Spatial Navigation Using Bio-Plausible Learning Rules

Mattia Assane Wade

September 30, 2025

Contents

Abstract

This work investigates navigation in maze environments using reinforcement learning (RL) agents guided by biologically plausible representation learning. While standard deep RL approaches rely on backpropagation and non-biological training schemes, recent studies suggest that three-factor learning rules can approximate plasticity observed in real neurons. However, these rules are limited when applied directly to single-layer networks and struggle with complex high-dimensional tasks. To address this, we evaluate the use of deep encoders pretrained with bio-plausible rules, such as Contrastive Learning with Associative Predictive Plasticity (CLAPP), as a means of providing structured representations for downstream RL. We implement and compare actor-critic (A2C) and proximal policy optimization (PPO) methods in custom T-maze and Four-Room environments, analyzing their convergence, stability, and sensitivity to architectural modifications. Furthermore, we explore enhancements through decorrelation layers, memory-based intrinsic motivation, and curiosity-driven mechanisms. Our findings demonstrate that biologically inspired representation learning can improve navigation performance over raw pixel inputs, though challenges remain in matching the efficiency of conventional multi-layer agents. This research highlights both the promise and limitations of bio-plausible approaches for scalable RL navigation tasks.

1 Introduction

Reinforcement learning has achieved remarkable success in solving complex control and navigation problems, yet many state-of-the-art methods rely on architectures and optimization techniques that diverge significantly from biological learning. In contrast, neuroscience evidence points to three-factor learning rules that couple synaptic plasticity with neuromodulatory signals, suggesting a natural correspondence between neural learning and RL. However, applying these rules directly in artificial agents is difficult, as single-layer formulations lack the representational power required for visual and spatial reasoning.

This thesis addresses the gap between biologically plausible learning and practical navigation performance. We investigate whether deep encoders trained with bio-inspired objectives can provide structured state representations suitable for lightweight RL algorithms. Specifically, we build upon CLAPP, a plasticity-driven contrastive learning method, to generate features that capture spatial regularities without backpropagation. These features are then used by reinforcement learning agents, implemented through A2C and PPO, to solve navigation tasks in challenging visual mazes.

2 Literature Review and Basics

2.1 Reinforcement Learning (RL) Foundations

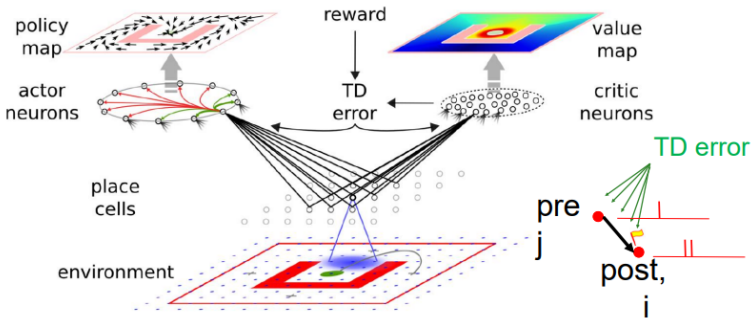


Figure 1: Illustration of how an actor-critic algorithm works.

Actor-Critic with Eligibility traces

Actor–Critic with Eligibility Traces (continuing), for estimating $\pi_\theta \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: $\lambda^{\mathbf{w}} \in [0,1]$, $\lambda^{\boldsymbol{\theta}} \in [0,1]$, $\alpha^{\mathbf{w}} > 0$, $\alpha^{\boldsymbol{\theta}} > 0$

Initialize state-value weights $\mathbf{w} \in \mathbb{R}^d$ and policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)
Initialize $S \in \mathcal{S}$ (e.g., to s_0)

$\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ (d-component eligibility trace vector)
$\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ (d'-component eligibility trace vector)
Loop forever (for each time step):
  $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
  Take action A, observe S', r
  $\delta \leftarrow \quad r + \gamma\, \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$

  $\mathbf{z}^{\mathbf{w}} \leftarrow \lambda^{\mathbf{w}} \mathbf{z}^{\mathbf{w}} + \nabla \hat{v}(S,\mathbf{w})$
  $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda^{\boldsymbol{\theta}} \mathbf{z}^{\boldsymbol{\theta}} + \nabla \ln \pi(A|S,\boldsymbol{\theta})$
  $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\, \delta\, \mathbf{z}^{\mathbf{w}}$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\, \delta\, \mathbf{z}^{\boldsymbol{\theta}}$
  $S \leftarrow S'$

Adapted from Sutton and Barto

Figure 2: Actor-Critic algorithm (adapted from Wulfram Gerstner's course).

