

# Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park<sup>1</sup>, Lisa Anne Hendricks<sup>1</sup>, Zeynep Akata<sup>2,3</sup>, Anna Rohrbach<sup>1,3</sup>,  
 Bernt Schiele<sup>3</sup>, Trevor Darrell<sup>1</sup>, and Marcus Rohrbach<sup>4</sup>

<sup>1</sup>EECS, UC Berkeley, <sup>2</sup>University of Amsterdam, <sup>3</sup>MPI for Informatics, <sup>4</sup>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

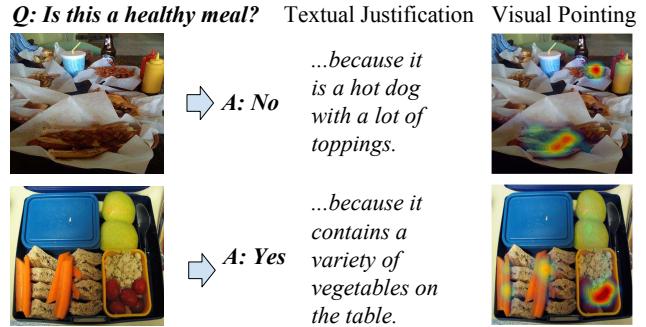


Figure 1: For a given question and an image, our Pointing and Justification Explanation (PJ-X) model predicts the answer and *multimodal* explanations which both point to the visual evidence for a decision and provide textual justifications. We show that considering multimodal explanations results in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

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.

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

## 멀티 모달 설명 : 결정을 정당화하고 증거를 가리키는 것

Dong Huk Park<sup>1</sup>, Lisa Anne Hendricks<sup>1</sup>, Zeynep Akata<sup>2,3</sup>, Anna Rohrbach<sup>1,3</sup>, Bernt Schiele<sup>3</sup>, Trevor Darrell<sup>1</sup> 및 Marcus rohrbach<sup>4</sup> <sup>1</sup> eecs, ucs} <sup>2</sup> 암스테르담 대학교, <sup>3</sup> 정보학을위한 MPI, <sup>4</sup> Facebook AI Research

### 추상적인

*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. 소개

결정을 설명하는 것은 인간 통신, 이해 및 학습의 필수 요소이며, 인간은 전형적인 설명에서 신적(포인팅)과 텍스트 양식을 모두 제공합니다. 우리는 시각적 및 텍스트 양식에서 유사한 유사성으로 결정을 설명 할 수 있는 심층 학습 모델을 구축하는 것을 목표로 합니다. 설명을 위한 이전의 기계 학습 방법은 작업과 관련하여 이미지에 조절 된 텍스트 전용 설명을 제공하거나 작업을 수행하는 깊은 네트워크에서 활성 인터페트 유닛을 시각화 할 수 있었지만 이미지에 기반을 둔 설명 텍스트를 제공 할 수 없었습니다.

우리는 텍스트 이론적 근거를 생성 할 때주의 마스크를 사용하여 주목 마스크를 사용하여 공동 및 텍스트 설명을 공동으로 생성 할 수있는 새로운 모델을 제안합니다. 우리는 효과적인 모델을 훈련시키기 위해 품질을 측정한다고 주장합니다.

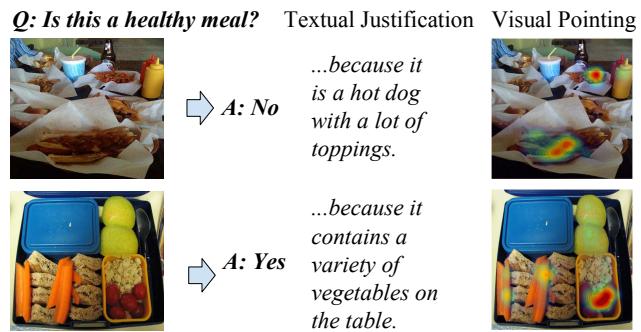


그림 1 : 주어진 질문과 이미지의 경우, 우리의 포인팅 및 정당화 설명 (PJ-X) 모델은 결정에 대한 시각적 증거를 지적하고 텍스트 정당성을 제공하는 answer 및 multimodal 설명을 예측합니다. 우리는 멀티 모달 설명을 고려하면 시각적 및 텍스트 구성 요소가 서로를 보완함에 따라 더 나은 설명이 나타납니다.

생성 된 설명 중, 다른 방법과 비교하고, 방법이 일반화 될 시기를 이해하면, 지상 진실 인간의 설명에 접근 할 수 있는 것은 중요합니다. 다행히도 인간이 특정 결정을 정당화하는 방법을 포함하는 데이터 세트가 부족합니다. 따라서 우리는 ACT-X와 VQA-X라는 두 가지 새로운 데이터 세트를 수집하여 새로운 모델을 교육하고 평가할 수 있습니다. 이 모델은 PJ-X (Pointing and Justification Description) 모델이라고 합니다. PJ-X는 명시적으로 멀티 모달입니다. 설명 적용 단계가 포함되어 있어 모델이 증거를 시각적으로 가리키고 텍스트로 모델 결정을 정당화 할 수 있습니다.

멀티 모달 설명의 유용성을 설명하기 위해, 그림 1을 고려하기 위해, 두 예 모두에서 “이것은 건강한 식사입니까?”라는 질문입니다. PJ-X 모델은 시각적 입력에 따라 “아니오” 또는 “예”를 올바르게 맡기고 있습니다. 이미지가 건강하지 않은 이유를 정당화하기 위해 생성 된 텍스트 정당은 이미지의 건강에 해로운 음식의 종류 (“핫도그” 및 “토핑”)을 언급합니다. 건강에 해로운 음식을 언급하는 것 외에도 우리의 모델은 이미지의 핫도그에 point를 할 수 있습니다. 마찬가지로, 오른쪽의 이미지가 건강한 이유를 정당화하기 위해 텍스트 설명은 “식물성”을 언급합니다. 그런 다음 PJ-X 모델은 채소를 가리킵니다

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 common-sense 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.

텍스트 설명에는 언급되어 있지만 빵과 같은 이미지의 다른 항목은 아닙니다.

우리는 설명을 위한 흥미로운 속성을 가지고 있는 도전적이고 중요한 시각적 작업이기 때문에 설명을 연구하기 위한 테스트 베드로서 VQA 및 활동 인식을 제안합니다. VQA는 일반적인 감각 지식뿐만 아니라 시각적 및 텍스트 이해를 반영하는 널리 연구된 다중 모드 작업입니다. 새로 수집된 VQA V2 데이터 세트 [16]에는 보완적인 질문과 답변이 포함되어 있습니다. 보완 VQA 쌍은 다른 답변을 가진 두 개의 의미적으로 유사한 이미지에 대해 동일한 질문을 합니다. 두 이미지가 의미적으로 유사하기 때문에 VQA 모델은 문제에 대한 답변을 올바르게 답변하기 위해 소포 추론을 사용해야 합니다. 이것은 전반적인 VQA 성능을 측정하기 위한 흥미롭고 유용한 설정 일뿐만 아니라 설명을 연구 할 때도 영향을 미칩니다. 상보적인 쌍의 계획을 비교함으로써, 우리의 설명이 결정을 내리는 데 중요한 요소에 초점을 맞추는 데 쉽게 결정할 수 있습니다.

