

Visual Dialog

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, Dhruv Batra

Presented by: Alan Luo

Introduction **Natural Language Processing + Computer Vision**

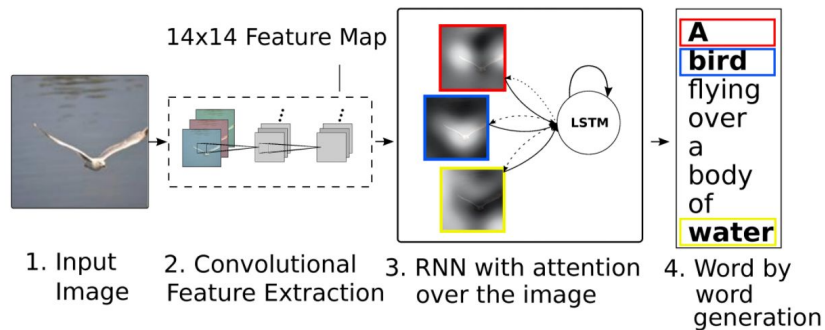
- Aiding visually impaired users in understanding their surroundings or social media content
- Interacting with an AI assistant

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 <p>A person riding a motorcycle on a dirt road.</p>	 <p>Two dogs play in the grass.</p>	 <p>A skateboarder does a trick on a ramp.</p>	 <p>A dog is jumping to catch a frisbee.</p>
 <p>A group of young people playing a game of frisbee.</p>	 <p>Two hockey players are fighting over the puck.</p>	 <p>A little girl in a pink hat is blowing bubbles.</p>	 <p>A refrigerator filled with lots of food and drinks.</p>
 <p>A herd of elephants walking across a dry grass field.</p>	 <p>A close up of a cat laying on a couch.</p>	 <p>A red motorcycle parked on the side of the road.</p>	 <p>A yellow school bus parked in a parking lot.</p>



Related Work Image/Video Captioning

Image Captioning



a man is throwing a frisbee in a park



a man riding a wave on top of a surfboard

Video Captioning

Input video:

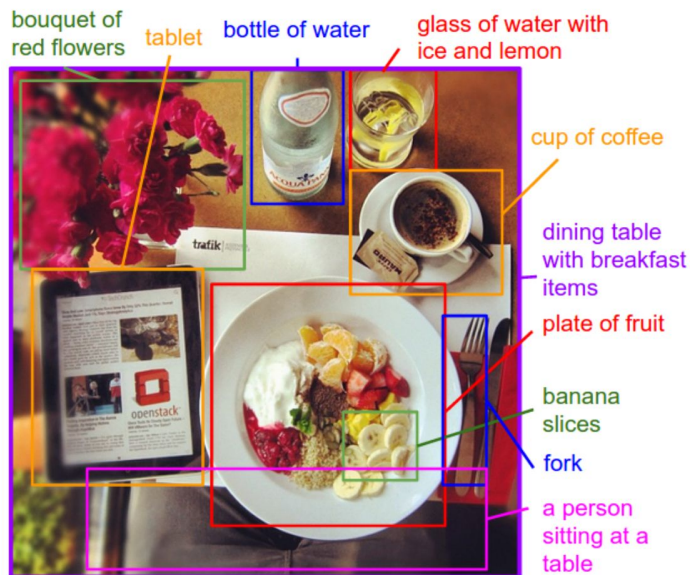


Our output: A cat is playing with a toy.

Humans: A Ferret and cat fighting with each other. / A cat and a ferret are playing. / A kitten is playing with a ferret. / A kitten and a ferret are playfully wrestling.

Related Work Visual-Semantic Alignments

Visual-Semantic Alignments



Datasets

Regions	Attributes	Relationships
the white bowls on the table	bowl is white	top
the woman teaching the little girl to cook	pan is silver	lady WEARING shirt
the liquid in the bowl	kitchen utensils is kitchen	woman helping girl
the utensil on the food	kitchen utensils is behind	sauce ON bread
the pink shirt on the woman	utensils is behind	tin foil under bread
the pink shirt on the little girl	pizza is unripe	paper hanging on wall
little girl with a pink top on next to	hand is white	lady helping girl
	shirt is pink	child standing at table
	girl is young	woman IN kitchen
	spoon is large	
	spoon is silver	

Question Answers

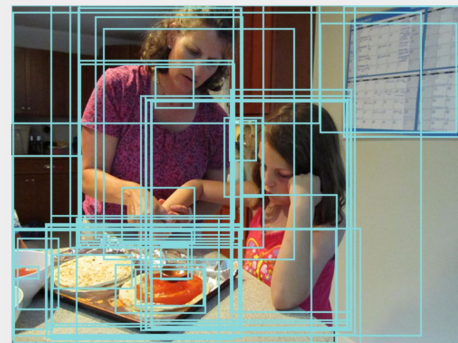
What color is the girl's shirt? Pink.

Where was the photo taken? Kitchen.

What are the people doing? Cooking.

How many adults in the kitchen? One.

What color is the sauce being put onto the food? Red.



