Competitive Reinforcement Learning in Atari Games

Mark McKenzie[1](✉), Peter Loxley[2], William Billingsley[2], and Sebastien Wong[1]

[1] Defence Science and Technology Group, Adelaide, SA, Australia
mark.colin.mckenzie@gmail.com
[2] University of New England, Armidale, NSW, Australia

Abstract. This research describes a study into the ability of a state of the art reinforcement learning algorithm to learn to perform multiple tasks. We demonstrate that the limitation of learning to performing two tasks can be mitigated with a competitive training method. We show that this approach results in improved generalization of the system when performing unforeseen tasks. The learning agent assessed is an altered version of the *DeepMind* deep Q–learner network (DQN), which has been demonstrated to outperform human players for a number of Atari 2600 games. The key findings of this paper is that there were significant degradations in performance when learning more than one game, and how this varies depends on both similarity and the comparative complexity of the two games.

Keywords: Reinforcement learning · DQN · Atari

1 Introduction

Reinforcement Learning is a distinct area within the broader fields of machine learning and artificial intelligence, utilising learning through action principles against its understanding of the environment, in order to choose the best course of action given some interpreted state, to achieve maximum long term reward [11]. Iteratively the reinforcement learning approach investigates its available actions and after some period of searching, refines a course of actions to achieve its objective. Recent progress in the application of reinforcement learning with deep networks has led to the *DeepMind* deep Q–learner network (DQN) [8], which can autonomously learn to play Atari video games [1,7] at or above the level of expert humans [10]. This system was able to consistently and contemptibly outperform a skilled human player across a variety of games with different game objectives, and was shown to learn the same high level strategies adopted by expert human players. As a result the DQN is considered to be the state of the art approach for reinforcement learning.

While DQNs were demonstrated to outperform human players at Atari games, one limitation is that the DQN agents received specialist training on

W. Peng et al. (Eds.): AI 2017, LNAI 10400, pp. 14–26, 2017.
DOI: 10.1007/978-3-319-63004-5_2

a single game. Whereas human agents are able learn and play multiple games. The problem of interest in our research is the ability of a DQN agent to perform multiple tasks, which we will assess through its ability to learn multiple Atari games.

Previous research into reinforcement learning agents performing multiple tasks has created the technique known as *distilling*, which was first presented by [2] and later adapted by [3] for neural networks. Using this approach DQN agents are first trained on individual games and then fused together to form one agent which applies to both games [9].

Recent work has demonstrated that when learning two tasks sequentially a DQN will suffer from catastrophic forgetting [4]. Where the process of learning a second task will destroy the network weights associated with the first task. This issue with sequential learning can be mitigated using elastic weight consolidation, where weights that are important for performing the first task are protected from large alterations when learning the second task [4].

The training method we propose is to simultaneously learn two games, effectively 'competing' for representational space within the neural network. This differs from the original *DeepMind* DQN [8] where the entire representational space of the neural network was *specialized* for a single game.

An interesting research question is how the representational space of the neural network will be allocated in a competitive learning environment; will the representational space be shared or segregated for the two tasks? An example of this being how well can the DQN agent perform at both Breakout, and Space Invaders; where one game is required to avoid falling objects, and another is required to hit and return falling objects. Our hypothesis is that the DQN will suffer significantly when learning competing tasks. The full extent of this detriment to performance being a function of the differences in the tasks at hand. A corollary supposition being that the architecture will increasingly segregate expressive power across the network for the alternate tasks as a function of the perceived differences in action for reward. This supposition is corroborated by success of the *distilling* approach [9] where a fuzed network is formed from two individual networks.

Another hypothesis is that a DQN agent that is trained in a competitive environment will generalise to new unforeseen tasks better than a specialist DQN agent that is trained on a single task.

Our contribution is an evaluation of one DQN agent acting on two environments, and this differs from [9] where two agents learning on separate environments are combined, and [4] where one agent acts and learns on a sequence of environments.