또한, 우리는 MPII Human Pose (MHP) 데이터 세트를 사용하여 활동 인식을 위한 주석을 수집합니다 [2]. 스틸 이미지에서의 인식은 포즈, 글로벌 맥락 및 인간과 대상 간의 상호 작용과 같은 다양한 신호에 의존합니다. 인식 모델은 활동을 올바르게 분류 할 수 있지만 결정 프로세스에 어떤 요인이 영향을 미치는지를 나타내는 것은 없습니다. 또한, 특정 활동을 분류하려면 지정적 인 차이가 필요합니다 (예 : "도로 자전거 타기" 및 "산악 자전거"는 "자전거" 및 "헬멧"과 같은 비슷한 물건을 포함하지만 도로 자전거는 산길에서 발생하는 반면 도로 자전거는 도로에서 발생합니다). 이러한 결합 된 차이는 신경망 결정을 발표 할 때 흥미롭지 만 캡처하기가 어렵습니다.

요약하면, 우리는 활동 인식 및 시각적 질문 답변을 위한 인간 주석이 달린 멀티 모달 설명의 두 가지 새로운 데이터 세트인 VQA-X와 ACT-X를 제시합니다. 이 데이터 세트를 통해 *multimodal* 설명을 생성하여 현재 시각적 설명 시스템을 넘어서서 시각적 지적 및 텍스트 정당화를 통해 예측된 답변을 정당화하여 현재 시각적 설명 시스템을 넘어서는 PJ-X (Pointing and Justification) 모델을 훈련시킬 수 있습니다. 우리의 데이터 세트를 통해 설명 모델을 효과적으로 평가할 수 있으며, PJ-X 모델이 강력한 기준선을 능가한다는 것을 보여줍니다. 중요하게도, 멀티 모달 설명을 생성함으로써 우리는 시각적 또는 텍스트 설명 만 생성하는 모델보다 성능이 우수합니다.

## 2. 관련 작업

설명. 초기 텍스트 설명 모델은 다양한 응용 분야 (예 : 의료 [31] 및 교육 프로그램 [19, 32, 9])에 걸쳐 있습니다. 보다 최근에, [17]는 자연스러운 분류의 자연 언어 정의를 생성하기 위해 깊은 네트워크를 발표했다. 우리 모델과 달리 그렇습니다

멀티 모달 설명을 제공하지 않으면, 그러한 데이터 세트가 존재하지 않았기 때문에 참조 인간 설명에 대해 훈련 되진 않았습니다. 다른 제품이나 결정을 시각적으로 설명하는 방법을 제안했습니다. 일부 방법은 차별적 시각적 패치 [7, 11]를 찾는 반면, 다른 방법은 특정 뉴런이 무엇을 나타내는지를 이해하는 것을 목표로 한다 [12, 38, 39]. 아마도 가장 널리 퍼진 시각적 설명 형태는 결정에 가장 중요한 이미지의 어느 영역을 나타내는 열 맵/주의 맵을 생성하는 데 의존 할 것입니다 [13, 29, 37, 41]. 우리의 PJ-X 모델은 이해 메커니즘을 통해 시각적 증거를 지적합니다.

설명 시스템은 *introspective* 시스템 일 수 있으며, 이는 심층 네트워크의 내부 작업 및 절제 프로세스를 반영하거나 *justification* 시스템을 반영하도록 설계되었으며, 어떤 시각적 평가가 결정을 지원하는지 전달하도록 설계되었습니다. 이 패러다임에서, 분류기 추론 프로세스를 모델링하려는 시도가 있는 차별적 이미지 속성을 강조하는 [17]와 같은 모델은 정당한 설명에 대한 설명으로 간주되는 반면, 심층 네트워크의 내적 인 추론 프로세스를 조명하는 것을 목표로 하는 [37, 12, 39]와 같은 모델은 내성적 설명으로 간주됩니다. 우리는 둘 다 유용하다고 주장합니다. Justifications는 AI 구성 요소를 디버깅하는 엔지니어가 반드시 도움이 될 필요는 없지만 정당화는 그 자체로 핵심 AI 문제라고 주장합니다.“이 이미지는 Calico Cat”이라고 대답하는 것이 AI 도전 일뿐 만 아니라“이것이 Calico 고양이의 이미지라고 말하는 이유는 무엇입니까?”라고 대답하는 것입니다. 우리는 이 작업에서 정당한 시스템을 훈련 시키지만, 우리가 수집 한 데이터는 인간의 주석이 달린 정당화와 내부적 인 설명이 얼마나 잘 일치하는지 이해하는 데 사용될 수 있습니다.

사전 연구는 시각 계획이 인간의 시선과 얼마나 잘 일치하는지 조사했다 [10]. 그러나 의문을 제기 할 때 인간은 결정을 설명하는 데 필요한 이미지 영역을 항상 보지 않습니다. “식당의 이름은 무엇입니까?”라는 질문이 주어지면 충분합니다. 인간의 시선은 식당에 정착하기 전에 다른 건물을 포착 할 수 있습니다. 주석을 수집 할 때 주석기는 전체 이미지를 보고 결정을 내리는데 가장 관련이 있는 시각적 증거를 가리킵니다. 또한, 멀티 모달 설명 모델을 구축하고 평가하기 위해 텍스트 설명과 함께 시각적 설명이 수집됩니다.

시각적 질문 답변과 관심. VQA에 대한 초기 예측은 전체 프레임 표현을 사용했지만 [22], 가장 최근의 접근법은 어떤 형태의 공간적 테일을 사용한다 [36, 35, 40, 8, 34, 30, 14, 18]. 우리는 VQA 2016 Challenge의 우승자인 [14]를 기반으로 하지만 소형 이중선 풀과는 달리 요소 별 제품을 사용합니다. [18] 또한 VQA의 요소 현저한 제품을 탐색하지만, 멀티 모달 풀링 후 쌍곡선 텐詹 (TANH)을 적용하여 [18]은 서명 된 정사각형 root 및 L2 정규화를 적용함으로써 임을 증명함으로써 성능을 향상시킨다.

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.

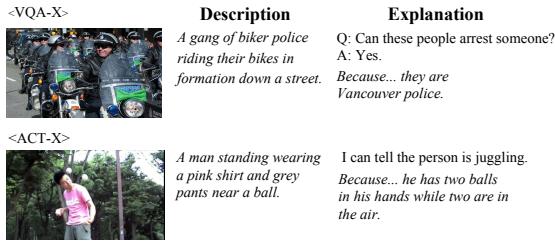


Figure 2: In comparison to descriptions, our VQA-X explanations focus on the evidence that pertains to the *question and answer* instead of generally describing the scene. For ACT-X, our explanations are task specific whereas descriptions are more generic. Images are from [21] and [2].

**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

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

표 1 : VQA-X (상단) 및 ACT-X (하단)에 대한 데이터 세트 통계. 고유 한 Q. = 고유 한 질문, 독특한 A. = 고유 한 답변, expl. = 설명, avg. #w = 평균 단어 수, comple. 쌍 = 보완 쌍, Visual Ann. = 시각적 주석.

<VQA-X>		Description	Explanation
		A gang of biker police riding their bikes in formation down a street.	Q: Can these people arrest someone? A: Yes. Because... they are Vancouver police.
<ACT-X>			
			I can tell the person is juggling. Because... he has two balls in his hands while two are in the air.

그림 2 : 설명과 비교할 때, 우리의 VQA-X 설명은 일반적으로 장면을 설명하는 대신 *question and answer*과 관련된 증거에 중점을 둡니다. ACT-X의 경우, 우리의 설명은 작업 지정이며 설명은 더 일반적입니다. 이미지는 [21]과 [2]에서 나온 것입니다.

