Jointly Learning to See, Ask and GuessWhat?!  
A. Venkatesh

Abstract

For a dialogue system to have a conversation with another speaker, it needs to understand and generate natural language, plan actions and keep track of the information exchanged. Amongst these abilities, tracking the dialogue state and maintaining a common ground established through the conversation are indispensable for the dialogue system to communicate effectively. The evaluation of general purpose dialogue systems is difficult as there is no quantitative evaluation measure for it. In contrast, within a task-oriented dialogue setting, it is possible to use task success as an evaluation metric.

We are interested in task-oriented visual dialogue, where a dialogue system has to ask a series of questions about a picture to obtain the necessary information to solve the task. More specifically, we develop a visual dialogue agent, Questioner, for the GuessWhat?! game. We model the agent using a modular neural network architecture which is capable of building a continuous representation of the dialogue state. The dialogue state of the agent combines the context from the linguistic and visual domains to condition the future actions. The agent is required to both ask informative questions about a picture and pick the target object in it. These two tasks are modelled by a single architecture which provides a multi-task learning setting to jointly learn the vision and language understandings with action and language generation. Thanks to its innovative design, this new architecture makes possible to fine-tune the parameters of its internal modules using cooperative learning. The proposed architecture for the Questioner agent trained by supervised and cooperative learning achieves state-of-the-art scores in the GuessWhat?! game. Furthermore, the modularity of the architecture allows us to add a decision-making component which in turn improves the quality of the dialogue by limiting the number of repeated questions.