Attentive 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 are the defacto standard in visual decision problems due to their impressive performance on a wide array of visual tasks. On the other hand, their opaqueness has led to a surge of interest in explainable systems. In this work, we emphasize the importance of model explanation in various forms such as visual pointing and textual justification. The lack of data with justification annotations is one of the bottlenecks of generating multimodal explanations. Thus, we propose two large-scale datasets with annotations that visually and textually justify a classification decision for various activities, i.e. ACT-X, and for question answering, i.e. VQA-X. We also introduce a multimodal methodology for generating visual and textual explanations simultaneously. We quantitatively show that training with the textual explanations not only yields better textual justification models, but also models that better localize the evidence that support their decision.

1 Introduction

Explaining decisions is an integral part of human communication, understanding, and learning. Therefore, we aim to build models that explain their decisions, something which comes naturally to humans. Explanations can take many forms. For example, humans can explain their decisions with natural language, or by pointing to visual evidence. We show here that deep models can demonstrate similar competence, and develop a novel multi-modal model which textually justifies decisions and visually grounds evidence simultaneously.

To measure the quality of the generated explanations, compare with different methods, and understand when methods will generalize, it is important to have access to ground truth human annotations. Unfortunately, there is a dearth of datasets which include examples of how humans justify specific decisions. We thus propose and collect explanation datasets for two challenging vision problems: activity recognition and visual question answering (VQA).

We confirm whether the model is actually attending to the discussed items when generating the textual justification (as opposed to just memorizing justification text) by comparing it to our visual pointing annotations. We also determine whether the model attends to the same regions when making a decision as it does when explaining its decision.

2 Multimodal Explanation Datasets

VQA Explanation Dataset (VQA-X). The VQA dataset [2] contains open-ended questions about images which require understanding vision, natural language, and commonsense knowledge to answer. We collected 1 explanation per data point for a subset of the training set and 5 explanations per data point for a subset of the validation and test sets, summing up to 30k justification sentences. The

Interpretable ML Symposium (NIPS 2017), Long Beach, CA, USA.

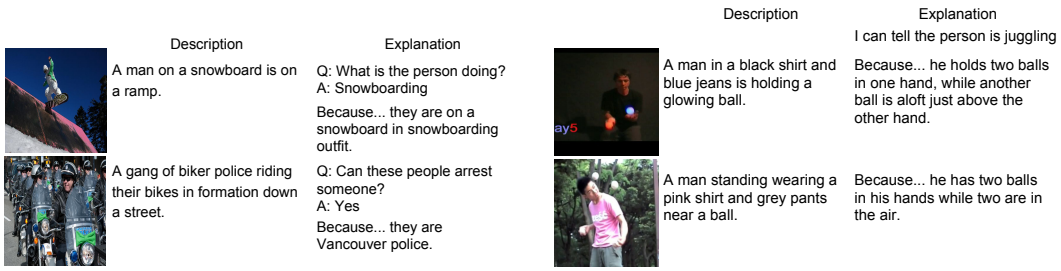
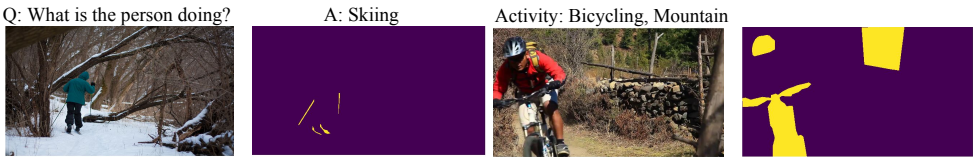


Figure 1: (Left) Our VQA-X explanations focus on the visual evidence that pertains to the question and answer instead of generally describing objects in the scene. (Right) Our ACT-X explanations are task specific whereas image descriptions are more generic.



(a) VQA-X (b) ACT-X

Figure 2: Example human visual explanations collected on: (left) VQA-X dataset, (right) ACT-X dataset. The visual evidence that justifies the answer is segmented in yellow.

annotators were asked to provide a proper sentence or clause that would come after the proposition "because" as explanations to the provided image, question, and answer triplet. Examples for both descriptions, i.e. from MSCOCO dataset, and our explanations are presented in Figure 1.