활동 인식. 스틸 이미지에서의 활동 인식에 대한 최근의 연구는 포즈 및 글로벌 맥락과 같은 다양한 신호에 의존 한다 [15, 23, 26]. 구체적으로, [15]는 전통적인 이미지 영 역을 고려하고 [23]은 활동이 발생하는 영역 외에도 글로벌 이미지 기능을 고려합니다. 일반적으로, MPII 인간 활동 데이터 세트는 시험 시간에 인간의 지상 진실 위치를 제공합니다 [15]. 대조적으로, 우리는보다 현실적인 시나리오를 고려하고 시험 시간에 인간의 지상 진실 위치를 제공하지 않습니다. 우리의 모델은 분류 및 설명을 위한 이미지의 중요한 부분에 집중하는 데주의를 기울입니다.

### 3. 멀티 모달 설명

시각적 질문 답변 및 활동 인식 테스트 베드 모두에 정의 된 시각적 및 텍스트 구성 요소가 있는 멀티 모달 설명 작업을 제안합니다. 이 작업에 대한 모델을 교육하고 평가하기 위해 우리는 두 가지 멀티 모달 실험 데이터 세트를 수집합니다. 시각적 질문 답변 설명(VQA-X) 및 활동 설명(ACT-X)(요약은 표 1 참조). 각 데이터 세트에 대해 인간 주석기에서 텍스트 설명(그림 2 참조)과 시각적 설명(그림 3 참조)을 수집합니다.

VQA 설명 데이터 세트 (VQA-X). 시각적 질문 응답 (VQA) 데이터 세트 [3]에는 개방형이 포함됩니다.

비전, 언어 및 상식 지식을 이해해야하는 이미지에 대한 질문에 대한 질문. VQA는 약 200k MSCoco 이미지 [21]로 구성되며 이미지 당 3 개의 질문과 질문 당 10 개의 답변으로 구성됩니다.

VQA의 많은 질문은 "바나나는 어떤 색입니까?" 기본적인 시각적 특성을 설명하기 때문에 설명하기가 어렵습니다 : 색상. 그러한 사례를 넘어서는 질문에 대한 텍스트 설명을 제공하기 위해, 우리는 [42]에서 수집 된 주석을 고려하여 인간이 몇 살이 퀴트에 응답해야한다고 생각합니다. 우리는 인간이 9 세 이상이어야하는 질문은 일반적으로 설명하기에 흥미 롭습니다.

또한 VQA V2 데이터 세트의 보완 쌍을 고려합니다 [16]. 보완적인 쌍은 질문과 두 개의 유사한 이미지를 구성되어 있으며 두 개의 다른 답변을 제공합니다. 보완적인 쌍은 설명 모델이 이미지 내용을 기반으로 올바른 증거의 이름을 지정하는지 여부를 이해할 수 있게되거나 특정 질문 유형을 기반으로 고려해야 할 구성 요소를 암기 할 수 있기 때문에 설명 작업에 특히 관심이 있습니다. 우리는 훈련 세트에서 QA 쌍에 대한 하나의 텍스트 설명과 Test/Val 세트에 대한 세 가지 텍스트 설명을 수집합니다.

액션 설명 데이터 세트 (ACT-X). MPII Human Pose (MPHP) 데이터 세트 [2]에는 YouTube 비디오에서 추출 된 25 K 이미지가 포함되어 있습니다. MHP 데이터 세트에서 우리는 397 개의 활동과 관련된 모든 일을 선택하여 총 18,030 Im-Ims 각 이미지에 대해 세 가지 설명을 수집합니다. 데이터 주석 중에, 우리는 주석기에 문장을 완성하도록 요청합니다. 우리는 또한 그들에게 최소 10 단어를 사용하고 문장에서 활동 클래스를 남성하는 것을 피하도록 요청합니다. MHP 데이터 세트에는 [27]가 제공 한 문장 설명도 함께 제공됩니다.

가리키기위한 근거 진실. 본문의 정당성 외에도, 우리는 우리 모델의 관심이 대답에 대한 증거를 어떻게 생각하는지에 얼마나 잘 일치하는지 평가하기 위해 VQA-X 및 ACT-X 데이터 세트 모두에 대한 인간의 시각적 설명을 수집합니다. 인간이 공개되었습니다



(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  $1 \times 1$  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  $1 \times 1$  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



(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.

그림 3 : 인간 주석이 달린 시각적 설명. 이미지는 [21]과 [2]에서 나온 것입니다.

시각적 설명은 OpenSurfaces 프로젝트에서 세분화 UI 인터페이스를 사용하는 Amazon Mechanical Turk를 통해 수집됩니다 [6]. 주석기에는 이미지와 답변이 제공됩니다 (VQA-X의 질문 및 답변 쌍, ACT-X의 클래스 레이블). 그들은 가장 눈에 띄게 정당화되는 분류 및/또는 영역을 분류하도록 요청받습니다. 일부 예는 그림 3에서 볼 수 있습니다.

vqa-hat과 비교. [10]의 데이터 세트와 vqa-hat 데이터 세트 사이의 철저한 비교는 두 개의 데이터 세트가 스플릿이 다르고 오버랩이 있기 때문에 실행 가능하지 않습니다. 그러나 우리는 3 (c)에서 자격 비교를 제시합니다. 첫 번째 행에서, 우리의 VQA-X an- 표기법은 VQA-Hat 주석보다 더 정확하게 관심있는 분야를 세분화하기 때문에 전세계적인 세분성을 가지고 있습니다. 두 번째 행에서, 우리의 주석에는 VQA-Hat 주석이 덜 자발적인 정보를 포함합니다. VQA-HAT 주석은 인간에게 이미지를 "무시"하여 수집되므로 관련이 없는 영역이 발견되면 소음을 소개 할 수 있습니다.

#### 4. 포인팅 및 정당화 모델 (PJ-X)

우리는 자연 언어로 결정을 내리고 evidence를 가리키는 멀티 모달 설명 시스템을 구현합니다. 우리의 포인팅 및 정당화 모델 (PJ-X)은 이 두 작업에 대해 잘 훈련되며 자연스러운 언어 정당화와 분류 라벨을 유일한 감독으로 사용합니다. PJ-X 모델은 시각적 표현의 공간 서브 세트에 집중할 수 있는 주의 메커니즘 [4]를 사용하여 잠재적인 방식으로 지적하는 법을 배웁니다.

응답 모델을 사용하여 이미지와 질문이 주어진 답변을 예측합니다. 그런 다음 답변, 질문 및 이미지가 주어지면 시각적 및 텍스트 설명을 생성합니다.

멀티 모달 설명 모델과 함께. 우리 모델의 개요는 그림 4에 나와 있습니다.

답변 모델. 시각적 질문에서 목표에 답하는 것은 질문과 이미지가 주어진 답을 예측하는 것입니다. 적성 인식을 위해 우리는 명백한 질문이 없습니다. 따라서 우리는 질문 표현을  $f^Q(Q) = 1$ 로 설정하는 것과 동등한 질문을 무시합니다.

우리는 MCB 모델 [14]의 전체 아키텍처를 기반으로 응답 모델을 기반으로 하지만 MCB 장치를 더 간단한 요소별 곱셈  $\odot$ 로 교체하여 멀티 모드 기능을 풀어줍니다. 이것은 비슷한 성능으로 이어지지 만 더 빨리 훈련합니다.

