Evidence-Driven Learning for Interpretability

Minwoo Lee & Giang Dao

May 5, 2022

BIRS

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Evidence-Driven Learning

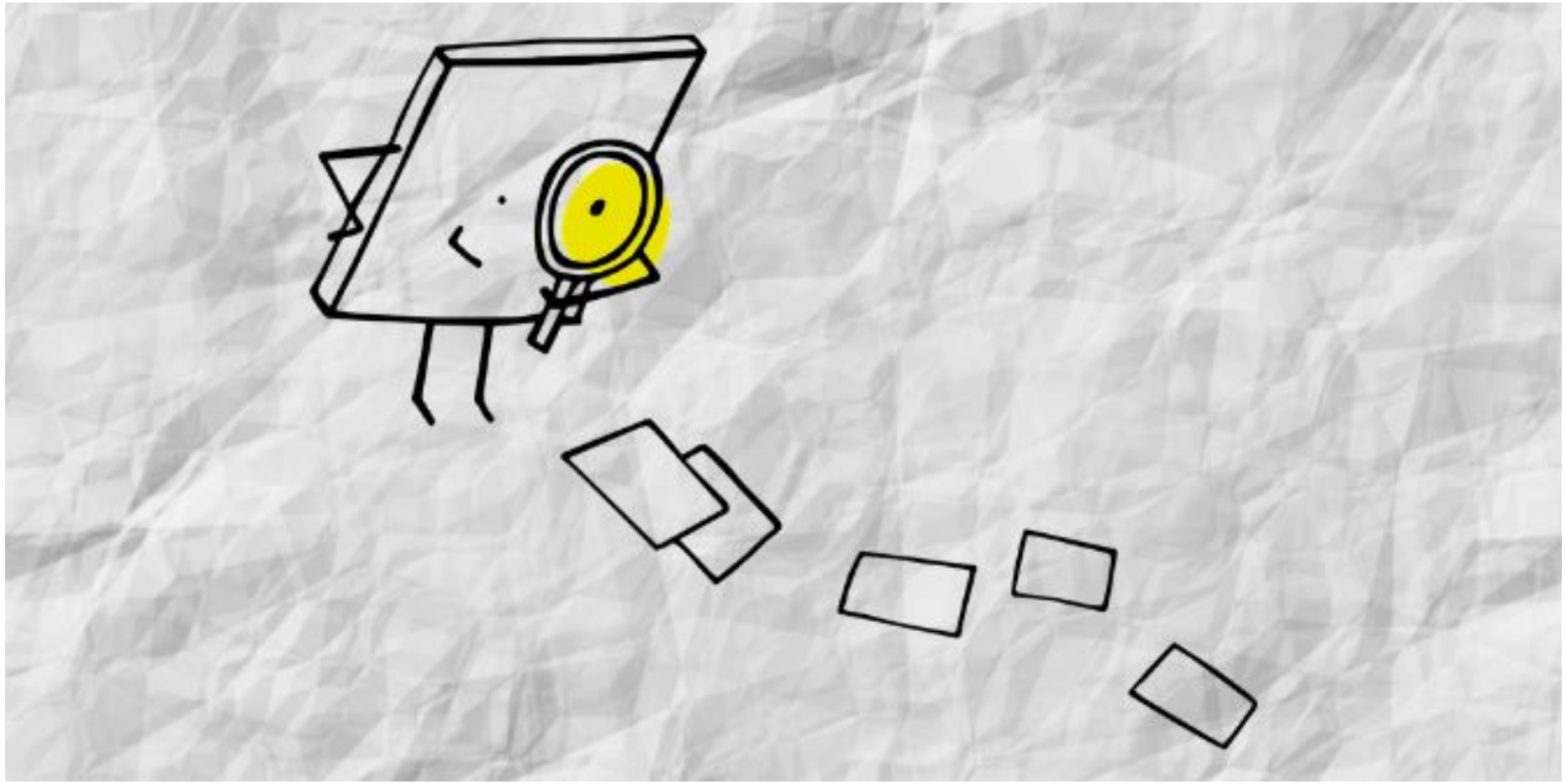
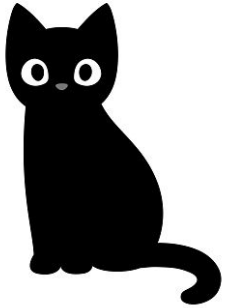
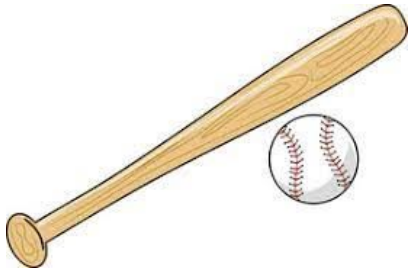


Image from: https://blogs.lse.ac.uk/impactofsocialsciences/2019/05/23/does-evidence-still-matter-10-strands-of-continuity-and-change-in-evidence-based-policy-and-practice/

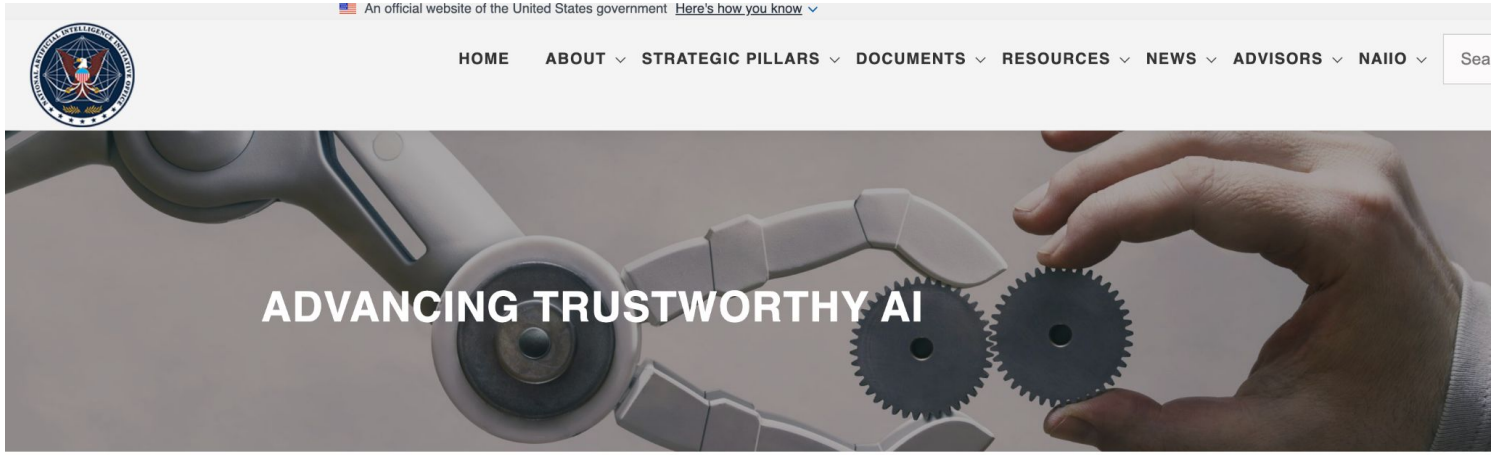
Evidence-Driven Learning





Why Interpretability or Explainability?

- Towards trustworthy AI (Kush)



Home » STRATEGIC PILLARS » *ADVANCING TRUSTWORTHY AI*

ADVANCING TRUSTWORTHY AI

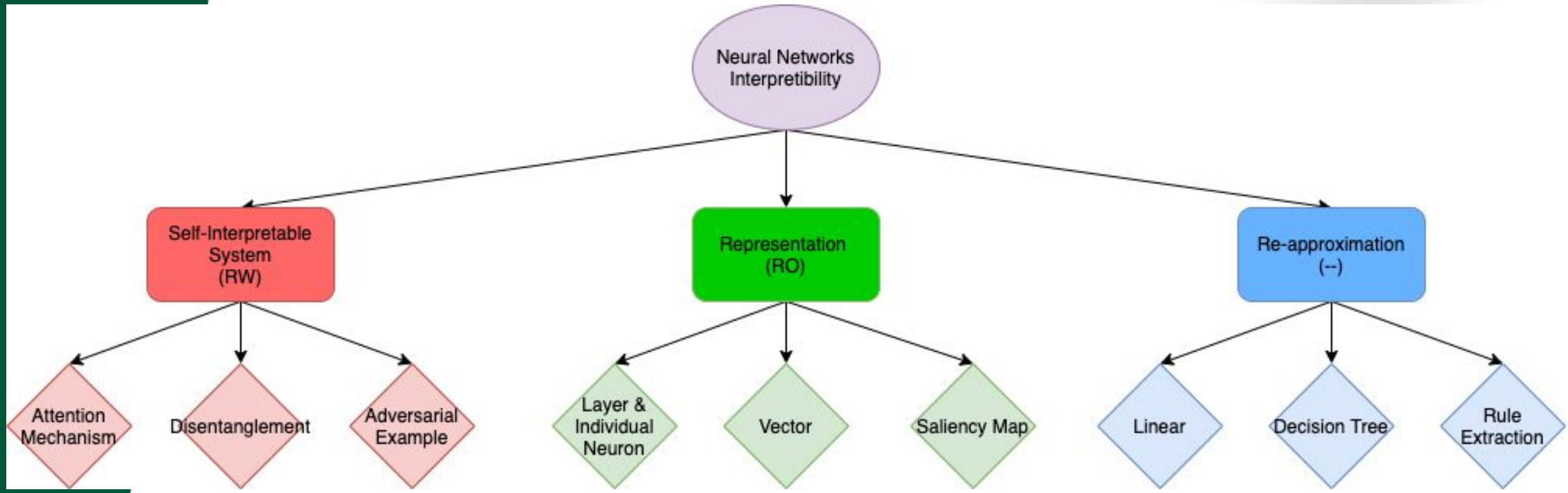
To be trustworthy, AI technologies must appropriately reflect characteristics such as accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security or resilience to attacks – and ensure that bias is mitigated. Factors such as fairness and transparency should be considered, particularly during deployment or use. In addition, the broader impacts of AI on society must be considered, such as implications for the workforce. Developing and using AI in ways that are ethical, reduce bias, promote fairness, and protect privacy is essential for fostering a positive effect on society consistent with core U.S. values.

ensure
ent and
sectors.
nder of
the rule
rights,
y; and
reflect

Contents [hide]

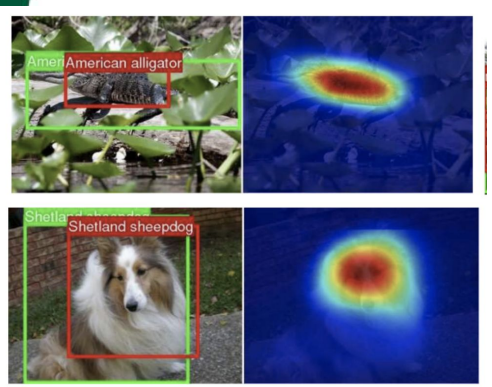
https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/

Opening the Black Box



Dao and Lee (2020). Demystifying Deep Neural Networks Through Interpretation: A Survey

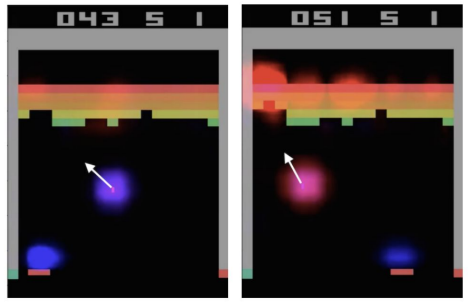
What Do We Need for Interpretation?



