



Dual Attention Networks for Visual Reference Resolution in Visual Dialog



Gi-Cheon Kang¹, Jaeseo Lim¹, Byoung-Tak Zhang^{1,2}

¹ Interdisciplinary Program in Cognitive Science

² School of Computer Science and Engineering
Seoul National University

What is Visual Dialog?

The logo for Visual Dialog is a large rectangle with a blue header bar at the top. The header bar contains the text "Visual Dialog" in white. The rest of the rectangle is white and empty.

Visual Dialog

What is Visual Dialog?

Visual Dialog



What is Visual Dialog?

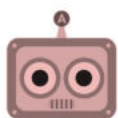
Visual Dialog



A man and a woman are holding umbrellas

What is Visual Dialog?

Visual Dialog



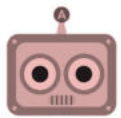
A man and a woman are holding umbrellas

What color is his umbrella?



What is Visual Dialog?

Visual Dialog



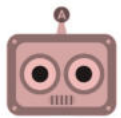
A man and a woman are holding umbrellas

What color is **his** umbrella?



What is Visual Dialog?

Visual Dialog



A man and a woman are holding umbrellas

What color is his umbrella?

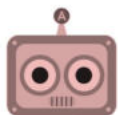


What is Visual Dialog?

Visual Dialog



A man and a woman are holding umbrellas



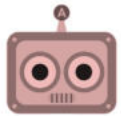
His umbrella is black

What color is his umbrella?

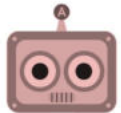


What is Visual Dialog?

Visual Dialog



A man and a woman are holding umbrellas



His umbrella is black

What color is his umbrella?




What about hers?



What is Visual Dialog?

Visual Dialog



A man and a woman are holding umbrellas


His umbrella is black

What color is his umbrella?

What about hers?

What is Visual Dialog?

Visual Dialog



A man and a woman are holding umbrellas

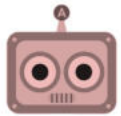
His umbrella is black

What color is his umbrella?

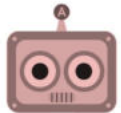
What about hers?

What is Visual Dialog?

Visual Dialog



A man and a woman are holding umbrellas



His umbrella is black



Hers is multi-colored

What color is his umbrella?



What about hers?



▪
▪
▪

What is Visual Dialog?

Given

- image I



What is Visual Dialog?

Given

- image I
- dialog history

$$H = (C, (Q_1, A_1^{gt}), \dots, (Q_{t-1}, A_{t-1}^{gt}))$$



C : A man and a woman are holding umbrellas.

Q_1 : What color is his umbrella?

A_1^{gt} : His umbrella is black.

Q_2 : How about hers?

A_2^{gt} : Hers is multi-colored

What is Visual Dialog?

Given

- image I
- dialog history

$$H = (C, (Q_1, A_1^{gt}), \dots, (Q_{t-1}, A_{t-1}^{gt}))$$

- follow-up question Q_t



C : A man and a woman are holding umbrellas.

Q_1 : What color is his umbrella?

A_1^{gt} : His umbrella is black.

Q_2 : How about hers?

A_2^{gt} : Hers is multi-colored

Q_t : How many other people are in the image?

What is Visual Dialog?

Given

- image I
- dialog history

$$H = (C, (Q_1, A_1^{gt}), \dots, (Q_{t-1}, A_{t-1}^{gt}))$$

- follow-up question Q_t

Predict natural language answer A_t^{pred}

- $A_t = \{A_t^1, \dots, A_t^{100}\}$

$$A_t^{gt}, A_t^{pred} \in A_t$$



C : A man and a woman are holding umbrellas.

Q_1 : What color is his umbrella?

A_1^{gt} : His umbrella is black.

Q_2 : How about hers?

A_2^{gt} : Hers is multi-colored

Q_t : How many other people are in the image?

A_t^{pred} : I think 3. They are occluded.

What is Visual Dialog?