자세히, 우리는 RESNET-152의 마지막 컨볼루션 층에서 공간 이미지 특징  $f^I(I, n, m)$ 과  $1 \times 1$  컨볼루션 ( $f^I$ )을 추출하여  $2048 \times N \times M$  스파이 이미지 기능을 제공합니다. 우리는  $Q$ 를 2 층 LSTM로  $Q$ 로 인코딩합니다.  $f^Q(Q)$ . 우리는 요소별 곱셈과 서명 된 제곱근, L2 정규화 및 드롭 아웃을 사용하여 이것과 공간 이미지 기능을 결합하고,  $1 \times 1$  컨볼루션의 두 층이 더 이상 릴루를 결합합니다. 이 과정은 우리에게  $N \times M$  주의 맵  $\bar{\alpha}_{n,m}$ 을 제공합니다. 우리는 부드러운 소프트주의 맵을 생성하기 위해 SoftMax를 적용합니다.

그런 다음주의 맵은 이미지 기능을 통해 각종 합계를 취하는 데 사용되며 표현은 LSTM 기능과 다시 결합되어 모든 답변  $Y$ 에 대한 분류 문제로 답변  $\hat{y}$ 를 예측합니다. 우리는 보충제에서 확장 공식화 된 버전을 제공합니다.

멀티 모달 설명 모델. 우리는 다중 모드 설명을 생성하기 위해 질문, 답변 및 이미지에 대한 계획을 조정해야한다고 주장합니다. 우리는 이미지, 질문 및 답변 표현을 모아서 이것을 모델링합니다.

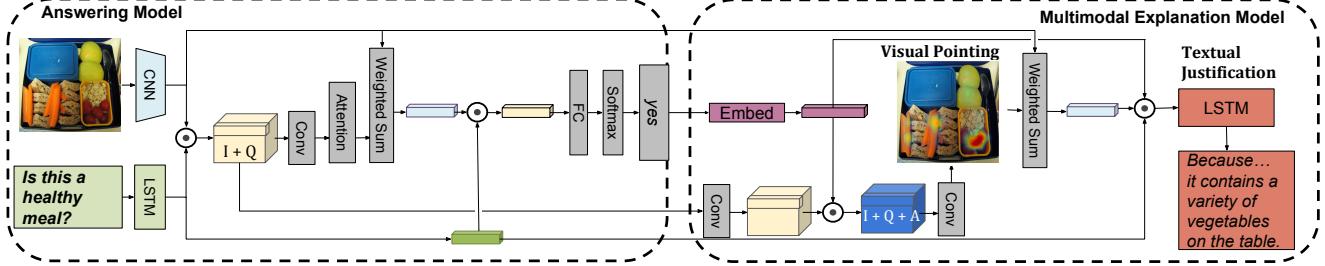


Figure 4: Our Pointing and Justification (PJ-X) architecture generates a multimodal explanation which includes textual justification (“it contains a variety of vegetables on the table”) and points to the visual evidence.

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 non-linearity and a fully connected layer:  $f^{y\text{Embed}}(\hat{y}) = W_6(\tanh(W_5\hat{y} + b_5)) + b_6$ . 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  $1 \times 1$  convolutions, element-wise multiplication followed by signed square-root, L2 normalization, and Dropout, results in a multimodal feature.

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

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

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

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*  $\alpha_{n,m}^{\text{point}X}$ , which aims to point at the evidence of the generated explanation:

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

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

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

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

Using  $\alpha_{n,m}^{\text{point}X}$ , 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:

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

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

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  $[w_1, w_2, \dots]$  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:

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

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

## 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

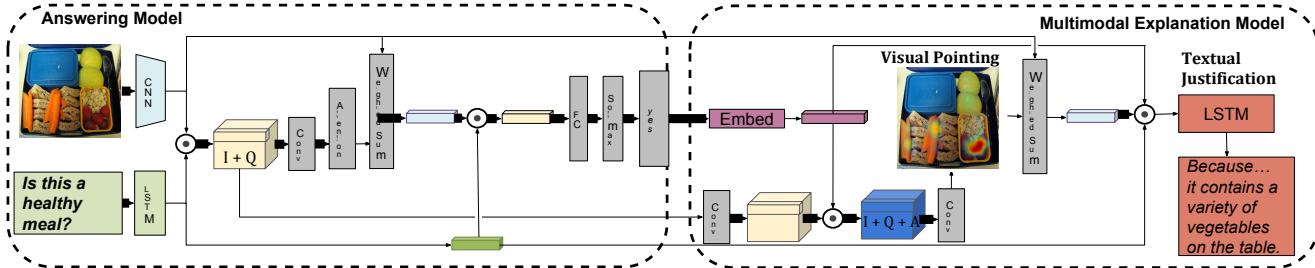


그림 4 : 우리의 포인팅 및 정당화 (PJ-X) 아키텍처는 텍스트 정당화 (“테이블에 다양한 야채를 포함 함”)를 포함하는 멀티 모달 설명을 생성하고 시각적 증거를 가리킵니다.

주의지도를 생성하려면 *Visual Pointing*. *Vi-sual Pointing*은 *Textual Justification*의 생성을 안내하는 주의 기능을 만드는 데 더 많이 사용됩니다.

보다 구체적으로, 답변 예측은  $d$ -차원 공간에 포함된 다음 TANH 비 선형성 및 완전히 연결된 층에 포함되어 있습니다.  $f^{y\text{Embed}}(\hat{y}) = W_6(\tanh(W_5\hat{y} + b_5)) + b_6$ . 모델이 An-Swer, Image 및 질문을 기반으로 관련 공간 위치에 참석하는 방법을 배울 수 있도록 답변 기능을 AN-Swering 모델의 질문 이미지 임베딩  $\bar{f}^{IQ}(I, Q)$ 과 결합합니다.  $1 \times 1$  컨볼루션 적용, 요소 별 곱셈 다음으로 서명된 제곱근, L2 정상-지화 및 드롭 아웃으로 인해 멀티 모달 특징이 발생합니다.

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

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

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

다음으로  $N \times M$  주의 맵  $\bar{\alpha}_{n,m}$ 을 예측하고 SoftMax를 적용하여 정규화된 소프트 주의 맵인 *Vi-sual Pointing*  $\alpha_{n,m}^{\text{point}X}$ 을 생성합니다.

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

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

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

$$\text{Relu } \rho(x) = \max(x, 0).$$

$\alpha_{n,m}^{\text{point}X}$ 을 사용하여 참석한 시각적 반복을 계산하고 이를 질문을 인코딩하는 LSTM 기능 및 An-Swer를 인코딩하는 임베딩 기능과 병합합니다.

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

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

그린 다음이 결합 된 기능은 LSTM 디코더로 공급되어 이미지, 질문 및 답변에 조절되는 텍스트 정당화를 생성합니다.

*Textual Justifications* 단어 시퀀스  $[w_1, w_2, \dots]$ 이며, 우리의 모델은 이전 단어와 lstm의 숨겨진 상태에 조절된 각 시간 단계  $t$ 에서  $w_t$ 을 예측합니다.

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

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

## 5. 실험

이 섹션에서는 텍스트 정당화 및 시각적 포인팅 작업을 위해 수행된 절제에 대한 정량적 결과를 제시하고 그 의미를 논의합니다. 또한 두 작업 모두에 대한 질적 결과를 제공하고 분석합니다.

### 5.1. 실험 설정

여기서는 모델 훈련, 하이퍼 파라미터 설정 및 평가 지표 측면에서 실험 설정을 자세히 설명합니다.