2 Methodology and Experimentation

The data used for training and testing was presented in [7,8] in which Atari 2600 games are used for DQN performance evaluation. For the purposes of this research pairs of games were selected according to different concepts of which

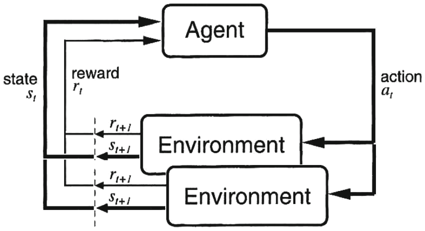

Fig. 1. The competitive reinforcement learning paradigm of interpreting the optimal action to receive reward for the current state, adapted from [11]

two are shown here; a pair of similar games i.e. functionally similar objectives and reward schemes, and a pair of high performing games i.e. those which the original code base easily converged on a suitable reward strategy. A common learning agent was then expected to learn both environment pairs competitively as depicted in Fig. 1. A baseline of performance was completed to ensure that the changes made to the code to assess multiple games did not alter its ability to converge to a similar performance as per the original research.

2.1 Alterations to the Deep Q-Learner Code

Due to the fundamental requirement to maintain similarity in learning approach with the original code base, a large portion of the code was necessarily kept identical to that provided [5]. However, some code alterations were made to functions responsible for training and testing of the architectures, and in some cases significant code changes were required to allow dual game learning. Fundamentally, the code base initialises two game environments and then alternates between their episodes to build the training database with equal weighting. The issue then arises where the two games selected do not necessarily have the same scoring range.

During the evaluation of performance for the delayed reward phase of learning, each competitive game is processed equally as was done in the original code. The total reward of each game is then normalized by the maximum score respectively observed as to remove bias on games with higher possible scores and accumulated for the decision as to improvement in the agent. This allows relative improvement of one game to supersede a smaller relative decrease in performance of another.

3 Experimental Results

The following section details the analysis results for the two presented scenarios consisting of a pair of similar games, and a pair of high performing games. The results presented show the relative performance decrease caused by competing environment learning as well as how well the learnt competitive agent then performed against unseen games. All comparisons are made relative to the performance measured by the original research presented by Mnih et al. [8].

3.1 Similar Games

Initially we will discuss the competitive training of two similar games within a common DQN architecture. It is noted that a neural network architecture is required to express a high level functional description of a task within the sequence of neurons of the network, and the success or otherwise in terms of performance lies in its ability to capture that high level function. Within this research, the fundamental question is how competing tasks affect that ability. As such, the first scenario to be applied in this competitive sense is the simultaneous learning of the classic game *Space Invaders* and a very similar game called *Demon Attack*. Both games require the player at the bottom of the screen to fire upwards to destroy descending targets, whilst avoiding being hit by return fire coming down the screen from only a subset of those targets. The motion of the targets and the absence of shields in *Demon Attack* are the only discernable differences between the games.

Under this applied competitive scenario it is hypothesised that there will be only a minor decrease in performance overall, due to the fact that the network is only required to represent some small high level differences between the games within the nodes of the network. The original code base was able to get a score of 121.5% that of human level performance for *Space Invaders*, and 294.2% for *Demon Attack*.

The scenario consisting of competitively playing the similar games of *Space Invaders* and *Demon Attack* within a single DQN architecture has been run, and the resulting performance against these two games is shown below in Figs. 2 and 3 respectively, at discrete stages of the learning process.

The network was able to achieve a mean score of 846.3 in the game *Space Invaders* whilst also achieving an average score of 4684.3 at *Demon Attack*. This is a reduction of approximately half for both games, 42.8% and 48.2% of the original scores respectively. Additionally the DQN agent was still able to beat the human score in *Demon Attack* by a reasonable margin (138%); however the agent was not able to beat the score of a human player in the game *Space Invaders*, only managing to achieve 51% that of the human. Despite this reduction in performance for these specific games, it is expected that the ability to generalise the solution to other games is improved given more diverse sampling of inputs.