2.2 Biologically Plausible Learning

Reinforcement learning provides a framework in which agents learn to act in an environment by maximizing cumulative rewards. Formally, an RL problem is modeled as a Markov Decision Process (MDP), defined by a state space, an action space, transition dynamics, and a reward function.

Modern RL has achieved strong results in high-dimensional control tasks, including Atari games and robotic locomotion. Among policy gradient methods, two families are central to this thesis:

- **Actor-Critic (A2C):** Combines a policy network (actor) with a value estimator (critic). While the actor improves the policy based on gradients of expected return, the critic reduces variance by approximating the value function. A2C updates online at each step, making it computationally lightweight but sensitive to critic quality.
- **Proximal Policy Optimization (PPO):** An extension of policy gradient methods that stabilizes updates by constraining the size of policy changes. PPO typically achieves faster convergence and better stability, especially when paired with Generalized Advantage Estimation (GAE). However, it is less biologically plausible, as it relies on batch updates and non-local optimization.

RL's effectiveness comes from its ability to scale with deep neural networks, but these implementations diverge from learning rules observed in biological systems.

2.3 Biologically Plausible Learning

Neuroscience experiments have shown that synaptic plasticity in the brain is often driven by **three-factor learning rules**, in which weight updates depend on:

1. Pre-synaptic activity,
2. Post-synaptic activity, and
3. A global modulatory signal (e.g., dopamine).

This mechanism maps naturally onto RL: the modulatory signal resembles a reward prediction error, shaping learning in a manner similar to policy gradients. However, traditional RL implementations rely heavily on **backpropagation**, which is not biologically feasible due to its requirement for symmetric weight transport and non-local computations.

Research into **bio-plausible alternatives** has proposed local learning rules that approximate gradient descent. Methods such as feedback alignment, predictive coding, and Hebbian-like rules attempt to reconcile neuroscience with machine learning. Still, scaling these methods to high-dimensional tasks remains challenging.

2.4 Representation Learning for RL

High-dimensional inputs, such as visual observations, make direct learning from pixels difficult for shallow networks. Effective RL therefore depends on good **representation learning**, where raw sensory data is transformed into compact, informative features.

In machine learning, self-supervised learning methods such as contrastive learning, autoencoding, and predictive modeling have advanced representation learning significantly. For bio-plausibility, contrastive Hebbian approaches are especially relevant.

Contrastive Learning with Associative Predictive Plasticity (CLAPP) is a recent biologically inspired method that trains representations without backpropagation. Instead, it relies on local Hebbian-like plasticity modulated by predictive coding signals. Encoders trained with CLAPP can capture visual and spatial structure in a way that aligns with biological principles, making them suitable candidates for downstream RL tasks.

2.5 Navigation in AI and Neuroscience

Navigation has long been a central problem in both artificial intelligence and neuroscience.

- **In AI**, maze-based tasks such as the T-Maze and Four-Rooms environments serve as standard benchmarks for evaluating spatial reasoning. Success requires the agent to integrate visual inputs, build internal state representations, and plan long-term actions.
- **In neuroscience**, the hippocampus and entorhinal cortex provide well-studied models of spatial representation. Place cells activate for specific spatial locations, while grid cells fire in periodic spatial patterns, together forming a cognitive map of the environment. These biological mechanisms inspire attempts to design artificial systems that learn **place-cell-like** or **grid-like** representations.

The parallels between neuroscience and AI suggest that bio-plausible representation learning could support navigation in artificial agents by providing structured spatial encodings analogous to hippocampal activity.

2.6 Intrinsic Motivation and Exploration

A key challenge in navigation is efficient exploration. Standard RL agents often fail to explore sufficiently in environments with sparse rewards. To address this, **intrinsic motivation mechanisms** have been introduced:

- **Curiosity-Driven Learning:** Encourages exploration by rewarding agents when they encounter novel or unpredictable states. The **Intrinsic Curiosity Module (ICM)**, for instance, provides rewards based on prediction errors between forward models of state transitions.

- **Novelty-Based Rewards:** Alternative approaches reward agents for encountering states dissimilar to previously visited ones, often computed using memory buffers or feature decorrelation.