모델 교육 및 하이퍼 파라미터. VQA의 경우, PJ-X의 스와이어 모델은 VQA V2 트레이닝 세트에서 미리 훈련된다 [16]. 그런 다음 vqa-x가 원래 VQA 데이터 세트보다 현저히 작기 때문에 테스트 주석에 대한 멀티 모달 설명 모델을 훈련시킬 때 응답 모델의 가중치를 동결하거나 미치십시오. 활동 인식을 위해 PJ-X의 답변 및 설명 구성 요소는 공동으로 교육을 받습니다. PJ-X의 공간 특징 크기는  $N = M = 14$ 입니다. VQA의 경우, 답변 공간은 훈련 세트에서 3000 개의 가장 빈번한 답변으로 제한되는 반면 (예:  $|Y| = 3000$ ) 활동 인식의 경우  $|Y| = 397$ .  $|Y| = 39$ .

평가 지표. 우리는 생성과 지상 진실의 유사성을 측정하는 텍스트 정당성 W.R.T Bleu-4 [24], Meteor [5], Rouge [20], Cider [33] 및 Spice [1] 메트릭을 평가합니다. 우리는 또한 자동으로 인간 평가를 포함합니다

Approach	GT-ans Conditioning	Train- ing Data	Att. Expl.	VQA-X						ACT-X					
				B	M	R	C	S	eval	B	M	R	C	S	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	<b>19.8</b>	<b>18.6</b>	<b>44.0</b>	<b>73.4</b>	<b>15.4</b>	<b>45.1</b>	<b>24.5</b>	<b>21.5</b>	<b>46.9</b>	<b>58.7</b>	<b>16.0</b>	<b>38.2</b>
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	<b>19.5</b>	<b>18.2</b>	<b>43.4</b>	<b>71.3</b>	<b>15.1</b>	–	<b>15.3</b>	<b>15.6</b>	<b>40.0</b>	<b>22.0</b>	<b>7.2</b>	–

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  $14 \times 14$ , 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

Approach	GT-ans	Train-	Att.	VQA-X						ACT-X					
	Condi-	ing	for	B	M	R	C	S	eval	B	M	R	C	S	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	<b>19.8</b>	<b>18.6</b>	<b>44.0</b>	<b>73.4</b>	<b>15.4</b>	<b>45.1</b>	<b>24.5</b>	<b>21.5</b>	<b>46.9</b>	<b>58.7</b>	<b>16.0</b>	<b>38.2</b>
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	<b>19.5</b>	<b>18.2</b>	<b>43.4</b>	<b>71.3</b>	<b>15.1</b>	—	<b>15.3</b>	<b>15.6</b>	<b>40.0</b>	<b>22.0</b>	<b>7.2</b>	—

표 2 : 텍스트 정당화의 평가 : 제안 된 모델은 BLEU-4 (b), 유성 (M), 루지 (R), 사이다 (C) 및 향신료 (S) 및 인간 평가의 기준선과 비교됩니다. 참조 문장은 항상 설명입니다. 모두 %.

메트릭이 항상 인간의 선호도를 반영하는 것은 아닙니다. 우리는 VQA-X 및 ACT-X 데이터 세트의 테스트 분할에서 각각 1000 개의 데이터 포인트를 선택하는데, 여기서 모델이 정답을 예측한 다음 각 데이터 포인트에 대해 3 인간의 대상에게 생성된 설명이 더 나은 것보다 더 나은지, 또는 그 이상의 설명과 동등한지 (우리는 인간의 판단이 무엇을 알지 못하는지 알지 못한다는 점에 주목합니다.) 우리는 3 명의 인간 판사 중 적어도 2 명이 동의 할 때 근거 진실 인간의 설명과 동등하거나 더 나은 생성된 설명의 비율을 보고합니다.

시각적 포인팅 작업을 위해, 우리는 지역에 대한 두 가지 확률 분포 사이의 거리를 측정하는 Earth Mover's Distance (EMD) [28]를 사용합니다. EMD를 계산하기 위해 우리는 [25]를 사용합니다. 우리는 또한 [10]에 사용된 순위 상관 관계에 대해서도 보고합니다. 순위 상관 관계를 계산하기 위해, 우리는 VQA-X/ACT-X/VQA-HAT 데이터 세트에서  $14 \times 14$ 로 생성된 주의 맵과 인간지면 진실 주석을 확장 한 다음, 이 두 순위가 매겨진 목록 사이의 상관 관계를 계산하는 경우 [10].

## 5.2. 텍스트 정당화

우리는 PJ-X를 제거하고 생성된 설명에 대한 자동 및 후원 평가를 통해 VQA-X 및 ACT-X 데이터 세트의 관련 접근법과 비교합니다.

비교 모델에 대한 세부 사항. 우리는 공개적으로 이용 가능한 코드를 사용하여 ART 상태와 비교하고 공정한 비교를 위해 RESNET 기능을 사용합니다. [17]의 생성된 문장은 이미지와 클래스 레이블 모두에 조절되며 차별적 손실을 사용합니다. 차별적 손실은 언어 생성기를 훈련 할 때 정책 구배를 분류하고 역행하는 문장을 훈련시킵니다. 우리의 모델은 차별적 손실/정책 그라디언트를 사용하지 않으며 보상을 정의 할 필요가 없습니다. [17]은 설명으로 훈련되었습니다. 마찬가지로, “우리의 설명에 대한 우리의 설명”

설명. “우리의주의가 없는 우리의주의”는 설명을 생성 할 때 관련된 주의 메커니즘이 없다는 점에서 [17]와 유사하지만, 범죄 적 손실을 사용하지 않으며 정책 대신 설명에 대한 교육을 받습니다. 모든 모델의 경우 지상 진실 레이블 또는 예측된 레이블에 컨디셔닝하여 설명을 생성 할 수 있습니다. 우리는 이전의 “GT-ANS 컨디셔닝”을 호출하고 표 2의 결과를 보여 주어 성능에 어떤 영향을 미치는지 확인합니다.

설명 대 설명. “우리의”는 설명을 생성하는 과정에 대한 설명이 필요하지 않기 때문에 두 데이터 세트에서 큰 마진으로 “설명과 함께 우리의 설명”을 크게 수행합니다. 또한, “우리의”는 “우리의”가 근거 진실 대답이 아니라 예측에 조절된 텍스트 정당성을 생성하는 경우에도 [17]와 호의적으로 비교됩니다. 이 결과는 설명이 있는 교육 설명 시스템의 한계를 보여 주므로 설명을 위해 특정으로 선별된 데이터 세트의 필요성을 지원합니다. “우리의 설명에 대한 우리의 설명”은 [17]로 구성된 특정 메트릭에서 더 나빠지는데, 이는 차별적 손실 및 정책 구배에서 생성된 추가 교육 신호에 기인 할 수 있지만 향후 작업을 위한 추가 조사가 남아 있습니다.

단단한 설명 대 다중 모드 설명. 텍스트 정당화를 생성 할 때주의를 포함하여 멀티 모달 설명 모델을 구축 할 수 있습니다. 결정에 대한 시각적 근거를 제공하는 즉각적인 이점 외에도 시각적 증거를 지적하는 법을 배우면 더 나은 텍스트 정당성을 생성하는 데 도움이 됩니다. TA-BLE 2에서 볼 수 있듯이, “우리의”는 두 데이터 세트에서 “주의가 없는 우리의주의”와 관련된 텍스트 정당화를 크게 향상시켜 멀티 모달 설명 시스템의 가치를 보여줍니다.

## 5.3. 시각적 포인팅

우리는 PJ-X의 시각적 포인팅 성능을 여러 기준선과 비교하고 정량적 결과를 보고합니다.

