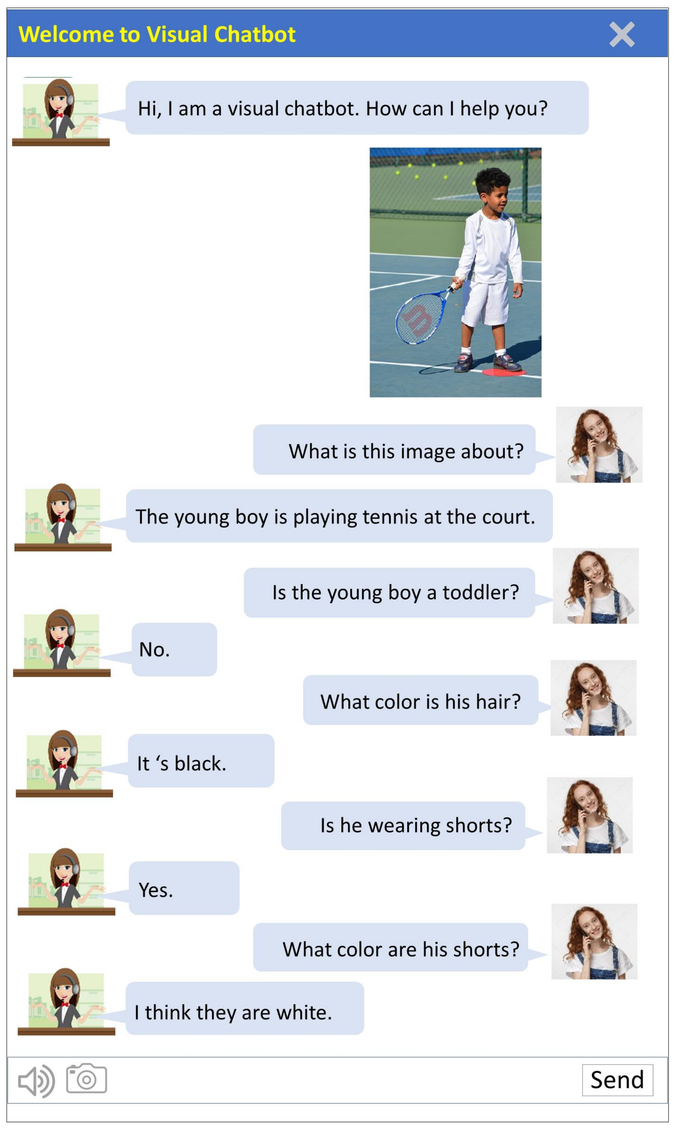
Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

