"Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos"

Зинов Александр

Высшая Школа Экономики

aazinov@edu.hse.ru

7 декабря 2022 г.

Почему Minecraft?

- Много обучающих видео
- Песочница с открытым миром больше похожа на сценарии из реальной жизни
- Популярная среди исследователей, но очень сложная для машинного обучения игра

Let's play

- Управление через обычный интерфейс – смотрим на кадры, управляем клавиатурой и мышкой
- 20 FPS, никаких макросов
- За счет этого легко собирать данные, идея более универсальна

Но ведь есть Reinforcement Learning!

Не получится исследовать пространство случайным образом:

- Добыть один блок дерева, на который агент уже смотрит – 60 последовательных действий
- Скрафтить доски и верстак – 50 секунд, $\sim$ 970 последовательных действий
- Каменная кирка – 2.3 минуты, $\sim$ 2790 последовательных действий
- Алмазная кирка (это новый рекорд) – 20 минут, $\sim$ 24000 последовательных действий

Ставим лайки, подписываемся на канал

- Можно не тыкаться "вслепую", потому что есть много видео базового геймплея
- Ищем на YouTube летсплеи по заданному списку тегов и скачиваем
- Вручную размечаем 8800 кадров для фильтрации видео с артефактами оверлея (камера, лого и т.п.), режимом, отличным от survival, или платформой, отличной от ПК
- Получается $\sim$ 70k неразмеченных часов видео

Behavioral cloning

- Есть датасет $D = \{(o_i, a_i)\}, i \in \{1 \dots N\}$ – пары (наблюдение, действие)
- Моделируем распределение $\pi \sim p_{BC}(a_t | o_1 \dots o_{\boldsymbol{t}})$ методом максимального правдоподобия
- Нужно очень много данных и времени, потому что модель каузальная

