Министерство науки и высшего образования Российской Федерации
Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Магистерская программа
«Методы и технологии искусственного интеллекта»

Разработка робастных методов схвата манипулятором произвольных объектов на основе обучения с подкреплением и методов оптимального управления

(Отчёт по научно-исследовательской работе)

Научный руководитель:
старший научный сотрудник, к.ф.-м.н.
Макаров Д.А.

Студент группы М05-404а:
Синицкий М.И.

Москва 2024

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ 3
2 МЕТОДЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ (Reinforcement Learning). 6
2.1 Имитационное обучение 6
2.2 Разделение стратегий и латентные пространства. 8
2.3 Многоцелевые задачи 10
2.4 Трансформеры и временные зависимости. 12
2.5 Диффузионные стратегии и управление в непрерывном пространстве действий. 14
2.6 Дискретизация непрерывного пространства 15
3 ЗАКЛЮЧЕНИЕ 16
СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ: 18

В данном обзоре исследуются передовые методы обучения с подкреплением (RL) и оптимального управления, применяемые для разработки робастных алгоритмов схвата манипулятора.

Рассматриваются основные теоретические понятия использования латентных пространств и механизмов внимания. Статьи о новейших методах в этой области включают детальную архитектуру и способы реализации, ориентированные на повышение эффективности и адаптивности систем манипуляции.

1 ВВЕДЕНИЕ

Роботы-манипуляторы все чаще используются. Применяются в быту (на кухне, в уборке, стирке), где они могут стать помощниками для облегчения повседневные задач.

Эти системы функциональны и предназначены для выполнения различных операций – от простого захвата предметов до сложных манипуляций, как расфасовка продуктов и приготовление пищи.

Основной целью таких манипуляторов является создание более удобных автоматизированных домашних помощников, что позволит людям освободить время от рутинных дел и сосредоточить своё внимание на учёбе и достижении в профессиональных и творческих навыках.

Методы обучения с подкреплением состоят в том, что у робота есть возможность взаимодействовать с окружающей средой и посредством этого улучшать свои действия, накапливая опыт.

В зависимости от успешности выполнения задач, манипулятор в процессе обучения наблюдает за результатами своих действий и получает положительную или отрицательную корректировку («вознаграждение» и «наказание»). Такой механизм постоянно совершенствуется и адаптируется к динамично изменяющимся условиям.

Обучение с подкреплением позволяет манипуляторам самостоятельно разработать оптимальные стратегии захвата и размещения объектов.

Например,

- Манипулятор может размещать продукты в холодильнике и в кухонных шкафах, учитывая вес, форму, различая долгохранящиеся и скоропортящиеся продукты.
- Алгоритмы движения и планирования позволят роботу оценивать способы размещения предметов и при этом избегать столкновения, учитывая пространство вокруг

В данной обзоре рассмотрим применение этих методов обучения с подкреплением для создания действенных операций в бытовом обиходе.

Особое внимание необходимо уделить разработке алгоритмов, которые обеспечат манипулятору возможность адаптироваться к изменяющимся условиям, убеждаясь в автономности и надежности при выполнении рутинных задач.

Целью является определение наиболее перспективных направлений исследования для разработки робастных методов манипуляции в условии динамично-изменяющейся среды.

Для достижения этой цели были сформулированы следующие задачи:

- Робастные методы захвата манипуляторами в робототехнике являются одной из интереснейших для рассмотрения в обучении искусственного интеллекта
- Развитие технологий RL в оптимальном управлении предлагает множество решений для задач, связанных с динамическими и сложными средами.
- Возможность создания моделей, способных адаптироваться к изменяющимся условиям и способных решать неопределённые моменты в динамике объекта.

- Разработка решений, помогающих снизить сложность при моделировании между физическими характеристиками объектов и движением манипулятора.

Постановка задачи состоит в рассматривании бесконечных процессов принятия решений Маркова (МППР), которые характеризуются набором

$$\left(S, A, T, R, \gamma, p_0\right),$$

где

$S \in R^n$ и $A \in R^m$ – непрерывные пространства состояний и действий;

$T : S \times A \times S \to R_+$ – функция переходов (динамики);

$R : S \times A \to R$ – функция вознаграждения;

$\gamma \in [0,1)$ – коэффициент дисконтирования;

p_0 – начальное распределение состояний.

