NorwAI
Norwegian Research Center
for AI Innovation

Deep Reinforcement Learning for International Diplomacy: Learning to Play Map Variants

Thomas Løkkeborg

NTNU

Abstract

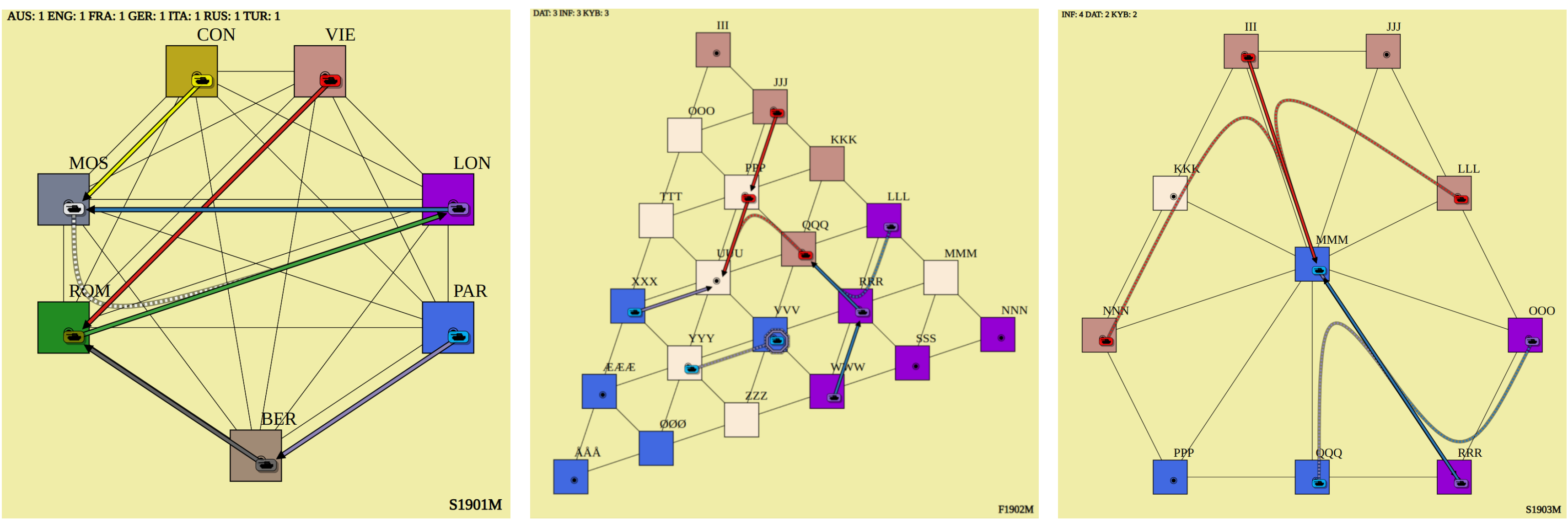
The thesis [1] investigates the generalisability of recent advancements in deep reinforcement learning techniques for the board game Diplomacy by training agents on three map variants of the classic game.

Introduction

The board game Diplomacy has received much attention in recent years as a benchmark problem for the field of artificial intelligence. Diplomacy features a massive combinatorial action space, a mix of cooperation and competition, negotiation in natural language, a deterministic ruleset, and simultaneous action selection. These traits pose a novel challenge to artificial intelligence research, and some are shared with real-life issues like negotiation, tactics, and coordination.

Recent research in artificial intelligence for Diplomacy has utilized deep reinforcement learning with (generalized) policy iteration, similar to AlphaGo Zero of Silver et al. [2]. State-of-the-art techniques have achieved success in the original formulation of the classic game. The success raises interest as to whether the techniques generalize to other problems.

The long lifespan and popularity of the game has spawned a culture of creating variants of the classic game. Game variants modify the ruleset, map topology, and player count to create new challenges for the player. As a step towards general applicability of state-of-the-art techniques, the thesis investigates their application in variants of the classic game.

