УДК 004.896, 004.932.2

Локализация по графу сцены

Т. Ю. Кондрашов, Д. А. Юдин, А. А. Мелехин
Московский физико-технический институт (Национальный исследовательский университет)
Центр Когнитивного Моделирования МФТИ

Задача визуального определения местоположения (Visual Place Recognition) является одной из ключевых для локализации и навигации автономных мобильных роботов. Использование визуальных сенсоров обладает преимуществом низкой стоимости и доступности, однако они уязвимы к различным видам изменений сцены - включая изменение ракурса камеры, освещения и изменение внешнего вида отдельных объектов.

Одним из решений указанных проблем является алгоритм SegVLAD [1], который формирует эмбеддинги не для всего изображения целиком, а для его отдельных сегментов. Метод строит граф на основе объектов, распознанных на изображении, где вершины соответствуют объектам, а рёбра строятся триангуляцией Делоне между соответствующими центроидами. Для каждого изображения выделяются сегменты - подмножества вершин графа, находящиеся на ограниченном расстоянии от выбранного центра. Далее для задачи сопоставления используется стратегия поиска ближайших эмбеддингов сегментов: эмбеддинги сегментов изображения-запроса сравниваются с эмбеддингами всех сегментов в базе, после чего отбираются наиболее похожие. Полученные кандидаты по локациям агрегируются, и итоговая оценка похожести между изображением-запросом и каждым кандидатом вычисляется как сумма сходства между соответствующими сегментами. Такой подход позволяет повысить устойчивость к локальным изменениям сцены, включая перемещения объектов и частичные окклюзии, а также обеспечивает более надёжное сопоставление при изменении ракурса камеры.

В оригинальной реализации SegVLAD визуальные признаки (visual features) извлекаются с помощью модели DINOv2 [3], а агрегация признаков осуществляется методом VLAD (Vector of Locally Aggregated Descriptors) (аналогично статье [4]).

В данной работе предпринимается попытка модификации SegVLAD, заключающаяся в замене связки DINOv2 + VLAD на более современный подход - Bag of Queries [2]. Bag of Queries опирается на набор глобальных, обучаемых запросов, которые более эффективно извлекают признаки с помощью механизма cross-attention.

Но в отличие от реализации с DINOv2, Bag of Queries не позволяет получать попиксельные признаки, поэтому в предоставляемом методе для каждого сегмента ищется bounding box, который потом целиком подается в Bag of Queries.

В докладе будут предоставлены результаты тестирования на открытом датасете 3RScan и датасете для задачи визуального определения местоположения, предоставленном центром робототехники Сбер.

Список литературы

[1] Garg K. et al. Revisit Anything: Visual Place Recognition via Image Segment Retrieval //European Conference on Computer Vision. – Cham : Springer Nature Switzerland, 2024. – С. 326-343.

[2] Ali-Bey A., Chaib-draa B., Giguère P. BoQ: A place is worth a bag of learnable queries //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2024. – С. 17794-17803.

[3] Oquab M. et al. Dinov2: Learning robust visual features without supervision //arXiv preprint arXiv:2304.07193. – 2023.

[4] Keetha N. et al. Anyloc: Towards universal visual place recognition //IEEE Robotics and Automation Letters. – 2023. – Т. 9. – №. 2. – С. 1286-1293.