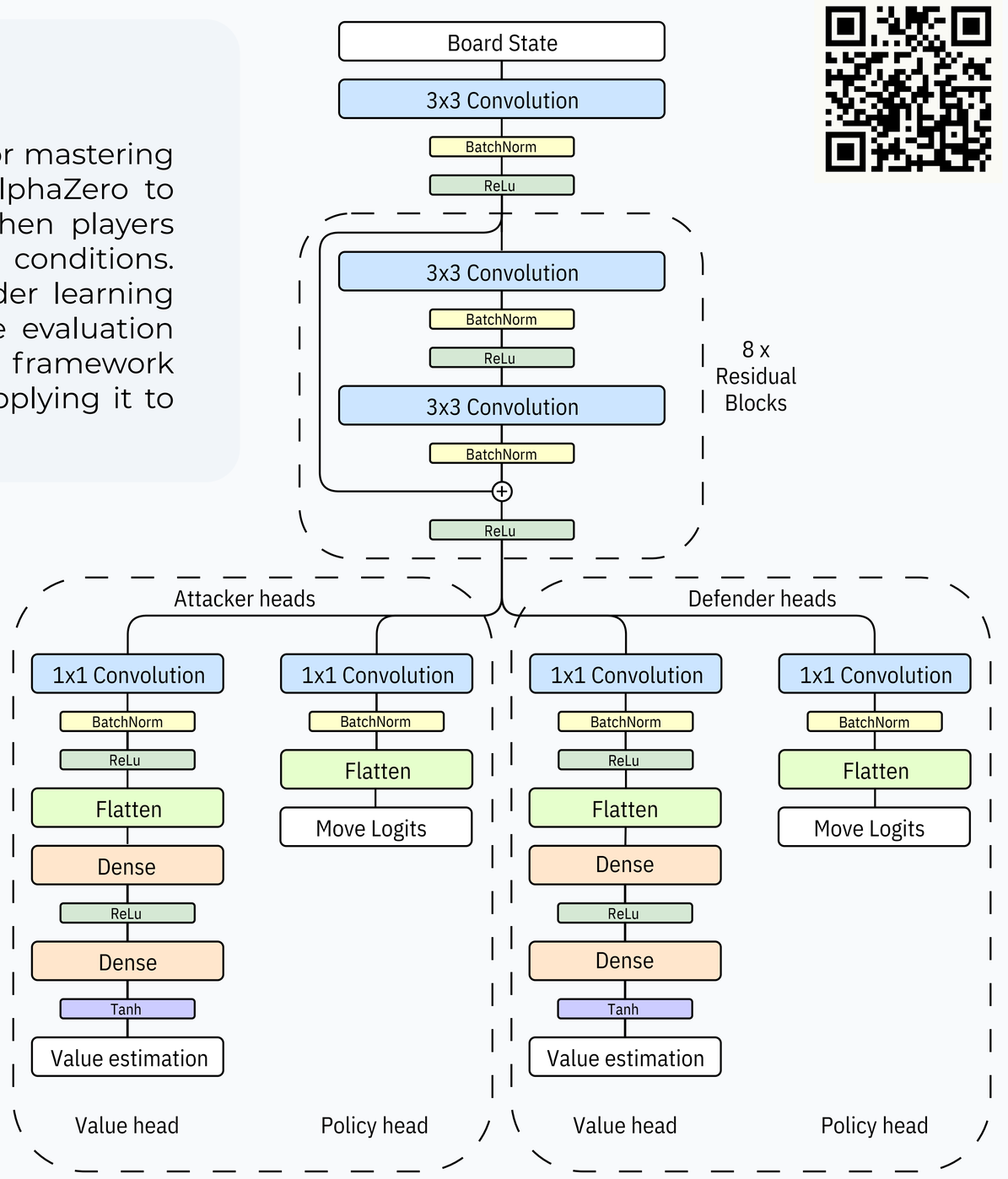
Reproducing AlphaZero on Tablut: Self-Play RL for an Asymmetric Board Game

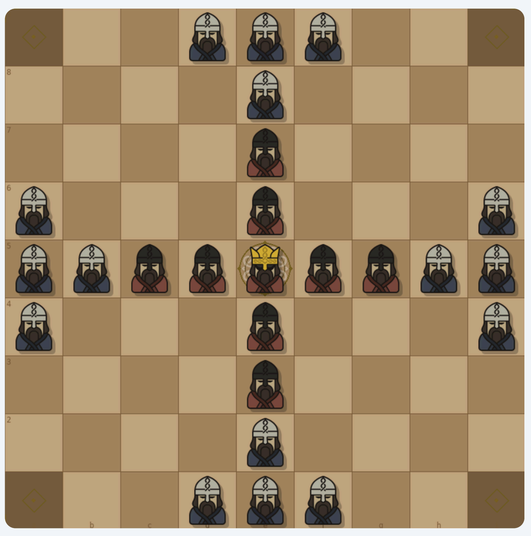
Author: Tõnis Lees
Supervisor: Tambet Matiisen
Computer Science BSc

UNIVERSITY OF TARTU
Institute of Computer Science

INTRODUCTION

AlphaZero is a general reinforcement learning algorithm renowned for mastering zero-sum symmetric games like Chess, Go, and Shogi. Applying AlphaZero to asymmetric games, however, presents a unique challenge, since then players have differing piece counts, opposing objectives, and distinct win conditions. Using a single neural network head for both perspectives can hinder learning efficiency and performance due to the need to learn two separate evaluation functions. This project investigates whether the AlphaZero self-play framework can be successfully transferred to an asymmetric environment by applying it to the historical board game Tablut.





