МЕТОДЫ УЧЕТА ТЕКСТОВОЙ ИНФОРМАЦИИ В ЗАДАЧАХ МАШИННОГО ЗРЕНИЯ

А.П. Попов

ИПМ им. М.В. Келдыша РАН, г.Москва

alekspopovr68@mail.ru

Разработка улучшенных методов решения задач машинного зрения – на сегодняшний день ключевая задача множества исследовательских лабораторий. Алгоритмы компьютерного зрения используются в автономных робототехнических системах, в системах видеонаблюдения и во множестве других прикладных систем. Важным трендом в развитии алгоритмов машинного зрения является изучение методов использования дополнительной не визуальной информации об объектах среды для решения задач описания и поиска этих объектов. В частности, широкое распространение получили методы использования текстовой информации в задачах компьютерного зрения.

В рамках данной работы проведен обзор способов учета текстовой информации в задачах машинного зрения и выбран способ учета дополнительной информации в задачах компьютерного зрения для дальнейшего исследования. Общая схема механизма учета текстовой информации изображена на рис. 1. Рассмотрены работы, посвященные анализу действий на видеопоследовательности и способам кодирования информации на видеопоследовательности [1, 2]. Например, в работе [1] при построении промежуточного представления учитываются части речи, используемые в описательных предложениях – такой подход позволяет получить более точное

Рис. 1. Общая схема учета текстовой информации для формирования выводов по наблюдаемой сцене в описанных алгоритмах.

представление описания для решения задачи сопоставления действий на видеопоследовательности и текстового предложения, описывающего это действие.

Кроме того, в обзор включены методы анализа отдельных кадров, в частности, алгоритмы детектирования объектов и их описаний на изображении [3–5]. Эти алгоритмы также используют методы анализа промежуточных нейросетевых представлений для решения задач детектирования объектов. Важно отметить, что в представленных подходах контроль генерации текстового описания осуществляется за счет учета данных, полученных на изображении, что позволяет уточнять как текстовое описание, так и детекции. В работе [5] предложен метод использования набора изображений и их описаний для обучения модели детектирования объектов произвольных классов и описаний этих классов. Набор детектируемых классов может быть ограничен для каждой конкретной задачи.

На основе анализа работ, посвященных задаче учета текстовой информации в контексте задач компьютерного зрения, был сделан вывод о перспективности нейросетевого метода генерации промежуточных представлений информации на основе данных из нескольких источников (текстовых и визуальных) и их дальнейшего использования для решения задач распознавания. С учетом сделанных выводов начато исследование возможности объединения этого метода с другими методами представления и обработки знаний в компьютерных системах.

Список литературы:

1. Wray M. et al. Fine-grained action retrieval through multiple parts-of-speech embeddings // Proceedings of the IEEE/CVF international conference on computer vision. 2019. P. 450–459.
2. Gabeur V. et al. Multi-modal transformer for video retrieval // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. Springer International Publishing. 2020. P. 214–229.
3. Gupta T. et al. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 16399–16409.
4. Maaz M. et al. Class-agnostic object detection with multi-modal transformer // European conference on computer vision. – Cham: Springer Nature Switzerland, 2022. P. 512–531.
5. Cheng T. et al. YOLO-World: Real-Time Open-Vocabulary Object Detection // arXiv preprint arXiv:2401.17270. 2024.