PAWS Navigation: Reinforcement Learning for Path Planning in Unknown Environment

Samantha Johnson

University of Maryland, College Park

ENPM690 – Robot Learning

# Abstract

words

# Introduction

PAWS (Psychological Assistance and Wellness System) is in development to be an in-home mental health robotic pet companion. This companion is geared toward helping individuals with moderate to severe mental health conditions who would benefit from constant supervision and stimulation from an external source. The target consumer is a young to middle aged adult, who may live alone or are in a situation where there is limited supervision, and who also suffers from a metal health disorder such as depression, anxiety, bipolar disorder, or PTSD. The goal of this product is to provide companionship, motivation, emotional stabilization, activity notifications, and emergency resources to aid in the recovery process for an individual suffering from a mental illness. The PAWS Bot hopes to speed up the recovery time and decrease the relapse rate for these individuals, as well as reduce the high number of suicides that occur amongst this group.

PAWS is a wheeled mobile pet-like robot that is capable of interacting with its human companion through various sensors. One of the main development points for the PAWS Bot is enabling its ability to safely navigate in the human’s home. The robot needs the capability to respond to a human’s distress signal and traverse to their location quickly to provide assistance. This navigation component is a complex problem to solve because the manufactured robot will have no information about the home in which it will exist. This means the robot needs to be trained ahead of time to be able to consistently react in an unknown environment. To address the complexity of needing to operate in an unknown environment, it was determined that a machine learning technique would be utilized to train the robot’s navigation component.

The machine learning method used for the assistance and navigation component of PAWS will be reinforcement learning. This will allow the robot to acquire feedback from its environment and human in order to make decisions on where it should go. The algorithm investigated for usage in the navigation component was Double-DQN, which combines the techniques of Q-learning with Neural Nets. The robot and simulation was created using Coppelia Sim (formerly VREM) and programmed using Python, through access with the simulation’s Remote API. It will be expected that in the real world the human the robot is assisting will be using a wearable technology to be able to transfer data that the robot will use for localization. The robot will also be equipped with several other sensors to facilitate its navigation and assistance abilities.

# Background and Related Work

Robot path planning is a highly researched field which has resulted in many path planning algorithms that enable a robot to autonomously navigate to a location. Some well known algorithms like Dijkstra and A\* utilize a discrete grid of the environment and mark the vertices in this map with a cost based on distance or a heuristic function. Both of these algorithms are complete, in that they are able to find the shortest path between the start and goal, which the robot can then travel. In contrast there are sampling based algorithms such as RRT (Rapidly-exploring Random Trees) which creates an expanding tree from the robot’s start position, until it reaches a goal. This algorithm does not necessarily find the optimal path, but will find a path to goal if one does exist. (<http://correll.cs.colorado.edu/?p=965> Path Planning page 75+)

The previously discussed path planning algorithms all have a major element in common; they require a known map of the environment. There are methods that can allow a robot to create a map of its environment for use in path planning, such as utilizing SLAM (Simultaneous Localization and Mapping), but for the PAWS robot it is desired that it can consistently act in an unknown environment without a map. There has been research done into methods that allow a robot to traverse an unknown environment but this field is a complicated and still evolving one. It has previously been proposed to use some application of Evolutionary and Hybrid algorithms to solve this path planning problem, such as Artificial Intelligence, Genetic Algorithms, and Fuzzy Logic, but there is minimal research in the application of Reinforcement Learning to solve this problem (mobile robots motion path planning in unknown environments.pdf).

Some research in path planning utilizing reinforcement learning has been done with the application of a DDQN (Double Deep Q-Network) as the planning algorithm. Originally proposed by DeepMind In the papers examined, the algorithm used for the DDQN was the same, but the main things that differed were their environments and reward functions. In one paper, “Dynamic Path Planning of Unknown Environment Based on Deep Reinforcement Learning”, a DDQN is applied to a mobile robot in an attempt to plan in an unknown, dynamic environment. The work is able to show that…. IDK FIX IT LATER

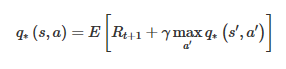
# Approach

Based on the promising results of research utilizing reinforcement learning with a DDQN for path planning in unknown environments, it seemed beneficial to utilize this algorithm for the development of the PAWS navigation system. The main objective for the PAWS Bot is to be able to navigate in an unknown environment to reach a goal in a timely manner. The secondary objective is to learn not to crash into obstacles. It is not the goal to necessarily find an optimal path, but a path that takes a shorter amount of time is preferable.