A comparison against other games which are compatible with the network are provided in Table 1. Here we are interested in the *generalisability* of an agent to perform an unforeseen task; the ability to receive reward on games that it was not trained on. Table 1 shows the generalisability of the competitive DQN agent (trained on *Space Invaders* and *Demon Attack*) has increased from the original specialist DQN agent (trained on *Space Invaders*). In many cases dramatic increases were observed, the largest was over 15000 times better performance. The remaining games also displayed significant improvement due to the competitive training method, at the expense of performance on the game of *Space*

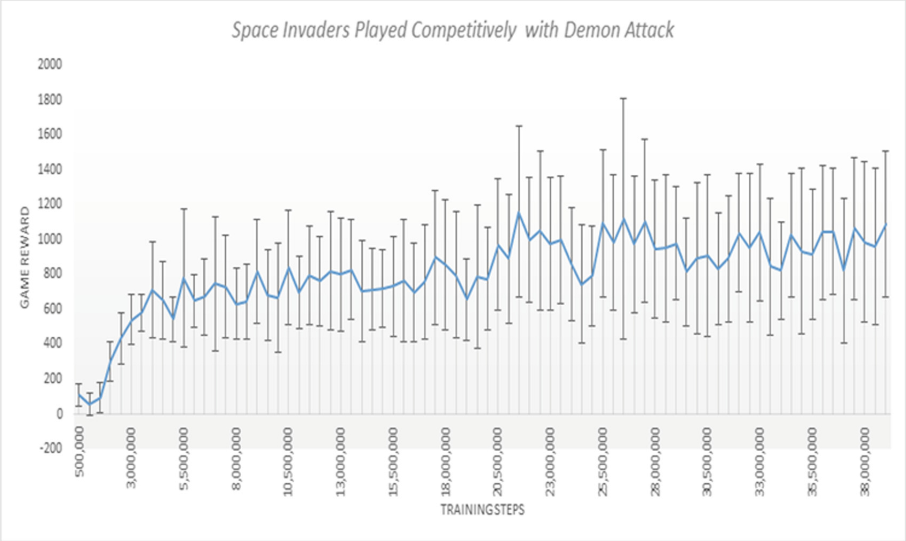

Fig. 2. Measured performance of the game Space Invaders across the competitive Reinforcement Learning process of Space Invaders and Demon Attack

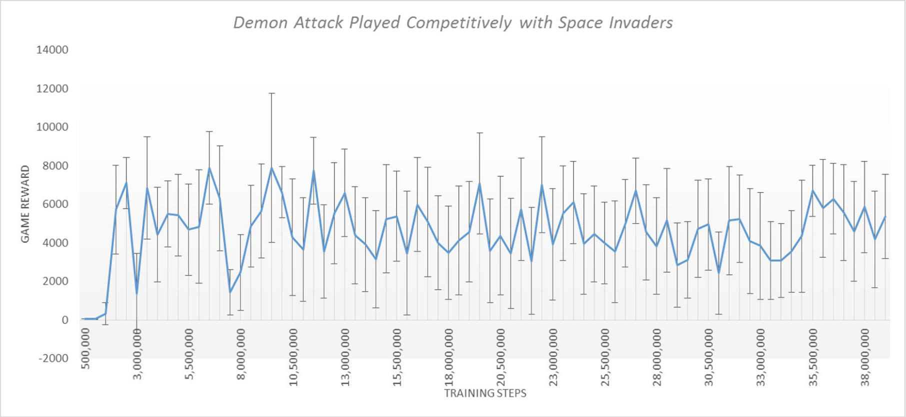

Fig. 3. Measured performance of the game Demon Attack across the competitive Reinforcement Learning process of Space Invaders and Demon Attack

Invaders. Hence, by sacrificing some performance, in this case approximately half, the overall ability to play many games is substantially improved. The only exceptions to this being the games *Breakout* and *Freeway*, which showed no ability to play the game by either network. As such the primary hypothesis of this research has been shown through this example, that competitive training of a DQN agent would result in better general performance at unseen tasks, at the expense of specific ability at a single trained task. Further to this, it is noted the

Table 1. Comparison of game scores obtained by different agents for several Atari games. The first column is a expert human agent. The second column is original specialist DQN agent when applied to the game it specializes in. The third column is a specialist Space Invaders DQN agent when applied to unforeseen games. The forth column is a competitive DQN agent (trained on Space Invaders and Demon Attack) when applied to other games. The fifth column is the performance of the competitive DQN compared to the original DQN. The sixth column is improvement of the competitive DQN over the specialist DQN in being able to generalise to unforeseen games.

