Decision Transformer with ViT State Encoder

Project by Akash

Overview

- ➔ Decision Transformer is a decoder-only transformer architecture that aims to generate actions in an auto-regressive manner.
- ➔ Learns from both expert and non-expert data thanks to reward conditioning.
- ➔ Empirically proven to be significantly better than vanilla Behavior Cloning methods.

Goals

- ➔ Train baseline DT, ViT-DT and MLP BC models on the dataset.
- ➔ Tune CNN, ViT and MLP hyperparameters for optimal performance.
- ➔ Run for 3 distinct Atari games: Breakout, Seaquest and Qbert.
- ➔ Average the results over multiple runs with different seeds.

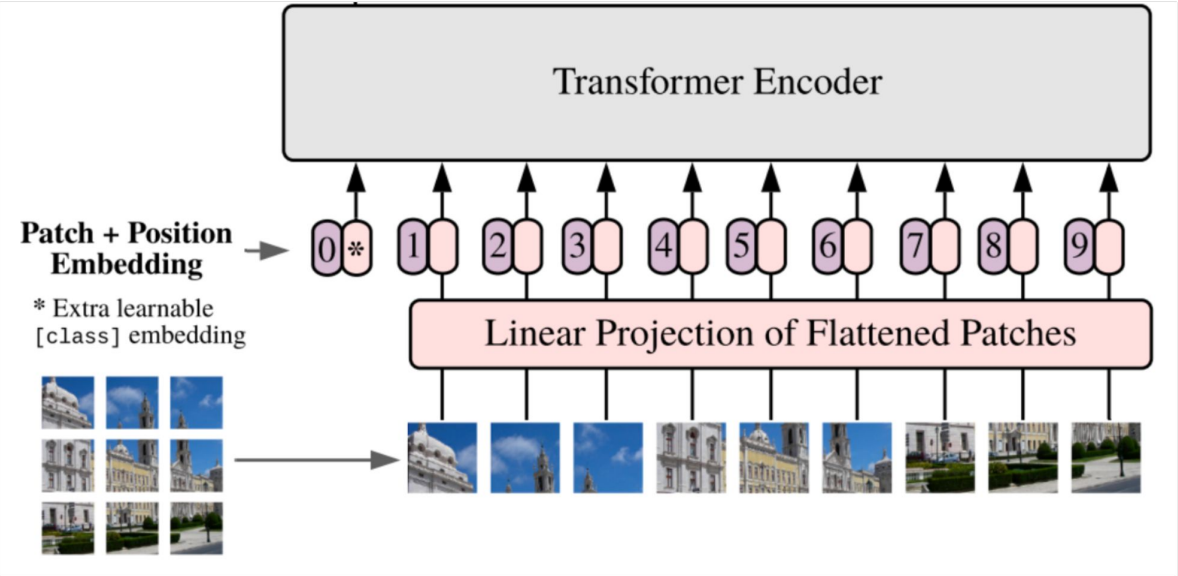
Dataset and Training

- ➔ DT is primarily trained in an offline learning environment.
- ➔ The agent is trained entirely on an existing dataset without any interaction with the environment.
- ➔ Dataset consists of 200M frames or 50M (state, actions and rewards) experience tuples.

DT Architecture

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. ArXiv. /abs/2106.01345

ViT State Representation

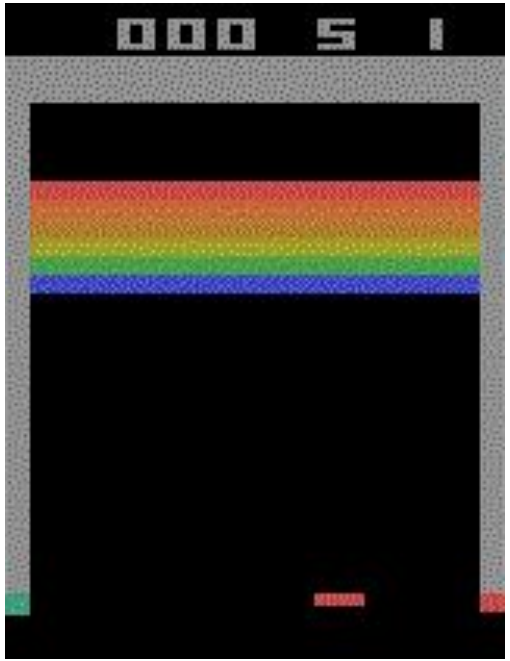


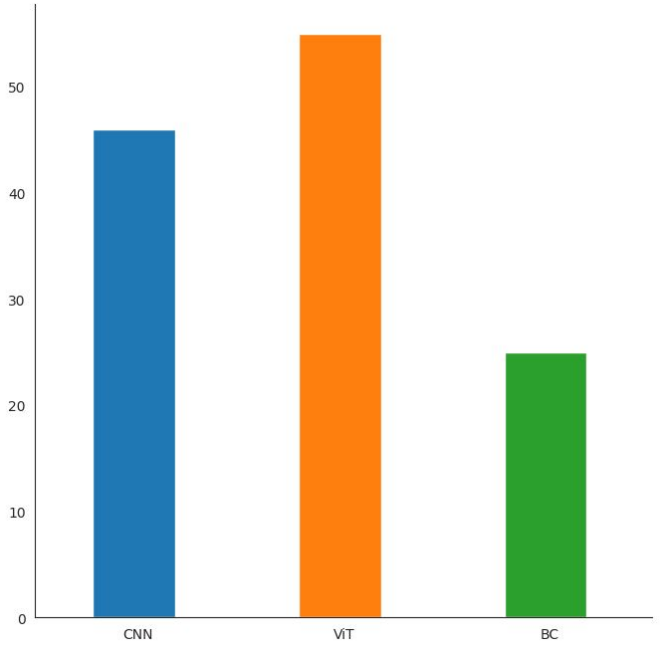
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv. /abs/2010.11929

DT Configuration

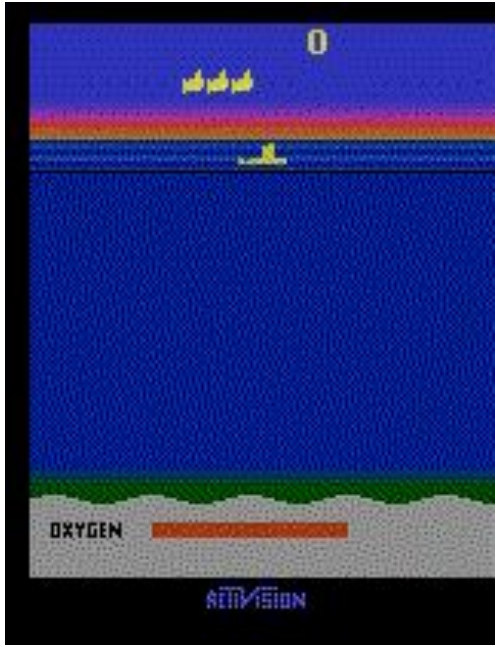
Hyperparameter	Value
Number of layers	6
Number of attention heads	8
Embedding dimension	128
Batch size	128 Breakout, Qbert, Seaquest
Context length K	30 Breakout, Qbert, Seaquest
Return-to-go conditioning	90 Breakout ($\approx 1\times$ max in dataset) 2500 Qbert ($\approx 5\times$ max in dataset) 1450 Seaquest ($\approx 5\times$ max in dataset)
Nonlinearity	ReLU, encoder GeLU, otherwise
Encoder channels	$32, 64, 64$
Encoder filter sizes	$8 \times 8, 4 \times 4, 3 \times 3$
Encoder strides	$4, 2, 1$
Max epochs	5
Dropout	0.1
Learning rate	$6 * 10^{-4}$
Adam betas	$(0.9, 0.95)$
Grad norm clip	1.0
Weight decay	0.1
Learning rate decay	Linear warmup and cosine decay (see code for details)
Warmup tokens	$512 * 20$
Final tokens	$2 * 500000 * K$

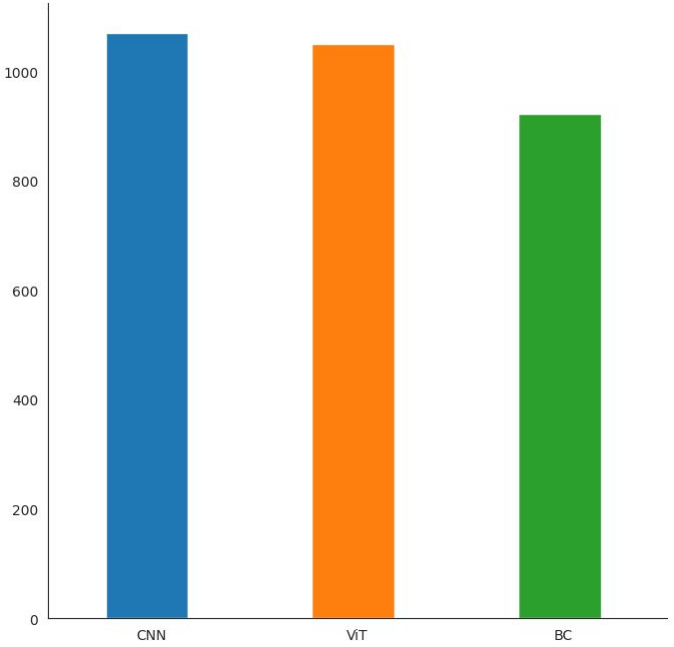
Breakout Results



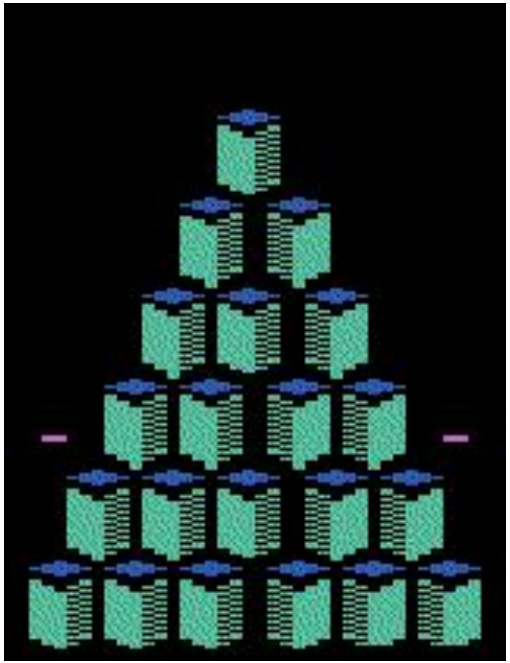


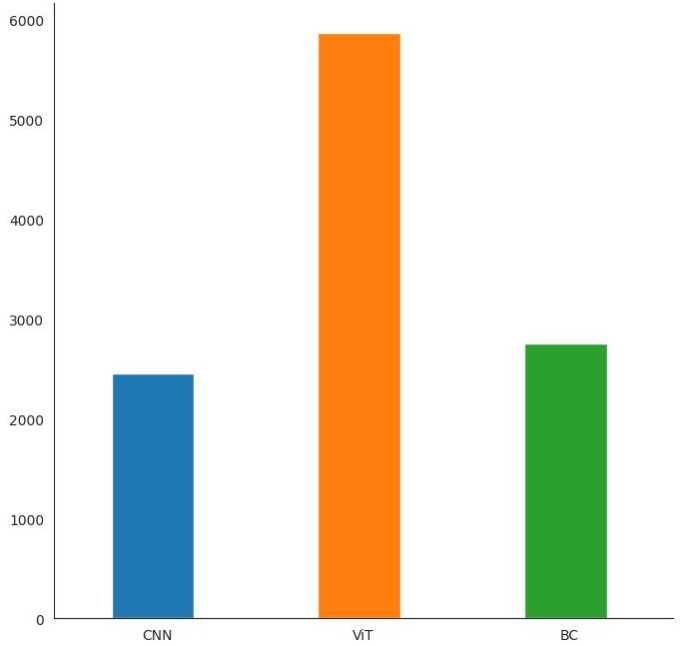
Seaquest Results



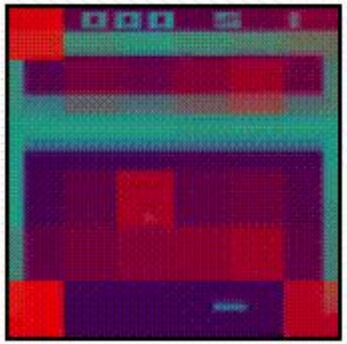


Qbert Results

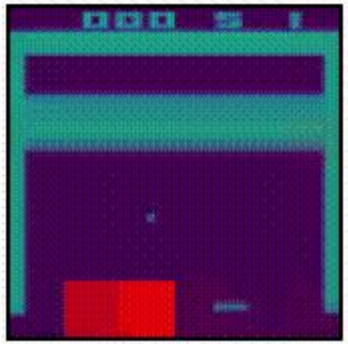




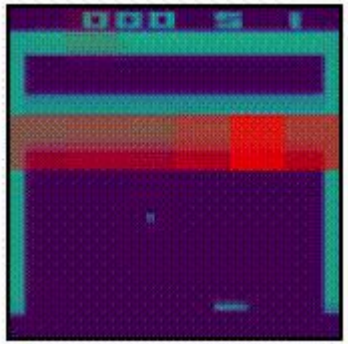
Breakout Attention Heads



Ball



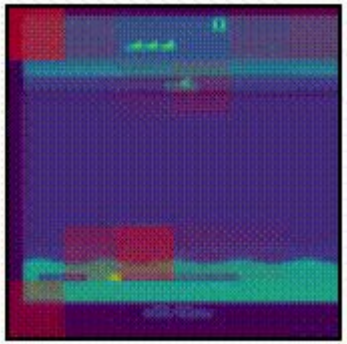
Free Move Space



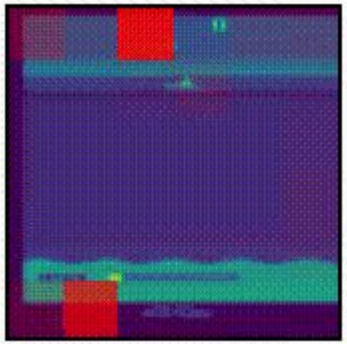
Unbroken Blocks

Lee, K., Nachum, O., Yang, M., Lee, L., Freeman, D., Xu, W., Guadarrama, S., Fischer, I., Jang, E., Michalewski, H., & Mordatch, I. (2022). Multi-Game Decision Transformers. *ArXiv*. /abs/2205.15241

Seaquest Attention Heads



Player



Bullets

Lee, K., Nachum, O., Yang, M., Lee, L., Freeman, D., Xu, W., Guadarrama, S., Fischer, I., Jang, E., Michalewski, H., & Mordatch, I. (2022). Multi-Game Decision Transformers. *ArXiv*. /abs/2205.15241

Further Developments

- ➔ Multi-game pre-training of DT shows human-level performance: https://arxiv.org/abs/2205.15241
- ➔ Online fine-tuning approach for DT that learns from replay rollouts (with hindsight experience replay): https://arxiv.org/abs/2202.05607
- ➔ Prompting DTs to allow better generalization capabilities: https://arxiv.org/abs/2206.13499

Questions?