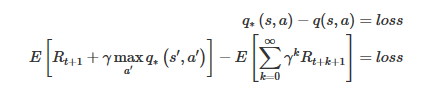
## The Robot

The PAWS Bot is a 2-wheeled circular robot that only performs planar motion which is limited to the cardinal directions; in other words, it is not allowed to move diagonally. The information the robot has access to is limited due to its unfamiliarity with the environment it is in. The robot can only sense locally and receives a type of beacon signal denoting the location to which it needs to travel. It is able to determine the distance vector between itself and the beacon location, but it has no other knowledge of the rest of the map. The robot is designed to gather the information about the environment through mainly proximity sensors, and feed this information into a decision making process to determine an appropriate action to take in that instance. There is no feed back once the robot takes its prescribed action so the control algorithm is open loop.

## Double Deep Q-Network

The algorithm for a DDQN combines the search for the optimal policy, that the Q-learning provides, with the generalization ability and the application to large state spaces that neural networks provide. Together this algorithm allows for Q-learning to be expanded from the tabular method, to something more complex. The fundamental aspect of DDQN is the Bellman Equation. This equation explains the optimal Q-function, which defines the optimal return for a relationship between states and actions.

Bellman Equation: <https://deeplizard.com/learn/video/rP4oEpQbDm4>

The neural network in this algorithm functions to approximate this optimal Q function by minimizing the loss between the values output from the network, and the target values.

Network Loss: <https://deeplizard.com/learn/video/0bt0SjbS3xc>

Calculating the target values for the loss function can be tricky because the target function is what is being predicted. Using the network being trained to also calculate the target values causes the neural net to essentially eat its own tail during training. To resolve this, it is common to introduce a target network which is used solely to calculate the target values of the Bellman Equation. This target network is then updated on some interval to match the weights of the trained network. The introduction of this target network is what makes this algorithm a *Double* DQN (<https://deeplizard.com/learn/video/xVkPh9E9GfE> ).

The DDQN also utilizes the Epsilon Greedy Strategy which is what enables the implementation of exploration vs. exploitation to the algorithm. This strategy pushes the agent to explore the environment early on in the training phase, to potentially find a more rewarding action than the one that is being chosen by the network. As the training phase goes on, the epsilon value decays, making it more likely to exploit the information it has learned through training ( <https://deeplizard.com/learn/video/mo96Nqlo1L8> ).

The last major component of the DDQN algorithm is called Experience Replay. This technique is accomplished by storing the state, action pairs, as well as some additional information from the environment, into Replay Memory, which is then accessed in random batches to train the network. This method is used to reduce the correlation caused by training the network on consecutive samples (<https://deeplizard.com/learn/video/Bcuj2fTH4_4> ).

The above components get applied to create the DDQN algorithm which was utilized for the PAWS navigation system. This algorithm is generalized below.

1. Initialize the environment and agent
2. Copy the policy network to the target network
3. For each episode:
   1. Initialize the simulation
   2. For each time in max time:
      1. Select an action based on e-greedy strategy
         1. Random or from neural net
      2. Execute action in simulation environment
      3. Return reward obtained and next state
      4. Store last experience in replay memory
      5. Perform neural net training
         1. Select batch of memories to replay through policy network
         2. Utilize the target net to obtain values for the next state
         3. Calculate the loss between the output Q values and target Q values
         4. Gradient descent to minimize loss
      6. Decay epsilon
      7. After a set time update the target net weights to match the policy net weights

<https://deeplizard.com/learn/video/xVkPh9E9GfE>

## The Neural Net

The neural network that is used in the DDQN algorithm has layers that are all densely connected and utilize ReLU activation functions. The input layer was created with an input size of 7, matching the state size for the agent. The output layer is composed of 4 outputs, which correspond to each of the 4 possible actions that the robot can take (forward, back, left, right). There are 2 hidden layers connecting the input to the output layers.

