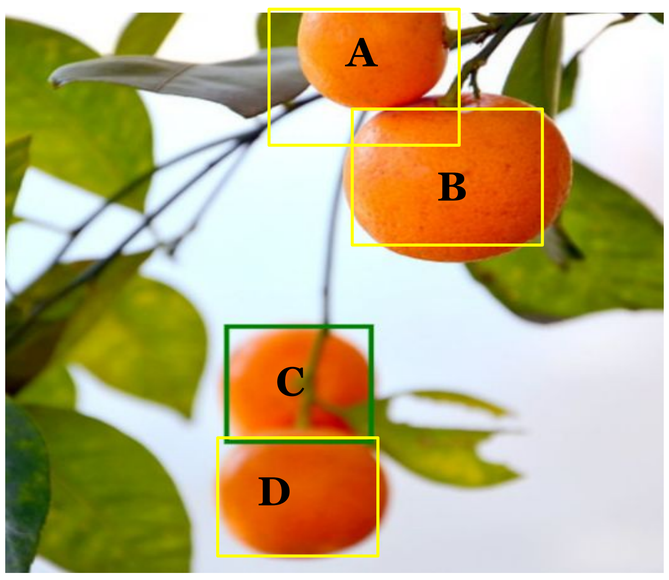
Visually Grounded Follow-up Questions: A Dataset of Spatial Questions Which Require Dialogue History

Dota - Tianai Dong[1], Alberto Testoni[2] , Luciana Benotti[3] , Raffaella Bernardi[1,2]
[1]CIMeC, University of Trento, Italy
[2] DISI, University of Trento, Italy
[3] Universidad Nacional de Córdoba, CONICET, Argentina

With the support of the Erasmus+ Programme of the European Union

Visually Grounded Follow-up Questions



Questioner (Oracle)

1. Is it a fruit? (Yes)
2. Is it in the foreground? (No)
3. Are there two of them on the branch? (Yes)
4. Is it the top one? (Yes)

How to collect more history-dependent questions?

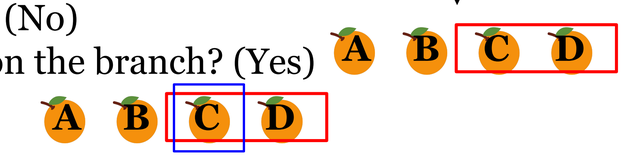
↓

Question under Discussion (QuD) [Ginzburg, 2012]

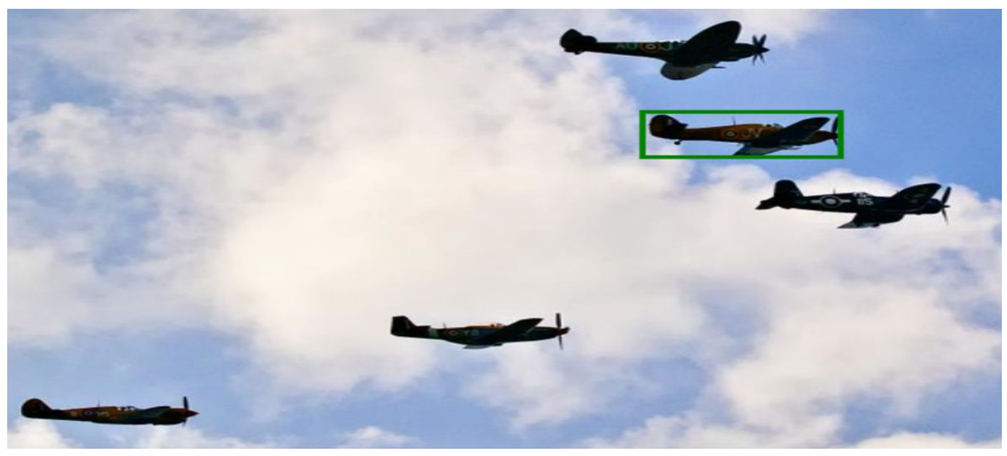
↓

Objects conjectured to be the target

↓



Q3 -> Trigger Q idenities a group of objects with shared property
Q4 -> Zoomer Q focuses on one of members in the group



Questioner (Oracle)

1. Is it a plane? (Yes)
2. One of the 3 planes in front? (Yes) [Trigger]
3. Is it the one in the middle? (Yes) [Zoomer]

To answer Q3 (the zoomer), "middle" needs to be interpreted relatively to the group of 3 planes defined by Q2 (the trigger), instead of the middle of image (without Q2)

