Spatial Navigation Using Bio-plausible Learning Rules

Mattia Assane Wade, Aurelien Cormorèche
Laboratory of Computational Neuroscience

INTRODUCTION

Most deep reinforcement learning algorithms rely on backpropagation, which is not biologically plausible.

A more biologically realistic model for learning in the brain is a three-factor rule, which is effective in simple networks but struggles with complex tasks like visual navigation.

The research investigates whether **biologically plausible representation learning (CLAPP)** can be used to generate features that allow a simple, one-layer reinforcement learning agent to solve complex tasks like **maze navigation**.

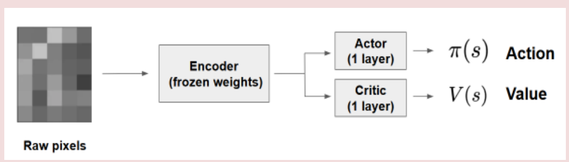
CLAPP - Bio-Plausible Encoder

- A bio-plausible deep learning algorithm (Illing & Gerstner, 2021).
- Learns representations with **local learning rules** without requiring backpropagation
- Training:
  - Pretrained on grayscale STL-10 dataset to extract general-purpose features.
  - Encoder weights frozen during RL tasks.
- Produces rich and structured representations suitable for downstream RL

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

ALGORITHM

- Reinforcement learning module: **Actor-Critic with eligibility traces**.
- Constraints: only **one-layer actor** and **one-layer critic** to preserve biological plausibility.
- Training signals:
  - Neuromodulation provided by reward/ temporal-difference error.
  - Local weight updates following the three-factor rule.



$\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ (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
$\delta \leftarrow r + \gamma\, \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$

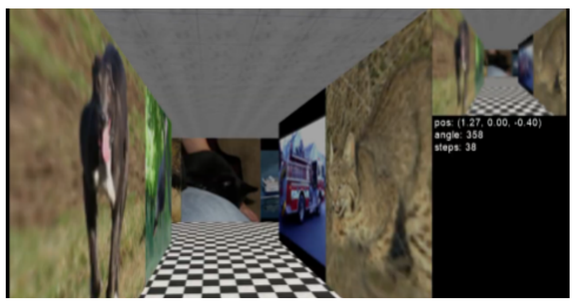
$\mathbf{z}^{\mathbf{w}} \leftarrow \lambda^{\mathbf{w}} \mathbf{z}^{\mathbf{w}} + \nabla \hat{v}(S, \mathbf{w})$
$\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda^{\boldsymbol{\theta}} \mathbf{z}^{\boldsymbol{\theta}} + \nabla \ln \pi(A|S, \boldsymbol{\theta})$
$\mathbf{w} \leftarrow \mathbf{w} +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

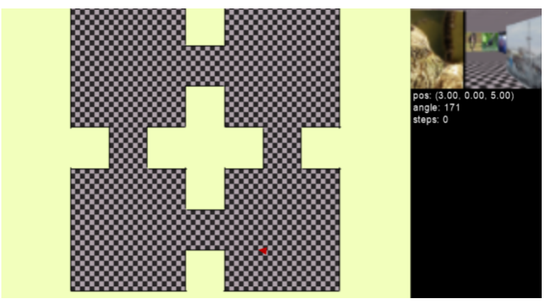
Adapted from Sutton and Barto

ENVIRONMENT

MiniWorld-Gymnasium framework: T-Maze and Four Rooms

- Visual walls textured with STL-10 images (not used in encoder pretraining).
- Agent spawns at random position/orientation, ensuring infinite state space.
- Actions: forward, left turn, right turn (fixed 45°), reward localized (e.g., left side of T-maze).
- Challenge: high-dimensional visual input without convolution → need expressive encoder.



