EXPLORING REINFORCEMENT LEARNING FOR GAME PLAYING AGENTS

BEKZAT ONGDASSYNOV, DILNAZ SEMBEKOVA, YULIYA BARKO, ASSYL RAKHMASHEV, YERZHAN YERBATYR

HTTPS://WWW.CANVA.COM/DESIGN/DAGW7PVYJMK/H5PUTSZLNL9KC3NFHWFA3Q/EDIT

MOTIVATION

- WHY RL FOR GAME PLAYING?
- WHY DONKEY KONG GAME?
  - UNIQUE CHALLENGES
  - UNDEREXPLORED GAME IN CONTRAST WITH HEAVILY STUDIED ATARI GAMES
- WHY BREAKOUT GAME?
  - SIMPLE GAME MECHANICS

REINFORCEMENT LEARNING

STATE

ACTION SPACE

POLICY

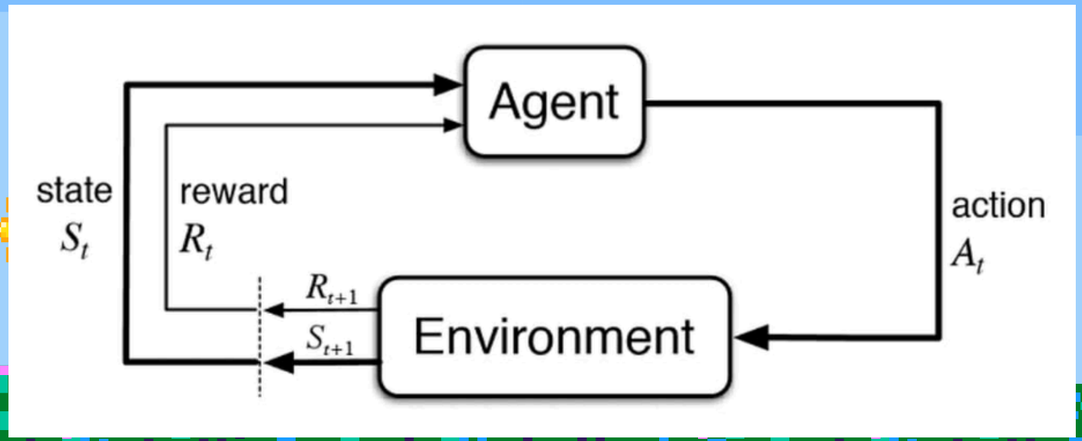
REWARD

RETURN

VALUE FUNCTION

EXPLORATION

EXPLOITATION



https://spinningup.openai.com/en/latest/spinningup/rl_intro.html

ENVIRONMENTS AND PREPROCESSING

GYMNASIUM ENVIRONMENT [GYM.MAKE("ALE/DONKEYKONG-V5")]

↓

STABLE-BASELINES3

↓

ATARI WRAPPER [PREPROCESSING]

↓

VECTORIZED ENVIRONMENT [STABLE-BASELINES3] - METHOD FOR STACKING MULTIPLE INDEPENDENT ENVIRONMENTS INTO A SINGLE ENVIRONMENT

↓

VECFRAMESTACK [VECENV]

↓

VECTRANSPOSEIMAGE [VECENV] - FROM HXWXC TO CXHXW. REQUIRED FOR PYTORCH CONVOLUTION LAYERS

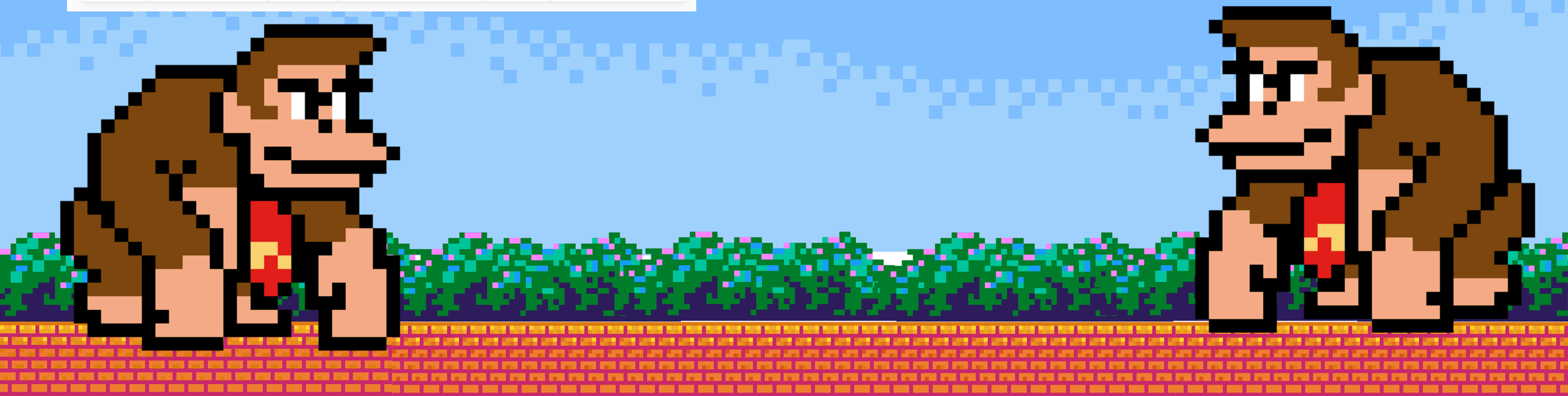
Name	Box	Discrete	Dict	Tuple	Multi Processing
DummyVecEnv	✔	✔	✔	✔	✘
SubprocVecEnv	✔	✔	✔	✔	✔