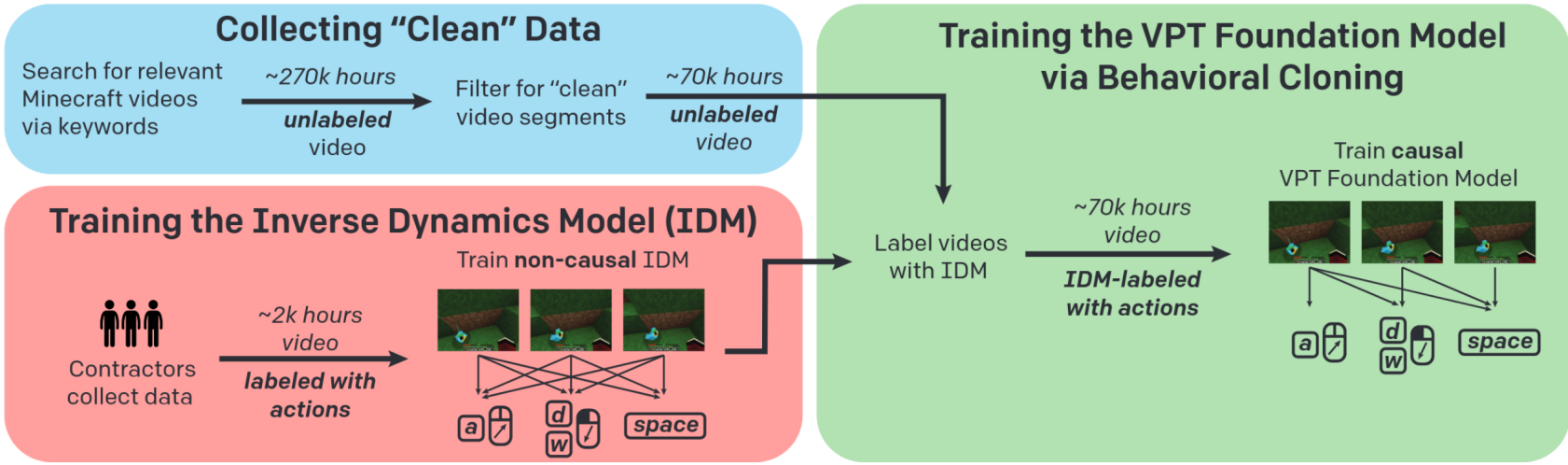
Inverse dynamics model

- Есть датасет $D = \{(o_i, a_i)\}, i \in \{1 \dots N\}$ – пары (наблюдение, действие)
- Моделируем распределение $p_{IDM}(a_t | o_1 \dots o_{\boldsymbol{T}})$ методом максимального правдоподобия
- Нельзя напрямую использовать, как политику, потому что модель не каузальная.
- Значительно проще учится – можно использовать для разметки

VPT Foundation model

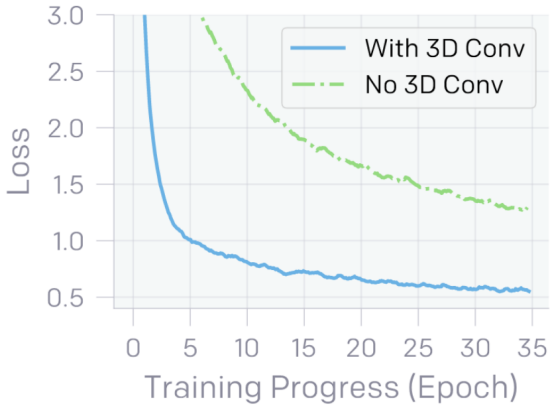
$$\min_{\theta} \sum_{t \in [1 \ldots T]} -\log \pi_\theta(a_t | o_1, \ldots, o_t)$$

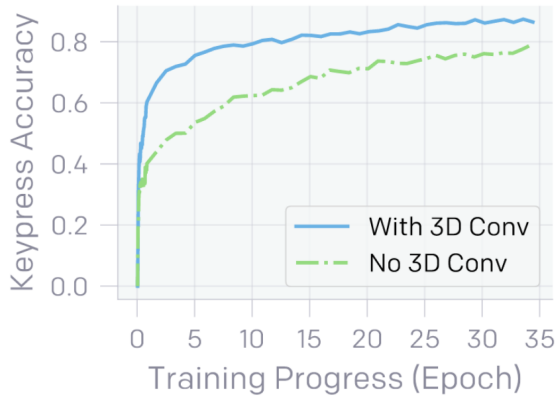
$$\text{where } a_t \sim p_{IDM}(a_t | o_1, \ldots, o_t, \ldots, o_T)$$

