УДК 004.89

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ ПРОМЕЖУТОЧНЫХ ПРЕДСТАВЛЕНИЙ ВИЗУАЛЬНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ (VLM) ДЛЯ ЗАДАЧ СЕГМЕНТАЦИИ И ДЕТЕКЦИИ ОБЪЕКТОВ С МАЛЫМ КОЛИЧЕСТВОМ ЭКЗЕМПЛЯРОВ НА СЛОЖНО СТРУКТИРОВАННЫХ ИЗОБРАЖЕНИЯХ

Топольницкий А.А. (ИТМО)

Научный руководитель – кандидат физико-математических наук, доцент Михайлова Е.Г.
(ИТМО)

Введение. Современные методы обработки изображений сталкиваются с вызовами, связанными со сложной структурой визуальных данных. Объекты с нестандартными формами, перекрытиями и разнообразными условиями освещения затрудняют точный анализ и распознавание. Поэтому возникает необходимость в разработке алгоритмов, способных эффективно работать с такими сложными изображениями, обеспечивая высокую точность и надежность результатов. Ещё одной проблемой является малое количество таких объектов и их уникальность, что затрудняет создание систем с высокой обобщающей способностью. К подобным объектам можно отнести медицинские снимки, произведения искусства, архитектурные элементы, научные и технические изображения, а также спутниковые фотографии.

Основная часть. В качестве объекта исследований выступают изображения архитектурных сооружений города Санкт-Петербурга и целью исследование стоит применение представлений из визуальных языковых моделей для выявления стадий эрозии и разрушений на фасадах сооружений. Большинство современных методов, решающих схожую задачу, имеют три основных ограничения:

1. Работа с очень крупными изображениями, содержащими незначительную часть фасада;
2. Работа и определение крайне заметных и явных повреждений, например больших трещин или трещин, которые занимают значительную долю изображений;
3. Работа с размеченными и собранными вручную наборами данных [1].

Также необходимо отметить, что при упомянутых выше ограничениях для достижения высокой точности и скорости обработки изображений вполне достаточно широко известных нейросетей из семейства YOLO [2] и SegFormer [3] и их модификаций. Однако для того, чтобы избавиться от вышеупомянутых ограничений имеет смысл рассмотреть другие подходы.

Одним из возможных подходов, который может позволить избавиться от ограничений при сохранении высокой точности, выступает метод на основе использования больших языковых моделей, в частности визуальных языковых моделей или VLM [4]. Vision-Language Models (VLM) объединяют обработку изображений и текста, создавая обобщённые скрытые (промежуточные) представления, позволяющие моделям понимать связь между визуальными и языковыми данными. В связи с широким развитием больших языковых и визуальных языковых моделей широкое распространение получил подход, называемый обучением с открытым словарём (Open Vocabulary Learning), который позволяет моделям распознавать и классифицировать объекты на изображениях без ограниченного набора классов, заданного на этапе обучения. Это достигается за счёт использования текстовых описаний объектов и мультимодальных представлений, позволяющих моделям интерпретировать новые категории через их семантическое сходство с уже известными [5]. Поскольку большие языковые модели обучены на огромных наборах данных и представляют из себя большие нейросетевые архитектуры, то существует целый ряд подходов в рамках обучения с открытым словарём, в той или иной форме использующих информацию либо на выходе из языковых моделей, либо из её промежуточных представлений.

Использование большого объёма знаний может быть крайне полезно, чтобы работать с неразмеченными данными или с данными, в которых у каждого класса объектов содержится

мало количество экземпляров. Также из-за большого набора знаний способность к сегментации и детекции сложных объектов, например статуй или барельефов, и их повреждений может быть достаточно высокой. Наконец, благодаря лучшему, по сравнению с классическими свёрточными нейросетями, понимаю окружающего реального мира вполне возможно, что точность сегментации у моделей, использующих большие объёмы знаний, будет относительно высокой при работе не с частями изображений, а с целыми фасадами зданий.

Выводы. Рассмотрены подходы к сегментации и детекции повреждений архитектурных объектов, а также выявлены ограничения традиционных методов. Проанализирована возможность применения визуальных языковых моделей и обучения с открытым словарём для обработки сложных и редких объектов, что может способствовать повышению точности и гибкости анализа не только фасадов зданий, но и других сложно структурированных изображений.

Список использованных источников:

1. Mishra, Mayank & Lourenco, Paulo. (2024). Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review. Journal of Cultural Heritage. 66. 536-550. 10.1016/j.culher.2024.01.005.

2. Redmon, Joseph, Santosh Kumar Divvala, Ross B. Girshick and Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 779-788.

3. Xie, Enze, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez and Ping Luo. “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers.” *Neural Information Processing Systems* (2021).

4. Bordes, Florian et al. “An Introduction to Vision-Language Modeling.” *ArXiv* abs/2405.17247 (2024): n. pag.

5. Wu, J., Li, X., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., Tao, D., & Xu, S. (2023). Towards Open Vocabulary Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*, 5092-5113.