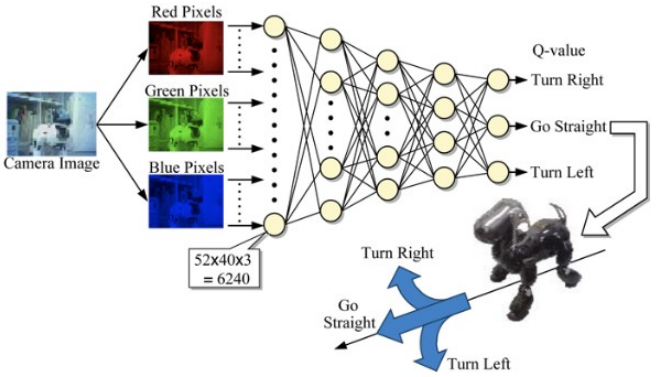
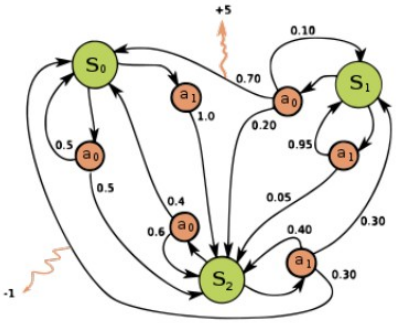
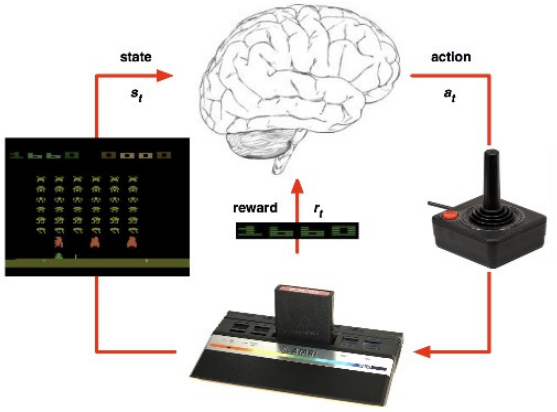
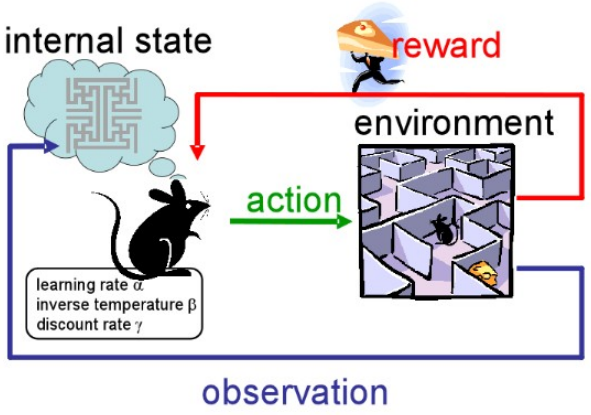
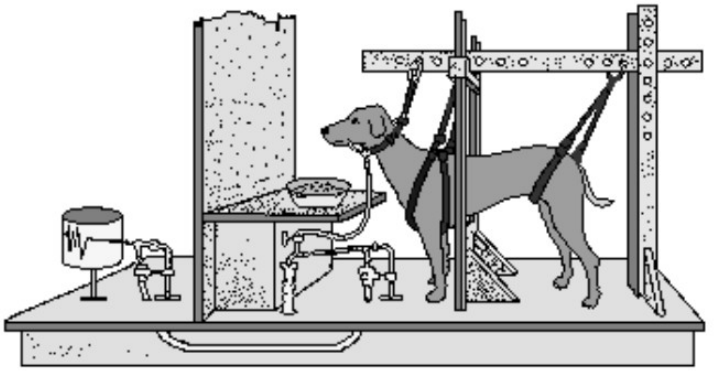
COMP 138: Reinforcement Learning











Instructor: Jivko Sinapov

Announcements

Announcements

- Office hours this week: Thursday at 2 pm

Reading Assignment

- Chapter 8 of Sutton and Barto

Research Article Topics

- Transfer learning
- Learning with human demonstrations and/or advice
- Approximating q-functions with neural networks

Research Paper

- Narvekar, S., Sinapov, J., Leonetti, M. and Stone, P. (2016) **“Source Task Creation for Curriculum Learning”**, *In proceedings of the 2016 ACM Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*
- Responses should discuss both readings

Programming Assignment #2

Team Formation: due next Tuesday

Moderated Discussion

Reading Discussion

“What makes Q-Learning off-policy? Is it because the agent estimates the value of a state-action pair based on the next action that would maximize reward, but doesn't necessarily take that next action in its behavior?”

