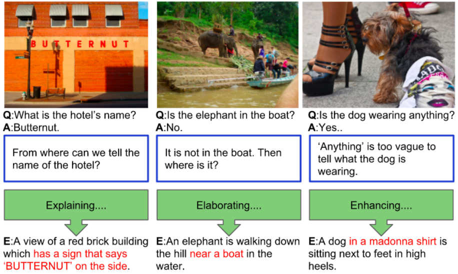
VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions

Qing Li[1], Qingyi Tao[2,3], Shafiq Joty[2], Jianfei Cai[2], and Jiebo Luo[4]
[1]University of Science and Technology of China, [2]Nanyang Technological University,
[3]NVIDIA AI Technology Center, [4]University of Rochester

Introduction

Goal: Generate Textual Justifications for Predicted Answers

- ➢ Be Accessible to visually impaired people.
- ➢ Provide beneficial feedbacks that enable the questioners to extend the conversation for effective communication.



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

From where can we tell the name of the hotel?

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?

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.

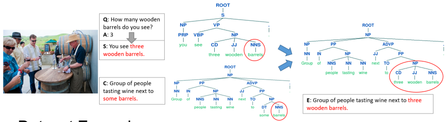
E:A dog in a madonna shirt is sitting next to feet in high heels.

Contributions:

- ➢ Constructed a new dataset with textual justifications for the answers, which is automatically derived from the VQA v2 dataset by intelligently exploiting the available captions.
- ➢ Proposed a novel multi-task learning framework which can generate a sentence to explain the predicted answer.
- ➢ Outperformed the state-of-the-art methods by a clear margin on the VQA v2 dataset.

VQA-E Dataset

➢ Explanation Synthesis



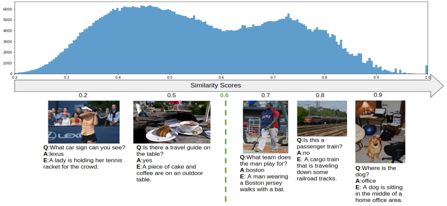
Q: How many wooden barrels do you see?
A: 3

S: You see three wooden barrels.

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

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

➢ Dataset Examples



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

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

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

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

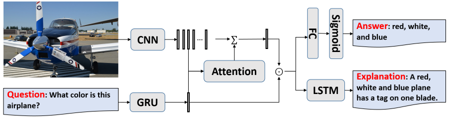
Q:Where is the dog?
A:office
E: A dog is sitting in the middle of a home office area.

➢ Dataset Analysis

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

Download: https://github.com/liqing-ustc/VQA-E

Multi-task VQA Model



Answer: red, white, and blue

Question: What color is this airplane?

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

$$L = L_{\mathrm{vqa}} + L_{\mathrm{vqe}}$$

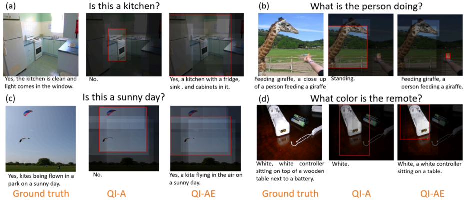
Experimental Results

Explanation Generation

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**

Answer Prediction

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



(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

Insight: the additional supervision from explanations helps the model better localize and understand the important image regions.