DONKEY KONG

Value	Meaning	Value	Meaning	Value	Meaning
0	NOOP	1	FIRE	2	UP
3	RIGHT	4	LEFT	5	DOWN
6	UPRIGHT	7	UPLEFT	8	DOWNRIGHT
9	DOWNLEFT	10	UPFIRE	11	RIGHTFIRE
12	LEFTFIRE	13	DOWNFIRE	14	UPRIGHTFIRE
15	UPLEFTFIRE	16	DOWNRIGHTFIRE	17	DOWNLEFTFIRE

```
Starting Bonus Value (each Screen)      5000 points
Jumping a barrel or fireball             100 points
Eliminating a Rivet                      100 points
Smashing a barrel or Fireball            800 points

The player recieves three Marios per game
```

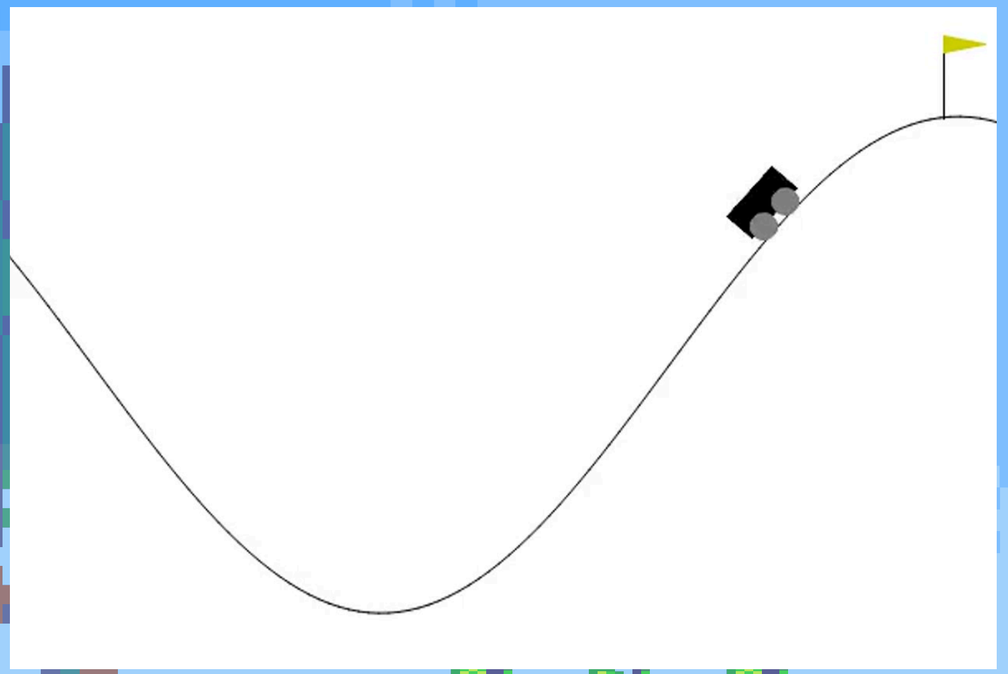



ALGORITHMS
1 DQN
2 A2C and A3C
3 PPO

DQN

Deep Q Network

Mountain car

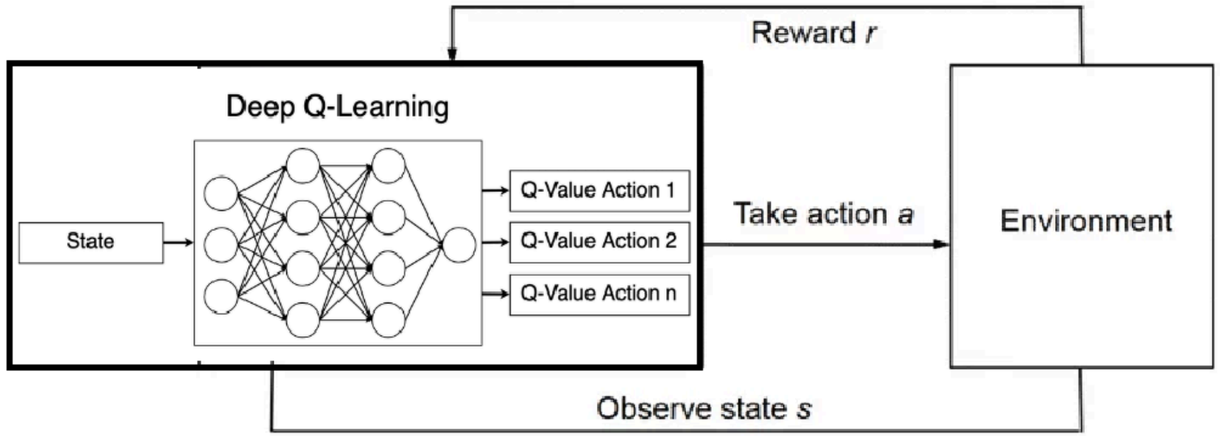


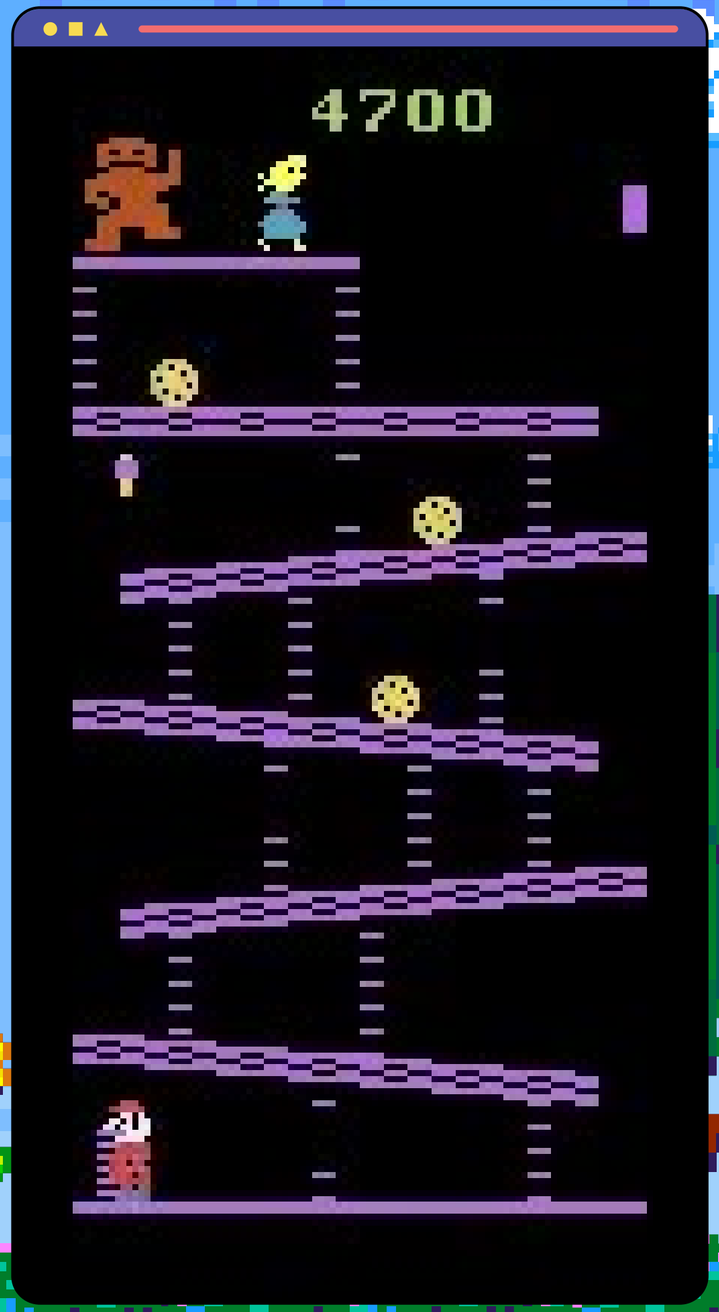
Breakout

DQN

$$Q(s, a; \theta) = r + \gamma \max_{a'} Q(s', a'; \theta')$$

Equation 1: Bellman's Equation for the DQN algorithm.



