand what specific neurons represent [12, 38, 39]. Perhaps the most prevalent form of visual explanation rely on producing heat maps/attention maps which indicate which region of an image is most important for a decision [13, 29, 37, 41]. Our PJ-X model points to visual evidence via an attention mechanism [4] which conveys knowledge about what evidence is important without requiring domain knowledge to understand.

Explanation systems can either be *introspective* systems, which are designed to reflect the inner workings and decision processes of deep networks, or *justification* systems, which are designed to communicate which visual evidence supports a decision. In this paradigm, models like [17] which highlight discriminative image attributes without attempting to model the classifiers reasoning process are considered justification explanations, whereas models like [37, 12, 39] which aim to illuminate the inner reasoning process of deep networks are considered introspective explanations. We argue that both are useful. Though justifications would not be necessarily helpful for an engineer debugging an AI component, we assert justification is a core AI problem in and of itself: not only is it an AI challenge to answer "is this image a calico cat," but also we claim it is a foundational AI challenge to answer "why would one say this is an image of a calico cat." Though we train justification systems in this work, the data we have collected could be used to understand how well introspective explanations align with our human annotated justifications.

Prior work investigated how well generated visual explanations align with human gaze [10]. However, when answering a question, humans do not always look at image regions which are necessary to explain a decision. For example, given the question "What is the restaurant's name?" human gaze might capture other buildings before settling on the restaurant. When we collect annotations, annotators view the entire image and point to the most relevant visual evidence for making a decision. Furthermore, visual explanations are collected in conjunction with textual explanations to build and evaluate multimodal explanation models.

Visual Question Answering and Attention. Initial approaches to VQA used full-frame representations [22], but most recent approaches use some form of spatial attention [36, 35, 40, 8, 34, 30, 14, 18]. We base our method on [14], the winner of VQA 2016 challenge, but use an element-wise product as opposed to compact bilinear pooling. [18] also explore the element-wise product for VQA, but [18] improves performance by applying hyperbolic tangent (TanH) after the multimodal pooling whereas we improve by applying signed square-root and L2 normalization.

Dataset	Split	#Imgs	#Q/A Pairs	#Unique Q.	#Unique A.	#Expl.	(Avg. #w) Expl.	Vocab Size	#Comple. Pairs	#Visual Ann.
VQA-X	Train	24876	29459	12942	1147	31536	(8.56)	12412	6050	–
	Val	1431	1459	813	246	4377	(8.89)	4325	240	3000
	Test	1921	1968	898	272	5904	(8.94)	4861	510	3000
	Total	28180	32886	13921	1236	41817	(8.64)	14106	6800	6000
ACT-X	Train	12607	–	–	397	37821	(13.96)	12377	–	–
	Val	1802	–	–	295	5406	(13.91)	4802	–	3000
	Test	3621	–	–	379	10863	(13.96)	6856	–	3000
	Total	18030	–	–	397	54090	(13.95)	14588	–	6000

Table 1: Dataset statistics for VQA-X (top) and ACT-X (bottom). Unique Q. = Unique questions, Unique A. = Unique answers, Expl. = Explanations, Avg. #w = Average number of words, Comple. Pairs = Complementary pairs, Visual Ann. = Visual annotations.

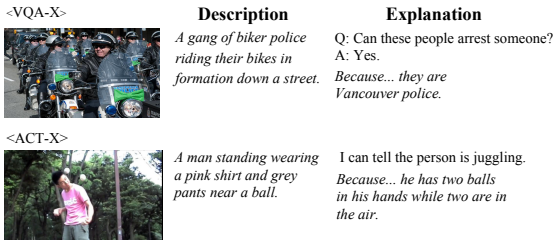

Figure 2: In comparison to descriptions, our VQA-X explanations focus on the evidence that pertains to the *question and answer* instead of generally describing the scene. For ACT-X, our explanations are task specific whereas descriptions are more generic. Images are from [21] and [2].