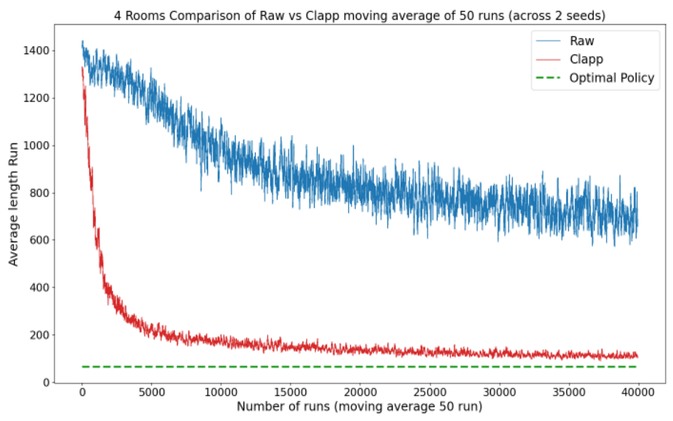


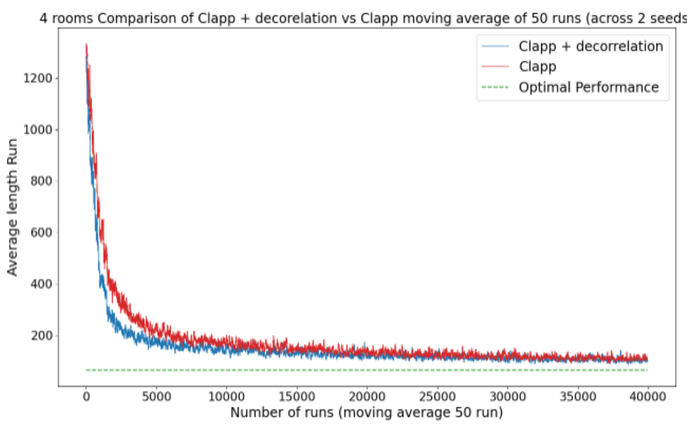
RESULTS

T-Maze:
- One-layer actor-critic with CLAPP features learns an **almost optimal policy**.
- Demonstrates end-to-end bio-plausible navigation from vision.

Four Rooms:
- CLAPP features generalize to more complex layouts.
- Raw pixel baselines fail due to lack of invariance and high correlation among features.

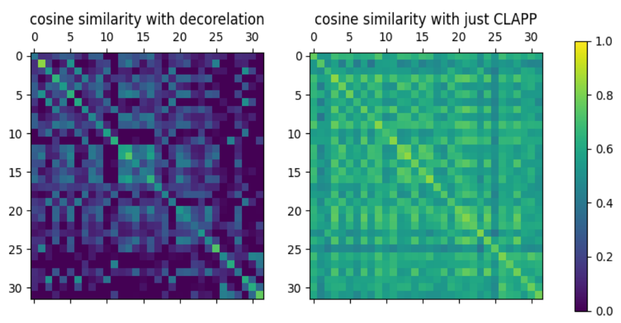
Limitation: Similarity matrices show **highly correlated representations** across states





After decorrelation layer:

- Performance improves compared to only CLAPP features.
- Still below PPO or multi-layer baselines, but more stable and efficient than without decorrelation.
- Contrastive decorrelation reduces feature redundancy.
- Supports one-layer actor-critic learning by making inputs more separable



DECORRELATION

CLAPP features are often too correlated, leading to critic instability and slow learning. Inspired by hippocampal place cells, we added a **decorrelation layer** trained with a contrastive InfoNCE loss. The agent explores the maze, where:

- Recent observations act as positives (similar)
- Older observations act as negatives (dissimilar)

A **cascade memory buffer** structures these comparisons, encouraging nearby states to share features while pushing apart distant ones.

$$\mathcal{L} = -\log\left(\frac{e^{\text{sim}(f_x, f_{y+})}}{e^{\text{sim}(f_x, f_{y+})} + \sum_{y^- \in Y^-} e^{\text{sim}(f_x, f_{y-})}}\right)$$

CONCLUSION

CLAPP and a one-layer actor-critic model successfully enabled near-optimal navigation from visual input.

A decorrelation layer was introduced to improve the stability and expressiveness of the features.

The decorrelation layer made the learned features more similar to place cell representations found in the hippocampus.