UR-78

Transforming Game Play: A Comparative Study of CNN and Transformer based Q-Networks in Reinforcement Learning

Abstract

In this study we investigate the performance of Deep Q-Networks utilizing Convolutional Neural Networks (CNNs) and Transformer architectures across 3 different Atari Games. The advent of DQNs have significantly advanced Reinforcement Learning, enabling agents to directly learn optimal policy from high dimensional sensory inputs from pixel or RAM data. While CNN based DQNs have been extensively studied and deployed in various domains Transformer based DQNs are relatively unexplored. Our research aims to fill this gap by benchmarking the performance of both DCQNs and DTQNs across the Atari games' Asteroids, Space Invaders and Centipede. We find that Transformer based Q-Networks

Introduction

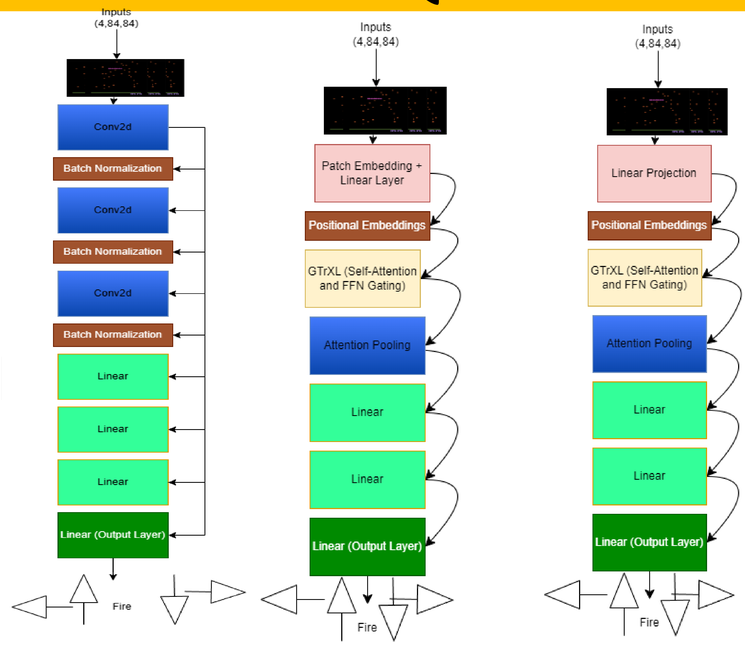
- Reinforcement Learning (RL) has undergone a revolution with Deep Q-Networks (DQNs), which enable agents to derive optimal strategies directly from complex, high-dimensional inputs like game pixels. Initially powered by Convolutional Neural Networks (CNNs) and achieving remarkable performance across a variety of Atari games, DQNs have now expanded to integrate Transformer architectures—still under-researched in RL.
- In our study, we leveraged the Arcade Learning Environment and OpenAI Gym to benchmarks to variants of the DQN; The common CNN-based DCQN and the Transformed-based DTQN; across three classic Atari games: Asteroids, Space Invaders and Centipede. Our aim is to explore the potential of Transformer architectures in RL without the aid of CNNs, RNNs or other recurrent structures like GRU.
- We used both a ViT and Linear Projection based Q-Networks

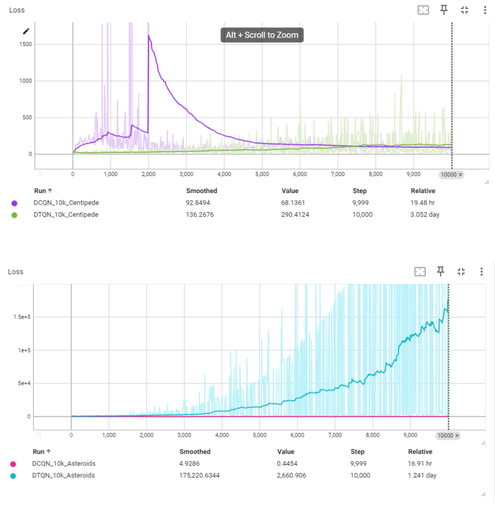
Research Question(s)

Why are Transformers rarely used as the architecture in Deep-Q-Networks? How do CNN and Transformer based DQNs match up in the 35-40 Million Parameter range?

