

Curiosity-Driven Learning

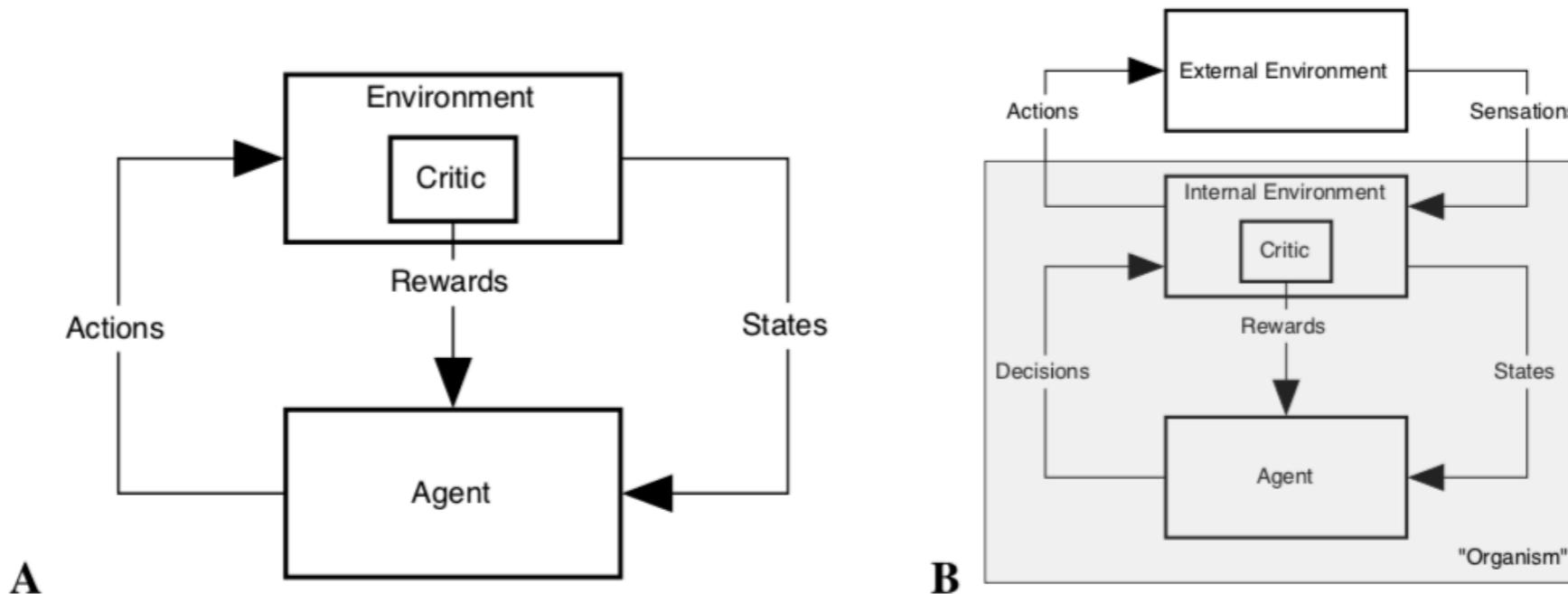
Куканов Виктор
НИУ ВШЭ
5 апреля, 2019

План

- Общее описание "внутренних наград"
- Мотивы применения
- Dynamics-based Curiosity-driven Learning
- Результаты применения подхода
- Ограничения

Intrinsic motivation

- **Intrinsic motivation** – способ мотивировать агента вне зависимости от наград из внешней среды



A – классическая схема

B – схема с *intrinsic motivation*

Зачем?

- RL алгоритмы сильно зависят от вручную созданной системы наград
- Требует много человека-часов для создания (плохо масштабируется)
- Для сложных игровых сред сложно создать сбалансированную систему наград
- Любой просчёт в балансе может привести к неработоспособности большинства RL алгоритмов