비교 모델에 대한 세부 사항. 우리는 모델을 비교합니다

	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	<b>2.64</b>	<b>2.54</b>	<b>+0.3423</b>	<b>+0.3933</b>	<b>+0.3964</b>

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  $14 \times 14$  grid. *Uniform Map* generates attention map that is uniformly distributed over the  $14 \times 14$  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.

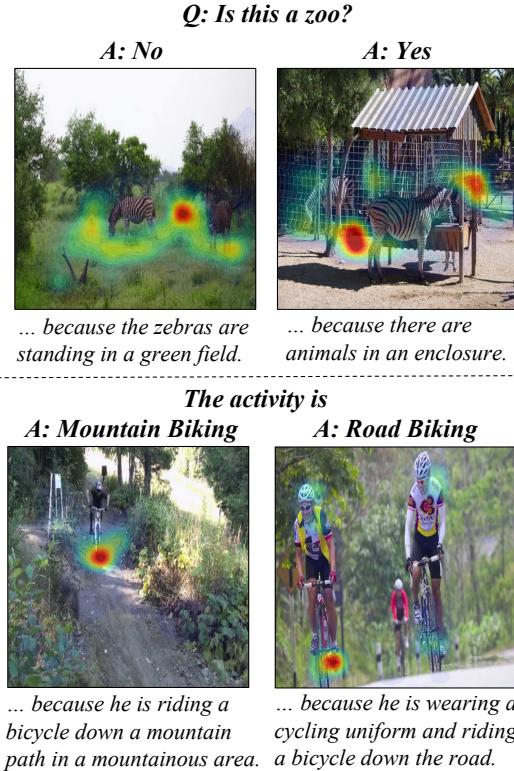


Figure 5: Qualitative results on VQA-X (top row) and ACT-X (bottom row): For each image the PJ-X model provides an answer and a justification, and points to the evidence for that justification. For VQA-X, we show complementary pairs. Images are from [21] and [2].

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

	Earth Mover's		Rank Correlation		
	(lower is better)	(higher is better)	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	<b>2.64</b>	<b>2.54</b>	<b>+0.3423</b>	<b>+0.3933</b>	<b>+0.3964</b>

표 3 : 시각적 포인팅 정당화의 평가. 순위 상관 관계의 경우 모든 결과에는 표준 오류 < 0.005가 있습니다.

다음 기준에 대해. Random Point  $14 \times 14$  그리드의 단일 지점에 무작위로 참석합니다. Uniform Map  $14 \times 14$  그리드에 균일하게 분포된 주의 맵을 생성합니다. 또한 PJ-X주의 맵을 최첨단 VQA 시스템에서 생성한 것과 비교합니다 ([10]).

텍스트 설명으로 현지화 향상. 표 3의 vqa-x 및 act-x에서 지구 발동기의 거리(더 낮음)와 순위 상관 관계(더 높음)를 사용하여 주의 맵을 평가합니다. 표 3에서, 우리는 “우리”가 기준선 Uni- 및 Uni-form Map뿐만 아니라 우리의 응답 모델과 [10]가 angishing and metets를 능가한다는 것을 관찰합니다. 응답 모델에서 생성된 주의지도는 정답을 예측하기 위해 훈련된 반면, PJ-X 다중 모드 설명 모델에서 생성된 주의지도는 텍스트 주석의 감독을 통해 학습됩니다. 이는 텍스트 설명을 생성하는 법을 학습하는 것이 시각적 포인팅 작업을 개선하는 데 도움이 되며, 다중 모드 설명의 장점을 추가로 구성한다는 것을 의미합니다.

#### 5.4. 질적 결과

이 섹션에서는 VQA-X 및 ACT-X 데이터 세트에 대한 정성적 결과를 제시하여 모델이 고품질 문장을 생성하고 주의지도가 이미지의 관련 위치를 가리킵니다.

VQA-X. 그림 5에서 볼 수 있듯이, 본문의 정당화는 상식을 포착하고 질문에 대답하는데 중요한 특정 부분에 대해 논의 할 수 있습니다. 예를 들어, “이것은 동물원입니다.”라고 물었을 때, 설명 모델은 “동물원”的 개념이 무엇을 나타내는지 논의 할 수 있으며(즉, “인클로저의 “아미드”)가 또한 동물원인지 여부를 결정하기 위해 특정 지역(즉, “녹색 필드”)에 대해 논의 할 수 있습니다.

시각적으로, 우리의 주의 모델은 중요한 시각적 증거도 지적 할 수 있음을 알 수 있습니다. 예를 들어 그림 5의 상단 행에서 시각적 설명은 한 경우에 필드에 초점을 맞추고 다른 경우에는 울타리에 중점을 둡니다.

act-x. 그림 5는 또한 ACT-X 데이터 세트의 결과를 보여줍니다.

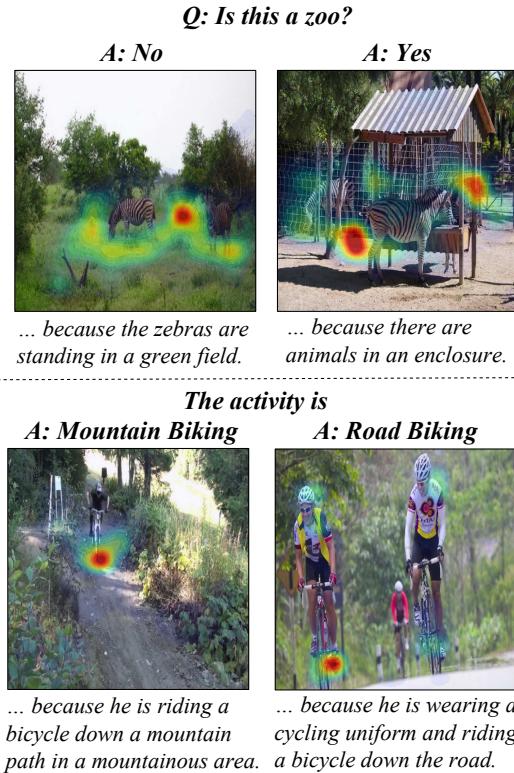


그림 5 : VQA-X (상단 행) 및 ACT-X (하단 행)의 질적 결과 : 각 이미지에 대해 PJ-X 모델은 답변과 정당화를 제공하며 해당 정당의 증거를 가리킵니다. VQA-X의 경우 상보적인 쌍을 보여줍니다. 이미지는 [21]과 [2]에서 나온 것입니다.

텍스트 설명은 글로벌 텍스트(예: “산악 지역”) 및 산악 자전거를 위한 사람·객체 상호 작용(예: “자전거를 타기”)과 같은 활동을 올바르게 분류하기 위한 다양한 시각적 신호에 대해 논의합니다. 이러한 설명은 특정 행동을 정당화하기 위해 많은 여러 큐 중 어느 것이 적합한지 결정해야 합니다.

우리의 모델은 각 인간 활동을 이해하지 못하는 데 중요한 시각적 증거를 지적합니다. 예를 들어 그림 5의 맨 아래 줄에서 “산악 자전거”를 분류하기 위해 모델은 자전거와 산악 경로에 중점을 둡니다. 우리의 모델은 또한 상황에 따라 유사한 활동을 구별 할 수 있습니다(예: “Mountain Biking”/“Road Biking”).