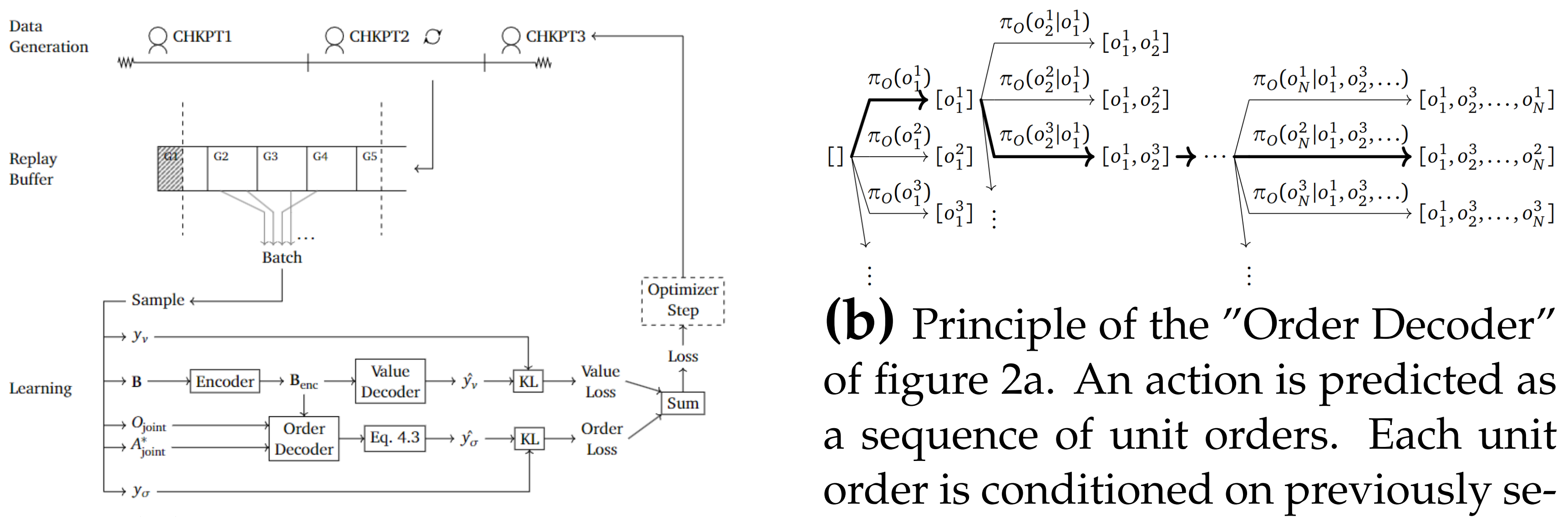


(a) Pure variant **(b)** Lattice variant **(c)** Hub variant

Figure 1: Visualizations of example game states for each of the three game variants. Each player (distinguished by color) seeks to control a majority of the dotted regions on the board by issuing orders to its units. Action selection is simultaneous, and consists of each player choosing one order per unit. Orders are shown here as solid arrows (move), dashed arrows (support another unit's order), or a hexagon (do nothing).

Methodology

Inspired by state-of-the-art, an agent is implemented that at each turn performs a game-theoretic search over a subset of the joint action space, with payoff given by next-state end-game score prediction from a neural network. Action subsets are generated by a neural network that decomposes the action space as a sequential selection of sub-actions. Over time, the accuracy of end-game score prediction is improved via bootstrapped score estimates, and the quality of the generated action subset is improved by making actions valued by search more likely for inclusion in the future. The neural networks train from scratch with no human data, and an action exploration procedure helps discover reasonable actions.