CoastRunners

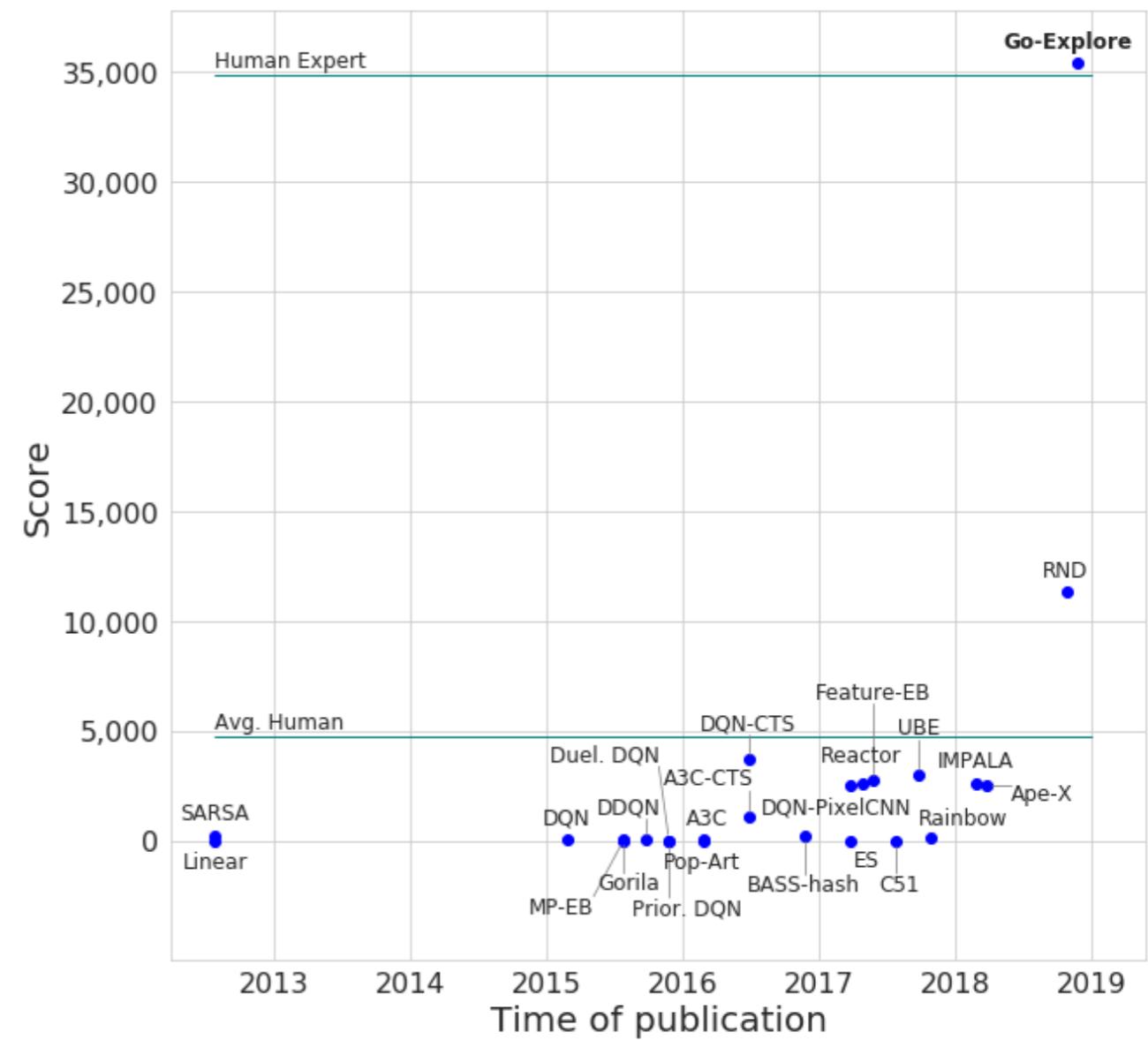
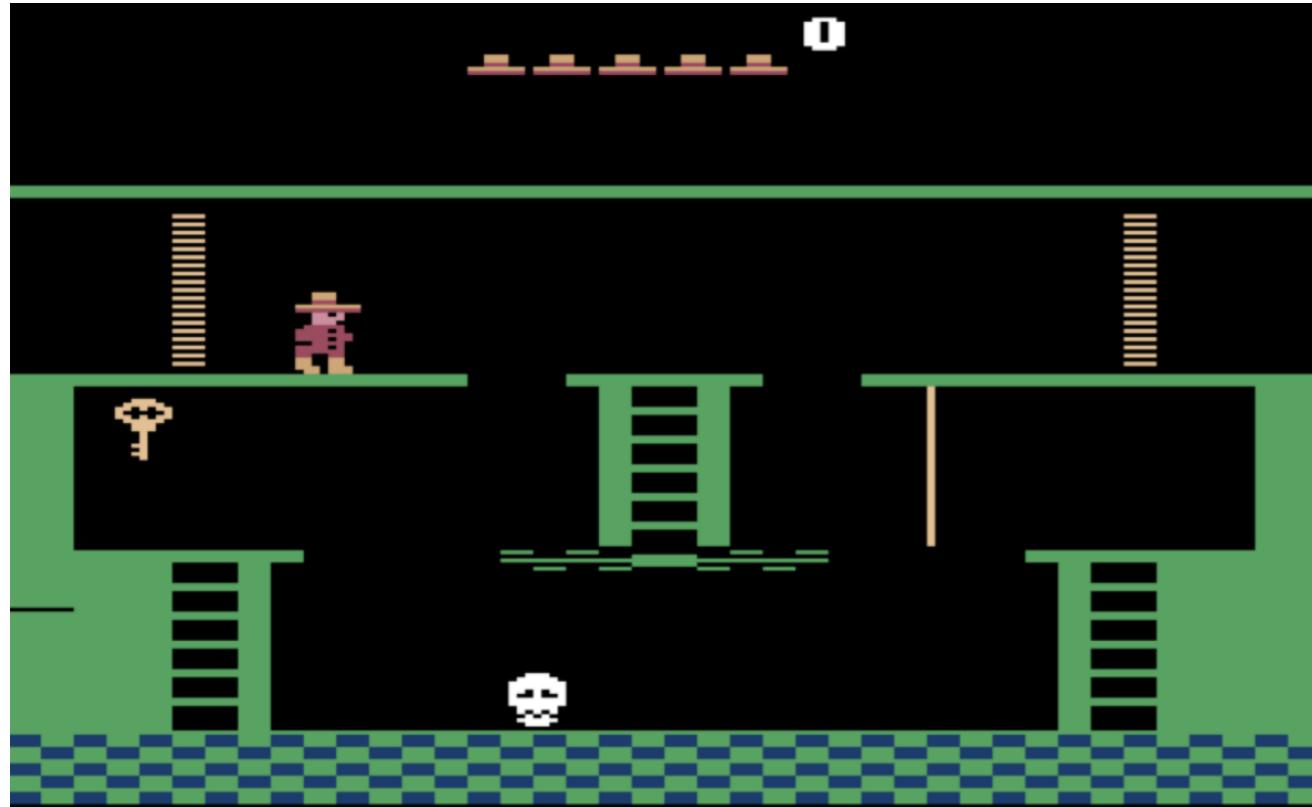


