# Searching by learning: Exploring artificial general intelligence on small board games by deep reinforcement learning

Wang, H.

# Chapter 8

# Conclusion

This work relies on a framework, as used by AlphaZero, that combines online searching and offline learning. This framework has become an effective approach in deep reinforcement learning since AlphaGo series algorithms achieve super human level performance on playing complex games. Within this framework, the offline learning model provides state values to guide MCTS search, and the neural network is trained by the self-play game records played by MCTS search results.

Before deep neural networks were common in reinforcement learning (due to the limits of hardware computation capacity), table based approaches of Q-learning were used, and MCTS, as online search methods to play small versions of Go. A combination of online search (MCS) and offline learning (table based Q-learning) in GGP was assessed. MCS was used to generate expert data for self-play at the beginning phase for the table based Q-learning. The results show that table based Q-learning converges in GGP and has potential to be improved by MCS techniques. Inspired by our work in Chapter 2, [60] establish their deep reinforcement learning GGP system.

Therefore, in this dissertation, an AlphaZero-like self-play framework was studied to further investigate the combination of online search and offline learning in a deep reinforcement learning context. A detailed analysis of 12 hyper-parameters was provided. Among these hyper-parameters, four interesting hyper-parameters are analyzed further, and several interesting correlations are presented in Chapter 3. Then the alternative loss functions were evaluated to see how value loss and policy loss contribute to training in Chapter 4.

AlphaZero-like self-play initializes its neural network randomly, and therefore suffers from a cold-start problem, just as table based Q-learning, which is initialized as empty. Applying MCTS enhancements to generate expert data at the start phase of training achieves better performance. This methods is called the warm-start method (Chapter 5). Furthermore, an adaptive warm-start method was proposed to control the necessary iteration length, which is more robust to different enhancements and different training runs (Chapter 6).

In Chapter 7, the ranked reward method was combined with AlphaZero-like self-play to tackle a complex single agent combinatorial game, Morpion Solitaire, and achieved a near human level grid. This chapter highlights the potential of the AlphaZero framework to provide competitive results in combinatorial searching games starting from tabula rasa setting.

Next, the main contributions of this dissertation will be presented in Sect 8.1 and directions for future work will be discussed in Sect 8.2.

## 8.1 Contributions

This dissertation mainly focus on applying searching and learning methods of reinforcement learning in GGP and AlphaZero-like self-play framework. The main contributions can be summarized as follows.

**Classical Q-learning can be used to play GGP games, although training is slow.** This finding provides a basis for applying deep neural networks to GGP. The MCS enhancement generates better training examples for Q-learning at the start phase of training, which also reveals a promising direction for deep reinforcement learning approaches like AlphaZero-like self-play to be further improved.

**For AlphaZero-like self-play, balancing the number of outer iteration loop and the inner epochs training is a key point to generalize more efficient training examples and to avoid useless and redundant training.** Since the overall epochs number for training is influenced by the outer iteration, the result is that the neural network is trained by the same training examples for too many epochs, if the epoch number is too big. Similarly, it is found that the MCTS search is costly, but that the improvements brought by more MCTS simulation time can be compensated by a better trained model, which also requires proper balancing method.

**The policy loss and value loss functions contribute differently to different games in an AlphaZero-like self-play framework, but a sum of these two losses is a reasonable compromise choice.** The neural network model of AlphaZero-like self-play has two heads, i.e. policy head and value head. This poses the question whether policy or value loss is necessary for training, and how they contribute to the training. Again, this confirms that the sum could be a compromise choice, but not necessary the best choice.

**In both general games frameworks (GGP and AlphaZero-like self-play), employing search enhancements at the start of training improves final training results. For self-play, learning from scratch is not the default best choice. Existing available human expert data and expert data from AI programs can be used to improve training, especially at the start phase of training.** Since the model is initialized randomly, the self-play agent based on such models performs nearly randomly which results in bad training examples. Therefore, search enhancements can be used to generate better examples until the model is improved enough to outperform the search enhancements. Our findings on both frameworks suggest boosting the training at start phase of training by search enhancements like MCS and MCTS with RAVE works better.

**The structure of combining online search and offline training can be improved by MCTS enhancements.** From table based Q-learning to neural network based reinforcement learning, search and learning are both important. It is also found that, just like MCTS with RAVE improves table based learning in [18], MCTS with RAVE can also be used to improve the performance of neural network based reinforcement learning, see Chapter 5 and 6.

**AlphaZero-like self-play can be transferred to solve a complex single agent combinatorial game when assisted by other techniques like ranked reward.** Since it is not possible to directly determine a win or loss for single player combinatorial games like Morpion Solitaire, methods like ranked reward can be used to set sub-goals for such sparse reward-long episode scheduling problems. It is found that AlphaZero-like self-play can be used to also solve complex combinatorial games. The result with Morpion Solitaire indicates a promising future for this approach.

## 8.2 Outlook

The research reported in this thesis has yielded many interesting results. We now enlist some promising avenues for further research.

Inspired by using table based Q-learning in GGP, Goldwasser et al. [60] have build a deep reinforcement learning framework, showing that the deep neural network can easily be embedded into it. The next step could be applying heuristic search enhancements to improve such deep reinforcement learning framework.

Other possible studies on warm-start enhancements of AlphaZero-like self-play have not been conducted yet. Thus, a number of interesting problems remain to be investigated.

- ★ Which enhancements will work best on which games? Does the hypothesis hold that games with more consistent opening plans benefit more from the warm-start?
- ★ How should the weight (weight $w$ is the parameter which is used to combine the different enhancements search results) be changed along with the training iteration progress? Linearly or non-linearly? In our experiments it simply decays linearly.
- ★ There are more parameters that are critical and that could not really be explored yet due to computational cost, but this exploration may reveal important performance gains. For example, the MCTS simulation count ($m$) and the step threshold ($T'$).
- ★ Other warm-start enhancements, e.g., built on variants of RHEA's or hybrids of it, can be explored.
- ★ All our current test cases are relatively small games. How do the results transfer to larger games, or to different applications?

For single agent problems, our first results on Morpion Solitaire give us reason to believe that there remain ample possibilities to improve the approach by investigating the following aspects:

- ★ Parameter Tuning: such as the Monte Carlo simulation times. Since good solutions are sparse in this game, maybe more exploration is beneficial?
- ★ Neural Network Design: It is reported that Pointer Networks perform better on combinatorial problems [125]. A next step could be to also make the neural network structure deeper.

- ★ Local Optima: By monitoring the reward list $B$, it can be enlarged to allow more exploration, once it gets stuck in a locally optimal solution.
- ★ By adding more computational resources and parallelization results can be enhanced.

In addition, it is also interesting to further study the importance of searching and learning in AlphaZero-like self-play, as it is still unclear if searching (like MCTS) is necessary for the last part of a long term training. Searching is quite expensive and normally the last part of training does not give too much improvement.

To further explore the AGI in deep reinforcement learning, curriculum learning [95], meta learning [130] and transfer learning techniques [46, 54] should be also combined to deal with more general tasks.

Besides, in our AlphaZero-like self-play experiments, we suffer from the shortage of computation resources. In the future, we can technically applying parallelization programming for the self-play phase of AlphaZero. And optimizing the neural network structure is also useful to speed up the training.

In consequence, this thesis should not only provide new methods and results but also encourage researchers to help explore these approaches and questions in future investigations.