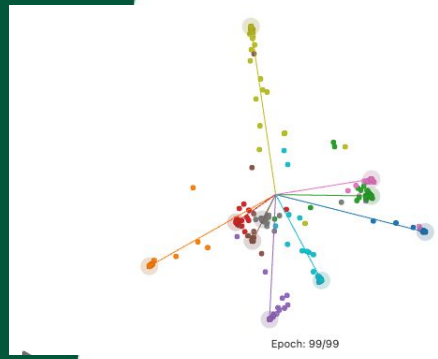
Zhou et al., 2016

is an arthritic attempt at directing by callie khouri. i had to look away - this was god awful	Negative
a visually seductive, unrepentantly trashy take on rices second installment of her vampire chronicles	Positive
could easily be called the best korean film of 2002	Positive
the best disney movie since the lion king	Positive
a cheerful enough but imminently forgettable rip-off of [bessons] earlier work	Negative

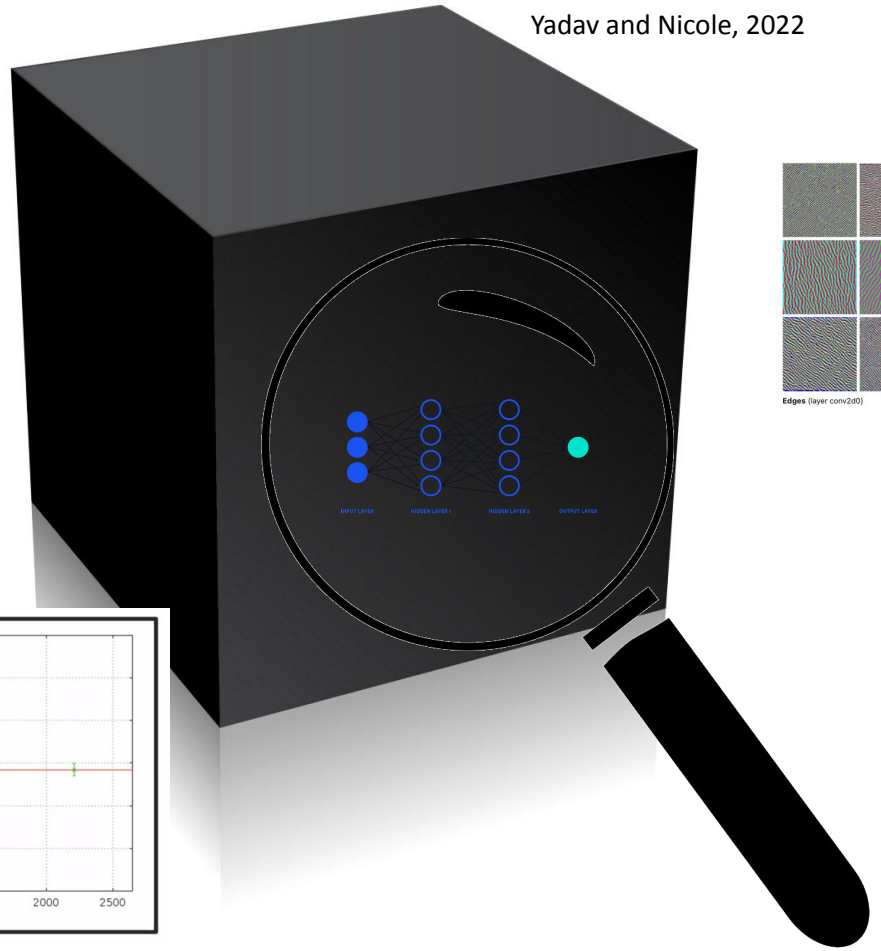
Yadav and Nicole, 2022

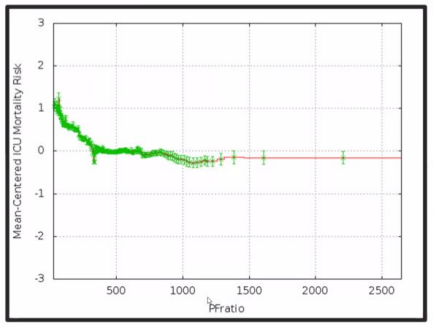


Graudanus, 2017

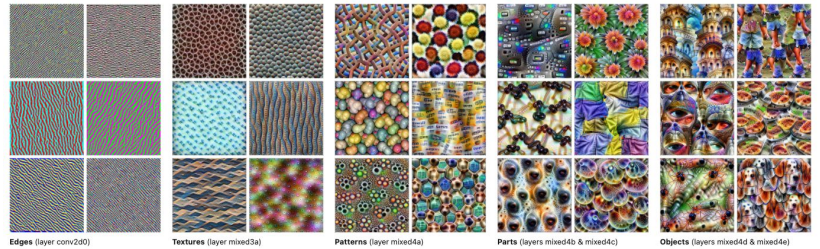


Li et al., 2020 (Grand Tour)

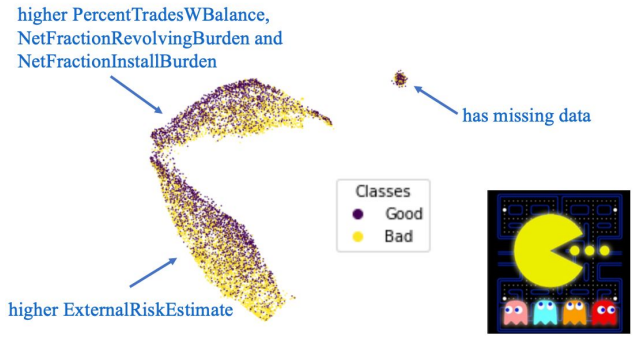




(Rich)

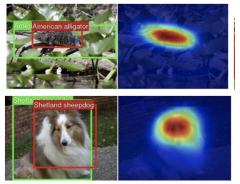


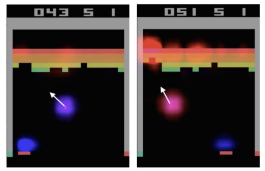
Olah et al., 2017



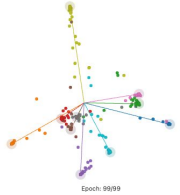
(Cynthia)

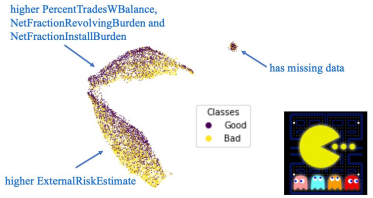
Sources of Interpretation

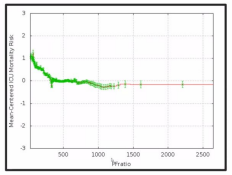




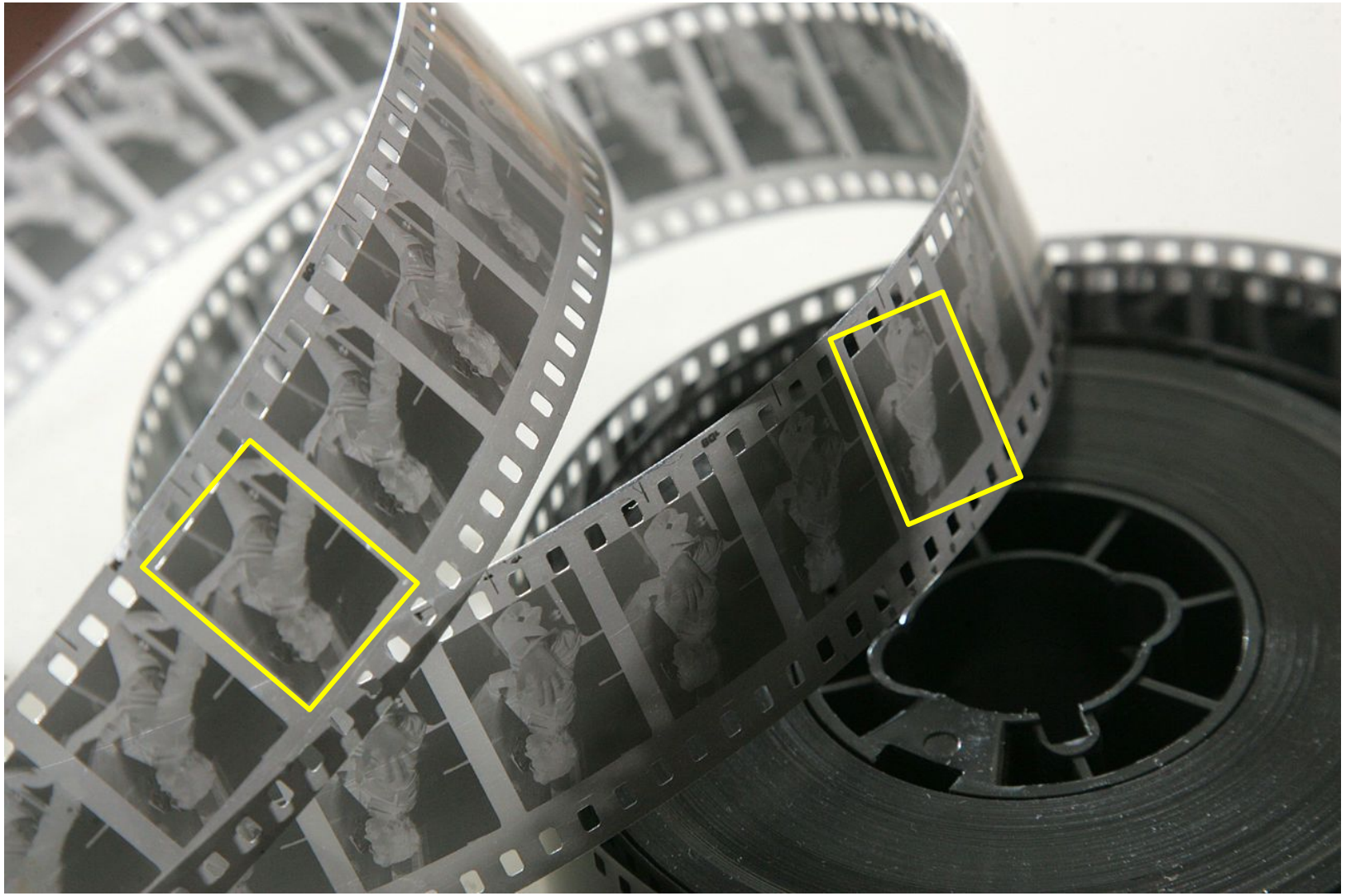
is an arthritic attempt at directing by callie khouri. i had to look away - this was god awful	Negative
a visually seductive, unrepentantly trashy take on rices second installment of her vampire chronicles	Positive
could easily be called the best korean film of 2002	Positive
the best disney movie since the lion king	Positive
a cheerful enough but imminently forgettable rip-off of [bessons] earlier work	Negative