Зачем?

- Основные результаты получены в средах с "плотными" наградами
- RL алгоритмы плохо справляются со средами, где награды приходят очень редко (например, только в конце игры)

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_b Q(s_{t+1}, b) - Q(s_t, a_t))$$

Montezuma's Revenge



Основная идея

- Нужно поощрять агента за посещение новых состояний
- Чем неожиданнее получился результат после действия агента, тем больше награда (**поощрение любопытства агента**)
- Подобная система наград стимулирует агента целенаправленно исследовать новые состояния

Основная идея

- Аналогия с человеком: каждый раз, когда мы узнаем новое, мы испытываем вброс дофамина в зависимости от степени новизны и от значимости новой информации
- Уровень гормонов = настроение = "награда"
- В детском возрасте люди исследуют мир из обычного любопытства, редко преследуя какие-либо внешние награды

Dynamics-based Curiosity-driven Learning

- Мотивируем агента за то, насколько "информационным" было действие
- Если в состоянии s_t совершено действие a_t которое перенесло агента в состояние s_{t+1} то награду можно начислить следующим образом:

$$r_t = -\log p(\phi(s_{t+1}) | \phi(s_t), a_t)$$

Dynamics-based Curiosity-driven Learning

- ϕ преобразует входное состояние (обычно, картинку) в более компактное представление
- Требования к ϕ :
 - компактность
 - сохранение информации
 - устойчивость к незначительным изменениям во входном изображении
 - устойчивость распределения выходов в ходе обучения

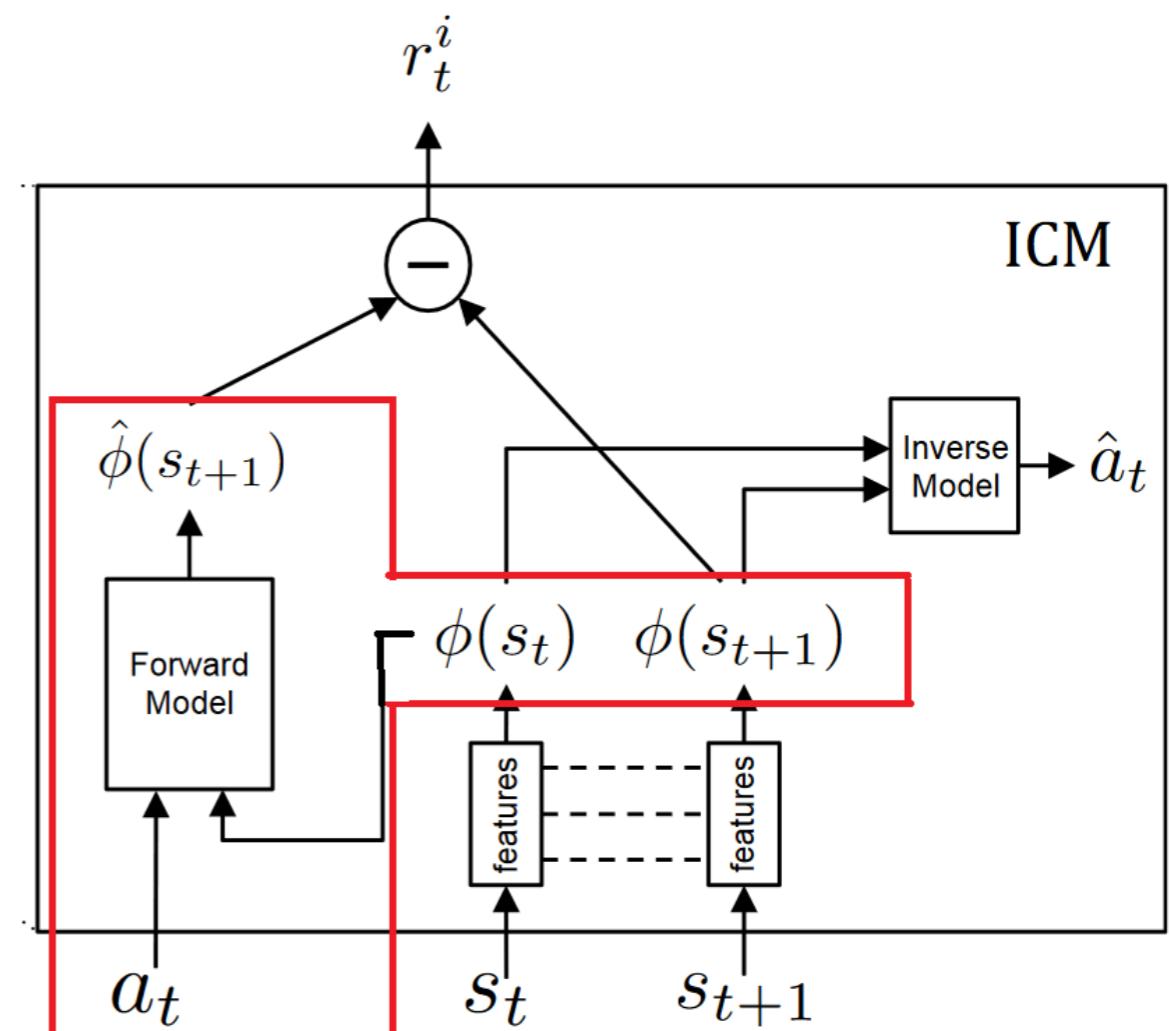
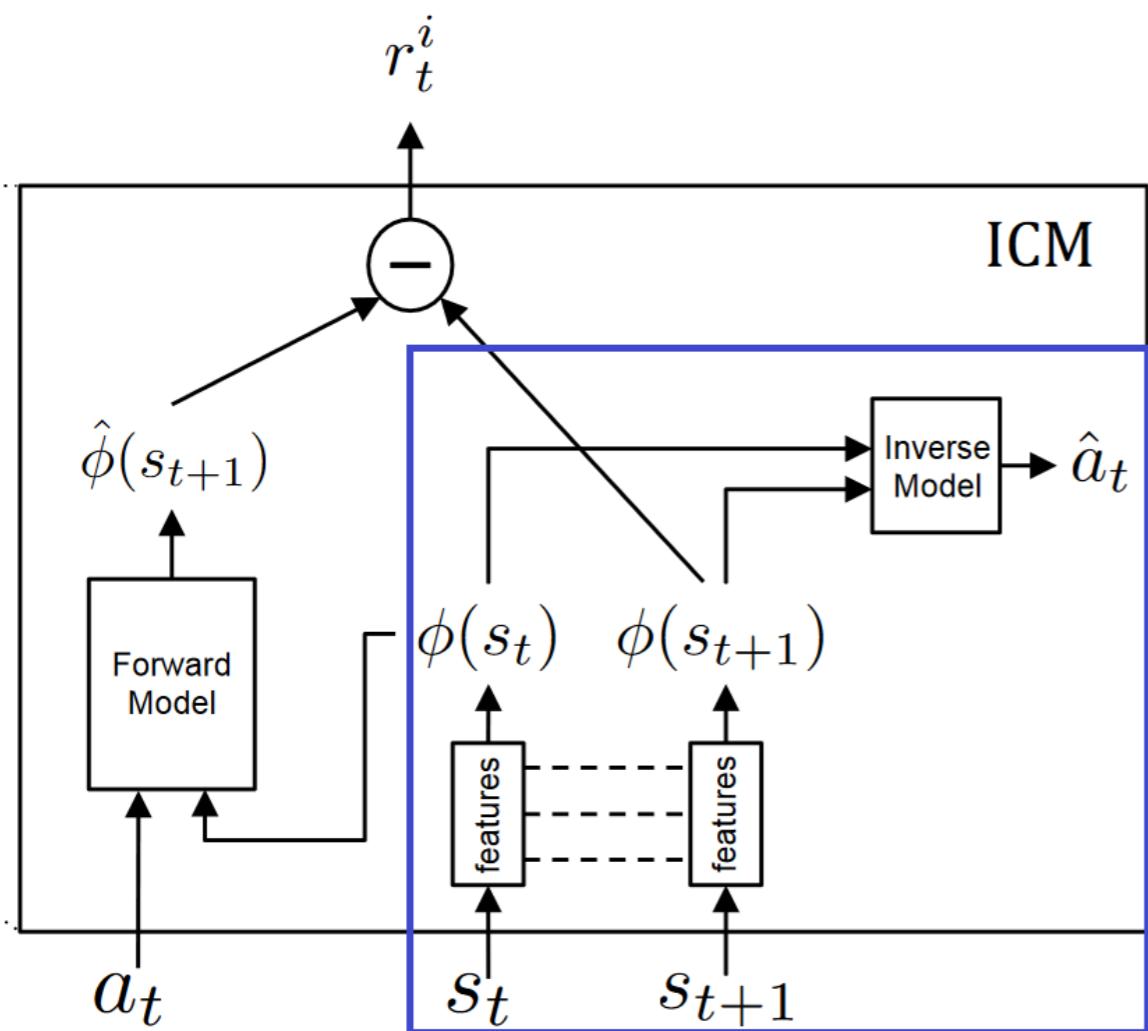
Варианты признаковых пространств

- Пиксели
- Random Features (выходы промежуточного слоя случайно инициализированной CNN)
- VAE Features
- Inverse Dynamics Features (выходы промежуточного слоя сети, которая по двум состояниям предсказывает совершённое действие)

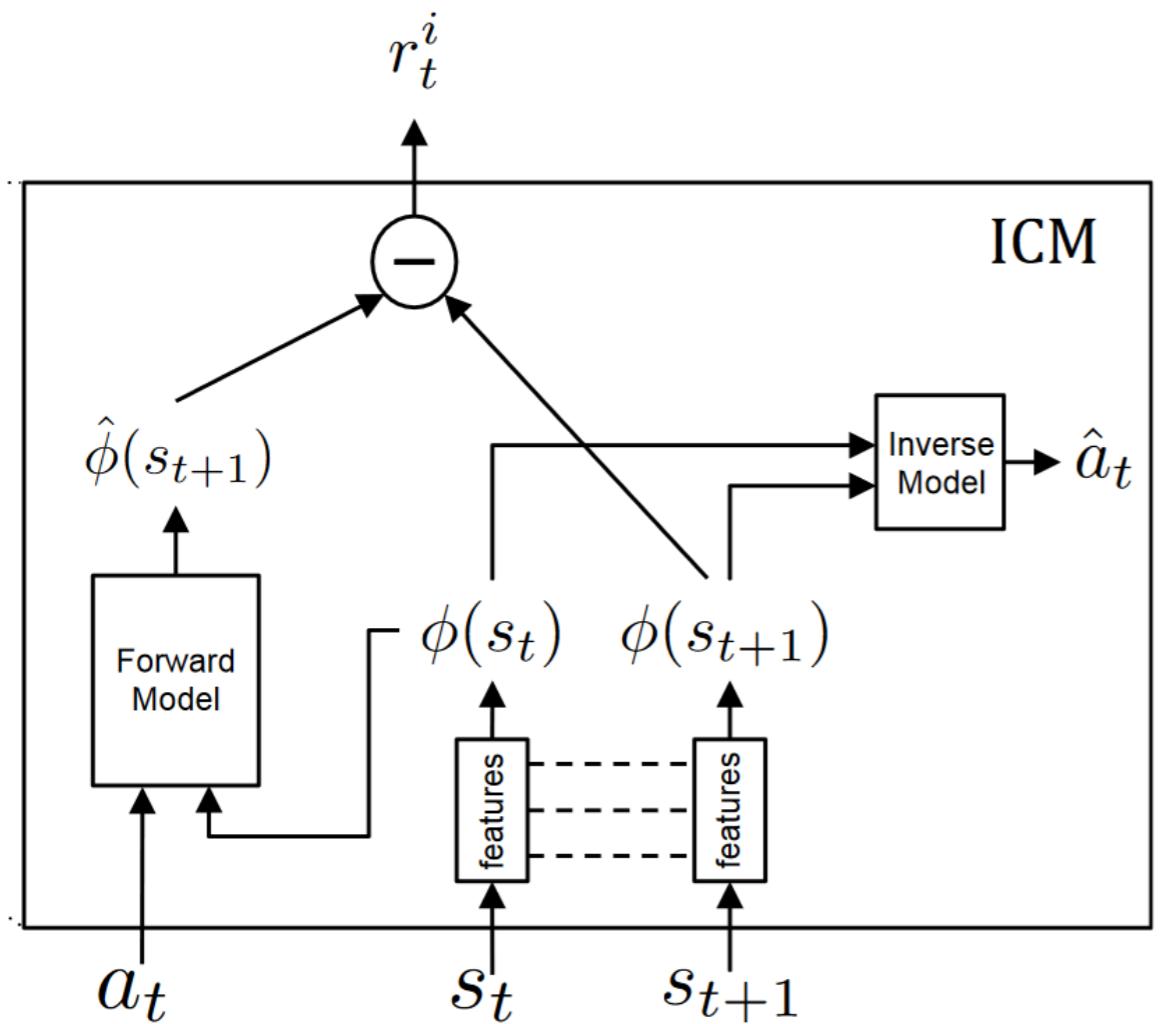
	VAE	IDF	RF	Pixels
Stable	No	No	Yes	Yes
Compact	Yes	Yes	Maybe	No
Sufficient	Yes	Maybe	Maybe	Yes

Inverse Dynamics

- Итоговая награда зависит от поведения двух сетей: первая (**синяя**) преобразует входные изображения в более компактное признаковое пространство; вторая (**красная**) предсказывает следующее состояние