(a) Training architecture

(b) Principle of the ”Order Decoder” of figure 2a. An action is predicted as a sequence of unit orders. Each unit order is conditioned on previously selected orders

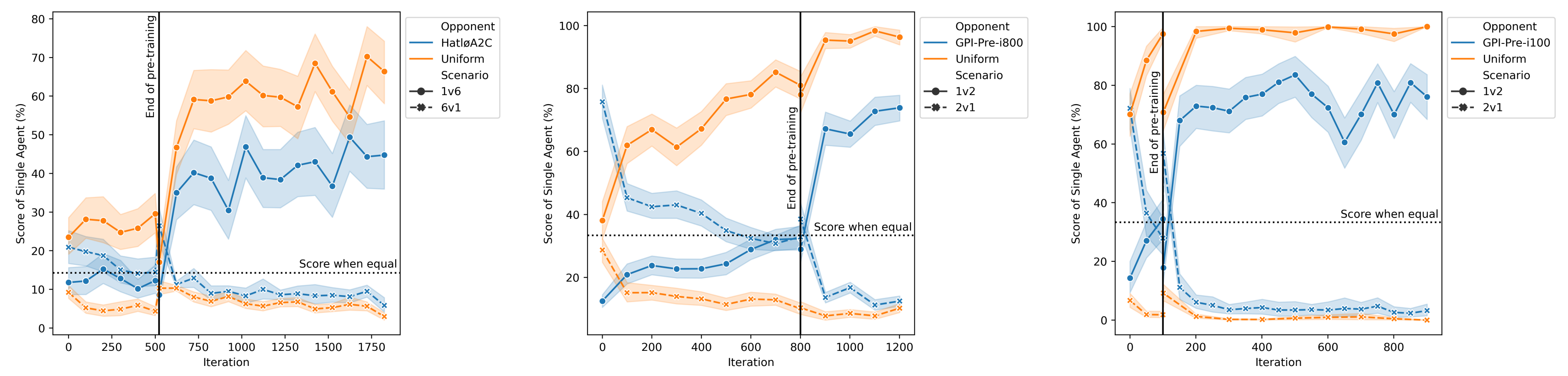
Figure 2: Key aspects of the agent implementation.

Figure 2a shows how a replay buffer of self-play games is used to train a new checkpoint of the agent. A training sample consists of the board state **B**, joint action space O_{joint} and the action subset that was searched over A^*_{joint} as input, and the discounted end-game scores y_v and probability distribution resulting from game-theoretic search y_σ as output targets. The network is trained to predict $\hat{y_v}$ and $\hat{y_\sigma}$ close to y_v and y_σ.

Results

Agents are trained through self-play on the three non-communicative Diplomacy map variants shown in figure 1, and evaluated through skill in tournaments with baseline agents: an actor-critic agent, early checkpoints of the agent under training, and a “uniform” agent acting at random.

Skill is measured by running “one-versus-all” tournaments between the agent under training and baseline agents. For an N-player Diplomacy map variant, the agent under training first plays games where it controls a single player and the $N-1$ other players are controlled by the baseline agent. The roles are then reversed, and more games are played. Figure 3 shows, as a function of training iterations, the average score of a single agent playing 100 games against a population of another, along with the 95% confidence interval. Also shown is the expected average score when an agent plays against copies of itself: $\frac{1}{N}$. Baseline agents are separated by color. Solid lines show the score of the agent under training playing as the single agent. These lines trend upward and away from $\frac{1}{N}$, showing that the agent under training is increasingly capable of outperforming a population of baseline agents. Dashed lines show the score of the baseline agent playing as the single agent. These lines trend downward and away from $\frac{1}{N}$, showing that the agent under training is increasingly capable of locking out the single baseline agent.



(a) Pure variant **(b)** Lattice variant **(c)** Hub variant

Figure 3: Skill of agent on each game variant across training iterations.

Conclusion

The work shows that state-of-the-art deep reinforcement learning techniques that have seen success on the classic Diplomacy map can be applied successfully in alternative map topologies, which hints at the generality of the techniques and acts as a step toward their application in real-life issues.

Acknowledgment

I would like to thank my supervisor Professor Keith. L. Downing for guiding me through this daunting task, and for giving me an opportunity to combine my master’s degree with my passion for Diplomacy. I would also like to thank my family and friends.

References

[1] T. Løkkeborg. Deep Reinforcement Learning for International Diplomacy: Learning to Play Map Variants. Master’s thesis, NTNU, 2023. Accepted: 2023-09-14T17:21:41Z.

[2] Silver, D., Schrittwieser, J., Simonyan, K. *et al.* Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, Oct. 2017. Number: 7676 Publisher: Nature Publishing Group.