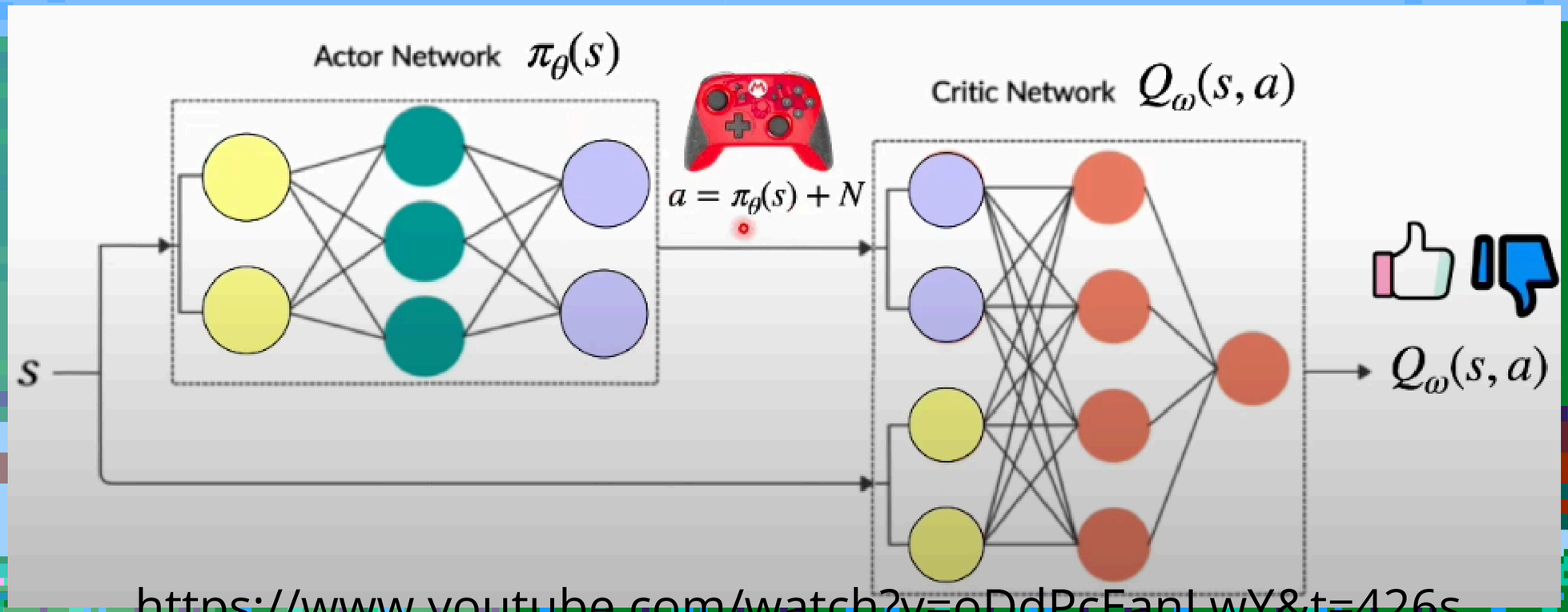


MNIH, V., KAVUKCUOGLU, K., SILVER, D. ET AL. HUMAN-LEVEL CONTROL THROUGH DEEP REINFORCEMENT LEARNING. NATURE 518, 529–533 (2015). HTTPS://DOI.ORG/10.1038/NATURE14236

A2C & A3C

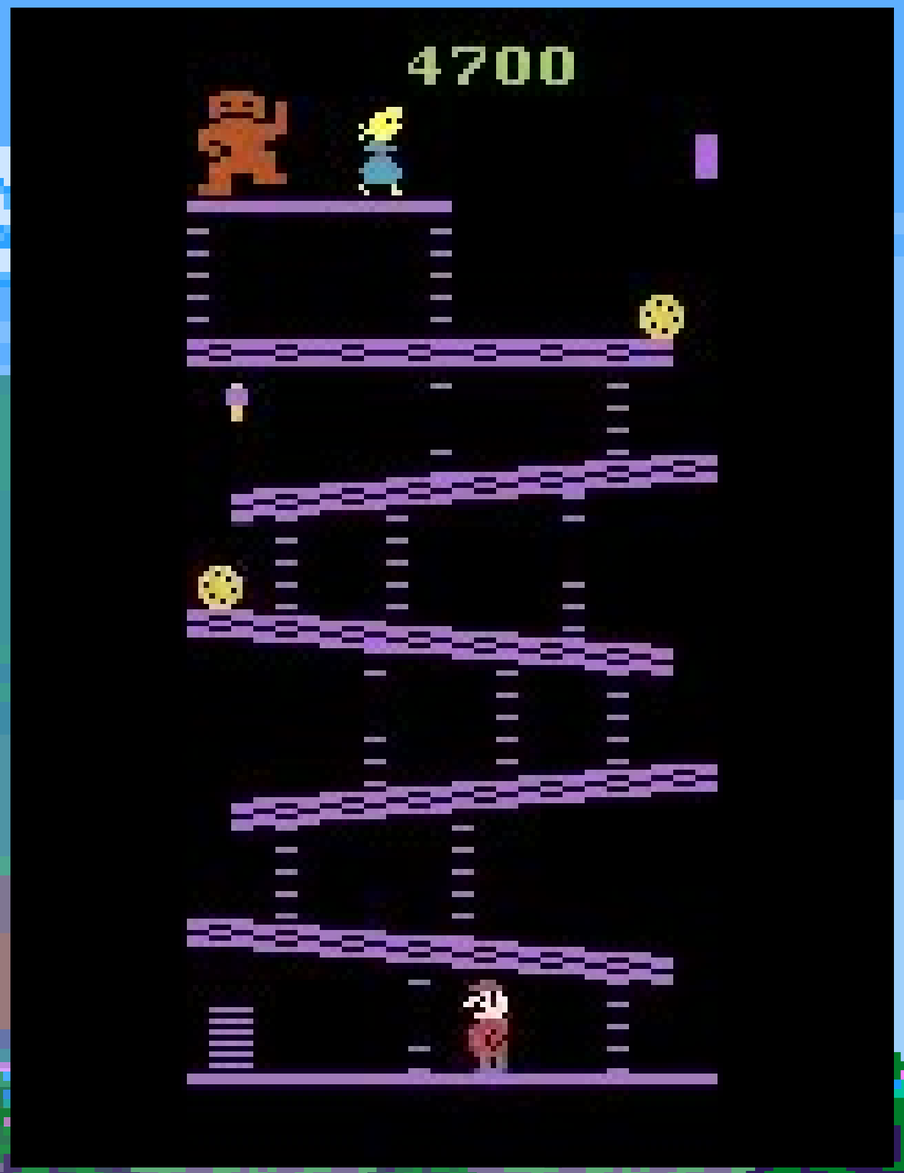
Actor => learns policy $\pi_\theta(a|s)$

