Jump, Run and Learn: Reinforcement Learning Take on SuperMario Bros

Team: P24

Parashara	A0285647M	E1216292
Sriram	A0276528R	E1132261
Xian He	A0268425X	E1101724
Grace Ngu Sook Ern	A0268281X	E1101580

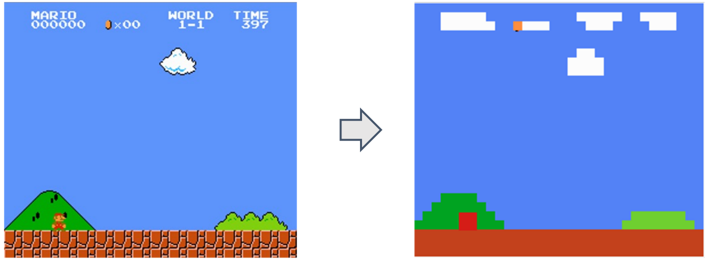
AY2023/24 Sem2
CS4246/CS5446

NUS Computing
National University of Singapore

Agenda

- SuperMario Bros Environment & Scope
- Project Objective
- Training Setup
- Understand DQN vs. PPO vs. A2C: Concept & Implementation
- Goal 1: Battle of the Models: DQN vs. PPO vs. A2C
- Goal 2: Custom Reward Function vs. Regular Reward Function
- Goal 3: Generalizability Test
- Conclusion

SuperMario Bros Environment & Scope



Mode	Coin-Collector (CC)	Regular
World-Stage	1-1, 1-2	
Version	3 (Rectangular mode is selected due to easier training)	
Action Space	COMPLEX_MOVEMENT from OpenAI Gym Example: {L, R, Jump, Down, No-Op, …}	
State	Pixels (Skip Frame, Grayscale, Resize, Stack Frame)	
Reward Function	How far right, speed, death, score (coins, enemy, etc.)	How far right, speed, death

Project Objective

- This project aims to study DQN vs. PPO vs. A2C on a single agent complex environment in 3 depths:
    - Goal 1: Battles of the models: DQN vs. PPO vs. A2C
        - Qualitative Analysis: Video Evaluation
        - Quantitative Analysis: Tensorboard Logs Evaluation
    - Goal 2: Custom reward function vs. Regular reward function
        - For each algo, compare behavioral differences between having environment with coin as a reward vs. no coins
    - Goal 3: Generalizability Test
        - Compare model's performance on unseen stage world 1-2

Training Setup

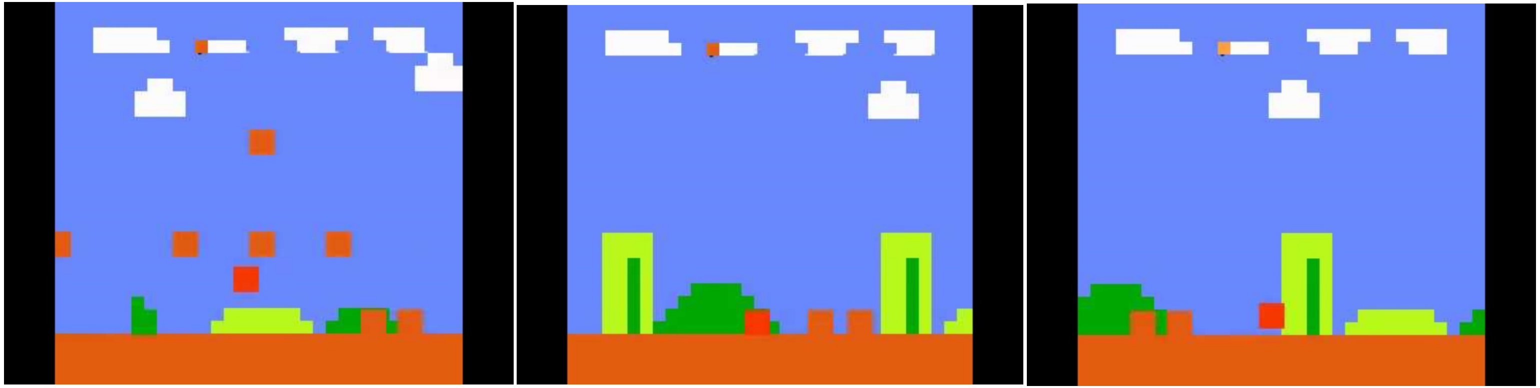
Mode	DQN	PPO	A2C
Package	Stable Baseline 3		
Modes Trained	Both CC and Regular Mode		
Training Steps	5 million	~2.5 million	600k
Video Evaluation	3 success plays, 3 failure plays		
Training Evaluation	Tensorboard Logs (Exploration rate, Entropy Loss, Policy Loss and Value Loss), Best Reward, Average Reward		
Testing Evaluation	Pass Count per 1000 plays, Coins Collected per 1000 plays		

