CS539 - Embodied AI

Paper Review - Vision-and-Dialog Navigation

The paper introduces Cooperative Vision and Dialog Navigation, a dataset of over 2k embodied, human-human dialogs situated in simulated, photorealistic home environments. Defines the task of Navigation from Dialog History, given a target object and dialog history between humans cooperating to find that object, an agent must learn to infer navigation actions towards the goal in unexplored environments. The paper proposes an initial multi-modal sequence-to-sequence model to approach the above mentioned task.
The main contributions of the paper are the introduction of the Vision-and-Dialog Navigation idea, the dataset of Cooperative human dialogs (2k) to support training agents for the navigation task. Approach has been to use a sequence-to-sequence model to encode the entire dialog history as the initial state of the LSTM and interpret visual inputs as feature vectors using pre-trained ResNet-152. An LSTM decoder has been used to interpret the navigation action to be taken.
Results have been that when compared to non-learning agent and uni-modal baselines the model performs better. But when tested on unseen environments the performance has been disappointing. Probably due to the model's inability to understand the visual scene or interpret navigation instructions into actions.

Some strength of the paper are that this paper is the first to introduce the task of Navigation from Dialog History. They are the first to formulate and collect the dataset required for the task and introduce the first model which performs the task reasonably. The area is interesting. The experiments of comparing to unimodal baselines provided some insights into importance of difference components of the model. The community could benefit from this new line of work and for sure would see more improvements in models in this direction.

Weakness of the paper were the experiments, they did not justify their claim of additional dialog history improving performance significantly, infact the improvements were just marginal. It would have been better if they could do experiments showing where the model understood the visual scene or not, since the performance gap was high from seen to unseen environments, supposing that model did not understand the scene and that the ResNet-152 feature vector was not helping.