Can Meta-Interpretive Learning Outperform Deep Reinforcement Learning of Evaluable Game Strategies?

Céline Hocquette
Imperial College London, Department of Computing
celine.hocquette16@imperial.ac.uk

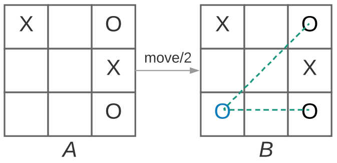
1 Introduction

World-class human players have been outperformed in a number of complex two person games such as Go by Deep Reinforcement Learning systems [Silver *et al.*, 2016]. However, several drawbacks can be identified for these systems: 1) The data efficiency is unclear given they appear to require far more training games to achieve such performance than any human player might experience in a lifetime. 2) These systems are not easily interpretable as they provide limited explanation about how decisions are made. 3) These systems do not provide transferability of the learned strategies to other games. We study in this work how an explicit logical representation can overcome these limitations.

For example, an applicable strategy for playing Noughts-and-Crosses is to lead double attacks when possible, an example of which is shown in Figure 1. Player O executes a move from board A to board B which creates two threats represented in green, and results in a forced win for O. The rules presented in Figure 1 describe such a strategy. A and B are variables representing states that encode both the board description and the active player. A move from A to B is a winning move if the opponent can not immediately win and cannot make a move to prevent an immediate win. These rules provide an understandable strategy for winning in two moves. Moreover, they are transferable to more complex games as they are generally true for describing double attacks.

We introduce a new logical system called *MIGO* [1] designed for learning two player game optimal strategies of the form presented in Figure 1. It benefits from a strong inductive bias which provides the capability to learn efficiently from a few examples of games played. Additionally, *MIGO*'s learned rules are relatively easy to comprehend, and are demonstrated to achieve significant transfer learning.

Owing to tractability considerations, minimax regret of a learning system cannot be evaluated in complex games. In this work, we consider a simple game (Noughts-and-Crosses) in which minimax regret can be efficiently evaluated. We use these games to compare Cumulative Minimax Regret for variants of both standard and deep reinforcement learning against two variants of *MIGO*.



```
win_2(A,B):-win_2_1_1(A,B),not(win_2_1_1(B,C)).
  win_2_1_1(A,B):-move(A,B),not(win_1(B,C)).
       win_1(A,B):- move(A,B),won(B).
```

Figure 1: Example of optimal move from board A to board B. For all moves of X from board B, O can win in one move. This statement can be expressed with the logic program presented: O makes a move such that X cannot immediately win nor make a move that blocks O.

2 Related Work

Various early approaches to game strategies [Shapiro and Niblett, 1982; Quinlan, 1983] used the decision tree learner ID3 to classify minimax depth-of-win for positions in chess end games. These approaches used a set of carefully selected board attributes as features. Conversely, MIGO is provided with a set of three relational primitives (move/2, won/1, drawn/1) representing the minimal information a human would expect to know before playing a two person game.

Classical reinforcement learning approaches, and more recently Deep Q-learning [Mnih *et al.*, 2015], are based upon the identifciation of a Q-function [Watkins, 1989]. The learned strategy is implicitly encoded into the Q-value parameters. Conversely, *MIGO* aims at deriving hypotheses describing an optimal strategy from examples of moves, which provides better understandability of the learned strategy.

In the relational reinforcement learning (RRL) framework [Džeroski *et al.*, 2001], states, actions and policies are represented relationally. The learning is also based upon the identification of Q-values whereas *MIGO* learns hypotheses from examples of moves. Both RRL and *MIGO* provide the ability to carry over the policies learned in simple domains to more complex situations. However, most RRL systems aim at learning single agent policy and, in contrast to *MIGO*, are not designed to learn to play two person games.

3 Completed Work

MIGO uses Meta-Interpretive Learning (MIL), a form of inductive logic programming which supports predicate inven-

[1] From the children's game-playing phrase *My go!* and the literal translation into English of the French word *Ordinateur* which means computer.

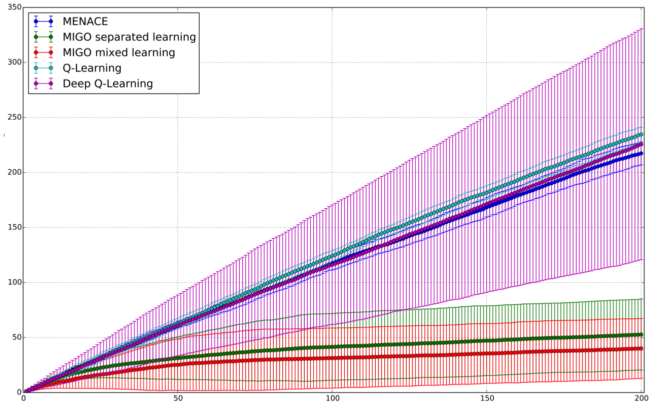


Figure 2: Cumulative regret versus the number of games played for Noughts-and-Crosses