Activity Explanation Dataset (ACT-X). The MPI Human Pose (MHP) dataset [1] contains images extracted from videos downloaded from Youtube. For each image we collected 3 explanations, totaling 54k sentences. During data annotation, we asked the annotators to complete the sentence "I can tell the person is doing X because..", where X is the ground truth activity label, see Figure 1.

Visual Pointing. In addition to textual justification, we collect visual explanations from humans for both VQA-X and ACT-X datasets. 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 2.

3 Pointing and Justification Model (PJ-X)

The goal of our work is to justify why a decision was made with natural language, and point to the evidence for both the decision and the textual justification provided by the model. We deliberately design our Pointing and Justification Model (PJ-X) to allow training these two tasks as well as the decision process jointly. Specifically we want to rely on natural language justifications and the classification labels as the only supervision. We design a model which learns "to point" in a latent way. For the pointing we rely on an attention mechanism [3] which allows the model to focus on a spatial subset of the visual representation. Our model uses two different attentions: one for making predictions and another for generating textual explanations. We first predict the answer given an image and a question. Then given the answer, question, and image, we generate the textual justification. In both cases we include a latent attention mechanism which allows to introspect where the model is looking. An overview of our double attention model is presented in Figure 3 and the detailed formulation is given in [7].

4 Experiments

Experimental Setup. For visual question answering, our model is pre-trained on the VQA training set [2] to achieve state-of-the-art performance, but we either freeze or finetune the weights of the

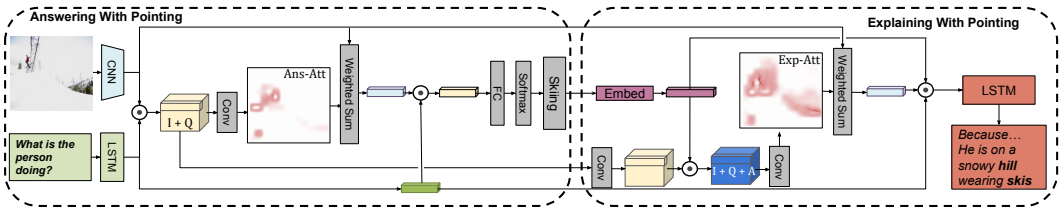


Figure 3: Our Pointing and Justification (PJ-X) architecture generates a multi-modal explanation which includes a textual justification ("He is on a snowy hill wearing skis") and points to the visual evidence. Our model consists of two "pointing" mechanisms: answering with pointing (left) and explaining with pointing (right).

Approach	Training Data	Att. for Expl.	Answer Conditioning	VQA-X Automatic eval. B	M	C	S	VQA-X Human eval.	ACT-X Automatic eval B	M	C	S	ACT-X Human eval
[6]	Desc.	No	Yes	–	–	–	–	–	12.9	15.9	12.4	12.0	7.6
Ours on Descript.	Desc.	Yes	Yes	8.1	14.3	34.3	11.2	24.0	6.9	12.9	20.3	7.3	18.0
Captioning Model	Expl.	Yes	No	17.1	16.0	43.6	7.3	19.2	20.7	18.8	40.7	11.3	20.4
Ours w/o Exp-Att.	Expl.	No	Yes	25.1	20.5	74.2	11.6	34.4	16.9	17.0	33.3	10.6	17.6
Ours	Expl.	Yes	Yes	25.3	20.9	72.1	12.1	33.6	24.5	21.5	58.7	16.0	26.4
Ours (Finetuned)	Expl.	Yes	Yes	27.1	20.9	77.2	11.8	–	–	–	–	–	–

Table 1: Evaluation of Textual Justifications. Evaluated automatic metrics: BLEU-4 (B), METEOR (M), CIDEr (C), SPICE (S). Human evaluation: 250 random images 3 judges rate whether a generated explanation is better than, worse than, or equivalent to a ground truth explanation. We report the % of generated explanations which are equivalent to or better than ground truth human explanations, when at least 2 out of 3 human judges agree.

prediction model when training on explanations as the VQA-X dataset is significantly smaller than the original VQA training set. We refer the finetuned model as 'Finetuned' throughout the paper. For activity recognition, prediction and explanation components of the model are trained jointly.