Data Extraction & Model Evaluation

Guesswhat!? Dataset → Keyword matching → Automatically extracted sets
Search for Trigger-Zoomer → Manually filtered sets

→ Models:
- LSTM & LSTM-DH (language-only baseline model)
- V-LSTM & V-LSTM-DH (LSTM with visual features)
- LXMERT & LXMERT-DH (**SOTA**; Multimodal Encoder)

Trigger Q: Color ["Is it blue?"] or Group ["One of the three oranges?"] (Shekhar et al. (2019))
Zoomer Q: Group or Absoulte ["In the middle"] (Testoni et al. (2020))

Results

Dataset

	Context-dependent Group		Context-dependent Absolute	
	Group-Group	Color-Group	Group-Absolute	Color-Absolute
Automatically Extracted	364	145	530	389
Manually Filtered	103	61	107	insufficient

Task Accuracies (Automatically extracted sets; Manually filtered sets)

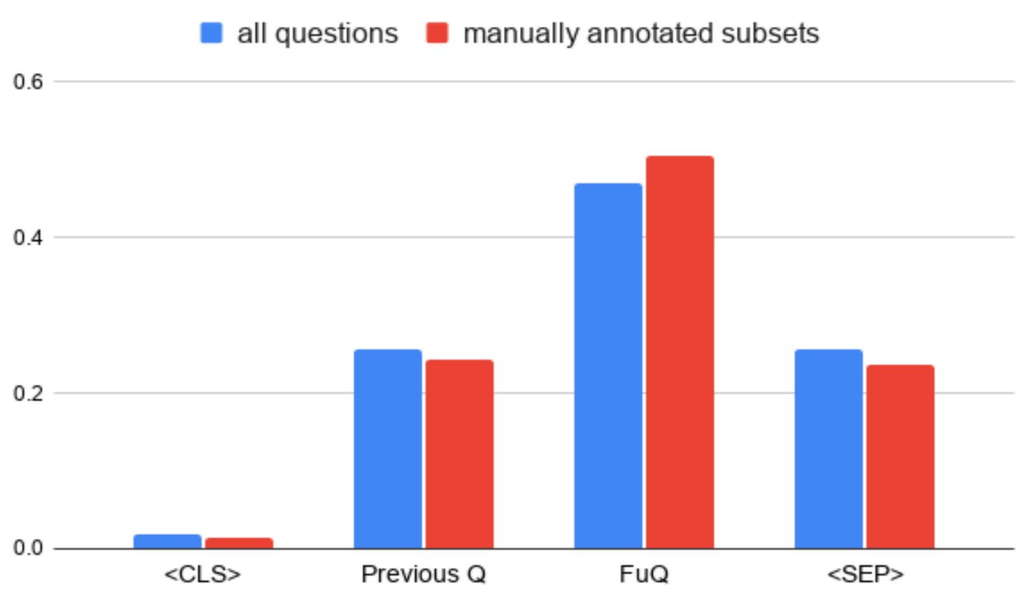
	Controlled sets			Context Depedent	
	All	Absolute	Group	Absolute	Group
LSTM	77.31	70.45	67.11	60.57*	59.39
LSTM-DH	77.88	71.03	67.75	64.09*	59.06
V-LSTM	74.65	70.87	67.42	58.43*	62.15
V-LSTM-DH	73.82	70.13	65.12	63.33*	62.27
LXMERT	82.40	79.42	74.48	74.21	70.01
LXMERT-DH	82.79	80.19	74.49	74.43	71.12

	Group-Group	Color-Group	Group-Absolute
LXMERT	65.84	71.58	76.95
LXMERT-DH	66.34	74.32	76.63

Confidence of models

	Group-Group	Color-Group	Group-Absolute
LXMERT	65.84	71.58	76.95
LXMERT-DH	66.34	74.32	76.63

LXMERT-DH attention



Conclusions

- We define a novel methodology for extracting history-dependent spatial questions from visual dialogues, "trigger-zoomer".
- We evaluate our "trigger-zoomer" methodology on the Guesswhat?! dataset.
- We use our dataset to evaluate SOTA multimodal encoders and show that there is for improvement for answering history-dependent questions

References

[1] Ginzburg,J. 2012. The Interactive Stance. Oxford Press.
[2] Testoni, A.,et al., *They are not all alike: answering different spatial questions requires different grounding strategies*
[3] Shekhar, R., et al, *Jointly learning to see, ask, decide when to stop, and then GuessWhat.*

VISPA

SCAN ME

Dataset