Deep RL Models with Raw Pixel Data

Evan Goldman (egold018) Nicko Martinez (nmart130)

Abstract

This project explores reinforcement learning (RL) from raw pixel data using deep learning architectures, specifically Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The goal is to enable agents to learn optimal policies directly from high-dimensional visual inputs, simulating real-world scenarios where sensory data is unprocessed. We implement and compare CNNs and ViTs to understand their respective capabilities in extracting meaningful representations from pixel-based environments, with a focus on Atari-like games. Key performance metrics include training stability, sample efficiency, and the convergence of rewards. Through this comparative study, we aim to identify the strengths and challenges of each architecture in handling spatial and temporal dependencies in visual RL tasks, providing insights into their potential applications in vision-dependent autonomous systems and robotic control.

Summary

Reinforcement learning with visual data has significant implications for fields like robotics, autonomous systems, and game AI, where agents need to act based on complex, high-dimensional sensory input. Traditionally, Convolutional Neural Networks have been employed in RL tasks for processing image data, but recent advancements in Vision Transformers suggest they could also excel in visual RL tasks by capturing long-range dependencies. This project aims to compare the effectiveness of CNNs and ViTs when fed raw pixel data, with the goal of evaluating their ability to learn robust policies directly from visual inputs.

We experiment with three different Atari games from the gym library: Pong, Breakout, and Space Invaders. We experiment with a CNN and ViT architecture as the neural network backbone for Double DQN and PPO. Through our experiments, we conclude that in their current state, Vision Transformers are not suitable as the neural network backbone of a reinforcement learning agent. We further hypothesize that transformers will not make good neural network backbones for any agents with states not in pixel values due to many of the reasons that ViTs fail.

Environment

Setup: The environments loaded from Atari often have meaningless inputs, such as the “FIRE” button in pong. We exclude these from the action space. We give a temporal context of 4 frames stacked on top of each other in the channel dimension.
Pong-v5: In Pong, the agent is a paddle that moves back and forth, trying to deflect a ball. The actions we restrict it to are no action, up, and down. The game ends when one player reaches 21 points.

Breakout-v5: In Breakout, the agent is a paddle that also tries to bounce a ball, but it tries to bounce it into bricks. When a brick is hit, the agent gets rewarded. The controls we give it are no action, up, down, and fire.
SpaceInvaders-v5: In SpaceInvaders, the agent is a ship that shoots at aliens before they get to the earth. The game ends when the ship runs out of health due to enemy fire or the aliens reach the bottom of the screen. Rewards are given when an alien is destroyed.

Method/Experiments

In all our experiments, the outcome of ViTs was the same in that it was unable to learn, so we will omit it from this part and discuss in the conclusion.

The Atari game environments are slightly more challenging to learn than the in class environments since the rewards are given well after the action leading to the reward was taken. For example in Breakout, the agent will line the paddle with the ball, then many frames later, the ball will collide with a brick.

We passed the frames into the model after performing augmentation on the frames to reshape, grayscale, and normalize. For our CNN with the best results, we use three convolutional layers, followed by a linear layer all separated by a ReLU. For PPO, this is followed by separate linear policy and value heads. Due to time constraints, we were unable to train our models to convergence and thoroughly explore hyperparameters, so given more time I'm sure our results could be improved substantially.

Due to the limited time, we were mainly doing work on PPO since the results were more stable and consistent, making it easier to analyze.

Pong: With PPO, we achieved a top score of 8. Since each pong game ran to 21 points, it had the longest training time, since a rollout was completed each episode. Towards the end of training, episodes were averaging around 6000 steps. We expect that CNNs have an inherent advantage over ViTs in Pong because there is little long range dependence. It achieves balanced performance on this game, indicating it understands the general concept of the game, but struggles to optimize strategy.

Breakout: Breakout had very long evaluation times towards the beginning due to having a large max steps value (27,000). If the network mapped the start state to anything that isn't the FIRE action, the game would never start, meaning that in an untrained model, there is a 75% chance it will sit around doing nothing until the max time steps are reached. In breakout, there is some long range dependence, so ViTs have an advantage in that sense.

SpaceInvaders: SpaceInvaders has both short and long range visual context. The agent needs to know where aliens are to line up the shot from far away, while also avoiding the shots, so it could also benefit from short range visual context bias. We thought if the ViTs could learn, it would be interesting to see the results of this one. We achieved not so great results on space

invaders, we believe because we weren't collecting enough samples and truncating rollouts early.

