UNIVERSITY OF MICHIGAN

Learning useful features through curiosity-driven exploration

Rajiv Govindjee | Priya Thanneermalai

2022-04-26

Motivation

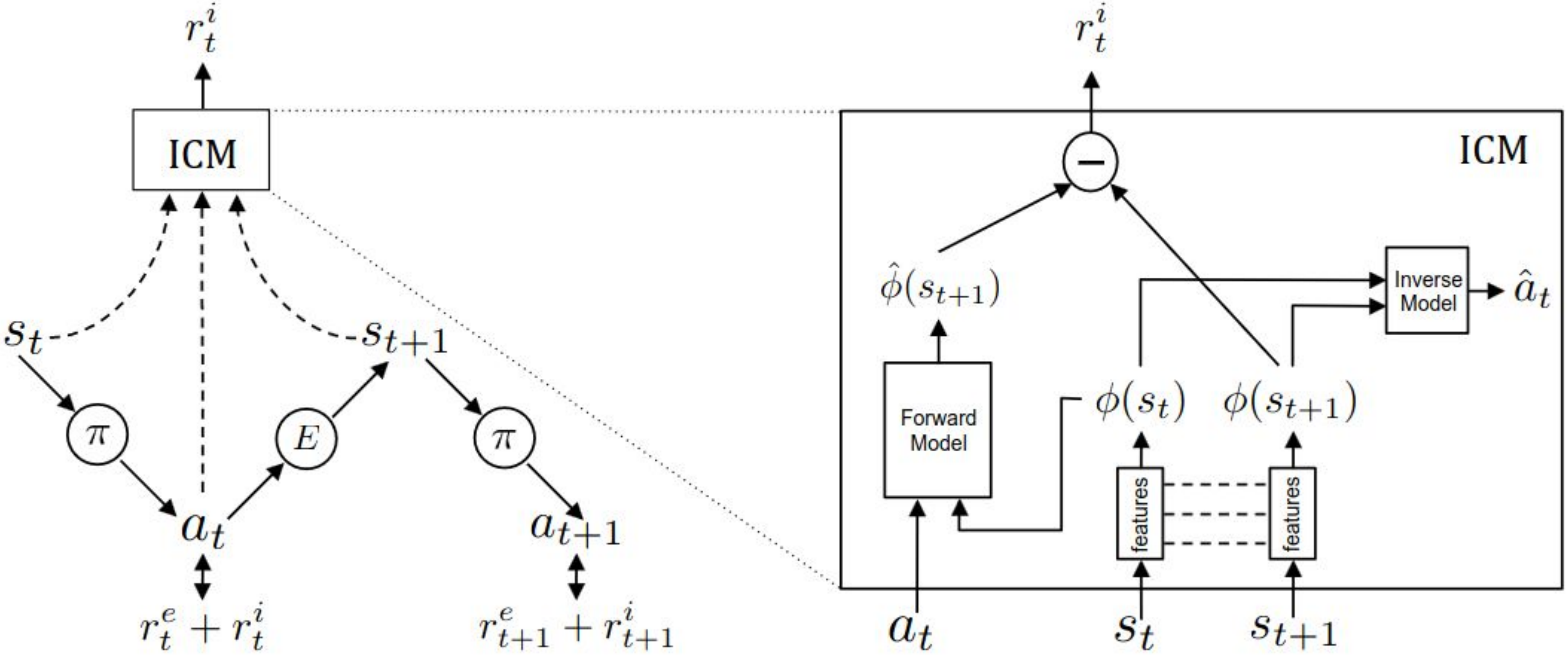
- Sparse rewards in real world settings
- Curiosity based learning in babies
- Potential applications: UAVs, legged robots, anything high-dimensional
- Get better policies by tweaking some technical details in Pathak et. al. paper

[113]