Activity Recognition. Recent work on activity recognition in still images relies on a variety of cues, such as pose and global context [15, 23, 26]. Specifically, [15] considers additional image regions and [23] considers a global image feature in addition to the region where an activity occurs. Generally, works on the MPII Human Activities dataset provide the ground truth location of a human at test time [15]. In contrast, we consider a more realistic scenario and do not provide the ground truth location of humans at test time. Our model relies on attention to focus on important parts of an image for classification and explanation.

3. Multimodal Explanations

We propose multimodal explanation tasks with visual and textual components, defined on both visual question answering and activity recognition testbeds. To train and evaluate models for this task we collect two multimodal explanation datasets: Visual Question Answering Explanation (VQA-X) and Activity Explanation (ACT-X) (see Table 1 for a summary). For each dataset we collect textual explanations (see Figure 2) and visual explanations (see Figure 3) from human annotators.

VQA Explanation Dataset (VQA-X). The Visual Question Answering (VQA) dataset [3] contains open-ended questions about images which require understanding vision, language, and commonsense knowledge to answer. VQA consists of approximately 200K MSCOCO images [21], with 3 questions per image and 10 answers per question.

Many questions in VQA are of the sort: "What color is the banana?" which is difficult to explain because it requires explaining a fundamental visual property: color. To provide textual explanations for questions that go beyond such trivial cases, we consider the annotations collected in [42] which say how old a human must be to answer a question. We find that questions which require humans to be of age 9 or higher are generally interesting to explain.

Additionally, we consider complementary pairs from the VQA v2 dataset [16]. Complementary pairs consist of a question and two similar images which give two different answers. Complementary pairs are particularly interesting for the explanation task because they allow us to understand whether explanatory models name the correct evidence based on image content, or just memorize which content to consider based off specific question types. We collect one textual explanation for QA pairs in the training set and three textual explanations for test/val set.

Action Explanation Dataset (ACT-X). The MPII Human Pose (MHP) dataset [2] contains 25K images extracted from Youtube videos. From the MHP dataset, we select all images that pertain to 397 activities, resulting in $18,030$ images total. For each image we collect three explanations. During data annotation, we ask the annotators to complete the sentence "I can tell the person is doing (action) because.." where the action is the ground truth activity label. We also ask them to use at least 10 words and avoid mentioning the activity class in the sentence. MHP dataset also comes with sentence descriptions provided by [27].

Ground truth for pointing. In addition to textual justifications, we collect visual explanations from humans for both VQA-X and ACT-X datasets in order to evaluate how well the attention of our model corresponds to where humans think the evidence for the answer is. Human-annotated

(a) Example annotations collected on VQA-X dataset. The visual evidence that justifies the answer is segmented in yellow.

(b) Example annotations collected on ACT-X dataset. The visual evidence that justifies the answer is segmented in yellow.

(c) VQA-HAT vs VQA-X. We aggregate all the annotations in each image and normalize them to create a probability distribution. The distribution is then visualized over the image as a heatmap.

Figure 3: Human annotated visual explanations. Images are from [21] and [2].

visual explanations are collected via Amazon Mechanical Turk where we use the segmentation UI interface from the OpenSurfaces Project [6]. Annotators are provided with an image and an answer (question and answer pair for VQA-X, class label for ACT-X). They are asked to segment objects and/or regions that most prominently justify the answer. Some examples can be seen in Figure 3.

Comparing with VQA-HAT. A thorough comparison between our dataset and VQA-HAT dataset from [10] is currently not viable because the two datasets have different splits and the overlap is small. However, we present qualitative comparison in 3(c). In the first row, our VQA-X annotation has a finer granularity since it segments out the objects in interest more accurately than the VQA-HAT annotation. In the second row, our annotation contains less extraneous information than the VQA-HAT annotation. Since the VQA-HAT annotations are collected by having humans "unblur" the images, they can introduce noise when irrelevant regions are uncovered.

