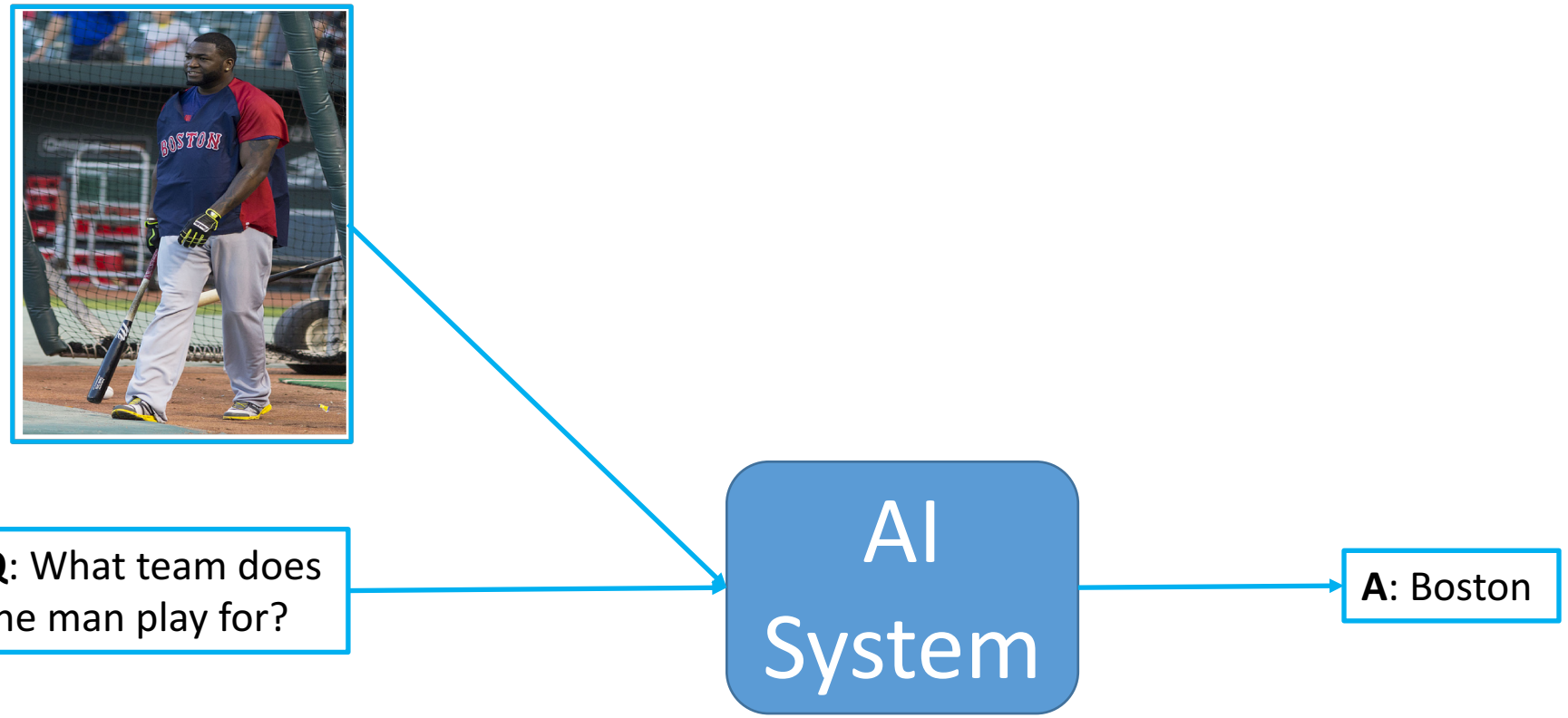
VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, Jiebo Luo