Background

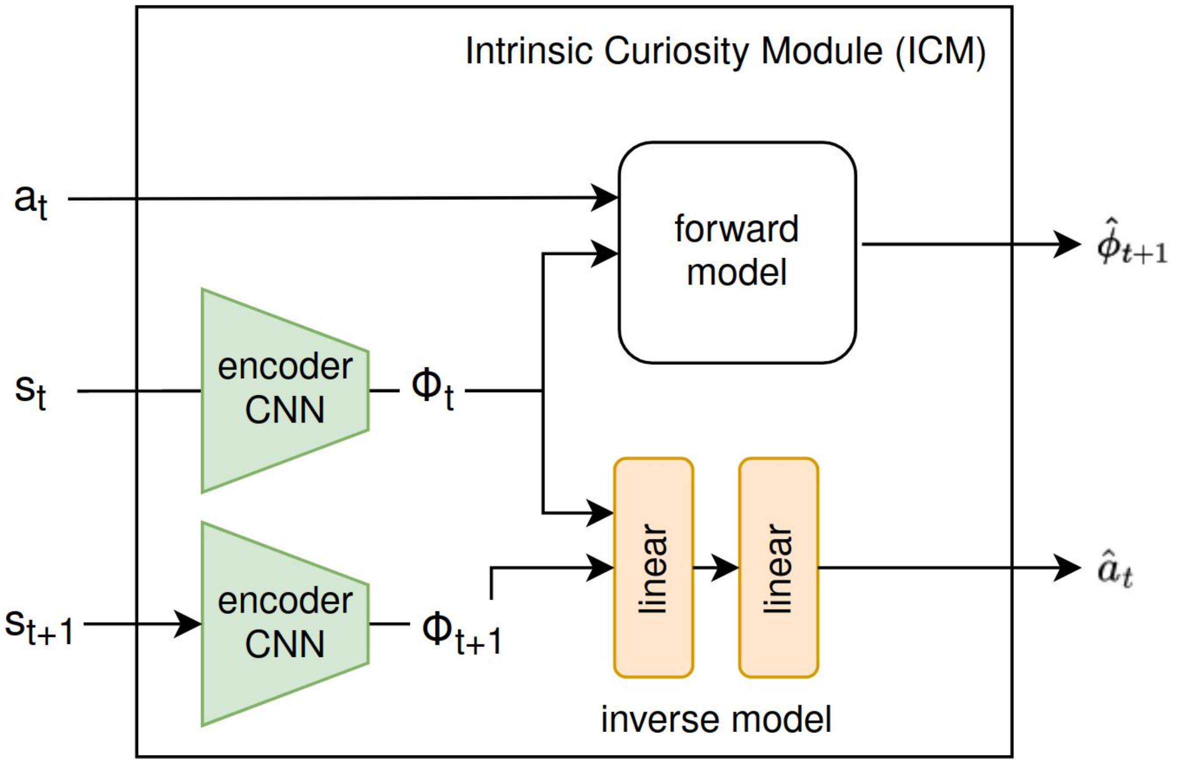
- Seek novel (difficult to predict) situations and learn dynamics
- **Problem**: Agent may simply seek out sources of random noise in the world
- **Solution**: Require agent to make predictions on features that encode information about what agent's own actions can influence.
- **ICM**:
  - **Encoder:** encode states $\mathbf{s}_t$ and $\mathbf{s}_{t+1}$ into features $\mathbf{\Phi}_{st}$ and $\mathbf{\Phi}_{st+1}$
    - Trained on both forward and inverse model loss
  - **Inverse:** use features $\mathbf{\Phi}_{st}$ and $\mathbf{\Phi}_{st+1}$ to predict $\mathbf{a}_t$
  - **Forward:** use $\mathbf{\Phi}_{\ t}$ and $\mathbf{a}_t$ as input to predict feature $\mathbf{\Phi}_{st+1}$ of state $\mathbf{s}_t$.
  - **Output:** curiosity-based reward signal $\mathbf{r}_t$ is prediction error of forward model
- A3C (**policy**) maximizes the reward signal $\mathbf{r}_t$ from ICM