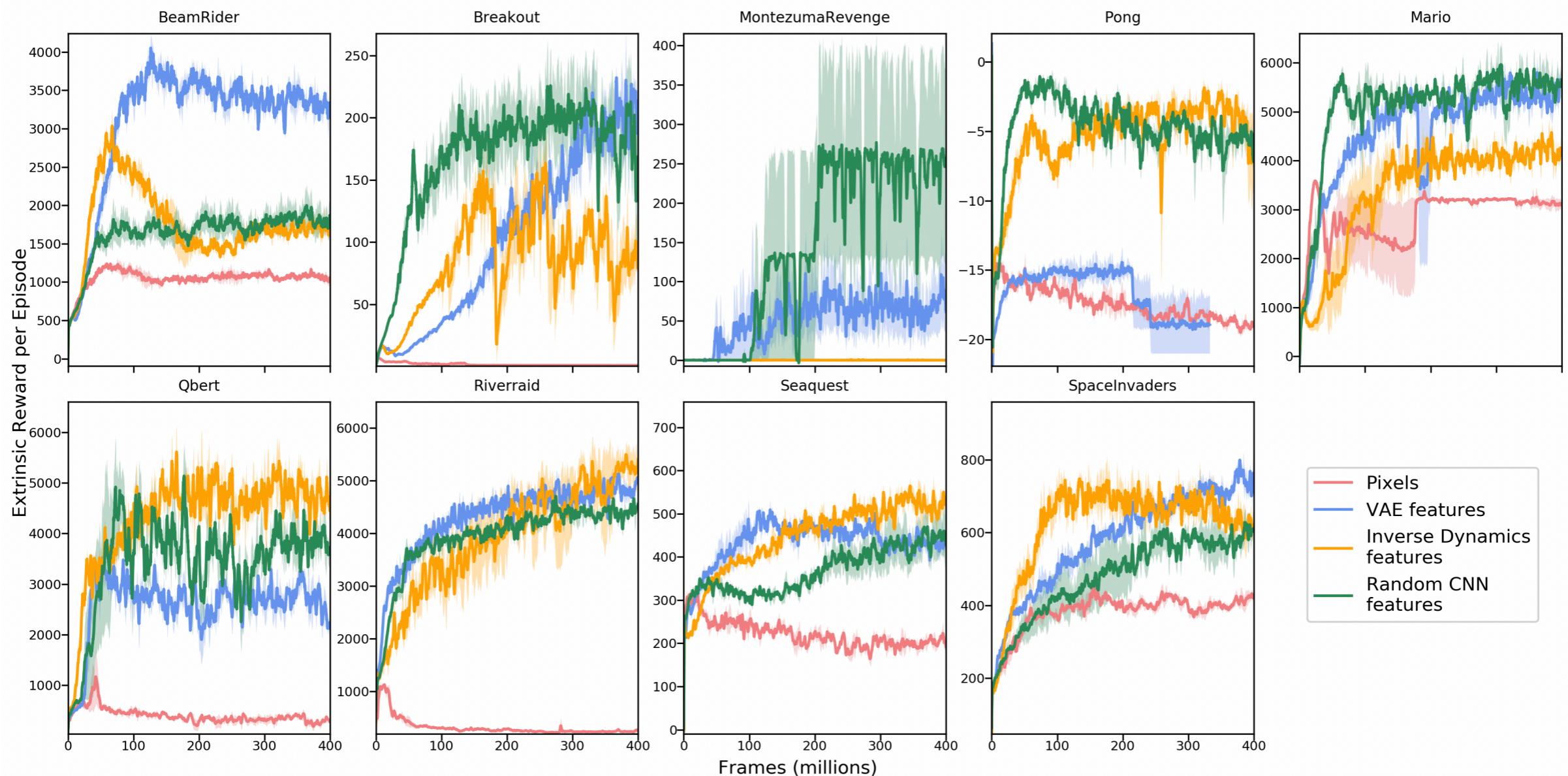
Inverse Dynamics



$$r_t^i = \frac{\eta}{2} \|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$$

$$L_F(\phi(s_t), \hat{\phi}(s_{t+1})) = \frac{1}{2} \|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$$

Эксперименты (Atari & Mario)



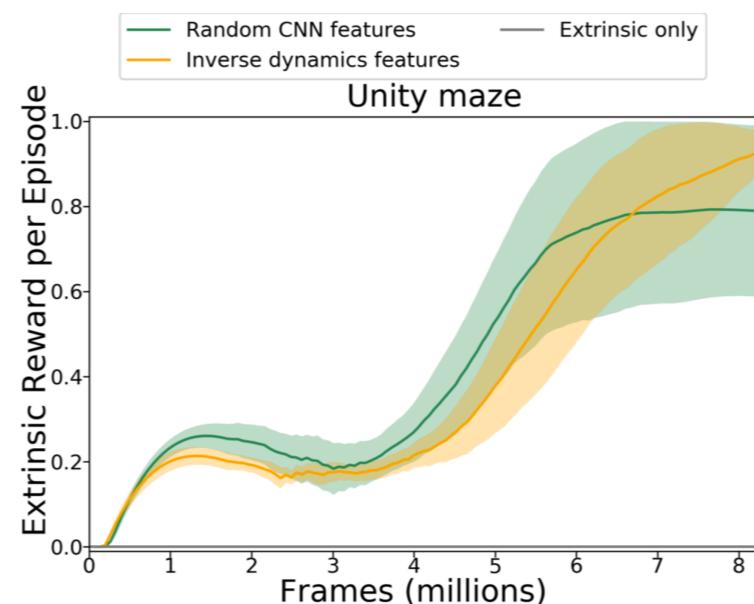
Агент, обученный на чистом любопытстве
(без наград от среды)

