Visual Dialog

Virginia Tech · Georgia Tech · Carnegie Mellon University

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

arxiv.org/abs/1611.08669
visualdialog.org

Goal

Captioning
Two people are in a wheelchair and one is holding a racket

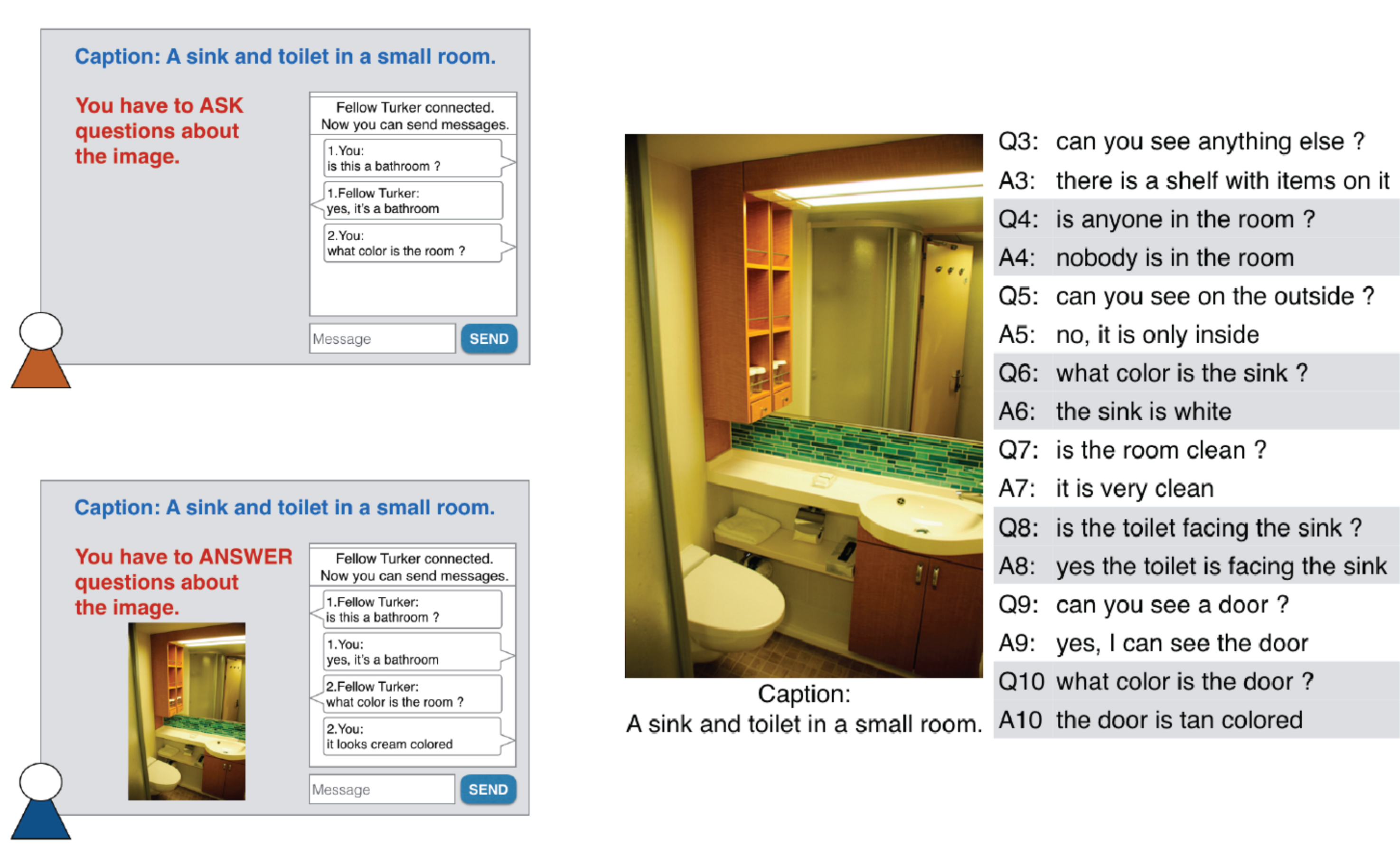
Visual Question Answering
How many people on wheelchairs? Two
How many wheelchairs? One

Visual Dialog
Q1: How many people on wheelchairs?
A1: Two
Q2: What are their genders?
A2: One male and one female
Q3: Which one is holding a racket?
A4: The woman
...