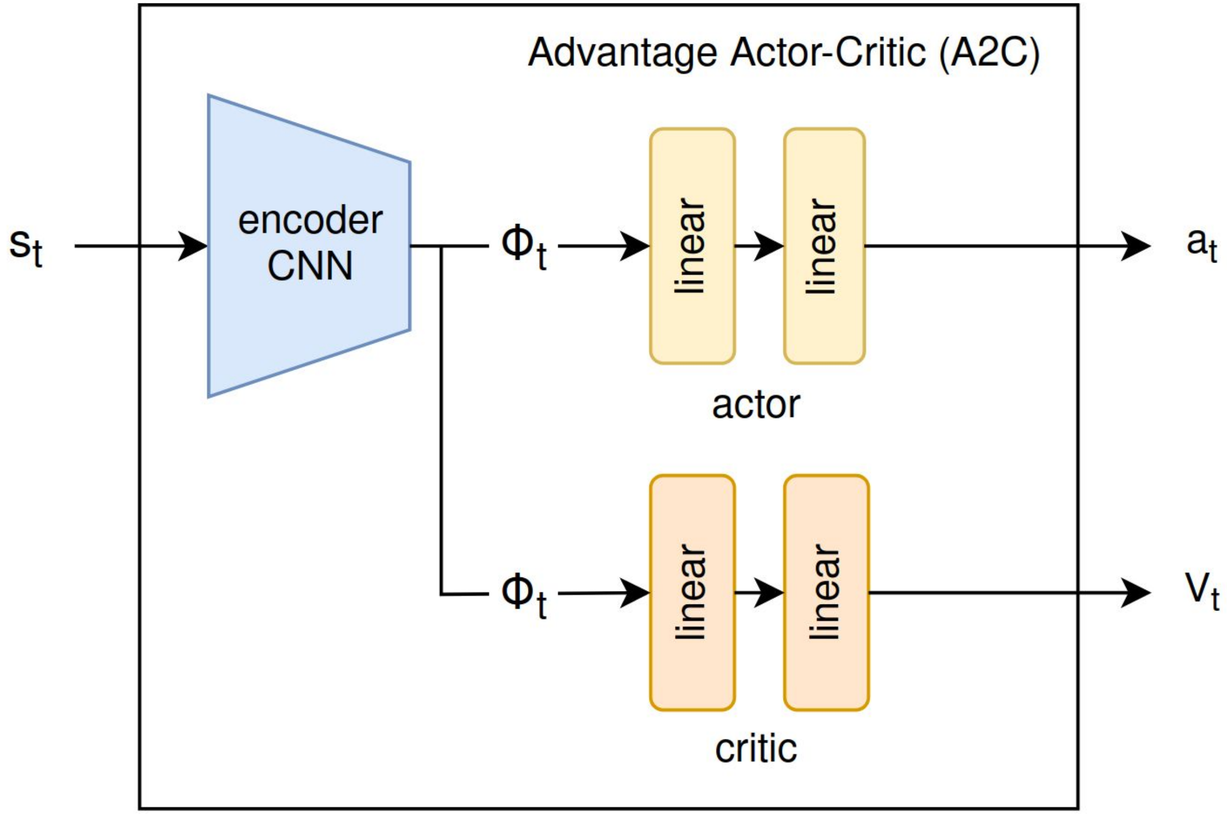
Pathak et. al. on intrinsic + extrinsic rewards