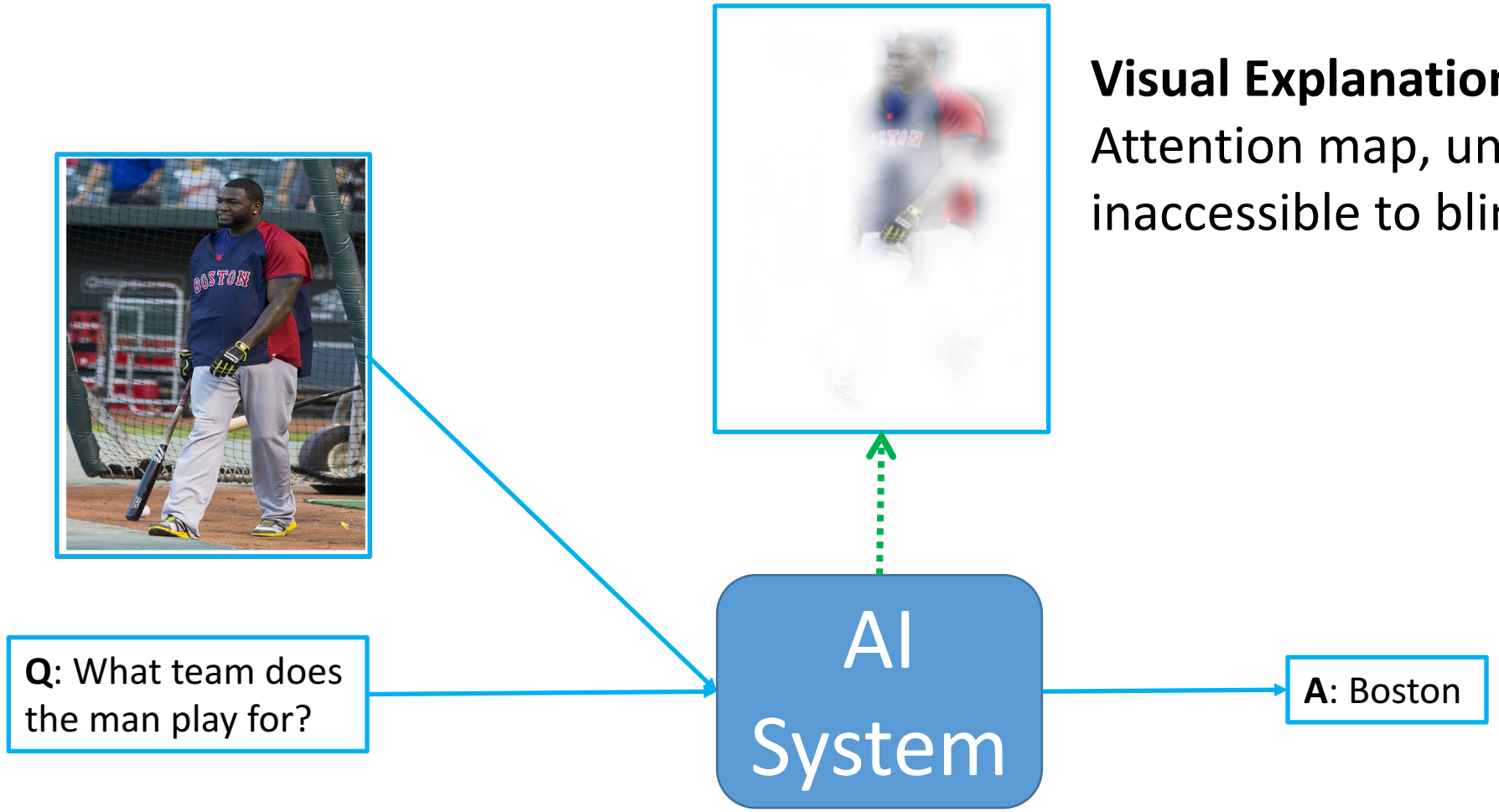
Outline

- Introduction
- VQA-E Dataset
  - Explanation Synthesis
  - Dataset Analysis
  - User Study
- Multi-task Model
- Experiments
- Conclusions

Visual Question Answering



Visual Question Answering with Explanation



Visual Question Answering with Explanation

Visual Explanation
Attention map, unsupervised, inaccessible to blind people.

Q: What team does the man play for?

AI System

Textual Explanation

A: Boston
***E**: A man wearing a Boston jersey.*

The Benefits of Textual Explanations

- Be accessible to visually impaired people.
- Provide beneficial feedbacks to extend the conversation.

Q:What is the hotel's name?
A:Butternut.

From where can we tell the name of the hotel?

Explaining....

E:A view of a red brick building which has a sign that says 'BUTTERNUT' on the side.

Q:Is the elephant in the boat?
A:No.