Task: Given image, dialog history, follow-up question — predict free-form answer

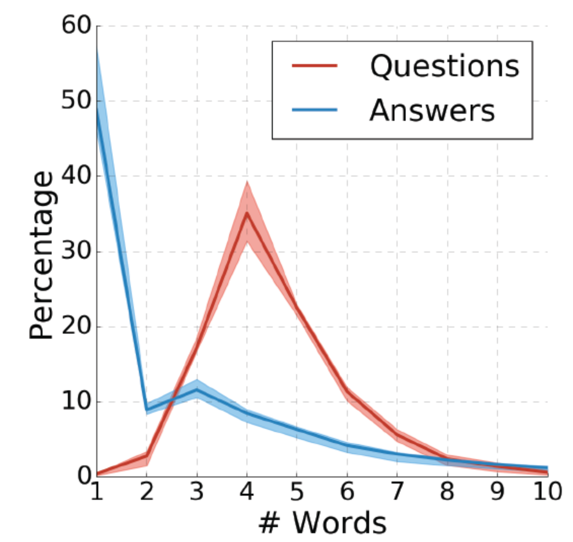
Build an agent capable of holding a meaningful dialog with humans in conversational language about visual content

VisDial Data Collection

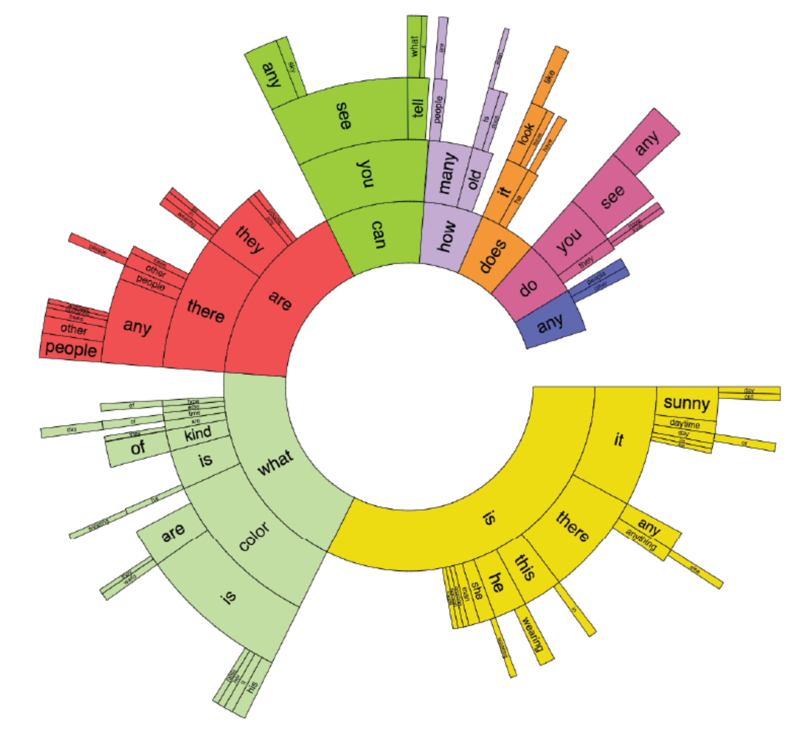


Two-person real-time chat on Amazon Mechanical Turk

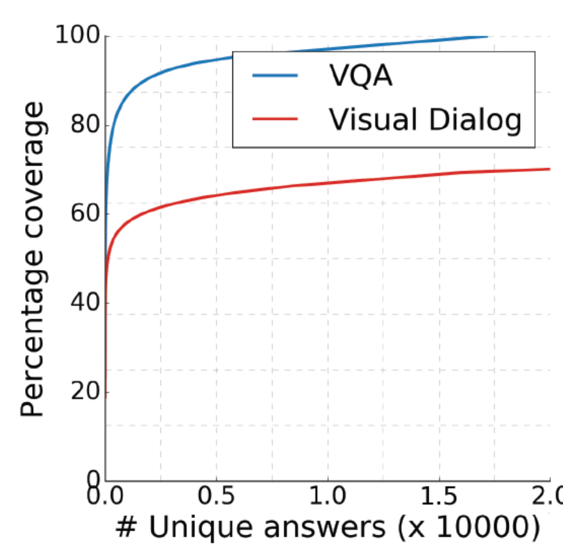
VisDial Dataset & Evaluation