Эксперименты (Atari)

- Для большинства игр добавление награды за любопытство приводило к увеличению скорости сходимости и более высокому итоговому качеству:

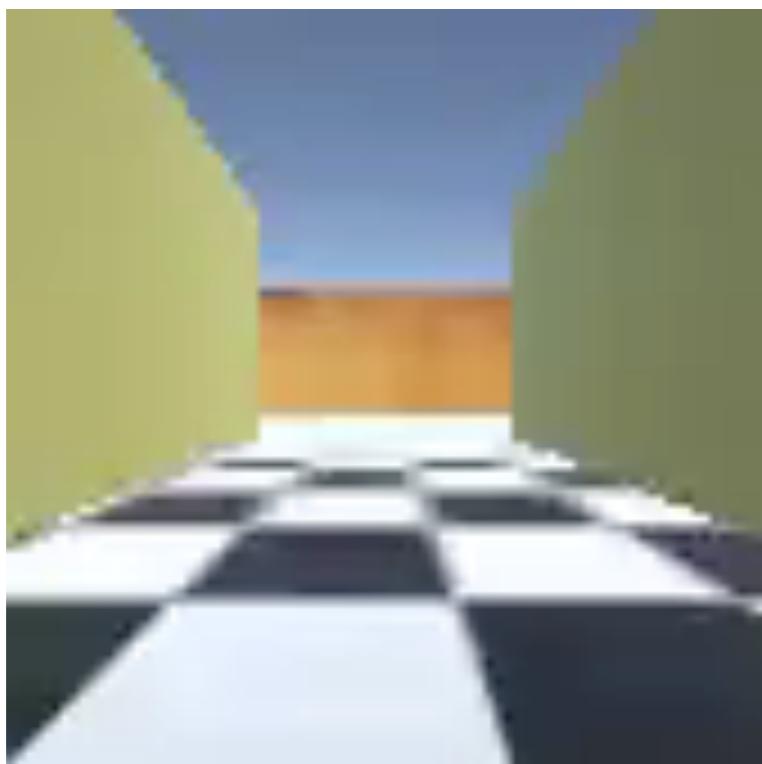
Reward	Gravitar	Freeway	Venture	PrivateEye	MontezumaRevenge
Ext Only	999.3 ± 220.7	33.3 ± 0.6	0 ± 0	5020.3 ± 395	1783 ± 691.7
Ext + Int	1165.1 ± 53.6	32.8 ± 0.3	416 ± 416	3036.5 ± 952.1	2504.6 ± 4.6

- В средах с очень разреженными наградами без curiosity агент совсем неправлялся:

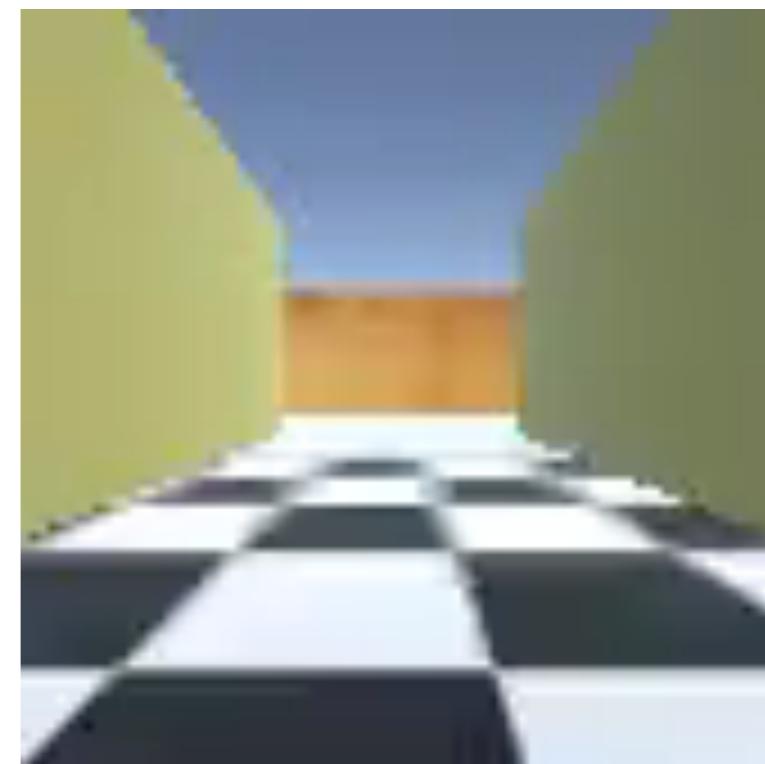


Ограничения

- Но существуют ограничения на степень "предсказуемости" среды
- Если среда абсолютно непредсказуема, то даже идеальная модель не сможет эффективно награждать агента



