A COMPARISON OF CONVOLUTIONAL NEURAL NETWORKS AND VISION TRANSFORMERS AS MODELS FOR LEARNING TO PLAY COMPUTER GAMES

By Adrien Dudon & Oisin Cawley

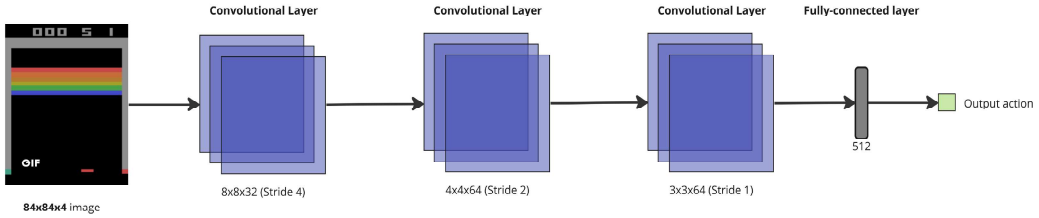
http://fyp.adriendudon.me/ | https://github.com/Deewens/FYP-DRL-Comparison

Introduction

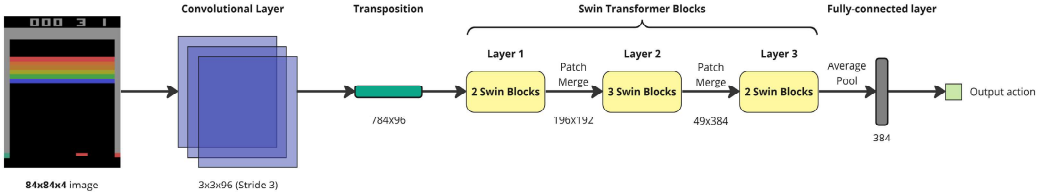
The Convolutional Neural Network (CNN) architecture, coupled with the Double Deep Q Network (DQN) (a Reinforcement Learning algorithm), has been extensively employed in solving complex video game environments. Nevertheless, the emergence of the Vision Transformer (ViT) architecture has demonstrated superior performance in various tasks previously dominated by CNNs. This research seeks to replicate the study conducted by Meng et al. and assess whether the Swin Transformer, a variant of ViT, can effectively learn to play video games using Reinforcement Learning and achieve comparable results within fewer training steps as compared to the same experiment with CNN.

Experimental Details

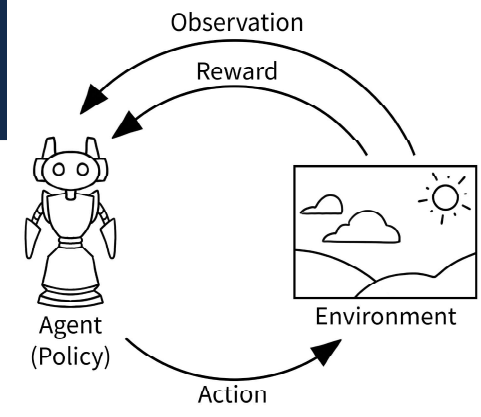
To replicate the approach of Meng et al. (2022), which replaces the traditional CNN with a Swin Transformer in Double DQN, identical algorithms and hyperparameter settings were employed.



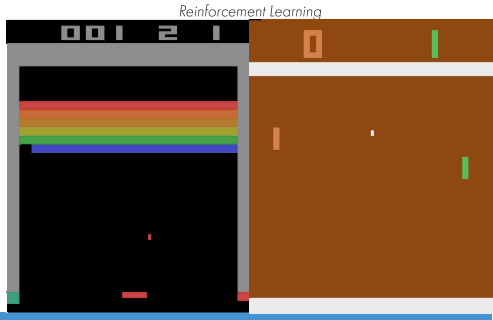
CNN Architecture



Swin Transformer Architecture

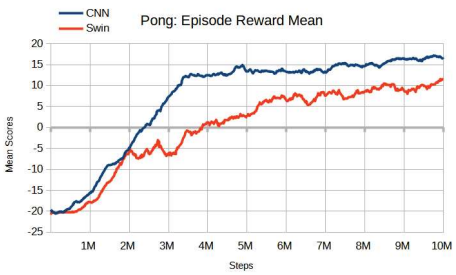


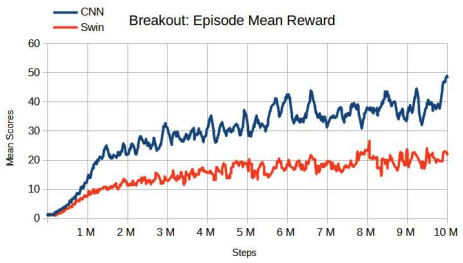
Reinforcement Learning

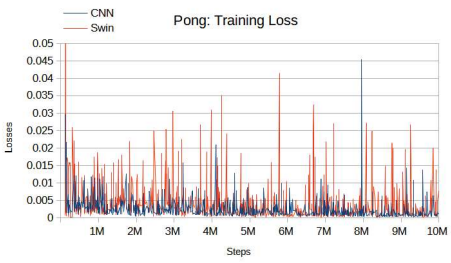


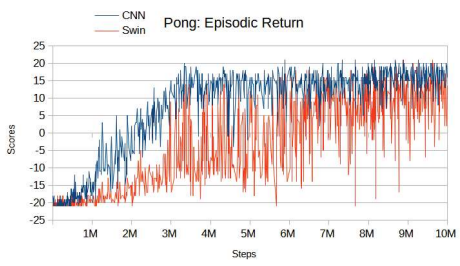
Results

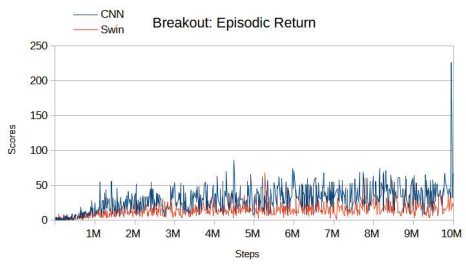
Results show that the agents trained with the CNN architecture outperformed the agents trained with Swin Transformer (for 10 million steps). However, in the Meng et al. study, Swin Transformed outperformed CNN, but only for a significant number of training steps (50 millions).

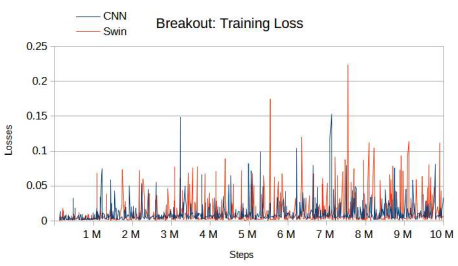












Conclusion

Training an agent with the Swin Transformer proved to be computationally intensive and demanded substantial GPU memory, making it a challenging task for an average computer and impractical for consumer video games, considering the limited ownership of such hardware. As a result, CNN remains a viable option for high-performance algorithms, considering the hardware constraints of end-user machines.