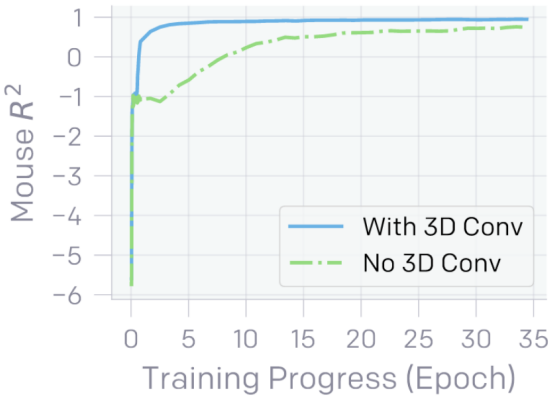


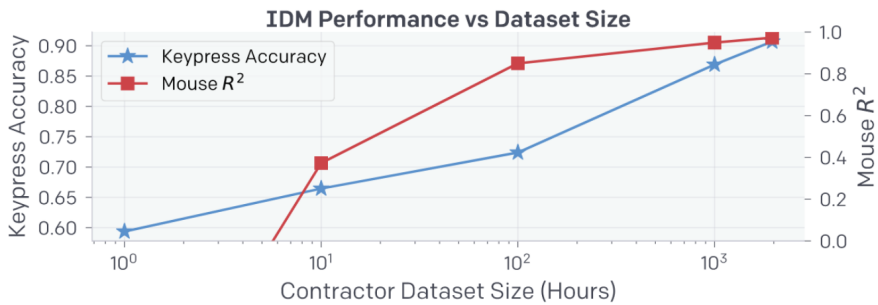
Подробнее про IDM

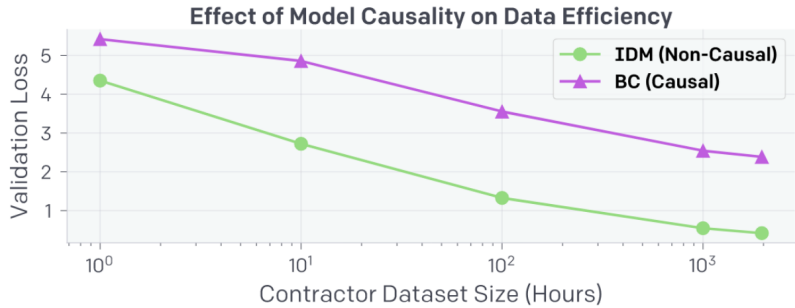
3D свертка → ResNet → перцептрон → трансформер → перцептрон → классификатор



