A
Presentation
on

INFOBOT
(Information Bottleneck)

Transfer and Exploration
via
The Information Bottleneck

Anirudh Goyal et al. “InfoBot: Transfer and Exploration via the Information Bottleneck”. In: arXiv e-prints, arXiv:1901.10902 (2019), arXiv:1901.10902. arXiv: 1901.10902 [stat.ML].

Introduction

Idea

- We usually follow **default action**

  Because there are a few **decision states**

- Use of **Multi-goal task** structure is easy

  Because **sub-goal to sub-goal transition** is easy to learn

  Decision states are inherently **sub-goals**

Where does it fit ?

- In **Exploration**
- In **Policy Transfer**

- For shaping

  Single-Goal to **Multi-goal** environment

  to learn **goal conditioned** policies

Expected Results

- **Goal Conditioned Policy with Information Bottleneck**

  leads to better policy transfer than

  standard RL training procedures

- **Decision state based exploration bonus**

  leads to better performance than

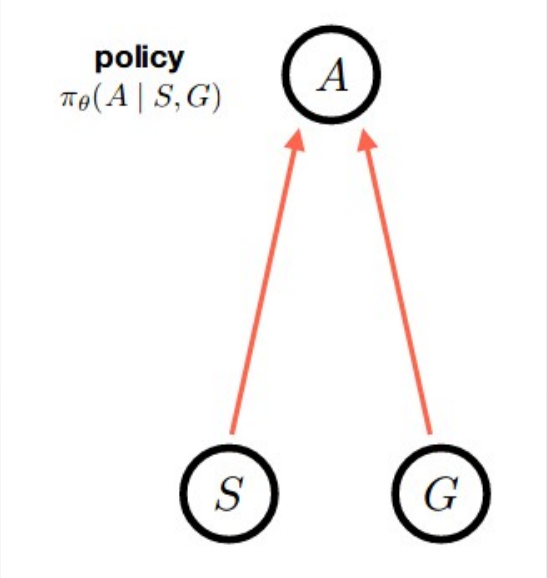
  standard task-agnostic exploration methods

Mathematical Formulation

Regularization

Objective Learning to Share and Hide Intentions
using Information Regularization
by Strouse et. Al (2018)

$$J(\theta) \equiv \mathbb{E}_{\pi_\theta}[r] - \beta I(A; G \mid S)$$



Objective Function

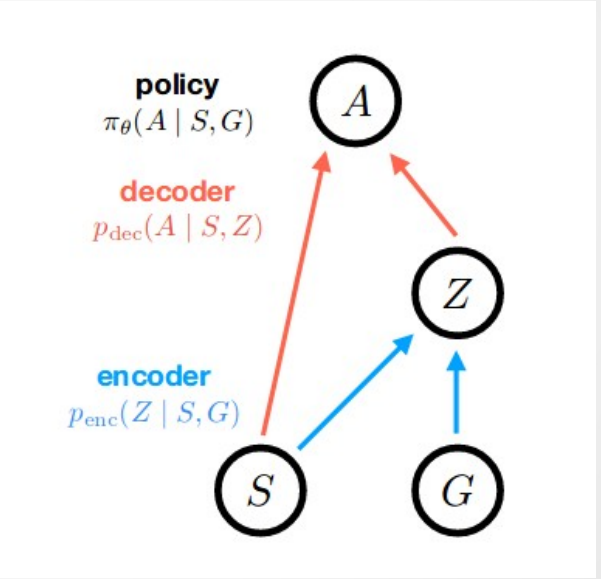
$$J(\theta) \equiv \mathbb{E}_{\pi_\theta}[r] - \beta I(A; G \mid S)$$

$$\pi_\theta(A \mid S, G) = \sum_z p_{\text{enc}}(z \mid S, G)\, p_{\text{dec}}(A \mid S, z)$$

