MeetUp! A Task For Modelling Visual Dialogue

1 Abstract

After achieving impressive success representing image content textually (as done by image captioning models (Fang et al., 2015; Devlin et al., 2015; Chen and Lawrence Zitnick, 2015; Vinyals et al., 2015; Bernardi et al., 2016); and research in referring expression resolution and generation (Kazemzadeh et al., 2014; Mao et al., 2015; Yu et al., 2016; Schlangen et al., 2016)), the Vision and Language community has recently established "Visual Dialogue" as the more challenging follow up task (Das et al., 2017; De Vries et al., 2017).

In that task, a Questioner, prompted by some textual information (a caption) can ask an Answerer questions about an image that only the latter sees. We argue here that this setup leads to an impoverished form of dialogue and hence to data that is not substantially more informative than captioning data, if the goal is to model visual *dialogue*.

In my talk I will describe our ongoing work on the MeetUp setting, where two players navigate separately through a visually represented environment, with the goal of being at the same location. This goal gives them a reason to describe visual content, leading to motivated descriptions, and the dynamic setting induces an interesting split between private and shared information.

Our pilot data collected in a small-scale pilot study indicates that MeetUp! dialogues consistently include plenty of phenomena found in human-human dialogue interaction: from crucial dialogue coherence / cohesion to factual descriptions of images containing many reasonable referring expressions.

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442, January.

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China, July. Association for Computational Linguistics.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*, Boston, MA, USA, June. IEEE.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg, 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing, Cham.