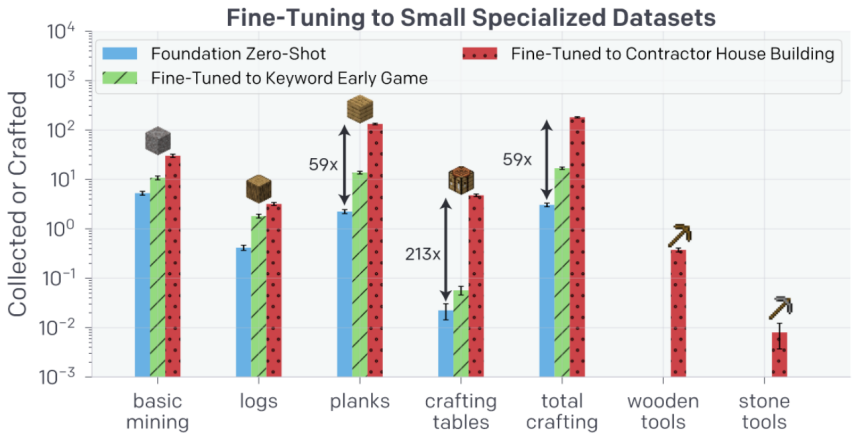


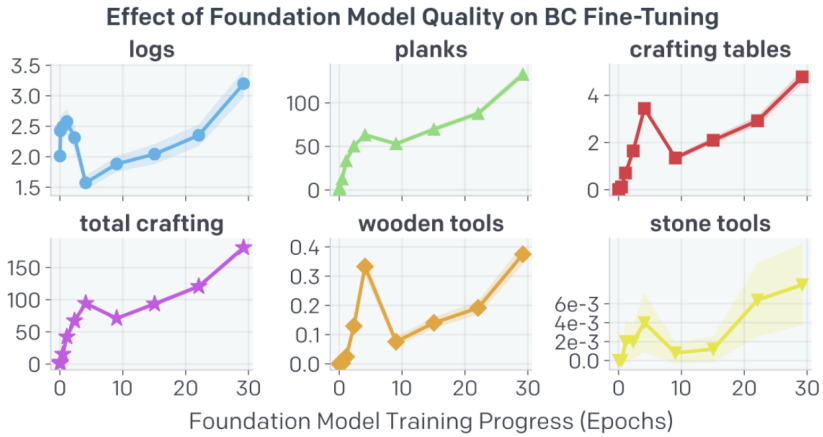




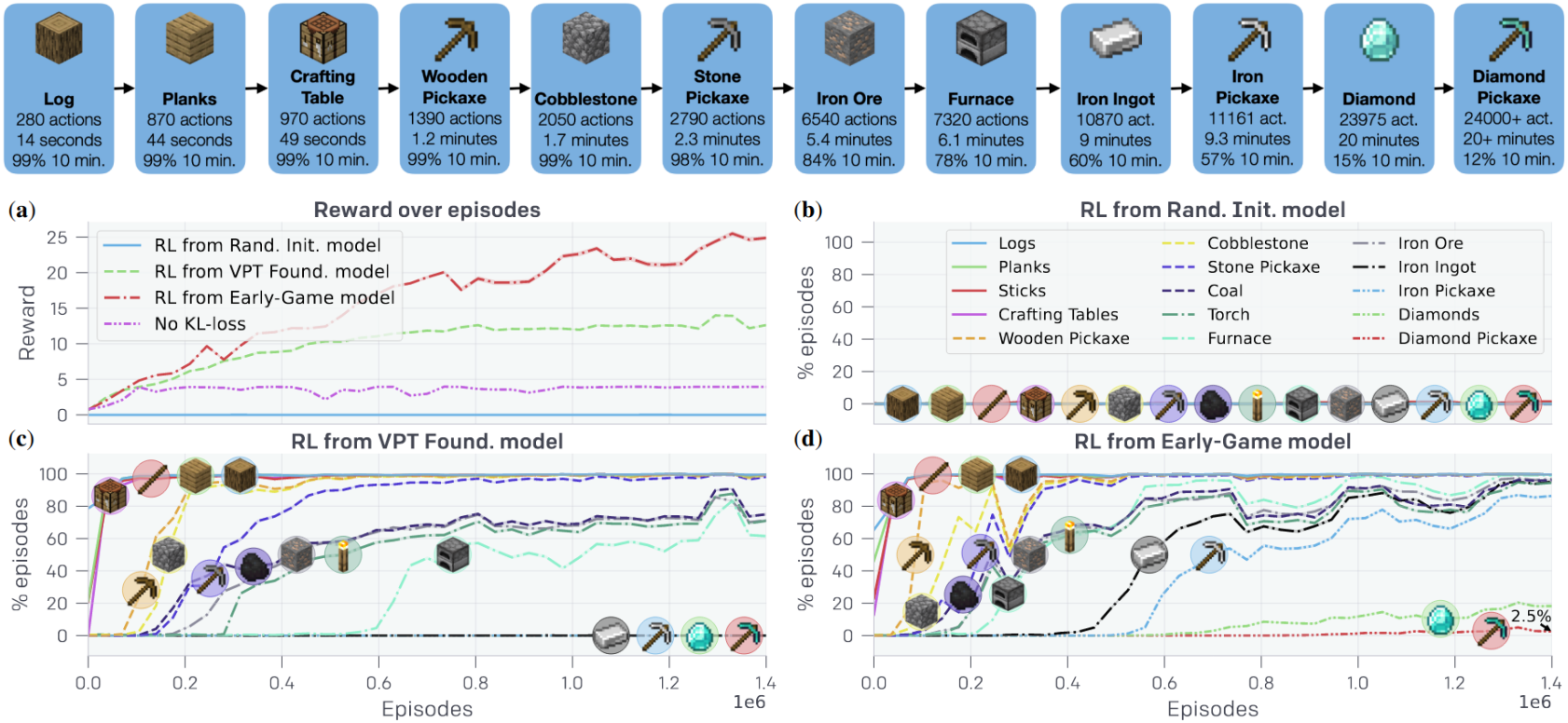


Fine-Tuning with Behavioral Cloning (Zero Shot)





Fine-Tuning with Reinforcement Learning (Long Term)



Заключение

- **Мы не использовали никаких эвристик, связанных с конкретной игрой**
- Аналогичный алгоритм можно использовать для обучения моделей, способных выполнять произвольные задачи в графическом интерфейсе компьютера
- Вопросы из чата