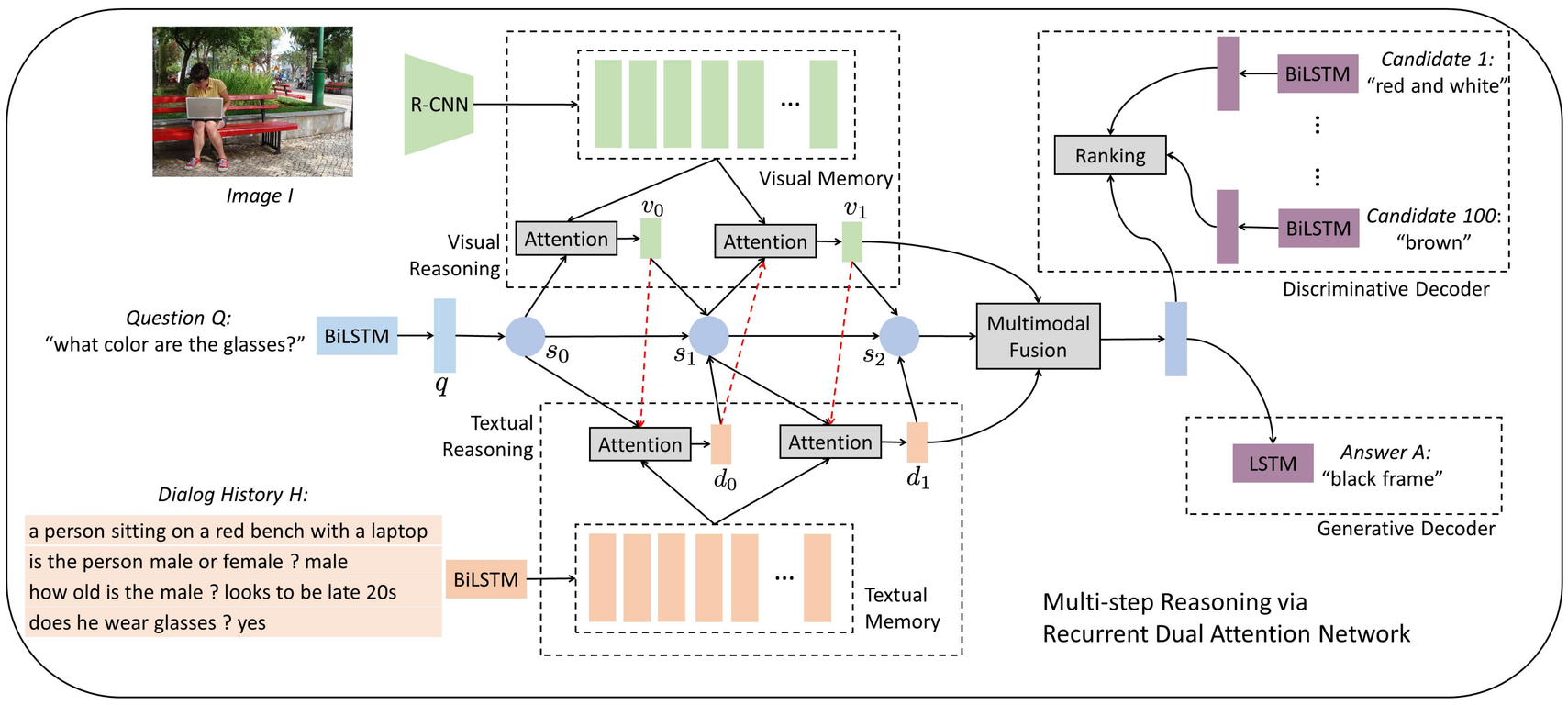
Zhe Gan[1], Yu Cheng[1], Ahmed El Kholy[1], Linjie Li[1], Jingjing Liu[1], Jianfeng Gao[2]
[1]Microsoft Dynamics 365 AI Research, [2]Microsoft Research

ACL 2019

Visual Dialog

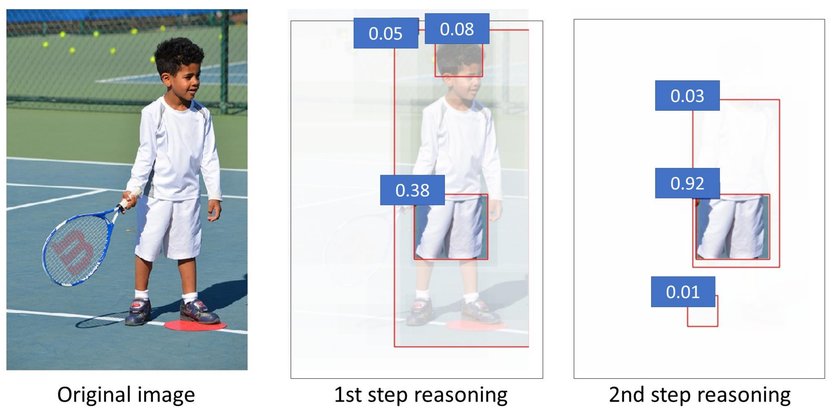


Recurrent Dual Attention Network (ReDAN)