Game	Human	Original DQN (±σ)	Specialist DQN (±σ)	Competitive DQN (±σ)	Performance (% Original)	Generalisability (% Increase)
Asteroids	13157	1629 (±542)	0.0 (±0.0)	188.7 (±255.6)	11.6%	100.0%
Atlantis	29028	85641 (±17600)	4050.0 (±1234.8)	5066.7 (±1411.1)	5.9%	25.1%
Bowling	154.8	42.4 (±88)	23.9 (±8.7)	29.9 (±0.6)	70.5%	24.9%
Breakout	31.8	401.2 (±26.9)	0.0 (±0.4)	0.0 (±0.2)	0.0%	0.0%
Demon attack	3401	9711 (±2406)	103.0 (±54.3)	4684.3 (±2547.3)	48.2%	4447.9%
Freeway	29.6	30.3 (±0.7)	0.0 (±0.0)	0.0 (±0.0)	0.0%	0.0%
Name this game	4076	7257 (±547)	287.0 (±269.9)	718.7 (±390.6)	9.9%	150.4%
Q'bert	13455	10596 (±3294)	1.7 (±9.1)	260.8 (±47.7)	2.5%	15549.8%
Space invaders	1652	1976 (±893)	1785.0 (±607.6)	846.3 (±279.7)	42.8%	−52.6%
Up'n down	9082	8456 (±3162)	1239.3 (±745.2)	2039.0 (±763.8)	24.1%	64.5%

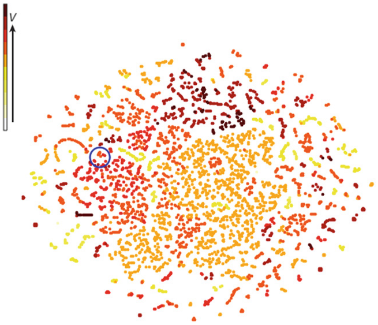

Fig. 4. Network analysis of estimated game reward using the t-SNE technique for the Neural Network trained competitively for the games Space Invaders and Demon Attack

increased ability at some of these games could be due to some small similarity in game play to both *Space Invaders* and *Demon Attack*. For example *Atlantis*; but other games such as *Bowling or Q'bert* which have no similarity to either game showed significant improvement due to competitive learning.

A corollary hypothesis specific to this example of *Similar Games* was that the representative space of the network was capable of capturing the required information to play each game due to the similarity of the games. Figure 4 shows that the network has in fact retained the ability to predict high reward states of the game, where the state information contained within the screen has been clustered by similarities, and predictive of reward according to the clustered region. This corresponds to the findings of the original *DQN* network, except what is also shown is that the network representation has also segregated between the two games, as seen in Fig. 5, where the graphical representations shown have

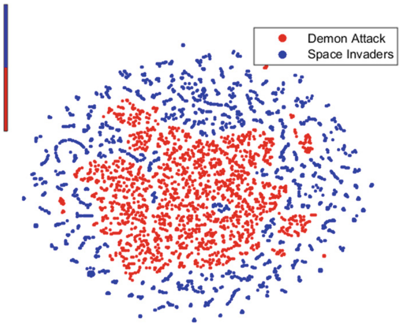

Fig. 5. Network analysis of game type using the t-SNE technique for the Neural Network trained competitively for the games Space Invaders and Demon Attack

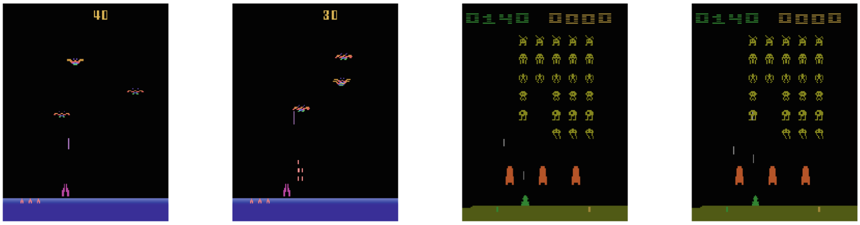

Fig. 6. High reward states for the games Space Invaders and Demon Attack trained under competitive DQN conditions, showing the similarity in game state for the high predicted reward

been generated using the visualization techniques presented by Van der Maaten and Hinton [6] as per the original research.

The final observation to note from the analysis of the competitive training of similar games is that qualitative similarities between the games has also been clustered, shown in Fig. 4 by the blue circle, according to similar predictions of reward as shown in Fig. 6, where the network has determined a game strategy of clearing an area on the left of screen and attacking from one side to minimise the chance of being hit by return fire. Given this similar strategy being applicable to both games, it is not surprising that the network was able to perform relatively well in both cases.

3.2 High Performing Games

Consider the case of two high performing games; games in which the original DQN easily learnt an optimal strategy and dramatically beat the scores of a human player. These games are performing at this high standard due to the inherent simplistic state-action relationship. Two such games being *Boxing* and *Robotank*, where the original code base was able to get a score of 1707.9% that of human level performance for *Boxing*, and 509.0% for *Robotank*. Under this condition of competitive training for the games *Boxing* and *Robotank* it is hypothesised that after competitive training a moderate decrease in performance will be observed, as there is still considerable space within the architecture to represent and interpret the state of the two games. However their actions space is significantly different as are the input signals, for example the game *Boxing* requires the DQN to position the player at a distance to the opponent by moving along two axes, and strike at an opportune moment to minimise the number of hits taken and maximise hits landed. The response has translational invariance, i.e. indifferent to where on the image the players may be, with the exception of being near the edge which restricts movement. In contrast, the game *Robotank* pans left and right in order to find the enemy based on a radar guidance of enemy locations, and fire upon them without being destroyed, i.e. quickly lining up the target with the crosshairs. These are very different game objectives, particularly when considering the previous scenario discussed. Given this, the hypothesis is that the network will be required to prioritise areas of the architecture towards specific tasks, effectively dividing up the expressive power of the network, and as a result, the performance of each game and the rate of convergence to an optimal solution will suffer.

Again the scenario consisting of competitively playing the games *Boxing* and *Robotank* within a single DQN architecture has been run, and the resulting performance against these two games are shown below in Figs. 7 and 8.

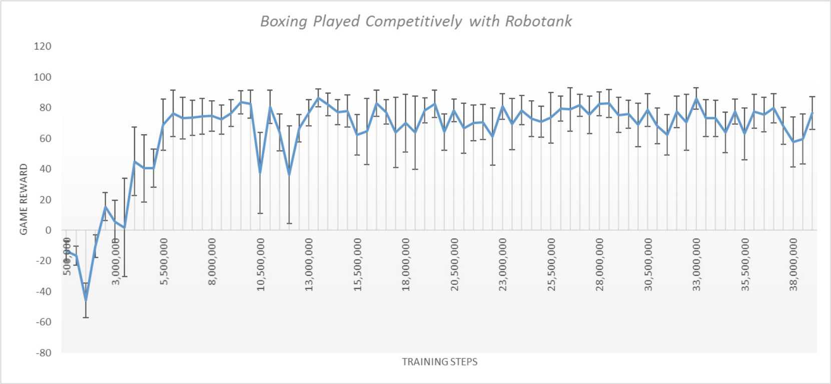

Fig. 7. Measured performance of the game Boxing across the competitive Reinforcement Learning process of Robotank and Boxing

The network was able to achieve a score of 73.7 in the game *Boxing* whilst also achieving a score of 1.4 at *Robotank*. What is of interest here is that the score achieved for *Boxing* is actually higher than what was reported by the original research, suggesting that the competitive training has actually improved the network at its ability to play *Boxing*. However the difference in score is within the confidence interval and hence is not a statistically significant difference. In

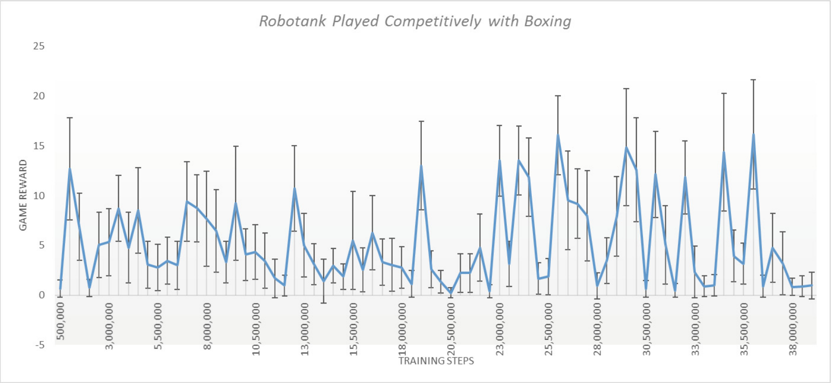

Fig. 8. Measured performance of the game Robotank across the competitive Reinforcement Learning process of Robotank and Boxing