잘못된 예측과 일치하는 설명. 정답에 대한 합리적인 설명을 작성하는 것은 중요하지만 예측이 잘못되었을 때 시스템이 어떻게 행동하는지 보는 것이 중요합니다. 이러한 분석은 모델의 설명 생성 회사가 답변 예측 구성 요소와 일치하는지 여부에 대한 통찰력을 제공합니다. 그림 7에서, 우리는 설명이 VQA-X와 ACT-X 모두에 대한 잘못된 예측 답변과 일치한다는 것을 알 수 있습니다. 예를 들어 올바른 예에서는 우리가

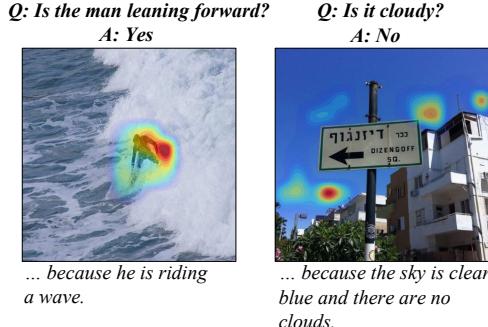


Figure 6: Qualitative results comparing the insightfulness of visual pointing and textual justification. The left example demonstrates how visual pointing is more informative than textual justification whereas the right example shows the opposite. Images are from [21].

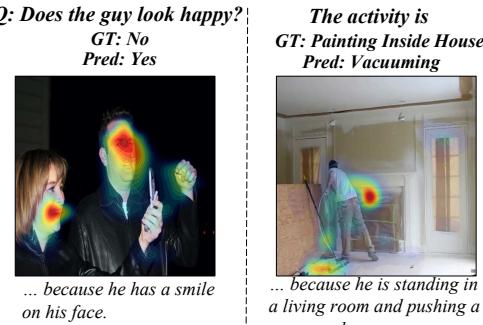


Figure 7: Visual and textual explanations generated by our model conditioned on incorrect predictions. Images are from [21] and [2].

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 sen-

	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	<b>70.0%</b>	<b>80.5%</b>

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

tence. 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.

**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.

*Q: Is the man leaning forward?*

*A: Yes*



*... because he is riding a wave.*

*Q: Is it cloudy?*

*A: No*

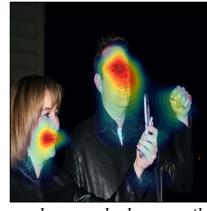


*... because the sky is clear blue and there are no clouds.*

그림 6 : 시각적 지적 및 텍스트 정당화의 통찰력을 비교하는 질적 결과. 왼쪽 시험은 시각적 포인팅이 텍스트 정당화보다 유익한 방법을 보여줍니다. 올바른 예는 그 반대를 보여줍니다. 이미지는 [21]에서 나온 것입니다.

*Q: Does the guy look happy?*

*GT: No  
Pred: Yes*



*... because he has a smile on his face.*

*The activity is  
Painting Inside House*

*Pred: Vacuuming*



*... because he is standing in a living room and pushing a vacuum cleaner.*

그림 7 : 잘못된 예측에 조절 된 모델에 의해 생성 된 시각적 및 텍스트 설명. 이미지는 [21]과 [2]에서 나온 것입니다.

이 모델은 진공과 같은 물체에 참석하고 예측은 "진공 청소기"를 정당화합니다. 응답 모델과 설명 모델 사이의 이러한 일관성은 표 2에도 표시되어 있으며, 여기서 설명이 지면 진실 답변(상단 행) 대신 예측(하단 행)에 조절될 때 성능이 떨어집니다.

### 5.5. 멀티 모달 설명의 유용성

이 섹션에서는 멀티 모달 설명을 생성하는 데 있어 몇 가지 장점을 다룹니다. 특히, 우리는 시각적 설명이 텍스트 설명보다 유익한 경우를 살펴보고 그 반대도 마찬가지입니다. 우리는 또한 멀티 모달 설명이 인간이 AI 시스템의 성능을 향상시키는 데 어떻게 도움이 될 수 있는지에 대한 정보를 제공합니다.

보완적인 설명. 멀티 모달 설명은 서로 다른 작업을 지원하거나 서로를 지원할 수 있습니다. 흥미롭게도, 그림 6에서, 우리는 시각적 포인팅이 텍스트 정당화보다 더 통찰력이 있고 그 반대도 마찬가지입니다. 그림 6의 왼쪽 예를 살펴보면 언어로 "기울기"를 설명하는 것은 다소 어렵습니다.

	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	<b>70.0%</b>	<b>80.5%</b>

표 4 : 모델이 질문에 올바르게 대답했는지 여부를 추측하는 인간의 정확성.

텐스. 그러나 시각적 포인팅 결과를 볼 때 개념이 쉽게 전달됩니다. 대조적으로, 오른쪽 외부는 반대를 보여줍니다. 시각적 포인팅 결과에 의해 제시된 하늘의 일부만을 보면 장면이 흐리거나 그렇지 않으면 반드시 하늘의 전체 지역에 참석하는 것이 바람직한 행동인지는 확실하지 않습니다. 그러나 텍스트 정당화는 간결하게 이론적 근거를 포착합니다. 이 예는 멀티 모달 설명을 생성하는 가치를 명확하게 보여줍니다.

진단 설명. 우리는 인간이 시스템이 질문에 올바르게 대답했는지 여부를 추측 해야하는 보조 작업을 평가합니다. 예측 된 답변은 표시되지 않습니다. 이미지, 질문, 정답 및 문자/시각적 설명 만. 이 세트에는 50%가 올바르게 문제가 있습니다. 우리는 표 2의 절제에 사용 된 모드와 우리의 모델을 비교합니다. 표 4는 설명이 설명이 없는 것보다 낫고 우리의 모델은 설명에 대해 훈련 된 모델과 텍스트 설명만으로 훈련 된 모델보다 더 도움이 된다는 것을 나타냅니다.

## 6. 결론

설명 가능한 AI 모델을 향한 단계로서, 우리는 실제 작업에 대한 멀티 모달 설명을 제안했습니다. 우리의 모델은 최초의 결정에 대한 자연 언어 정당성을 제공하고 이미지의 증거를 지적 할 수 있는 것입니다. 우리는 시각적 질문 답변 및 활동 인식을 위해 군중 소싱을 통해 두 가지 새로운 설명 데이터 세트, 즉 VQA-X 및 ACT-X를 수집했습니다. 우리는 포인트를 배우는 것이 고품질 텍스트 설명을 달성하는 데 도움이 된다는 것을 정량적으로 입증했습니다. 또한 모델을 훈련시키기 위해 참조 텍스트 설명을 사용하면 더 나은 시각적 포인팅을 달성하는 데 도움이 된다는 것을 정량적으로 보여줍니다. 더욱이, 우리는 우리의 모델이 인간과 비슷한 자연 문장 정당성을 제공 할뿐만 아니라 자연스러운 문장을 제공 할 수 있음을 보여주었습니다. 우리의 모델은 두 번째 인간의 행동을 설명하도록 요청할 때 한 사람이 생산하는 것과 유사한 3 인칭, 사후 합리화 유형의 설명입니다. 3 인칭의 설명은 제 1 인분의 설명과는 분명히 다르지만, 우리는 두 가지 형태의 설명이 가치가 있다고 생각합니다.

감사의 말. 이 작업은 DARPA XAI 프로그램에 의해 부문적으로 지원되었습니다.