It is not in the boat. Then where is it?

Elaborating....

E:An elephant is walking down the hill near a boat in the water.

Q:Is the dog wearing anything?
A:Yes..

'Anything' is too vague to tell what the dog is wearing.

Enhancing....

E:A dog in a madonna shirt is sitting next to feet in high heels.

Current VQA Datasets

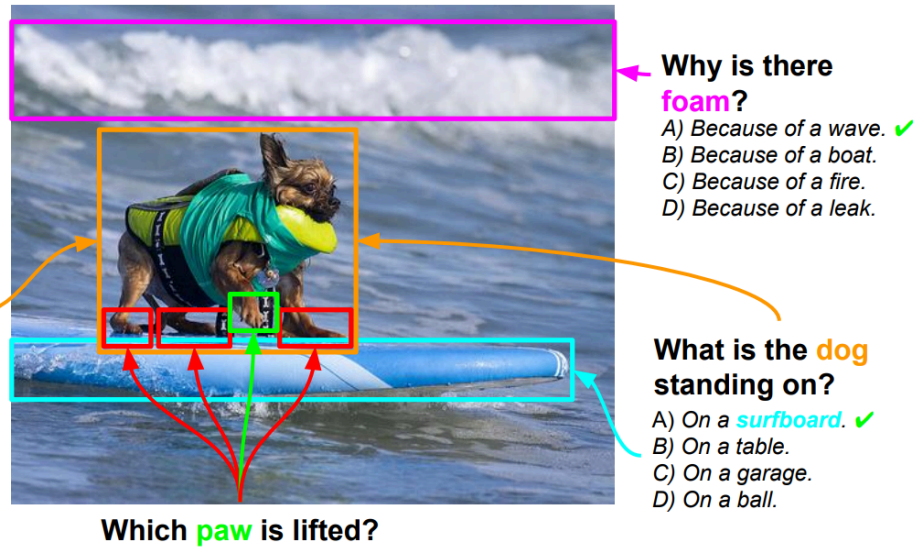
Short answers and
No explanations!

Toronto COCO-QA (NIPS'15)

COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three

COCOQA 1238
What is the color of the tee-shirt?
Ground truth: blue

Visual7W (CVPR'16)



VQA (ICCV'15, CVPR'17)

Proposed VQA-E Dataset

- > 108K images (from COCO), ~270K <Q, A, E> pairs
- Explanations are *synthesized*, paired with confidence scores.

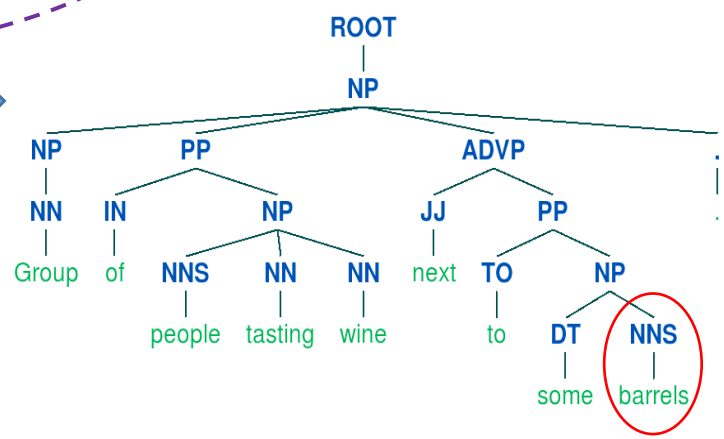
Dataset	Split	#Images	#Q&A	#E	#Unique Q	#Unique A	#Unique E
VQA-E	Train	72,680	181,298	181,298	77,418	9,491	115,560
	Val	35,645	88,488	88,488	42,055	6,247	56,916
	Total	108,325	269,786	269,786	108,872	12,450	171,659
VQA-v2	Train	82,783	443,757	0	151,693	22,531	0
	Val	40,504	214,354	0	81,436	14,008	0
	Total	123,287	658,111	0	215,076	29,332	0

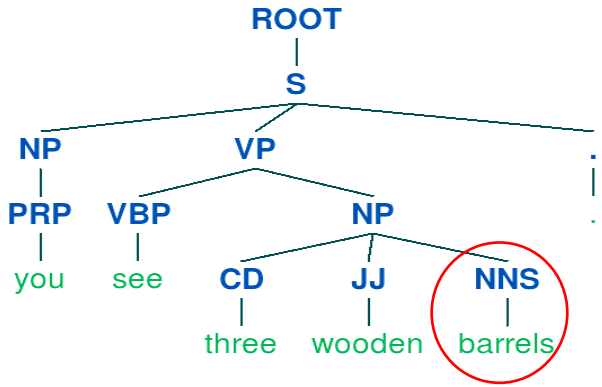
Explanation Synthesis

C: Group of people tasting wine next to some barrels.

