Motivation and Overview

• Consider the reinforcement learning task where an agent with no prior knowledge is given a first-person view of a 3D environment and a task to execute.
• Training the agent to successfully carry out the task is called task-oriented language grounding.
• Because this is a sparse problem, we introduce curiosity pretraining into our model.
• Curiosity is an intrinsic reward signal that rewards the agent based on discovering new objects and parts of the environment.

Data

• We use a custom environment developed by Chaplot et al., which is based on VizDoom, a 3D game engine for reinforcement learning.
• We use 70 manually generated commands as our dataset, each containing a directive involving information from the tuple (object, attributes of object, action). 55 commands are used for the training set and 15 are used for the test set.
• The agent and five objects are randomly spawned in a small room, only one of which is the objective object.
• Because objects may spawn out of the field of view of the agent, the agent needs to be able to explore the environment in order to find the correct object.
• The test set contains commands which weren’t trained on, so the agent also needs to be able to extrapolate object attribute information such as color to new commands.

Models

• We built primarily on top of the gated-attention architecture of Chaplot et al.
• They use a CNN to process the input image and a GRU to process the natural language instruction and combine these two modalities using the GRU output as the attention weights for the CNN output.
• This output is then fed into the policy learning module for A3C, which consists of an FC layer, an LSTM, and two more FC layers which predict the value function for A3C and a distribution over the action space (turn left, turn right, move forward).

Results & Discussion

• We introduced a curiosity module similar to that in Pathak et al.
• The curiosity module consists of two separate neural networks.
• One predicts a distribution of the action taken at the last step given the last frame and current frame, while the other predicts the current frame given the previous frame and previous action.
• The difference between the predicted values and observed values serves as an intrinsic reward signal during pretraining.

References: