Human-level Control through Deep Reinforcement Learning

Volodymyr Mnih, Koray Kavukcuoglu, David Silver et al.

For space purposes, it is permitted to list a subset of the authors on the cover page. Here we listed the paper's three first autho

Keywords: Deep reinforcement learning, DQN, Arcade Learning Environment, Atari 2600.

This is a compressed version of the original abstract, published in Nature in 2015. Some sentences were omitted, including the stated contributions because they are described and contextualized below.

Summary

The theory of reinforcement learning (RL) provides a normative account, deeply rooted in psychological and neuroscientific perspectives on animal behaviour, of how agents may optimize their control of an environment. To use RL successfully in situations approaching real-world complexity, however, agents are confronted with a difficult task: they must derive efficient representations of the environment from high-dimensional sensory inputs, and use these to generalize past experience to new situations. While reinforcement learning agents have achieved some successes in a variety of domains, their applicability has previously been limited to domains in which useful features can be handcrafted, or to domains with fully observed, low-dimensional state spaces. Here we use recent advances in training deep neural networks to develop a novel artificial agent, termed a deep Q-network, that can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning. This work bridges the divide between high-dimensional sensory inputs and actions, resulting in the first artificial agent that is capable of learning to excel at a diverse array of challenging tasks.

Contribution(s)

Contribution makes the ke properties of the algorithm clear. Here, the contributic is both algorithmic (1st sentence) and in terms of the capabilities DQN unlocks (2nd sentence).

1. We introduce deep Q-network (DQN), a more stable RL algorithm based on Q-Learning (Watkins, 1989) that performs non-linear function approximation (FA) through deep convolutional neural networks. DQN can learn successful policies directly from high-dimensional sensory inputs using end-to-end RL.
   Context: The main components of DQN are a slower-moving target, which we call a target network, and an experience replay buffer (Lin, 1991). DQN belongs to the family of fitted value iteration (FVI) algorithms (Gordon, 1995). Most similarly, Neural Fitted Q-Iteration (Riedmiller, 2005) uses experience replay with shallow networks (two-layer multi-layer perceptron), but it achieves stability by fitting the network *de novo* at each iteration from all past experience. In contrast, DQN achieves scalability by sampling batches uniformly at random from buffered recent experience. Alternative existing approaches that perform non-linear FA in RL were either evaluated in a few high-dimensional environments (Koutník et al., 2013) or require domain-specific knowledge (Hausknecht et al., 2012; 2014).

Context places the algorithmic contribution into the existing body of literature, distinguishing the new algorithm from existing approaches. The last sentence adds precision to the claim, clarifying how the claimed results are different from those in related work.

Prioritize providing context where caveats are more informative; for example, no context is provided on experience replay.

2. We show that DQN can learn meaningful representations and that they generalize to data generated from policies other than its own.
   Context: We inspect where representative game states are placed in a two-dimensional t-SNE embedding (van der Maaten & Hinton, 2008) of the representations in the last hidden layer assigned by DQN. We do the same for states generated by the human player.

Context that further elaborates on how the claimed contribution was validated.

3. We demonstrate that DQN, with the same network architecture and hyperparameters, can learn effective control policies in various Atari games, receiving only the pixels and game score as inputs.
   Context: As this is a demonstration, DQN was trained once in each game, and we report the performance of evaluating the best checkpoint obtained during learning. We use 75% of a professional human tester's performance as a baseline. Performance from related work (Bellemare et al., 2012; 2013) is provided as a reference but is not directly comparable since DQN was given access to additional information (e.g., loss-of-life) and many more samples.

Context clarifying the point of the experimental results (i.e., a demonstration, not a benchmark), which justifies some methodological choices. Context is also provided on how empirical oractices differ from existing literature, to allow for better interpretation of the results.

References

Marc G. Bellemare, Joel Veness, and Michael Bowling. Investigating contingency awareness using atari 2600 games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 864–871, 2012.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Geoffrey J. Gordon. Stable function approximation in dynamic programming. In Armand Prieditis and Stuart Russell (eds.), *Proceedings of the International Conference on Machine Learning*, pp. 261–268, 1995.

Matthew J. Hausknecht, Piyush Khandelwal, Risto Miikkulainen, and Peter Stone. HyperNEAT-GGP: a hyperNEAT-based Atari general game player. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 217–224, 2012.

Matthew J. Hausknecht, Joel Lehman, Risto Miikkulainen, and Peter Stone. A neuroevolution approach to general Atari game playing. *IEEE Transactions on Computational Intelligence and AI Games*, 6(4):355–366, 2014.

Jan Koutník, Giuseppe Cuccu, Jürgen Schmidhuber, and Faustino J. Gomez. Evolving large-scale neural networks for vision-based reinforcement learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1061–1068, 2013.

Long Ji Lin. Self-improvement Based on Reinforcement Learning, Planning and Teaching. In *Proceedings of the International Workshop on Machine Learning (ML)*, pp. 323–327, 1991.

Martin A. Riedmiller. Neural Fitted Q Iteration - First Experiences with a Data Efficient Neural Reinforcement Learning Method. In *European Conference on Machine Learning*, pp. 317–328, 2005.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, 1989.