## The Environment

The environment for the PAWS Bot consists of free space, obstacles, and a goal. To simulate a changing environment the obstacles and initial robot location were kept static, but the goal location changed around the map. This allowed for a new look at the environment every episode iteration. The environment is confined to a set space and closed off by walls.

The reward structure for the environment was complicated to devise. It was determined that the robot should learn to avoid obstacles, traverse to a goal, and reach the goal in a timely manner. So it seemed a simple 1 or 0 reward function would not be sufficient to encompass the scope of the problem. To accomplish the determined goals a function was created that incorporates aspects of distance and time to calculate the reward as shown below.

To summarize, if the robot reaches the goal state, it receives the goal reward, which is a large positive value to entice completion of the task. If the robot encounters an obstacle it receives the obstacle reward, which is a moderate negative value to encourage not crashing into obstacles. For all other spaces the reward is a composite of , which creates a positive value for moving towards the objective and a negative value if moving away. , which creates a higher value for being in spaces nearer to the objective and a negative value for being far away. And , which is a negative value that increases as the time limit reaches expiration, which encourages completing the task quickly. The α are weights which sum to one and allow for tuning of the reward function.

## The Agent State

The robot agent needs the capability to observe its local area for obstacles and needs knowledge of the goal it is trying to reach. Due to these needs a state was created that attempts to coalesce this information into a succinct state. The first component it 4 Boolean proximity sensor inputs that sense at north, south, east, and west based on the robot’s point of view. They are either 1 or 0 depending on whether an object is sensed by that sensor. The second component is the vector pointing between the human location and robot’s location. This vector is rounded to 1 decimal place in an attempt to batch the state space into somewhat of a grid. The Z coordinate of the pointing vector is excluded due to the environment being planar in motion. The last component is the normalized time elapsed since the episode began which is a decimal between 0 and 1. The state can be represented as the below, where

# Implementation

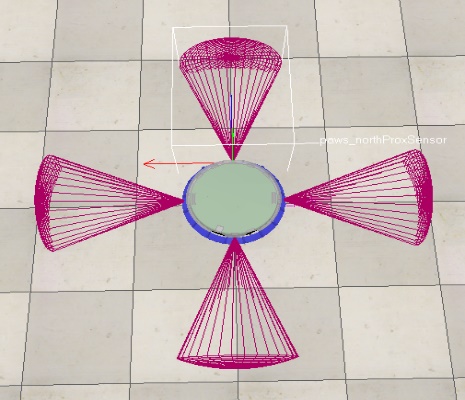
To design the PAWS navigation system the robot and algorithms were implemented via simulation, which allowed a controlled environment for training and testing. The original plan for these tests was to utilize the CoppeliaSim software to simulate the robot and the environment. But due to several issues outlined below the use of this software became nearly impossible. After much tweaking and testing it was eventually decided to abandon CoppeliaSim and instead utilize a simulation library that was better suited for the fast pace of machine learning algorithms. The OpenAI Gym was chosen to use for simulation due to it being designed for reinforcement learning problems. The work done and data collected utilizing CoppeliaSim is still captured in the remaining sections for reference.

## Original Design - CoppeliaSim

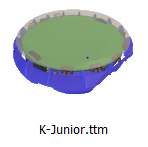
The simulation software utilized was CoppeliaSim (formerly V-REP). The algorithms were developed using Python which implemented the Keras library for the neural network portions which runs on a TensorFlow backend. Various other libraries were used such as numpy and scipy for some of the calculations, as well as matplotlib for data plotting.

### Robot Design and Control

The designed robot was developed using an available model from the CoppeliaSim library. The K-Junior model shown below was modified to be slightly larger and the sensor placement was arranged so that the sensors pointed in the North, South, East, and West directions from the robot’s point of view. The range that the sensors could detect was modified to 0.7m, and the line of sight of the sensors was adjusted so they would each have sensing perimeters that were independent from one another.