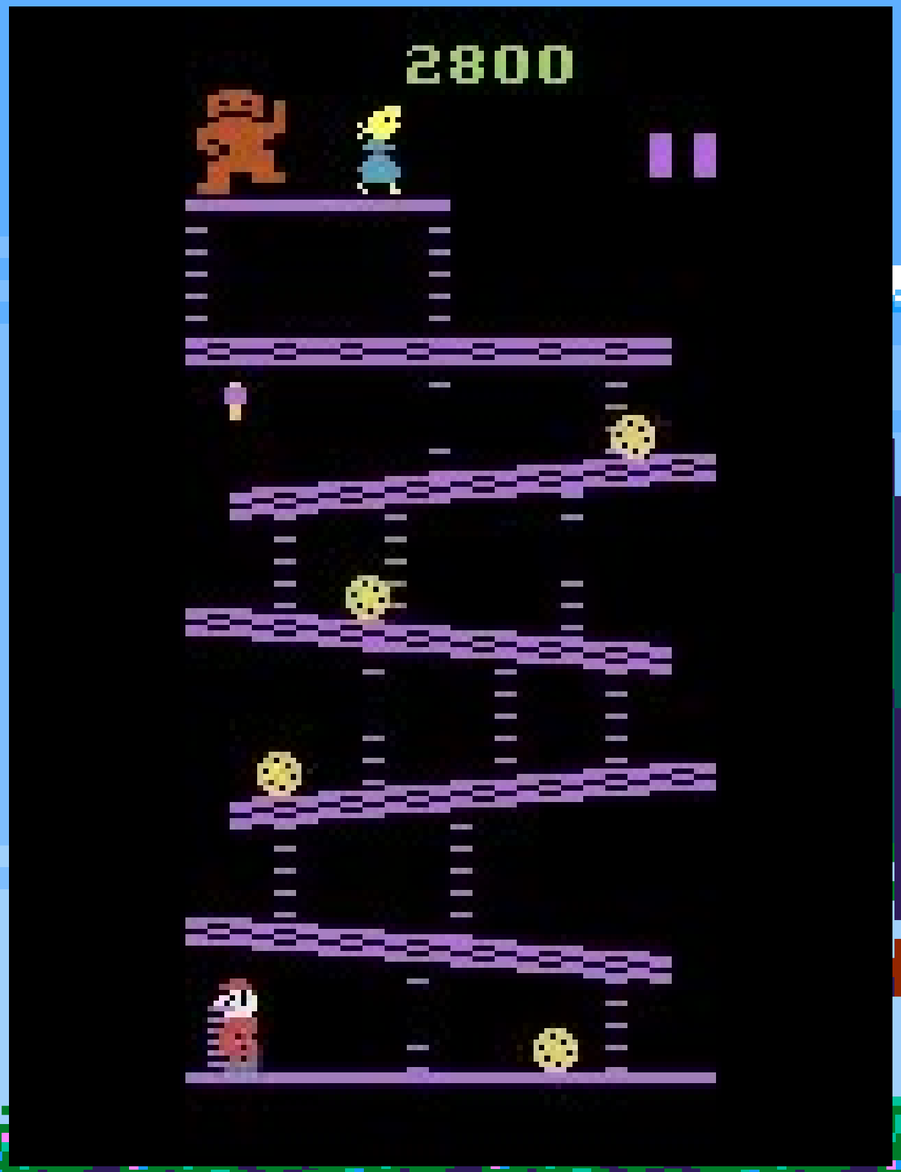
Critic => learns value function $Q_\omega(s,a)$



https://www.youtube.com/watch?v=oDdPcEanLwY&t=426s

FROM ZERO TO HERO





EXPLORATION VS EXPLOITATION

Trying new actions to discover their rewards.

Ensures the agent learns about the environment.

Choosing known actions to give the best rewards.

Focuses on maximizing immediate gains.

Trade-off

Balancing

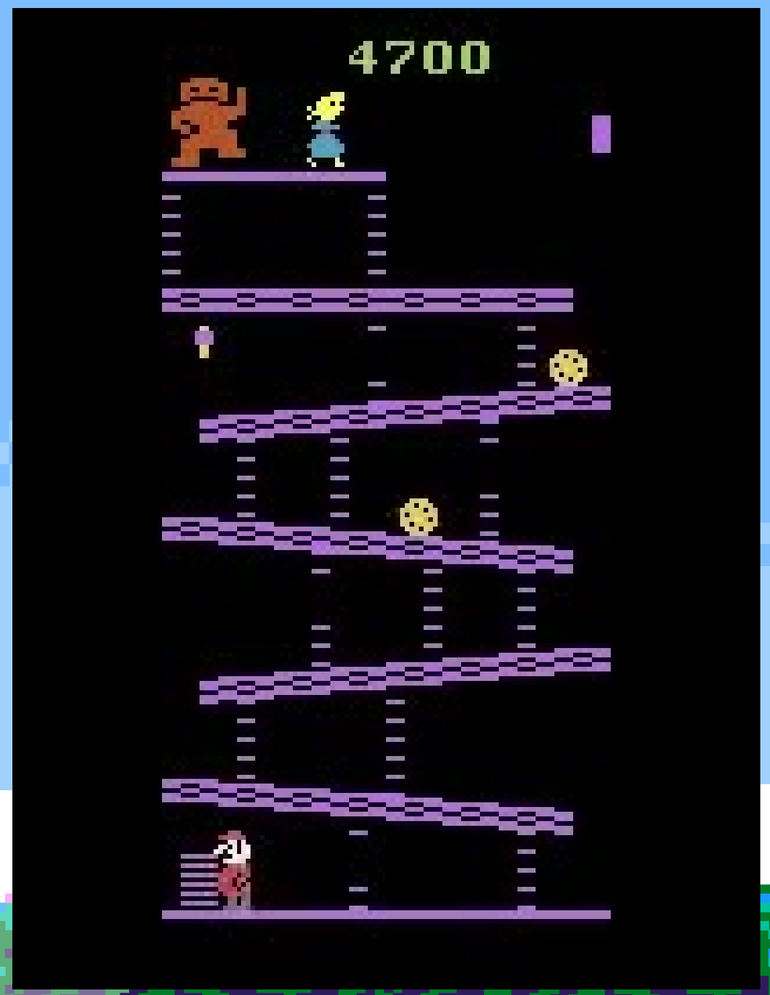
Too much exploration:
Slow convergence.