Необходимо обучить параметрическое отображение $\Pi_\theta : S \to A$ с параметрами θ, такое, что дисконтированный возврат

$$E_{\Gamma \sim \Pi_\theta}\left[\sum_{t=1}^{\infty} \gamma^t r_t\right]$$

где

$r_t \sim R\left(\cdot \middle| s_t, a_t\right)$ достигается максимума вдоль траектории $\Gamma = \left(s_0, a_0, s_1, a_1, \ldots, s_\infty\right)$ следуя за стратегией Π_θ.

Для того чтобы решить задачу улучшения управления, необходимо получить локальное решение задачи корректирования траектории на каждом шаге t, оценивая оптимальные действия $a_{t:t+H}$ на конечном горизонте H и выполняя действие a_t.

2 МЕТОДЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ (Reinforcement Learning).

Существующие на данный момент методы обучения с подкреплением подстраиваются под окружающую среду и анализируют разные подходы, помогая разработке новых методов для улучшения оптимального управления манипулятором.

Рассмотрим принципы и подходы RL, применяемые в задачах манипуляции.

2.1 Имитационное обучение

Имитационное обучение является перспективным в решении задач манипулирования роботами. Его алгоритмы предлагают успешную, с точки зрения обработки данных, основу где приобретаются сенсомоторные навыки на основе небольшого набора демонстраций, часто проводимых непосредственно на реальных роботах.

Имитационное обучение (ИО) действует, используя экспертные данные как эталон. Использование обучающих экспертных примеров позволяют быстрее распознавать и повторять сложные действия. Система корректирует свои действия в реальном времени, минимизирует риск ошибок для критически важных задач. Методы имитационного обучения расширяют навыки для решения долгосрочных задач манипулирования, и ориентированы на длительное взаимодействие при разнообразных поведенческих реакциях.

В статье [1] представлен алгоритм непрерывного имитационного обучения, который позволяет роботу успешно решать задачи

манипулирования. Основная идея заключается в создании постоянно растущей библиотеки навыков на основе последовательности новых задач с использованием модели, которая извлекает навыки в виде повторяющихся шаблонов и обновления существующих (рис.1).

