16-824 Project Proposal: Visual Navigation for the THOR Dataset and Challenge

Anirban Ghosh
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
anirbang@andrew.cmu.edu

Siwei Zhu
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
szhu1@andrew.cmu.edu

Abstract

The AI2-THOR is a framework provides a platform to train machine learning algorithms intended to recognize and navigate real-world like environments. The platform provides high quality, photo-realistic renderings of common household scenes along with an accurate representation of the underlying physics phenomena that are found in the real world. This gives us an opportunity to train a machine learning model to be able to navigate a scene by interacting (issuing commands to move/turn) with the framework with the goal of obtaining a target.
This kind of interactive application is well suited for deep reinforcement learning. However, the challenge with existing methods mentioned in [1] is that these networks embed the target within the model, making the model target specific and therefore, the model will be unable ot generalize across targets. [1] proposes a method using an actor-critic model to overcome this problem by taking both the target and the scene as inputs to the network and mapping this to an action.
In this project we will try to implement a similar architecture using a convolution neural network for scene and target recognition and fully connected layers for generating the actions.

1. Introduction

We plan to participate in the THOR challenge (vuchallenge.org) for this project. The dataset used will be from the THOR dataset which consists of common household photo-realistic scenes generated by the THOR framework. The targets and scenes will be fed to our model, which will then generate a command based on these inputs. This command, which fed to the THOR framework will casuse the framework to render the next scene based on the movement command. The agent will then process the next frame and this process goes on iteratively till the agent arrives at the right location and orientation of the target image.
The THOR engine fixes the step size of forward and backward movements to 0.5m and the left and right movements to 90 degree turns. In order to model the randomness of the real world, gaussian noise is added to the inputs to the engine before generating the next frame for the agent to process. The framework also models the physics of the scene, so that the movements and interactions with objects (such as collisions) are accurately modeled. The model is written in TensorFlow, which is what we will be using for our development as well.

1.1. Implementation Plan

The main purpose for the project THOR is to train an agent to first recognize a "target" that we want it to find, then try to locate the target in different scenarios and finally navigate to the target as fast as possible. This makes the task well suited for a Deep Reinforcement Learning (DRL) approach, but it is hard to train a real robot in a real world scenario. Therefore, in [1], the authors developed one of the first realistic simulation frameworks with high-quality photo-realistic 3D scenes, called The House Of inteRactions (AI2-THOR). The dataset is basically a collection of high-quality 3D environments depicting common household scenes on which our agent can be trained in simulation. In this framework, we can collect and update parameters , which saves a lot of time and money during the training phase.

The training flow is described as follows: After we successfully get our agent to recognize the target object, the next phase will be to train the agent to locate the target and finally, to navigate in the direction of the target in the minimum number of steps possible. The task of location and navigation are linked to each other since the starting point of the simulation is unknown. Because the agent relies on visual data alone, this means that there is a chance that the target object is not in the field of view of the agent when the simulation begins. Therefore, in order to bring the target in the agent's field of view, it will have to move around the scene, searching for the target. Once the target is located, the agent will then have to navigate to the target in the shortest possible path, avoiding obstacles in the process. In order

to encourage the agent to minimize the path length and to avoid bumping into obstacles, we plan to penalize the agent if it collides with an object and also make each step it takes carry a penalty.

We foresee some potential challenges that we will have while coming up with an implementation for this project. The first is the challenge of designing a network structure that generalizes well over targets and scenes and effectively maps a scene to a motion. Since the next frame depends on the input of the agent in the current frame, there may also be a requirement for the network to take previous state information before making the next decision. Another challenge is how to properly define the reward function for deep reinforcement learning algorithm and, for other deep learning layers (such as convolution layers), how to define the loss function for back-propagation.

Our first step in this project will be to understand how to combine deep learning for visual data with reinforcement learning for interacting with the environment. We plan to speak with the professor regarding possible approaches and architectures for this network. Next, we will map the incoming target and scene image features to motion commands. This will trigger the THOR engine to render the next frame. Based on the outcome of this action, the reward will be updated with the appropriate penalties that were generated. Whether the network parameters are updated at each step of the training phase or after each training iteration has completed (i.e., either the agent has located the target or, possibly, when the total penalty has exceeded a certain threshold) is a choice that we are yet to make and need to gather more information for before we make. We plan to meet with the TA's and the professor in the coming weeks to get further clarity on the approaches and key design decisions.

1.2. References

[1] Y. Zhu, R. Mottaghi, E. Kolve, JJ. Lim, A. Gupta, L. Fei-Fei, A. Farhadi, "Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning", ICRA 2017