Given

- image I
- dialog history

$$H = (C, (Q_1, A_1^{gt}), \dots, (Q_{t-1}, A_{t-1}^{gt}))$$

- follow-up question Q_t

Predict natural language answer A_t^{pred}

- $A_t = \{A_t^1, \dots, A_t^{100}\}$

$$A_t^{gt}, A_t^{pred} \in A_t$$

Key challenges

- capturing temporal topics
- finding visual groundings of linguistic expressions



C : A man and a woman are holding umbrellas.

Q_1 : What color is his umbrella?

A_1^{gt} : His umbrella is black.

Q_2 : How about hers?

A_2^{gt} : Hers is multi-colored


Q_t : How many other people are in the image?

A_t^{pred} : I think 3. They are occluded.

What is visual reference resolution?

Visual reference resolution

Image



Question

Does *it* look like a nice one?

Dialog History

A young man jumping his skateboard on a ramp

Where is he? In a skate park

Can you see his *skateboard*? Yes

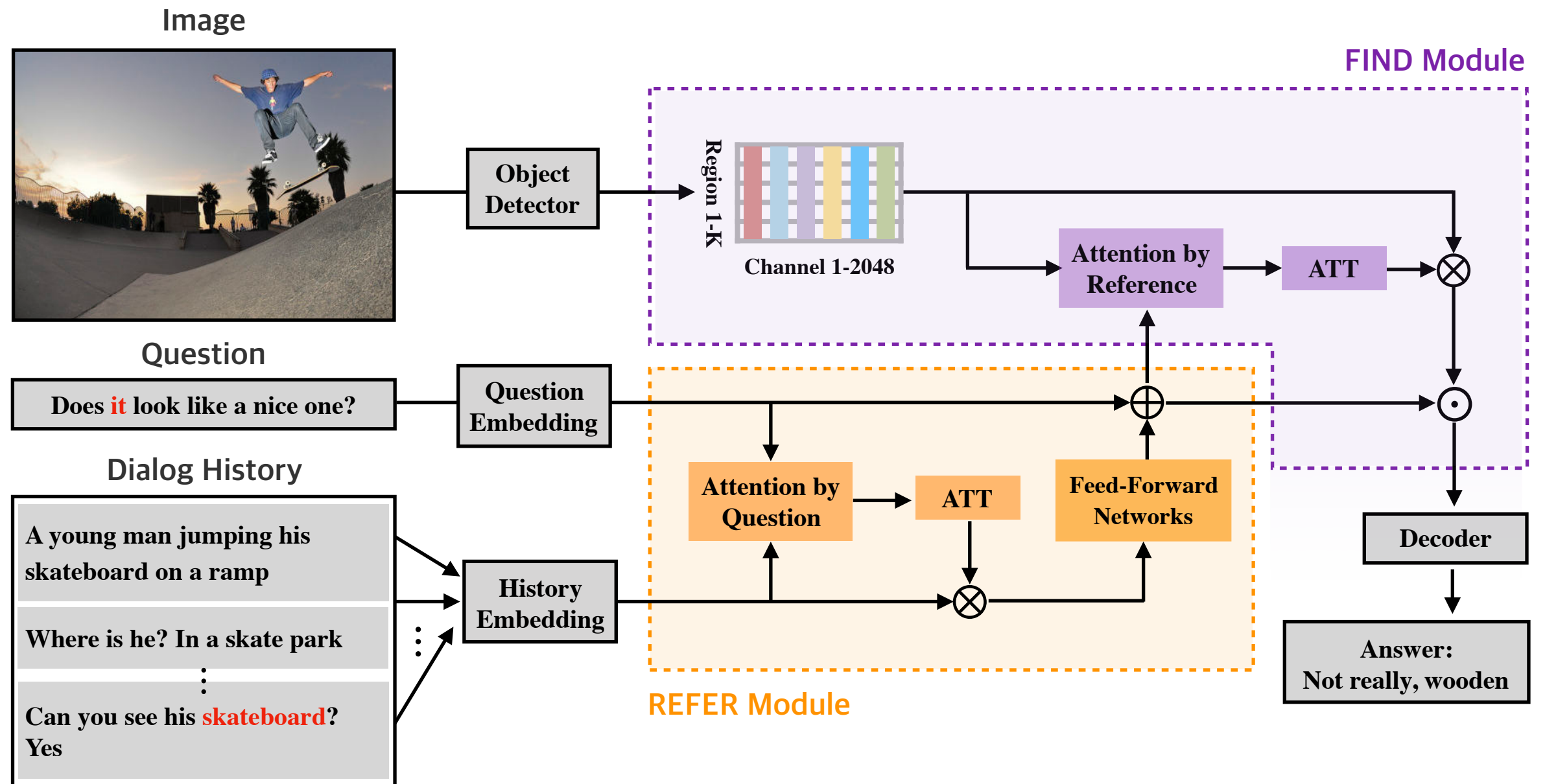
- links the reference to an entity in the visual source (e.g., image and video).
- source is the natural language and the target is an image (multi-modal).
- dialog history gives a cue to refer.

