Visual Dialog

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

Presented by Wei-Chieh Wu

Visual Dialog

- Requires an AI agent to hold a meaningful dialog with humans about visual content.
- Input:
  - Image
  - Dialog history
  - Question
- Output:
  - Answer to the question

VQA vs Visual Dialog

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

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

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

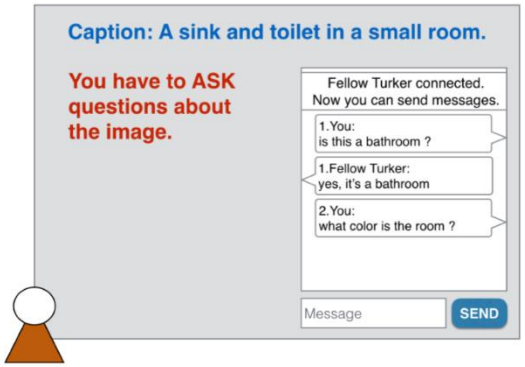
A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

VisDial Dataset

- Contains ~123k images and 10 question-answer pairs for each image
- Images are from COCO dataset
- Question-answer pairs are collected on AMT with human dialog



(a) What the 'questioner' sees.

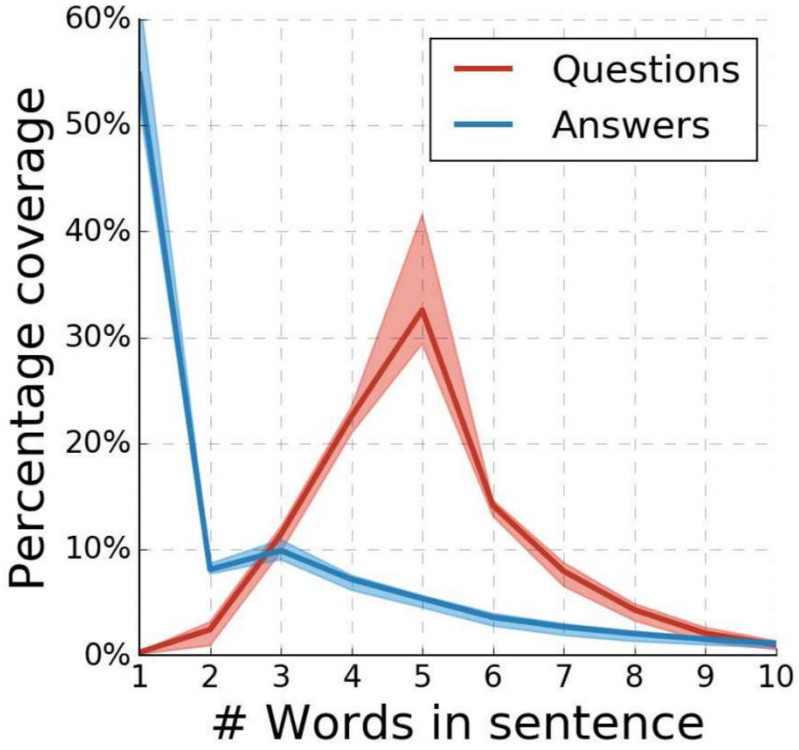
(b) What the 'answerer' sees.

