Derek Paterson  
October 23, 2022

CS-370

Design Defense

A human being generally solves a maze with trial and error. Humans can process spatial relationships and visualize general distance, so a human can use informed trial and error by estimating path length and referring back to sections of the maze they have already seen. This would allow them to build a spatial model of where they are going and perhaps let them solve the maze with minimal errors, but it could also vary drastically depending on the person. Another strategy they could take is following the entirety of the maze, i.e. always following the wall with one hand touching it, but this is guaranteed to be the longest possible solution as it is an exhaustive exploration of all sections of the maze.

The intelligent agent, however, solves this problem rather differently, and there are two key methods to how it solves the maze. The first is that the intelligent agent is capable of learning, so every time it makes a decision in regards to the direction it is headed in, it remembers that decision flawlessly. The second is that it takes random actions to try to navigate the maze, but this randomness is weighted against its prior experiences. So, the steps the agent takes are as follows, assuming a fixed layout and endpoint:

* Agent is assigned a random starting point
* Agent runs through maze, making decisions randomized   
  based on experiential weighting
* Once agent reaches the end of the maze, the experience is evaluated and stored
* Changes to decision weighting are implemented
* Repeat until optimized

The important thing to consider here is that the agent is always placed in a random point while it is being trained, and that the decisions it makes are weighted with point values to determine the optimal pathway.

There are only a couple of ways these approaches are similar, chiefly that both a human and an intelligent agent are aware of the parameters of the maze, i.e. that there is a fixed layout and an endpoint, and that in both cases the optimal way to navigate the maze is by using an educated guess. They are different in that the intelligent agent is drawing on much more overall experience specific to the maze, as it is running it many many times in sequence to try to find the optimal path. Every action the intelligent agent takes is done via careful analysis of point values (rewards) in that it is trying to get the biggest reward for the task by making the least amount of mistakes. A human being navigating a maze just doesn’t consciously think this way. Their way their educated guesses are used also differs, a human being is making guesses based on current visual input and spatial awareness, not on past experiences of trial and error like the intelligent agent. Ultimately, for the task of navigating a specific maze, an intelligent agent will quickly be able to navigate a maze faster than a human.

When developing an intelligent agent, exploitation is the part of the decision making system that decides on whether to repeat previous decisions that have earned a high reward, while exploration is making new decisions in hopes of gaining even better rewards (Lindwurm, 2019). The ideal proportion of exploitation to exploration is difficult to balance, in this case we are using a discount factor of 0.95 which means our intelligent agent is much more predisposed towards exploring based on future rewards (C. K., 2022). The reason this works is we’re using a very large number of training runs, so we can predict and optimize over relatively long timespans, and be largely unconcerned with short term results.  
 That ties in with the overall method we use for pathfinding, which is that of reinforcement learning. Reinforcement learning helps an agent accomplish its task by establishing an environment, a state for the agent, rewards, policy, and value (which is related to our discount factor mentioned previously) (Bhatt, 2019). The way this works in our favor is that we’re constantly training the policy of the agent to make better decisions in moving towards its ultimate goal (the end of the maze). When we do this, we’re prioritizing future points because of our high discount factor, which means initially the agent will focus on exploring the maze and building out methods to decide where to go, but eventually this lets us optimize the decision making and have the agent navigate the path quite adeptly.

The Treasure Hunt Maze problem was solved using deep Q-learning, which means we have our agent refer to a Q-table that lists actions it can take and calculates the future rewards based on what it has done and the current state of its environment (Loeber, 2022). The process is relatively simple, it is a constant iterative analysis loop. The Q-table starts with all zeros for Q-values (Loeber, 2022). From there, the intelligent agent chooses an action for its current location based on the best Q-value it can determine, in this case we are leaning towards exploration with our high discount factor, and so early in the training the agent is likely to choose random directions (Loeber, 2022). The agent performs whatever action (movement within the maze) is deemed appropriate and the outcome is observed, then the reward for this outcome is calculated and used to update the Q-table (Loeber, 2022). Over time the agent learns to optimize itself based on the Q-values and we see less and less exploration, leading to a point where the agent no longer is capable of meaningfully improving its performance and is navigating the maze as best as it can (Loeber, 2022).

References

Bhatt, S. (2019, April 19). *Reinforcement learning 101*. Medium. Retrieved October 23, 2022, from <https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292>

C. K., S. (2022, June 3). *Discount factor in reinforcement learning*. Intuitive Tutorials. Retrieved October 23, 2022, from <https://intuitivetutorial.com/2020/11/15/discount-factor/>

Lindwurm, E. (2019, November 3). *Intuition: Exploration vs exploitation*. Medium. Retrieved October 23, 2022, from <https://towardsdatascience.com/intuition-exploration-vs-exploitation-c645a1d37c7a>

Loeber, P. (2022, February 10). *Reinforcement learning with (deep) Q-learning explained*. News, Tutorials, AI Research. Retrieved October 23, 2022, from <https://www.assemblyai.com/blog/reinforcement-learning-with-deep-q-learning-explained/>