$$I(Z; G \mid S) \geq I(A; G \mid S)$$

$$\begin{aligned} J(\theta) &\geq \mathbb{E}_{\pi_\theta}[r] - \beta I(Z; G \mid S) \\ &= \mathbb{E}_{\pi_\theta}[r - \beta D_{\text{KL}}[p_{\text{enc}}(Z \mid S, G) \mid p(Z \mid S)]] \end{aligned}$$

$$p(Z \mid S) = \sum_g p(g)\, p_{\text{enc}}(Z \mid S, g)$$



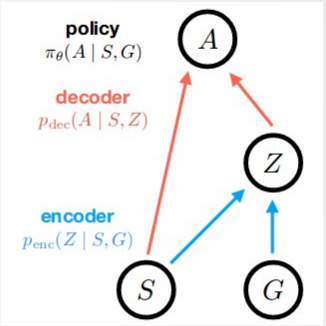
Objective function

$$J(\theta) \geq \tilde{J}(\theta) \equiv \mathbb{E}_{\pi_\theta}[r - \beta D_{\text{KL}}[p_{\text{enc}}(Z \mid S, G) \mid q(Z \mid S)]]\,.$$

Final functions

Objective function

$$J(\theta) \geq \tilde{J}(\theta) \equiv \mathbb{E}_{\pi_\theta}[r - \beta D_{\mathrm{KL}}\left[p_{\mathrm{enc}}(Z \mid S, G) \mid q(Z \mid S)\right]] .$$

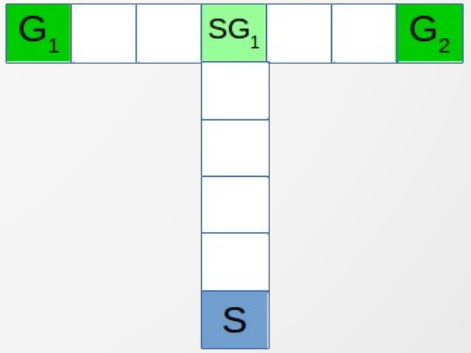


Learning Step

$$\nabla_\theta \tilde{J}(t) = \tilde{R}_t \log(\pi_\theta(a_t \mid s_t, g_t)) - \beta \nabla_\theta D_{\mathrm{KL}}\left[p_{\mathrm{enc}}(Z \mid s_t, g_t) \mid q(Z \mid s_t)\right]$$

$$\tilde{R}_t \equiv \sum_{u=t}^{T} \gamma^{u-t} \tilde{r}_u$$

$$\tilde{r}_t \equiv r_t + \beta D_{\mathrm{KL}}\left[p_{\mathrm{enc}}(Z \mid s_t, g) \mid q(Z \mid s_t)\right]$$



Policy Transfer with Visitation Penalty

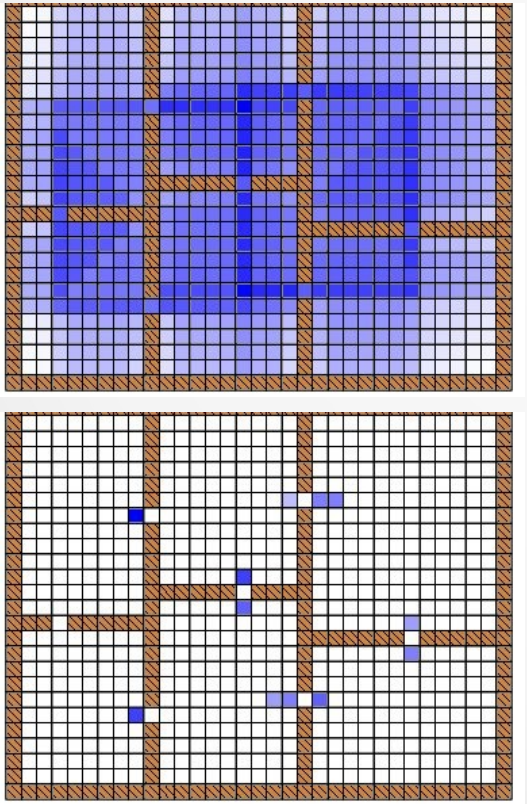
$$r_t = r_e(t) + \frac{\beta}{\sqrt{c(s_t)}} D_{\mathrm{KL}}\left[p_{\mathrm{enc}}(Z \mid s_t, g_t) \mid q(Z \mid s_t)\right] .$$