Pathak et. al. on intrinsic + extrinsic rewards

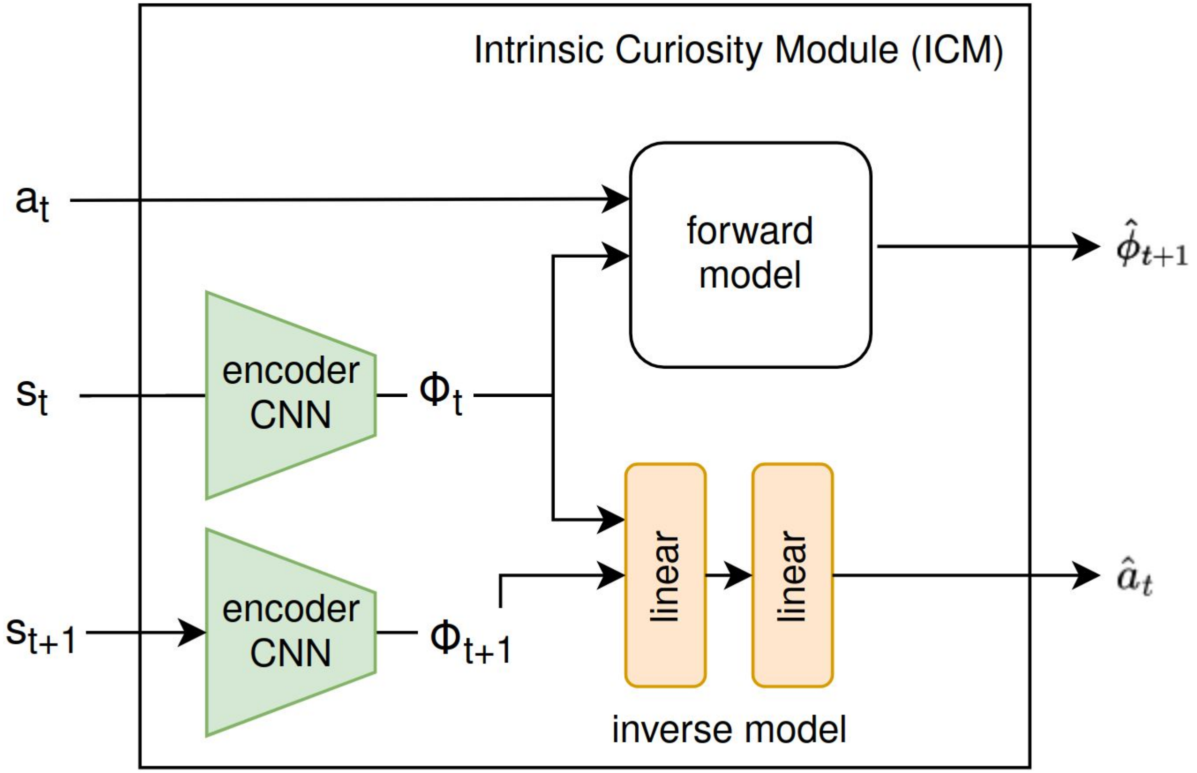
- Separate embeddings for A2C (policy) and ICM (rewards)
- A2C is trained on curiosity reward + extrinsic reward simultaneously

