Video Dialog as Conversation about Objects Living in Space-Time: Supplementary

Hoang-Anh Pham[1], Thao Minh Le[1], Vuong Le[1], Tu Minh Phuong[2], Truyen Tran[1]

[1]Applied Artificial Intelligence Institute, Deakin University, Australia
[2]Posts and Telecommunications Institute of Technology, Vietnam
[1]{phamhoan, thao.le, vuong.le, truyen.tran}@deakin.edu.au
[2]phuongtm@ptit.edu.vn

1 Additional Details

1.1 DGCN

The DGCN introduced in [2] is a multi-layer graph neural network that aims to refine an input matrix $Z \in \mathbb{R}^{N\times d}$ given the adjacency matrix $\mathcal{K}_t$:

$$\bar{Z} = \text{DCGN}\left(Z; \mathcal{K}_t\right) \in \mathbb{R}^{N\times d} \tag{1}$$

Starting with the initialization $H^0 = Z \in \mathbb{R}^{N\times d}$, the new representations are refined by:

$$H^i = \text{ReLU}\left(H^{i-1} + \text{GCN}_i\left(H^{i-1}; \mathcal{K}_t\right)\right) \in \mathbb{R}^{N\times d}, \tag{2}$$

where GCN is standard Graph Convolutional Network [1]:

$$\text{GCN}_i\left(H^{i-1}; \mathcal{K}_t\right) = \left(\text{ReLU}\left(\mathcal{K}_t H^{i-1} W_A^i\right)\right) W_2^{i-1}. \tag{3}$$

The skip connection in Eq. (2) allows multiple layers of GCN without the gradient vanishing problem with out it, similar to what is found in ResNet. After a fixed number of I GCN iterations, we have the refined representation of Z: $\bar{Z} \leftarrow H^I$.

1.2 Gating in Answer Generator

Recall from the main text that v_4 is partial semantic embedding of the partial answer, and $Q_t \in \mathbb{R}^{L_Q\times d}$ is the embedding of the question. We extract the relevant information from the question w.r.t the current partial answer as

$$v_5 = \sum_{i=1}^{L_Q} Q_{i,t} * \text{Attn}\left(v_4, Q_t, Q_t\right) \in \mathbb{R}^{1\times d},$$

where $Q_{i,t}$ is the i-th question word embedding, and $*$ is element-wise multiplication. The combination of attention and multiplicative interaction amplifies the question words that are related to the current answer semantic.

Table 1. Hyper parameters have tried.

	Settings
Representation dims (d)	$\{\mathbf{128}, 256, 512\}$
Number DGCN Block (N_{dgcn})	$\{1, \mathbf{3}, 6\}$
Dropout rate ($dropout$)	$\{0.1, \mathbf{0.15}, 0.2\}$
Self-Attention in Answer Generator	
Hidden representation dims(d_{ff})	$\{\mathbf{512}, 1024, 2048\}$
Number stacked blocks (N_b)	$\{1, \mathbf{3}, 6\}$
Number head (N_h)	$\{2, \mathbf{4}, 8\}$

This is then combined with the current semantic to estimate the gating function:

$$\alpha = \text{sigmoid}\left(\text{Linear}\left([v_4, v_5, a_l]\right)\right).$$

where a_l is the embedding of last word in the current answer.

2 Additional Implementation Details

2.1 Data Preparation and Pre-processing

As described in the main part, we employed Faster R-CNN (code and pre-trained weight are from Detectron2 [5]) for object detection and combined with DeepSORT[4] for objects tracking through each video. With AVSD dataset, we set the threshold for NMS to 0.3, and the minimum confidence of detected object to 0.2.

2.2 Training and Validation

To improve generalization, Label Smoothing [3] is applied on the ground-truth answer $A^*_{1:T}$. For hyperparameter tuning we use the loss score on the validation set. The loss is fast to compute compared to accuracy because generating free-form response takes time producing one word at a time. Table 1 show hyperparameters that we have tried, the bold number has the best score. The chosen model has approximately 4M parameters in total.

2.3 Testing and Inference

In testing, we use the beam search strategy with beam size of 3 for the reported results. Table 2 shows the results on AVSD@DSTC7 with different beam sizes when inferencing our model.

3 Additional Ablation Studies

Table 3 and Table 4 compare the results between our model and other baselines on the FVS and LDS subsets that show performance degradation as reported in the main paper.

Table 2. Impact of beam size on AVSD@DSTC7.

beam size	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
1 (greedy)	0.715	0.580	0.476	0.395	0.260	0.558	1.075
3	**0.723**	**0.589**	**0.483**	**0.400**	**0.266**	**0.561**	**1.085**
5	0.719	0.586	0.481	0.399	0.2665	0.560	1.079

Table 3. Experimental results on FVS

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
MTN	0.575	0.394	0.262	0.170	0.183	0.417	0.412
BiST	0.609	0.433	0.292	0.193	0.191	0.438	0.456
COST	**0.627**	**0.447**	**0.309**	**0.211**	**0.201**	**0.452**	**0.545**

4 Additional Qualitative Results

We demonstrated more examples of the attended interactions between objects from our model in Fig. 1. Consider the first row consisting of three pictures of a scene in which a man interacts with two objects – the picture and the broom. The first question "How does the man start the video" has groundtruth answer "The man walks in holding a picture". Although our COST doesn't know the answer, it does highlight the relationship between "person" and "picture". Likewise, the second question "What does he do with the picture" has the groundtruth answer "The man sets the picture down and grabs the broom". Here the relationships are highlighted between "person" and "picture", and between "person" and "broom". The correct recognition of relationships, on which COST is never be trained explicitly, is important for selecting the right answer.