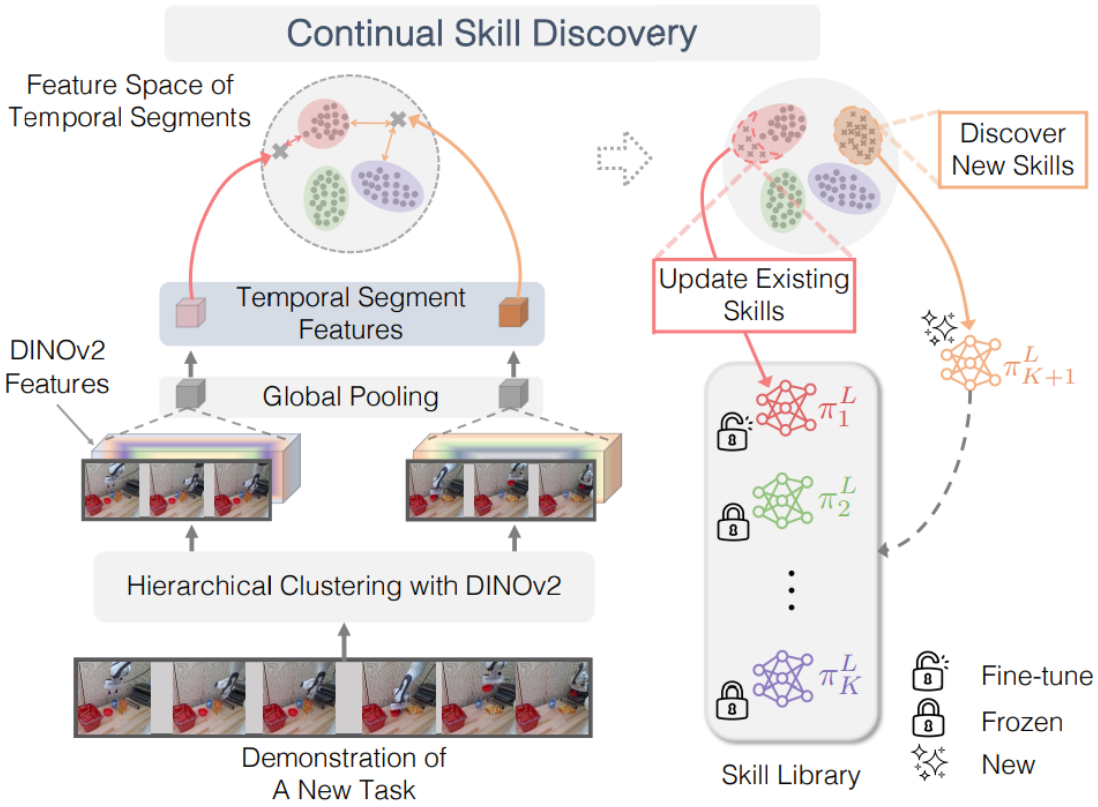


Рис.1: постоянное выявление навыков

Рассматривается проблема непрерывного имитационного обучения для управления роботами в реальном мире. Разрабатывается практический алгоритм, который обучается на основе последовательности новых задач, с которыми может столкнуться робот.

Алгоритм создает постоянно растущую библиотеку навыков, добавляя новые навыки и обновляя старые.

