УДК 004.852

Сиамский визуальный трансформер с перекрёстным вниманием для инверсного моделирования действий в визуальной среде

Дьячков Д.А. (Университет ИТМО, ООО «SpaceThink»)

Левин А.А. (Университет ИТМО, ООО «SpaceThink»)

Кудашов А.С. (ООО «SpaceThink»)

Научный руководитель – к.т.н. Ефимова В.А. (Университет ИТМО)

Введение. В последние годы методы обучения с подкреплением (Reinforcement Learning, RL) и трансформерные архитектуры продемонстрировали высокую эффективность в задачах управления агентами, работающими в визуальной среде. Классический подход к обучению агентов взаимодействию с окружающей средой был сформулирован в рамках марковских процессов принятия решений (MDP) и градиентных методов политики, таких как REINFORCE [1]. Параллельно развитие архитектур самовнимания, начиная с работы «Attention Is All You Need» [2], и последующее появление Vision Transformer (ViT) [3] создали мощный инструмент для извлечения высокоуровневых признаков из изображений.

Несмотря на это, задача планирования параметризованных действий, эффект которых непосредственно проявляется в изменении растрового изображения (например, смещение, поворот или масштабирование объекта), остаётся недостаточно формализованной.

Основная часть. В данной работе предлагается метод планирования действий агентных систем на основе глубоких нейронных сетей (далее агентов), работающих с визуальными данными и совершающих действия, эффект которых проявляется в растровых изображениях.

Предлагаемый метод заключается в двухэтапном обучении сети ViT. На первом этапе обучается сиамская сеть ViT, состоящая из сетей Net1 и Net2, одной из которых (Net1) подаётся исходное изображение (I_before), а второй сети (Net2) – целевое изображение (I_after), которое необходимо получить, совершив действие. Каждая из сетей предсказывает действие (одно из какого-то количества) и n параметров для этого действия. К примеру, сеть может предсказать перемещение курсора вверх, а параметром будет количество пикселей, на которое необходимо переместить курсор. Net1 и Net2 являются копиями друг друга и имеют одинаковые веса. В процессе обучения Net1 и Net2 имеют единый градиент и обучаются параллельно. Изображения I_before и I_after обрабатываются параллельно в рамках одного батча на протяжении всего прямого прохода, за исключением слоёв cross-Attention, на которых единый батч разделяется на два части (I_before и I_after). Для сети Net1 матрица Q вычисляется на основе I_before, а матрицы K и V вычисляются на основе соответствующего изображения I_after. Для сети Net2 cross-attention вычисляется соответствующим способом. После блоков ViT полученное скрытое представление (отдельно для I_before и I_after) передается в голову классификации (какое действие необходимо совершить) и в голову для предсказания параметров этого действия.

Ошибка состоит из двух компонентов:

- Ошибка RL (на основе награды, полученной из среды) (Supervised learning по псевдо-таргету, полученному на основе награды от среды для предсказания типа действия (*) и Monte Carlo Policy Gradient with Action Enumeration (**))
- Ошибка между двумя предсказаниями голов (для I_before и I_after). Для головы классификации это мера похожести между предсказаниями классов действий, а для головы предсказания параметров это степень похожести для симметричности действий, т.е. если Net1 предсказала перемещение курсора мыши по вертикали на 50

пикселей, то Net2 должна предсказать обратное число, то есть перемещение курсора мыши по вертикали на -50 пикселей.

$$\min \mathrm{CE}(\pi\theta(\mathrm{s}), \mathrm{argmax}[\mathrm{R}(\mathrm{s},\mathrm{a})]) \quad (*)$$

$$\nabla\phi J \approx K1k = 1\sum KM1m = 1\sum MRk, m\nabla\phi log\pi\phi(\theta k, m \mid s, ak)$$

После первого этапа обучения сети, описанного выше, происходит второй этап обучения, на котором сеть перестает быть сиамской, а выходы cross-attention принимают информацию от LLM в виде скрытого представления, по своей сути аналогичной текстовому промпту с указанием ожидаемых действий. Таким образом второй этап обучения представляет из себя адаптацию к новой модальности данных из cross-attention.

Выводы. В работе предложен метод обучения агентной системы, сочетающий сиамскую архитектуру Vision Transformer и методы обучения с подкреплением для восстановления параметризованных действий в визуальной среде. В отличие от стандартной схемы RL, где политика обучается исключительно на основе скалярной награды [1], в предлагаемом подходе используется структурированная награда. Использование механизма cross-attention между состояниями до и после действия расширяет классическую ViT-архитектуру [3] и обеспечивает явное моделирование оператора перехода в пространстве скрытых представлений. Дополнительная симметричная регуляризация между Net1 и Net2 формирует инвариантность к направлению преобразования (прямое/обратное действие), что можно интерпретировать как приближение к обучению латентного оператора динамики среды.

Представленный подход может быть применён в задачах автономного управления интерфейсами, визуального манипулирования объектами и обучении embodied-агентов, где требуется планирование действий, непосредственно изменяющих пиксельное представление среды.

Список использованных источников:

[1] Williams R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine Learning. — 1992. — Vol. 8, № 3–4. — P. 229–256.

[2] Vaswani A., Shazeer N., Parmar N., et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). — 2017. — Vol. 30. — P. 5998–6008.

[3] Dosovitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale // International Conference on Learning Representations (ICLR). — 2021. — 16 p.