УДК 004.896

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО АГЕНТА ДЛЯ ЭФФЕКТИВНОГО УПРАВЛЕНИЯ МОБИЛЬНЫМ РОБОТОМ С ПРИМЕНЕНИЕМ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Лапин М.В. (ИТМО)

Научный руководитель – доцент, кандидат физико-математических наук Трифанов А.И. (ИТМО)

Введение. В современном мире представлено множество различных робототехнических моделей, большая часть из которых предназначена для выполнения строго запрограммированных и ограниченных задач. Тем не менее на сегодняшний день существует потребность в разработке робота, способного выполнять ряд неявных комбинированных задач, которые строго не зафиксированы и не описаны. Для достижения данной цели предлагается использовать методы машинного обучения и искусственного интеллекта в процессе моделирования поведения робота. Это позволит создать систему, способную самообучаться и адаптироваться к изменяющимся условиям, а также успешно воспринимать сложные инструкции и принимать автономные решения, основываясь на контексте и поставленных целях. В последние годы разработке агентов искусственного интеллекта для управления мобильными роботами уделяется значительное внимание в связи с их потенциальным применением в различных областях. Основные работы, развивающие данное направление:

1. Eureka [6];
2. PaLM-E [3];
3. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances [1];
4. DeepMimic [8].

Первые три представленные выше модели объединяют RL (обучение с подкреплением) и LLM (большие языковые модели) подходы для достижения SOTA результатов. Последняя работа основывается на RL и имитационном обучении, что позволяет роботам выполнять реалистичные движения, наподобие человеческих. При этом на текущий момент нет работ, которые пытались бы внедрить мультимодальные диалоговые модели с генерацией движения в RL алгоритмы.

Основная часть. Для эффективного управления роботом агент должен иметь возможность распознавать и классифицировать объекты в окружающей среде, обнаруживать препятствия и принимать решения о дальнейших действиях на основе этой информации, распознавать звуки и речь, а также понимать взаимосвязь между IMU и изображением с камер робота. В связи с чем довольно интересным выглядит направление по применению мультимодальных диалоговых моделей с единым эмбеддинговым пространством для:

- текста;
- звука и речи;
- изображения;
- видео;
- карты глубины;
- генерации движения;
- IMU.

Одним из потенциально успешных подходов является интеграция подобной мультимодальной диалоговой LLM в dreamerV3 [5] в качестве компонента отвечающего за сбор и объединение всех доступных агенту наблюдений и прогнозирование следующего состояния мира. Это позволит создать более гибкую и адаптивную робототехническую систему, способную эффективно взаимодействовать с окружающей средой и принимать решения на основе совокупности различных модальностей. Тем самым это открывает новые возможности для развития автономных роботов, способных адаптироваться к различным

ситуациям и выполнять сложные задачи.
В данной работе разрабатывается мультимодальная диалоговая модель на основе предобученной модели из семейства GPT [2]. В базовую модель было добавлено несколько компонентов для улучшения процесса генерации, а именно:

- предобученная ImageBind [4], которая предоставляет совместные эмбеддинги для текста, изображений, звука и IMU;
- VQ-VAE [7] - вариационный автоэнкодер с дискретным векторным квантованием для описания движения человека путем преобразования трехмерного движения в токены движения.

В дальнейшем обученная мультимодальная модель внедряется в dreamerV3 в качестве входного компонента, отвечающего за обработку сырых данных, приходящих агенту на вход.

Выводы. В результате проведенной работы разработан агент для виртуальной модели робота, который может распознавать и выполнять набор неявных команд через текстовый интерфейс. Данная работа может быть полезна в области на пересечении NLP и робототехники и имеет потенциал для применения в различных сферах деятельности, таких как управление умным домом, образование, медицина и промышленность. Внедрение мультимодальной диалоговой модели в RL-агента, управляющего роботом, может упростить координацию и планирование пути в сложных средах со множеством недетерминированных препятствий за счет получения единого векторного представления о всех наблюдениях из окружающей среды. Также это может упростить взаимодействие между человеком и робототехнической системой.

Список использованных источников:

1. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K. and Herzog, A., 2022. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691.
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.
3. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T. and Huang, W., 2023. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378.
4. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A. and Misra, I., 2023. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15180-15190).
5. Hafner, D., Pasukonis, J., Ba, J. and Lillicrap, T., 2023. Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104.
6. Ma, Y.J., Liang, W., Wang, G., Huang, D.A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L. and Anandkumar, A., 2023. Eureka: Human-level reward design via coding large language models. arXiv preprint arXiv:2310.12931.
7. Oord, A.V.D., Vinyals, O. and Kavukcuoglu, K., 2017. Neural discrete representation learning. arXiv preprint arXiv:1711.00937.
8. Peng, X.B., Abbeel, P., Levine, S. and Van de Panne, M., 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG), 37(4), pp.1-14.