Transformers for Reinforcement Learning in Strategy Games

PIC2 - Master in Computer Science and Engineering
Instituto Superior Técnico, Universidade de Lisboa

Advisors: Pedro A. Santos, João Dias

João Santos

February 5, 2026

Introduction - Motivation

- **Reinforcement Learning Breakthroughs:** Alphazero demonstrates the power of combining deep RL with search, especially in perfect-information games.
- **Complex Representations:** Wargames involve stacked units, spatial reasoning, and partial observability, making them an ideal testbed for advanced RL models.
- **Generalization:** Learning-based approaches can potentially adapt across different maps, scenarios, and even other strategy games.

Introduction - Objectives

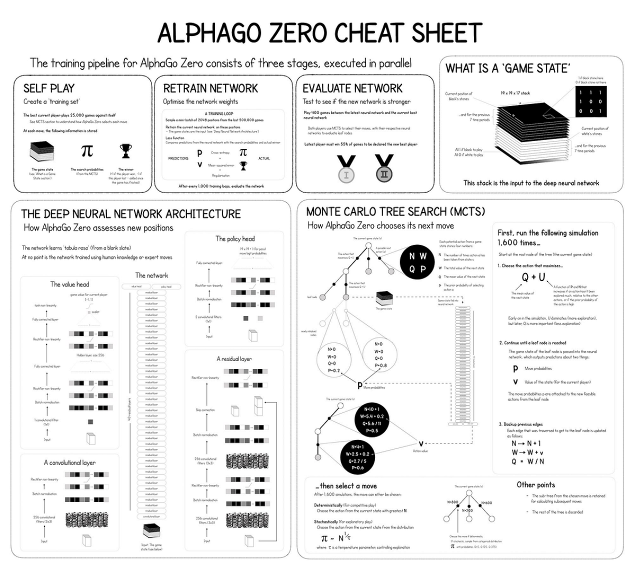
- **Main Objective:** Develop an RL agent capable of learning strategies in the wargame *Hispania*.
- **Specific Goals:**
  - Implement the proposed Transformer-based architecture.
  - Train and optimize the model through self-play, on small game scenarios to later extrapolate to real game states.
  - Evaluate, study and assess the AI's performance, analyzing its outcomes compared to the benchamrks established.

Introduction - Problem

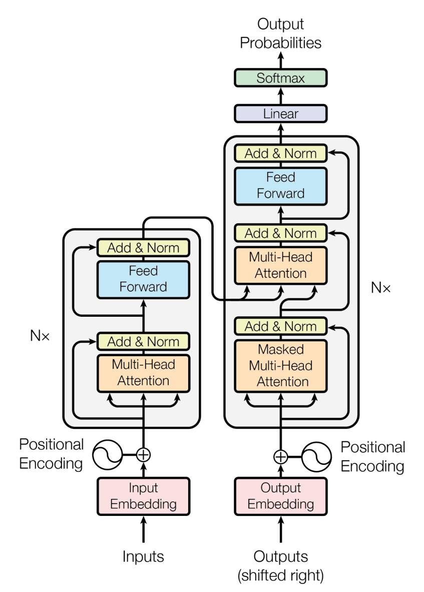
- **State Space Complexity:** Modern strategy games exhibit extremely large and structured state spaces, making learning and generalization challenging for reinforcement learning agents.
- **Imperfect Information and Stochasticity:** Actions may lead to different outcomes due to probabilistic events, such as dice rolls, introducing uncertainty that is absent in deterministic, perfect-information games.
- **Unit Stacking:** Multiple units occupying the same region introduce interactions between units and regions that few existing models are explicitly designed to represent.