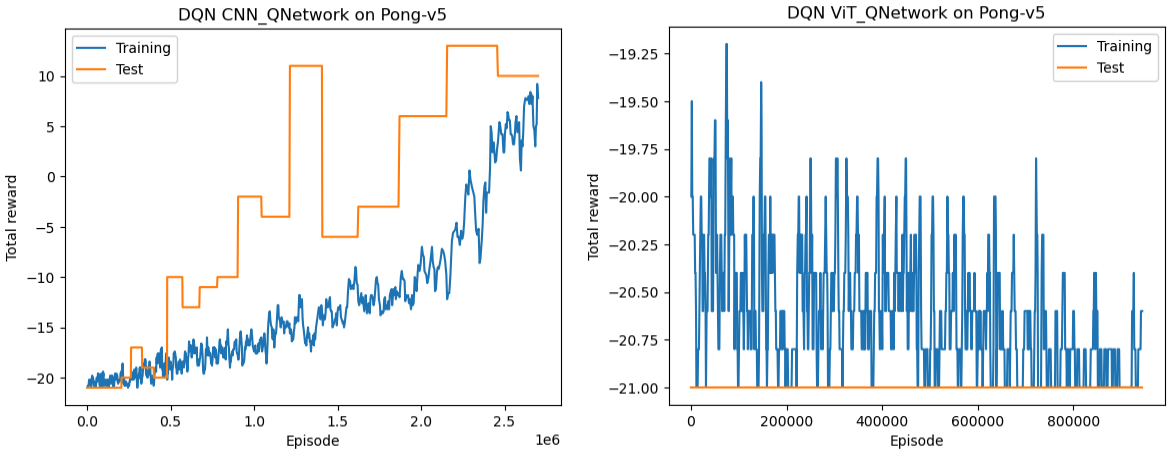

Figure: CNN and ViT with Double DQN trained for 1000 episodes. (Note that the x axis should be samples not episodes)

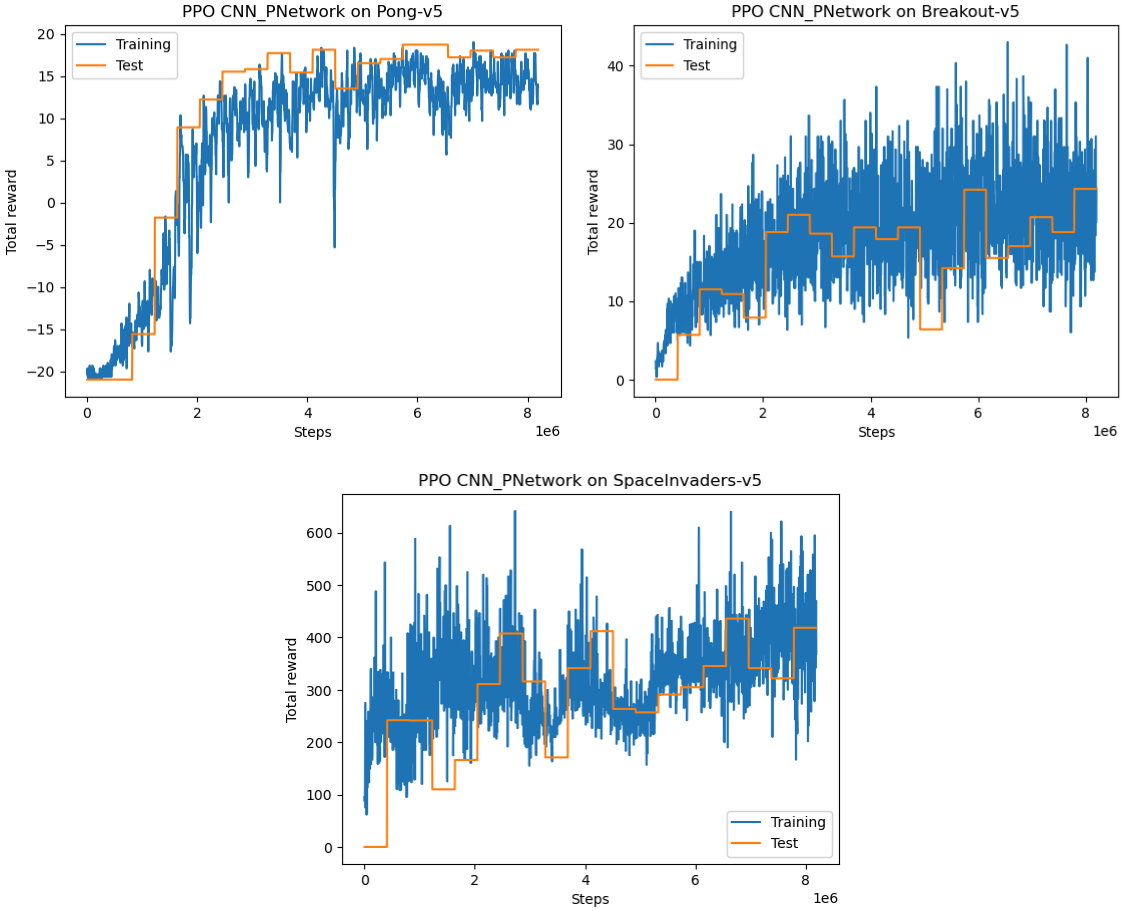

Figure: CNN with PPO trained for 8,192,000 episodes.

Conclusion

We experimentally found that ViTs are inferior to CNNs in these Atari games. Upon further research and observations, it is apparent why they fail so badly, and can extend our results to assume that transformers would also be very poor networks for agents for non-pixel observations. Given more time, we would verify this hypothesis.

Transformers share much of the same weaknesses that reinforcement learning does. Namely, sample/computational efficiency, and hyperparameter tuning. Many reinforcement learning algorithms aim to maximize sample or computational efficiency with DQN being more sample efficient and PPO being more computationally efficient. Transformers are both sample and computationally inefficient, causing them to exacerbate the intricacies of reinforcement learning.

CNNs also perform superior to ViTs due to their improved ability to capture temporal differences between the stacked frames. The complexity of ViTs quickly causes them to abstract away the stacked frames.

Future Steps

We weren't able to investigate this as thoroughly as we would have liked due to the unfortunate timing of the final exam. We have a lot of future work that we would have liked to show.

We would like to try our transformers with non pixel inputs since we suspect they will also perform poorly. We would also like to try pretrained CNN and ViT backbones. We believe that a pre-trained ViT might help with sample inefficiency.