CS539 - Embodied AI

Paper Review - Gated-Attention Architectures for Task-Oriented Language Grounding

The paper addresses the problem of task-oriented language grounding, in which an agent is required to extract semantically meaningful representations of language and map it to visual elements, actions in a given environment to perform tasks specified via natural language instructions. The paper proposes a novel end-to-end trainable neural architecture for the mentioned task, develops a novel Gated-Attention mechanism for multimodal fusion of representations from verbal and visual modalities, paper also introduces new environment for training and evaluating models for the specified task which is built over ViZDoom. Experiments demonstrate that the Gated-Attention mechanism helps the model learn to associate attributes of the object mentioned in the instruction with the visual representations. Effectiveness of the proposed model to generalize well on unseen instructions as well as unseen maps has also been shown.

Strengths of the paper are that it proposes a novel architecture for task-oriented language grounding, the Gated-Attention mechanism which learns a joint state representation of language instruction, input image. The novelty of the gated-attention mechanism has been supported with experiments, comparison to baselines which just concatenate the individual representations rather than learning a joint state, results show a significant improvement when using gated-attention on episodes with a 'hard' setting, where in the agent has no view of the object initially forcing it to explore. Gated-Attention provides efficient exploration for the model and a better understanding of language instruction and its correlation to visual inputs. Attention vectors also support a learning of object attributes like color, size and type which is great.

Weakness of the paper were, they could have shown the path length comparisons to baselines, showing how effective their policy learning module is. The current graphs are based on accuracy which isn't always a good indicator of model performance. Experiments/plots supporting their argument of generalization to unseen environments is also missing, only training set comparison is shown to baselines of concat vs gated attention.

Reflections are that the paper is good, the novelty of the gated-attention mechanism to have a unified state representation of language instruction, visual input is novel and shows great promise. Improvements to the fusion module could help other works as well as it stores a better mapping relation.