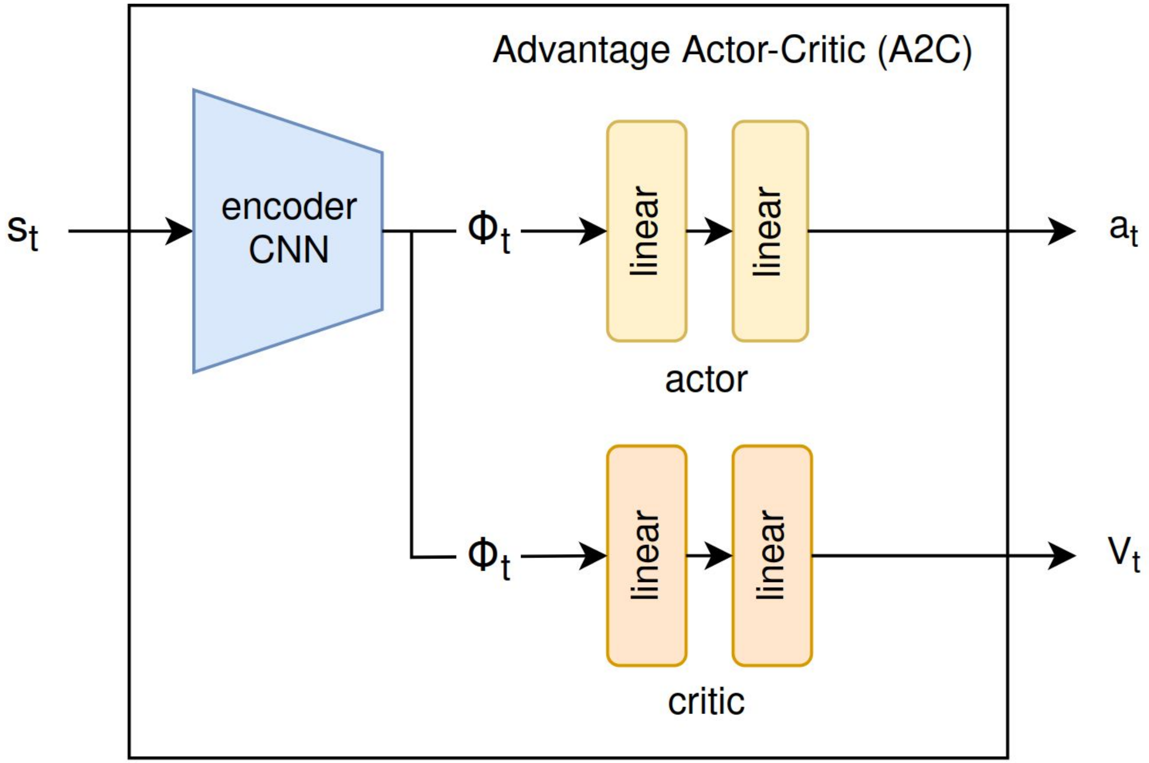




Ours on intrinsic rewards

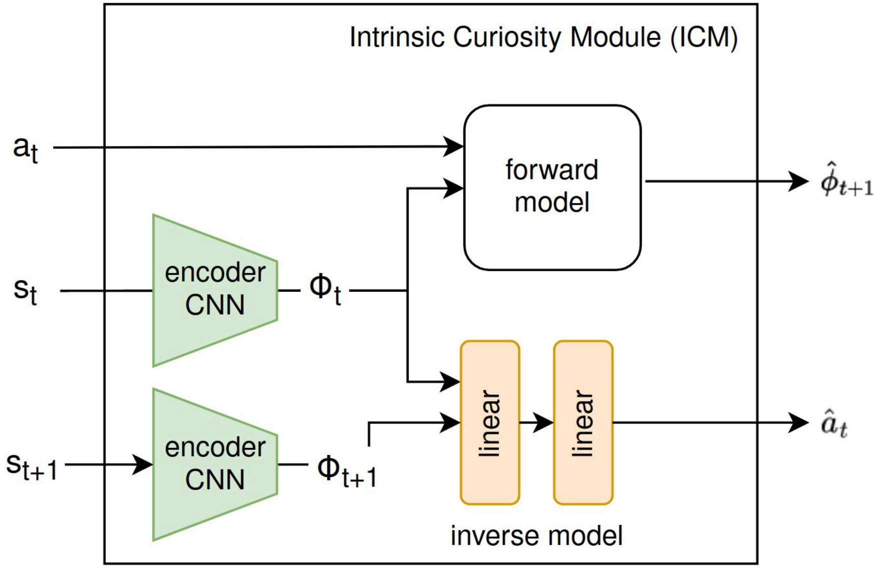
- Separate embeddings for A2C (policy) and ICM (rewards)