4. Pointing and Justification Model (PJ-X)

We implement a multimodal explanation system that justifies a decision with natural language and points to the evidence. Our Pointing and Justification Model (PJ-X) is explicitly trained for these two tasks and relies on natural language justifications and the classification labels as the only supervision. The PJ-X model learns to point in a latent way using an attention mechanism [4] which allows it to focus on a spatial subset of the visual representation.

We first predict the answer given an image and question using the answering model. Then given the answer, question, and image, we generate visual and textual explanations with the multimodal explanation model. An overview of our model is presented in Figure 4.

Answering model. In visual question answering the goal is to predict an answer given a question and an image. For activity recognition we do not have an explicit question. Thus, we ignore the question which is equivalent to setting the question representation to $f^Q(Q) = 1$, a vector of ones.

We base our answering model on the overall architecture from the MCB model [14], but replace the MCB unit with a simpler element-wise multiplication $\odot$ to pool multimodal features. This leads to similar performance, but trains faster.

In detail, we extract spatial image features $f^I(I, n, m)$ from the last convolutional layer of ResNet-152 followed by 1×1 convolutions ($\bar{f}^I$) giving a $2048 \times N \times M$ spatial image feature. We encode the question Q with a 2-layer $LSTM$, which we refer to as $f^Q(Q)$. We combine this and the spatial image feature using element-wise multiplication followed by signed square-root, L2 normalization, and Dropout, and two more layers of 1×1 convolutions with ReLU in between. This process gives us a $N \times M$ attention map $\bar{\alpha}_{n,m}$. We apply softmax to produce a normalized soft attention map.

The attention map is then used to take the weighted sum over the image features and this representation is once again combined with the LSTM feature to predict the answer $\hat{y}$ as a classification problem over all answers Y. We provide an extended formalized version in the supplemental.

Multimodal explanation model. We argue that to generate multimodal explanation, we should condition the explanation on question, answer, and image. We model this by pooling the image, question, and answer representations

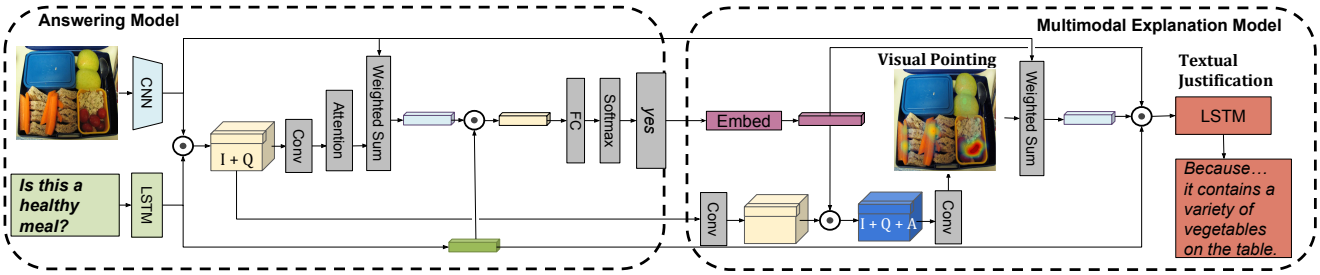

Figure 4: Our Pointing and Justification (PJ-X) architecture generates a multimodal explanation which includes textual justification ("it contains a variety of vegetables on the table") and points to the visual evidence.

to generate an attention map, our *Visual Pointing*. The *Visual Pointing* is further used to create attention features that guide the generation of our *Textual Justification*.