Co-reference resolution

"I voted for Donald because he was most aligned with my values," she said.

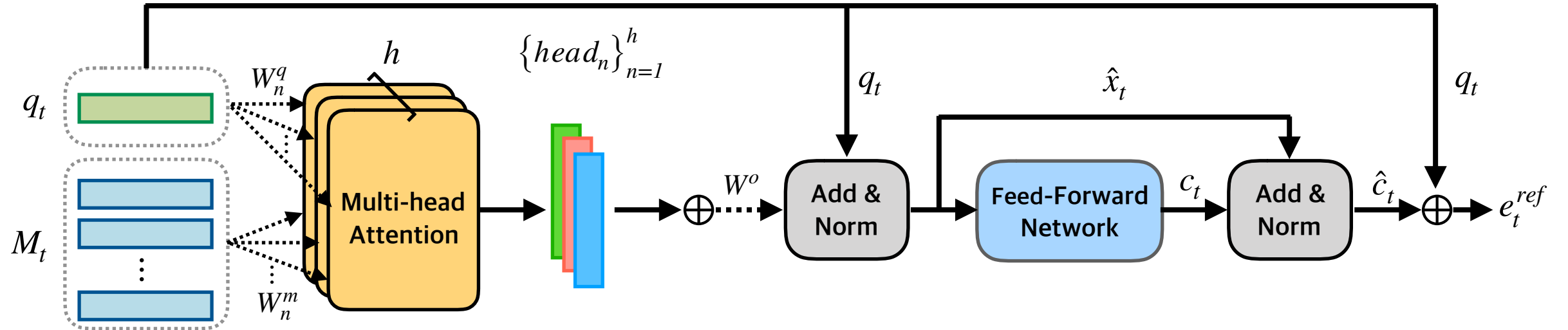
- clustering noun phrases and pronouns, which refer to the same entity.
- source and target are both natural language (uni-modal).

Dual Attention Networks



- We propose two attention modules, REFER and FIND.
- REFER retrieves relevant dialog history using multi-head attention mechanism, and FIND performs visual grounding using bottom-up attention mechanism.

REFER module



$$head_n = \text{Attention}(\underbrace{q_t W_n^q}_{1 \times d_{\text{ref}}}, \underbrace{M_t W_n^m}_{t \times d_{\text{ref}}})$$

$$\text{Attention}(a, b) = \text{softmax}\left(\frac{ab^\top}{\sqrt{d_{\text{ref}}}}\right)b$$

$$x_t = (head_1 \oplus \dots \oplus head_h) W^o$$

$$\hat{x}_t = \text{LayerNorm}(x_t + q_t)$$

$$c_t = \text{ReLU}(\hat{x}_t W_1^f + b_1^f) W_2^f + b_2^f$$

$$\hat{c}_t = \text{LayerNorm}(c_t + \hat{x}_t)$$

$$e_t^{\text{ref}} = \hat{c}_t \oplus q_t$$

- REFER module computes multi-head attention over all previous dialogs in a sentence-level fashion, followed by feed-forward networks to get the *reference-aware representations*.
- From this pipeline, we expect REFER to be capable of *question disambiguation*.
- q_t and M_t denote question feature and dialog history feature, respectively. W_n^q , W_n^m , W^o and W^f denote linear projection matrices. \oplus denotes concatenation operation.