tion and learning recursive theories [Muggleton *et al.*, 2014; 2015]. *MIGO* extends MIL to additionnally support Dependent Learning [Lin *et al.*, 2014]. The idea is to first learn low-level predicates from single examples and with limited complexity. The definitions are added into the background knowledge such that they can be used in further definitions. The process iterates until no further predicates can be learned. Practically, for successive values of k a series of interrelated definitions are learned for predicates win_$k(A, B)$ and draw_$k(A, B)$. These predicates define maintenance of minimax win and draw in k-ply when moving from position A to B. For instance, *MIGO* first learns a simple definition of *win_1/1* for winning in one move. Next, a predicate *win_2/1* describing the action of winning in two moves can be built from *win_1/1* as shown in Figure 1.

MIGO distinguishes itself from classical reinforcement learning approaches by the way it addresses the Credit Assignment Problem. We identify examples of moves that necessarily are positive examples for the task of winning or drawing. We assume the learner plays against an optimal opponent and that the game starts from a randomly chosen initial board B and can demonstrate that two categories of moves are necessarily positive examples for *win/2* or *draw/2* under these assumptions. However, no negative examples can be identified under these assumptions. Therefore, the learning protocol is based upon learning from positive examples only.

The reinforcement learning systems considered for comparison are MENACE [Michie, 1963] which is the world's first reinforcezment learning system, Tabular Q-learning [Watkins, 1989] and Deep Q-learning [Mnih *et al.*, 2015]. In our experiment all tested variants of both normal and deep reinforcement learning have worse performance (higher cumulative minimax regret) than both variants of *MIGO* on Noughts-and-Crosses as shown in Figure 2.

4 Conclusion and Future Work

This work introduces a novel logical system named *MIGO* for learning two-player-game strategies and based upon the MIL framework. Our experiment have demonstrated that *MIGO* achieves lower Cumulative Minimax Regret compared to Deep and classical Q-Learning. Moreover, strategies learned with *MIGO* are general enough to be transferable to more complex games. Learned strategies are also relatively easy to comprehend.

One current limitation of *MIGO* is the limited scalability. The execution of learned strategies is computationally expensive as it browses the minimax tree to evaluate whether a move is a winning move. Therefore the running time increases rapidly with the state dimensions. The scalability is also limited by initial assumptions: the current version of *MIGO* requires a minimax player as opponent which is intractable in large dimensions. We further plan to extend this framework by relaxing our credit assignment protocol and weakening the optimal opponent assumption. A solution would be to learn from self-play.

Despite these limitations, we believe the novel approach introduced in this work opens exciting new avenues for machine learning game strategy.

References

[Džeroski *et al.*, 2001] Sašo Džeroski, Luc De Raedt, and Kurt Driessens. Relational reinforcement learning. *Machine Learning*, 43(1):7–52, Apr 2001.

[Lin *et al.*, 2014] Dianhuan Lin, Eyal Dechter, and Kevin M. et al Ellis. Bias reformulation for one-shot function induction. In *In Proceedings of the 23rd European Conference on Artificial Intelligence (ECAI 2014)*, pages 525–530,. IOS Press, 2014.

[Michie, 1963] Donald Michie. Experiments on the mechanization of game-learning part i. characterization of the model and its parameters. *The Computer Journal, Volume 6, Issue 3*, pages 232–236, 1963.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, and et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 02 2015.

[Muggleton *et al.*, 2014] Stephen H. Muggleton, Dianhuan Lin, Niels Pahlavi, and Alireza Tamaddoni-Nezhad. Meta-interpretive learning: application to grammatical inference. *Machine Learning 94*, pages 25–49, 2014.

[Muggleton *et al.*, 2015] Stephen H. Muggleton, Dianhuan Lin, and Alireza Tamaddoni-Nezhad. Meta-interpretive learning of higher-order dyadic datalog: predicate invention revisited. *Machine Learning*, 100(1):49–73, Jul 2015.

[Quinlan, 1983] J. Ross Quinlan. *Learning Efficient Classification Procedures and Their Application to Chess End Games*, pages 463–482. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[Shapiro and Niblett, 1982] Alen Shapiro and Timothy Niblett. Automatic induction of classification rules for a chess endgame. In M.R.B. Clarke, editor, *Advances in Computer Chess*, volume 3, pages 73–91. Pergammon, Oxford, 1982.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, and et al. Mastering the game of go with deep neural networks and tree search. 529:484–489, 01 2016.

[Watkins, 1989] Christopher Watkins. Learning from delayed rewards, phd thesis. 1989.