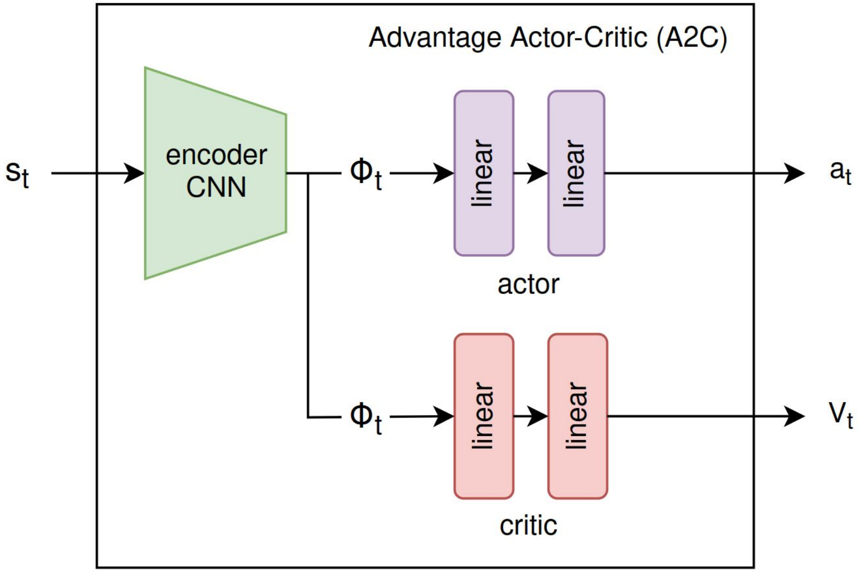


Ours on extrinsic rewards

- Use embeddings (green) from ICM exploration phase
  - These extract useful information from the environment
  - Relevant to predicting agent-related dynamics
- Train new policy to maximize some extrinsic reward (no curiosity reward)



(ICM shown for reference, not used with extrinsic rewards)