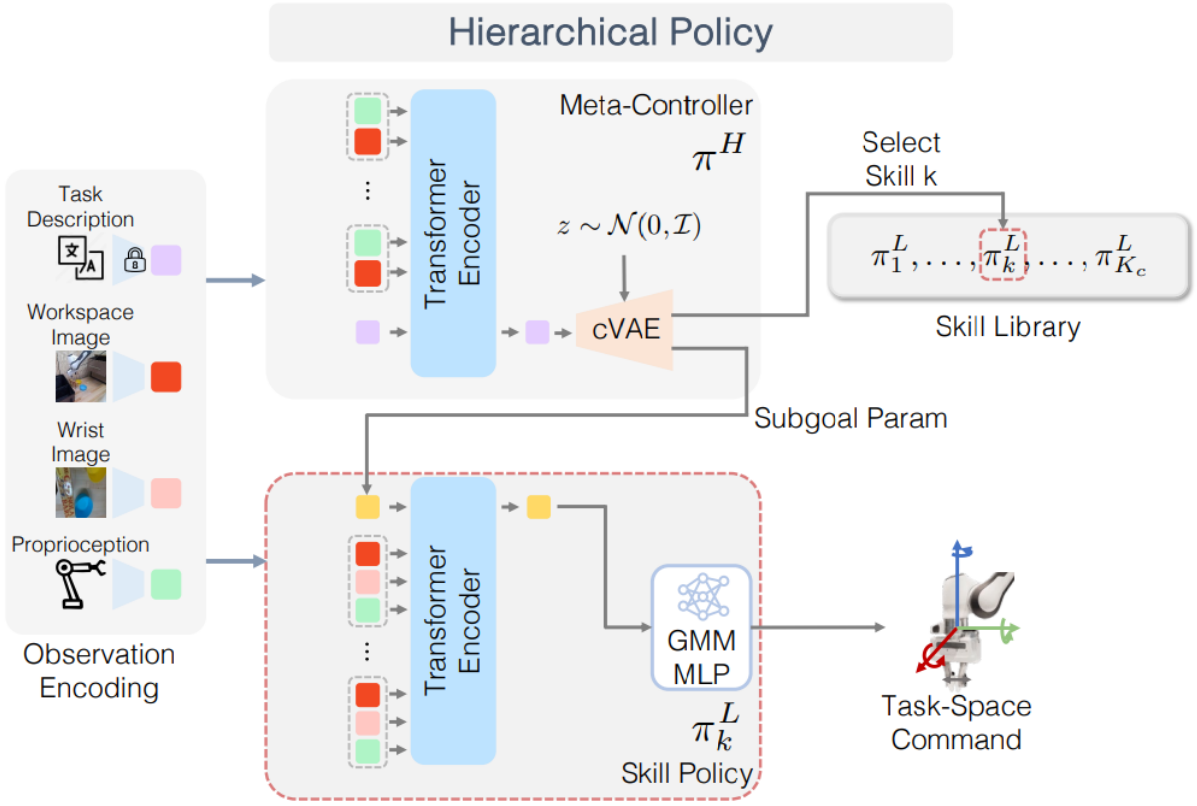


Рис.2: Иерархическая стратегия

Была поставлена цель (рис.2) успешно освоить стратегию, которая использует навыки, полученные из прошлого опыта, для достижения более высокой производительности при выполнении новых задач, сохраняя при этом конкурентоспособные результаты при выполнении ранее изученных задач.

2.2 Разделение стратегий и латентные пространства.

В статье [2], уделяется особое внимание достижениям в комбинировании высокоуровневых и низкоуровневых стратегий с переходом в латентное пространство.

В данной работе рассматривается подход, который предполагает, что высокоуровневая стратегия определяет общий план действий, в то время как низкоуровневая – конкретные шаги для её реализации.

Переход в латентное пространство осуществляется с помощью автоэнкодеров, помогая выявлять скрытые признаки.

Использование такой стратегии требуется где возникает необходимость приспособиться к разнообразным средам, обеспечить быстрый отклик системы управления и минимальную задержку в принятии решений.

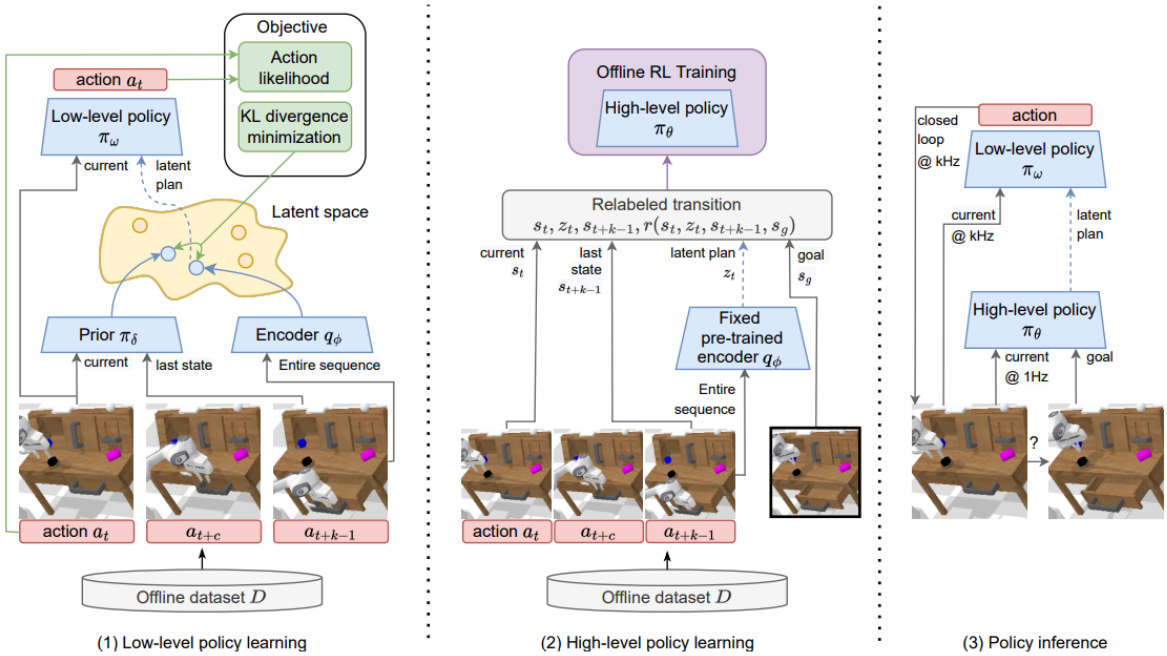


Рис.3: Архитектура обучения высокоуровневой и низкоуровневых стратегий

Стратегии низкого уровня (рис.3-1) распознают и объединяют похожие действия манипулятора (поведения) на основе неразмеченного и неупорядоченного набора данных в пространстве скрытого плана. Стратегия высокого уровня (рис.3-2) создает план, который выделяет области действий для низких стратегий.