Background - Hispania

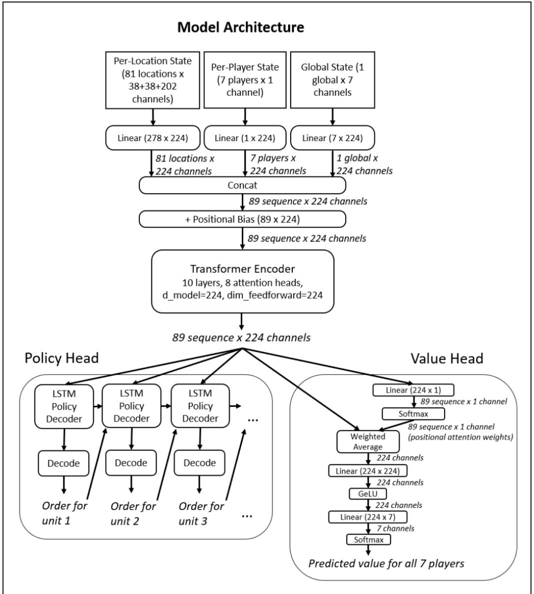
Background - AlphaZero



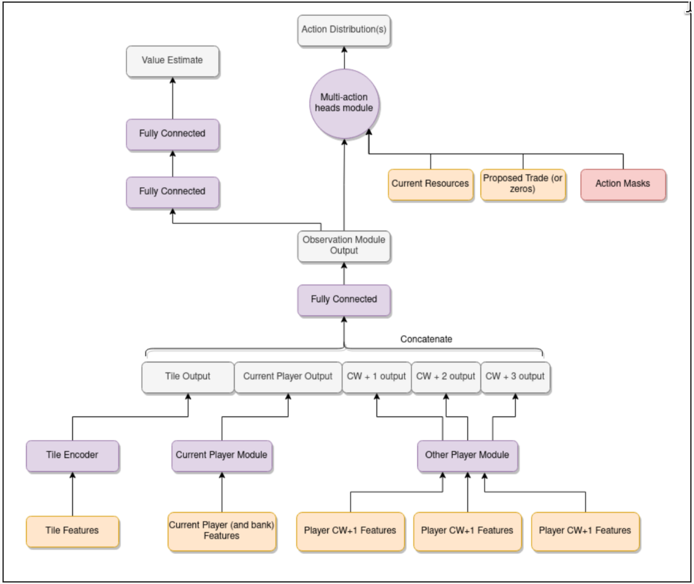
Background - Transformer



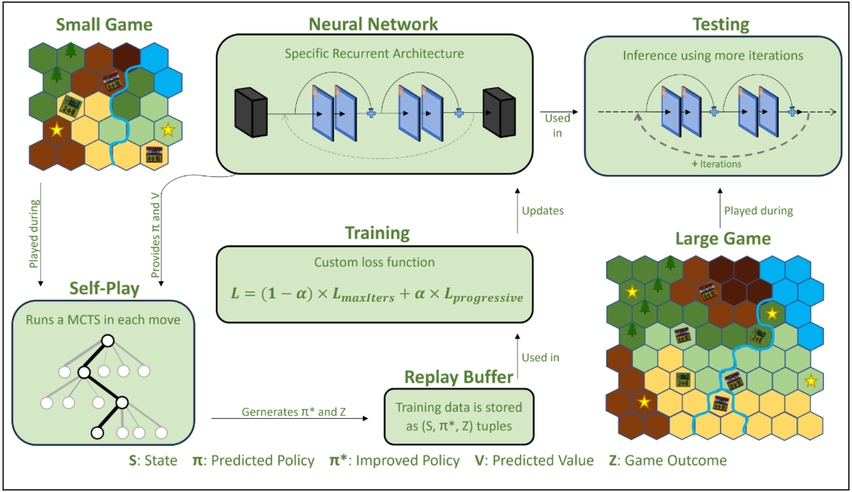
Related Work - Diplodocus (No-Press Diplomacy)



Related Work - Settlers of Catan



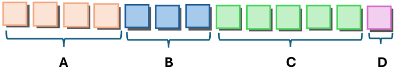
Related Work - Hex-and-Counter Wargames



Solution Proposal

- **Tile Encoder:** Encodes 55 regions using learned positional embeddings and adjacency bias.
- **Piece Encoder:** Encodes each unit with reference to their respective map location.
- **Game-State Encoder:** Fuses tile and unit information into a latent state tensor.

Solution Proposal - Tile Feature Vector

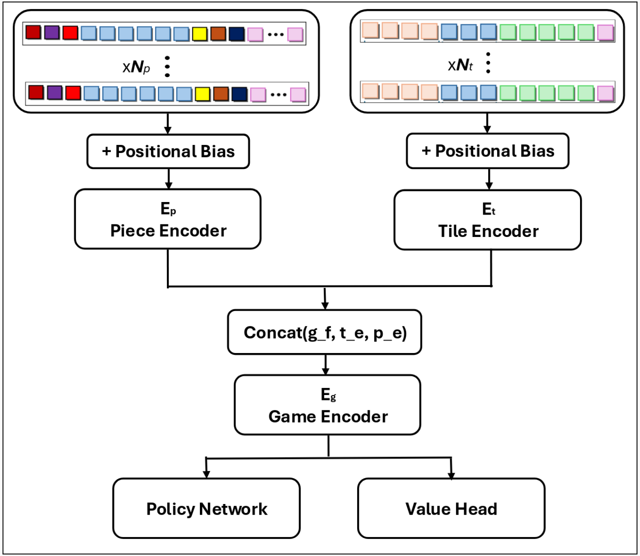


A. Presence of a building: each entry corresponds to a building type (Castle, Fort, City, None).
B. Terrain type: each entry corresponds to a terrain type of the province (clear, difficult, impassable)
C. Region section: each entry zone the province is a part of (North, West, East, South, Off-map)
D. Population points: the population points accumulated by that province.

Solution Proposal - Piece Feature Vector

A. Health
B. Damage
C. Defense
D. Minimum roll required to be killed
E. Movement Points
F. Purchase price
G. Owning nation

Solution Proposal - Architecture



Solution Proposal - Policy and Value Heads

- **Policy Network:**
  - Conditions each prediction on the shared game-state embedding and previous actions using a transformer decoder.
  - Each action is composed with: action type, source region, unit selection, target region, and combat order.
  - Predicts each component of an action sequentially, via a respective action type head, tile head, unit head and battle head.
  - Constrained by legal action masks to ensure only valid moves.
- **Value Head:**
  - Predicts the expected game outcome from the current state.
  - Used to guide learning and stabilize training during self-play.

Solution Proposal - Training

- **Learning:** Training starts on simplified maps and progressively scales to full game scenarios.
- **Self-Play:** The agent learns exclusively from self-play, iteratively updating the policy and value heads from game outcomes.
- **Iterative Evaluation:** Performance is periodically assessed on larger maps to measure generalization and learning stability.

Solution Proposal - Evaluation

- **Baseline Validation:**
    - Evaluation against a random decision agent to verify correct rule learning and reward propagation.
- **Self-Play Progression:**
    - Periodic evaluation against earlier versions of the model to assess learning stability and strategic improvement.
- **Final Benchmark:**
    - Direct comparison against the existing heuristic-based Hispania AI in full scale games.

Work Schedule