Understanding DQN vs. PPO vs. A2C: Concept & Implementation

Model	DQN	PPO	A2C
Model type	Value-based	Policy-based	Actor-Critic
Algorithm Concept	Estimate Q-value of taking an action	Directly learns policy mapping states to action	Estimate the advantage of taking certain action by learning an actor (policy) and a critic (value) b
Sample efficiency & Stability Technique	Experience replay & target network	Clipped surrogate objective & entropy regularization	Parallel environments & Advantage loss

Goal 1: Battle of the Models (Video Evaluation)

DQN	PPO	A2C



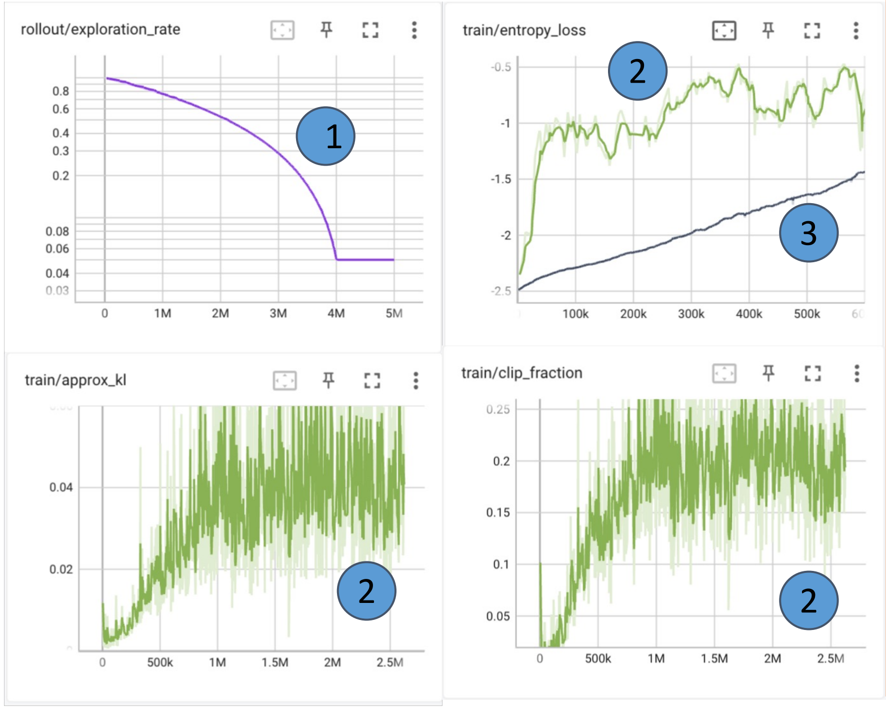
- **DQN** (20s): Conservative, trapped in local optima (pauses long before pipes and stairs) but occasional exploration help agent jump over obstacles but possibly run into monsters or cliffs
- **PPO** (16s): Moderately conservative, does not get trapped in local optima as much as DQN but tends to avoid monsters -> Lesser coins collected
- **A2C** (16s): Risk seeking, maximizes the coin gathered by crushing monsters and prefers to time the jumps