5 Source Code

The source code of our model including preprocessing, training, inferencing, evaluation code is attached with this document. Please refer to README in order to know how to use of our code. This will be made public in case the paper is accepted.

References

1. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. ICLR (2017)

Table 4. Experimental results on a LDS

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
MTN	0.664	0.525	0.419	0.341	0.246	0.526	0.920
BiST	0.717	0.578	0.473	0.389	0.258	0.552	1.025
COST	**0.723**	**0.590**	**0.485**	**0.402**	**0.266**	**0.560**	**1.086**

2. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Dynamic language binding in relational visual reasoning. In: IJCAI. pp. 818–824 (2020)
3. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
4. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP. pp. 3645–3649. IEEE (2017)
5. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)

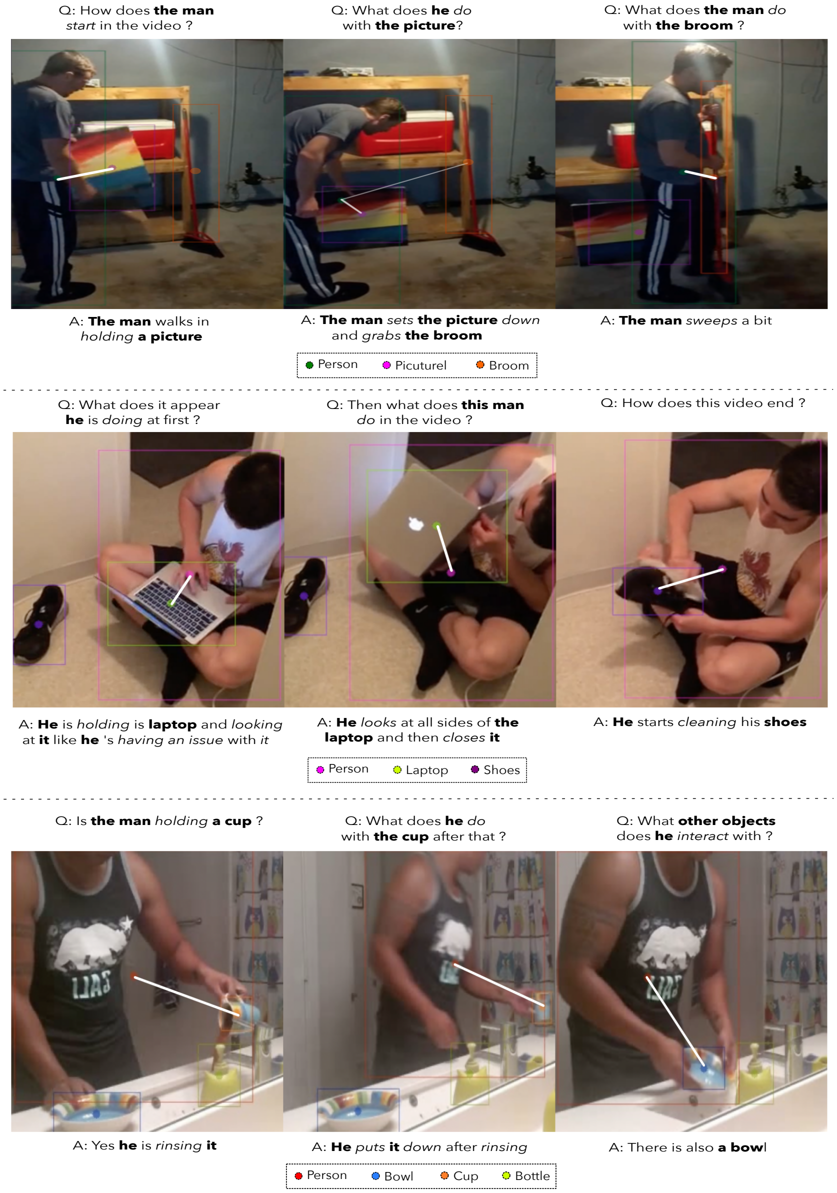
Q: How does the man start in the video ?
Q: What does he do with the picture?
Q: What does the man do with the broom ?
A: The man walks in holding a picture
A: The man sets the picture down and grabs the broom
A: The man sweeps a bit
● Person ● Picuturel ● Broom
Q: What does it appear he is doing at first ?
Q: Then what does this man do in the video ?
Q: How does this video end ?
A: He is holding is laptop and looking at it like he 's having an issue with it
A: He looks at all sides of the laptop and then closes it
A: He starts cleaning his shoes
● Person ● Laptop ● Shoes
Q: Is the man holding a cup ?
Q: What does he do with the cup after that ?
Q: What other objects does he interact with ?
A: Yes he is rinsing it
A: He puts it down after rinsing
A: There is also a bowl
● Person ● Bowl ● Cup ● Bottle

Fig. 1. Each frame/question pair (left to right) represents a dialog turn, where the frame is selected to reflect the moment being queried. The visibility of edges denotes the relevance of visual objects' relations to the questions and desired answers. Detected objects are named by Faster R-CNN. Samples are taken from DSTC7 test split – Best viewed in colors.