These mechanisms are not only useful for RL but also align with biological principles, where novelty and surprise are key drivers of exploration in animals.

While conventional RL methods are highly effective, they rely on optimization techniques far removed from biological learning. Conversely, biologically plausible learning rules capture neural mechanisms but struggle to scale to complex tasks.

Bridging this gap requires **integrating bio-inspired representation learning with reinforcement learning**, enabling agents to learn structured, navigationally relevant features without backpropagation. Previous work has not fully explored how methods such as CLAPP can enhance RL agents in maze-based navigation tasks, nor how additional mechanisms (decorrelation, memory, curiosity) can extend their capabilities.

3 Methodology

3.1 Overall Framework

The objective of this work is to examine whether biologically plausible representation learning can support reinforcement learning in navigation tasks. The methodology integrates three main elements: carefully designed environments, reinforcement learning algorithms that serve as baselines, and biologically inspired extensions that address limitations of standard approaches. A dedicated code base was implemented and versioned to allow reproducibility and modular experimentation.

The implementation was carried out in Python, with models trained on GPU-enabled hardware. The full code base is available at:

https://github.com/Aurelien-Cormoreche/-RL_Using_CLAPP.

3.2 CLAPP Encoder for Representation Learning

A central component of this project is the use of a biologically plausible encoder trained with **Contrastive Learning with Associative Predictive Plasticity (CLAPP)**. The encoder provides structured state representations that serve as input to the reinforcement learning agents, replacing direct raw pixel observations.

CLAPP is motivated by the observation that biological neurons adapt through local plasticity rules modulated by global signals, rather than through backpropagation. It implements a contrastive learning framework that encourages similar representations for temporally or spatially related inputs, while pushing apart representations of unrelated ones. Unlike standard deep learning methods, CLAPP does not rely on weight symmetry or non-local gradient computations, making it more aligned with known mechanisms of synaptic plasticity.

In practice, the CLAPP encoder was pretrained on visual inputs before being coupled with the RL algorithms. The walls of the T-Maze and Four-Rooms environments were textured with STL-10 images that had not been part of the encoder's training data. This ensured that the agent encountered novel visual stimuli, testing the encoder's ability to generalize representations. During reinforcement learning, the encoder's weights were frozen, and only the downstream agent (A2C or PPO) was trained. This setup isolated the effect of bio-plausible representations on navigation performance.

The encoder outputs vectors of dimension 128, which were sufficient to capture rich visual and spatial features.

$$\Delta W_{ji} \quad \propto \quad \underbrace{\text{modulators}}_{\text{broadcast factors}} \cdot \underbrace{(\boldsymbol{W}^{\text{pred}}\boldsymbol{c}^{t_1})_j}_{\text{dendritic prediction}} \cdot \underbrace{\text{post}_j^{t_2} \cdot \text{pre}_i^{t_2}}_{\text{local-activity}} .$$

Figure 3: Learning rule of CLAPP -" *Wulfram Gerstner et al. - 2021*"

3.3 Environments

Two navigation environments were developed in order to evaluate the agents' ability to learn spatial representations and policies. The first environment is a custom T-Maze. Here, the agent is placed in a corridor that branches into two paths, with the starting location and orientation randomized at the beginning of each run. The walls of the maze are covered with images drawn from the STL-10 dataset, ensuring that the encoder cannot exploit memorized patterns. Random spawning and perceptual variation create an effectively infinite state space, forcing the agent to rely on robust representations rather than memorization.

The second environment is the Four-Rooms maze, which introduces additional complexity by requiring the agent to traverse longer corridors and move between interconnected rooms. Compared with the T-Maze, this setting demands more strategic exploration and longer-term decision-making. To accommodate the increased difficulty, the maximum number of steps per episode was increased from 1000 to 1500. Together, the two environments provide complementary challenges for testing the scalability of biologically inspired methods.

Figure 4: View of the agent at the start of the T-Maze

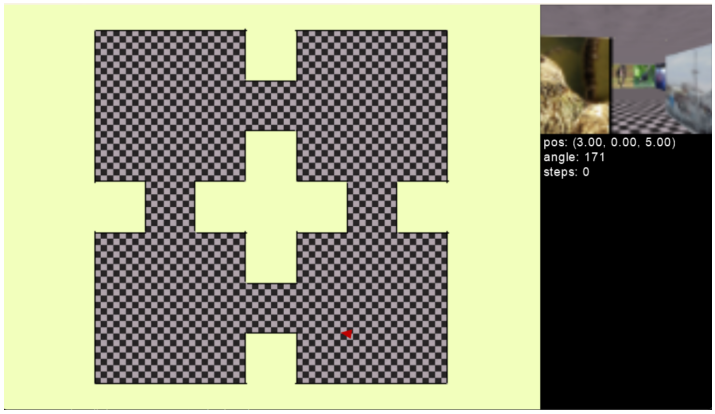


Figure 5: Top view of the Four Rooms Maze

3.4 Reinforcement Learning Algorithms

The first reinforcement learning algorithm implemented was the Advantage Actor-Critic (A2C) method. In this approach, the actor learns a policy while the critic estimates the value of states. Both networks are updated online at each step, and eligibility traces can be incorporated to capture temporal dependencies. Early experiments with A2C revealed several challenges. The policy often collapsed when the critic produced inaccurate value estimates, which caused advantages to converge near zero. Learning was further destabilized by the use of high learning rates, the absence of gradient clipping, and the lack of sufficient exploration. To mitigate these problems, an entropy term was introduced into the loss function to encourage exploration. The actor and critic were also trained with different learning rates, gradient clipping was applied, and the eligibility trace parameter was tuned. After these modifications, A2C was able to learn near-optimal navigation policies in the T-Maze and generalized reasonably well to the Four-Rooms environment.

Proximal Policy Optimization (PPO) was implemented as a second baseline. PPO was chosen because of its strong performance in benchmark reinforcement learning tasks and its more stable training dynamics. The algorithm constrains policy updates, preventing drastic changes that destabilize learning. In practice, PPO achieved faster convergence than A2C and benefited from the use of Generalized

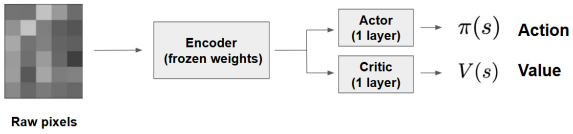


Figure 6: Model of our A2C pipeline

Advantage Estimation. However, its reliance on batch updates makes it less compatible with biological plausibility. In fact, when PPO was forced to update online at every step, it collapsed entirely, confirming that its strength derives from mechanisms that do not resemble neural learning.

3.5 Addressing Critic Limitations

A central issue observed during experimentation was the weakness of the critic in A2C. The critic was implemented as a single linear layer, which struggled to assign accurate values when the encoded features were highly correlated. Similarity analysis of the representations confirmed that many states were mapped to vectors with high cosine similarity, effectively collapsing the state space. This resulted in unstable value estimation and extremely slow convergence. To address this, the project investigated ways of reducing feature correlation, drawing inspiration from biological systems where hippocampal place cells and entorhinal grid cells form highly diverse spatial codes.

3.6 Biologically Inspired Extensions

To overcome the critic's limitations and increase the biological plausibility of the agents, several extensions were introduced.

The first was a decorrelation layer inserted between the encoder and the reinforcement learning agent. This layer was trained contrastively so that representations corresponding to temporally close states were encouraged to be similar, while those corresponding to temporally distant states were pushed apart. In a preliminary step, the linear decodability of spatial positions was tested by discretizing the maze and training a one-layer classifier. The classifier achieved 97% accuracy, demonstrating that spatial information was in principle present in the representations. However, when a two-layer actor-critic was trained directly, no strong clustering reminiscent of place-cell activity emerged. The full decorrelation layer was therefore implemented and trained with a cascade memory mechanism.

Cascade memory provided a more structured approach. The agent was allowed to explore the maze with an intrinsic reward that encouraged the discovery of new rooms. During this phase, representations were collected and trained with an InfoNCE contrastive loss, which encouraged the system to form distinct spatial encodings. After training, the weights of the decorrelation layer were frozen, and the enriched representations were concatenated with the original encoder features. This method produced modest improvements compared to raw inputs, though it remained below the performance of PPO or deeper conventional agents.