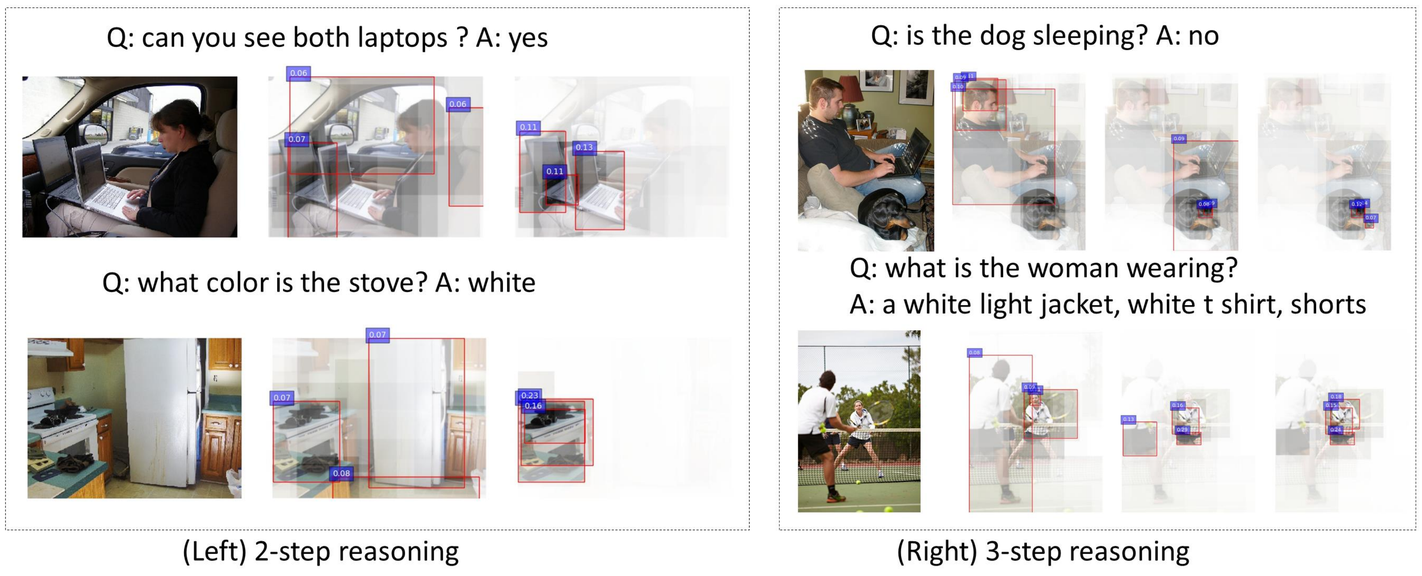


- Visual Memory: Faster R-CNN is used to extract image features
- Textual Memory: Each history snippet is encoded via BiLSTM
- Query: Self-attended question
- Visual Reasoning: Query and history attending to visual memory
- Textual Reasoning: Query and image attending to textual memory
- Multi-step Reasoning: Image and history updating RNN state
- Answer Decoding:
  - Discriminative (Dis.): Retrieval
  - Generative (Gen.): Text generation

Visualization of Multi-step Reasoning



Attention maps become sharper through more reasoning steps



Experimental Results

- Dataset: VisDial v1.0, containing 130K images and 1.3M QA pairs in total
- Multi-step reasoning lifts both Dis. & Gen. model performance
- Gen. models achieve higher NDCG than Dis. Models (performing better on Yes/no questions), but with much worse MRR & Mean Rank

Model	NDCG	MRR	Mean	NDCG	MRR	Mean
Decoder Type	Discriminative (Dis.)			Generative (Gen.)		
MN	55.13	60.42	4.63	56.99	47.83	18.76
HCIAE	57.65	62.96	4.24	59.70	49.07	18.43
CoAtt	57.72	62.91	4.21	59.24	49.64	17.86
ReDAN (T=1)	58.49	63.35	4.19	59.41	49.60	17.79
ReDAN (T=2)	59.26	63.46	4.15	60.11	49.96	17.53
ReDAN (T=3)	**59.32**	**64.21**	**4.05**	**60.47**	**50.02**	**17.40**

- Rank aggregation between Dis. & Gen. models boosts performance
- 2nd place in Visual Dialog Challenge 2019

Model	Ens. Method	NDCG	MRR	Mean
4 Dis.	Average	60.53	65.30	3.82
4 Gen.	Average	61.43	50.41	17.38
1 Dis. + 1 Gen.	Average	63.85	53.53	9.00
1 Dis. + 1 Gen.	Reciprocal	63.18	59.03	4.88
4 Dis. + 4 Gen.	Average	65.13	54.19	8.74
4 Dis. + 4 Gen.	Reciprocal	64.75	61.33	4.41
Diverse Ens.	Average	**67.12**	**56.77**	**5.96**

Question Type	All	Yes/no	Number	Color	Others
Percentage	100%	75%	3%	11%	11%
Dis.	59.32	60.89	44.47	58.13	52.68
Gen.	60.47	63.49	41.09	52.16	51.45