Too much exploitation:
Suboptimal solutions.

Epsilon-greedy strategy

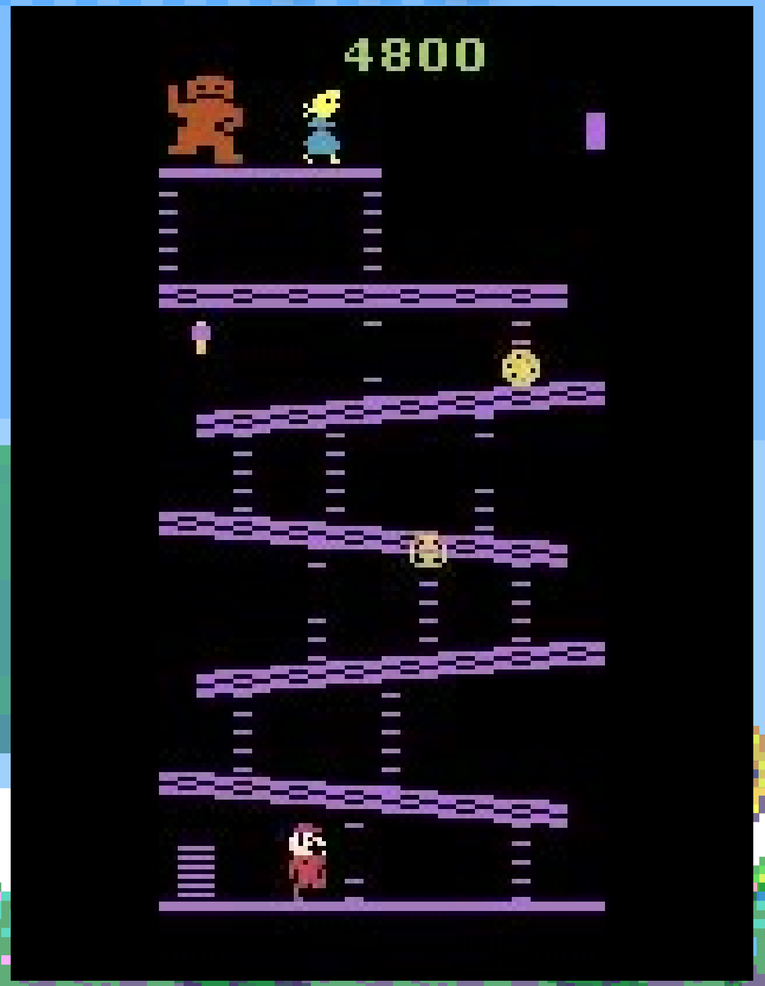
QUIZ-TIME

Where is the highest entropy? Lowest?