Непрерывное пространство примитивных действий изучается с помощью стратегии низкого уровня $\pi_\omega\left(a|s_c,z\right)$

D, который преобразует скрытый план z в соответствующие действия по управлению двигателем a. После обучения можно использовать скрытые планы в качестве пространства действий для стратегии высокого уровня, чтобы научиться переходить из одного низкоуровневого действия в другое

В статье [3] также предлагается автономный иерархический алгоритм имитационного обучения, который может использоваться в сочетании с любым низкоуровневым алгоритмом.

Так как сложные задачи разбиваются на относительно простые с краткосрочным достижением целей и более приспособлены к иерархической структуре. Вместо детерминированного действия используется вероятностная подцель $z \sim q_{\psi}\left(\cdot \middle| s_t, s_{t+c}\right)$.

Энкодер и декодер используются для представления текущего состояния и целевого действия. Модель обучается распределять действия, которые она может достичь из её текущего состояния и понимать, какие из них достижимы.

2.3 Многоцелевые задачи

Исследуется, как несколько независимых агентов могут работать в координации, минимизируя потребности в централизованном управлении. Это снижает риски, связанные с отказами индивидуальных элементов в системе, и позволяет лучше распределять задачи.

Роботизированные системы общего назначения должны обладать широким спектром разнообразных навыков, чтобы быть полезными в решении целого ряда повседневных задач [4].

В этой статье рассказывают об исследовании, как крупномасштабная коллективная роботизированная обучающая система может одновременно осваивать набор моделей поведения, обмениваясь опытом и представлениями о задачах (рис.4).

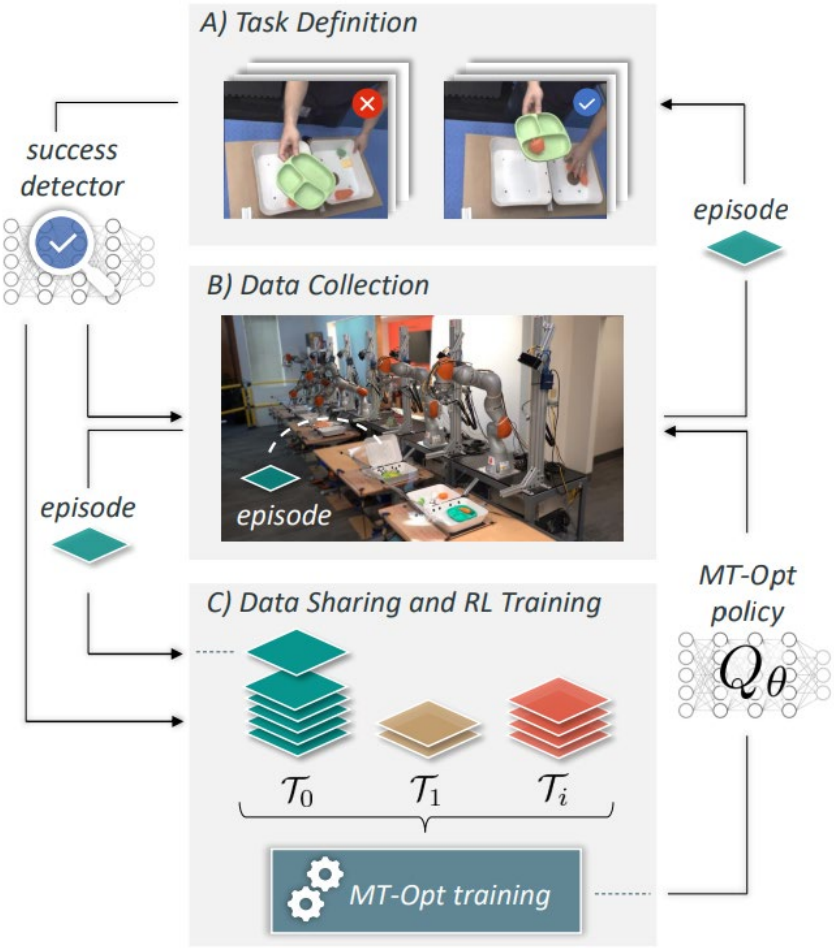


Рис.4: Передача успешных действий между манипуляторами

Новые навыки могут непрерывно создаваться на основе ранее изученных, повышая общую производительность и возможности системы. Метод обучения представленный в статье помогает расширить возможности поставленных задач, таких как выбор предмета из определённой категории, размещение, выравнивание и перестановка их в различных местах:

- Эксперт выявляет примеры успешных задач в группе манипуляторов

- Происходит обмен данными между моделями, что расширяет опыт, используемый для обучения
- Дополнительный опыт может быть использован для определения новых задач и подготовки будущих стратегий

2.4 Трансформеры и временные зависимости.

В архитектуре трансформеров для манипуляций основное внимание уделяется механизму внимания. Трансформеры облегчают согласованность и точность в подходах управления, позволяя учитывать, как предыдущие, так и текущие данные в процессе захвата. Это учитывается для задач, где важна точность, например, при выполнении манипуляций в медицинской робототехнике или при перемещении ценных объектов.

Перспективность применения трансформеров заключается в их способности помочь манипуляторам приспосабливаться к переменным условиям и предсказывать цепочку действий.

Исследование в статье [5], демонстрирует, как трансформеры справляются с моделированием длительных зависимостей во временных рядах данных.

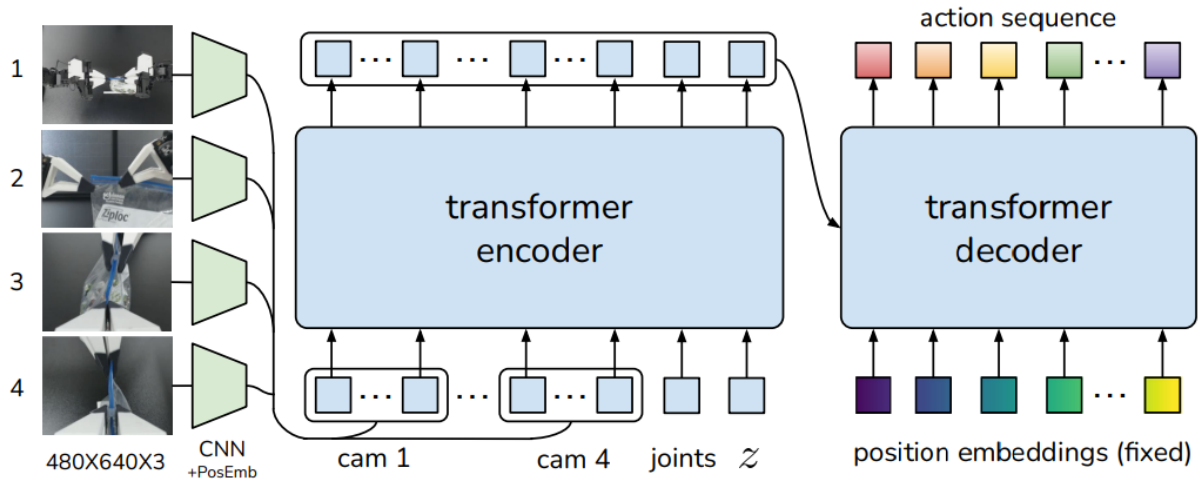


Рис.5: Архитектура разделения действий на последовательности с помощью трансформера

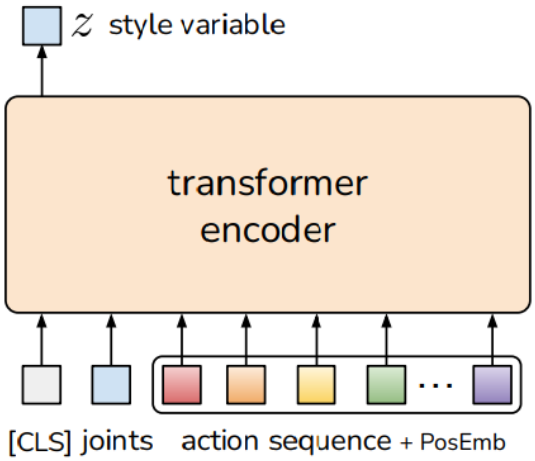


Рис 6: Энкодер, сжимающий последовательность действий, обученный как авто кодировщик

Через каждые k шагов робот получает текущее состояние, генерирует следующие k последовательных действий (рис.5 и рис. 6). Другими словами формула стратегии имеет вид $\pi_\theta\left(a_{t:t+k}\middle|s_t\right)$ вместо $\pi_\theta\left(a_t\middle|s_t\right)$. Это также моделирует немарковское поведение человека.

Применение механизмов внимания в трансформерах улучшает прогноз и координацию задач, особенно в системах, требующих высокой автоматизации и точности, таких как робототехника и автономные транспортные системы.

Исследования, которые касаются направления механизмов внимания, представленные в статье. Рассматриваются методы улучшения модулей внимания для наиболее значимых элементов сложных сигналов.

Внимание применимо в манипуляциях объектов разнообразных размеров, предоставляя механизмы приспособления манипуляторов к точным манёврам и условиям, ухудшающим видимость.

2.5 Диффузионные стратегии и управление в непрерывном пространстве действий.

Методом под названием Diffusion Policy генерируют действия робота, с помощью показаний датчиков. Обычно диффузионные модели используют для генерации изображений.

В статье [6] показано, что использование диффузии для генерации траектории робота преимущественно по сравнению с другими подходами:

- Многомерные пространства действий. Робот планирует будущие действия.
- Постоянное и надёжное обучение. Предполагается, что обучение не требует ручной настройки и манипулятору не придётся искать конкретной оптимальной точки

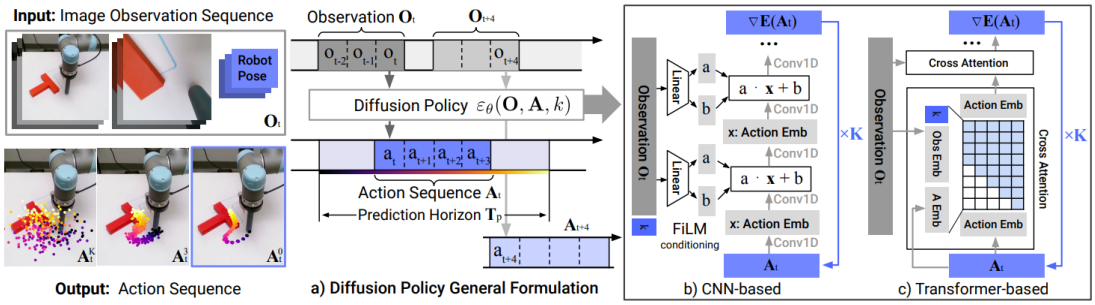


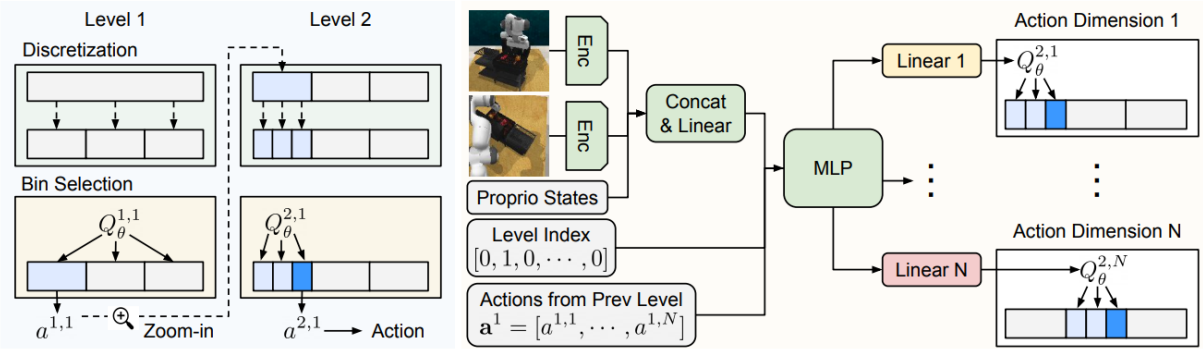
Рис.7: Построение траектории, с помощью диффузионной стратегии

В момент времени t диффузионная стратегия получает T_0 шагов последних наблюдений и выдаёт T_a шагов следующих действий (рис.7).

Это перспективное направление пытается решить ряд вопросов, связанных с недостаточно точной работой при малом количестве демонстрационных данных, а также с большими вычислительными затратами по сравнению с другими методами.

2.6 Дискретизация непрерывного пространства

Идея заключается в обучении агентов, которые дискретизируют непрерывное пространство действий в несколько интервалов и выбирают наиболее лучший (с большим значением Q) [7]. В последующем, выбранный интервал делиться на более мелкие. На последней шаге, выбирается центроид в качестве действия. Сохраняется интервальный путь, чтобы агент понимал в какой области пространства и на каком уровне он находится и куда необходимо увеличить масштаб.



(a) Coarse-to-fine inference procedure (b) Coarse-to-fine critic architecture

Рис.8: Увеличение масштаба дискретизации

В статье утверждается (рис.8), что проблемы управления возникают из-за использования дискретных алгоритмов в непрерывном пространстве, которое они предлагают разделять на несколько интервалов (дискретизировать). Однако разделение на большое количество интервалов ухудшает точность обучения алгоритмов с подкреплением. Поэтому предлагают иерархическое разделение от более крупных к мелким.

3 ЗАКЛЮЧЕНИЕ

Обзоры подчеркивают не только текущие достижения в области RL и оптимального управления, но и описывают возможности их дальнейшего развития и встраивания в робототехнические системы.

Перспективы для будущих исследований включают создание гибридных алгоритмов, комбинирующих самые сильные стороны каждого подхода, а также оценку их применимости в условиях повышенной сложности.

Глубокая интеграция трансформеров и методов децентрализованного управления открывает новые горизонты в создании универсальных, устойчивых операционных систем, что как никогда востребовано при росте разнообразия и сложности задач, стоящих перед современными роботами.

Различные подходы представляют результативные механизмы для увеличения важных характеристик сложных объектов, жертвуя незначительными из них для существенного улучшения производительности. Подразумевается, что использованные подходы помогут роботам-манипуляторам быть не только полезными, но и уметь выполнять свои задачи как можно быстрее.

Объединяя методы машинного обучения с многомерными данными из различных сенсорных потоков, при этом снижая время обработки, чтобы помочь лучшему обучению модели для управления манипуляторами в реальном времени. Это важно при обработки ранее не встречающихся данных для увеличения восприимчивости моделей и позволяет быть независимыми от предварительной обработки.

От улучшенных методов имитационного обучения до использования автокодов и латентных пространств, каждый из рассмотренных подходов способствует разработке более устойчивых, продуктивных и разнообразных решений.

Это создаст предпосылки для разработки более умных и отзывчивых роботизированных помощников, способных активно участвовать в повседневной жизни человека.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ:

1. https://arxiv.org/pdf/2311.02058
2. http://tacorl.cs.uni-freiburg.de/paper/taco-rl.pdf
3. https://www.ijcai.org/proceedings/2023/0469.pdf
4. https://arxiv.org/pdf/2104.08212
5. https://arxiv.org/pdf/2304.13705
6. https://www.roboticsproceedings.org/rss19/p026.pdf
7. http://younggyo.me/cqn/static/paper/cqn_paper.pdf