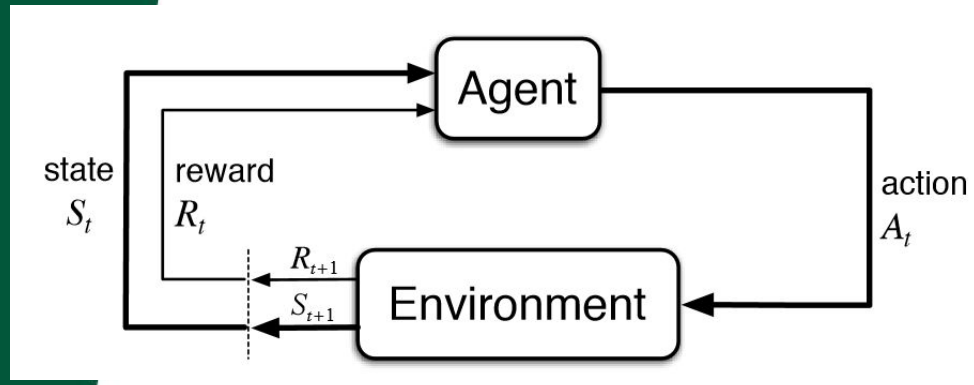


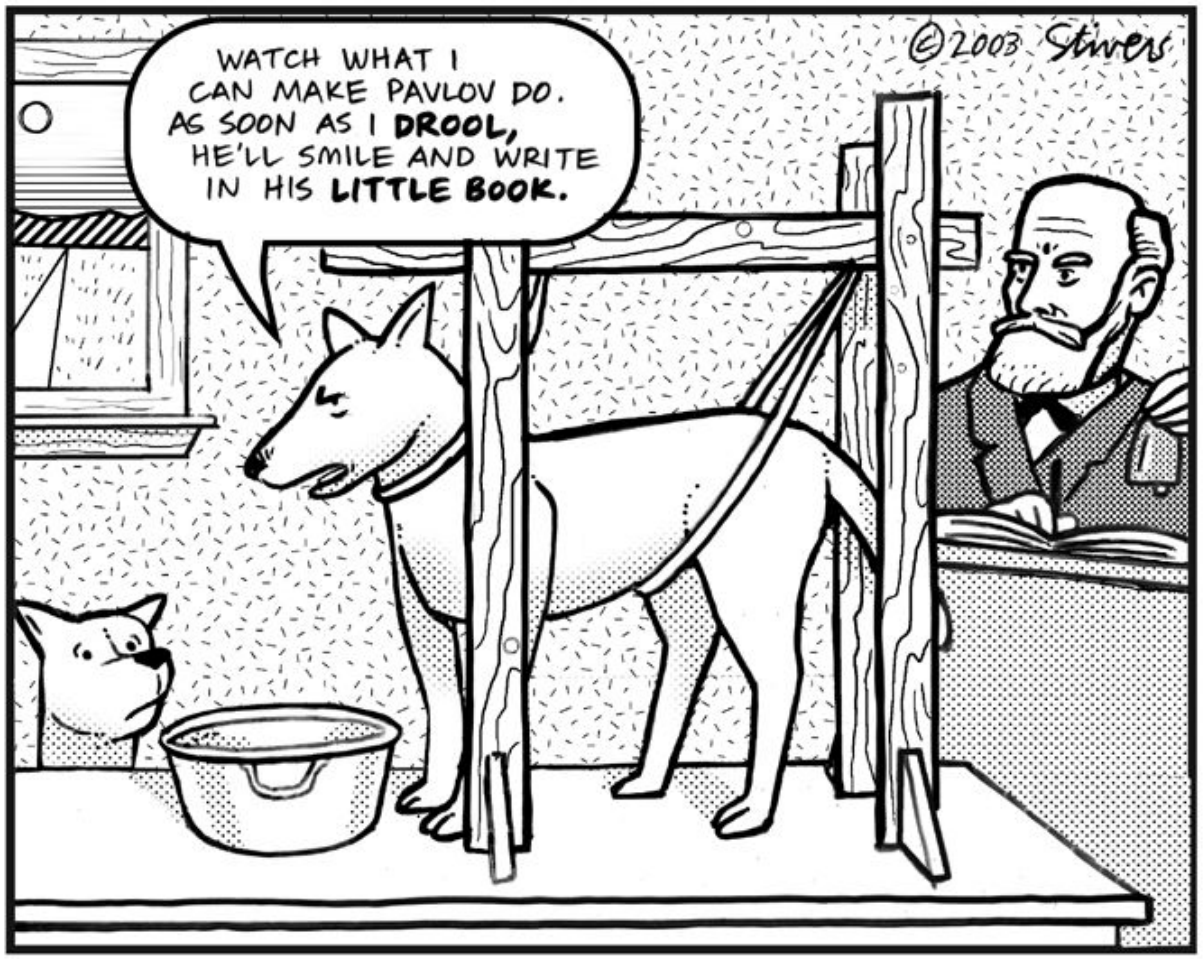
Sources of Interpretation => Evidence



https://en.wikipedia.org/wiki/35_mm_movie_film#/media/File:35mm_movie_negative.jpg

Reinforcement Learning





RL Applications

Reinforcement Learning in Healthcare: A Survey

Chao Yu, Jiming Liu, *Fellow, IEEE*, and Shamim Nemati

***Abstract*—As a subfield of machine learning, *reinforcement learning* (RL) aims at empowering one's capabilities in behavioural decision making by using interaction experience with the world and an evaluative feedback. Unlike traditional supervised learning methods that usually rely on one-shot, exhaustive and supervised reward signals, RL tackles with sequential decision making problems with sampled, evaluative**

feedback and the new state from the environment. The goal of the agent is to learn an optimal policy (i.e., a mapping from the states to the actions) that maximizes the accumulated reward it receives over time. Therefore, agents in RL do

nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open Access | Published: 16 February 2022

Magnetic control of tokamak plasmas through deep reinforcement learning

Jonas Degrave, Federico Felici, ... Martin Riedmiller + Show authors

Nature **602**, 414–419 (2022) | Cite this article

126k Accesses | **8** Citations | **2309** Altmetric | Metrics

Spotify Reinforcement Learning Recommendation System

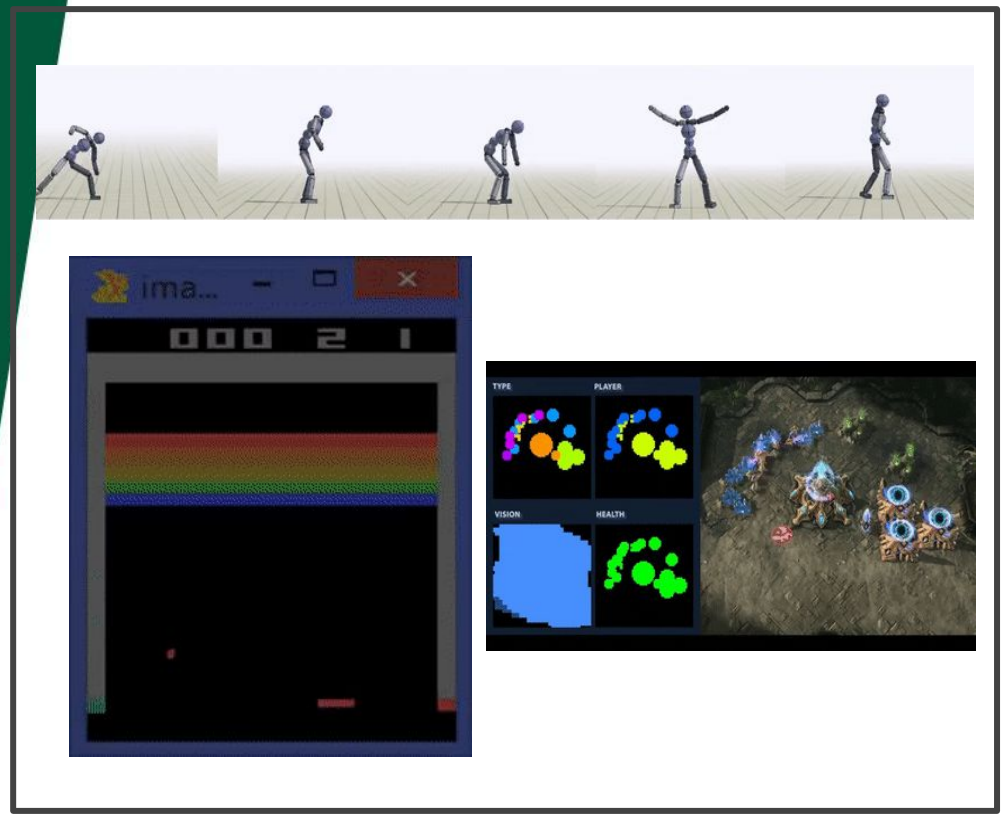
Posted on December 14, 2019

IACS HARVARD School of Engineering and Applied Sciences Spotify

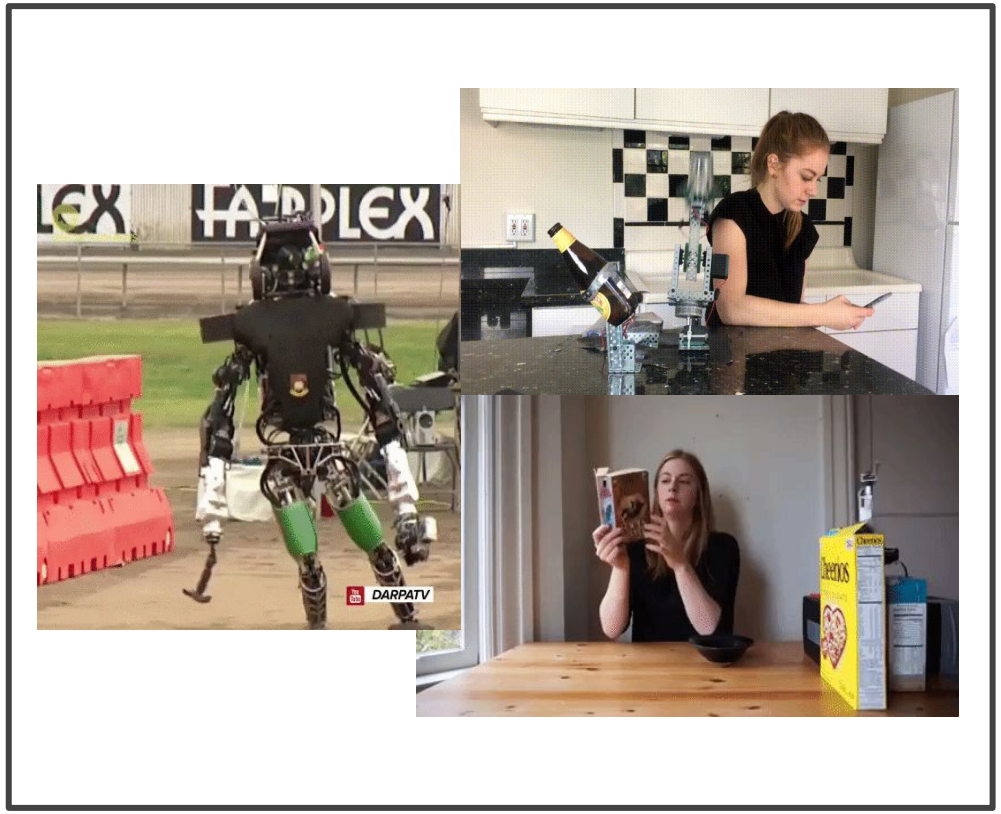
Team members: Feng Qian, Sophie Zhao, Yizhou Wang

images from: https://www.automationworld.com/factory/robotics/article/21759681/3-robotics-industry-predictions
https://chatbotslife.com/deep-learning-in-finance-learning-to-trade-with-q-rl-and-dqns-6c6cff4a1429