Understanding DQN vs. PPO vs. A2C: Behavioural traits

Model	DQN	PPO	A2C
Risk level	Conservative	Moderately conservative	Risk-seeking
Movements	Simple maneuvers (Non-precise)	Smoother and fluid movement	Complex maneuvers
Repetitive Behavior	Yes	No	No
Sample efficiency	High	Moderate	Low

AY2023/24 Sem2
CS4246/CS5446

Goal 1: Battle of the Models (Exploration vs. Exploitation)

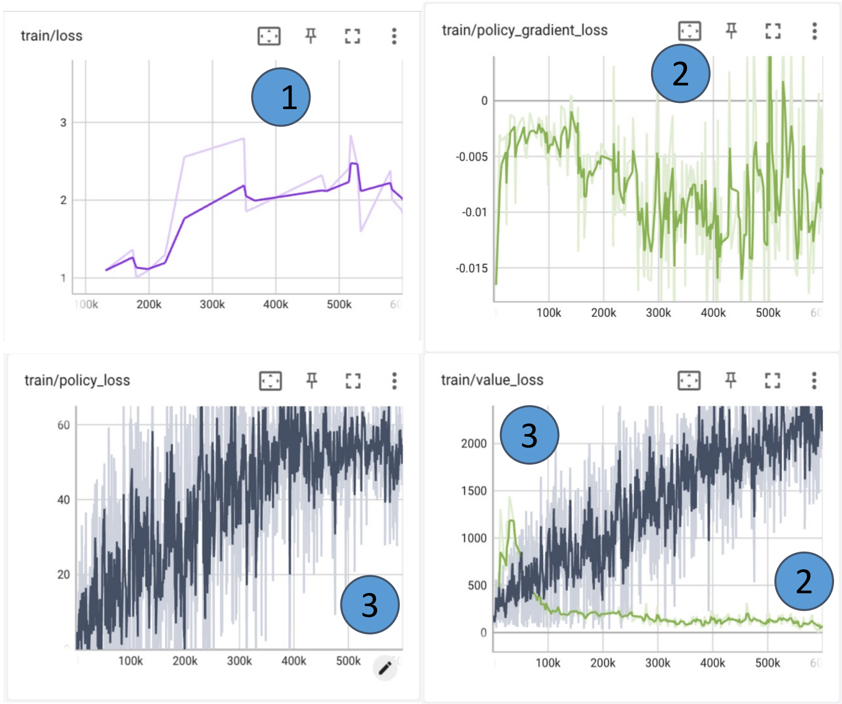


1. DQN: starts with high exploration and then to exploitation due to ε-greedy strategy
2. PPO: increases sharply to achieve high exploration then gradually increase later.PPO's entropy loss fluctuates as it has a clipping mechanism that clips it's policy changes. It encourages taking large updates without going too far.
3. A2C: slow increases exploration

Legend: — DQN — PPO – A2C

AY2023/24 Sem2
CS4246/CS5446

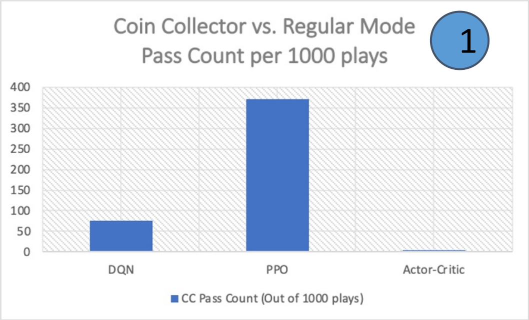
Goal 1: Battle of the Models (Policy & Value)