More specifically, the answer predictions are embedded in a d-dimensional space followed by $\tanh$ nonlinearity and a fully connected layer: $f^{yEmbed}(\hat{y}) = W_6(tanh(W_5\hat{y} + b_5)) + b_6$. To allow the model to learn how to attend to relevant spatial location based on the answer, image, and question, we combine this answer feature with Question-Image embedding $\bar{f}^{IQ}(I, Q)$ from the answering model. Applying 1×1 convolutions, element-wise multiplication followed by signed square-root, L2 normalization, and Dropout, results in a multimodal feature.

$$\bar{f}^{IQA}(I, n, m, Q, \hat{y}) = (W_7 \bar{f}^{IQ}(I, Q, n, m) + b_7) \quad (1)$$

$$\odot f^{yEmbed}(\hat{y})) \quad (2)$$

$$f^{IQA}(I, Q, \hat{y}) = L2(signed\_sqrt(\bar{f}^{IQA}(I, Q, \hat{y}))) \quad (3)$$

Next we predict a $N \times M$ attention map $\bar{\alpha}_{n,m}$ and apply softmax to produce a normalized soft attention map, our *Visual Pointing* $\alpha^{pointX}_{n,m}$, which aims to point at the evidence of the generated explanation:

$$\bar{\alpha}_{n,m} = f^{pointX}(I, n, m, Q, \hat{y}) \quad (4)$$

$$= W_9 \rho(W_8 f^{IQA}(I, Q, \hat{y}) + b_8) + b_9 \quad (5)$$

$$\alpha^{pointX}_{n,m} = \frac{\exp(\bar{\alpha}_{n,m})}{\sum_{i=1}^{N}\sum_{j=1}^{M}\exp(\bar{\alpha}_{n,m})} \quad (6)$$

with Relu $\rho(x) = max(x, 0)$.

Using $\alpha^{pointX}_{n,m}$, we compute the attended visual representation, and merge it with the LSTM feature that encodes the question and the embedding feature that encodes the answer:

$$f^X(I, Q, \hat{y}) = (W_{10} \sum_{x=1}^{N}\sum_{y=1}^{M} \alpha^{pointX}_{n,m} f^I(I, n, m) + b_{10}) \quad (7)$$

$$\odot (W_{11} f^Q(Q) + b_{11}) \odot f^{yEmbed}(\hat{y}) \quad (8)$$

This combined feature is then fed into an LSTM decoder to generate our Textual Justifications that are conditioned on image, question, and answer.

Textual Justifications are a sequence of words $[\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots]$ and our model predicts one word w_t at each time step t conditioned on the previous word and the hidden state of the LSTM:

$$h_t = f^{LSTM}(f^X(I, Q, \hat{y}), w_{t-1}, h_{t-1}) \quad (9)$$

$$w_t = f^{pred}(h_t) = Softmax(W_{pred} h_t + b_{pred}) \quad (10)$$

5. Experiments

In this section, we present quantitative results on ablations done for textual justification and visual pointing tasks, and discuss their implications. Additionally, we provide and analyze qualitative results for both tasks.

5.1. Experimental Setup

Here, we detail our experimental setup in terms of model training, hyperparameter settings, and evaluation metrics.

Model training and hyperparameters. For VQA, the answering model of PJ-X is pre-trained on the VQA v2 training set [16]. We then freeze or finetune the weights of the answering model when training the multimodal explanation model on textual annotations as VQA-X is significantly smaller than the original VQA dataset. For activity recognition, answering and explanation components of PJ-X are trained jointly. The spatial feature size of PJ-X is $N = M = 14$. For VQA, the answer space is limited to the 3000 most frequent answers on the training set (i.e. $|Y| = 3000$) whereas for activity recognition, $|Y| = 397$. The answer embedding size is $d = 300$ for both tasks.