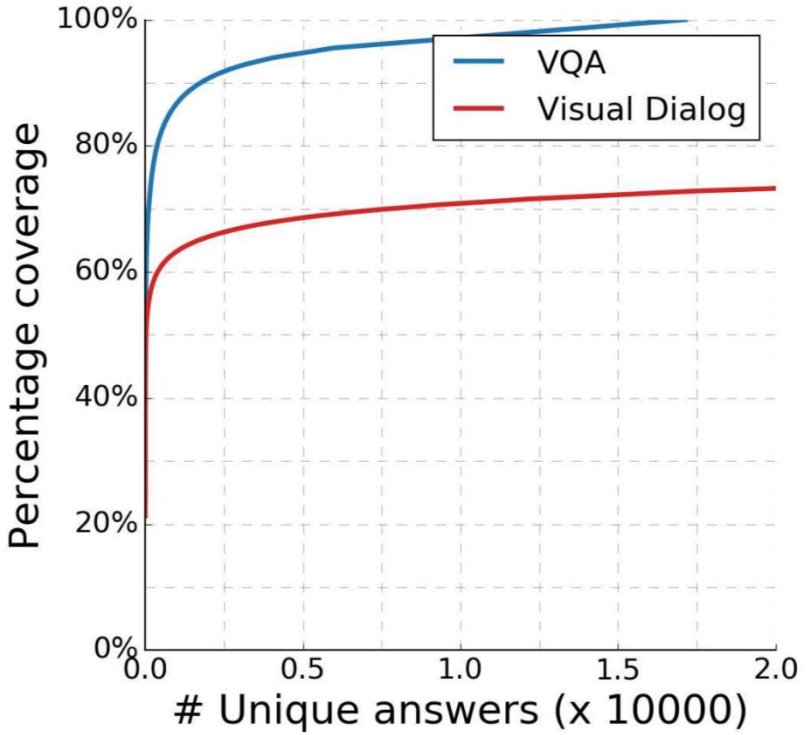
Caption: A sink and toilet in a small room.

Q3:	can you see anything else ?
A3:	there is a shelf with items on it
Q4:	is anyone in the room ?
A4:	nobody is in the room
Q5:	can you see on the outside ?
A5:	no, it is only inside
Q6:	what color is the sink ?
A6:	the sink is white
Q7:	is the room clean ?
A7:	it is very clean
Q8:	is the toilet facing the sink ?
A8:	yes the toilet is facing the sink
Q9:	can you see a door ?
A9:	yes, I can see the door
Q10	what color is the door ?
A10	the door is tan colored

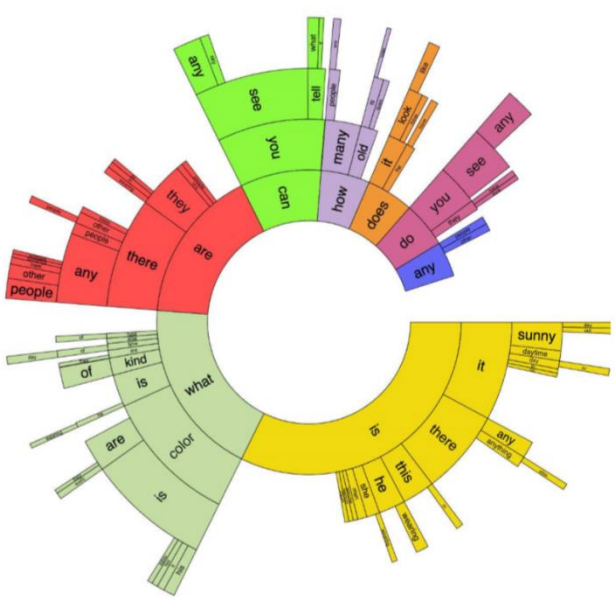
(c) Example dialog from our VisDial dataset.