RL Challenges



RL in simulations



RL in the real world

Poole and Lee (2021). Towards Intrinsic Interactive Reinforcement Learning

RL Challenges

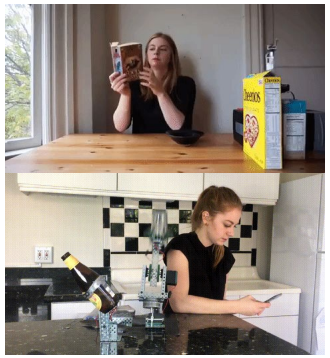
RL Challenges

Robustness/ Scalability

Focuses on being able to learn

Alignment

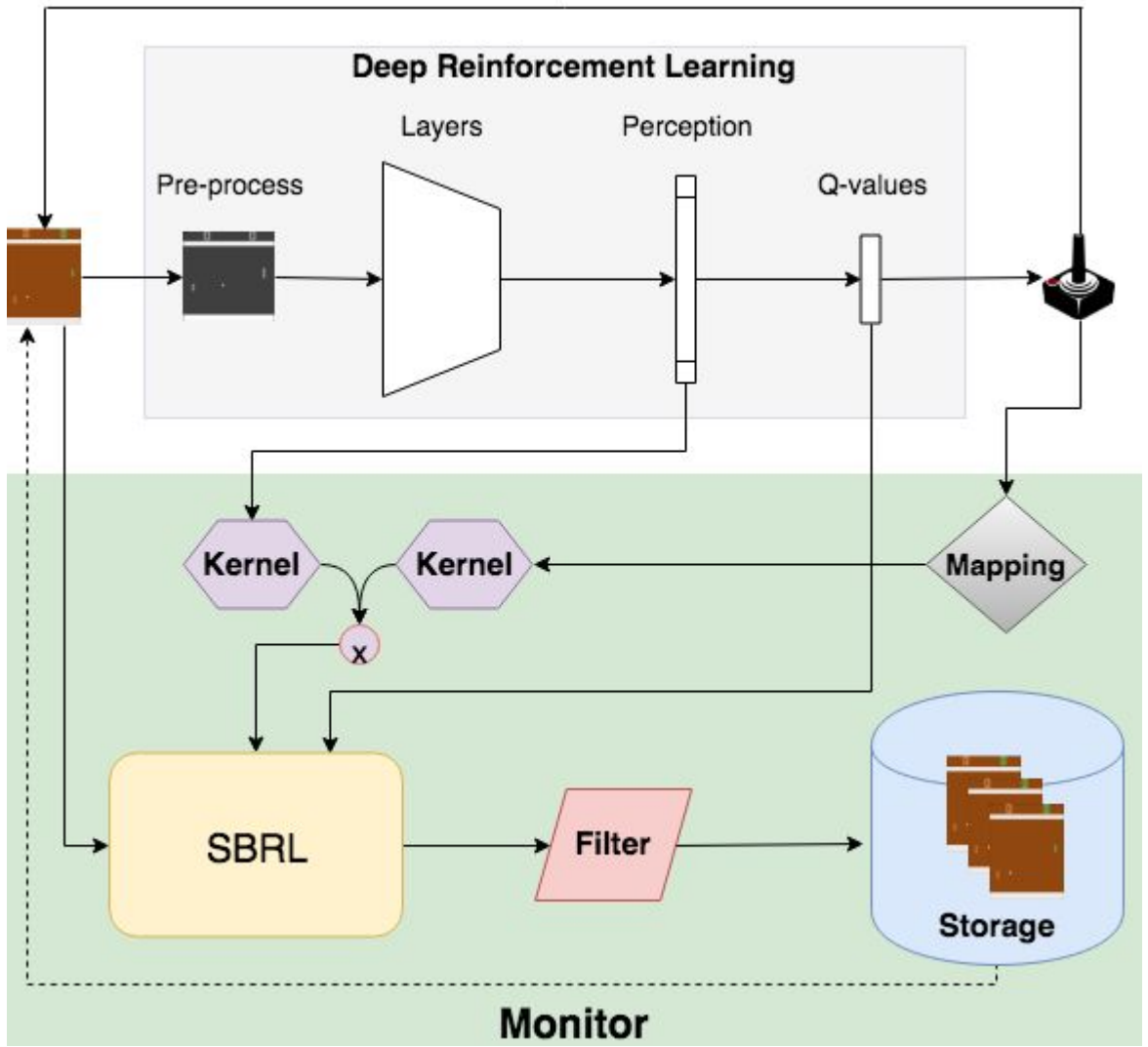
Focuses on learning as intended



Poole and Lee (2021). Towards Intrinsic Interactive Reinforcement Learning

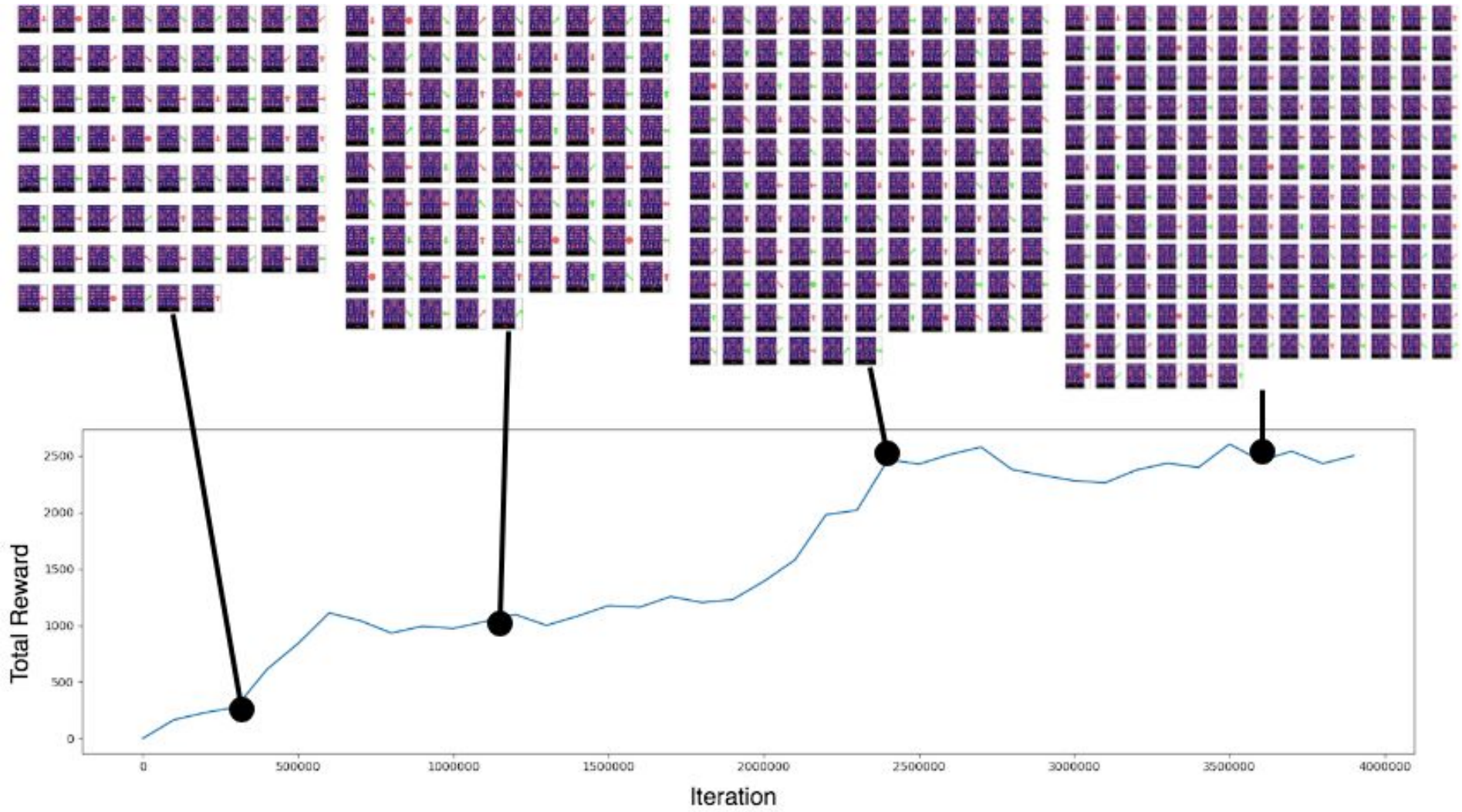
Let us collect “Evidence”

DRL-Monitor



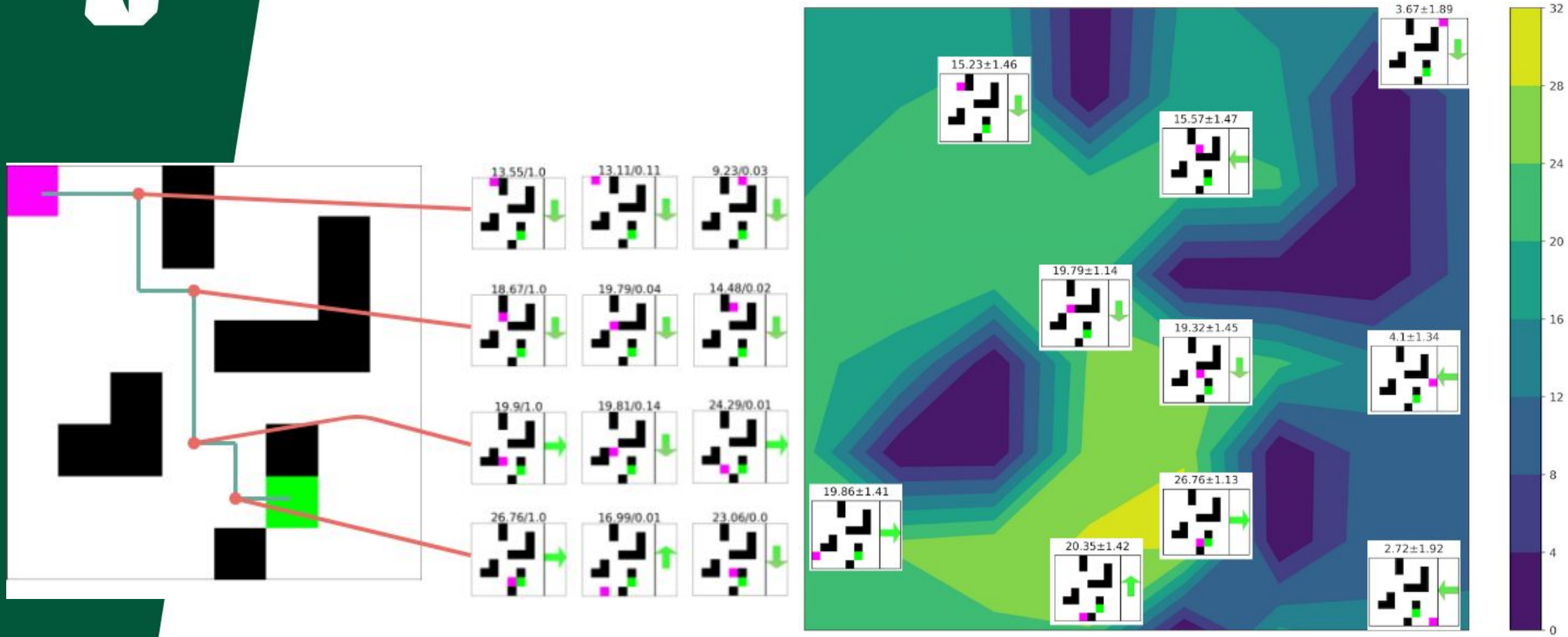
Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

DRL-Monitor



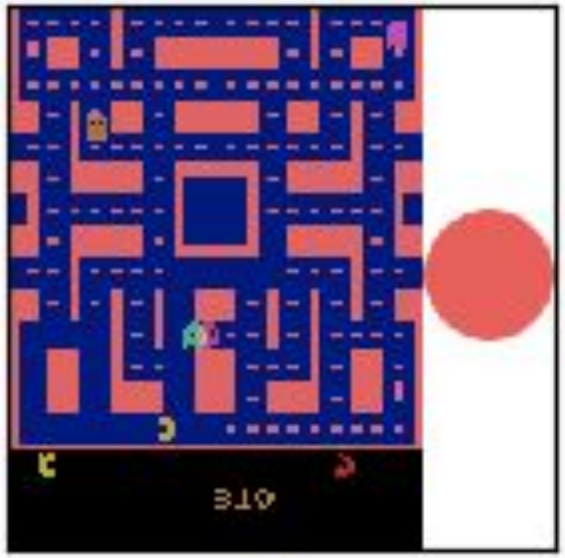
Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

Maze



Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

Ms PacMan (after 300K iterations)



1

2

3

Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

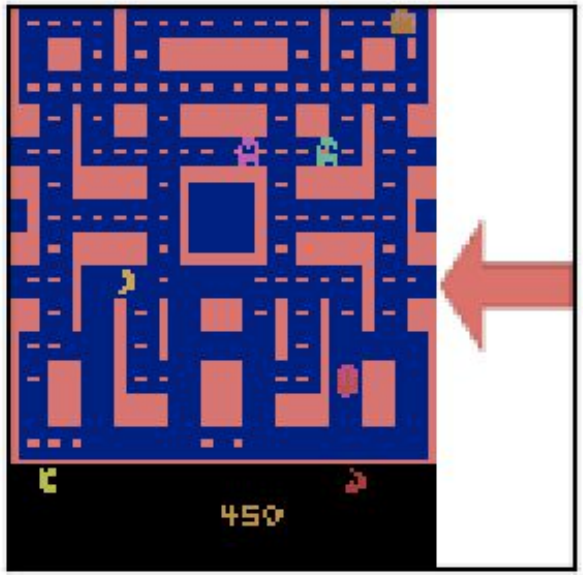
Ms PacMan (after 300K iterations)

1 2 3

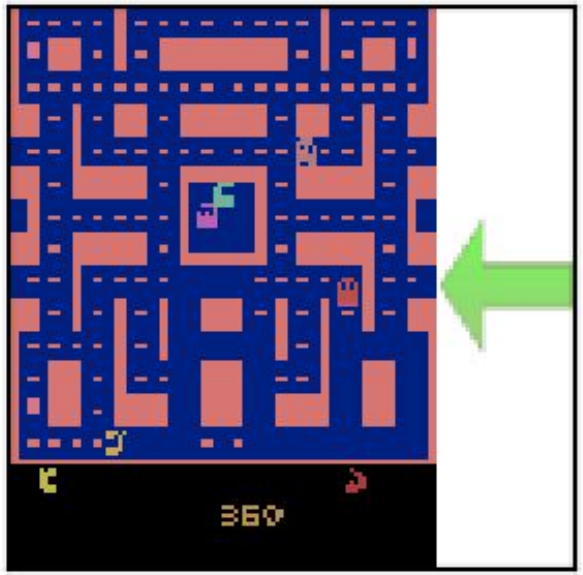
No action (not moving) does not help!

Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

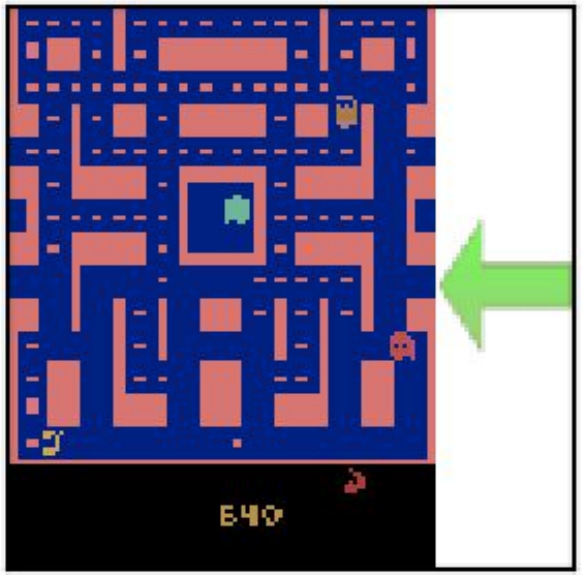
Ms PacMan (after 1.2M iterations)



1



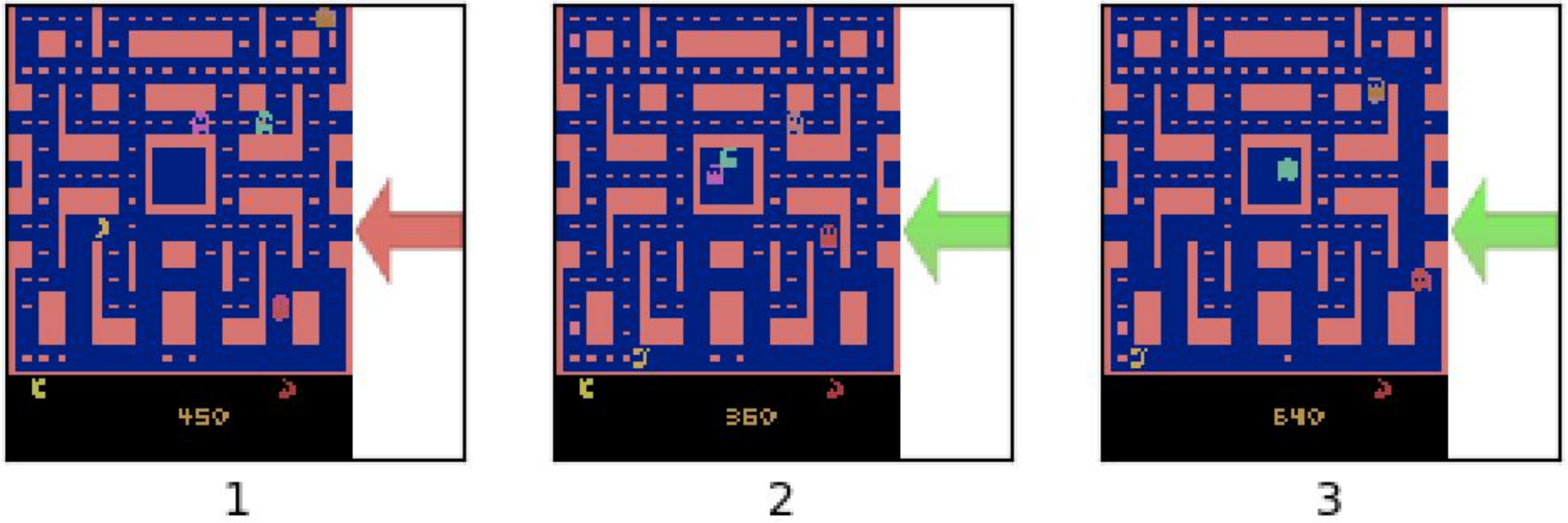
2



3

Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

Ms PacMan (after 1.2M iterations)

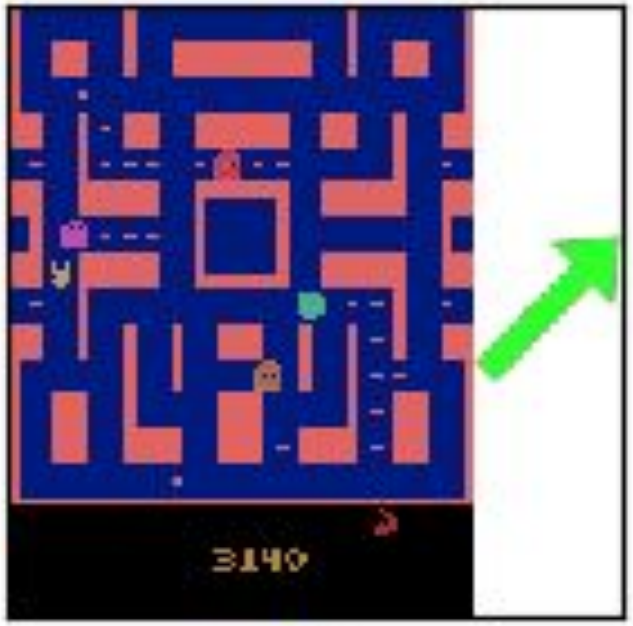


Craving for food. Heading towards the food is good, and moving away from it is bad.

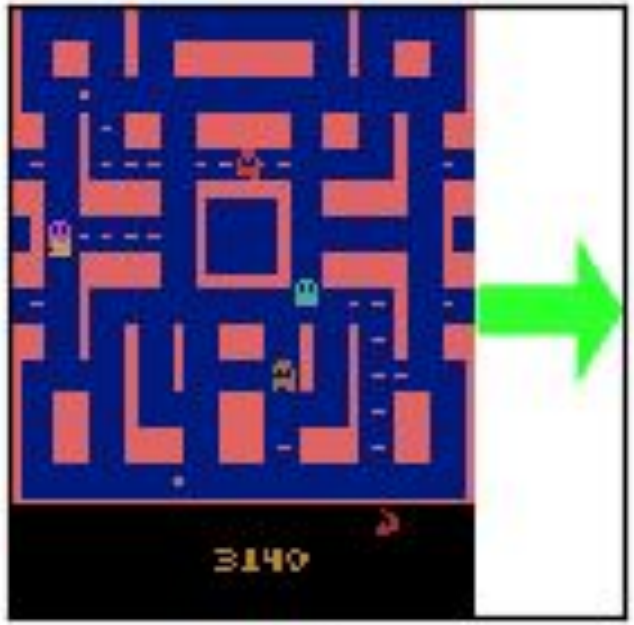
Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

Ms PacMan (after 2.4M iterations)

1



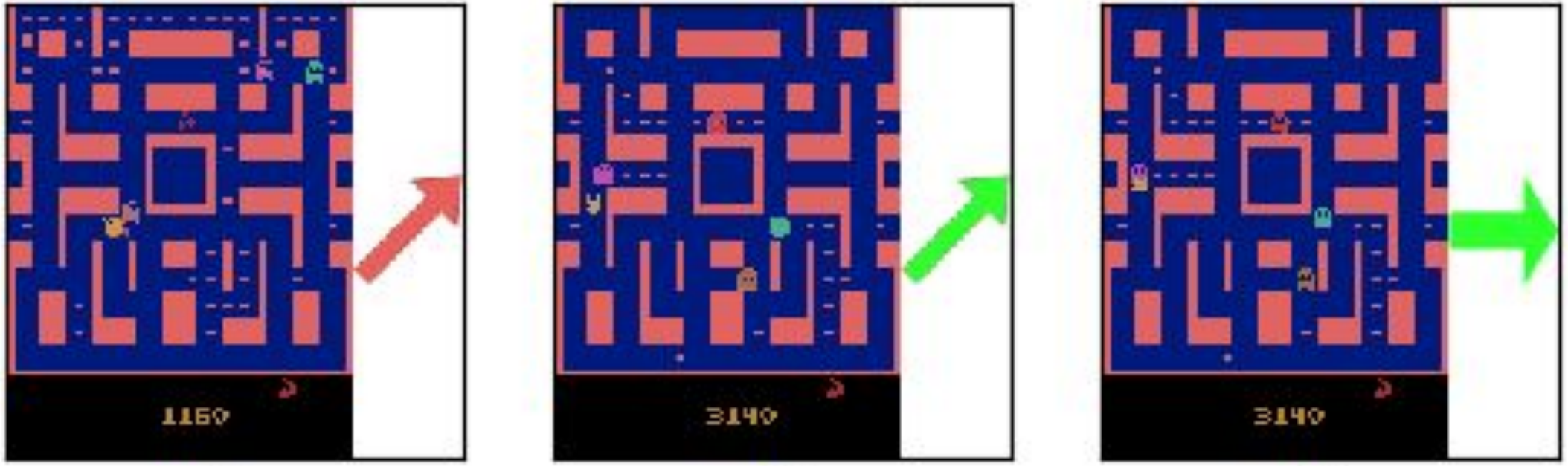
2



3

Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

Ms PacMan (after 2.4M iterations)



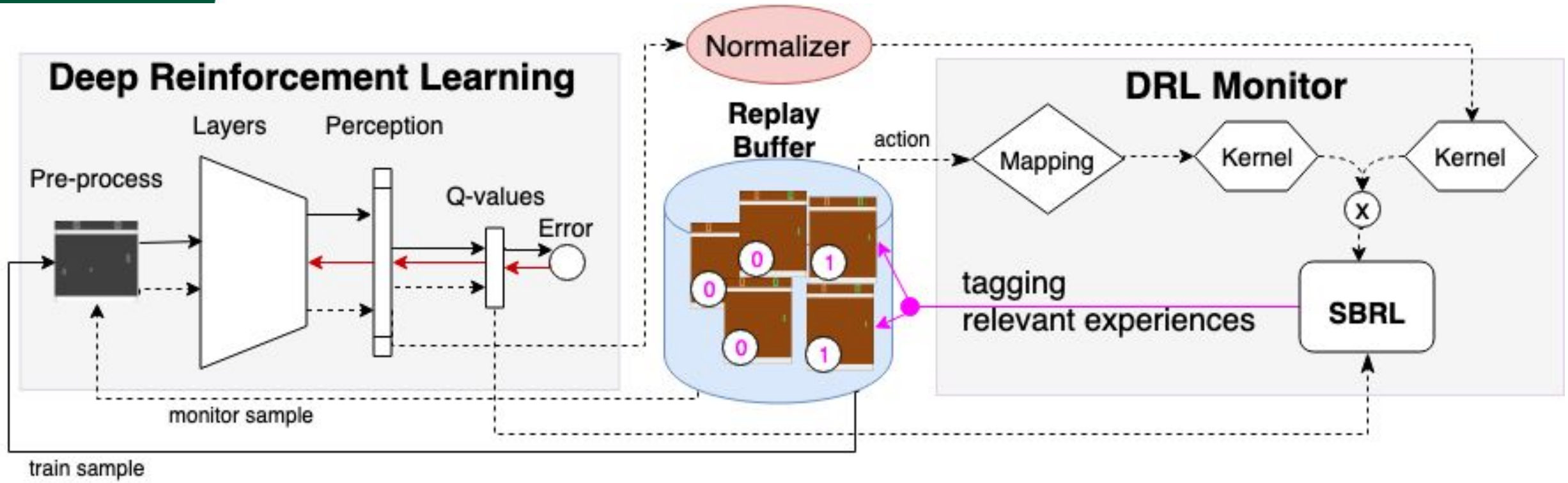
1 2 3

Starting to be scared with a ghost, but still craving for food to make a wrong decision.

Dao and Lee (2018). Deep Reinforcement Learning Monitor for Snapshot Recording.

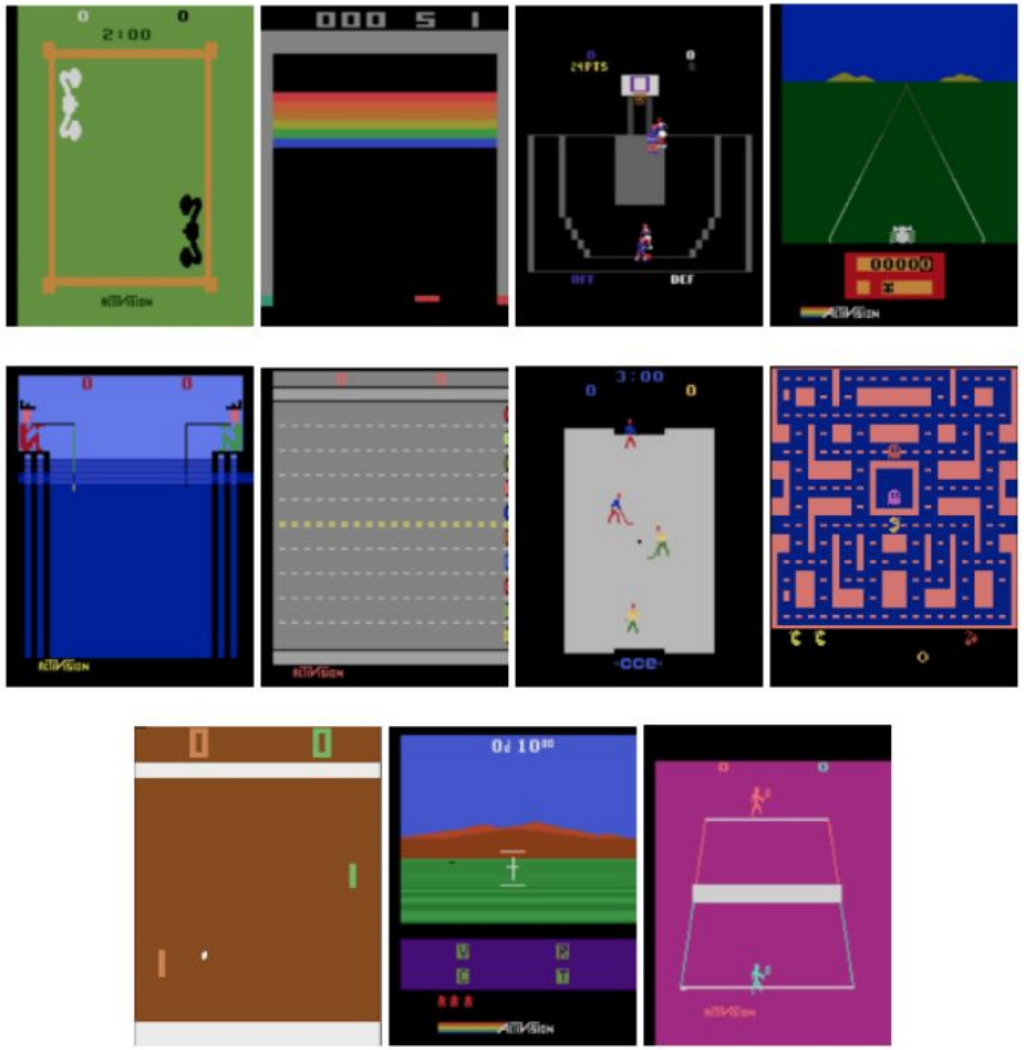
Overcoming Interpretability-Performance Tradeoff