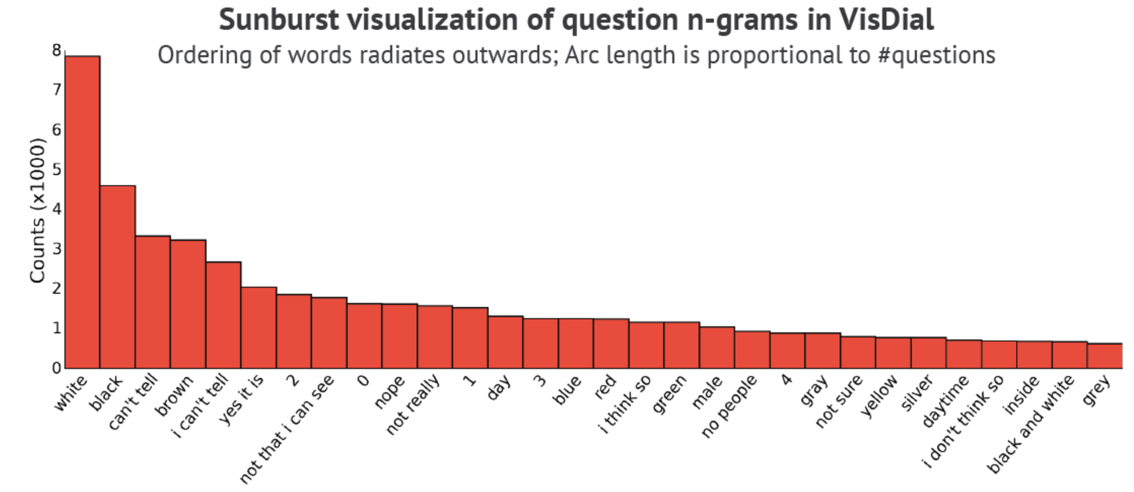
Answers in VisDial are longer
3.1 words (VisDial) vs. 1.1 (VQA) vs. 2.0 (Visual 7W)



Sunburst visualization of question n-grams in VisDial
Ordering of words radiates outwards; Arc length is proportional to #questions



VisDial captures a rich heavy tail
top-1000 answers cover ~58% of the dataset



Most frequent answer responses in VisDial except for "yes"/"no"