FIND module

- FIND module performs visual grounding with respect to the reference-aware representations (i.e., the output of REFER). Also, FIND returns the joint embeddings of the multi-modal inputs. We used Faster-RCNN (Ren et. al., 2015) pre-trained with Visual Genome (Krishna et. al., 2017) to extract the object-level image features.
- $f_v(\cdot)$, $f'_v(\cdot)$, $f_{ref}(\cdot)$ and $f'_{ref}(\cdot)$ denote the two-layer MLP.

Visual attention

$$r_t = \underbrace{f_v(v)}_{K \times d_{\text{find}}} \odot \underbrace{f_{ref}(e_t^{ref})}_{1 \times d_{\text{find}}}$$

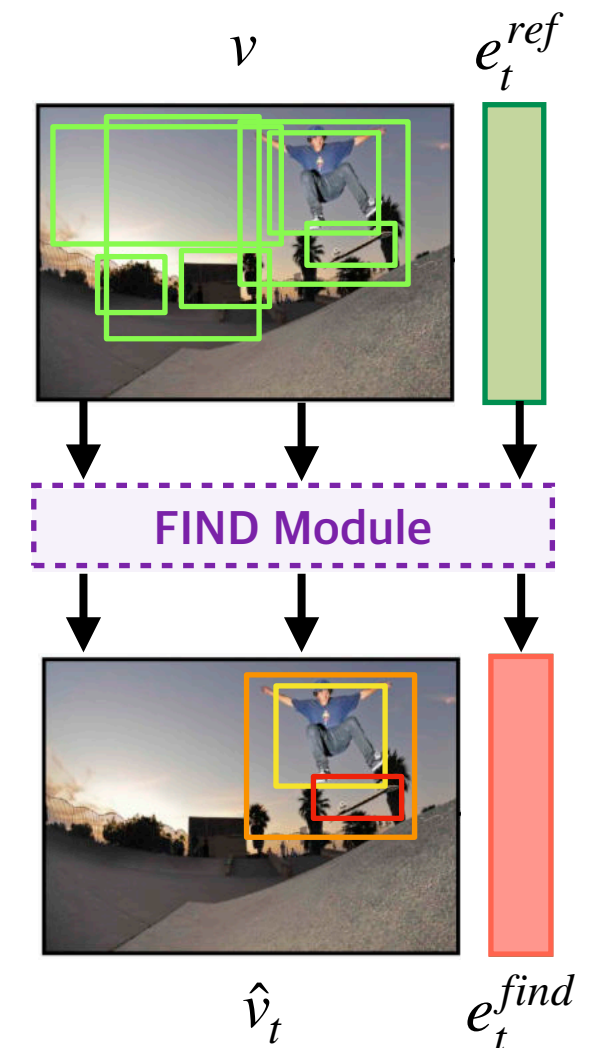
$$\alpha_t = \text{softmax}(r_t \underbrace{W^r}_{d_{\text{find}} \times 1} + b^r)$$

$$\hat{v}_t = \sum_{j=1}^K \alpha_{t,j} v_j$$

Joint embeddings

$$z_t = \underbrace{f'_v(\hat{v}_t)}_{1 \times d_{\text{find}}} \odot \underbrace{f'_{ref}(e_t^{ref})}_{1 \times d_{\text{find}}}$$

$$e_t^{find} = z_t W^z + b^z$$



Answer decoder

- Answer decoder computes each score of candidate answers via a dot product with the embedded representation e_t^{find} , followed by a softmax activation to get a categorical distribution over the 100-candidates.

$$p_t = \text{softmax}(\overset{1 \times L}{\underbrace{e_t^{find}}_{1 \times L}} \overset{L \times 100}{\underbrace{O_t^T}_{L \times 100}})$$

- In training phase, DAN is optimized by minimizing the *cross-entropy loss* between the one-hot encoded label vector (i.e., y_t) and the probability distribution (i.e., p_t).

$$\mathcal{L}(\theta) = - \sum_k y_{t,k} \log p_{t,k}$$

- In test phase, the list of candidate answers is ranked by the distribution p_t , and evaluated by the given metrics.