VisDial Dataset

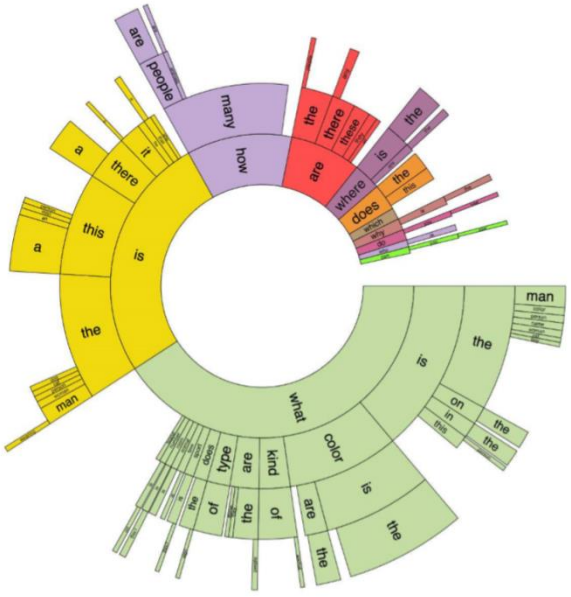




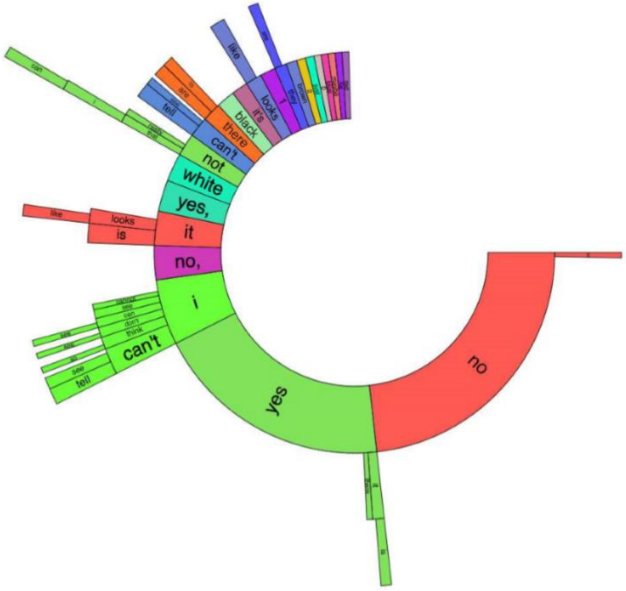
VisDial Dataset



(a) VisDial Questions



(b) VQA Questions



(c) VisDial Answers

Evaluation

- Given N = 100 candidate answers, return a sorting of them
- Candidate answers:
  - The human response
  - Answers to 50 most similar questions
  - 30 most popular answers from the dataset
  - 19 random answers
- Retrieval metrics:

  MRR, recall@k, average rank of the human response

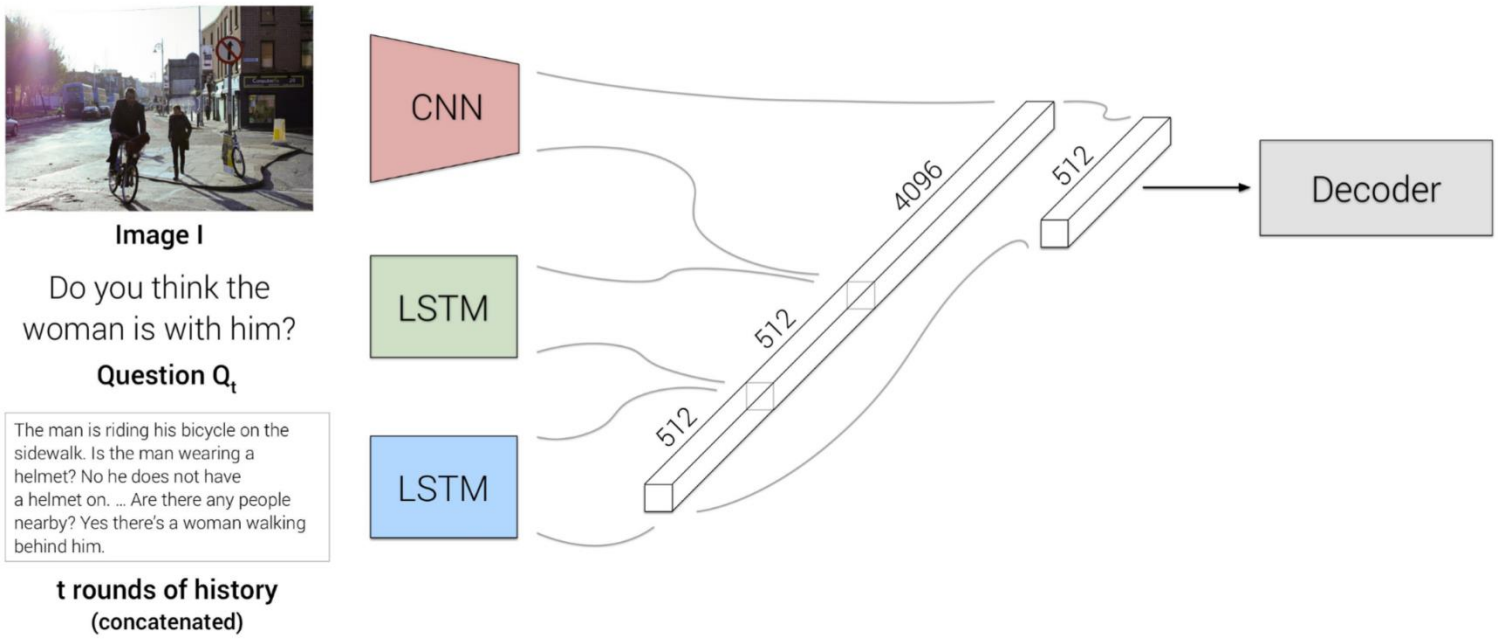
Models

- Following the encoder-decoder framework
- 2 kinds of decoder
  - Generative Decoder
  - Discriminative Decoder
- 3 kinds of encoder
  - Late Fusion Encoder
  - Hierarchical Recurrent Encoder
  - Memory Network Encoder

