Abstract

A robot that can perform a task described using natural language instructions in a real-world environment has been a core robotics challenge. Vision-and-Language(VLN) is the task of guiding the robot across the real-world environment using natural language instructions. Navigational robots have a limited number of actions available such as Forward, Backward, Turn Right and Turn Left. The VLN agent has to interpret the natural language instructions and perform the actions while reasoning over the images gathered via the camera and avoid collisions while navigating. To follow the instructions, the robot needs to map the instructions to the relevant camera images using object detection. For example, to take right after the table, the robot has to find the table in the images using an object detection algorithm. The RxR Habitat Competition is a challenge that requires building the VLN agents using Habitat-Lab to follow natural language instructions to go across the rooms and reach the goal in a simulated environment using the Habitat Sim. Habitat-Sim is a simulator designed to simulate a 3D environment and can be accessed by Habitat-Lab API.
This thesis describes a baseline model and a state-of-the-art model for the development of a VLN agent which follows the instructions provided and navigates in the simulated scene. We also propose a VLN agent using a Transformer model which may work better than the state-of-the-art VLN agent. This thesis even explains the system constraints for the training environment of the VLN Task.
The VLN agents were implemented and tested using the Seq2Seq Model which is a baseline model containing a single recurrent network and the CMA Model which is more complex and includes two recurrent networks with attention layers in a 3D simulated environment. Usually, these kinds of agents are trained for a large number of episodes on the systems having 40+GB RAM and multiple 16GB GPU. For this dissertation, an average system having 25GB RAM and 16GB single GPU is used and agents were trained by reducing the number of episodes. The results show that even with reduced training, the agents can only carry out the VLN tasks which contain shorter instructions. The experiment also showed that to train the agent with the transformer model, a high-end system with more than 40GB RAM with multiple 16GB GPU is required.