5‑: Robot with Sensor Locations



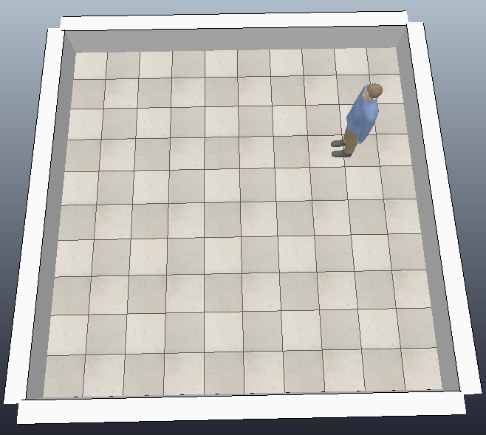
‑:K-Junior Model CoppeliaSim

The robot is controlled via Python modules that interface with the CoppeliaSim Remote API. The possible actions the robot can take are coded as forward or backward, in which the robot moves straight ahead or in reverse, and left or right, where the robot first rotates 90 degrees in the appropriate direction and then moves straight ahead. The necessary motor commands to complete these actions are transmitted to the robot via the API using the simxSetJointTargetVelocity command. The robot was designed to move in steps set to 0.5 m to emulate a grid like state space. The robot can detect when it collides with an object by querying the API for simxReadProximitySensor which returns a vector to the nearest object. If that distance is less than a set tolerance, the robot is determined to be in a collision state.

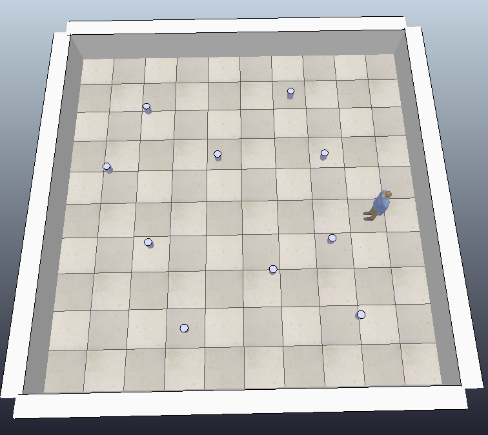
There were some difficulties in developing the robot control interface. Controlling the robot’s turn radius was exceptionally difficult due to the information that is available via the API. The only information about rotation is accessible via the simxGetObjectQuaternion command, which then has to be translated into angles via a set of transformations. Another issue arose with needing to reset the robot to the zero location at the beginning of every episode. The API allows an object location to be set using the simxSetObjectPosition command, which essentially teleports an object to a new location. Due to the dynamics that are imposed on the robot model, and teleportation not being physically possible, the robot would break into pieces when trying to relocate it. This issue was resolved by removing and reloading the robot model into the simulation every episode.

### The Environment

The environment was created in CoppeliaSim and was chosen to be a simple planar floor. The size was either 5m x 5m or 2.5m x 2.5m depending on the tests being run. The obstacles are composed of walls to set the boundaries of the state space and cylindrical posts that the robot can detect and collide with. There were 2 versions of the environment created, a larger one with obstacles, and a smaller one that contained only the goal and walls. The goal that the robot needs to travel to is marked by a human figure that is moved around the environment every new simulation episode.



5‑3: Map 2.5m x 2.5m



5‑4: Map 5m x 5m

### The Algorithm

The DDQN was implemented via Python and followed the algorithm outlined in the previous section. The original plan for the algorithm was to develop the pieces from scratch; this includes the algorithm for the neural network portion of the DDQN. The code for the neural net was originally written using Python with ReLU activation functions and fully connected layers. Due to many issues encountered during the training of the DDQN, it was opted to use the open source library Keras to handle the neural network component instead. This helped eliminate one of the variables of the DDQN algorithm in hopes that it would be easier to narrow down what was going wrong in the training process. The originally developed neural net remains in the code base for reference.