- Randy

Reading Discussion

“Can we conclude that for TD prediction, state-value function is used; while for TD control, action-value function is used?”

- Qing

Reading Discussion

“What is the derivation of Q-learning? Where is the maximum term derived from? For off-policy MC methods, there was an exploratory policy, where is that here?”

- Chami

Reading Discussion

“...why are we using the best action value of future state to estimate the current action value?”

- Qidi

Reading Discussion

“The book specifies that the start states are arbitrary. Is there a start state that is slightly more preferable to one or another?”

- Daniel

Reading Discussion

“Can there be situations in RL where prioritizing immediate rewards might mislead the agent away from an optimal long-term strategy?”

- Jianan

Reading Discussion

“Throughout this chapter, disadvantages of TD methods are not introduced, so it is hard to imagine why we still use other methods like Monte Carlo or DP. Is it having a disadvantage of in perspective of computational resource? Or do other methods have advantages over some specific tasks or in environments?”

- Changgyu

Reading Discussion

“The question I have is how do we choose between on-policy and off-policy methods”

- Winnie

Reading Discussion

“What exactly is a function approximator? I didn't quite understand the notion of tiling. I think perhaps this topic gets into neural networks, which I do not have much experience with.”

- Andrew

Reading Discussion

“Is there a benchmark or systematic way that could guide the selection ‘n’ in n-step bootstrapping problems? Regarding n-step off-policy Learning, will the choice of ‘n’ be related to the importance sampling ratio?”

- Weishi

Reading Discussion

“At this point I wonder is there another method that’s a step up from n-step TD?”

- Prithvi

Breakout: Finding Project Partners

Research Article Questions

Reading Discussion

“how can the relationship/similarity between the source task and target task influence the performance before and after transfer learning?”

- Jesus

Reading Discussion

“I'm wondering if any work has been done to learn the transfer function?”

- Jacob

Reading Discussion

“Is there a benchmark for the mapping method that we choose? Is it even possible to set up a more generalized benchmark across different environments?”

- Weishi

Reading Discussion

“Is it possible to extend the knowledge transfer process to be even more general regardless of the task and learning policy?”

- Chami

Open Questions about Transfer Learning

- What are some unanswered research questions posed by the article?

Further Reading on Transfer Learning

Reading Discussion

“In real world settings, human feedback can be sparse; in order for human feedback to be useful, is there a way that optimum amount of human feedback can be decided? Also, human feedback is very expensive when it comes to scaling up the system, yet the dilemma is that less expensive human feedback can be a lot less accurate. How would companies balance between the amount of feedback and quality of feedback?”

Reading Discussion

“How exactly did they "combine" the policies of the agent and that of choosing the optimal action based on human feedback?”

- Jacob

Reading Discussion

“I have one question about the importance of the .5 C value. Why does having a C < .5 actually work well? If c < .5, does the agent "learn to assume" that the human is lying most of the time, and it should treat the correct feedback as the opposite of the actual human's feedback?”

- Andrew

Reading Discussion

“I have some questions about the paper. What is the actual experiment step? Is it the case that the oracle is given one picture of the game and it needs to decide which action should the character take? I also wonder how consistency and feedback likelihood is expressed in a human-interaction scenario. “

- Zixiao

Reading Discussion

“Can this Advise strategy be used in the situation when there is not only one “right” action?”

Reading Discussion

“How can this method avoid potential bias within the human feedback?”

- Rafeed

Reading Discussion

“Is it possible to integrate advise with other human feedback RL methods that affect the rewards?”

- Caleb

https://www.ijcai.org/proceedings/2021/0599.pdf

THE END