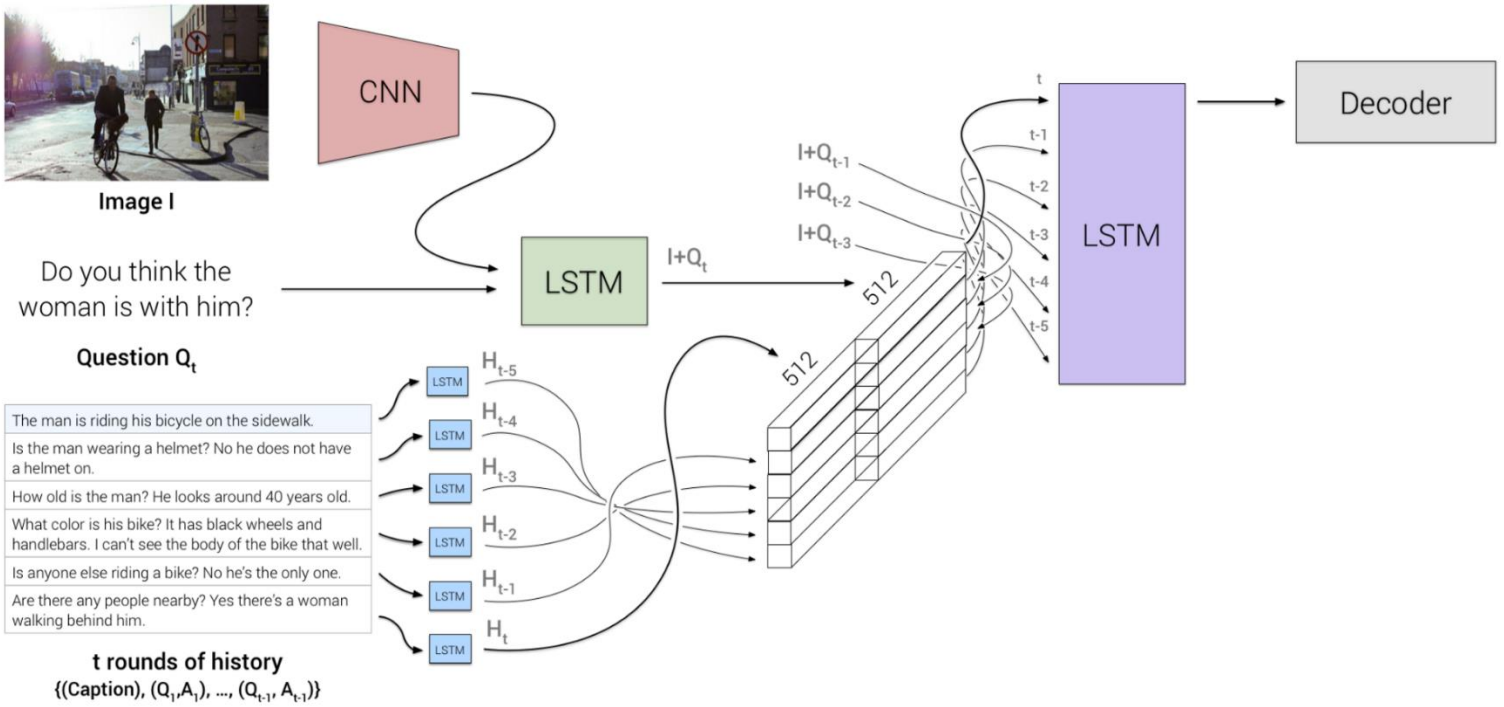
Decoders

- Generative Decoder
  - LSTM decoder
  - Maximize the log-likelihood of the ground truth answer
  - Use the model's log-likelihood scores for ranking
- Discriminative Decoder
  - Compute similarity between the input encoding and LSTM encoding for candidate answers
  - Maximize softmax score of the ground truth answer
  - Use the similarities for ranking