The parameters used in the DDQN are as follows:

**Training Settings**

EPISODES = 500

TIME\_LIMIT = 100

MEMORY\_CAPACITY = 1000

BATCH\_SIZE = 35

TARGET\_UPDATE\_COUNT = 25

**Hyperparameters**

ALPHA = 0.05

DISCOUNT\_RATE = 0.95

EPSILON = 1

EPSILON\_DECAY = 0.995

EPSILON\_MIN = 0.01

One of the things to note about these parameters is the high learning rate (ALPHA). This was chosen due to issues with training which prevented large numbers of episodes from being performed. The higher learning rate was chosen to help the algorithm reach a solution more quickly so more iterations would not be necessary.

The reward weights and values were iterated many times throughout training, but they generally settled around the following values:

**Reward Settings**

GOAL\_REWARD = 100 (or 10)

NOT\_SAFE\_REWARD = -1

REWARD\_DISTANCE\_WEIGHT = 0.4

REWARD\_CLOSE\_WEIGHT = 0.6

REWARD\_FREE\_SPACE = 1

REWARD\_TIME\_DECAY = -0.5

### Data Collection

Throughout training data was being collected and plotted at set intervals. The values that were collected are as follows:

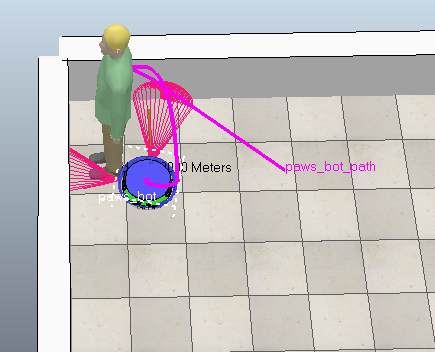
* ***Distance vs. Time****:* Shows the path that the robot took over the course of training. A separate plot was saved for every episode.
* ***Error vs. Time:*** Shows the average RMS error calculated every training loop. A separate plot was saved for every episode.
* ***Time vs.*** ***Episode:*** Shows the amount of time it took the robot to reach the done state in each episode. A plot was saved at the end of training.
* ***Reward vs.*** ***Episode:*** Shows the average reward accumulated per episode. A plot was saved at the end of training.

To allow for easier testing, the values of the weights in the neural net were saved off once training was complete. This allowed for any net to be loaded during the testing phase so comparison of results could be made

### Problems

Training utilizing CoppeliaSim as the simulation software proved to be extremely difficult. One of the downsides to using this software was the amount of time it took to run the episodes. The steps applied to the robot happen in real time so each time step took seconds, which made training a high number of iterations very time consuming. The ability to vary the training and hyperparameters was limited due to the amount of time it took to run. The max run time achieved from the simulation was 5 hours which only amounted to 50 episodes. Increasing the motor speed did not help speed up the time because it instead caused the robot to travel much farther than anticipated due to the rate at which commands are communicated through the API.

Another huge issue experienced with this software was its tendency to introduce anomalies into the physics that caused the robot to get flung around the map when the model was loaded, which ended with the robot being flipped upside down. This resulted in the training halting unexpectedly which ruined several overnight training runs that were planned. It took quite a bit of time to figure out what was causing the anomaly since it happened randomly, but the flipping was eventually observed and was attempted to be mitigated through code (i.e. killing the episode if this happens, changing the order and timing of moving models, etc.). These methods were all unsuccessful in resolving the issue. A snapshot of this anomaly is shown below. It shows the robot resting upside down after the resetting of the environment, as well as the erratic path it took, where it bounced off the wall, which is shown in pink.



‑:Error Anomaly

## Modified Design – OpenAI Gym

# Results and Analysis

## CoppeliaSim – Original

Test accuracy

Random errors with simulation

Remove obstacles and reduce map size

Goes in squares ☹

Why didn’t it work

Too large of a state space… (source)

Changing map

## Gym – Modified

Is there time to run this in Gym? Maybe future work

# Conclusions

fml

# Future Work

# Bibliography

Dynamic Path Planning of Unknown Environment Based on Deep Reinforcement Learning (PDF)