Evaluation metrics

Mean Reciprocal Rank (MRR) -
$$\text{MRR} = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{\text{rank}_i^{gt}}$$

Recall@k, $k \in \{1, 5, 10\}$ - existence of ground truth answer in top-k ranked list

Mean Rank (Mean) - mean rank of the ground truth answer

Normalized Discounted Cumulative Gain (NDCG) - answer *relevance*

Answer options : ["two", "yes", "probably", "no", "yes it is"]

Ground-truth relevances : [0, 1.0, 0.5, 0, 1.0] (collecting dense annotations)

Ideal ranking of answer options : ["yes", "yes it is", "probably", "two", "no"]

Submitted ranking of answer options : ["yes", "yes it is", "two", "probably", "no"]

$$\text{NDCG} = \frac{DCG_{submitted}}{DCG_{ideal}} \approx \frac{1.63}{1.88} \approx 0.87 \quad DCG = \sum_{j=1} \frac{\text{relevance}_j}{\log_2(j+1)}$$

NDCG penalizes the lower rank of candidates with high relevance scores

Quantitative results

Results on semantically complete (SC) & incomplete (SI) questions

- pronouns : it, its, they, their, them, these, those, this, that, he, his, him, she, her
- check the contribution of the *reference-aware representations* for the SC, SI questions, respectively.

Performance

Model		MRR	R@1	R@5	R@10	Mean
SC	No REFER	61.85	47.80	79.10	88.43	4.49
	DAN	64.81	51.22	81.63	90.19	4.03
	Improvements	2.96	3.42	2.53	1.76	0.46
SI	No REFER	58.44	44.38	75.36	85.48	5.36
	DAN	61.77	48.13	78.43	87.81	4.70
	Improvements	3.33	3.75	3.07	2.33	0.66

Table 3: VisDial v1.0 validation performance on the semantically complete (SC) and incomplete (SI) questions. We observe that SI questions obtain more benefits from the dialog history than SC questions.

Observations

1. DAN shows significantly better results than No REFER model for SC questions.
2. SI questions obtain more benefits from the dialog history than SC questions.
3. A dialog agent faces greater difficulty in answering SI questions than SC questions

Quantitative results

Results on VisDial v1.0 and v0.9 datasets

- DAN outperforms all other approaches on NDCG, MRR, and R@1, including the previous state-of-the-art method. The result indicate that our proposed model ranks higher than all other methods on both single ground-truth answer (R@1) and all relevant answers on average (NDCG)

Performance

	VisDial v1.0 (test-std)						VisDial v0.9 (val)				
	NDCG	MRR	R@1	R@5	R@10	Mean	MRR	R@1	R@5	R@10	Mean
LF (Das et al., 2017)	45.31	55.42	40.95	72.45	82.83	5.95	58.07	43.82	74.68	84.07	5.78
HRE (Das et al., 2017)	45.46	54.16	39.93	70.45	81.50	6.41	58.46	44.67	74.50	4.22	5.72
MN (Das et al., 2017)	47.50	55.49	40.98	72.30	83.30	5.92	59.65	45.55	76.22	85.37	5.46
HCIAE (Lu et al., 2017)	-	-	-	-	-	-	62.22	48.48	78.75	87.59	4.81
AMEM (Seo et al., 2017)	-	-	-	-	-	-	62.27	48.53	78.66	87.43	4.86
CoAtt (Wu et al., 2018)	-	-	-	-	-	-	63.98	50.29	80.71	88.81	4.47
CorefNMN (Kottur et al., 2018)	54.70	61.50	47.55	78.10	88.80	4.40	64.10	50.92	80.18	88.81	4.45
RvA (Niu et al., 2018)	55.59	63.03	49.03	80.40	89.83	4.18	66.34	52.71	82.97	90.73	3.93
Synergistic (Guo et al., 2019)	57.32	62.20	47.90	80.43	89.95	4.17	-	-	-	-	-
DAN (ours)	57.59	63.20	49.63	79.75	89.35	4.30	66.38	53.33	82.42	90.38	4.04

Table 1: Retrieval performance on VisDial v1.0 and v0.9 datasets, measured by normalized discounted cumulative gain (NDCG), mean reciprocal rank (MRR), recall @k (R@k), and mean rank. The higher the better for NDCG, MRR, and R@k, while the lower the better for mean rank. DAN outperforms all other models across NDCG, MRR, and R@1 on both datasets. NDCG is not supported in v0.9 dataset.