contrast to this, the score achieved at the game *Robotank* has suffered considerably; an approximate reduction of 97%. As was seen in previous analysis, the ability to determine a strategy is critical to overall reward obtained by the network. Comparing the game play of this competitively trained network against *Robotank* and *Boxing* has shown that the same optimal strategy of pushing the opponent into a corner for *Boxing* is found, and hence the high reward obtained. However, this has resulted in a strategy for *Robotank* of simply going around in circles and firing repeatedly, with no attempts to line up a target, with the opponents eventually landing a hit. This is a poor strategy and results in the exceedingly low score. What this suggests is that the network has been unable to learn different strategies for these games, and perhaps that it is unable to differentiate between the two states, or that the short term reward of the game *Boxing* has dominated over the slightly longer term reward mechanism of *Robotank*. Despite this reduction in performance for *Robotank*, it is again expected that the ability to generalise the solution to other games is improved given more diverse sampling of inputs.

A comparison against all other games are provided in Table 2. What the analysis of the remaining games, not seen during training for the competitive case of the high performing games *Boxing* and *Robotank*, has shown is a highly polarized generalisability. In many cases, a dramatic improvement in the ability of the network to gain reward was found; for example the games *Atlantis, Battle Zone, Bowling, Crazy Climber, Gravitar, James Bond, Private Eye, Tutankham and Wizard of Wor*. In contrast, the games *Beam Rider, Centipede, Double Dunk, Frostbite, H.E.R.O., Kung Fu Master, Name This Game, Q'Bert and Seaquest* all showed solid decreases in generalisability, with the remaining games showing minor change to generalisability. What is immediately noticed from this list is the fact that these games are fundamentally different, neither the improved or detrimental generalisability cases follow a pattern, simply that the learnt strategy of pushing to one side of the screen and pushing the fire button either applies

Table 2. Comparison of game scores obtained by the competitively learnt agent against Boxing and Robotank with the original single game learning scenario of Boxing