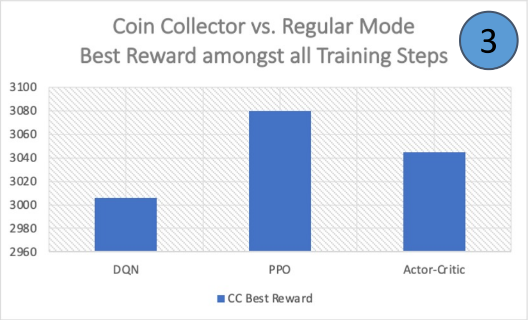


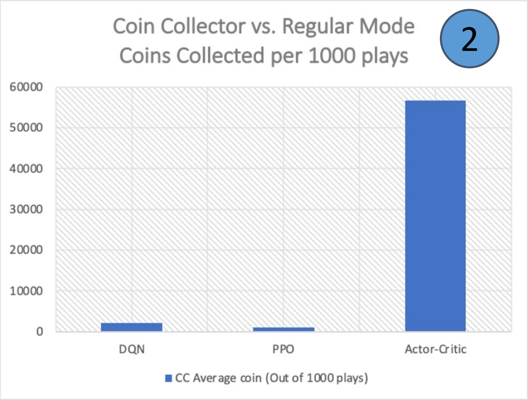
1. DQN: Overall loss increases over time. It may indicate agent has not explored enough in the early stages and make poor actions later.
2. PPO: Policy gradient loss decreases but increases again while value loss decreases indicate policy updates are too aggressive, causing policy to move away from optimal policies.
3. A2C: Policy and value loss increases over time. Model is diverging from the optimal policy.

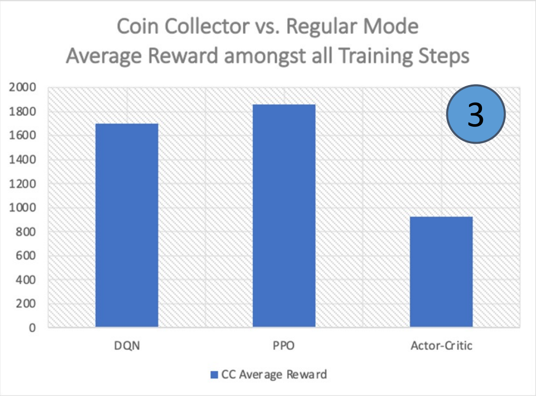
Legend: — DQN — PPO – A2C

Goal 1: Battle of the Models (Pass Count, Coins Collected, Training Best & Average Reward)