Evaluation metrics. We evaluate our textual justifications w.r.t BLEU-4 [24], METEOR [5], ROUGE [20], CIDEr [33] and SPICE [1] metrics, which measure the degree of similarity between generated and ground truth sentences. We also include human evaluation since automatic

Approach	GT-ans Conditioning	Training Data	Att. for Expl.	VQA-X Automatic evaluation B	M	R	C	S	VQA-X Human eval	ACT-X Automatic evaluation B	M	R	C	S	ACT-X Human eval
[17]	Yes	Desc.	No	–	–	–	–	–	–	12.9	15.9	39.0	12.4	12.0	17.4
Ours on Descriptions	Yes	Desc.	Yes	6.1	12.8	26.4	36.2	12.1	34.5	6.9	12.9	28.3	20.3	7.3	22.9
Ours w/o Attention	Yes	Expl.	No	18.0	17.6	42.4	66.3	14.3	40.1	16.9	17.0	42.0	33.3	10.6	21.4
Ours	Yes	Expl.	Yes	**19.8**	**18.6**	**44.0**	**73.4**	**15.4**	**45.1**	**24.5**	**21.5**	**46.9**	**58.7**	**16.0**	**38.2**
Ours on Descriptions	No	Desc.	Yes	5.9	12.6	26.3	35.2	11.9	–	5.2	11.0	26.5	10.4	4.6	–
Ours w/o Attention	No	Expl.	No	18.0	17.3	42.1	63.6	13.8	–	11.9	13.6	37.9	16.9	5.7	–
Ours	No	Expl.	Yes	**19.5**	**18.2**	**43.4**	**71.3**	**15.1**	–	**15.3**	**15.6**	**40.0**	**22.0**	**7.2**	–

Table 2: Evaluation of Textual Justifications: Our proposed model compares favorable to baselines on BLEU-4 (B), METEOR (M), ROUGE (R), CIDEr (C), and SPICE (S) and human eval. Reference sentence is always an explanation. All in %.

metrics do not always reflect human preference. We randomly choose 1000 data points each from the test splits of VQA-X and ACT-X datasets, where the model predicts the correct answer, and then for each data point ask 3 human subjects to judge whether a generated explanation is better than, worse than, or equivalent to the ground truth explanation (we note that human judges do not know what explanation is ground truth and the order of sentences is randomized). We report the percentage of generated explanations which are equivalent to or better than ground truth human explanations, when at least 2 out of 3 human judges agree.

For visual pointing task, we use Earth Mover's Distance (EMD) [28] which measures the distance between two probability distributions over a region. To compute EMD, we use [25]. We also report on Rank Correlation which was used in [10]. For computing Rank Correlation, we follow [10] where we scale the generated attention map and the human ground-truth annotations from the VQA-X/ACT-X/VQA-HAT datasets to 14×14, rank the pixel values, and then compute correlation between these two ranked lists.

5.2. Textual Justification

We ablate PJ-X and compare with related approaches on our VQA-X and ACT-X datasets through automatic and human evaluations for the generated explanations.

Details on compared models. We compare with the state-of-the-art [17] using publicly available code and use ResNet features for fair comparison. The generated sentences from [17] are conditioned on both the image and the class label and uses a discriminative loss. The discriminative loss requires training a sentence classifier and back-propagating policy gradients when training the language generator. Our model does not use discriminative loss/policy gradients and does not require defining a reward. Note that [17] is trained with descriptions. Similarly, "Ours on Descriptions" is an ablation in which we train PJ-X on descriptions instead of explanations. "Ours w/o Attention" is similar to [17] in the sense that there is no attention mechanism involved when generating explanations, however, it does not use the discriminative loss and is trained on explanations instead of descriptions. For all models, explanations can be generated either by conditioning on ground-truth labels or on predicted labels. We call the former "GT-ans Conditioning" and show results in Table 2 to see how it affects the performance.

Descriptions vs. Explanations. "Ours" significantly outperforms "Ours with Descriptions" by a large margin on both datasets which is expected as descriptions are insufficient for the task of generating explanations. Additionally, "Ours" compares favorably to [17] even in the case when "Ours" generates textual justifications conditioned on the prediction, not the ground-truth answer. These results demonstrate the limitation of training explanation systems with descriptions, and thus support the necessity of having datasets specifically curated for explanations. "Ours on Descriptions" performs worse on certain metrics compared to [17] which may be attributed to additional training signals generated from discriminative loss and policy gradients, but further investigation is left for future work.