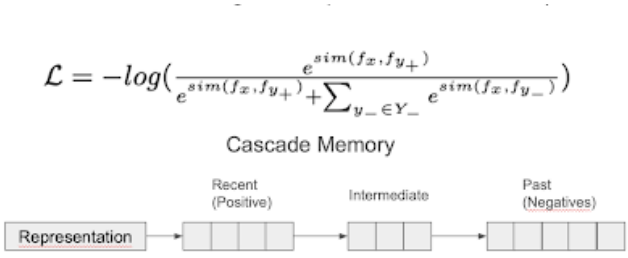


Figure 7: cascade

A second extension involved a memory buffer and comparator system. The buffer stored previously encountered representations, provided they were sufficiently different from those already in memory. New states were compared against this buffer, and the agent received novelty rewards when the representation was dissimilar from its stored history. This mechanism served as a biologically inspired analogue of novelty detection in the brain. In practice, the memory buffer accelerated exploration in PPO, though it introduced computational overhead in larger environments.

A third extension implemented an Intrinsic Curiosity Module (ICM). The idea was to provide intrinsic rewards proportional to the prediction error of a forward model that estimated the next state based on the current state and action. A simplified version was tested first, focusing only on the forward model. However, this implementation did not yield improvements. The prediction task proved too difficult for the shallow networks available, as they were required to predict high-dimensional vectors

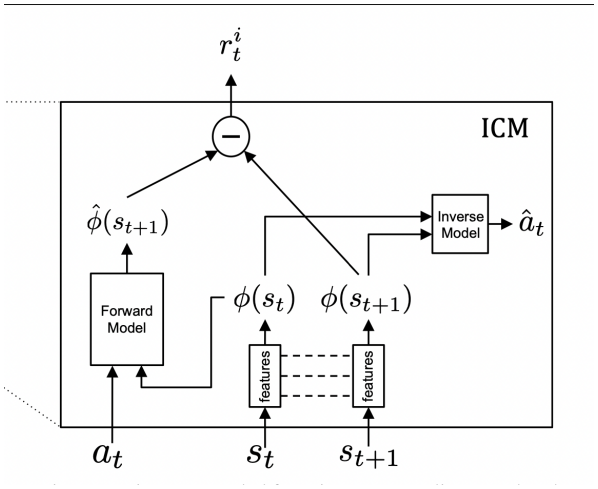
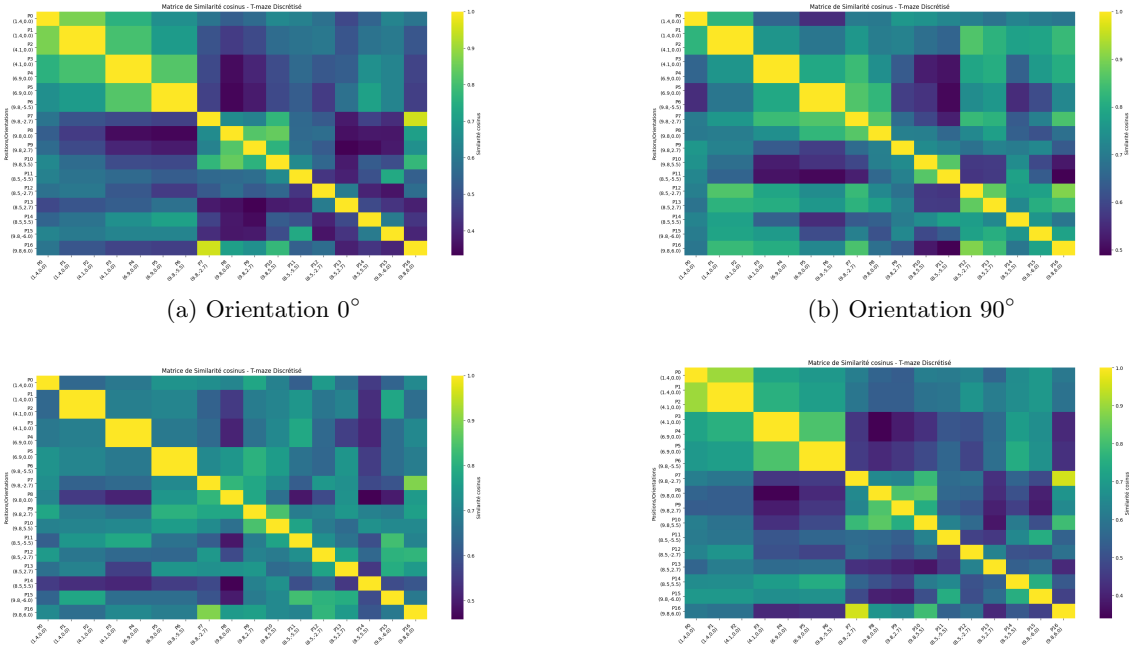


Figure 8: icm

from nearly equally large inputs. As a result, the intrinsic reward was largely uninformative and sometimes even degraded learning speed.