Q: How many wooden barrels do you see?
A: 3

S: You see three wooden barrels.





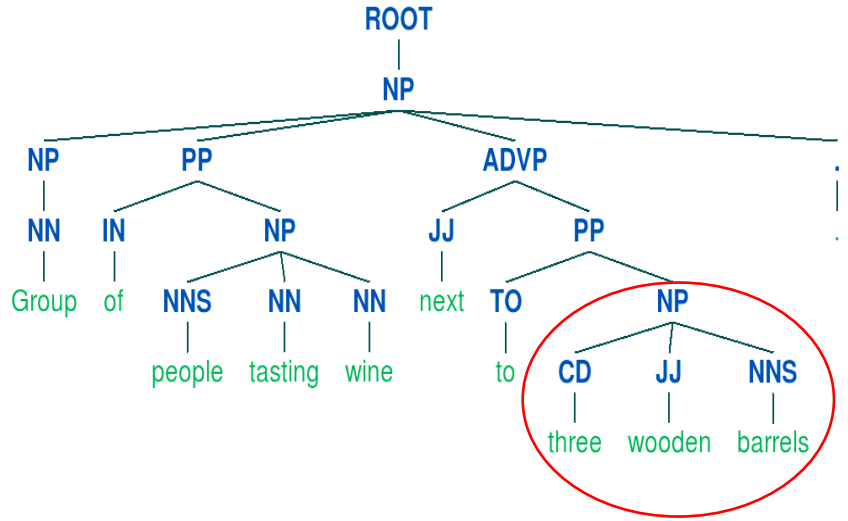
❖ Find the most relevant caption:

$$s(\boldsymbol{w}_i, \boldsymbol{w}_j) = \frac{1}{2}(1 + \frac{\boldsymbol{w}_i^T \boldsymbol{w}_j}{||\boldsymbol{w}_i|| \cdot ||\boldsymbol{w}_j||})$$

$$S(\mathcal{Q}, \mathcal{C}) = \frac{1}{T_q} \sum_{\boldsymbol{w}_i \in W_q} \max_{\boldsymbol{w}_j \in W_c} s(\boldsymbol{w}_i, \boldsymbol{w}_j)$$

$$S(\mathcal{A}, \mathcal{C}) = \frac{1}{T_a} \sum_{\boldsymbol{w}_i \in W_a} \max_{\boldsymbol{w}_j \in W_c} s(\boldsymbol{w}_i, \boldsymbol{w}_j)$$

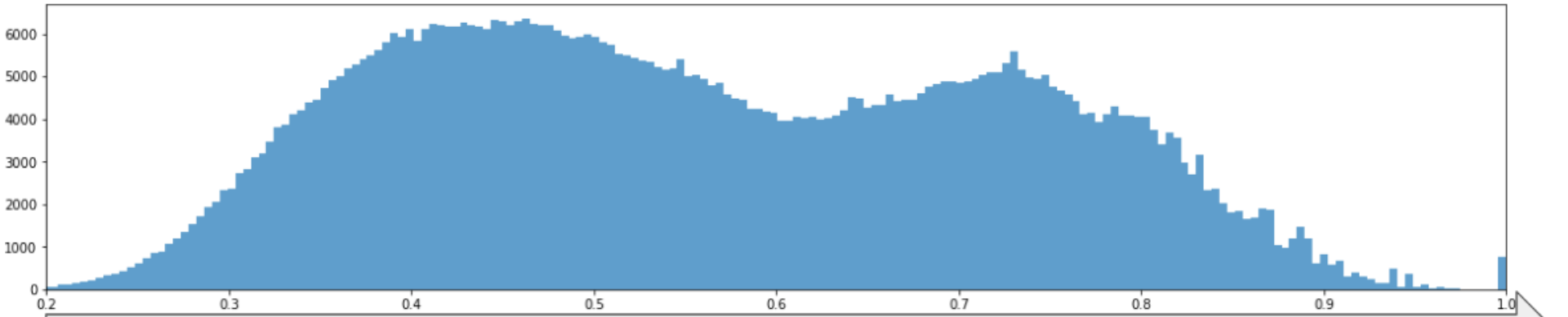
$$S(< \mathcal{Q}, \mathcal{A} >, \mathcal{C}) = \frac{1}{2}(S(\mathcal{Q}, \mathcal{C}) + S(\mathcal{A}, \mathcal{C}))$$