Related Work Visual Q&A



How many pickles
are on the plate?

1
1
1

1
1
1

What is the shape
of the plate?

circle
round
round

circle
round
round



What does
the sign say?

stop
stop
stop

stop
stop
yield

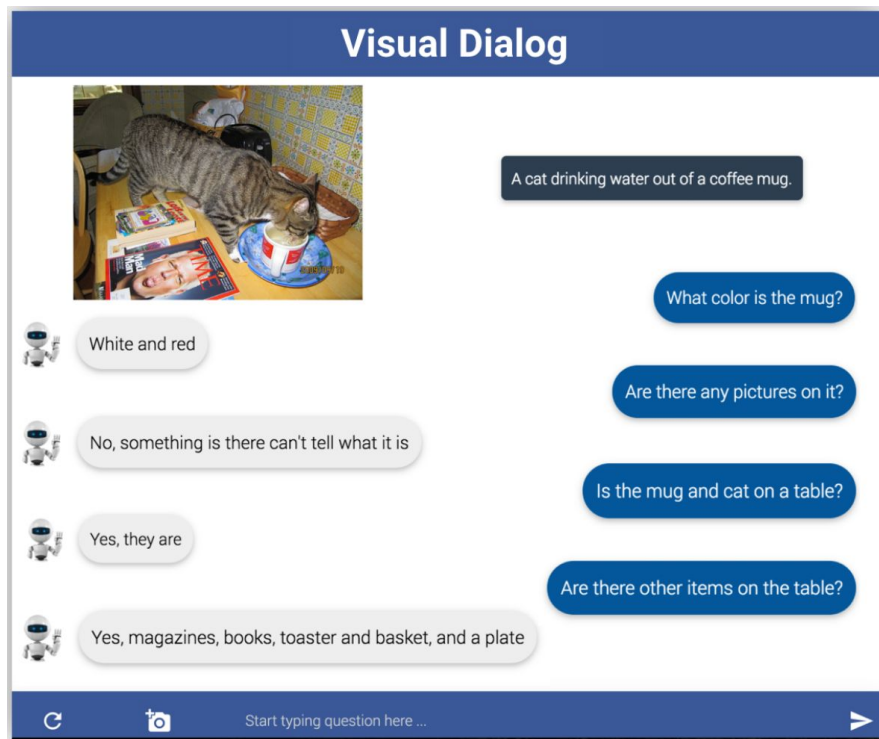
What shape is
this sign?

octagon
octagon
octagon

diamond
octagon
round

Contributions

1. Propose a new AI task: Visual Dialog
2. Develop a novel two-person chat data-collection protocol and introduce a new dataset
3. Introduce a family of neural encoder-decoder models for Visual Dialog



Technical Details With Late Fusion Encoder



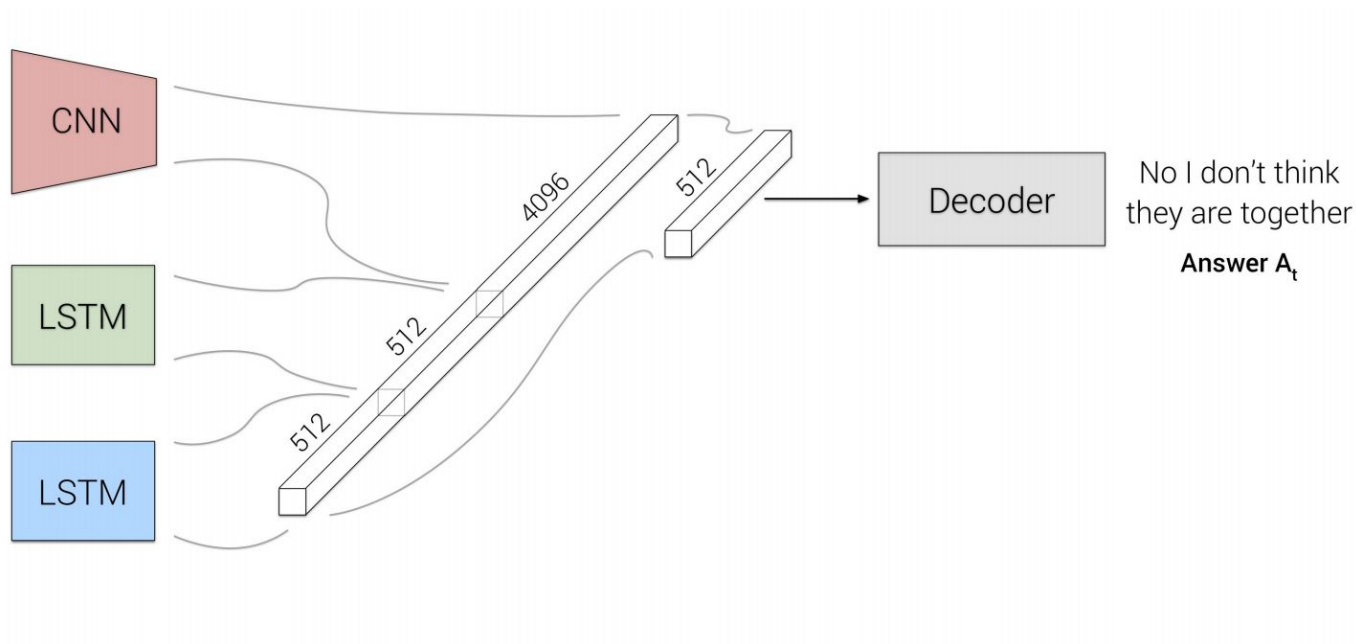
Image I

Do you think the woman is with him?