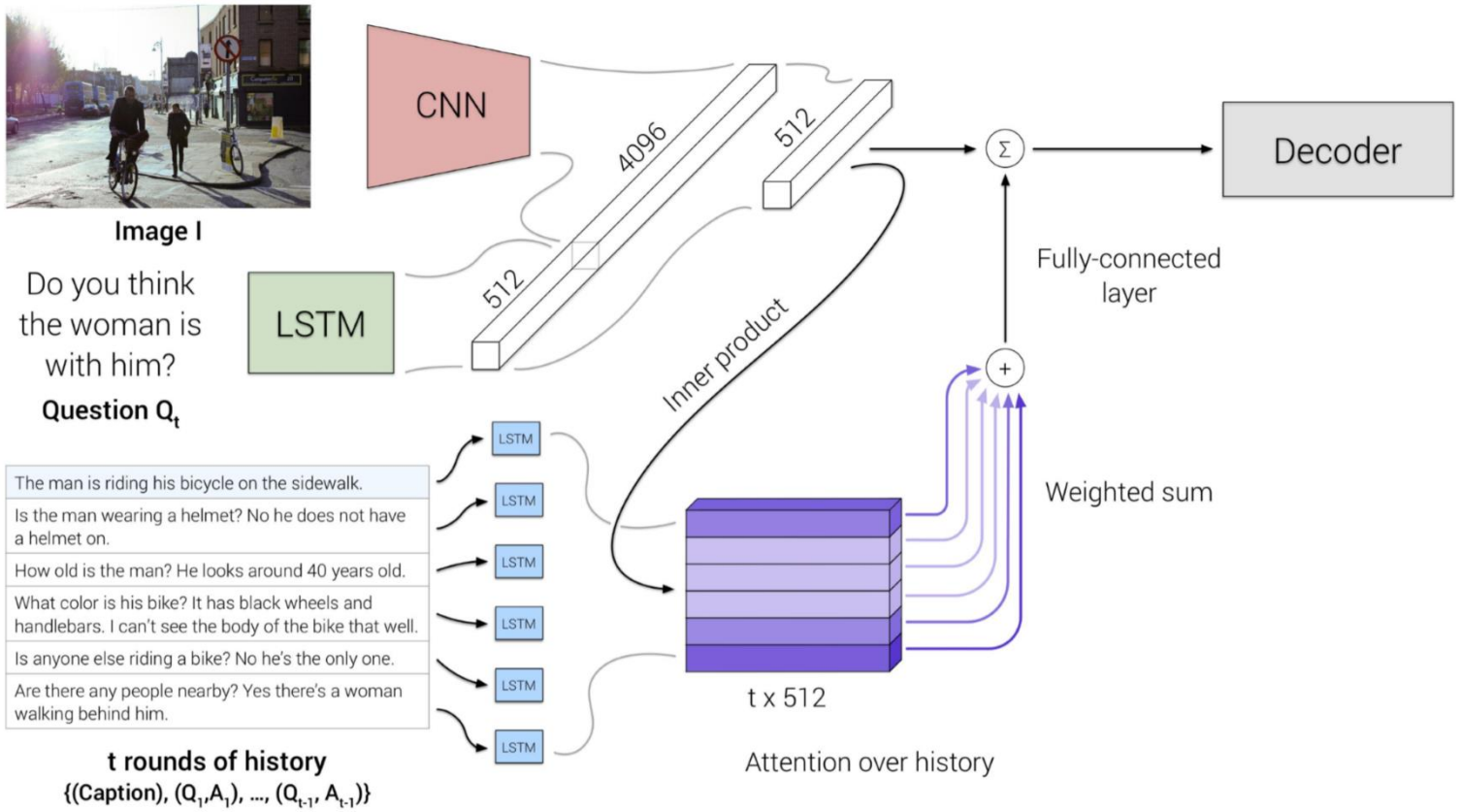
Late Fusion (LF) Encoder



Hierarchical Recurrent Encoder (HRE)



Memory Network (MN) Encoder



Experiments

- Dataset: VisDial v0.9
- Baseline
  - NN-Q:

    Find k nearest neighbor questions for a test question, and score answers by their mean similarity with these k answers
  - NN-QI:

    Find K nearest neighbor questions for a test question, then find a subset of size k based on image feature similarity. Score answers by their mean similarity with these k answers
- VQA models
  - SAN and HieCoAtt
  - Feed VQA outputs to their discriminative decoder, and train end-to-end on VisDial

Results

	Model	MRR	R@1	R@5	R@10	Mean
Baseline	Answer prior	0.3735	23.55	48.52	53.23	26.50
	NN-Q	0.4570	35.93	54.07	60.26	18.93
	NN-QI	0.4274	33.13	50.83	58.69	19.62
Generative	LF-Q-G	0.5048	39.78	60.58	66.33	17.89
	LF-QH-G	0.5055	39.73	60.86	66.68	17.78
	LF-QI-G	0.5204	42.04	61.65	67.66	16.84
	LF-QIH-G	0.5199	41.83	61.78	67.59	17.07
	HRE-QH-G	0.5102	40.15	61.59	67.36	17.47
	HRE-QIH-G	0.5237	**42.29**	62.18	67.92	17.07
	HREA-QIH-G	0.5242	42.28	62.33	68.17	**16.79**
	MN-QH-G	0.5115	40.42	61.57	67.44	17.74
	MN-QIH-G	**0.5259**	**42.29**	**62.85**	**68.88**	17.06
Discriminative	LF-Q-D	0.5508	41.24	70.45	79.83	7.08
	LF-QH-D	0.5578	41.75	71.45	80.94	6.74
	LF-QI-D	0.5759	43.33	74.27	83.68	5.87
	LF-QIH-D	0.5807	43.82	74.68	84.07	5.78
	HRE-QH-D	0.5695	42.70	73.25	82.97	6.11
	HRE-QIH-D	0.5846	44.67	74.50	84.22	5.72
	HREA-QIH-D	0.5868	44.82	74.81	84.36	5.66
	MN-QH-D	0.5849	44.03	75.26	84.49	5.68
	MN-QIH-D	**0.5965**	**45.55**	**76.22**	**85.37**	**5.46**
VQA	SAN1-QI-D	0.5764	43.44	74.26	83.72	5.88
	HieCoAtt-QI-D	0.5788	43.51	74.49	83.96	5.84