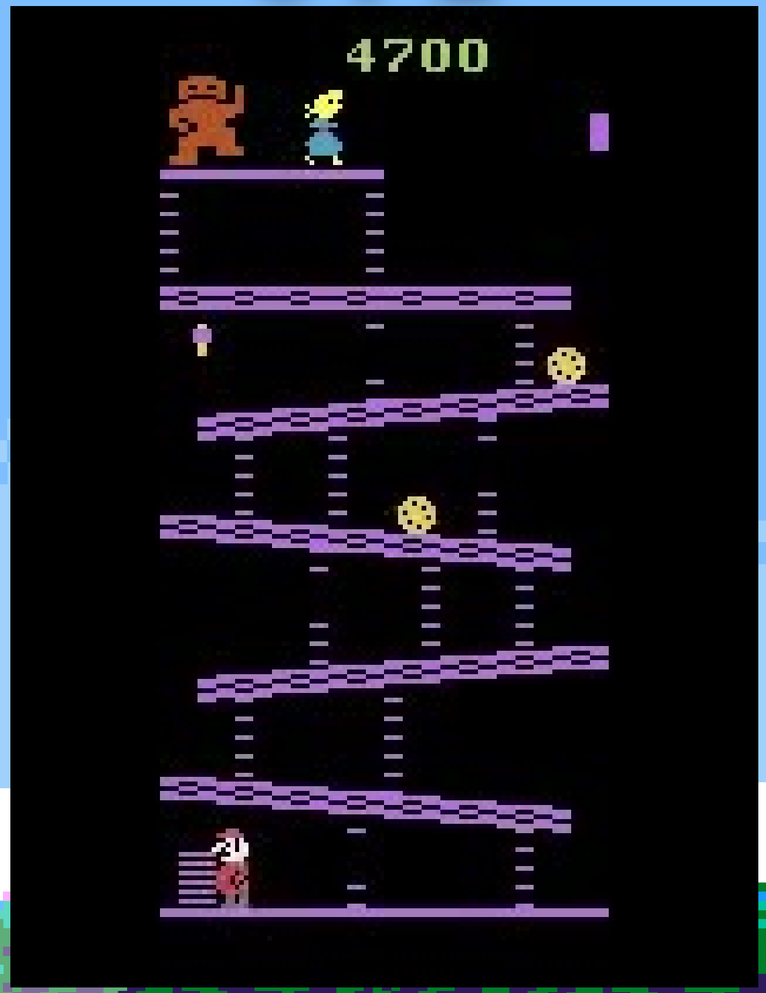


ANSWERS

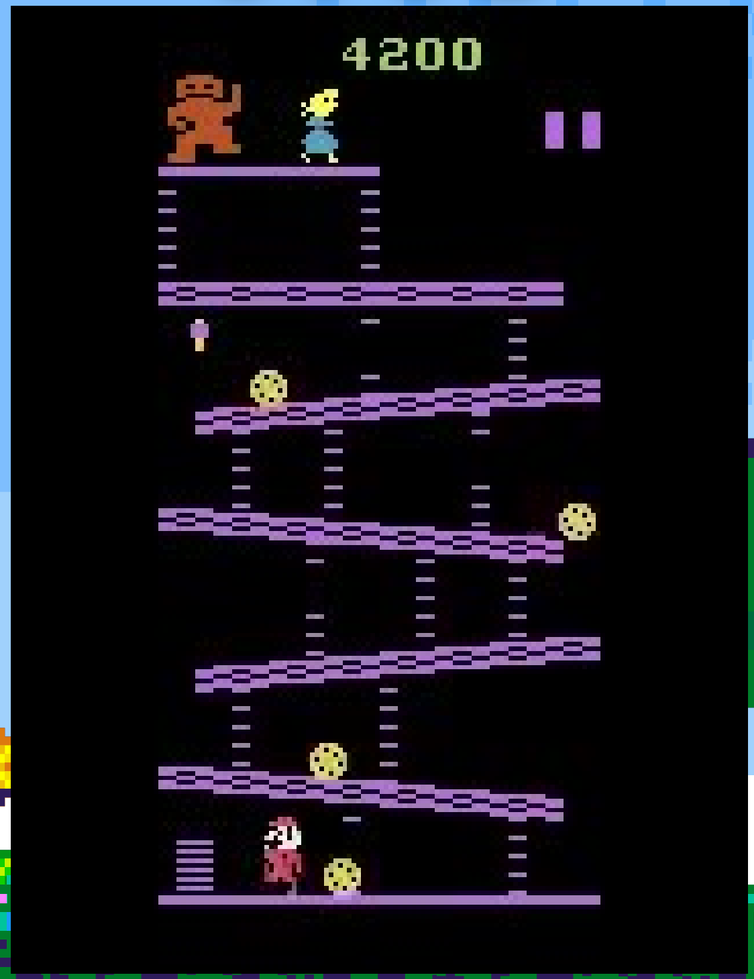
0.5



0.1

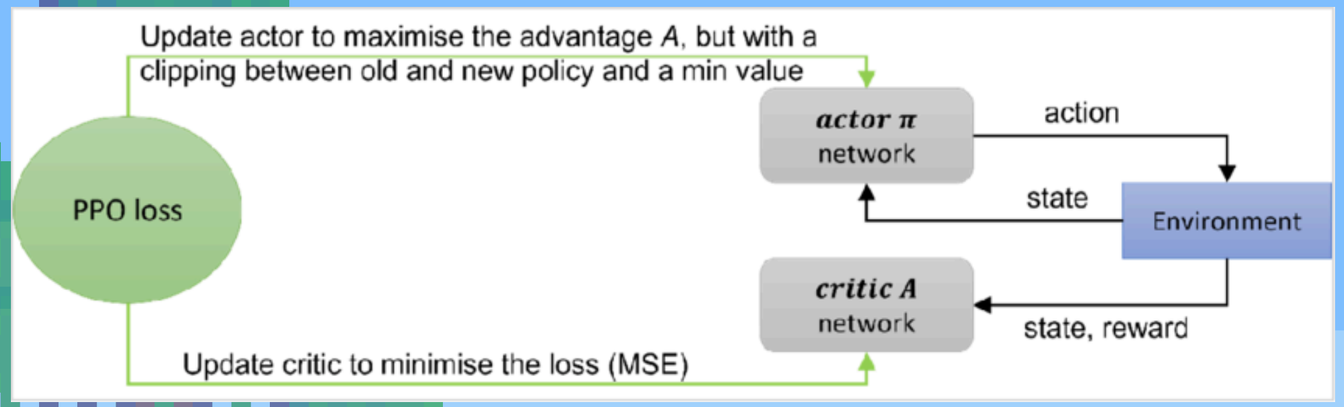


0.05



PPO

- IMPROVE THE AGENT'S POLICY GRADUALLY WHILE ENSURING IT DOESN'T CHANGE TOO MUCH AT ONCE
- WORKS WELL FOR TASKS WITH DISCRETE OR CONTINUOUS ACTION SPACES
- USES CLIPPING TO LIMIT LARGE CHANGES TO THE POLICY