Algorithm

Sub-Goal Identification

Grounding subgoals in information transitions

by Van Dijk et. al. (2011)



Learn policy π_g over $\mathcal{O}$ for g
$< \mathcal{I}_o, \beta_o, \pi_o >$

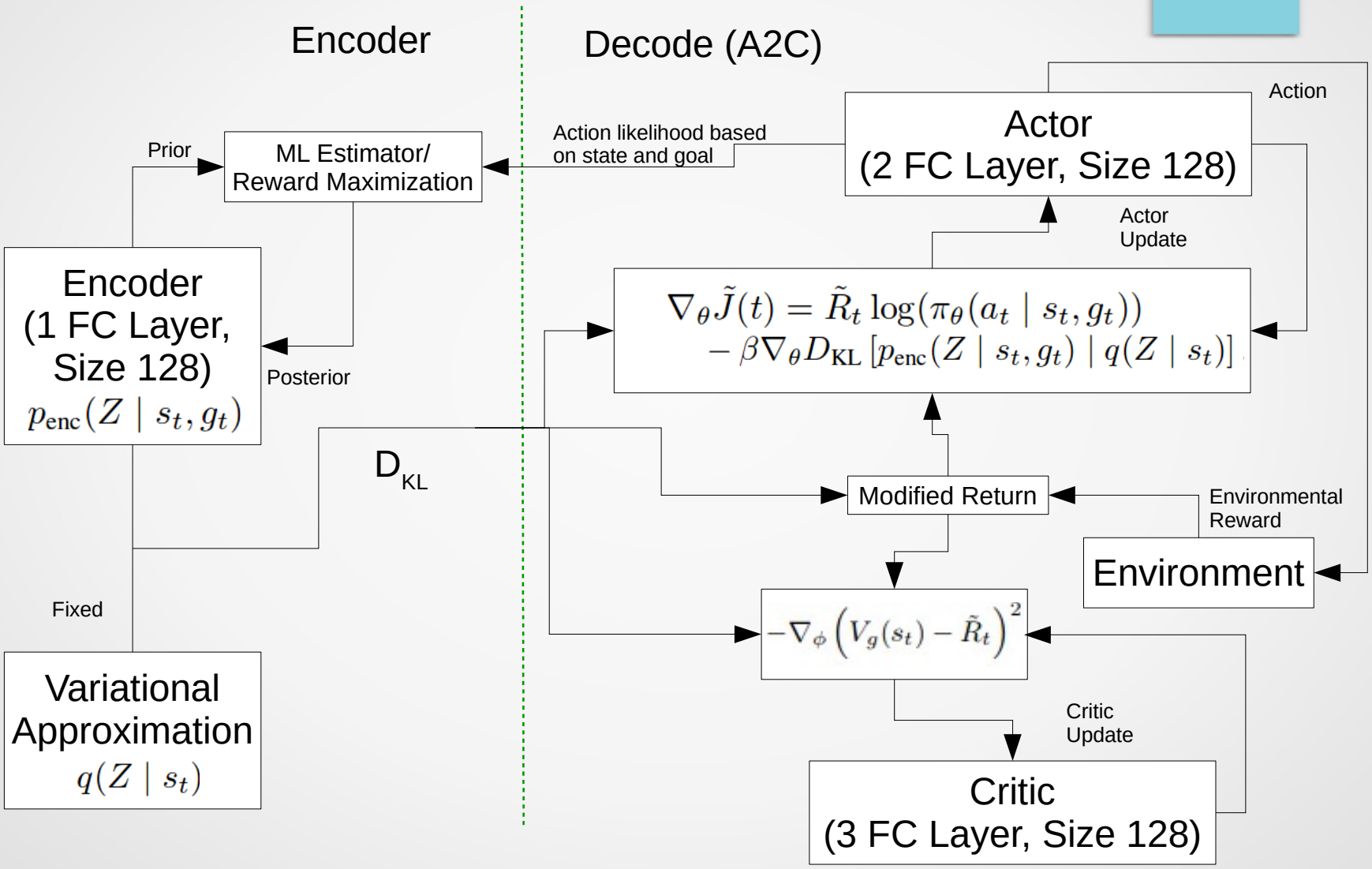
Initialise $p(g|\mathbf{e}_0)$ uniformly

$$p_v(g|\mathbf{e}_{t+1}) \leftarrow \frac{1}{Z} p(a_t|s_t, g) p_v(g|\mathbf{e}_t)$$

$$\delta(s_t) \leftarrow \delta(s_t) + H(G|\mathbf{e}_t) - H(G|\mathbf{e}_{t+1})$$

Learn policy π_s over $\mathcal{A}$ for s

Learning


Encoder
Decode (A2C)
Action
Actor
(2 FC Layer, Size 128)
Action likelihood based on state and goal
Prior
ML Estimator/
Reward Maximization
Actor
Update
Encoder
(1 FC Layer,
Size 128)
$p_{\text{enc}}(Z \mid s_t, g_t)$
Posterior
$\nabla_\theta \tilde{J}(t) = \tilde{R}_t \log(\pi_\theta(a_t \mid s_t, g_t))$
$- \beta \nabla_\theta D_{\text{KL}}\left[p_{\text{enc}}(Z \mid s_t, g_t) \mid q(Z \mid s_t)\right]$
D_{KL}
Modified Return
Environmental
Reward
Environment
Fixed
$-\nabla_\phi \left(V_g(s_t) - \tilde{R}_t\right)^2$
Critic
Update
Variational
Approximation
$q(Z \mid s_t)$
Critic
(3 FC Layer, Size 128)

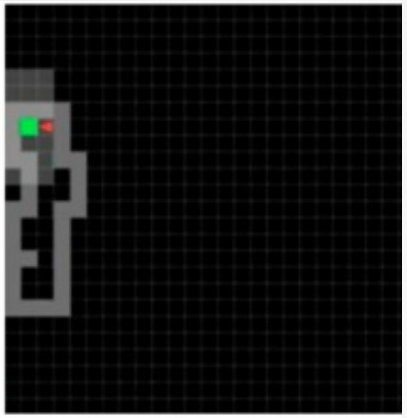
Algorithm

Algorithm 1 Transfer and Exploration via the Information Bottleneck

Require: A policy $\pi_\theta(A \mid S, G) = \sum_z p_{\text{enc}}(z \mid S, G)\, p_{\text{dec}}(A \mid S, z)$, parameterized by θ
Require: A variational approximation $q(Z \mid S)$ to the goal-marginalized encoder
Require: A regularization weight β
Require: Another policy $\pi_\phi(A \mid S, G)$, along with a RL algorithm $\mathcal{A}$ to train it
Require: A set of training tasks (environments) $p_{\text{train}}(T)$ and test tasks $p_{\text{test}}(T)$
Require: A goal sampling strategy $p(G \mid T)$ given a task T
  for episodes = 1 to N_{train} **do**
    Sample a task $T \sim p_{\text{train}}(T)$ and goal $G \sim p(G \mid T)$
    Produce trajectory τ on task T with goal G using policy $\pi_\theta(A \mid S, G)$
    Update policy parameters θ over τ using Eqn 5
  end for
  Optional: use π_θ directly on tasks sampled from $p_{\text{test}}(T)$
  for episodes = 1 to N_{test} **do**
    Sample a task $T \sim p_{\text{test}}(T)$ and goal $G \sim p(G \mid T)$
    Produce trajectory τ on task T with goal G using policy $\pi_\phi(A \mid S, G)$
    Update policy parameters ϕ using algorithm $\mathcal{A}$ to maximize the reward given by Eqn 6
  end for