## 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. Open-surfaces: 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

## 참조

- [1] P. Anderson, B. Fernando, M. Johnson 및 S. Gould. 향신료 : 시맨틱 제안 이미지 캡션 평가. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 5 [2] M. Andriluka, L. Pishchulin, P. Gehler 및 B. Schiele. 2D 인간 포즈 추정 : 새로운 벤치 마크 및 최신 기술 분석. *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 및 D. Parikh. VQA : 시각적 질문 답변. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3 [4] D. Bahdanau, K. Cho 및 Y. Bengio. 공동으로 조정하고 늦게 까지 학습함으로써 신경 마인트 번역. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2, 4 [5] S. Banerjee 및 A. Lavie. Meteor : 인간 판단과의 상관 관계가 향상된 MT 평가를 위한 자동 대표. *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 및 K. Bala. 오픈 표면 : 풍부하게 주석이 달린 표면 모양의 카탈로그. *SIGGRAPH*, 2013. 4 [7] T. Berg 및 P. N. Belhumeur. 까마귀의 흑인 새에게 어떻게 말합니까? *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2 [8] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu 및 R. Nevatia. ABC-CNN : 시각적 질문 답변을 위한 주의 기반의 컨벤션 신경 네트워크. *arXiv:1511.05960*, 2015. 2 [9] M. G. Core, H.C. Lane, M. Van Lent, D. Gomboc, S. Solomon 및 M. Rosenberg. 설명 가능한 인공 지능 시스템 구축. *Proceedings of the national conference on artificial intelligence*에서. 케리포니아 주 멘로 파크; 케임브리지, MA; 런던; AAAI Press; MIT Press; 1999, 2006. 2
- [10] Das, H. Agrawal, C. L. Zitnick, D. Parikh 및 D. Batra. 시각적 질문에 대한 인간의 관심 : 인간과 깊은 네트워크는 같은 지역을 보입니다? *CoRR*, abs/1606.03556, 2016. 2, 4, 6, 7
- [11] Doersch, S. Singh, A. Gupta, J. Sivic 및 A. Efros. 파리가 파리처럼 보이게 하는 이유는 무엇입니까? *ACM Transactions on Graphics*, 31 (4), 2012. 2
- [12] Escorcia, J. C. Niebles 및 B. Ghanem. 시각적 속성과 컨볼루션 사이의 재건에 네트워크. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [13] R. C. Fong 및 A. Vedaldi. 의미 있는 섭동에 의한 블랙 박스의 해석 가능한 설명. *arXiv preprint arXiv:1704.03296*, 2017. 2
- [14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell 및 M. Rohrbach. 시각적 질문 응답 및 시각적 접지를 위한 멀티 모달 소형 빌린 이어 풀링. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2, 4
- [15] G. Gkioxari, R. Girshick 및 J. Malik. r\* cnn을 통한 상황 행동 인식. *Proceedings of the IEEE international conference on computer vision*, 페이지 1080–1088, 2015. 3 [16] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra 및 D. Parikh. VQA 문제에서 V를 만들기 : 시각적 질문 답변에서 이미지 이해의 역할 향상. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 5 [17] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele 및 T. Darrell. 시각적 설명 생성. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 6
- [18] J. Kim, K. W. On, J. Kim, J. Ha 및 B. Zhang. 저급 이중선 풀링을 위한 Hadamard 제품. *CoRR*, abs/1610.04325, 2016. 2
- [19] H.C. Lane, M. G. Core, M. Van Lent, S. Solomon 및 D. Gomboc. 훈련 및 지도를 위한 설명 가능한 인공 지능 기술 보고서, DTIC DUCUMINT, 2005. 2
- [20] C.-Y. Lin. Rouge : 요약 자동 평가 패키지. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004. 5
- [21] T.-Y. Lin, M. Maire, S. Picture, J. Hays, P. Per-Ona, D. Ramanan, P. Dollár 및 C. L. Zitnick. Mi-Crosoft Coco : 맥락에서 일반적인 객체. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3, 4, 7, 8
- [22] M. Malinowski, M. Rohrbach 및 M. Fritz. 뉴런에게 물어보십시오 : 이미지에 대한 질문에 대한 신경 기반 접근 방식. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2
- [23] A. Mallya와 S. Lazebnik. 질문 답변으로의 전송과의 행동 및 개인 객체 상호 작용에 대한 학습 모델. *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 Conference 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/1610.02391v3>, 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

- [24] K. Papineni, S. Roukos, T. Ward 및 W.-J. Zhu. BLEU : 기계 번역의 자동 평가 방법. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318, 2002. 5
- [25] O. Pele 및 M. Werman. 빠르고 강력한 지구 이동 거리. *2009 IEEE 12th International Conference on Computer Vision*, 페이지 460–467. IEEE, 2009년 9월. 6
- [26] L. Pishchulin, M. Andriluka 및 B. Schiele. 전체 론적 및 포즈 기반 기능으로 미세한 활동 인식. *Proceedings of the German Conference on Pattern Recognition (GCPR)*, 678–689 페이지. Springer, 2014. 3 [27] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele 및 H. Lee. 무엇을 그리고 어디에서 그리는지 배우기. *Advances in Neural Information Processing Systems (NIPS)*, 2016. 3
- [28] Y. Rubner, C. Tomasi 및 L. J. Guibas. 이미지 데이터베이스에 응용 프로그램이 있는 분포 측정 항목. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1998. 6
- [29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh 및 D. Batra. 졸업식 : 그라디언트 기반 지역화를 통해 깊은 네트워크의 시각적 설명. *See <https://arxiv.org/abs/1610.02391> v3*, 7 (8), 2016. 2
- [30] K. J. Shih, S. Singh 및 D. Hoiem. 볼 수 있는 곳 : 시각적 질문에 대한 초점 지역. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [31] E. H. Shortliffe와 B. G. Buchanan. 의학에서의 경험적 추론 모델. *Mathematical biosciences*, 23 (3) : 351–379, 1975. 2
- [32] M. Van Lent, W. Fisher 및 M. Mancuso. 소규모 단위 전술 행동을 위한 예외적인 인공 지능 시스템. *NCAI*, 2004. 2
- R. Vedantam, C. Lawrence Zitnick 및 D. Parikh. 사이다 : 컨센서스 기반 이미지 설명 평가. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 페이지 4566–4575, 2015. 5
- C. Xiong, S. Merity 및 R. Socher. 시각적 및 텍스트 질문 답변을 위한 동적 딥 네트워크. *Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2
- [35] H. Xu와 K. Saenko. 묻고 참석하고 답변하십시오 : 시각적 질문에 대한 질문 유도 공간주의를 탐구하십시오. *. Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2
- [36] Z. Yang, X. He, J. Gao, L. Deng 및 A. Smola. 이미지 질문에 대한 쌍인주의 네트워크. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [37] M. D. Zeiler와 R. Fergus. 컨볼루션 네트워크를 시각화하고 이해하지 못합니다. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2
- [38] B. Zhou, D. Bau, A. Oliva 및 A. Torralba. 네트워크 해부를 통해 깊은 시각적 표현을 해석합니다. *arXiv preprint arXiv:1711.05611*, 2017. 2 [39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva 및 A. Torralba. 물체 감지기는 깊은 장면 CNN에서 나타납니다. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2 [40] Y. Zhu, O. Groth, M. Bernstein 및 L. Fei-Fei. Vi-sual7w : 이미지에서 대답하는 질문. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2 [41] L. M. Zintgraf, T. S. Cohen, T. Adel 및 M. Welling. 깊은 신경망 결정 시각화 : 예측 차이 분석. *arXiv preprint arXiv:1702.04595*, 2017. 2 [42] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra 및 D. Parikh. 시각적 질문 응답을 통해 기계 양도 측정. *CoRR*, abs/1608.08716, 2016. 3