Materials and Methods

- We streamlined the action space for each game using OpenAI Gym, which simplifies the decision-making process for RL agents and provides already defined rewards. Frame stacking was used to provide temporal context for the Transformer Architecture.
- Game frames were converted to grayscale and resized before feeding them into either model, input shape and frame skipping were kept constant to ensure parity even at the expense of computational load.
- The DCQN Architecture employs three convolutional layers with batch normalization, followed by a series of fully connected layers.
- The DTQN's structure is inspired by Vision Transformers, utilizing patch embeddings to process game frames as sequences capturing both spatial and temporal features. We also draw heavy inspiration from the GTrXL opting for gated self-attention
- We implemented a memory buffer to store experiences, with a minibatch strategy for training. An epsilon-greedy approach guided the agents' exploration vs exploitation, with epsilon decay to gradually decrease the amount of exploration the Agent performs.
- To optimize our agents' decision making, we employed Q-learning with separate policy and target networks. The policy network proposes actions, while the slower-updating target network stabilizes training by providing consistent future value estimates.





Results

Game	Model	Lowest Reward	Highest Reward	Average Reward
Centipede	DCQN	7714	13853	10152.4
Centipede	DTQN	3505	27908	13266.2
Asteroids	DCQN	1300	680	920
Asteroids	DTQN	140	140	140
Space Invaders	DCQN	285	285	285
Space Invaders	DTQN	245	535	411.7

Conclusions

We can determine that Transformers struggle in the 35-40 million parameter range because of the dimensionality of the input frames. With a strict restriction on using convolutions, it becomes difficult to reduce the features to the embedding dimension without significant loss of information.

Over 10000 Episodes it appears that the Transformer model only recognizes changes in the pixels on the linear plane directly above the position of the Agent. This can be evidenced by the tendency to shoot faster when there is a target overhead but not dodging, which's effects are exasperated by the reward function not applying explicit penalties for losing lives/taking damage.

We trained the DTQN Centipede Agent using the huber loss function and the DTQN Asteroids Agent using MSE, we predict that huber loss is not valid for Transformer-based Q-Networks without pretraining on object recognition.

The Transformer only outperforms the CNN in the game of Centipede over 10000, based on the reward curve and gameplay observation without significant policy change the CNN will converge on a similar strategy to the Transformer, as standing still proves to yield the highest points.

The ViT implementation is 5 times slower than the Traditional CNN implementation, and the Linear Projection model is twice as slow as the Traditional CNN.

The Transformer model is disadvantaged at the 35-40 million parameter range because patch embedding creates a dimension of 52000 to be reduced to the embedding size of the Gated Transformer. Linear Projection results in a dimension of 28224. We defined a Transformer based Q-Network with CNN and GRU gating and trained an additional Centipede Agent, outperforming both Agents across all metrics except 3 times slower than the CNN Implementation.

Contact Information

Scan the QR code in references to visit the website, to access the paper, videos, presentation and other details related to the project..

References

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279, jun 2013.
[2] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021.
[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
[4] Kevin Esslinger, Robert Platt, and Christopher Amato. Deep transformer q-networks for partially observable reinforcement learning, 2022.
[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS'12, pages 1097–1105, Red Hook, NY, USA, 2012. Curran Associates, Inc.
[6] Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. Journal of Artificial Intelligence Research, pages 523–562, 2018.
[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.
[8] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, and Shane Legg. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015
[9] Emilio Parisotto, H. Francis Song, Jack W. Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant M. Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, Matthew M. Botvinick, Nicolas Heess, and Raia Hadsell. Stabilizing transformers for reinforcement learning, 2019.
[10] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay, 2016.
[11] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning, 2015.
[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
[13] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: William A. Stigall
Advisors: Jitendra Sai Kota