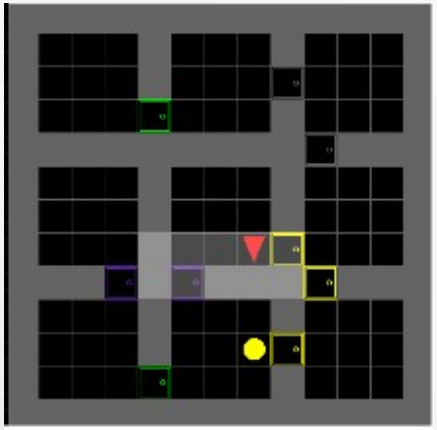
Experiments

Environment

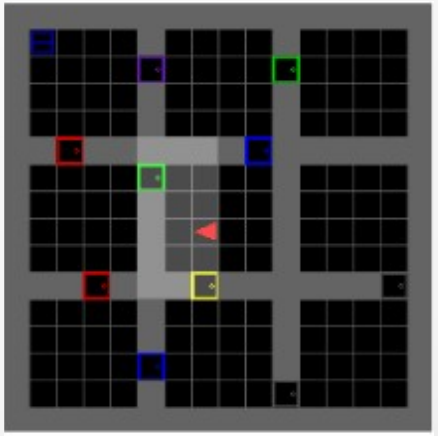


(a) MultiRoomN4S4

(b) MultiRoomN12S10



(c) FindObjS5



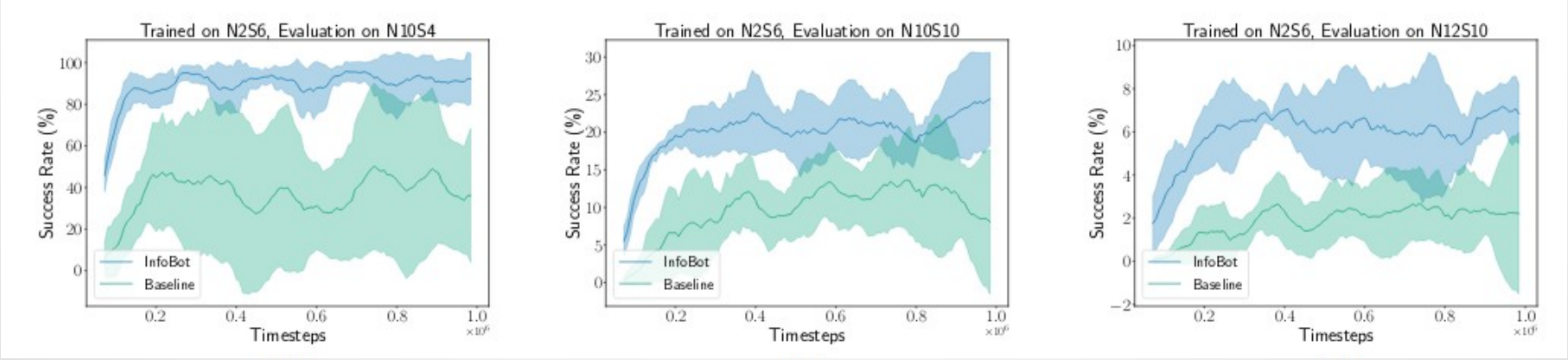
(d) FindObjS6

Direct Policy Generalization

Training on FindObjS5

Method	FindObjS7	FindObjS10
Goal-conditioned A2C	56%	36%
InfoBot with $\beta = 0$	44%	24%
InfoBot	**81%**	**61%**