Game	Human	Original DQN (±σ)	Specialist DQN (±σ)	Competitive DQN (±σ)	Performance (% Original)	Generalisability (% Increase)
Alien	6875	3069 (±1093)	140.0 (±0)	130.0 (±23.3)	4.2 %	−7.1 %
Amidar	1676	739.5 (±3024)	34.8 (±7.5)	2.1 (±0.3)	0.3 %	−94.0 %
Asteroids	13157	1629 (±542)	0.0 (±0)	0.0 (±0.0)	0.0 %	-
Atlantis	29028	85641 (±17600)	1793.3 (±488.4)	2786.7 (±1096.9)	3.3 %	55.4 %
Bank heist	734.4	429.7 (±650)	0.0 (±0)	0.7 (±2.5)	0.2 %	100.0 %
Battle zone	37800	26300 (±7725)	3666.7 (±3507)	6666.7 (±3526.6)	25.3 %	81.8 %
Beam rider	5775	6846 (±1619)	748.3 (±123.2)	452.3 (±176.5)	6.6 %	−39.6 %
Bowling	154.8	42.4 (±88)	0.8 (±1.6)	5.6 (±1.9)	13.1 %	595.8 %
Boxing	4.3	71.8 (±8.4)	78.7 (±15.8)	73.7 (±9.7)	102.6 %	−6.4 %
Breakout	31.8	401.2 (±26.9)	0.4 (±0.9)	0.2 (±0.6)	0.0 %	−53.8 %
Centipede	11963	8309 (±5237)	8388.2 (±3484.6)	3874.9 (±1941.6)	46.6 %	−53.8 %
Chopper command	9882	6687 (±2916)	666.7 (±260.4)	526.7 (±216.4)	7.9 %	−21.0 %
Crazy climber	35411	114103 (±22797)	3.3 (±18.3)	2123.3 (±750.9)	1.9 %	63600.0 %
Demon attack	3401	9711 (±2406)	111.3 (±37.8)	66.7 (±17.1)	0.7 %	−40.1 %
Double dunk	−15.5	−18.1 (± − 2.6)	−23.5 (±1)	−21.4 (±1.5)	−18.2 %	8.8 %
Enduro	301.6	301.8 (±24.6)	0.0 (±0)	0.3 (±1.6)	0.1 %	100.0
Fishing derby	5.5	−0.8 (±19)	−99.0 (±0)	−98.4 (±0.9)	−12200.0 %	0.6 %
Freeway	29.6	30.3 (±0.7)	0.0 (±0)	0.0 (±0.0)	0.0 %	-
Frostbite	4335	328.3 (±250.5)	83.0 (±13.2)	0.0 (±0.0)	0.0 %	−100.0 %
Gopher	2321	8520 (±3279)	16.7 (±37.9)	16.7 (±34.9)	0.2 %	0.0 %
Gravitar	2672	306.7 (±223.9)	6.7 (±36.5)	55.0 (±115.5)	17.9 %	725.0 %
H.E.R.O	25763	19950 (±158)	92.5 (±47)	0.0 (±0.0)	0.0 %	−100.0 %
Ice hockey	0.9	−1.6 (±2.5)	−21.2 (±3.5)	−22.4 (±2.4)	−1302.1 %	−5.8 %
Jamesbond	406.7	576.7 (±175.5)	6.7 (±17.3)	35.0 (±37.5)	6.1 %	425.0 %
Kangaroo	3035	6740 (±2959)	0.0 (±0)	0.0 (±0.0)	0.0 %	-
Krull	2395	3805 (±1033)	622.3 (±227.2)	549.3 (±304.7)	14.4 %	−11.7 %
Kung fu master	22736	23270 (±5955)	126.7 (±161.7)	13.3 (±50.7)	0.1 %	−89.5 %
Montezuma's R	4367	0 (±0)	0.0 (±0)	0.0 (±0.0)	-	-
Ms Pacman	15693	2311 (±525)	374.3 (±212.8)	328.7 (±309.1)	14.2 %	−12.2 %
Name this game	4076	7257 (±547)	1789.3 (±658.1)	333.0 (±212.1)	4.6 %	−81.4 %
Private eye	69571	1788 (±5473)	−599.3 (±308.7)	6.7 (±25.4)	0.4 %	101.1 %
Q'bert	13455	10596 (±3294)	136.7 (±32.7)	2.5 (±7.6)	0.0 %	−98.2 %
River raid	13513	8316 (±1049)	402.7 (±54.5)	244.0 (±50.9)	2.9 %	−39.4 %
Road runner	7845	18257 (±4268)	0.0 (±0)	823.3 (±135.7)	4.5 %	100.0
Robotank	11.9	51.6 (±4.7)	10.9 (±3.5)	1.4 (±1.8)	−97.2 %	−86.8 %
Seaquest	20182	5286 (±1310)	125.3 (±35.2)	12.0 (±12.4)	0.2 %	−90.4 %
Space invaders	1652	1976 (±893)	180.0 (±15.3)	155.7 (±68.7)	7.9 %	−13.5 %
Tennis	−8.9	−2.5 (±1.9)	−24.0 (±0)	−16.8 (±2.4)	−572.0 %	30.0 %
Time pilot	5925	5947 (±1600)	733.3 (±590.9)	1066.7 (±1166.3)	17.9 %	45.5 %
Tutankham	167.6	186.7 (±41.9)	0.1 (±0.4)	15.9 (±9.2)	8.5 %	23799.9 %
Up'n down	9082	8456 (±3162)	623.3 (±268.5)	670.0 (±548.7)	7.9 %	7.5 %
Venture	9083	8457 (±3163)	0.0 (±0)	0.0 (±0.0)	0.0 %	-
Video pinball	9084	8458 (±3164)	0.0 (±0)	8166.3 (±8100.1)	96.6 %	100.0
Wizard of wor	9085	8459 (±3165)	273.3 (±267.7)	483.3 (±136.7)	5.7 %	76.8 %
Zaxxon	9086	8460 (±3166)	0.0 (±0)	0.0 (±0.0)	0.0 %	-

to the specific game to some extent, or it does not. To further analyse what the competitive training of two dissimilar, but high performing games, consider the visualisation of the network as shown in Figs. 9 and 10 for displaying the estimated reward for the given game state, and the game itself respectively.

