Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230668

Applications of deep reinforcement learning — Alphago

Yingchen Liu

Lassonde School Of Engineering, York University, Toronto, Ontario, Canada, M3J2S5

lyc99@my.yorku.ca

Abstract. With the progress of the times, the field of artificial intelligence (AI) has become one of the hottest fields in the 21st century. Currently, artificial intelligence is successfully used in the retail, financial, and medical industries. Especially in 2016, Google's DeepMind used deep reinforcement learning to train AlphaGo and defeated Lee Sedol, which propelled the field into the public eye. Most people are aware of artificial intelligence, but few understand it. This article will focus on analyzing the literature "Mastering the game of Go with deep neural networks and tree search" and other related articles to introduce the basics of deep reinforcement learning and AlphaGo. Finally, readers will understand how artificial intelligence can successfully imitate humans and defeat humans in Go.

Keywords: Deep Reinforcement Learning, Artificial Intelligence, Deep Learning, Reinforcement Learning, AlphaGo.

1. Introduction

In the 21st century, AI-related technologies have developed rapidly. First, the AlphaGo AI robot developed by Google's DeepMind has defeated the best human players in Go, followed by Boston Dynamics' robot dogs. Now, Tesla's autonomous driving has entered daily life. So Artificial Intelligence has become a popular topic in the 21st century. In this article, we apply the literature analysis method to analyze the basic concepts of deep reinforcement learning. In this article, we take AlphaGo as a case study and explore the principles of AlphaGo in terms of policy and value networks. Through this article, any scholar interested in AI can quickly gain a basic understanding of the field. For researchers who have no relevant knowledge, this article is a helpful way to understand AlphaGo quickly.

2. Preparatory knowledge for deep reinforcement learning

2.1. Deep learning (DL)

The development of deep learning derives from the study of neural networks. Neural networks are constructed based on the principle of biological neurons. Since humans have not fully understood biological neural networks, they can only abstractly construct them as hierarchical structures. Therefore, we use complex algebraic circuits to simulate and construct neural networks. The term "depth" in this context refers to the fact that circuits are usually organized into many layers, which means that there are many steps in the computational path from input to output [1].

DOI: 10.54254/2755-2721/5/20230668

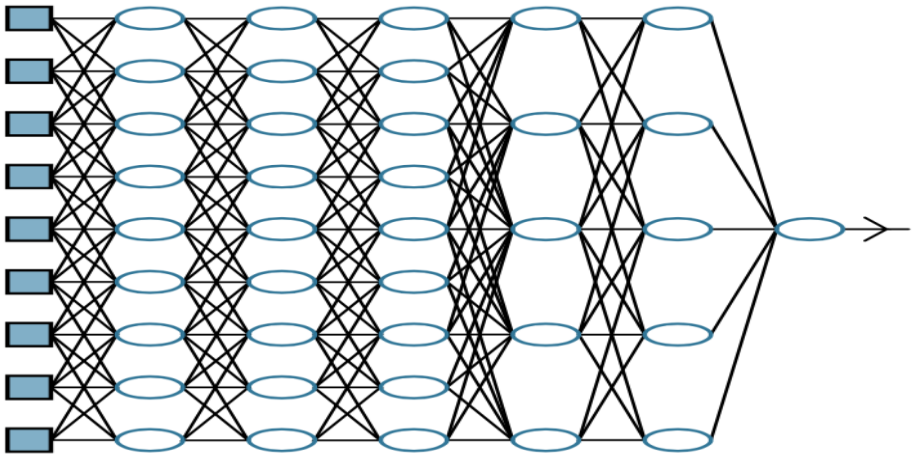

Figure 1. The general structure of a neural network [1].

For Figure 1, we call the leftmost "input layer," the rightmost "output layer," and the rest is called the "hidden layer ."In this process, the system takes the output of the current layer as the input of the next layer and, in this way, learns abstract feature representations from a large amount of training data automatically to discover distributed features of the data [2]. Then the more layers needed means the problems are more complex. For example, AlphaGo's strategy network is 13 layers, with 192 neurons per layer.

2.2. Reinforcement learning (RL)

Reinforcement learning is inspired by behaviorist theory: animals or humans are stimulated by rewards or punishments given by the environment and gradually develop expectations of the stimulus, which results in habitual behavior to obtain the maximum benefit.

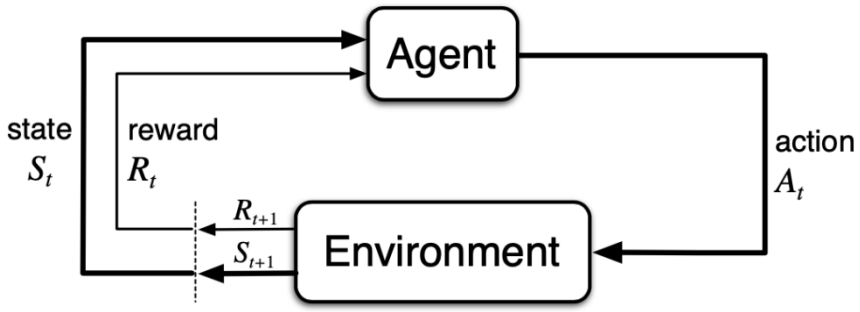

Figure 2. The agent – environment interaction in reinforcement learning [3].

Then we can obtain a process like this: after an agent decides on an action, the environment receives the action and transfers it to a new state, then a reward is given for the action, and after the agent receives the reward, it corrects its decision and continues to decide in the new state. The Markov decision process (MDP) for reinforcement learning combines several processes. We assign a probability that they may be chosen for different actions, which is γ.

2.3. Deep reinforcement learning (DRL)

Deep reinforcement learning, as the name implies, combines deep learning and reinforcement learning because deep learning has strong perceptual ability but lacks specific decision-making abilities. At the

DOI: 10.54254/2755-2721/5/20230668

same time, reinforcement learning has strong decision-making ability but is powerless against perceptual problems. So scholars have combined these two approaches to make them complementary to solve more complex problems [4]. The real beginning of deep reinforcement learning was DeepMind's DQN (deep Q network) algorithm which was published at NIPS in 2013. It learns strategies directly from pixel images to play Atari games [5]. After this, DeepMind created AlphaGo, based on deep reinforcement learning and Monte Carlo tree search, and defeated the top human Go player [6].

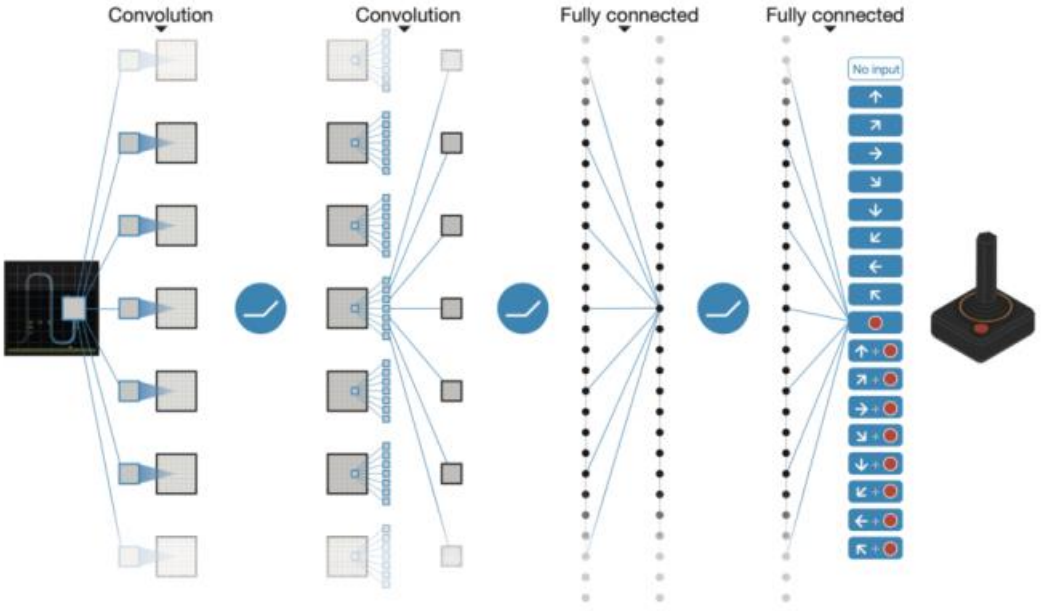

Figure 3. The Schematic illustration of the convolutional neural network[7].

For the DQN algorithm, the application in the Atari game successfully combines neural networks and Q Learning. DQN directly takes the original images of the game as input rather than relying on manual extraction. Then passes through multiple convolutional and fully connected layers, outputting the Q-value of the optional actions in the current state, thus achieving end-to-end learning control. In this approach, the machine can be manipulated like a human player [4].

3. How AlphaGo works and how it is implemented

3.1. Principle of AlphaGo

Before AlphaGo, humans had used Deep Blue to defeat grandmasters in chess, but AlphaGo did not copy its method and use it directly because of the vast difference in the total number of states between the two games. Each chess move has a search breadth of 35 and a depth of 80; however, Go has a breadth of 250 and a depth of 150 [8]. This results in a search space of approximately 1e50 for chess but 1e170 for Go [9]. For Go, a violent method of searching the entire strategy space is impossible. AlphaGo first imitates a human player's feeling of chess to deduce the best position for the current step and then extrapolates this step to determine the best position for the current move, thus reducing the breadth of the search. For backward extrapolation, AlphaGo does not extrapolate the whole game directly; it imitates human players and extrapolates only a few, or a dozen moves backward to reduce the search depth [9]. Finally, the Monte Carlo tree search combines the two to obtain the final board victory [10].