3.7 Spatial Representation Analysis

To evaluate whether the representations captured spatial structure, similarity matrices were computed across positions and orientations within the environments. These analyses revealed that encoded features often displayed high similarity, regardless of position. This explained the difficulty faced by the linear critic and underscored the importance of decorrelation mechanisms. While the additional layers and memory systems improved differentiation to some extent, the representations remained less structured than biological place-cell or grid-cell activity.



(a) Orientation 0°

(b) Orientation 90°

(c) Orientation 180°

(d) Orientation 270°

Figure 9: Cosine similarity matrices for different orientations in the T-Maze environment.

4 Experiments and Results

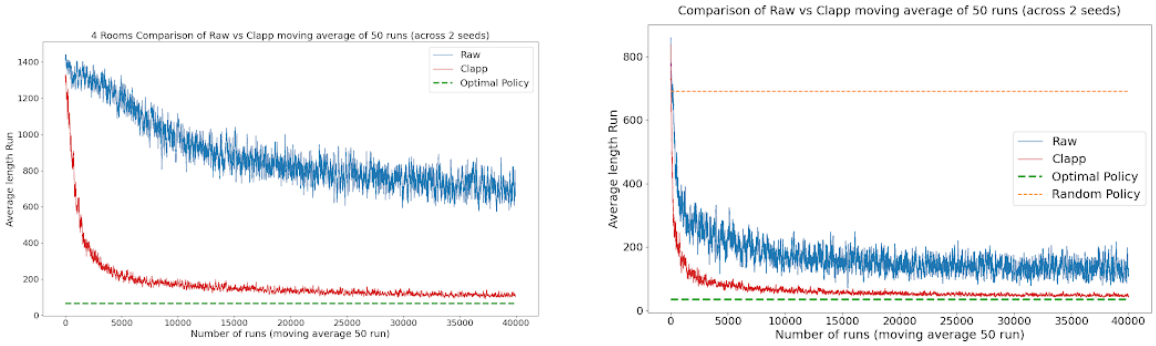
This chapter presents the experimental results obtained in the T-Maze and Four-Rooms environments using A2C and PPO as baseline reinforcement learning algorithms. It also examines the effects of biologically inspired extensions introduced during the project, such as the decorrelation layer, cascade memory, and intrinsic motivation modules. The analysis addresses both the convergence of policies and the quality of the learned spatial representations.

4.1 Results with A2C

The first series of experiments involved training an A2C agent in the T-Maze environment. Initial results showed partial convergence followed by policy collapse. Several causes were identified: inaccurate value estimates from the critic, excessively high learning rates, the absence of entropy regularization, and insufficient exploration. Under these conditions, the estimated advantage quickly converged to zero, effectively neutralizing the actor's updates.

After introducing a series of fixes, the performance improved substantially. These modifications included the addition of an entropy bonus in the loss function, separate learning rates for actor and critic, gradient clipping, and the tuning of eligibility traces. Using grayscale inputs, a frame skip of three, and controlled weight initialization, A2C was able to learn near-optimal navigation policies in the T-Maze. The trajectories exhibited robustness and demonstrated that the agent had developed a generalizable navigation strategy.

The same adjustments were applied to the Four-Rooms environment, which is more complex due to longer corridors and multiple decision points. Despite the increased difficulty, A2C generalized well and successfully learned effective policies. In contrast, learning directly from raw pixel inputs performed poorly, confirming the importance of representation learning for visual navigation. A supplementary test fixed the agent's initial orientation. In this restricted setting, the state space became finite, and even raw pixels sufficed to learn nearly optimal policies. This experiment highlighted the critical role of invariance and generalization when dealing with environments of higher variability.



(a) Four Rooms with Actor-Critic

(b) T-Maze with Actor-Critic

Figure 10: Agent trajectories in different environments using the Actor-Critic algorithm.

4.2 Results with PPO

PPO was evaluated as a comparative baseline. With conservative hyperparameters, it displayed stable but relatively slow convergence. When tested with a single environment and forced to update online at each step, PPO collapsed entirely. This behavior was expected, as PPO relies heavily on Generalized Advantage Estimation (GAE), which requires access to full episodes for stable value estimation.