Textual Justification. We ablate our model and compare with related approaches on our VQA-X and ACT-X datasets based on automatic and human evaluation for the generated explanations.

We re-implemented the state-of-the-art captioning model [5] with an integrated attention mechanism which we refer to as "Captioning Model". This model only uses images and does not use class labels (i.e. the answer in VQA-X and the activity label in ACT-X) when generating textual justifications. We also compare with [6] using publicly available code. Note that [6] is trained with image descriptions and justifications are generated conditioned on both the image and the class predictions. "Ours on Descriptions" is another ablation in which we train the PJ-X model on descriptions instead of explanations. "Ours w/o Exp-Attention" is similar to [6] in the sense that there is no attention mechanism for generating explanations, however, it does not use the discriminative loss and is trained on explanations instead of descriptions.

Our PJ-X model performs well when compared to the state-of-the-art on both automatic evaluation metrics and human evaluations (Table 1). "Ours" model significantly improves "Ours with description" model by a large margin on both datasets. Additionally, our model outperforms [6] which learns to generate explanations given only description training data. These results confirm that our new datasets with ground truth explanations are important for textual justification generation.

Comparing "Ours" to "Captioning Model" shows that conditioning explanations on a model decision is important (human evaluation score increases from 20.4 to 26.4 on ACT-X and 19.2 to 33.6 on VQA-X). Thus it is important for our model to have access to questions and answers to accurately generate the explanation. Finally, including attention allows us to build a multi-modal explanation model. On the ACT-X dataset, it is clear that including attention (compare "Ours w/o Exp-Attention" to "Ours") greatly improves textual justifications. On the VQA-X dataset, "Ours w/o Attention" and

	Earth Mover's distance (lower is better)		Rank Correlation (higher is better)		
	VQA-X	ACT-X	VQA-X	ACT-X	VQA-HAT
Random Point	9.21	9.36	-0.0010	+0.0003	-0.0001
Uniform	5.56	4.81	-0.0002	-0.0007	-0.0007
HieCoAtt-Q [4]	–	–	–	–	0.2640
Ours (ans-att)	4.24	6.44	+0.2280	+0.0387	+0.1366
Ours (exp-att)	4.31	**3.8**	+0.3132	**+0.3744**	+0.3988
Finetuned (ans-att)	**4.24**	–	+0.2290	–	+0.2809
Finetuned (exp-att)	**4.25**	–	**+0.3152**	–	**+0.5041**

Table 2: Evaluation of visual pointing. Ours (ans-att) denotes the attention map used to predict the answer whereas Ours (exp-att) denotes the attention map used to generate explanations.

"Ours" are comparable, however, the latter also produces a multi-modal explanation that offers us an added insight about a model's decision.

Visual Pointing. We compare our generated attention maps to the following baselines and report quantitative results with corresponding analysis. *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 denote the attention map used to predict the answer as *ans-att*, whereas *exp-att* denotes the attention map used to generate 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 datasets in Table 2. We observe that our exp-att outperforms baselines, indicating that exp-att aligns well with human annotated explanations. For ACT-X, our exp-att also outperforms all the baselines. The exp-att significantly outperforms the ans-att, indicating that the regions the model attends to when generating an explanation agree more with regions humans point to when justifying a decision. This suggests that whereas ans-att attention maps can be helpful for understanding a model and debugging, they are not necessarily the best option when providing visual evidence which agrees with human justifications.

5 Conclusion

As a step towards explainable AI models, we introduced two novel explanation datasets collected through crowd sourcing for visual question answering and activity recognition, i.e. VQA-X and ACT-X. We also proposed a multimodal explanation model that is capable of providing natural language justifications of decisions and pointing to the evidence. We quantitatively demonstrated that both attention and supervision from the reference justifications help achieve high quality textual and visual explanations.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[4] 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? In *EMNLP*, 2016.

[5] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *TPAMI*, 2016.

[6] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *ECCV*, 2016.

[7] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv:1612.04757*, 2016.