Qualitative results



Question : Does he have a racket?

Answer : Yes

Dialog History

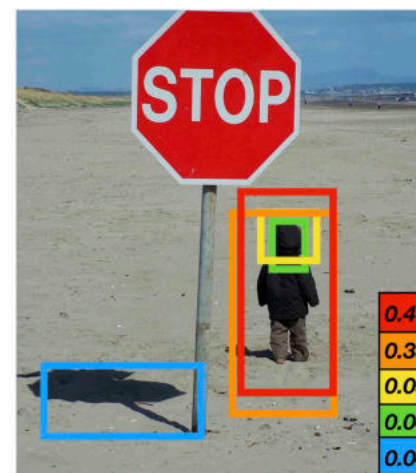
The tennis player is sitting next to the court

Is the tennis player male or female? Male

Are they sitting on a bench? Yes

How old is the male? 30s

What color is the bench? Green



Question : Is she alone or with someone?

Answer : Alone

Dialog History

A small child on the beach, walking past a stop

How old is the child? I 'd say 2

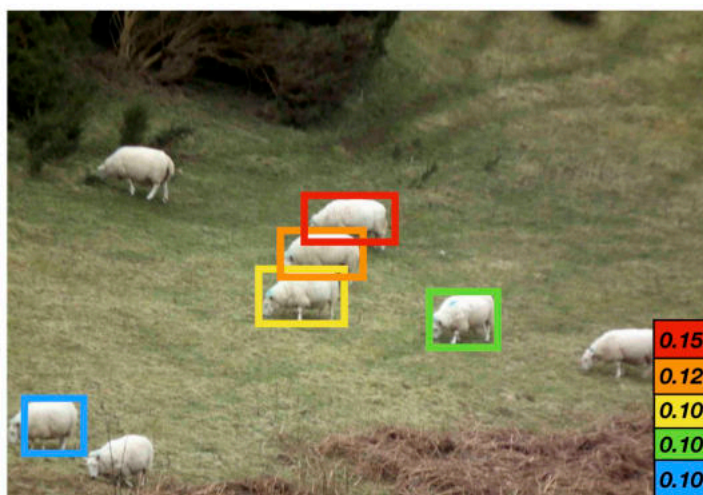
Is the child wearing a bathing suit?

No, a winter coat

What color hair does she have?

Covered by the coat hood

What color is the winter coat? Black



Question : Are they all white?

Answer : Yes

Dialog History

A herd of sheep grazing in a grassy field

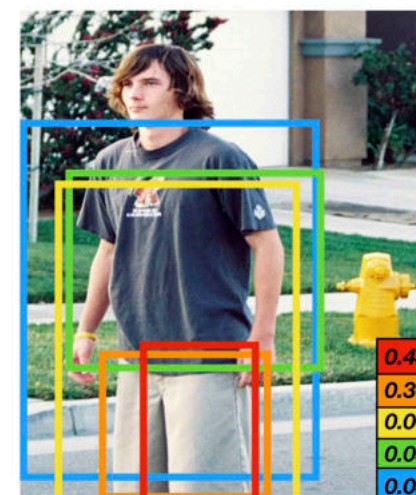
do you see any people? No

Do you see trees? Yes

Is this a big field? Yes

Do you see a fence? No

How many sheep are there? 8



Question : Is he wearing jeans?

Answer : Yes, he is

Dialog History

A man walks down the street, pass a yellow fire hydrant

Is this man young? I'd say early 20s

Is he dressed casual? Yes he is

Is his hair short or longish? It is shoulder length

what color is his shirt? It is medium brown

Thank You

Questions?

paper: arxiv.org/pdf/1902.09368.pdf
code: github.com/gicheonkang/DAN-VisDial