The comparison between A2C and PPO revealed key differences. PPO converged more rapidly under conventional training conditions, but its dependence on batch updates makes it less suitable in a biologically plausible setting. A2C, by contrast, though slower, maintained viability under conditions more aligned with neural plasticity rules.

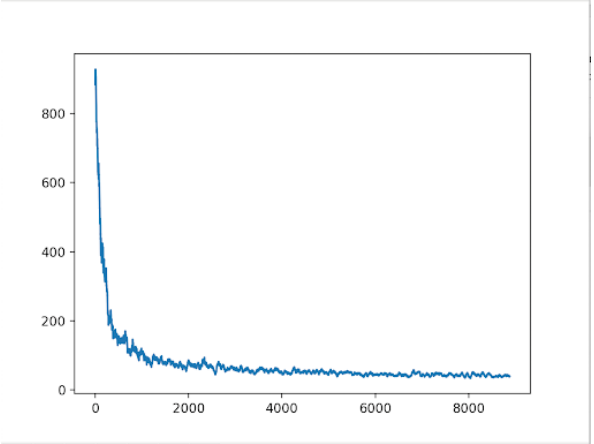


Figure 11: PPO in T-Maze

4.3 Critic Analysis and Convergence Issues

An inspection of the critic values in A2C revealed highly inaccurate estimations. Even in frequently visited position–direction pairs, the critic produced inconsistent predictions. This weakness explains the slow convergence and instability, at times reducing the algorithm to behavior resembling a basic REINFORCE strategy.

Further analysis showed that the root cause was the high similarity between encoded features. The critic, restricted to a single linear layer, was unable to distinguish between such correlated vectors, leading to poor value learning. This finding motivated the development of decorrelation mechanisms to diversify the representations.

4.4 Decorrelation Layer and Cascade Memory

The first step toward decorrelation was to test the linear decodability of positions in the T-Maze. By discretizing the maze into 32 cells and training a one-layer classifier, an accuracy of 97% was achieved, confirming that spatial information was present in the encoded features.

However, training a two-layer actor-critic did not reveal the spontaneous emergence of place-cell-like clustering. To address this, a decorrelation layer was implemented, trained with a cascade memory mechanism using an InfoNCE loss. In this setup, the agent explored the maze with intrinsic rewards for discovering new rooms. The collected representations were then used to train the decorrelation layer, whose weights were subsequently frozen and concatenated with the encoder's features.

The results showed that the decorrelation layer improved performance relative to raw inputs, but remained below the level achieved by PPO or deeper networks. Nevertheless, the approach produced more structured representations and moved the agent's learning dynamics closer to those observed in biological systems.

4.5 Intrinsic Motivation and Exploration

Two bio-inspired mechanisms were introduced to encourage exploration. The first was a memory buffer and comparator system. Representations were added to the buffer if their cosine similarity to stored vectors was below a threshold. The agent received novelty rewards when encountering states sufficiently different from those in memory. This mechanism accelerated exploration in PPO with a single environment but introduced computational overhead in larger environments.

The second mechanism was the Intrinsic Curiosity Module (ICM). A simplified implementation included only the forward model, which attempted to predict the next latent state from the current state and action. The results showed no improvement, and in some cases, convergence was slower. The likely explanation is that predicting high-dimensional representations with a shallow network is too difficult, rendering the intrinsic reward ineffective and noisy.