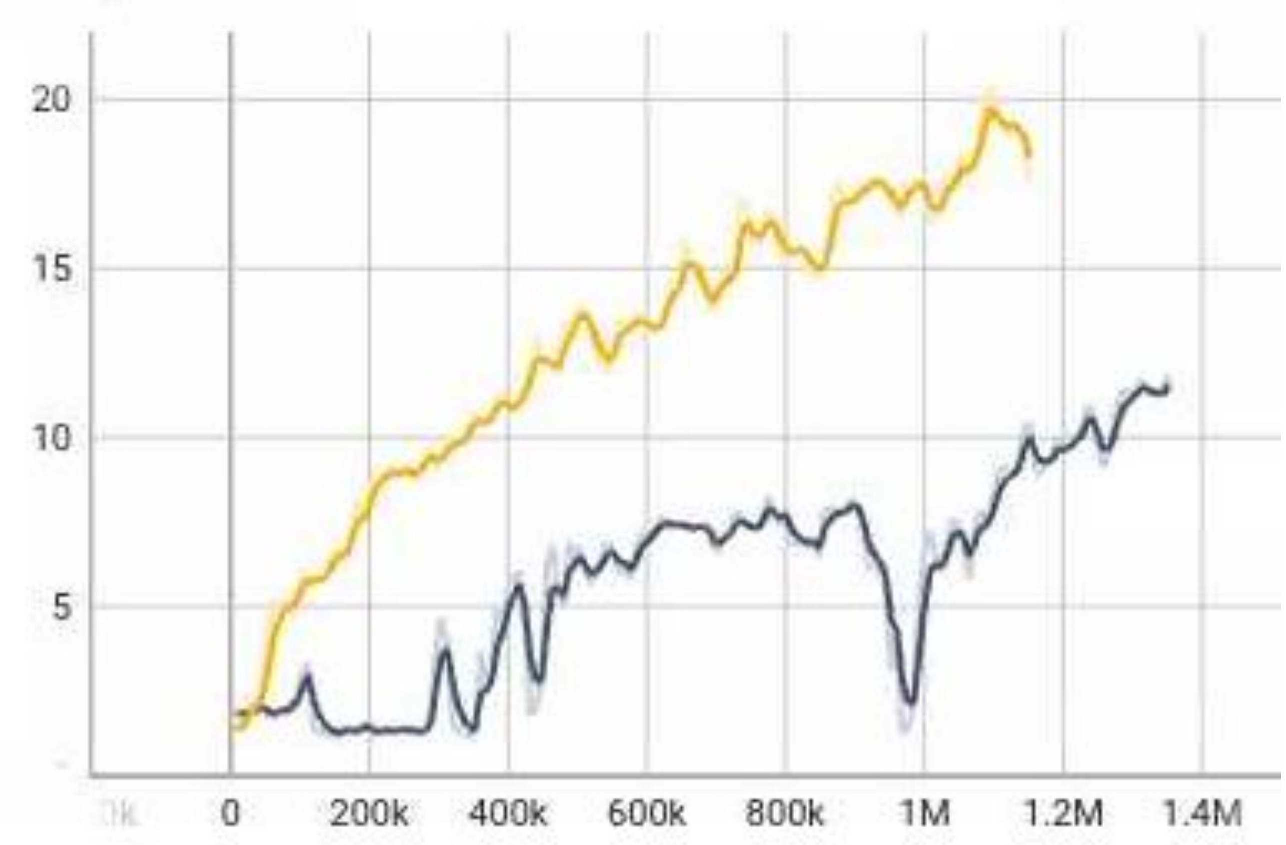
Tech Stack

- OpenAI Gymnasium
  - Atari Games (Breakout)
- PyTorch
- Stable Baselines 3
- Tensorboard
- Docker/Singularity

Experiments

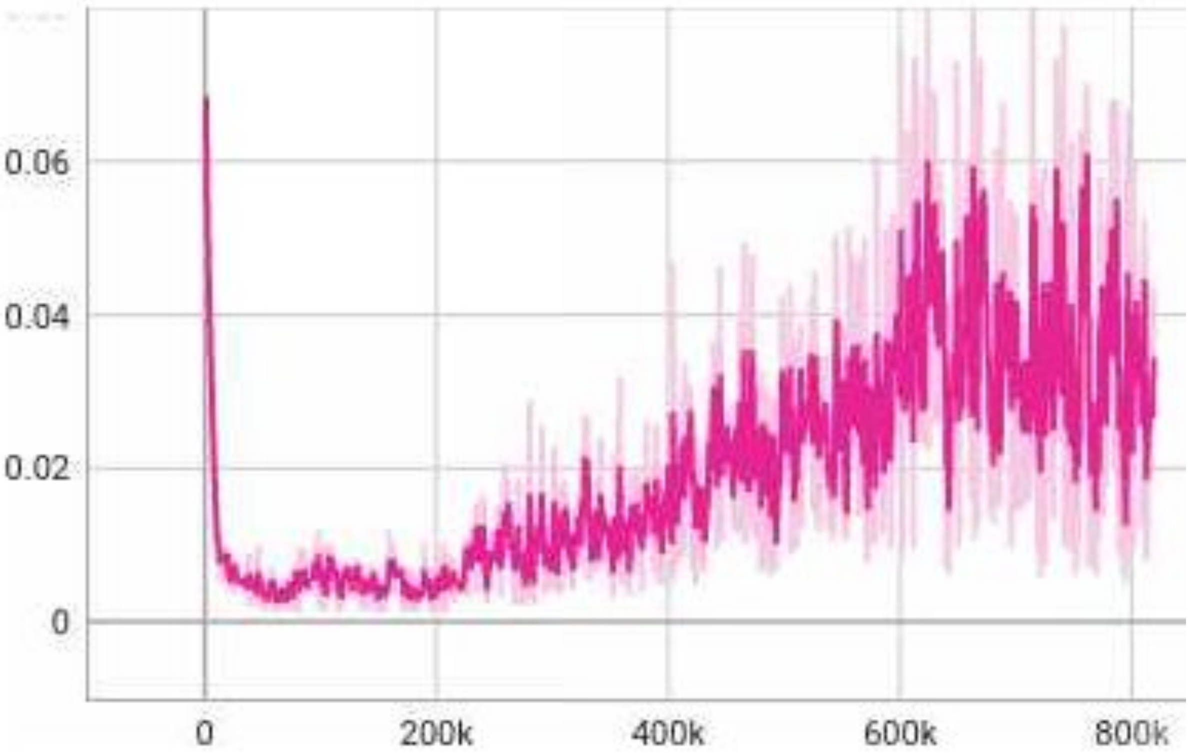
1. Train A2C policy on extrinsic rewards from Break-out. Rewards provided frequently but not for every time step.
2. Train A2C policy to maximize curiosity rewards as calculated by an ICM.
3. Train A2C policy on extrinsic rewards from the game, but using the learned feature extractor from the ICM with feature extractor parameters still being updated at each training step.
4. Train A2C policy on extrinsic rewards from the game, but using the learned feature extractor from the ICM with parameters frozen (so that only the actor and critic heads are being updated)