3.2. Policy network and value network

The process of reducing the search space size can be divided into three parts: supervised learning of policy networks, reinforcement learning of policy networks, and reinforcement learning of value

DOI: 10.54254/2755-2721/5/20230668

networks [6]. AlphaGo directly learns the moves of human masters for supervised learning of policy networks to gain chess sense. AlphaGo uses a 13-layer strategy network trained from 30 million Go games on the KGS Go server [6]. In this model, the policy network predicts the human player's moves with a 57% accuracy rate and takes only 3 seconds. AlphaGo also uses an additional "fast move" method, which takes only 2 microseconds to make a move, but the correct rate drops to 24.2% [6].

The structure of reinforcement learning of policy networks is similar to supervised learning of policy networks. However, rather than being based on learning human experts alone, it becomes an iterative game using the present policy network and a random selection of policy networks. During this time, AlphaGo uses a reward function r(s) and a reward rule to maximize the expected result [6]. The results also proved the effectiveness of reinforcement learning, with the reinforcement learning of policy network having an 80% win rate against the supervised learning of policy network in the self-gaming process [6].

The purpose of reinforcement learning for value networks is to evaluate the merits of the current move to evaluate the game. During this time, AlphaGo uses a value function for training, which unfortunately leads to overfitting. To alleviate this problem, the experts switched to self-gaming, with 30 million games generated in reinforcement learning of the policy network, which finally succeeded in reducing the mean square error (MSE) to the normal range [6].

3.3. Monte Carlo Tree search (MCTS)

Because the main feature of the policy network and value network is to reduce the search width and depth of the game tree through pruning to reduce the search space size. However, simply reducing the search space is not enough, so AlphaGo uses Monte Carlo tree search to implement the game tree search [9]. The Monte Carlo tree search algorithm is based on the principle of randomizing draws and then updating the value of the current draw by the final win or loss. In this way, AlphaGo is trained repeatedly to obtain a good drop plan [9]. Finally, AlphaGo can think like a human by combining the strategy and value networks through Monte Carlo tree search

4. Conclusion

Through AlphaGo's design principles, this paper analyzes how artificial intelligence can be trained quickly through deep learning and reinforcement learning. After humans give them a general guideline, they can quickly learn and train themselves and defeat human players in a brief period. We used to think it would take at least ten years for artificial intelligence to defeat a human player at Go, but it has proven that it does not require that long. AlphaZero came after AlphaGo, and AlphaZero defeated AlphaGo in only three days. Therefore, more and more excellent AI technologies have been successfully developed, designed, and applied in recent years, such as autonomous driving, which has improved from one-dimensional learning to multi-dimensional learning. Therefore, this article has a positive attitude towards the future of artificial intelligence and believes that in the right direction, artificial intelligence will enable a long-term positive cycle in the world.

References

[1] Russell, Stuart J. (Stuart Jonathan) et al. Artificial Intelligence: a Modern Approach. Fourth edition. Hoboken, NJ: Pearson, 2021. Print.

[2] LeCun, Bengio, Y., & Hinton, G. (2015). Deep learning. Nature (London), 521(7553), 436–444. https://doi.org/10.1038/nature14539

[3] Sutton, & Barto, A. G. (1998). Reinforcement learning an introduction. MIT Press.

[4] Tang, Shao, Zhao, et al. Advances in deep reinforcement learning: from AlphaGo to AlphaGo Zero[J]. Control Theory and Applications, 2017, 34(12):1529-1546. doi:10.7641/CTA.2017.70808.

[5] Li, Cao, Zhang, et al. A review of knowledge-based deep reinforcement learning research[J]. Systems Engineering and Electronics Technology,2017,39(11):2603-2613. doi:10.3969/j.issn.1001-506X.2017.11.30.

[6] Silver, Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J.,

DOI: 10.54254/2755-2721/5/20230668

Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature (London), 529(7587), 484–489. https://doi.org/10.1038/nature16961

[7] Mnih, Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature (London), 518(7540), 529–533. https://doi.org/10.1038/nature14236

[8] Allis, L. V. Searching for Solutions in Games and Artificial Intelligence. Ph.D. thesis, University of Limburg, Maastricht, The Netherlands (1994).

[9] Tao, Wu, Xiao. Analysis of the principle of AlphaGo technology and the prospect of military application of artificial intelligence[J]. Journal of Command and Control,2016,2(2):114-120. doi:10.3969/j.issn.2096-0204.2016.02.0114.

[10] Chen M.Y. Research on principle and future application of AlphaGo and AlphaZero[J]. Communication World,2019,26(12):22-23. doi:10.3969/j.issn.1006-4222.2019.12.012.