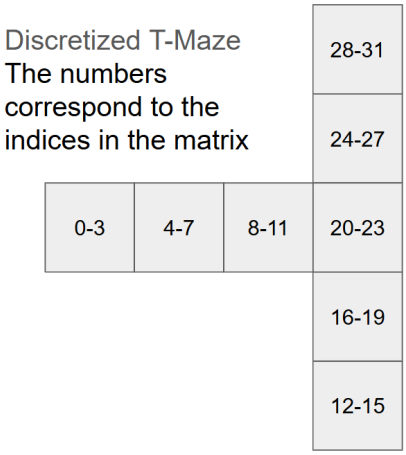


Figure 12: Positions of the discretized T-Maze

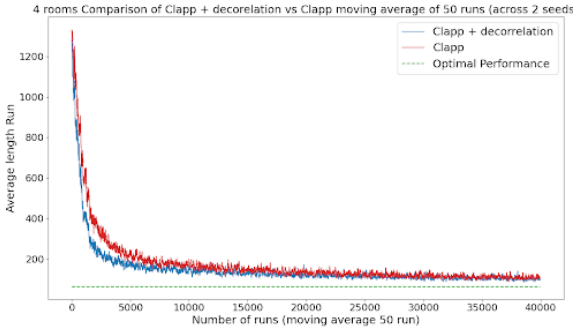


Figure 13: Enter Caption

4.6 Spatial Representation Analysis

The quality of spatial representations was further analyzed using similarity matrices based on cosine similarity across positions and orientations. The results revealed that many states were mapped to highly correlated vectors, confirming the critic's difficulties. Although the decorrelation layer and cascade memory partially reduced these correlations, the emergence of structured place-cell- or grid-cell-like activity remained limited.

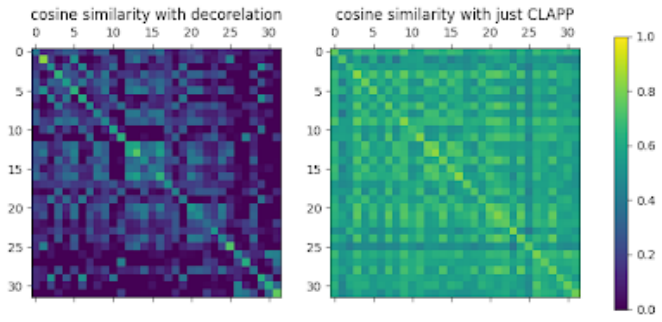


Figure 14: Comparison between the cosine similarities between encoded features in the discretized T-maze before and after the decorelation layer

5 Conclusion and Future Work

This thesis has explored the integration of biologically plausible representation learning into reinforcement learning for navigation tasks. The motivation stemmed from the gap between state-of-the-art deep reinforcement learning methods, which rely heavily on backpropagation and global optimization, and biological learning mechanisms, which are governed by local plasticity rules and neuromodulatory signals. By leveraging encoders trained with CLAPP, and by extending reinforcement learning agents with mechanisms inspired by place cells, novelty detection, and intrinsic motivation, this work has sought to bridge these two domains.

The experiments in the T-Maze and Four-Rooms environments demonstrated both the promise and the limitations of this approach. With appropriate stabilization techniques, A2C achieved near-optimal policies, confirming that lightweight actor-critic algorithms can succeed when supported by meaningful representations. PPO, while more efficient in conventional settings, collapsed under online training conditions, highlighting its incompatibility with biological plausibility. The use of CLAPP encodings consistently improved performance over raw pixel inputs, providing evidence that bio-inspired self-supervised learning can generate useful representations for reinforcement learning.

The exploration of biologically motivated extensions yielded mixed results. The decorrelation layer and cascade memory improved the diversity of learned features and partially addressed the weaknesses of the critic, though they did not yet match the efficiency of deeper artificial networks. The novelty buffer accelerated exploration in PPO but introduced computational trade-offs, while the intrinsic curiosity module failed to provide significant gains under shallow architectures. Together, these findings suggest that bio-plausible mechanisms can complement reinforcement learning but require further refinement to achieve competitive scalability.

Several contributions of this work stand out. First, it provided an empirical comparison between A2C and PPO under bio-inspired constraints, clarifying their respective strengths and weaknesses. Second, it demonstrated the feasibility of coupling CLAPP-trained encoders with reinforcement learning agents in complex navigation tasks. Third, it introduced and tested biologically motivated modifications — such as decorrelation, memory, and intrinsic motivation — that advance the dialogue between neuroscience and artificial intelligence.

Looking forward, there are several directions for future research. One priority is to design more powerful yet biologically grounded representation models that move closer to the richness of hippocampal and entorhinal coding. Another is to improve intrinsic motivation mechanisms so that they provide robust exploration signals without excessive computational cost. Extending this framework to more challenging three-dimensional environments would also test its scalability and ecological validity. Finally, combining CLAPP with other self-supervised paradigms may yield hybrid models that balance biological plausibility with efficiency.

In summary, this thesis has shown that biologically inspired representation learning can support reinforcement learning in navigation tasks, offering a meaningful step toward aligning artificial agents with principles of neural computation. While challenges remain, the results point to a promising direction where insights from neuroscience and machine learning can converge to produce agents that are both effective and biologically grounded.