Question Q_t

The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on. ... Are there any people nearby? Yes there's a woman walking behind him.

**t rounds of history
(concatenated)**



Dataset VisDial

Qualitative



Caption: A statue depicting a bear breaking into a car.

Person A (1): how big is statue

Person B (1): about size of real full grown bear

Person A (2): so is car full size then as well

Person B (2): yes replica of car

Person A (3): is statue all 1 color

Person B (3): no brown and black

Person A (4): what color is car

Person B (4): dark red

Person A (5): where is this, do you think

Person B (5): in wooded area someplace

Person A (6): do you see any people in image

Person B (6): yes 1 man

Person A (7): how old is man

Person B (7): 35-40

Person A (8): what is man doing

Person B (8): sitting in car behind replica

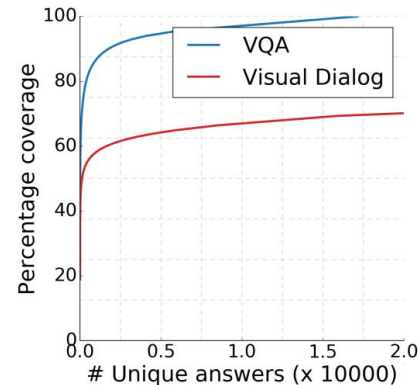
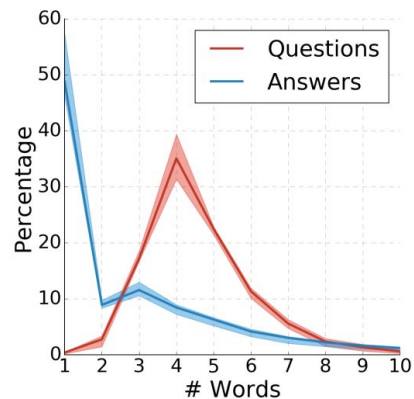
Person A (9): do you see any signs

Person B (9): yes, on car door warning sign

Person A (10): what else can you tell me about this image

Person B (10): there are many trees in background

Quantitative



Results

Qualitative Results

Visual Chatbot

Hi, I am a visual chatbot, capable of answering a sequence of questions about images. Please upload an image and fire away!



Caption: A group of bikers parked in a parking lot under large lights

moped

on street

cloudy sky looks clear

What is this?

Where is this?

how is the weather?

Quantitative Results

	Model	MRR	R@1	R@5	R@10	Mean
Baseline	Answer prior	0.311	19.85	39.14	44.28	31.56
	NN-Q	0.392	30.54	46.99	49.98	30.88
	NN-QI	0.385	29.71	46.57	49.86	30.90
Generative	LF-Q-G	0.403	29.74	50.10	56.32	24.06
	LF-QH-G	0.425	32.49	51.56	57.80	23.11
	LF-QI-G	0.437	34.06	52.50	58.89	22.31
	LF-QIH-G	0.430	33.27	51.96	58.09	23.04
	HRE-QH-G	0.430	32.84	52.36	58.64	22.59
	HRE-QIH-G	0.442	34.37	53.40	59.74	21.75
	HREA-QIH-G	0.442	34.47	53.43	59.73	21.83
	MN-QH-G	0.434	33.12	53.14	59.61	22.14
	MN-QIH-G	0.443	34.62	53.74	60.18	21.69
Discriminative	LF-Q-D	0.482	34.29	63.42	74.31	8.87
	LF-QH-D	0.505	36.21	66.56	77.31	7.89
	LF-QI-D	0.502	35.76	66.59	77.61	7.72
	LF-QIH-D	0.511	36.72	67.46	78.30	7.63
	HRE-QH-D	0.489	34.74	64.25	75.40	8.32
	HRE-QIH-D	0.502	36.26	65.67	77.05	7.79
	HREA-QIH-D	0.508	36.76	66.54	77.75	7.59
	MN-QH-D	0.524	36.84	67.78	78.92	7.25
	MN-QIH-D	0.529	37.33	68.47	79.54	7.03
VQA	SAN1-QI-D	0.506	36.21	67.08	78.16	7.74
	HieCoAtt-QI-D	0.509	35.54	66.79	77.94	7.68
Human Accuracies						
Human	Human-Q	0.441	25.10	67.37	-	4.19
	Human-QH	0.485	30.31	70.53	-	3.91
	Human-QI	0.619	46.12	82.54	-	2.92
	Human-QIH	0.635	48.03	83.76	-	2.83