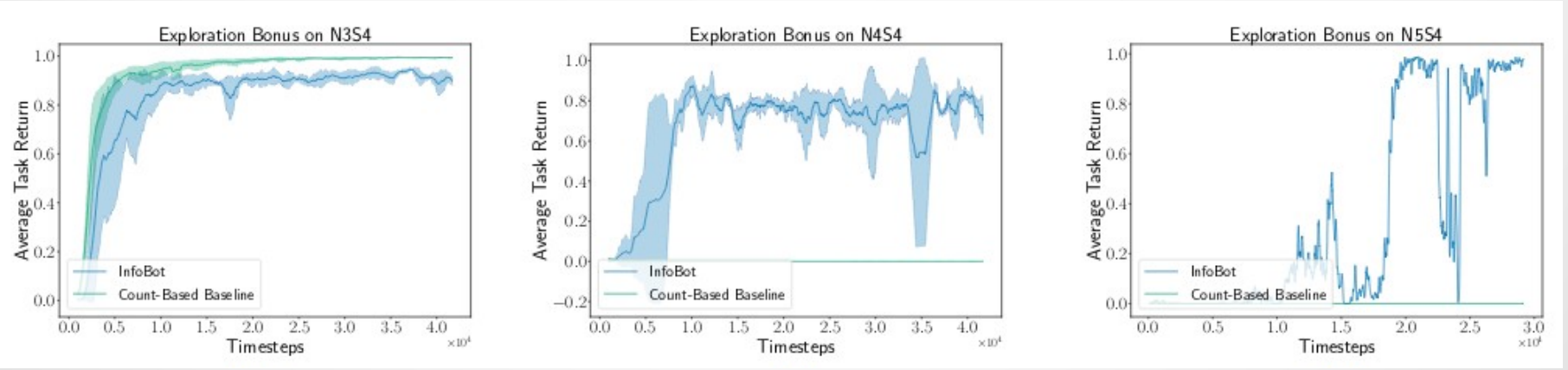
Baseline – Goal Conditioned A2C Agent



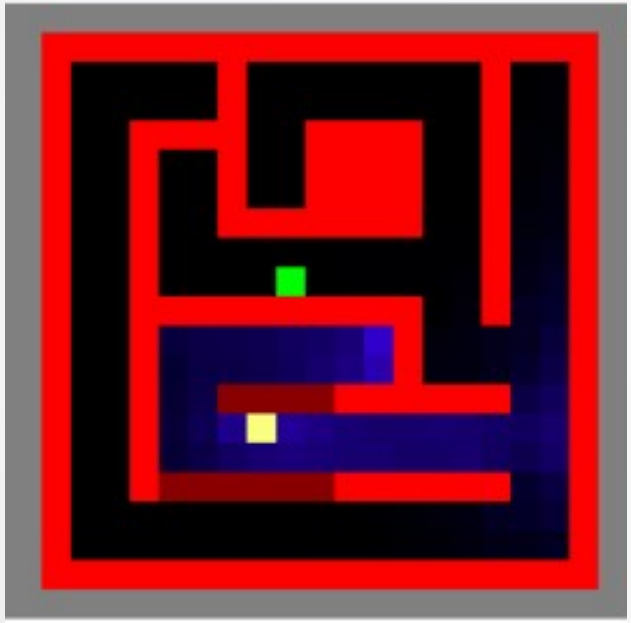
Transferable Exploration Strategies

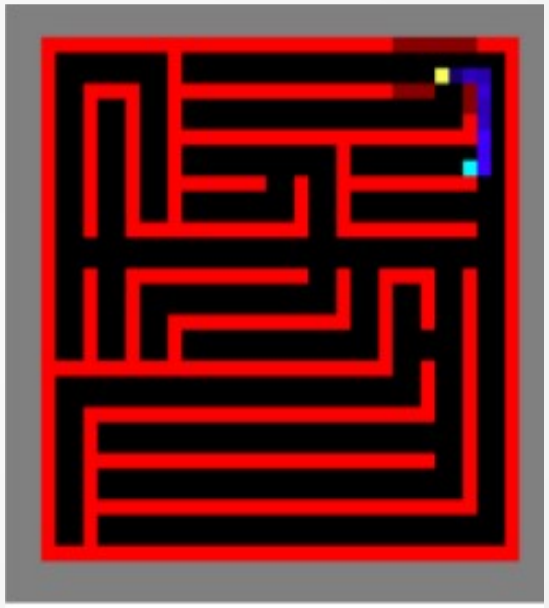
Trained on MultiRoomN2S6

Method	MultiRoomN3S4	MultiRoomN5S4
Goal-conditioned A2C	0%	0%
TRPO + VIME	54%	0%
Count based exploration	**95%**	0%
Curiosity-based exploration	**95%**	54%
InfoBot (decision state exploration bonus)	90%	**85%**