Unimodal explanations vs. Multimodal explanations. Including attention when generating textual justifications allows us to build a multimodal explanation model. Aside from the immediate benefit of providing visual rationale about a decision, learning to point at visual evidence helps generate better textual justifications. As can be seen in Table 2, "Ours" greatly improves textual justifications compared to "Ours w/o Attention" on both datasets, demonstrating the value of multimodal explanation systems.

5.3. Visual Pointing

We compare the visual pointing performance of PJ-X to several baselines and report quantitative results.

Details on compared models. We compare our model

	Earth Mover's (lower is better)		Rank Correlation (higher is better)		
	VQA-X	ACT-X	VQA-X	ACT-X	VQA-HAT
Random Point	6.71	6.59	+0.0017	+0.0003	-0.0001
Uniform	3.60	3.25	+0.0003	-0.0001	-0.0007
HieCoAtt-Q [10]	–	–	–	–	+0.2640
Answering Model	2.77	4.78	+0.2211	+0.0104	+0.2234
Ours	**2.64**	**2.54**	**+0.3423**	**+0.3933**	**+0.3964**

Table 3: Evaluation of Visual Pointing Justifications. For rank correlation, all results have standard error < 0.005.

against the following baselines. *Random Point* randomly attends to a single point in a 14×14 grid. *Uniform Map* generates attention map that is uniformly distributed over the 14×14 grid. We also compare PJ-X attention maps with those generated from state-of-the-art VQA systems ([10]).

Improved localization with textual explanations. We evaluate attention maps using the Earth Mover's Distance (lower is better) and Rank Correlation (higher is better) on VQA-X and ACT-X in Table 3. From Table 3, we observe that "Ours" outperforms baselines *Random Point* and *Uniform Map*, as well as our answering model and [10] on both datasets and on both metrics. The attention maps generated from our answering model and [10] do not receive training signals from the textual annotations as they are only trained to predict the correct answer, whereas the attention maps generated from PJ-X multimodal explanation model are latently learned through supervision of textual annotations. This implies that learning to generate textual explanations helps improve visual pointing task, and further confirms the advantages of multimodal explanations.

5.4. Qualitative Results

In this section we present our qualitative results on VQA-X and ACT-X datasets demonstrating that our model generates high quality sentences and the attention maps point to relevant locations in the image.

VQA-X. As seen in Figure 5, our textual justifications are able to both capture common sense and discuss specific image parts important for answering a question. For example, when asked "Is this a zoo?", the explanation model is able to discuss what the concept of "zoo" represents (i.e. "animals in an enclosure") and also discuss specific regions (i.e. "green field") to determine whether it is a zoo or not.

Visually, we notice that our attention model is able to point to important visual evidence as well. For example in the top row of Figure 5, the visual explanation focuses on the field in one case, and the fence in another.

ACT-X. Figure 5 also shows results on our ACT-X dataset.