Comparing the high reward states in Fig. 9 with the network differentiation of the respective games in Fig. 10, the reason for the disparate performance

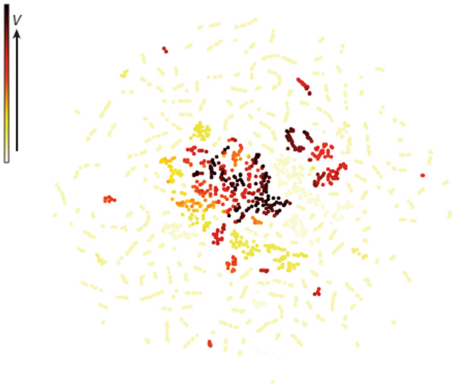

Fig. 9. Network analysis of estimated game reward using the t-SNE technique for the Neural Network trained competitively for the games Boxing and Robotank

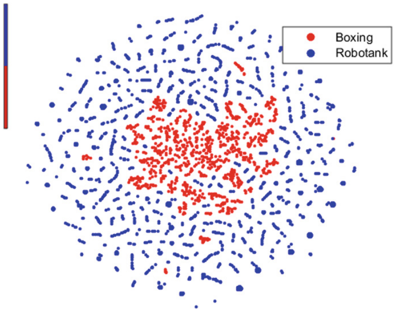

Fig. 10. Network analysis of game type using the t-SNE technique for the Neural Network trained competitively for the games Boxing and Robotank

becomes clear. The node activations for *Boxing* show disparate levels of reward respective of different states of the game. The game *Robotank* does not display this, all game states are representing mid to low reward levels. This is a critical component of the networks ability to play any game, the ability to actively pursue high reward states over low reward states; without this, the game is essentially playing with random actions.

4 Conclusions

We have adapted the original state-of-the-art DQN [8] to simultaneously learn two competing tasks. Considering the case of learning two similar games, it was found that the network could learn to gain reward at both tasks, however this ability rapidly diminished for dissimilar games, even simple ones. The ability to learn multiple examples was completely eroded when considering difficult tasks. Further to this, it was found that a fundamental limitation of this approach was the nature of the game, where the performance of the network relied on the ability to determine a strategy, and strategies which were not conducive to both games resulted in weaker performance.

A major aspect of this research is related to the ability of a reinforcement learning agent to perform multiple tasks well. We have demonstrated that, despite how well a specialist DQN agent performs at a singular specific task, it shows limited ability to generalize to unforeseen tasks. This ability to generalize learning to unforeseen tasks can be improved by competitively training the DQN agent on two tasks simultaneously.

What was largely seen through the analysis of the ability to generalise to unseen tasks, was that training against multiple objectives resulted in improved performance at those unseen tasks, at the expense of performance against specifically trained tasks. In each case of game scenarios applied to the network, be it *Similar, High Performing* games, the network segmented its representational space between the games. Further to this, that segmentation became more pronounced as the difficulty of the training scenario increased, with an increased partitioning clearly observed. What this shows is that the network is devoting representative power to each task, and as hypothesised, that representative power is being saturated as the difficulty increases.

References

1. Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M.: The arcade learning environment: an evaluation platform for general agents. arXiv preprint arXiv:1207.4708 (2012)
2. Bucila, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 535–541. ACM (2006)
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
4. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. arXiv preprint arXiv:1612.00796 (2016)
5. Kuzovkin, I.: Deepmind-atari-deep-q-learner (2015). https://www.github.com/kuz/DeepMind-Atari-Deep-Q-Learner.git
6. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
7. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
8. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
9. Rusu, A.A., Colmenarejo, S.G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., Hadsell, R.: Policy distillation. arXiv preprint arXiv:1511.06295 (2015)

10. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, vol. 1. MIT Press, Cambridge (1998)