E: Group of people tasting wine next to three wooden barrels.

Similarity Distribution

❖ We set a threshold of ***0.6*** to remove bad explanations.



0.2

Q:What car sign can you see?
A:lexus
E:A lady is holding her tennis racket for the crowd.

0.5

Q:Is there a travel guide on the table?
A:yes
E:A piece of cake and coffee are on an outdoor table.

0.6

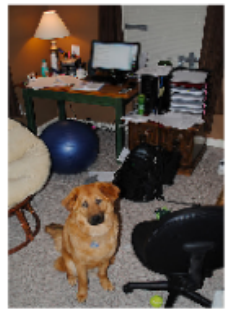
0.7

Q:What team does the man play for?
A:boston
E: A man wearing a Boston jersey walks with a bat.

0.8

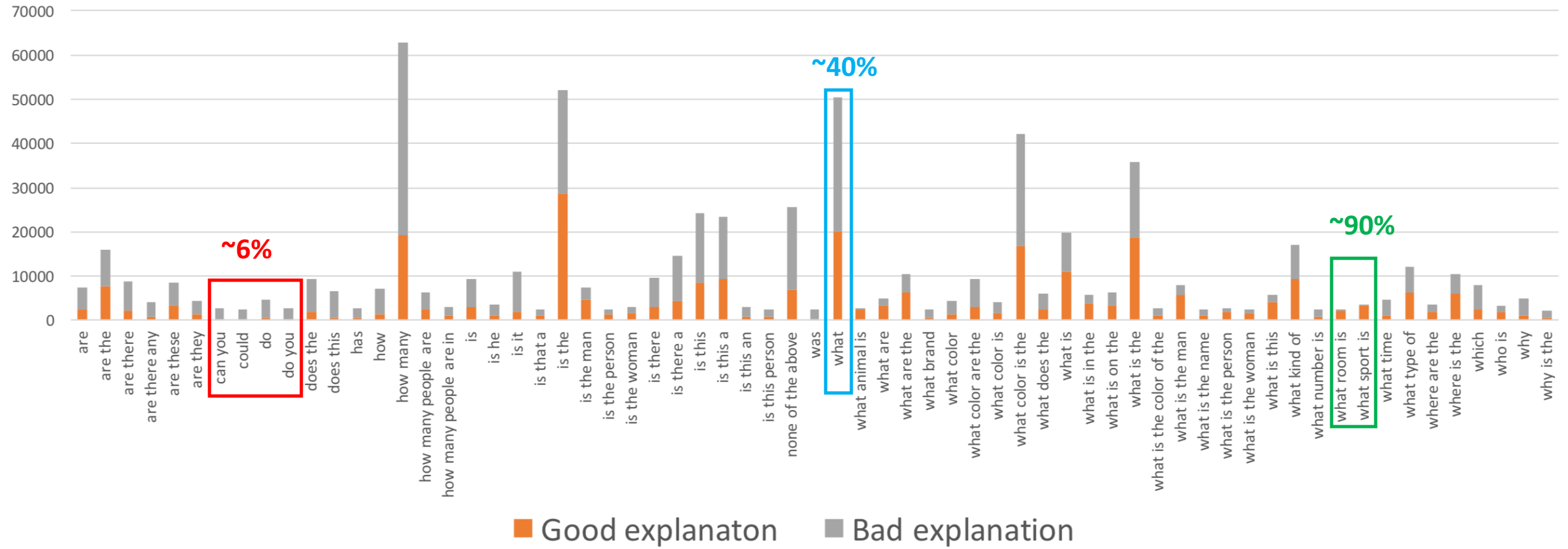
Q:Is this a passenger train?
A:no
E: A cargo train that is traveling down some railroad tracks.