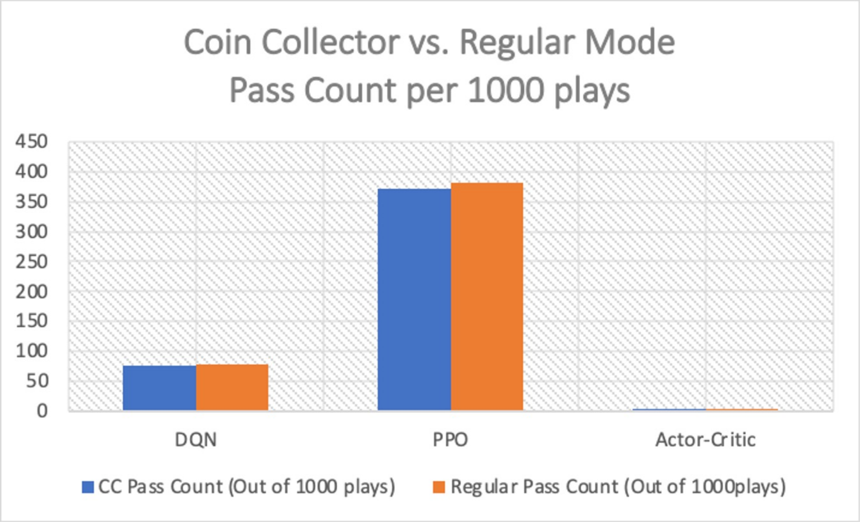


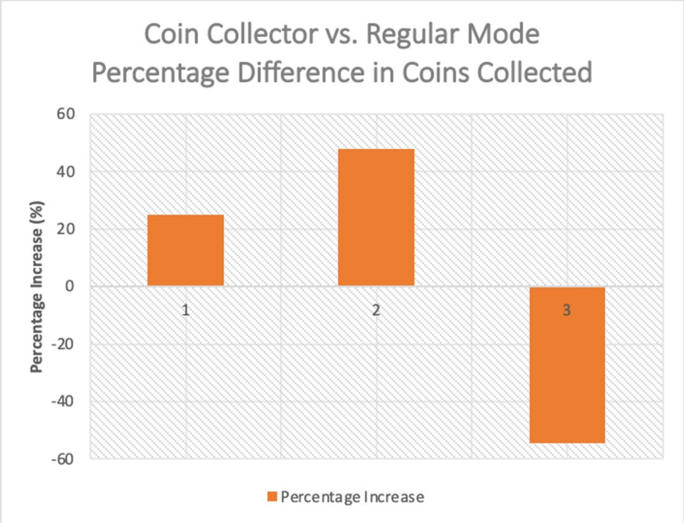


1. Most rounds completed: PPO > DQN > A2C
2. Coins collected: A2C > DQN > PPO
3. PPO outperforms DQN with lesser training steps (2.5 mil vs. 5 mil)
4. A2C showing higher best reward as compared to DQN due to the large number of coins collected

AY2023/24 Sem2
CS4246/CS5446

Goal 2: Custom Reward Function vs. Regular Reward Function

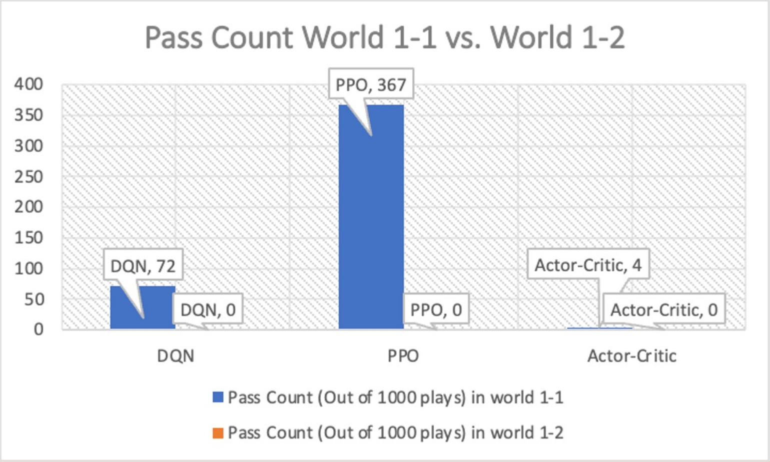


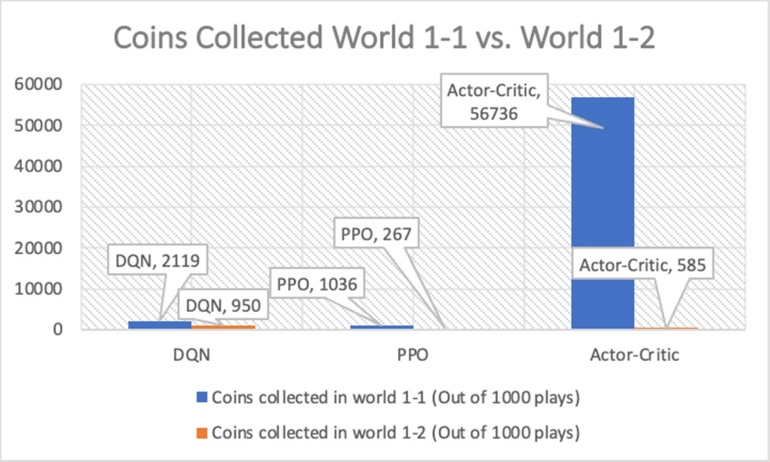


1. Using the CC mode led to higher coins collected as compared to regular mode when the agent is trained on CC mode.
2. While the pass rate stays almost the same.