Goal-Based MiniPacMan Navigation

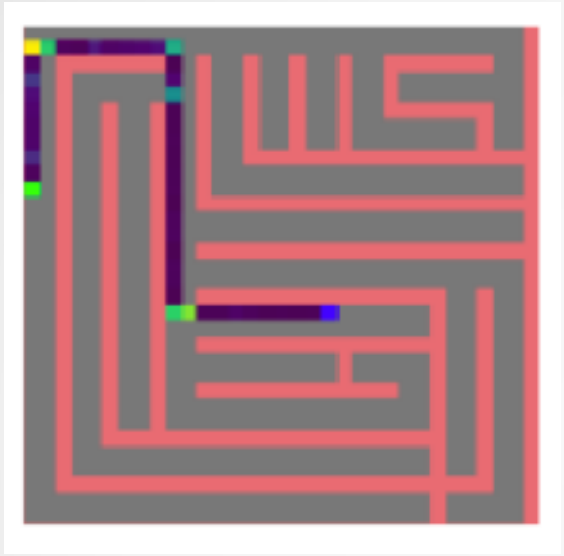


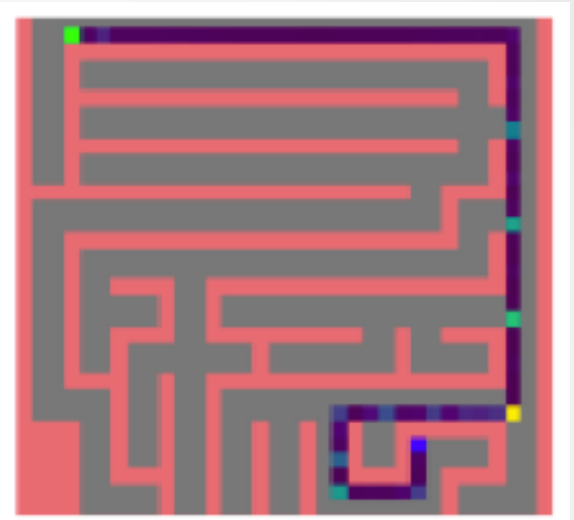


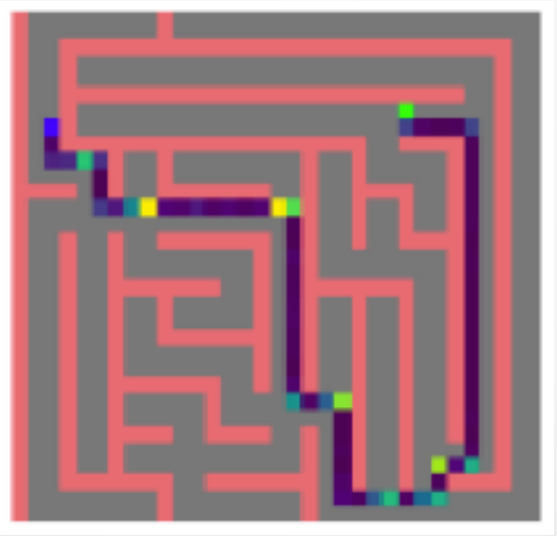
Goal Based Navigation Results

Algorithm (Train on 6×6 maze)	**Evaluate on 11×11 maze**
Actor-Critic	5%
PPO (Proximal Policy Optimization	8%
Actor-Critic + Count-Based	7%
Curiosity Driven Learning (ICM)	47%
Goal Based (UVFA) Goal - TopDownImage of the goal	7%
Goal Based (UVFA) Goal - Relative Dist	15%
Feudal RL	37%
InfoBot (proposed)	64%

Middle Ground – Model-Free and Model-Based RL







Conclusion

- Develop a default behavior
- Knowledge of when to break those behaviors

Thank You