Results

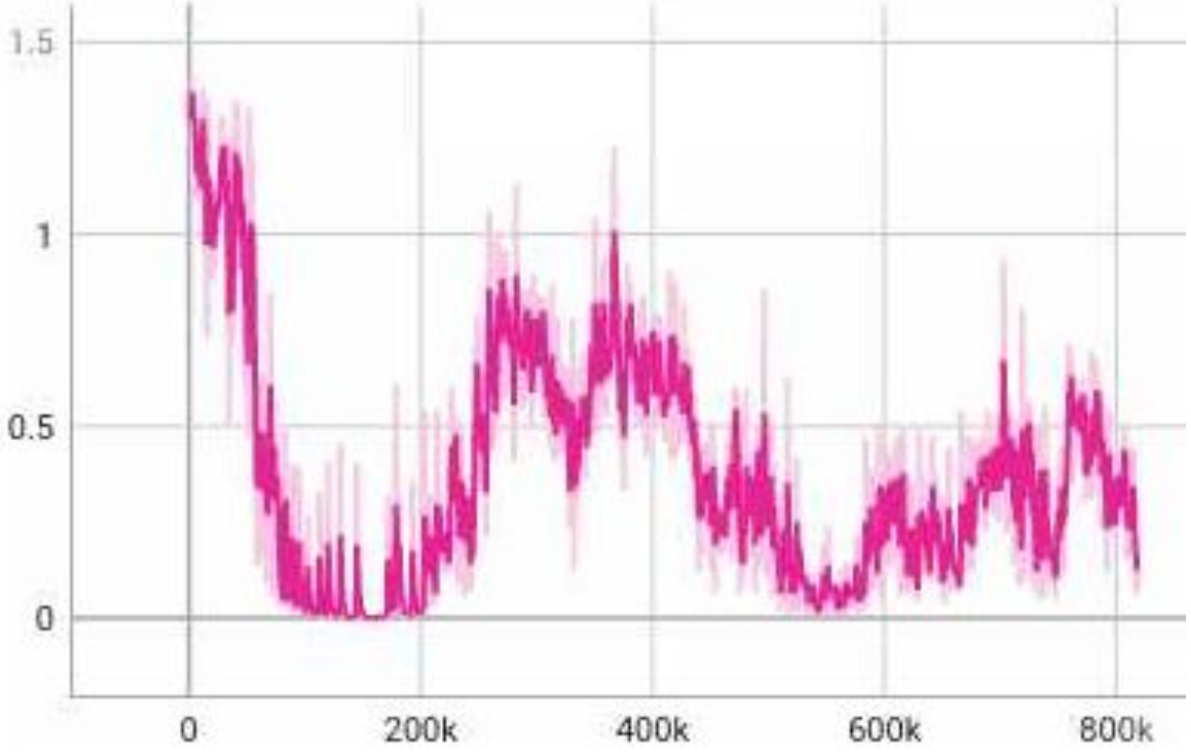


Extrinsic rewards (points) vs environment steps.
Yellow: fine-tuning by using ICM embeddings
Grey: training policy and embeddings from scratch

Results Graphs



Forward loss vs environment steps



Inverse loss vs environment steps

References

- Curiosity-driven Exploration by Self-supervised Prediction
- Learning Gentle Object Manipulation with Curiosity-Driven Deep Reinforcement Learning
- Learning Novel Objects Continually Through Curiosity
- Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions
- Intrinsically Motivated Reinforcement Learning
- Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play

- VIME: Variational Information Maximizing Exploration
- RMA: Rapid Motor Adaptation for Legged Robots
- Attention Is All You Need
- Exploration by Random Network Distillation
- Large-Scale Study of Curiosity-Driven Learning

Questions and Discussion