Relevance Experience Replay



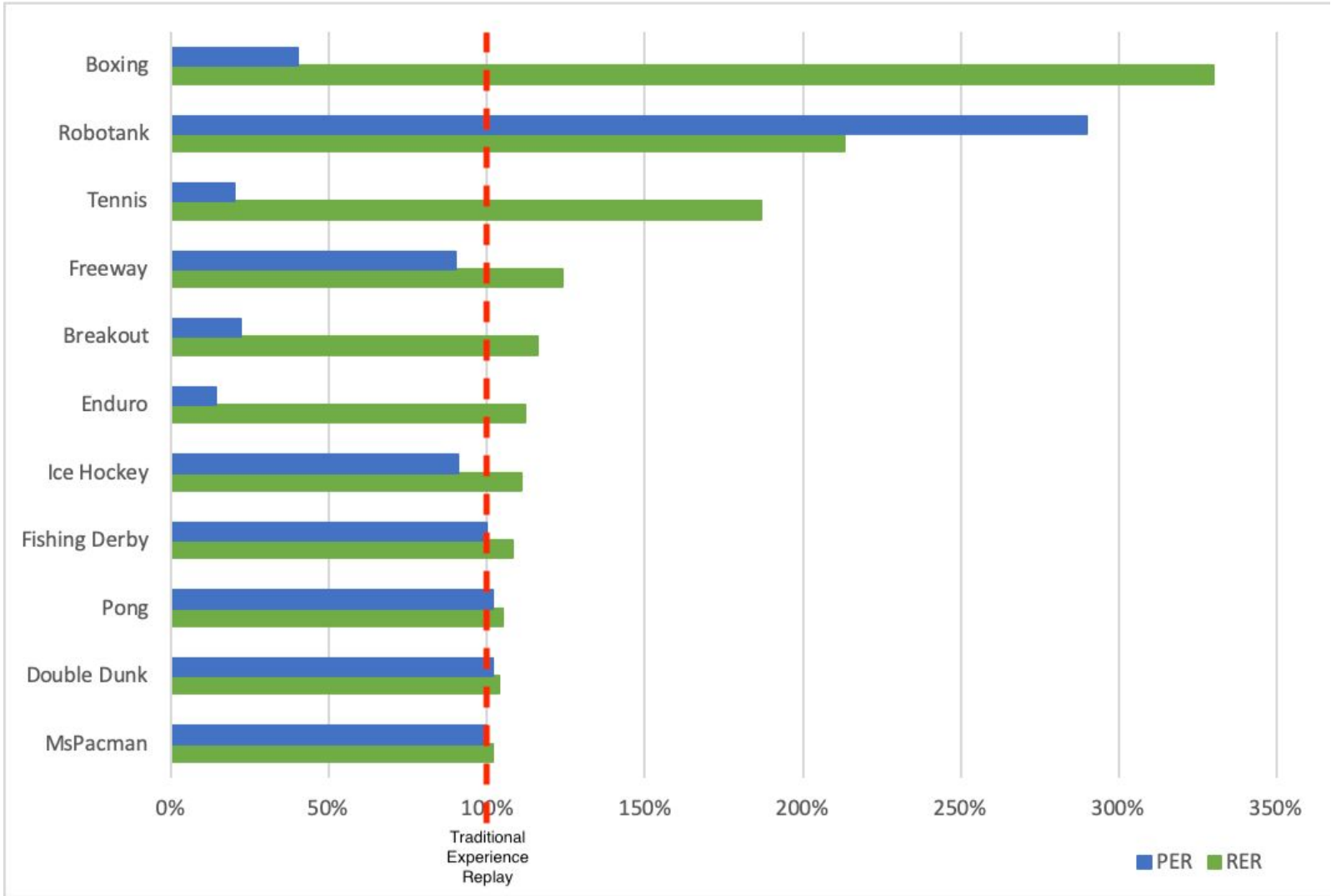
Dao and Lee (2019). Relevant Experiences in Replay Buffer

Atari Games



Dao and Lee (2019). Relevant Experiences in Replay Buffer

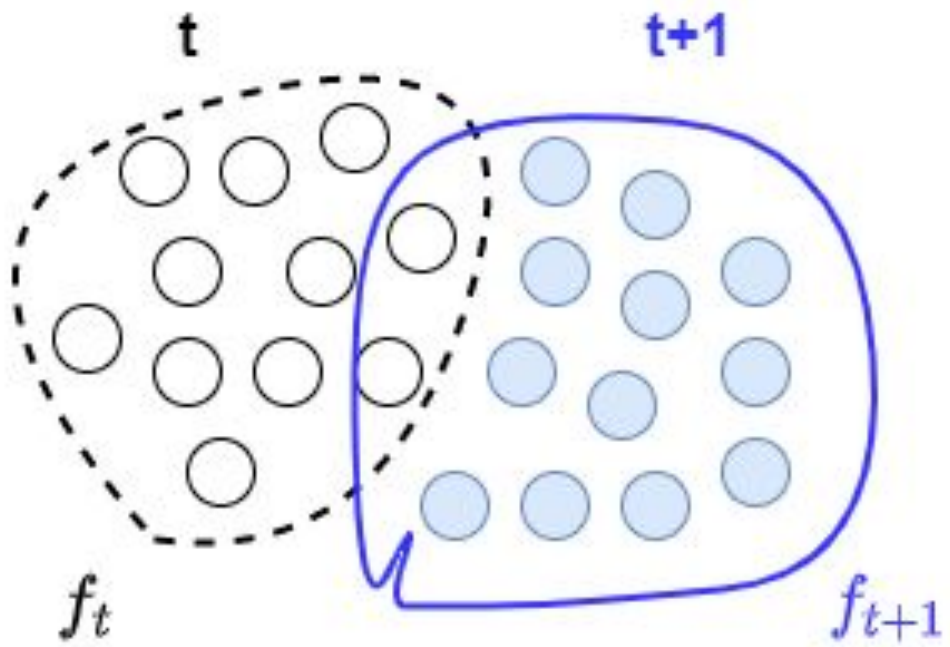
Atari Games



Dao and Lee (2019). Relevant Experiences in Replay Buffer

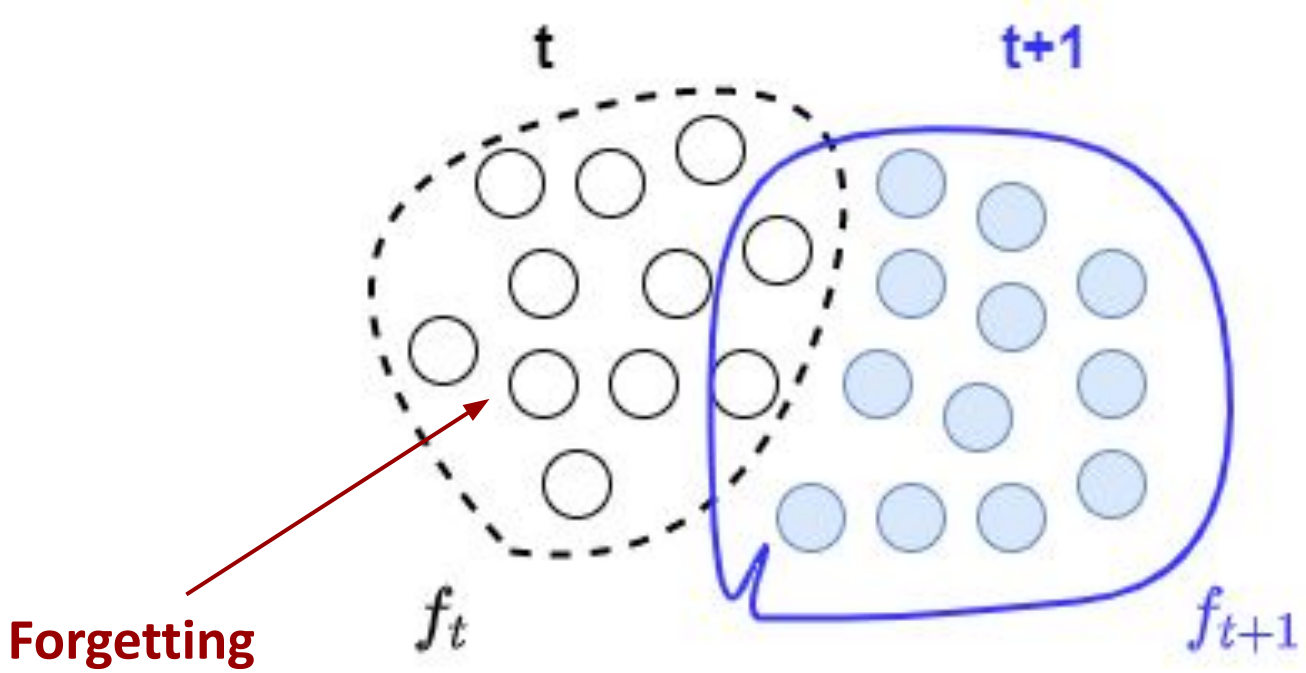
Continual / Lifelong Learning

Continual Deep Reinforcement Learning



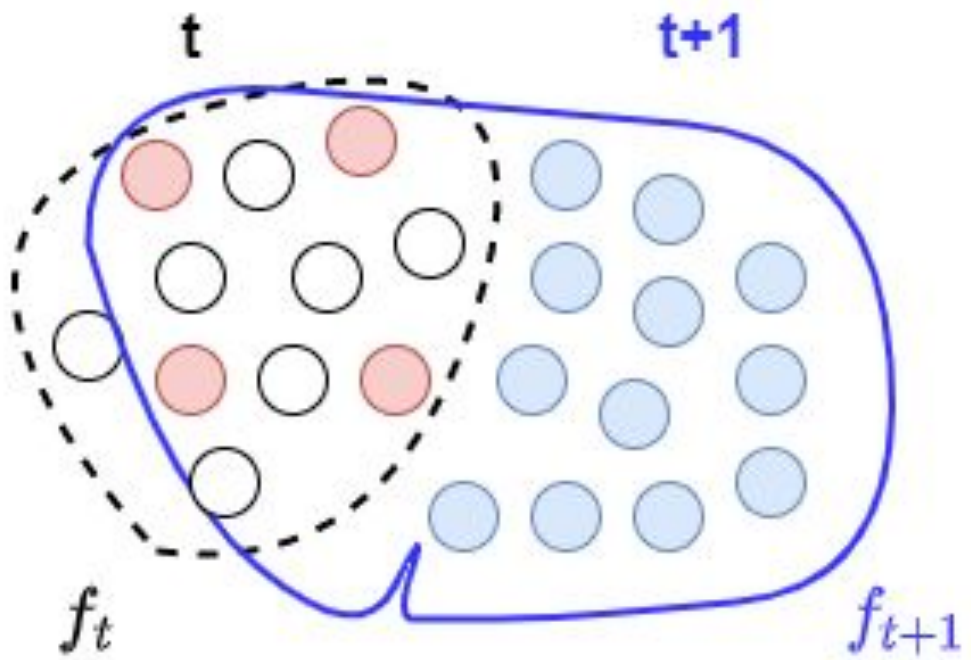
Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