AY2023/24 Sem2
CS4246/CS5446

Goal 3: Generalizability Test



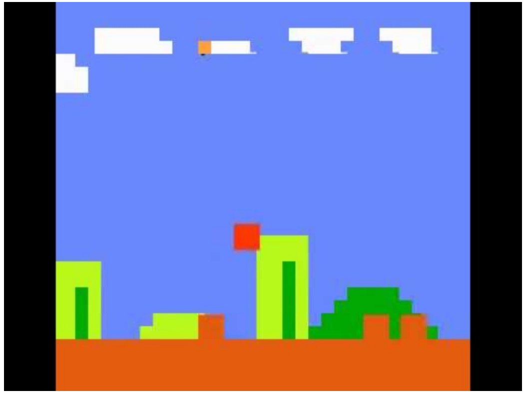
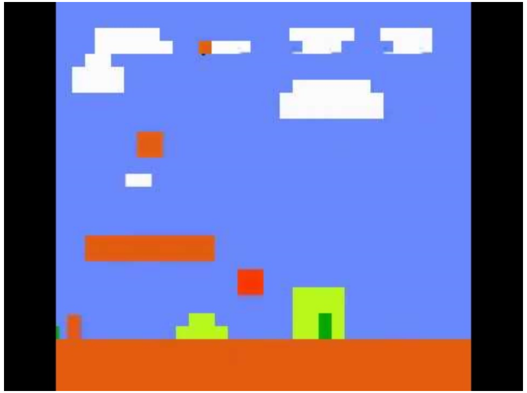
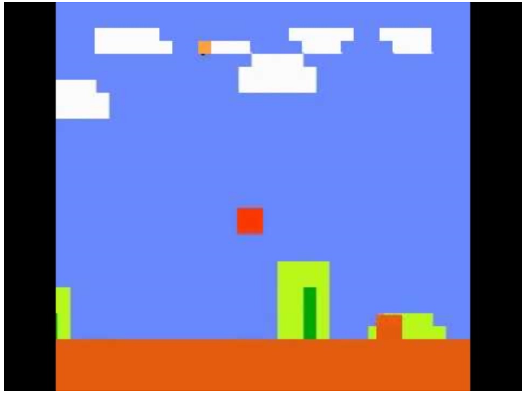


- All algorithms did not generalize well to other stages,
  - All algos have a pass counts of 0/1000 in unseen terrains
  - All algos have a drastic reduction of coins collected in unseen terrains
- Coins collected: A2C > PPO > DQN

AY2023/24 Sem2
CS4246/CS5446

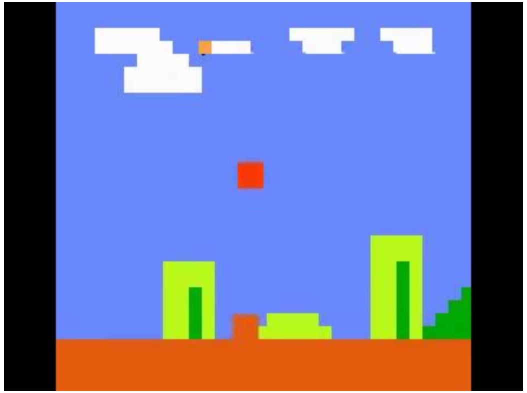
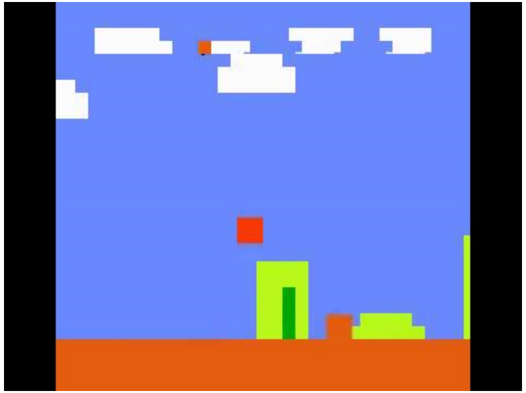
Conclusion

- Overall the behavior of the models match the theoretical concepts of the models:
  - Goal 1 Outcome (Compare between models):
    - Conservative: DQN > PPO > A2C
    - Most rounds completed: PPO > DQN > A2C
    - Coins collected: A2C > DQN > PPO
  - Goal 2 Outcome (CC mode vs. Regular mode):
    - CC reward altered agent’s behaviors, encouraging them to be more active in collecting coins and defeating enemies
    - CC reward maintained similar pass rate
  - Goal 3 Outcome (Generalizability):
    - Moderate generalizability in seen terrain and enemies
    - Did not generalize to unseen enemies

Thank you! Any Q&A?

**(Here are some bloopers)*

AY2023/24 Sem2
CS4246/CS5446