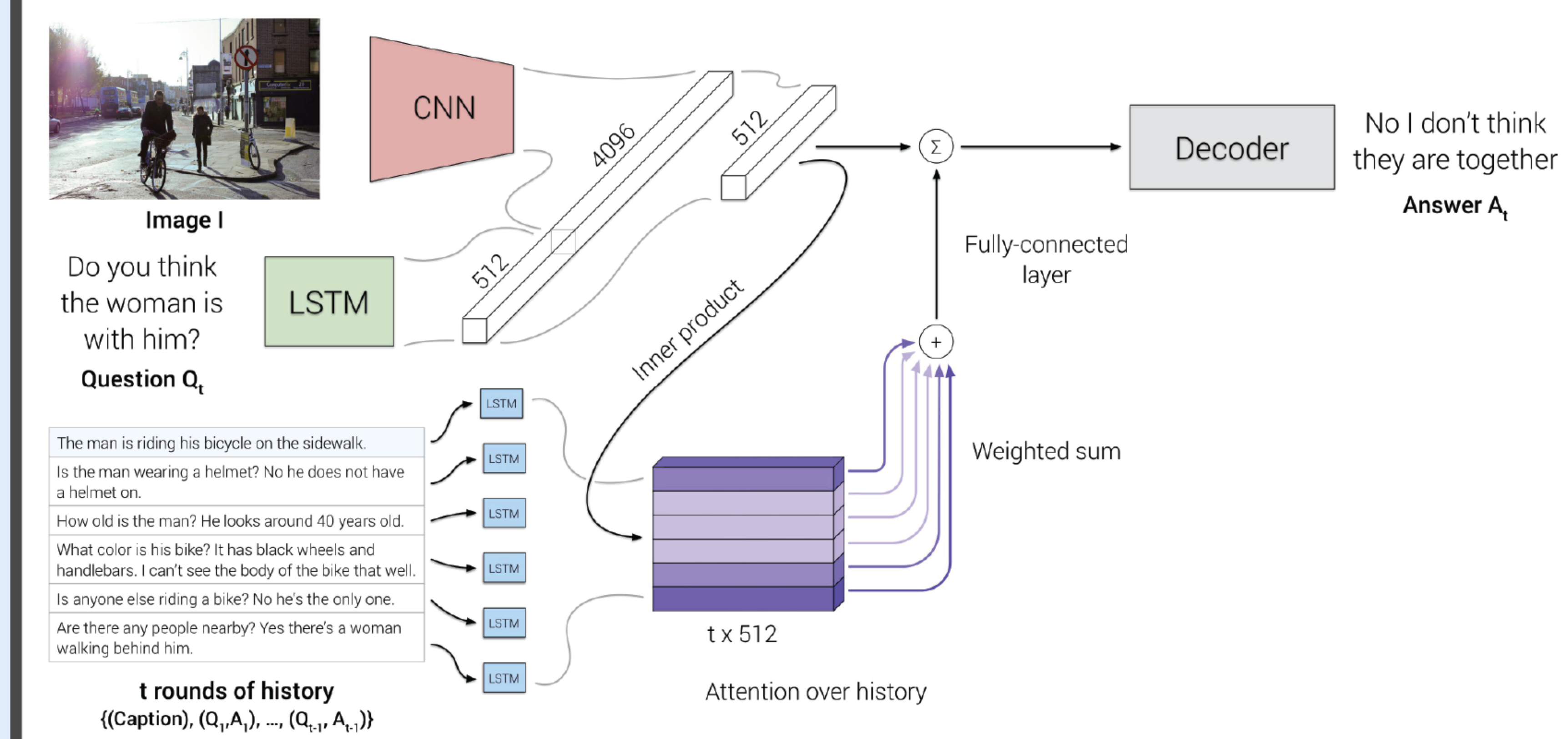
Evaluation: Given **image**, **dialog history**, **question**, **100 candidate answers** — evaluate model on **retrieval of ground-truth human response**
Metrics: mean reciprocal rank, recall@k, mean rank

>140k dialogs on COCO images; >1.4M dialog question-answers
Evaluation by retrieval of ground truth human response

Encoder-Decoder Models

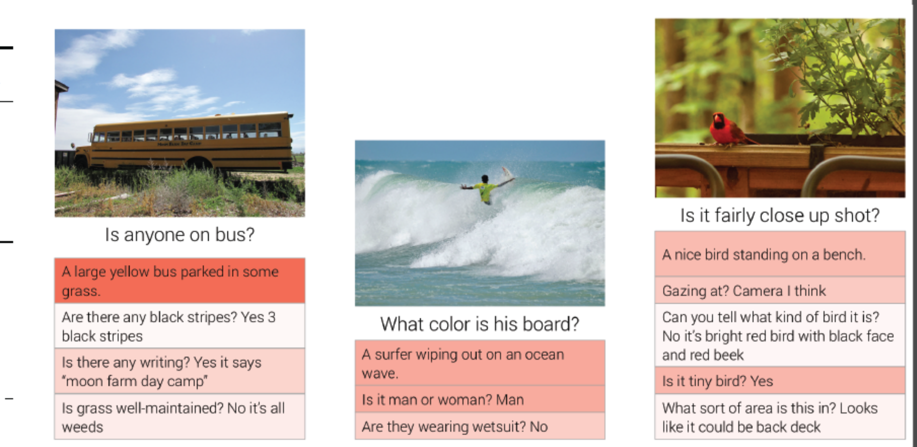
Late Fusion (LF) Encoder: Naive embedding of image, history, question
Hierarchical Recurrent Encoder (HRE): Dialog-level recurrent block on top of QA-level recurrent block
Memory Network (MN) Encoder (Similar to Weston et al., 2014)**:** Builds context vector from previous QA facts in memory

Generative (G) Decoder: RNN initialized with input encoding, predicts response word-by-word; Trained to maximize LL of ground truth response
Discriminative (D) Decoder: Dot product between input encoding and RNN encoding of 100 candidate answers + 100-way softmax



