Object-level Non-local operation for Visual Dialog

Sungjin Park, Heuiseok Lim,

Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea,
genom1324@gmail.com, limhseok@korea.ac.kr

Abstract. Visual dialog is a challenging vision-language task in which a series of questions visually grounded by a given image are answered. In this paper, we propose object-level non-local operation for visual dialog.

Keywords: Visual Dialog

1 Introduction

Visual Dialog is a task proposed by [1], requiring the dialog agent to answer the current question exploiting the image and previous dialog history. A dialog agent is given a set of multimodal inputs for each dialog turn, which consists of an image ***I***, a current question ***Q***, a set of dialog history ***H***, which consists of image caption and consecutive question-answer pairs, and a set of answer candidates ***A***. The agent then is required to answer the question by either discriminating or generating a correct answer.

2 Object-wise non-local operation (ONO)

Late-fusion model which is the baseline of Visual Dialog does not use relation information between visual features. We assumed that the underlying relations between image features could serve as an important factor. Specifically, in the field of NLP, just as the same word has different meanings depending on position and context, the image object may also have different meanings by the different objects that make up the scene. Therefore, to implicitly learn the relations between image features, we experimented with non-local operation [2] additionally.

3 Experimental Setup

Datasets We use visdial v1.0 dataset to evaluate our proposed method.
Evaluation Metrics. We evaluate our proposed model on retrieval metrics, following the work of [1]: (1) mean rank of ground truth response, (2) recall at k (k= {1, 5, 10}); denoted as R@k, evaluating where ground truth is positioned in the sorted list; and (3) mean reciprocal rank (MRR) [3]. NDCG is also introduced as a main metric in VisDial v1.0 dataset, and penalized when the model predicts candidate answers with high relevance scores as low rankings.

4 Results

Table 1. Results of ONO on VisDial v1.0(val).

Model	NDCG	MRR	R@1	R@5	R@10	Mean
LF [1]	55.05	60.69	46.48	78.23	87.59	4.72
LF + ONO	*58.76*	63.05	49.20	80.05	89.37	4.27

We conducted ablation studies on the visdial v1.0 to evaluate the influence of the additional method in LF [1]. ONO brings significant performance improvements in the LF model. These results indicate the relation between visual contents is helpful for answering the questions.

5 Conclusion

In this paper, we introduced an additional method for improving the Visual Dialog task. ONO is model-agnostic, and thus can be applied to most existing methods and boost significant improvement.

6 Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICT Creative Consilience program(IITP-2020-0-01819) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

References

1. Das, Abhishek, et al. "Visual dialog." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
2. Wang, Xiaolong, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
3. Voorhees, Ellen M. "The TREC-8 question answering track report." Trec. Vol. 99. 1999.