0.9



Q:Where is the dog?
A:office
E: A dog is sitting in the middle of a home office area.

Specific v.s. Abstract v.s. Subjective Questions



Subjective Examples

Q:Do you think this pony is cute?
A:Yes
E:A horse that is standing in the grass. (similarity score = 0.48, not a good explanation)

Q: Can you cross the street?
A: No
E: A dog and people are on a dark city street. (similarity score = 0.47, not a good explanation)

Q: Could you eat all these bananas by yourself?
A: Yes
E: A bowl that has slices of bananas in it. (similarity score = 0.46, not a good explanation)

Require commonsense knowledge!

Dataset Assessment – User Study

- 20 Subjects, 2,000 <Q, A, E> pairs
- *fluent, correct, relevant, complementary*
- 1-very poor, 2-poor, 3-barely acceptable, 4-good, 5-very good.

Evaluation # 3 | Q-Question, A-Answer, E-Explanation. Help

Q: Where is the donuts?

A: plate

E: A table with a donut on a plate

Please evaluate E (explanation) in following aspects.

Fluent:

1 2 3 4 5

Correct:

1 2 3 4 5

Relevant:

1 2 3 4 5

Complementary:

1 2 3 4 5

Next

User Study Results

	Fluent	Correct	Relevant	Complementary
Synthesized Explanation	4.89	4.78	**4.23**	**4.14**
Most Similar Caption	**4.97**	4.91	2.72	2.87
Random Caption	4.93	**4.92**	1.91	2.12
Generated Explanation (QI-AE)	3.89	3.67	3.24	3.11

Multi-task VQA-E Model

$$\tau_i = \boldsymbol{w}^T \left(\text{Relu}(W_v \boldsymbol{v}_i) \odot \text{Relu}(W_q \boldsymbol{q})\right)$$

$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{\tau})$$

$$\phi = \text{CNN}(\mathcal{I}) = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_P\}$$

$$\boldsymbol{v} = \sum_{i=1}^{P} \alpha_i \boldsymbol{v}_i$$

CNN

Attention

GRU

FC

Sigmoid

Answer: red, white, and blue

Question: What color is this airplane?

LSTM

Explanation: A red, white and blue plane has a tag on one blade.

$$W_q = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_{T_q}\}$$

$$\boldsymbol{q} = \text{GRU}(W_q)$$

$$\boldsymbol{h} = \text{Relu}(W_{qh} \boldsymbol{q}) \odot \text{Relu}(W_{vh} \boldsymbol{v})$$

Variant Image Features

- Global
  - ‘pool5’ in ResNet-152.
  - $P = 1$
- Grid
  - ‘res5c’ in ResNet-152.
  - $P = 7 \times 7 = 49$
- Bottom-up [1]
  - Features of object detection proposals. (Faster R-CNN)
  - $P = 36$

[1]. Anderson *et al.* “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering” (CVPR'18)

Soft Target in Answer Prediction

$$\text{Accuracy}(a) = \frac{1}{K}\sum_{k=1}^{K}\min\left(\frac{\sum_{1\le j\le K, j\neq k}\mathbb{1}(a=a_j)}{3}, 1\right)$$

$$= \begin{cases} 0.0, & \text{if } \sum_{1\le j\le K}\mathbb{1}(a=a_j)=0 \\ 0.1, & \text{if } \sum_{1\le j\le K}\mathbb{1}(a=a_j)=1 \\ 0.3, & \text{if } \sum_{1\le j\le K}\mathbb{1}(a=a_j)=2 \\ 0.9, & \text{if } \sum_{1\le j\le K}\mathbb{1}(a=a_j)=3 \\ 1.0, & \text{if } \sum_{1\le j\le K}\mathbb{1}(a=a_j)\ge 4 \end{cases}$$

Multi-label regression problem:

$$\hat{s} = \text{sigmoid}\ (W_o\ \text{Relu}\ (W_f\ \boldsymbol{h}))$$

Model Variants

- **Q-E:** generating explanation from question only.
- **I-E:** generating explanation from image only.
- **QI-E:** generating explanation from question and image and only training the branch of explanation generation.
- **QI-A:** predicting answer from question and image and only training the branch of answer prediction.
- **QI-AE:** predicting answer and generating explanations, training both branches.
- **QI-AE(relevant):** predicting answer and generating explanation and training both branches. The explanation used in this variant is the relevant caption obtained in the process of explanation synthesis.
- **QI-AE(random):** predicting answer and generating explanation and training both branches. The explanation is randomly selected from the ground-truth captions for the same image except the relevant caption.