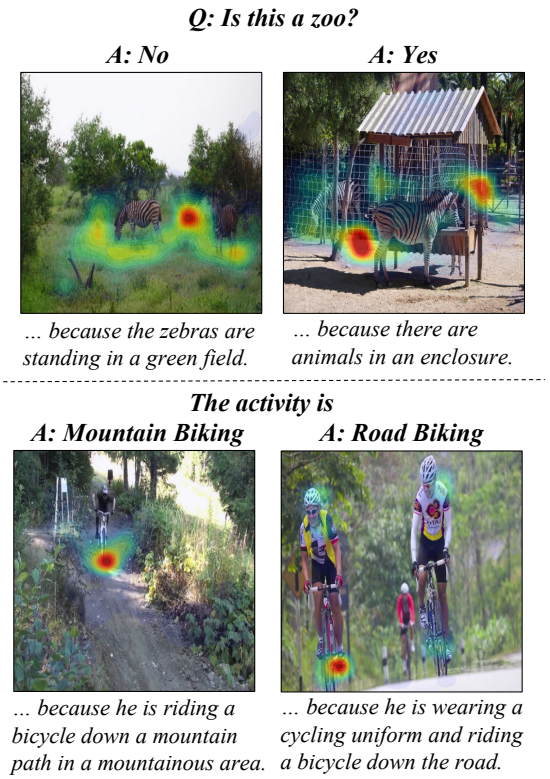

Figure 5: Qualitative results on VQA-X (top row) and ACT-X (bottom row): For each image the PJ-X model provides an answer and a justification, and points to the evidence for that justification. For VQA-X, we show complementary pairs. Images are from [21] and [2].

Textual explanations discuss a variety of visual cues important for correctly classifying activities such as global context (e.g. "a mountainous area"), and person-object interaction, (e.g. "riding a bicycle") for mountain biking. These explanations require determining which of many multiple cues are appropriate to justify a particular action.

Our model points to visual evidence important for understanding each human activity. For example to classify "mountain biking" in the bottom row of Figure 5 the model focuses both on the bicycle as well as the mountainous path. Our model can also differentiate between similar activities based on the context, e.g."mountain biking"/"road biking".

Explanation Consistent with Incorrect Prediction. Generating reasonable explanations for correct answers is important, but it is also crucial to see how a system behaves when predictions are incorrect. Such analysis would provide insights into whether the explanation generation component of the model is consistent with the answer prediction component. In Figure 7, we can see that the explanations are consistent with the incorrectly predicted answer for both VQA-X and ACT-X. For instance in the right example, we

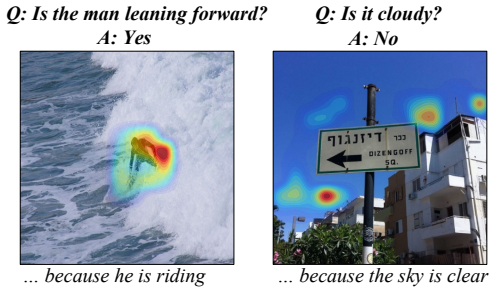

Figure 6: Qualitative results comparing the insightfulness of visual pointing and textual justification. The left example demonstrates how visual pointing is more informative than textual justification whereas the right example shows the opposite. Images are from [21].

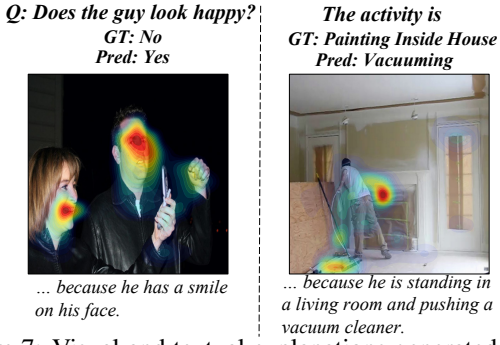

Figure 7: Visual and textual explanations generated by our model conditioned on incorrect predictions. Images are from [21] and [2].

see that the model attends to a vacuum-like object and textually justifies the prediction "vacuuming". Such consistency between the answering model and the explanation model is also shown in Table 2 where we see a drop in performance when explanations are conditioned on predictions (bottom rows) instead of the ground-truth answers (top rows).

5.5. Usefulness of Multimodal Explanations

In this section, we address some of the advantages of generating multimodal explanations. In particular, we look at cases where visual explanations are more informative than textual explanations, and vice versa. We also investigate how multimodal explanations can help humans diagnose the performance of an AI system.

Complementary Explanations. Multimodal explanations can support different tasks or support each other. Interestingly, in Figure 6, we present some examples where visual pointing is more insightful than textual justification, and vice versa. Looking at the left example in Figure 6, it is rather difficult to explain "leaning" with language and the model resorts to generating a correct, yet uninsightful sentence. However, the concept is easily conveyed when looking at the visual pointing result. In contrast, the right example shows the opposite. Looking at only some patches of the sky presented by the visual pointing result does not necessarily confirm if the scene is cloudy or not, while it is also unclear if attending to the entire region of the sky is a desired behavior. Yet, the textual justification succinctly captures the rationale. These examples clearly demonstrate the value of generating multimodal explanations.