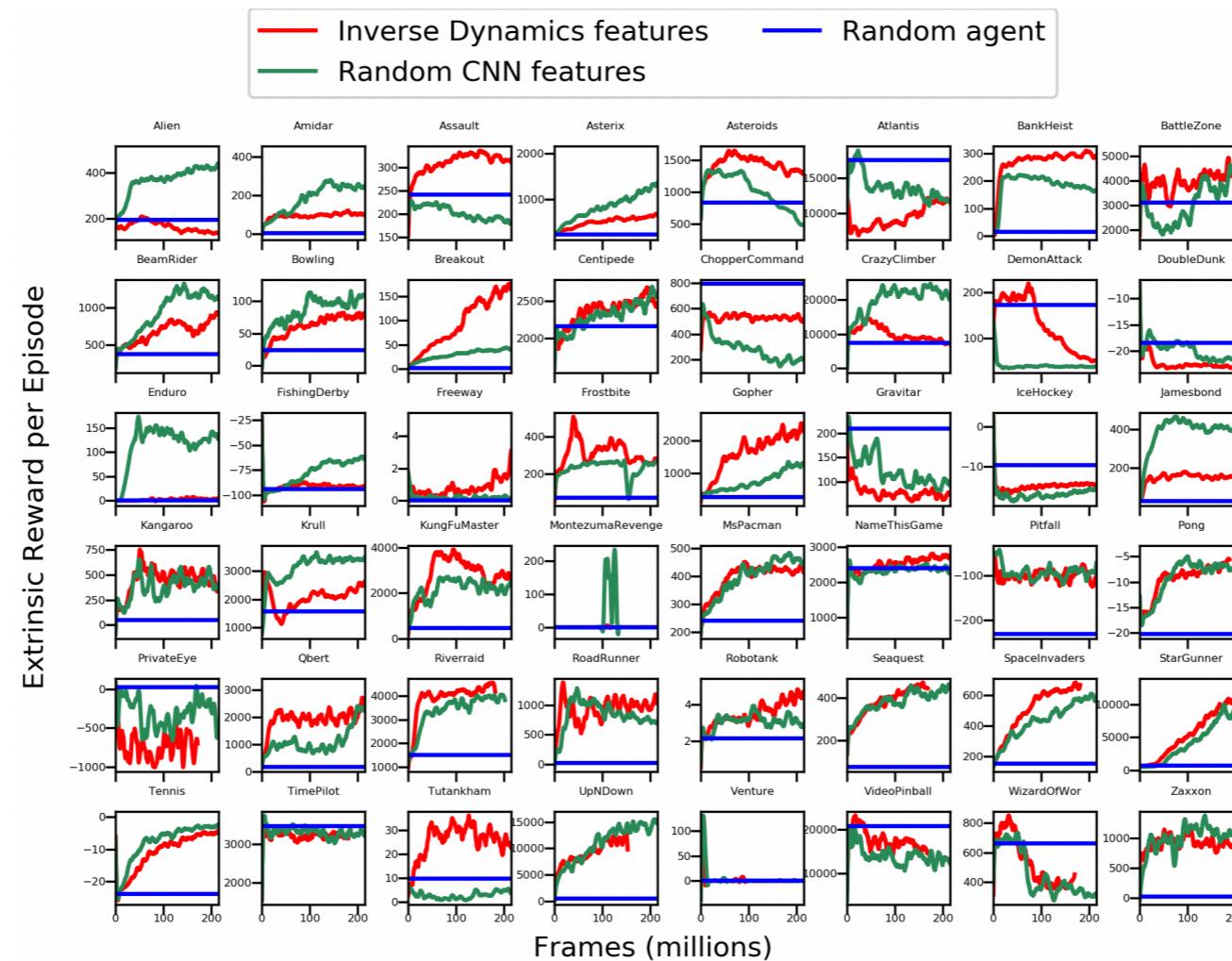
Лабиринт без TV



Лабиринт с TV

Ограничения

- Иногда любопытство не приводит ни к чему хорошему
- В ~30% игр из Atari любопытство не коррелировало с наградами среды (обученный агент играл хуже случайного)



Выводы

- Intrinsic motivation – эффективный подход к моделированию наград в разреженных средах
- Можно совмещать внешние награды из среды и внутренние награды от любопытства для достижения более высоких результатов
- У метода существуют ограничения на область применения

Источники

- <http://www.cs.cornell.edu/~helou/IMRL.pdf>
- <https://pathak22.github.io/noreward-rl/resources/icml17.pdf>
- <https://pathak22.github.io/large-scale-curiosity/resources/largeScaleCuriosity2018.pdf>