Evaluation of Explanation Generation

Table 3. Performance of explanation generation task on the validation split of the proposed VQA-E dataset, where B-N, M, R, and C are short for BLEU-N, METEOR, ROUGE-L, and CIDEr-D. All scores are reported in percentage (%).

Model	Image Features	B-1	B-2	B-3	B-4	M	C	R
Q-E	-	26.80	10.90	4.20	1.80	7.98	13.42	24.90
I-E	Global	32.50	17.20	9.30	5.20	12.38	48.58	29.79
QI-E	Global	34.70	19.30	11.00	6.50	14.07	61.55	31.87
	Grid	36.30	21.10	12.50	7.60	15.50	73.70	34.00
	Bottom-up	38.00	22.60	13.80	8.60	16.57	84.07	34.92
QI-AE	Global	35.10	19.70	11.30	6.70	14.40	64.62	32.39
	Grid	38.30	22.90	14.00	8.80	16.85	87.04	35.16
	Bottom-up	**39.30**	**23.90**	**14.80**	**9.40**	**17.37**	**93.08**	**36.33**

Evaluation of Answer Prediction

Table 4. Performance of the answer prediction task on the validation split of VQA v2 dataset. Accuracies in percentage (%) are reported.

Model	Image features	All	Yes/No	Number	Other
QI-A	Global	57.26	77.19	39.73	46.74
	Grid	59.25	76.31	39.99	51.38
	Bottom-up	61.78	78.63	41.30	52.54
QI-AE	Global	57.92	78.01	40.46	47.25
	Grid	60.57	78.35	39.36	52.66
	Bottom-up	**63.51**	**80.85**	**43.02**	**54.16**
QI-AE(random)	Bottom-up	58.74	78.75	40.79	48.26
QI-AE(relevant)	Bottom-up	62.18	79.02	41.07	53.26

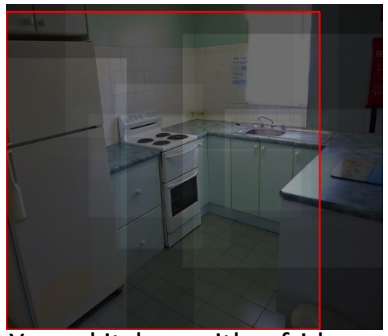
Qualitative Examples

(a)

Is this a kitchen?

Yes, the kitchen is clean and light comes in the window.

No.



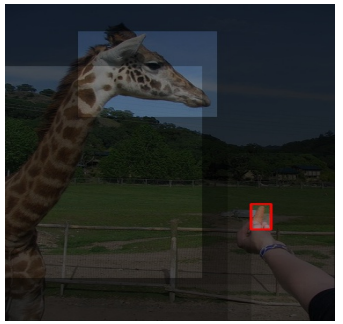
Yes, a kitchen with a fridge, sink , and cabinets in it.

(b)

What is the person doing?

Feeding giraffe, a close up of a person feeding a giraffe

Standing.



Feeding giraffe, a person feeding a giraffe.

(c)

Is this a sunny day?

Yes, kites being flown in a park on a sunny day.

No.

Yes, a kite flying in the air on a sunny day.

(d)

What color is the remote?

White, white controller sitting on top of a wooden table next to a battery.

White.

White, a white controller sitting on a table.

Ground truth QI-A QI-AE Ground truth QI-A QI-AE

Conclusions

- We propose a new dataset (VQA-E) to promote research on justifying answers for visual questions
  - Synthesized explanations are of high quality for visually-specific questions.
  - Inadequate for subjective questions which require commonsense.
- We propose a novel multi-task learning framework for VQA-E
  - The additional supervision from explanations helps the model *better localize* and *understand* the important image regions.