Continual Deep Reinforcement Learning



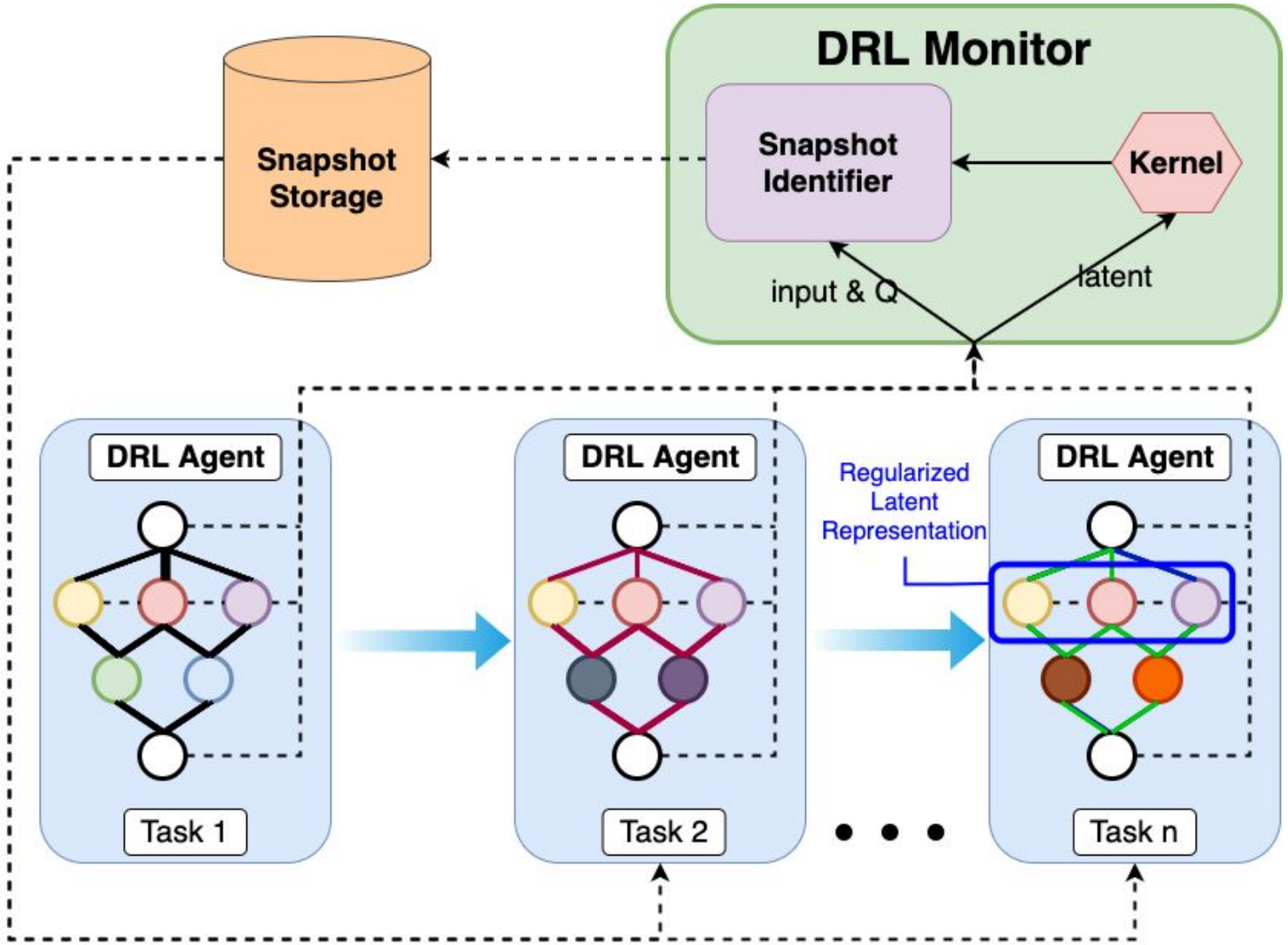
Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

Continual Deep Reinforcement Learning



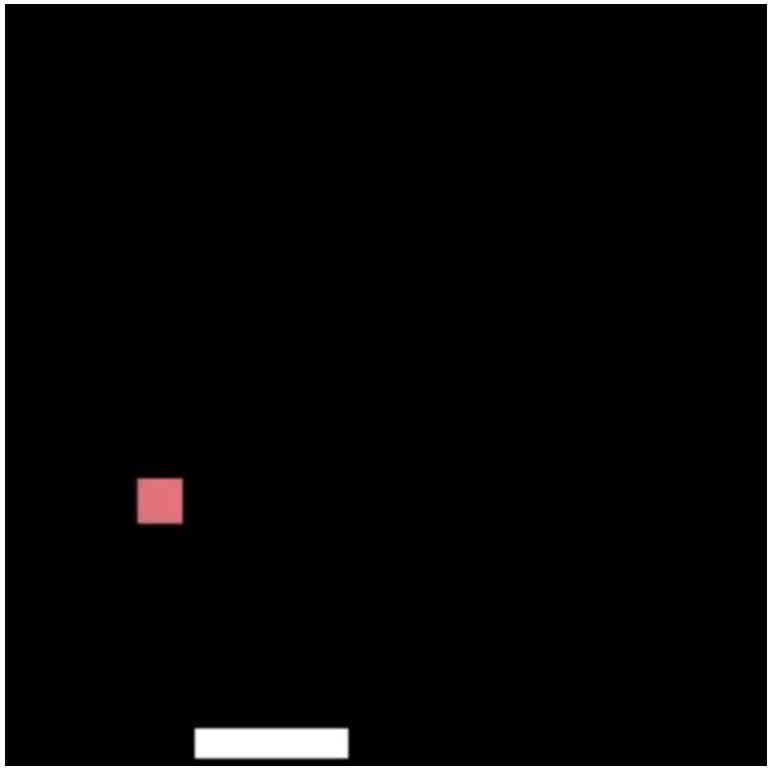
Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

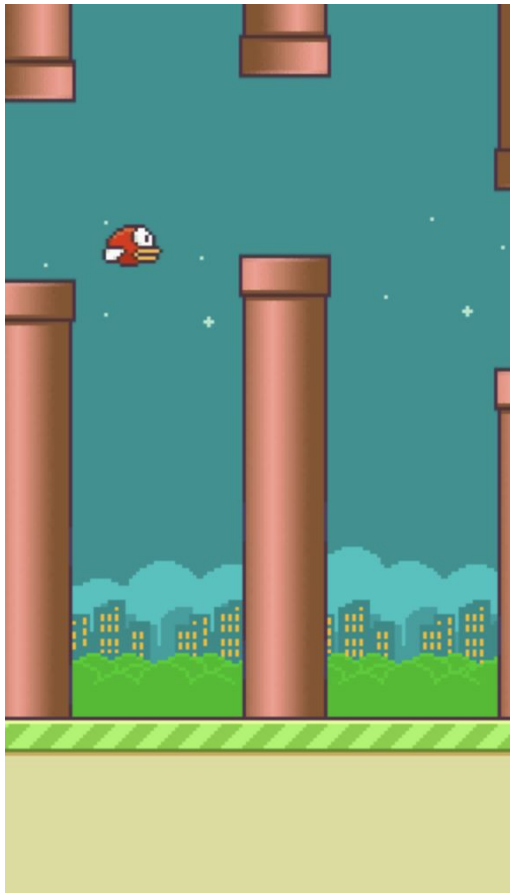
Continual Deep Reinforcement Learning



Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

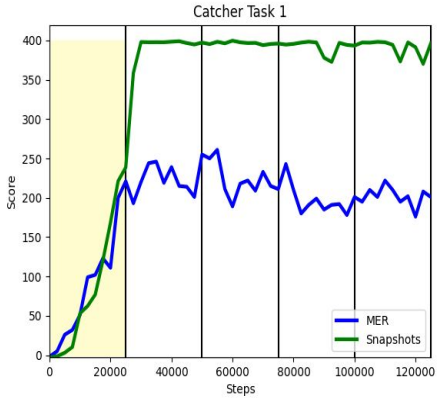
Catcher and Flappy Bird

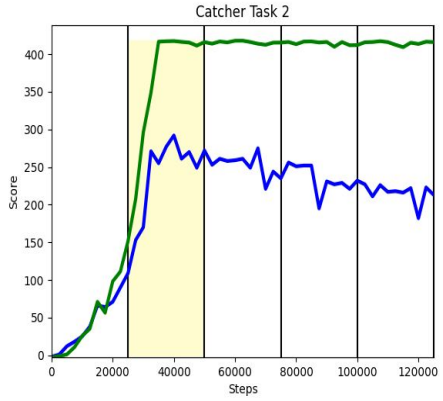


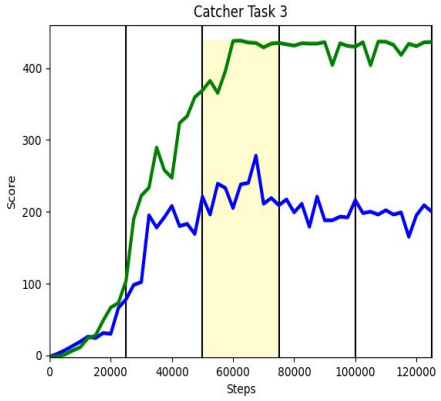


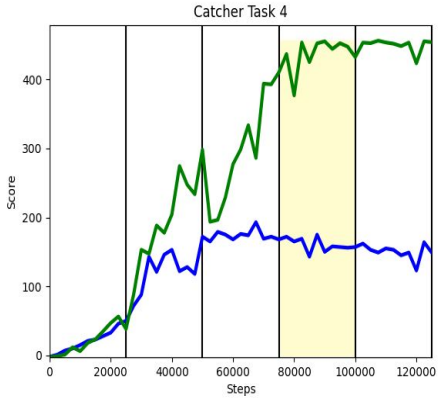
Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