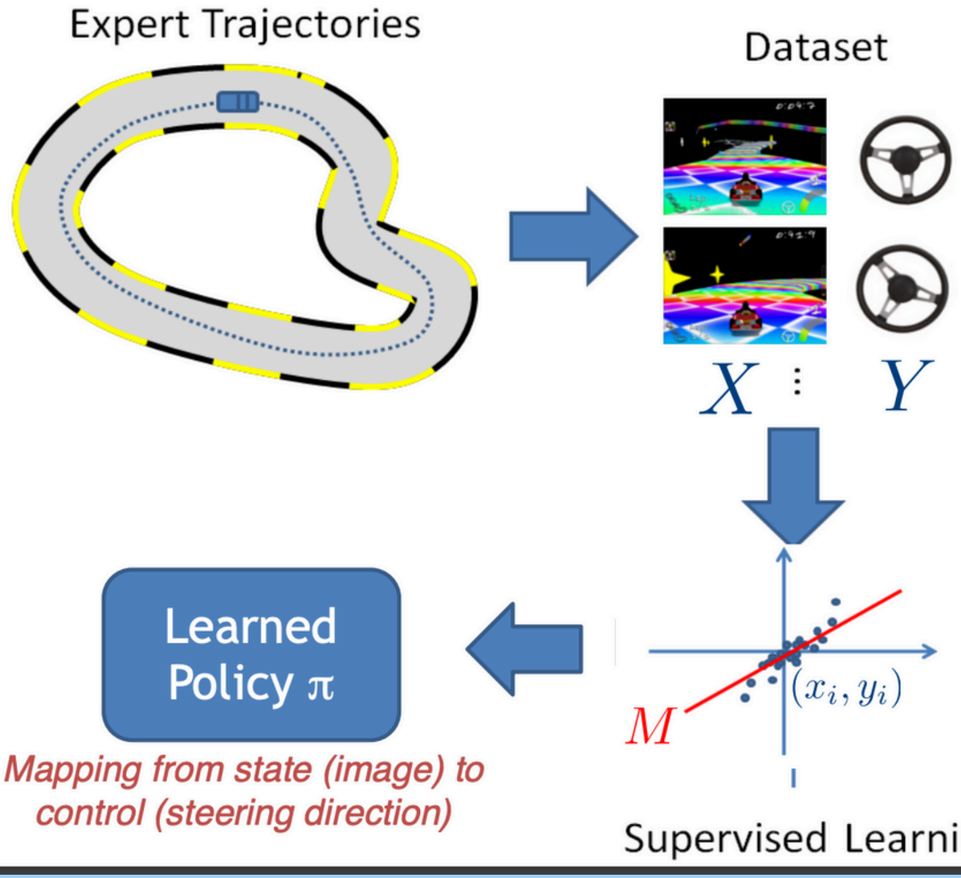


UNDERSTANDING REINFORCEMENT LEARNING ALGORITHMS: THE PROGRESS FROM BASIC Q-LEARNING TO PROXIMAL POLICY OPTIMIZATION - SCIENTIFIC FIGURE ON RESEARCHGATE. AVAILABLE FROM: HTTPS://WWW.RESEARCHGATE.NET/FIGURE/PROXIMAL-POLICY-OPTIMIZATION-PPO_FIG4_369759263 [ACCESSED 19 NOV 2024]

HOW TO MAKE HIM TAKE THE STAIRS???

[Widrow64, Pomerleau89]

Supervised Learning Approach: Behavior Cloning



10

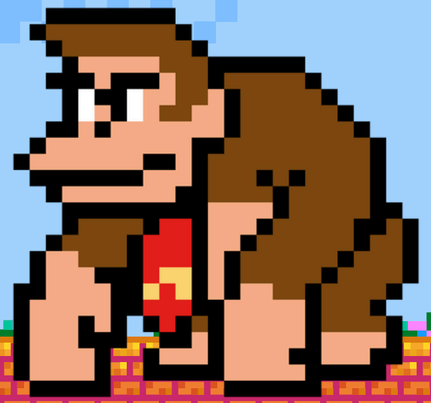
dataset:

[(state, action)]

https://wensun.github.io/CS4789_data/Imitation_Learning_April_8_annotated.pdf

CONCLUSION

1. DONKEY KONG ENVIRONMENT IS COMPLEX DUE TO REWARDS
2. DON'T USE DQN FOR COMPLEX ENVIRONMENTS, USE AC AND PPO INSTEAD
3. TRAINING TAKES HOURS
4. INTERPRETATION OF TRAINING IS HARD




THANK
YOU