УДК 004.93

Повышение эффективности больших визуально-языковых моделей для обнаружения объектов на изображениях без обучения

Хуссейн Авин (ИТМО)
Научный руководитель – кандидат технических наук, доцент, Кашевник А.М. (ИТМО)

Введение. Современные большие визуально-языковые модели (VLM) демонстрируют высокие результаты в вопросно/ответных задачах работы с изображениями, а также генерации описаний изображений. Однако, они часто показывают ограниченную эффективность в задачах обнаружения объектов на изображениях при ограниченности ресурсов. Несмотря на развитые мультимодальные возможности VLM, большая часть из них не обладают механизмом прямого сопоставления текстовых понятий с областями изображения без дополнительной адаптации под конкретную задачу. Существующие подходы, как правило, добавляют специализированные детекционные декодеры или переобучают их для каждой конкретной модели, что снижает их обобщающую способность и масштабируемость. В зарубежной и отечественной практике активно исследуются методы использования мультимодальных представлений и адаптации моделей, однако универсальные, мало ресурсоёмкие механизмы декодирования остаются недостаточно изученными. Научная проблема данного исследования заключается в разработке универсального механизма декодирования, позволяющего выполнять обнаружение объектов с использованием предобученных VLM без необходимости обучения под конкретную модель.

Основная часть. Детекторы с открытым словарём формируют эмбеддинги классов с использованием текстового энкодера (например, CLIP или BERT) [1], что позволяет обнаруживать категории, отсутствующие в обучающей выборке. Однако, использование полного текстового энкодера на этапе инференса значительно увеличивает вычислительную сложность. В случае фиксированного набора целевых классов, например 80 классов COCO 2017, эти эмбеддинги можно аппроксимировать более эффективным способом. Модели VLM, такие как Qwen3-VL-4B [2], имеют семантически насыщенный слой эмбеддингов, отображающий текстовые токены в непрерывное векторное пространство. Предлагаемый метод использует это свойство: эмбеддинги названий классов извлекаются из слоя эмбеддингов VLM и сопоставляются с текстовым пространством эмбеддингов детектора на основе модели DETR через использование линейных проекций, которое выполняется на фиксированном наборе целевых классов с использованием регуляризованной задачи наименьших квадратов для вычисления оптимального линейного преобразования между пространствами эмбеддингов. Линейная проекция заменяет исходный текстовый энкодер на этапе инференса. Данный подход является оптимальным с точки зрения вычислительной эффективности, оригинальным благодаря центрированию пространств эмбеддингов и экономически целесообразным, так как снижает нагрузку на память и время выполнения инференса, сохраняя возможность обнаружения новых объектов.

Выводы. В результате исследования создан универсальный механизм декодирования, обеспечивающий обнаружение объектов с использованием предобученных VLM без переобучения под конкретную модель. Практическая значимость работы заключается в возможности оснащения любой VLM модели разработанным механизмом, что позволяет выполнять задачи обнаружения объектов и ответы на запросы вида «найди объект на изображении». В рамках дальнейших исследований планируется разработка рабочего прототипа детектора и проверена его эффективность качественным анализом обнаружения объектов. Реализацию и тестирование планируется провести на валидационном наборе COCO 2017 с

визуальной проверкой обнаруженных объектов, а при наличии разметки с вычислением количественных метрик, таких как mAP@0.5.

Список использованных источников:

1. Wang J., Chen B., Kang B., Li Y., Chen Y., Xian W., Chang H., Xu Y. Ov-DQUO: Open-vocabulary DETR with denoising text query training and open-world unknown objects supervision // arXiv preprint. 2024. arXiv:2405.17913.

2. Wang P., Bai S., Tan S., Wang S., Fan Z., Bai J., Chen K., Liu X., Wang J., Ge W., Fan Y., Dang K., Du M., Ren X., Men R., Liu D., Zhou C., Zhou J., Lin J. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution // arXiv preprint. 2024. arXiv:2409.12191.