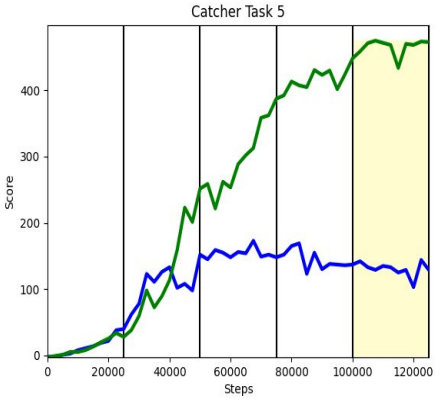
Catcher and Flappy Bird

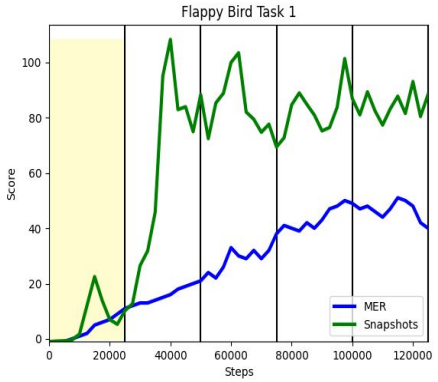


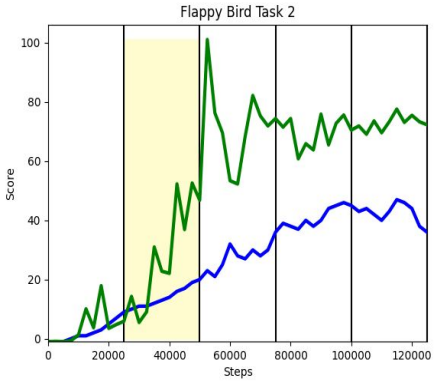


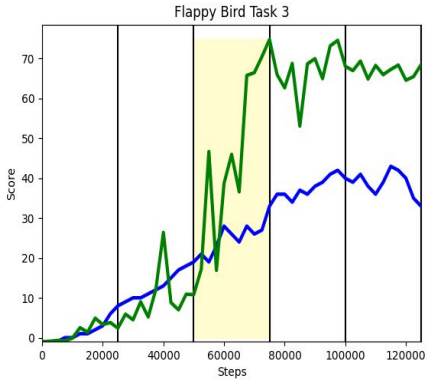


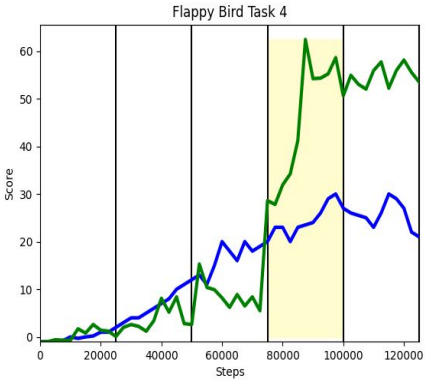


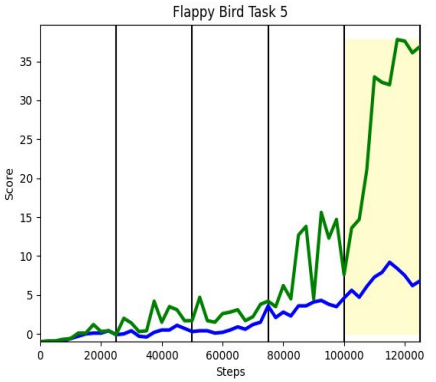






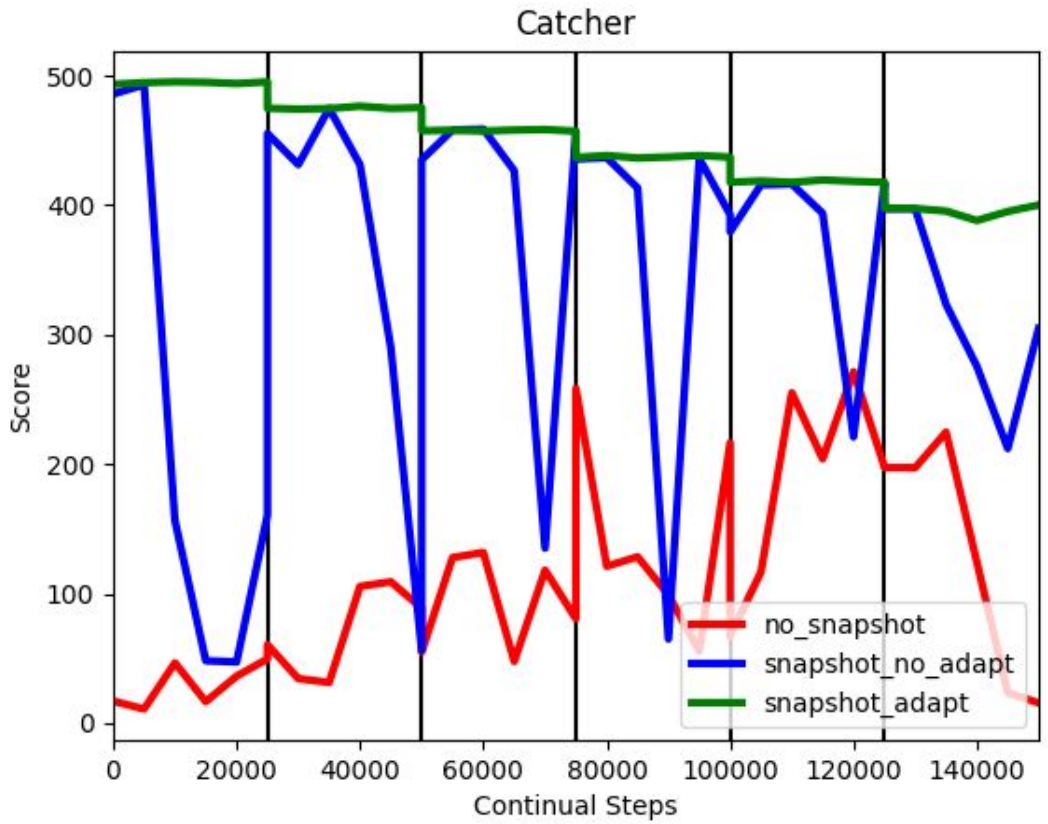


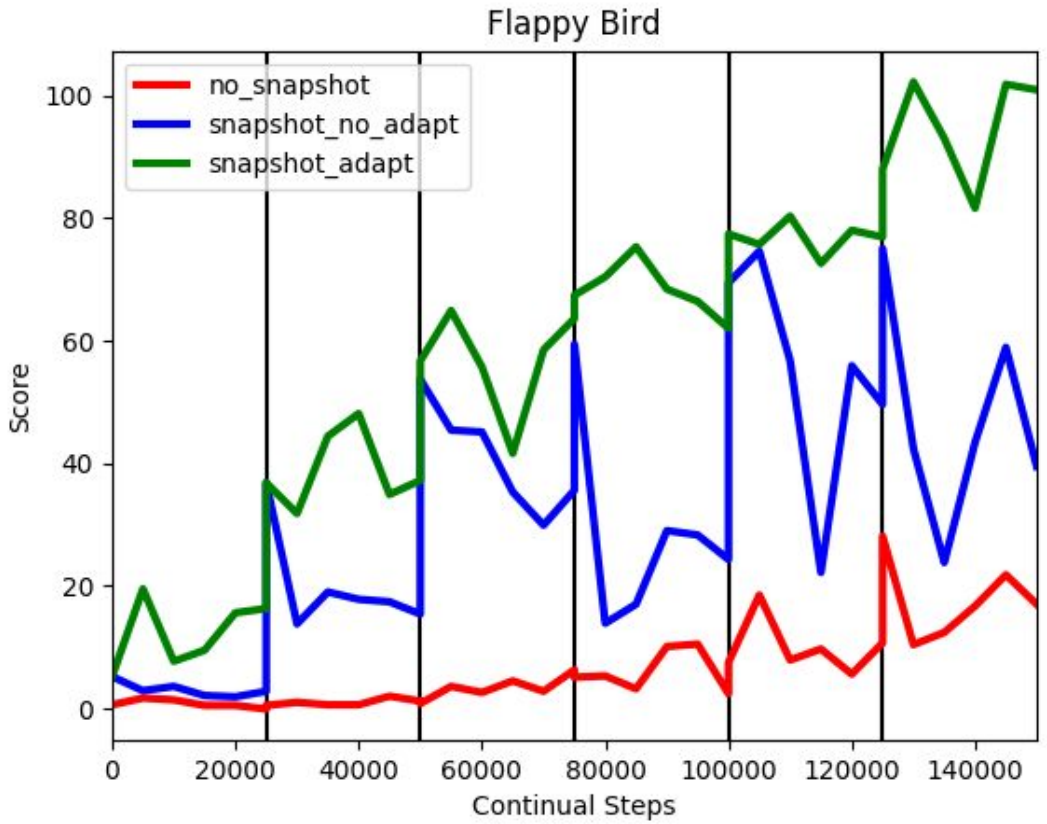




Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

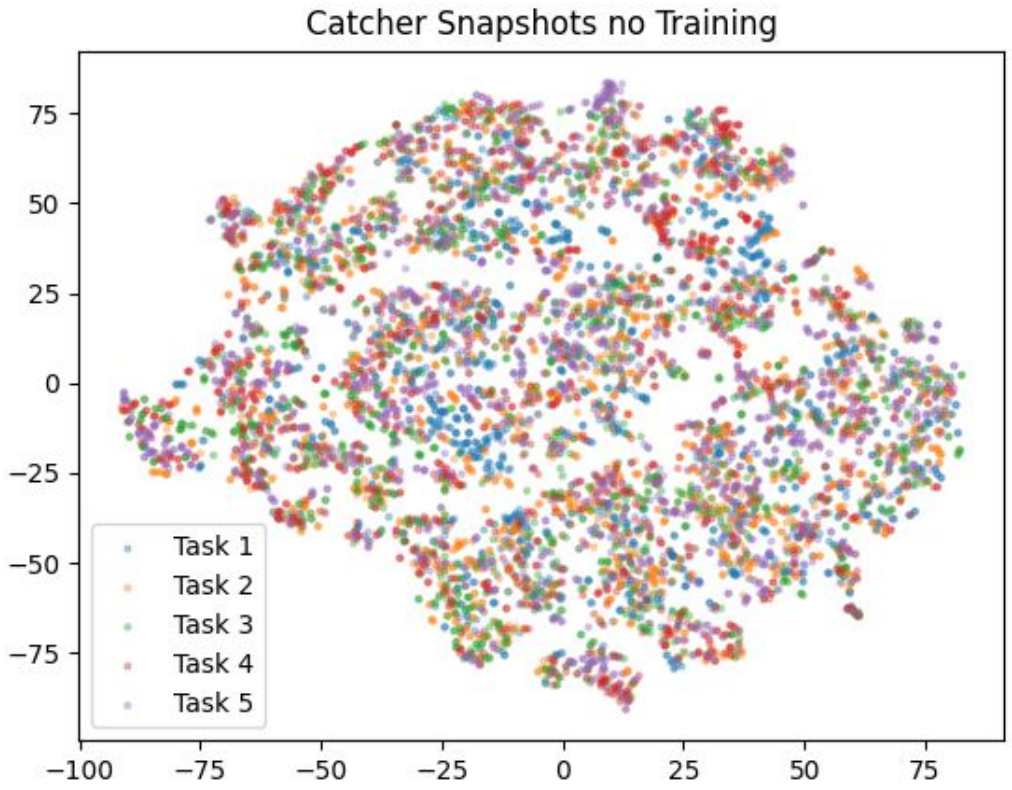
Catcher and Flappy Bird

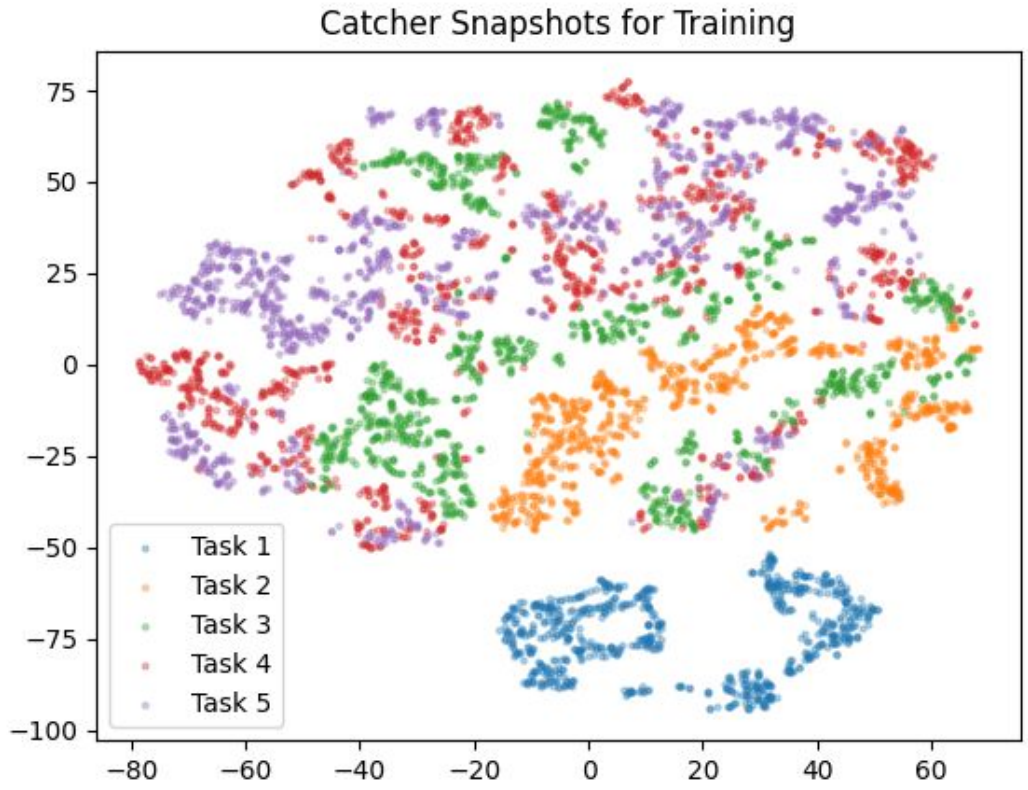




Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

Catcher and Flappy Bird





Dao and Lee (2021). Snapshots Storage for Continual Deep Reinforcement Learning

Closing…

- Evidence-Driven Reinforcement Learning
- Interpretation of Evidence
- Use of Evidence for Efficient Learning
- Evidence for Continual Learning
  - Mitigate forgetting problem
- Can be combined with existing methods
- Can collect evidence in all training stages of DRL

Are the most representative samples are evident?

Closing…

- Evidence-Driven Reinforcement Learning
- Interpretation of Evidence
- Use of Evidence for Efficient Learning
- Evidence for Continual Learning
  - Mitigate forgetting problem
- Can be combined with existing methods
- Can collect evidence in all training stages of DRL

Further exploration:

- Evidence-driven communication with Human
- Kernel tricks for stronger evidence selection
- Non-Bayesian Approximation
- Sparse Evidence / Local Evidence

Sparse Bayesian Reinforcement Learning
Knowledge Representation

Intrinsic Interactive Reinforcement Learning
Brain-Computer Interfaces
Human + AI

Giang Dao
PhD student

Benjamin Poole
PhD student

@ SHAIR Lab

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE