

Deep Reinforcement Learning with Swin Transformers

LI MENG, Department of Technology Systems, University of Oslo, Norway

MORTEN GOODWIN, Centre for Artificial Intelligence Research, University of Agder, Norway

ANIS YAZIDI, Oslo Metropolitan University, Norway

PAAL ENGELSTAD, Department of Technology Systems, University of Oslo, Norway

Transformers are neural network models that utilize multiple layers of self-attention heads and have exhibited enormous potential in natural language processing tasks. Meanwhile, there have been efforts to adapt transformers to visual tasks of machine learning, including Vision Transformers and Swin Transformers. Although some researchers use Vision Transformers for reinforcement learning tasks, their experiments remain at a small scale due to the high computational cost. This article presents the first online reinforcement learning scheme that is based on Swin Transformers: Swin DQN. In contrast to existing research, our novel approach demonstrate the superior performance with experiments on 49 games in the Arcade Learning Environment. The results show that our approach achieves significantly higher maximal evaluation scores than the baseline method in 45 of all the 49 games (~ 92 %), and higher mean evaluation scores than the baseline method in 40 of all the 49 games (~ 82 %). The code can be found on <https://github.com/mengli11235/SWIN-DQN>.

CCS Concepts: • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Reinforcement Learning, Vision Transformers

ACM Reference Format:

Li Meng, Morten Goodwin, Anis Yazidi, and Paal Engelstad. 2024. Deep Reinforcement Learning with Swin Transformers. In *2024 8th International Conference on Digital Signal Processing (ICDSP) (ICDSP 2024), February 23–25, 2024, Hangzhou, China*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3653876.3653899>

1 INTRODUCTION

Reinforcement learning (RL) and natural language processing (NLP) are considered sub-fields of Artificial Intelligence (AI). They both deal with temporal sequences with the possibility of using neural networks (NNs). State-of-the-art solutions to RL rely on the Markov decision process (MDP) to build the temporal decision-making model, which describes a discrete-time stochastic control process that can be solved mathematically through dynamic programming, while transformers are a common solution to handle NLP tasks.

Prior to the introduction of transformers, the most common deep learning (DL) solution to handle NLP tasks was to utilize recurrent neural networks (RNNs). Transformers use a self-attention network and demonstrate more robust capabilities than previous attention-based RNNs [27]. One main improvement from the Transformer is to encode each hidden state with its attention-based context vector, which yields significantly richer contextual representations when stacked over multiple NN layers. The contextual information is also exempted from sequential strictness. Another main improvement is to use positional embeddings to encode relative positions of the input tokens.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2024 Copyright held by the owner/author(s).

Manuscript submitted to ACM

Recent research has demonstrated the strength of convolutional neural networks (CNNs) on real-world applications [11–17, 23, 30], and focused on the viability of replacing CNNs with transformers in other domains of machine learning (ML). Vision Transformers (ViT) successfully adapted transformers from the NLP domain into the computer vision (CV) field [5]. Swin Transformers further improved ViT in terms of both computational overhead and accuracy [18].

RL studies intelligent behaviors in environments that typically require an agent to maximize the obtained rewards. One of RL’s main challenges is to balance the rate of exploration over exploitation in the agent’s planning. The outcomes of past exploration need to be stored, and the RL agent then uses those outcomes to predict the expected rewards in future runs. Q-values are values that record such outcomes of actions at a certain state. Q-learning is an RL algorithm that keeps track of Q-values and uses them to update the policy according to the Bellman Optimality Equation (Eq. 1) [29]. In this equation, γ is the discounting factor, $Q^*(s, a)$ is the Q-value at state s of action a , r_{t+1} is the result value obtained by advancing to state s_{t+1} at time instant $t + 1$.

$$Q^*(s, a) = E\{r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a' | s_t = s, a_t = a)\} \quad (1)$$

Combining RL policies with NNs has also become a common practice in RL. In particular, CNNs have been heavily applied to games with image displays in Deep reinforcement learning tasks (DRL). Deep Q-learning Network (DQN) is a DRL method that stores Q-values in NNs instead and produces the Q-value predictions as the network output. DQN has achieved human, or superhuman-level performances on many Atari games [20]. There were researches using ViT in RL from image pixels [8, 9, 25]. However, training ViT in RL can be costly due to its quadratic complexity relative to the input image size.

This paper introduces the Swin DQN, an online RL scheme with Swin Transformers. This method extends the well adopted Double Q-learning [6] with recently introduced Swin Transformers. The heart of the method is splitting groups of image pixels into small tokenized patches and applying local self-attention operations inside the (shifted) windows of fixed sizes as an extension.

2 RELATED WORK

DQN with ViT Kalantari et al. [8] applied ViT to DQN for the purpose of improving sample complexity. To reduce the costs of ViT, they used the Linformer architecture [28] and discarded all the tokens except value tokens at the end of the encoder. Nonetheless, they could not obtain complete results at 200M frames on the Atari games they had selected due to limited resources.

Self-supervised Vision Transformers Self-supervised Vision Transformers [4] combine ViT with self-supervised methods such as Momentum Contrast (MoCo) [7]. MoCo builds dynamic dictionaries for contrastive learning in pre-training and the learned representations are to be transferred to downstream tasks. It was observed that instability of ViT is a major issue when used together with MoCo. More generally, there is evidence that self-supervised pre-training of ViT provides clearer semantic segmentation information of images than supervised learning [2].

Decision Transformers Decision Transformers view RL problems as the same sequence modeling problems as NLP does, which abandon the traditional value function and policy gradient methods in RL and replace them with causally masked transformers [3]. One of the major limitations of Decision Transformers is that it only works on offline RL. States, actions and returns must be known prior to the training of the model. Online Decision Transformers (ODT) tackle this issue by integrating offline pretraining and online finetuning in a unified framework [31] and achieve state-of-the-art performance on the D4RL benchmark.

Algorithm 1 Double Q-learning

```

1: Input:  $\epsilon, \gamma, maxFrames, syncFrames, L()$ 
2: Parameter:  $D, Q_\theta^A, Q_\theta^B$ 
3: Output:  $Q_\theta^A$ 
4:  $frames \leftarrow 0$ 
5: while  $frames < maxFrames$  do
6:   Initialize the environment
7:   while Game not finished do
8:      $frames += 4$ 
9:     if  $random() < \epsilon$  then
10:      Select an action randomly
11:     else
12:      Select an action through  $Q_\theta^A$ 
13:     end if
14:     Store  $s, a, s', r, terminal$  into replay buffer D
15:     Draw minibatch of  $s, a, s', r, terminal$  from D
16:     Update  $Q_\theta^A$  according to Algorithm 2
17:   end while
18:   if  $frames \% syncFrames == 0$  then
19:      $Q_\theta^B \leftarrow Q_\theta^A$ 
20:   end if
21: end while
22: return  $Q_\theta^A$ 

```

Algorithm 2 Network Update

```

1: Input:  $s, a, s', r, terminal, L()$ 
2: Parameter:  $Q_\theta^A, Q_\theta^B$ 
3: Output:
4:  $target = r + \gamma Q_\theta^B(s', a^*) * (1 - terminal)$ 
5: Compute  $loss$  by  $L(Q_\theta^A(s, a), target)$ 
6: Update  $Q_\theta^A$  by  $loss$  with gradient descent
7: return

```

3 METHOD

Double Q-learning [6] is a widely adopted algorithm that improves upon Q-learning. Q-learning tends to overestimate the Q-values because the policy updates on the target values selected by the max operator, which perpetuates the approximation errors in the direction of overestimation. Overestimation in Q-learning can have undesired effects and lead to suboptimal performances of an agent [19, 26]. Double Q-learning avoids overestimation by having two separate Q-value functions. Only one of them is updated at once according to either (Eq. 2) or (Eq. 3). $\alpha(s, a)$ is the learning rate, a^* is $\arg\max_a Q_\theta^A(s', a)$, and b^* is $\arg\max_a Q_\theta^B(s', a)$ at the next state s' . This update rule also indicates that Double Q-learning suffers from underestimation instead of overestimation because the target approximation is a weighted estimate of unbiased expected values, which is lower or equal to the maximum expected values [6].

$$Q^A(s, a) = Q^A(s, a) + \alpha(s, a)(r + \gamma Q^B(s', a^*) - Q^A(s, a)) \quad (2)$$

$$Q^B(s, a) = Q^B(s, a) + \alpha(s, a)(r + \gamma Q^A(s', b^*) - Q^B(s, a)) \quad (3)$$

Algorithm 1 shows the pseudo-code of our Double DQN. Here, γ is the discounting factor, ϵ is the exploration ratio, $maxFrames$ is the maximal number of total frames, and $L()$ is the loss function, i.e., Smooth L1 loss in this case.

Q_θ^A stands for the policy network and Q_θ^B the target network, which is a copy of Q_θ^A . $syncFrames$ is the number of frames between the synchronizations of two networks. s is the current state, s' the next state, a the action, r the reward, and $terminal$ is a flag indicating if the game terminates or not. In Algorithm 2, a^* is $\arg\max_a Q_\theta^A(s', a)$.

Attention is implemented in transformers as the contextual embeddings of the 'query' (Q), 'key' (K) and 'value' (V), as in (Eq. 4). B is the relative position bias that improves over absolute position embedding and d_k is the dimension

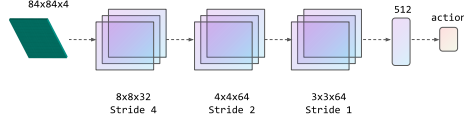


Fig. 1. The structure of our DQN. It consists of three convolutional layers, and a fully connected layer, followed by an output layer.

of the query and key. Transformers allow the re-combination of attention information from different layers, and the processing of all inputs at once, which are more convenient than RNNs when dealing with a large number of data.

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k} + B)V \quad (4)$$

We improve DQN by replacing CNNs with an adaptation of Swin Transformers, Swin MLP, which implements multi-head self-attention operations as grouped 1D convolutions. We name this new method as Swin DQN. Fig. 2 shows the architecture. Image pixels are grouped into small patches by one layer of 2D convolutions so that the output channels become hidden embeddings. Each basic layer contains a number of Swin blocks, illustrated on the bottom left in Fig. 2. Patches are grouped into local windows prior to the self-attention operation, which is done through a grouped 1D convolutional layer. This exploitation of locality reduces the computational complexity from quadratic to linear when the window size is fixed. Afterwards, there are two linear layers with the number of hidden units proportional to the embedding dimensions, followed by reshaping operations that reverse the window partition.

One shortcoming of utilizing locality is the lack of self-attentions among different local windows. To effectively overcome this, the window partition is shifted in successive blocks to introduce cross-window connections. Namely, the first and third blocks displace the windows by a number of patches, so that the windows are overlapped over different neighboring blocks. Neighboring patches are merged after all but the last basic layers in order to build hierarchical feature maps. Patch merging also reduces the dimensions of embeddings by half by adding an additional linear layer.

4 EXPERIMENTAL DETAILS

Our experiment is conducted among 49 Atari games of Arcade Learning Environment (ALE). Performances are evaluated based on both the highest evaluation scores and mean evaluation scores. Fig. 1 shows the NN architecture of our Double DQN. The list of chosen parameters for Double DQN is available in Table 1. We use Adam [10] as our optimizer and the learning rate remains the same for both Double DQN and Swin DQN.

Fig. 2 shows a full illustration of our Swin DQN. The list of parameters specified for Swin DQN is included in Table 2. The rest of the hyper-parameters are kept the same as in Double DQN. We use three layers of Swin blocks in our NN. Those layers contain 2, 3, 2 Swin blocks and 3, 3, 6 attention heads, respectively. The patch size is set to 3×3 , which yields 28×28 patches since the input size is 84×84 for each channel. The embedding dimension for each patch is 96. This suggests that the token size after patch embedding is 784×96 . The matrix operation of self-attentions is conducted through a grouped 1D convolutional layer, where the number of groups is equal to the number of attention heads. The drop path rate 0.1 states that there is a 10% chance that the input is kept as it is in skip connections.

There are random no-operation (no-op) steps of $[0, 30]$ at the start of each game to help introduce stochasticity into the environment. We set the seed of the Atari environment fixed for the purpose of training, but randomly select seeds for a more concrete evaluation phase. Randomly seeded games are initialized for every single episode of the evaluation.

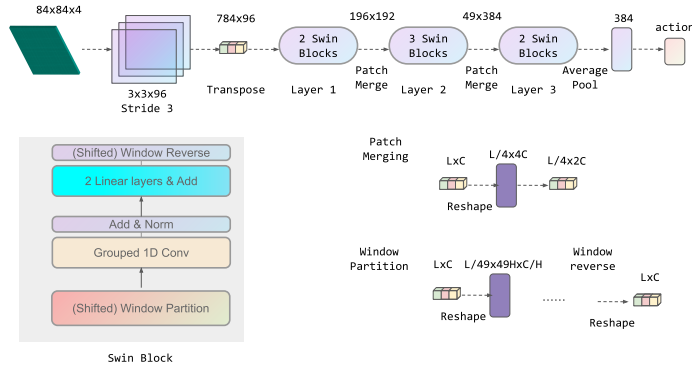


Fig. 2. The architecture of our Swin DQN. The top shows the step-by-step procedure. The bottom left box contains structures inside a Swin block. The details of patch merging, window partition and window merging are illustrated on the bottom right.

Table 1. Parameters used in Double DQN

Input	$84 \times 84 \times 4$
Optimizer	Adam
Adam learning rate	0.0000625
γ	0.99
Initial ϵ	1
Final ϵ	0.01
ϵ decay frames	1M
SyncFrames	40000
Frames per step	4
Steps per update	4
Steps per evaluation	250000
MaxFrames	200M
Replay size	1M
Batch size	32

Table 2. Parameters specified for Swin DQN

Layers	3
Blocks each layer	2, 3, 2
Heads each layer	3, 3, 6
Patch size	3×3
Window size	7×7
Embedding dimension	96
MLP ratio	4
Drop path rate	0.1

5 RESULTS

In this section, we consider the mean evaluation scores, together with the evaluation curves for a more detailed examination. Complete results over 49 games are included in Table 3 and Table 4. The mean evaluation scores and confidence intervals for 10 games are shown in Fig. 3. Each evaluation score is a mean of 5 runs from randomly seeded environments in a single training round.

Table 3 contains the maximal evaluation scores, whereas Table 4 contains the full mean evaluation scores, standard deviations, human normalized scores and area under curve (AUC) [24]. AUCs are the cumulative evaluation scores normalized by the final scores from [21] over the entire training period. Algorithms with the best performance for each game are highlighted. The key results are the comparisons between Swin DQN and Double DQN, since they have no algorithmic difference and the only change is that Swin DQN utilizes the Swin backbone instead of CNNs.

Table 3. Maximal Evaluation Scores

Game	Boot-DQN	Double DQN	(Normalized)	Swin DQN	(Normalized)	Nature
Alien	2436.6	5820	(0.81)	16670	(2.38)	3069
Amidar	1272.5	1421	(0.83)	2665	(1.55)	739.5
Assault	8047.1	4130	(7.52)	8582	(16.09)	3359
Asterix	19713.2	42900	(5.15)	981500	(118.32)	6012
Asteroids	1032	2100	(0.03)	11360	(0.23)	1629
Atlantis	994500	1012300	(61.78)	1006000	(61.39)	85641
Bank Heist	1208	1260	(1.69)	1230	(1.65)	429.7
Battle Zone	38666.7	57000	(1.57)	149000	(4.21)	26300
Beam Rider	23429.8	18150	(1.07)	60936	(3.66)	6846
Bowling	60.2	30	(0.05)	125	(0.74)	42.4
Boxing	93.2	100	(8.33)	100	(8.33)	71.8
Breakout	855	431	(14.91)	864	(29.94)	401.2
Centipede	4553.5	14297	(1.23)	23448	(2.15)	8309
Chopper Command	4100	3400	(0.39)	7200	(0.97)	6687
Crazy Climber	137925.9	218700	(8.3)	246900	(9.43)	114103
Demon Attack	82610	28225	(15.43)	134465	(73.84)	9711
Double Dunk	3	2	(9.36)	2	(9.36)	-18.1
Enduro	1591	2265	(2.63)	2361	(2.74)	301.8
Fishing Derby	26	29	(2.28)	43	(2.54)	-0.8
Freeway	33.9	26	(0.88)	34	(1.15)	30.3
Frostbite	2181.4	3770	(0.87)	6140	(1.42)	328.3
Gopher	17438.4	83360	(38.56)	108640	(50.3)	8520
Gravitar	286.1	1100	(0.29)	950	(0.24)	306.7
Hero	21021.3	13885	(0.43)	31475	(1.02)	19950
Ice Hockey	-1.3	3	(1.17)	5	(1.34)	-1.6
Jamesbond	1663.5	5700	(20.71)	9850	(35.87)	576.7
Kangaroo	14862.5	14800	(4.94)	14500	(4.84)	6740
Krull	8627.9	10647	(8.48)	19962	(17.2)	3805
Kung Fu Master	36733.3	62200	(2.76)	117700	(5.22)	23270
Montezuma Revenge	100	0	(0)	2600	(0.55)	0
Ms Pacman	2983.3	7250	(1.04)	13911	(2.05)	2311
Name This Game	11501.1	14470	(2.12)	16820	(2.52)	7257
Pong	20.9	21	(1.18)	21	(1.18)	18.9
Private Eye	1812.5	200	(0)	14000	(0.2)	1788
Qbert	15092.7	18475	(1.38)	32425	(2.43)	10596
Riverraid	12845	18320	(1.08)	27790	(1.68)	8316
Road Runner	51500	67500	(8.62)	69500	(8.87)	18257
Robotank	66.6	62	(6.16)	81	(8.12)	51.6
Seaquest	9083.1	30380	(0.72)	32840	(0.78)	5286
Space Invaders	2893	2250	(1.38)	3150	(1.97)	1976
Star Gunner	55725	54000	(5.56)	83400	(8.63)	57997
Tennis	0	24	(3.08)	24	(3.08)	-2.5
Time Pilot	9079.4	12300	(5.26)	132100	(77.37)	5947
Tutankham	214.8	394	(2.45)	396	(2.46)	186.7
Up N Down	26231	38600	(3.41)	174820	(15.62)	8456
Venture	212.5	1700	(1.43)	2000	(1.68)	380
Video Pinball	811610	970906	(54.95)	986525	(55.84)	42684
Wizard of Wor	6804.7	25100	(5.85)	59900	(14.15)	3393
Zaxxon	11491.7	20300	(2.22)	28300	(3.09)	4977

In Table 3, we compare Swin DQN, Double DQN, Bootstrapped DQN (Boot-DQN) [22] and the Nature DQN [21] because they share the most similarities in experimental design and hyper-parameter choices. It is clear that Swin DQN performs the best out of the four algorithms in terms of maximal evaluation scores. The highest evaluation scores of Swin DQN are higher or equal than those of Double DQN in 45 out of 49 games (91.84%), better than the mean evaluation scores. Its human normalized scores also suggest that our algorithm is capable of human or even superhuman level performances in most of the games. The mean evaluation scores of Swin DQN are higher or equal than those of Double DQN in 40 out of 49 games (81.63%). AUCs also entail that the gap between their score curves remain at a significant level for most of the training process for each game.

In many games, the improvements are as significant as multiple times the evaluation scores of Double DQN, but for few games, the improvement is less prominent. This discrepancy may stem from the variety of Atari games. Different games have various unique settings and require distinguished sets of skills. Entity/pixel attention connections are

Table 4. Mean Evaluation Scores, Standard Deviations, Human Normalized Scores and AUCs

Game	Double DQN	Std.	Normalized	AUC	Swin DQN	Std.	Normalized	AUC
Alien	1664.0	739.96	0.21	0.52	5324.0	3868.63	0.74	1.08
Amidar	604.0	63.5	0.35	0.68	513.6	247.56	0.3	1.2
Assault	1712.2	345.37	2.87	0.47	3798.8	2432.75	6.88	0.53
Asterix	25220.0	11338.15	3.02	1.68	596300.0	253477.93	71.88	27.84
Asteroids	1102.0	303.87	0.01	0.49	5046.0	2509.91	0.09	1.14
Atlantis	753880.0	1327.25	45.8	7.08	835940.0	20888.43	50.88	7.51
Bank Heist	696.0	63.75	0.92	1.71	552.0	100.88	0.73	1.62
Battle Zone	27000.0	7099.3	0.71	0.87	61600.0	13749.18	1.7	1.48
Beam Rider	7952.4	2875.24	0.46	0.86	24140.0	10361.26	1.44	2.2
Bowling	30.0	0.0	0.05	0.4	19.8	18.27	-0.02	1.33
Boxing	91.2	1.6	7.59	1.23	99.2	0.75	8.26	1.27
Breakout	184.2	159.12	6.34	0.8	857.8	6.97	29.73	1.1
Centipede	3933.0	1059.55	0.19	0.38	7071.0	1382.07	0.5	0.73
Chopper Command	620.0	146.97	-0.03	0.17	2120.0	584.47	0.2	0.22
Crazy Climber	103320.0	15244.2	3.69	0.94	118160.0	5443.01	4.29	1.12
Demon Attack	5533.0	3037.95	2.96	0.39	89254.0	49893.56	48.99	1.69
Double Dunk	0.4	0.8	8.64	-0.08	0.4	1.5	8.64	-0.12
Enduro	1038.6	189.68	1.21	3.06	1901.4	461.78	2.21	4.67
Fishing Derby	-43.4	20.41	0.91	-51.59	17.4	8.62	2.06	-34.48
Freeway	21.6	1.2	0.73	0.69	34.0	0.0	1.15	1.09
Frostbite	1002.0	929.83	0.22	3.77	2036.0	1024.04	0.46	2.26
Gopher	21496.0	12332.76	9.86	1.89	12216.0	3511.93	5.55	1.59
Gravitar	210.0	139.28	0.01	0.99	70.0	140.0	-0.03	0.43
Hero	12716.0	192.0	0.39	0.38	27819.0	176.42	0.9	0.85
Ice Hockey	-8.6	2.58	0.21	-6.1	-0.6	0.8	0.88	-2.27
Jamesbond	680.0	97.98	2.38	0.99	1980.0	697.57	7.13	2.55
Kangaroo	8660.0	1539.61	2.89	0.98	8280.0	1729.05	2.76	1.36
Krull	8385.0	771.32	6.36	1.76	6861.6	1060.26	4.93	1.77
Kung Fu Master	21300.0	4285.79	0.94	1.15	70640.0	15710.46	3.13	1.9
Montezuma Revenge	0.0	0.0	0.0	-	100.0	0.0	0.02	-
Ms Pacman	2408.0	811.75	0.32	1.08	5878.0	2426.4	0.84	2.13
Name This Game	9140.0	1193.05	1.19	1.22	9670.0	2263.6	1.28	1.28
Pong	21.0	0.0	1.18	1.01	21.0	0.0	1.18	1.05
Private Eye	0.0	0.0	-0.0	0.03	100.0	0.0	0.0	0.03
Qbert	15195.0	185.34	1.13	1.15	25365.0	1660.17	1.9	1.61
Riverraid	8106.0	224.64	0.43	1.11	20180.0	5523.52	1.19	1.37
Road Runner	46560.0	6021.49	5.94	2.72	56900.0	6829.35	7.26	2.97
Robotank	40.8	10.26	3.98	0.67	62.6	7.55	6.23	1.03
Seaquest	7852.0	1442.97	0.19	1.05	7480.0	1617.65	0.18	1.01
Space Invaders	1041.0	379.5	0.59	0.43	2113.0	879.69	1.29	0.54
Star Gunner	44980.0	8585.66	4.62	0.36	58480.0	11361.94	6.03	0.65
Tennis	20.6	1.02	2.86	1.66	23.8	0.4	3.07	5.77
Time Pilot	9100.0	3919.69	3.33	0.72	51920.0	20264.79	29.11	4.2
Tutankham	216.8	64.39	1.31	1.04	251.6	42.32	1.54	1.21
Up N Down	7876.0	4448.75	0.66	1.11	34650.0	8173.73	3.06	1.75
Venture	940.0	120.0	0.79	1.83	500.0	154.92	0.42	1.43
Video Pinball	207074.6	201917.13	11.72	2.62	256106.4	179668.88	14.5	2.05
Wizard Of Wor	1900.0	252.98	0.32	1.1	14700.0	6867.02	3.37	3.38
Zaxxon	7020.0	406.94	0.76	1.48	13820.0	936.8	1.51	2.1

crucial in order to solve puzzles in some games, but they might contribute trivially to other games. Swin Transformers appear to have a forceful impact in games that have a high complexity of features or require subtle world modeling.

Swin DQN also shows drastically improved scores over Double DQN throughout the training process except for the beginning in Fig. 3. For instance, we can see that there is a significant difference of scores for those two models from 1.5M to 3M steps in Tennis from Fig. 3. Swin DQN converges earlier at around 2M steps and Double DQN at around 4M steps. Double DQN being stuck in suboptimal performances longer indicates a more robust modeling capability of Swin Transformers on adversity. In most of the games, Swin DQN does not show faster converging speed than Double DQN in the initial training frames. This observation corresponds to the higher complexity of the Swin model. The Swin transformers designed in our experiment have a larger network structure and, in turn, requires more input data than CNNs. Thus, the strength of Swin DQN is demonstrated only after a proper amount of training updates.

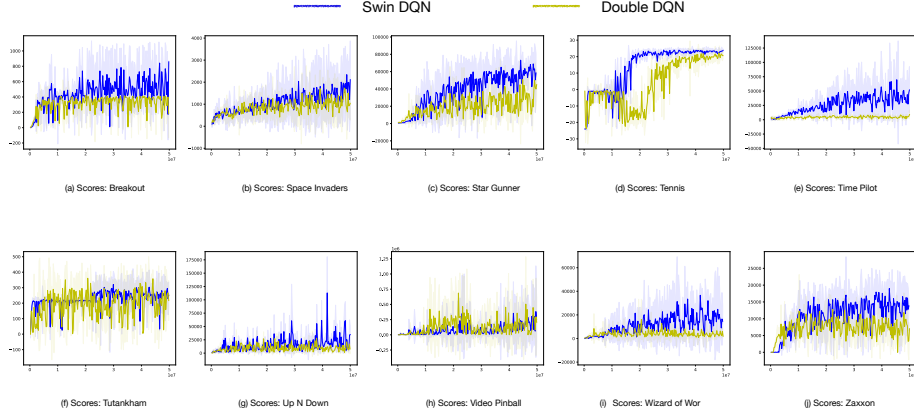


Fig. 3. The mean evaluation scores together with 95% confidence intervals during 200M frames (50M training steps). The blue line is Swin DQN. The orange line is Double DQN. The interval between every evaluation is 1M frames (250000 steps).

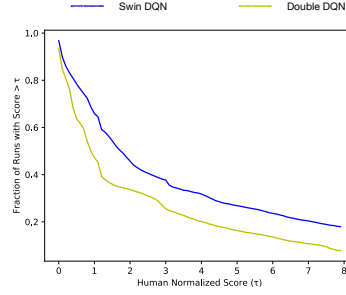


Fig. 4. The performance profiles up to $\tau = 8$ that shows the percentage of the score larger than τ . The blue line is Swin DQN and the orange line is Double DQN.

We also include performance profiles (score distributions) described in [1] with overall human normalized scores in Fig. 4, which allow an aggregated qualitative comparison between Swin DQN and Double DQN. The gap between two lines is non-trivial as τ increases, entailing that the improvement of human normalized scores among 49 games is salient when replacing CNNs with Swin Transformers.

In a few of the games (e.g., Video Pinball, Tutankham), Swin DQN shows tiny improvements compared to Double DQN during the training process, but still reaches slightly higher scores than Double DQN at the end. It is typical that Double DQN performs reasonably well in those games. Thus, the advantages of better feature representations through introducing Swin Transformers are relatively obscure. To further examine this observation, activation maps of images from four different games fed into trained models are illustrated in Fig. 5. Among the four games, Breakout and Time Pilot are representative games in which Swin DQN shows significant improvements, while Video Pinball and Tutankham are games in which the improvements are relatively small.

As expected, Swin DQN possesses significantly richer representations than Double DQN especially in layer-2 and layer-3 activation maps from Fig. 5. Moreover, Swin DQN have great advantage when representing entity/pixel relationships, while Double DQN is only capable of reasonably capturing part of the scoreboard or stationary objects.

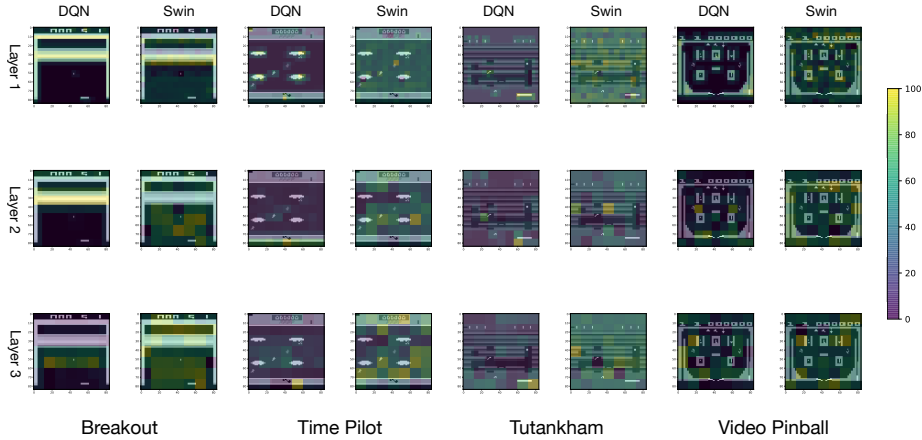


Fig. 5. The activation maps of our Double DQN and Swin DQN. All models are trained for 200M frames. The image background is in gray scale. As the color bar suggests, the more yellow the pixel gets, the higher the activation.

In Breakout and Time Pilot, Swin DQN successfully captures some possible moving trajectories of the target objects, but Double DQN is not excelled at this, or constantly activates wrong trajectories.

6 DISCUSSION

We have successfully demonstrated that Swin Transformers can improve the performances of DQN significantly in Atari games. Spatial self-attentions benefit not only CV tasks but also image-based RL tasks, and introducing spatial self-attentions alone is sufficient to improve the performances of current DRL models drastically. The structure of Swin Transformers also makes local attentions more affordable in DRL.

Our hyper-parameters of the Swin Transformer backbone are chosen by grid search, yet its structure might not be with the least necessary training parameters. Swin DQN is still computationally more costly (3 to 4 times in wall clock time) than DQN. It is appealing to seek for the smallest Swin model without jeopardizing its performances. Thus, experimenting Swin DQN with less training parameters can be crucial future work.

The positive effects of using self-attentions in Atari games lead us to the expectation that real-world control tasks can also exploit self-attentions. We humans often need to direct our attention to a small area in our visual field when solving physical control tasks, and this leaves room for transformers. DRL can be potentially more applicable in solving physical problems with the help of Swin Transformers in the future.

ACKNOWLEDGMENTS

This work was performed on the [ML node] resource, owned by the University of Oslo, and operated by the Department for Research Computing at USIT, the University of Oslo IT-department.

REFERENCES

- [1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems* 34 (2021).
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9650–9660.

- [3] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021).
- [4] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9640–9649. 10.1109/ICCV48922.2021.00950.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).
- [6] Hado Hasselt. 2010. Double Q-learning. In *Advances in neural information processing systems* 23. Curran Associates, Inc., Red Hook, NY, 2613–2621.
- [7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738. 10.1109/cvpr42600.2020.00975.
- [8] Amir Ardalan Kalantari, Mohammad Amini, Sarath Chandar, and Doina Precup. 2022. Improving Sample Efficiency of Value Based Models Using Attention and Vision Transformers. *arXiv preprint arXiv:2202.00710* (2022). 10.48550/arXiv.2202.00710.
- [9] Eshagh Kargar and Ville Kyrki. 2021. Vision transformer for learning driving policies in complex multi-agent environments. *arXiv preprint* (2021).
- [10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 10.48550/arXiv.1412.6980.
- [11] Changyang Li, Wanwan Li, Haikun Huang, and Lap-Fai Yu. 2022. Interactive augmented reality storytelling guided by scene semantics. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.
- [12] Wanwan Li. 2021. Image Synthesis and Editing with Generative Adversarial Networks (GANs): A Review. In *2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4)*. IEEE, 65–70.
- [13] Wanwan Li. 2022. Musical Instrument Performance in Augmented Virtuality. In *Proceedings of the 6th International Conference on Digital Signal Processing*. 91–97.
- [14] Wanwan Li. 2023. Synthesizing 3D VR Sketch Using Generative Adversarial Neural Network. In *Proceedings of the 2023 7th International Conference on Big Data and Internet of Things*. 122–128.
- [15] Wanwan Li. 2023. Terrain synthesis for treadmill exergaming in virtual reality. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 263–269.
- [16] Wanwan Li, Changyang Li, Minyoung Kim, Haikun Huang, and Lap-Fai Yu. 2023. Location-Aware Adaptation of Augmented Reality Narratives. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [17] Wanwan Li, Biao Xie, Yongqi Zhang, Walter Meiss, Haikun Huang, and Lap-Fai Yu. 2020. Exertion-aware path generation. *ACM Trans. Graph.* 39, 4 (2020), 115.
- [18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.
- [19] Li Meng, Anis Yazidi, Morten Goodwin, and Paal Engelstad. 2021. Expert Q-learning: Deep Q-learning With State Values From Expert Examples. *CoRR* (2021). 10.7557/18.6237.
- [20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *arXiv:1312.5602 [cs.LG]* 10.48550/arXiv.1312.5602.
- [21] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.
- [22] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems* 29 (2016), 4026–4034. <https://doi.org/10.5555/3157382.3157548>
- [23] B. Hui S. Gao and Li W. 2024. Image Generation of Egyptian Hieroglyphs. In *ICMLC 2024*.
- [24] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814* (2015). 10.48550/arXiv.1507.00814.
- [25] Tianxin Tao, Daniele Reda, and Michiel van de Panne. 2022. Evaluating Vision Transformer Methods for Deep Reinforcement Learning from Pixels. *arXiv preprint arXiv:2204.04905* (2022). 10.48550/arXiv.2204.04905.
- [26] Sebastian Thrun and Anton Schwartz. 1993. Issues in using function approximation for reinforcement learning. In *Proceedings of the 4th Connectionist Models Summer School*. Hillsdale, NJ, 255–263.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017). 10.5555/3295222.3295349.
- [28] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and H Linformer Ma. 2020. Self-attention with linear complexity. *arXiv preprint* 8 (2020).
- [29] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3–4 (1992), 279–292. 10.1007/BF00992698.
- [30] Wang Y., Liang W., Li W., Li D., and Yu L.F. 2020. Scene-aware background music synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1162–1170.
- [31] Qingqing Zheng, Amy Zhang, and Aditya Grover. 2022. Online decision transformer. In *International Conference on Machine Learning*. PMLR, 27042–27059.

Received 5 January 2024; revised 22 January 2024; accepted 22 January 2024