	VQA-X	ACT-X
Without explanation	57.5%	51.5%
Ours on Descriptions	66.5%	72.5%
Ours w/o Attention	61.5%	76.5%
Ours	**70.0%**	**80.5%**

Table 4: Accuracy of humans guessing whether the model correctly or incorrectly answered the question.

Diagnostic Explanations. We evaluate an auxiliary task where humans have to guess whether the system correctly or incorrectly answered the question. The predicted answer is not shown; only image, question, correct answer, and textual/visual explanations. The set contains 50% correctly answered questions. We compare our model against the models used for ablations in Table 2. Table 4 indicates that explanations are better than no explanations and our model is more helpful than models trained on descriptions and also models trained to generate textual explanations only.

6. Conclusion

As a step towards explainable AI models, we proposed multimodal explanations for real-world tasks. Our model is the first to be capable of providing natural language justifications of decisions as well as pointing to the evidence in an image. We have collected two novel explanation datasets through crowd sourcing for visual question answering and activity recognition, i.e. VQA-X and ACT-X. We quantitatively demonstrated that learning to point helps achieve high quality textual explanations. We also quantitatively show that using reference textual explanations to train our model helps achieve better visual pointing. Furthermore, we qualitatively demonstrated that our model is able to point to the evidence as well as to give natural sentence justifications, similar to ones humans give. Our model is a third-person, post-hoc rationalization type of explanation, akin to what one human produces when asked to explain the actions of a second human. A third-person explanation is clearly different from a first-person explanation, but we believe both forms of explanation are valuable.

Acknowledgements. This work was partially supported by the DARPA XAI program.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 5

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3, 4, 7, 8

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2, 4

[5] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, volume 29, pages 65–72, 2005. 5

[6] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Opensurfaces: A richly annotated catalog of surface appearance. In *SIGGRAPH*, 2013. 4

[7] T. Berg and P. N. Belhumeur. How do you tell a blackbird from a crow? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[8] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv:1511.05960*, 2015. 2

[9] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg. Building explainable artificial intelligence systems. In *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006. 2

[10] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CoRR*, abs/1606.03556, 2016. 2, 4, 6, 7

[11] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 2

[12] V. Escorcia, J. C. Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[13] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017. 2

[14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2, 4

[15] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015. 3

[16] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 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)*, 2017. 2, 3, 5

[17] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 6

[18] J. Kim, K. W. On, J. Kim, J. Ha, and B. Zhang. Hadamard product for low-rank bilinear pooling. *CoRR*, abs/1610.04325, 2016. 2

[19] H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, and D. Gomboc. Explainable artificial intelligence for training and tutoring. Technical report, DTIC Document, 2005. 2

[20] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004. 5

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3, 4, 7, 8

[22] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[23] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. 5

[25] O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, September 2009. 6

[26] L. Pishchulin, M. Andriluka, and B. Schiele. Fine-grained activity recognition with holistic and pose based features. In *Proceedings of the German Confeence on Pattern Recognition (GCPR)*, pages 678–689. Springer, 2014. 3

[27] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 3

[28] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1998. 6

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3*, 7(8), 2016. 2

[30] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[31] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3):351–379, 1975. 2

[32] M. Van Lent, W. Fisher, and M. Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *NCAI*, 2004. 2

[33] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 5

[34] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2

[35] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[36] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[37] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2

[38] B. Zhou, D. Bau, A. Oliva, and A. Torralba. Interpreting deep visual representations via network dissection. *arXiv preprint arXiv:1711.05611*, 2017. 2

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[40] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[41] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017. 2

[42] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh. Measuring machine intelligence through visual question answering. *CoRR*, abs/1608.08716, 2016. 3