TABLUT

Tablut is a historic Northern European board game played on a 9x9 grid. The attacker fields 16 pieces with the objective of capturing the opposing king. The defender has 8 pieces and a king, attempting to escort the king safely to any of the board's four corners. All pieces move orthogonally like rooks in chess, and captures are executed by "sandwiching" an enemy piece between two opposing pieces.

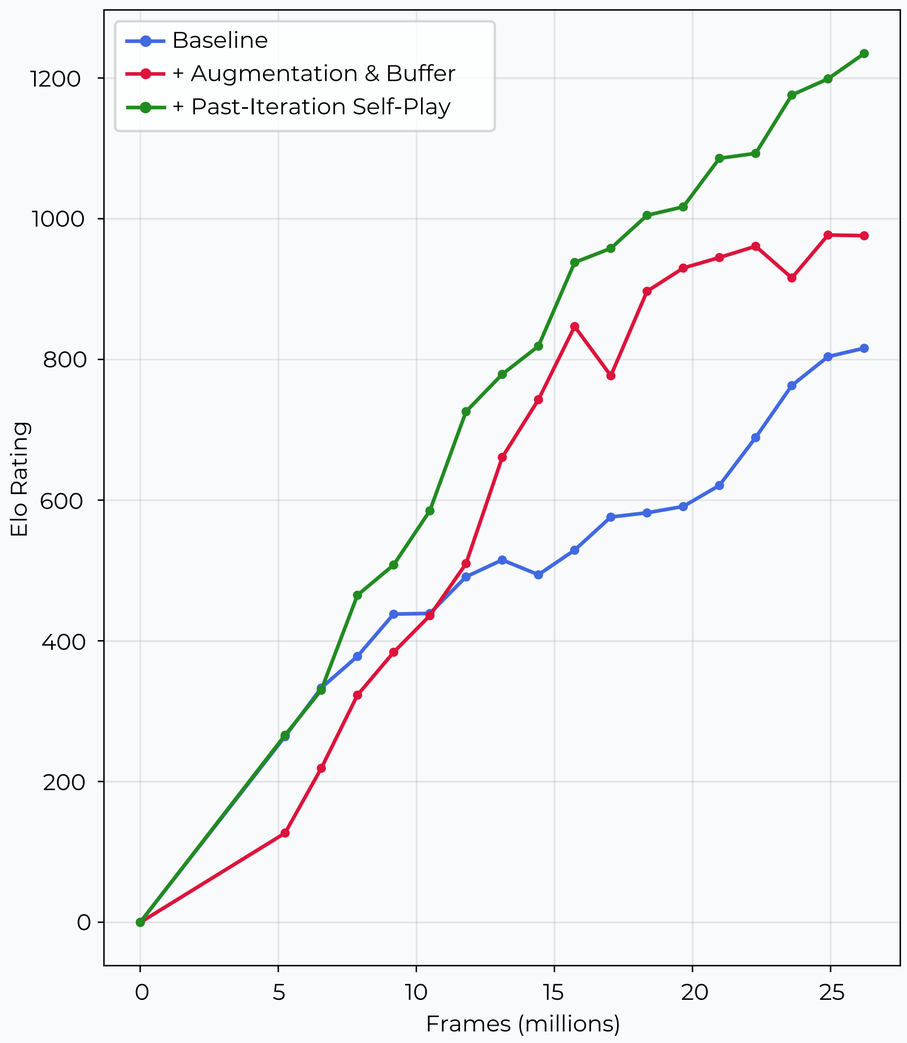
OVERCOMING INSTABILITY

The asymmetric nature of Tablut introduced severe training instabilities, most notably "catastrophic forgetting" between the alternating roles during self-play. To stabilize learning, three cumulative mitigation strategies were implemented:

- C_4 Data Augmentation (random 90-degree board rotations)
- Expanded Replay Buffer: The replay buffer capacity was doubled from 8 to 16 self-play iterations, storing roughly 4.2 million states.
- Past-Iteration Self-Play: 25% of training games were played against randomly sampled past checkpoints

METHODOLOGY

- MCTS & Gumbel MuZero: Limits massive search spaces by utilizing 128 simulations per move, focusing compute solely on promising trajectories.
- Dual-Head Architecture: Replaces the standard symmetric single-head network with separate, distinct policy and value heads for the Attacker and Defender roles.
- Shared Trunk: An 8-block residual trunk with 128 filters extracts underlying board features common to both perspectives.
- Framework: Fully vectorized, hardware-accelerated self-play environment built using the JAX ecosystem.



RESULTS

The model was trained for 100 self-play iterations, totaling 26 million frames, reaching a BayesElo rating of 1235 relative to its baseline. Play style became rapidly decisive: policy entropy fell from 3.05 to 1.47, and the average number of remaining pieces dropped from 22 to 15. Interestingly, the learning curves diverged after iteration 75. While both sides initially maintained a 70–80% win rate against past checkpoints, by iteration 100 the attacker's win rate surged to 86% while the defender's dropped to 52%. This divergence indicates that Tablut's ruleset heavily favors the attacker, or that defender strategies are inherently harder for the network to master.