POLITECNICO DI MILANO
Laurea Magistrale in Ingegneria Informatica
Dipartimento di Elettronica, Informazione e Bioingegneria

FAST REINFORCEMENT LEARNING USING DEEP STATE FEATURES

AI & R Lab
Laboratorio di Intelligenza Artificiale
e Robotica del Politecnico di Milano

Supervisor: Prof. Marcello Restelli
Co-supervisors: Dott. Carlo D'Eramo, Matteo Pirotta, Ph.D.

Master's Thesis by:
Daniele Grattarola (student ID 853101)

Academic Year 2016-2017

Contents

Chapter 1

State Of The Art

The integration of RL and neural networks has a long history. Early RL literature [5, 7, 1] presents *connectionist* approaches in conjunction with a variety of RL algorithms, mostly using dense ANNs as approximators for the value functions from low-dimensional (or engineered) state spaces. The recent and exciting achievements of DL, however, have caused a sort of RL *renaissance*, with DRL algorithms outperforming classic RL techinques on environments which were considered intractable. Much like the game of Chess was considered out of the reach of machines until IBM's *Deep Blue* [2] won against the world champion Garry Kasparov, DRL has paved the way to solve a wide spectrum of complex tasks which were previously considered a stronghold of humanity.

In this chapter we present the most important and recent results in DRL research, as well as some work related to the approach that is proposed in this thesis.

1.1 Value-based Deep Reinforcement Learning

In 2015, Mnih et al. [4] introduced the *deep Q-learning* (DQL or, more commonly, DQN) algorithm which we detailed in Section **??**, and basically ignited the field of DRL. The important contributions of DQN consisted in providing an end-to-end framework to train an agent starting from the pixel-level representation of the *Atari* games environments, which proved to be more stable than previous approaches thanks to the use of *experience replay* [3]. Moreover, the same architecture was reused to solve many different games without performing *hyperparameter tuning*, which proved the effectiveness of the method. From this work (which we could call *introductory*), many improvements have been proposed in the literature.

In 2016, Van Hasselt et al. proposed *Double DQN* (DDQN) [8] to solve an overestimation issue in DQN due to the max operator used in the parameters update (see Algorithm **??**). This approach used two separate CNNs: an *online network* to select the action for the collection of samples, and a *target network* to produce the update targets. DDQN performed better than DQN on the *Atari* games.

Also in 2016, Schaul et al. [6] introduced the concept of *prioritized experience replay*, which replaced the uniform sampling from the replay memory of DQN with a sampling strategy weighted by the *TD errors* committed by the network. This improved the performance of both DQN and DDQN.

Wang et al. introduced a slightly different end-to-end *dueling architecture* [9], composed of two different deep estimators: one for the state-value function V and one for the *advantage function* $A : S \times A \to \mathbb{R}$ defined as:

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s) \tag{1.1}$$

In dueling architectures, the two networks share the same convolutional layers but use two separate dense layers. The two streams are then combined to estimate the optimal action-value function as[1]:

$$Q^\pi(s,a) = V^\pi(s) + (A^\pi(s,a) - \max_{a'} A^\pi(s,a')) \tag{1.2}$$

Several other improvements have been proposed to the DQN and DDQN algorithms Munos, RÃ©mi, et al. "Safe and efficient off-policy reinforcement learning. Harutyunyan, Anna, et al. "Q(Î») with Off-Policy Corrections."

1.2 Other approaches

1.2.1 Neural Episodic Control

Differentiable neural computer (DNC) Pritzel, Alexander, et al. "Neural Episodic Control."

1.2.2 AlphaGo

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. (2016a). Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489.

1.2.3 A3C

Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning."

[1] In the original paper, the authors explicitly indicate the dependence of the estimates on different parameters (e.g. $V^\pi(s,a;\phi,\alpha)$ where ϕ is the set of parameters of the convolutional layers and α of the dense layers). For coherence in the notation of this thesis, here we report the estimates computed by the network with the same notation as the estimated functions (i.e. the network which approximates V^π is indicated as V^π, and so on...)

1.2.4 Transfer Learning

PathNet

1.3 Related work

FE: Deep Auto-Encoder Neural Networks in Reinforcement Learning Predict dynamics: Faster Reinforcement Learning After Pretraining Deep Networks to Predict State Dynamics

Bibliography

[1] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.

[2] Murray Campbell, A Joseph Hoane, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[3] Long-H Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3/4):69–97, 1992.

[4] 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. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–33, 2015.

[5] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering, 1994.

[6] T Schaul, J Quan, I Antonoglou, and D Silver. Prioritized experience replay. In *Proceeding of the International Conference on Learning Representations (ICLR)*, 2016.

[7] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[8] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, pages 2094–2100, 2016.

[9] Z Wang, T Schaul, M Hessel, H van Hasselt, M Lanctot, and N. de Freitas. Dueling network architectures for deep reinforcement learning. In *Proceeding of the International Conference on Machine Learning (ICML).*, 2016.