Memory Network (MN) Encoder

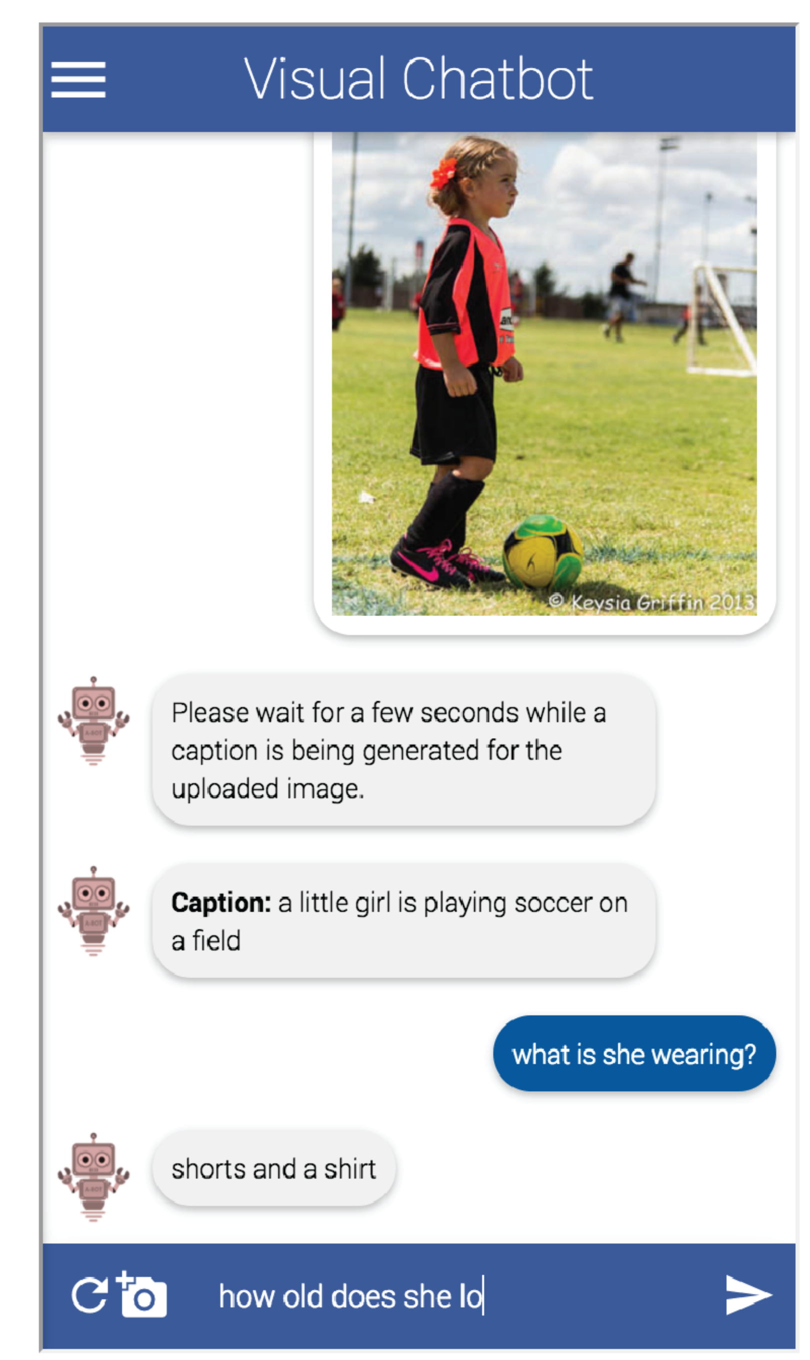
	Model	MRR	R@1	R@5	R@10	Mean
Baseline	Answer prior	0.311	19.85	39.14	44.28	31.56
	NN-Q	0.392	30.54	46.99	49.98	30.88
	NN-QI	0.385	29.71	46.57	49.86	30.90
Generative	LF-Q-G	0.403	29.74	50.10	56.32	24.06
	LF-QH-G	0.425	32.49	51.56	57.80	23.11
	LF-QI-G	0.437	34.06	52.50	58.89	22.31
	LF-QIH-G	0.430	33.27	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



Selected examples of attention from our Memory Network
Intensity of color indicates strength of attention placed on that fact

demo.visualdialog.org

Real-time visual chatbot hosted on CloudCV



Discriminative models work better than generative models
MN works best in both generative and discriminative settings
Human performance, topic transition studies, dialog perplexities, etc. in paper