МЕТОД ОБНАРУЖЕНИЯ ОБЪЕКТОВ НА ИЗОБРАЖЕНИИ С ОТКРЫТЫМ СЛОВАРЕМ

Верхов К.А.[1], ассистент, k.verkhov@bsuir.by
Ларькин А.Д.[2], ассистент, a.larkin@bsuir.by

2025

1. Белорусский государственный университет информатики и радиоэлектроники
2. Белорусский государственный университет информатики и радиоэлектроники

Ключевые слова: искусственный интеллект, компьютерное зрение, обнаружение объектов, визуально-языковые модели, открытый словарь.

Аннотация: Статья посвящена новым методам обнаружения объектов с открытым словарём (Open-Vocabulary Detection, OVD) в области компьютерного зрения, которые используют визуально-языковые модели (VLM). Эти методы позволяют моделям обнаруживать объекты, не входящие в заранее определённые категории, основываясь на взаимосвязи между визуальными признаками и текстовыми описаниями. Это кардинально отличается от традиционных подходов, которые ограничиваются заранее размеченными наборами данных с фиксированным числом классов. В статье подчеркивается, что методы OVD значительно повышают адаптивность и точность моделей в реальных приложениях, позволяя работать с новыми объектами, которые не были заранее размечены.

Обнаружение объектов с открытым словарем (open-vocabulary detection, OVD) – это новая техника обнаружения объектов в области компьютерного зрения, внедрённая благодаря визуально-языковым моделям (vision-language models), таким как CLIP. Модели теперь могут находить объекты, не входящие в заранее определённый список категорий. Они обучаются обнаруживать объекты, основываясь на взаимосвязи между визуальными признаками и текстовыми описаниями. Это основное отличие от традиционных методов, которые ограничиваются большими размеченными наборами данных с фиксированным набором классов. Использование этой технологии повышает адаптивность модели в таких областях, как автономная навигация, робототехника и поиск изображений [1].

Можно выделить следующие модели, использующие технику OVD:

1. GroundingDINO.
2. OWL-ViT.
3. RegionCLIP.

GroundingDINO – это улучшенная версия модели для обнаружения объектов, которая основана на технологии DINO (Detection Transformer with Improved DeNoising Anchor Boxes). Эта модель включает в себя объединение визуальных и текстовых данных, что даёт возможность обнаруживать объекты, указанные в текстовых запросах. В отличие от обычных моделей, ограниченных жёсткими категориями, GroundingDINO использует подход,

основанный на текстовых подсказках, чтобы обеспечить обнаружение объектов без предварительной подготовки [2]. Это не только повышает точность, но и позволяет использовать такие модели в областях, где сложно составить определённый список категорий. Модель проходит два этапа: сначала генерируются региональные предложения на основе визуальных данных, а затем эти предложения уточняются через сопоставление с текстовыми запросами.

OWL-ViT (Open-World Learning with Vision Transformers) – это ещё одна модель для обнаружения объектов, основанная на визуальных трансформерах. Вместо того чтобы полагаться на фиксированные метки классов, как это делают традиционные модели, OWL-ViT использует метод контрастного обучения, подобный CLIP, чтобы соотнести текстовые и визуальные данные. Это позволяет пользователю вводить текстовые запросы и сопоставлять их с визуальными данными. Особенностью этой модели является возможность обучаться на неразмеченных наборах данных что позволяет работать с теми объектами, которые не были заранее размечены [3]. Всё это делает OWL-ViT идеальной моделью для применения в таких областях, как ИИ-помощники и модерация контента, где нужна гибкость в распознавании объектов без жесткой привязки к заранее подготовленным категориям.

RegionCLIP – это усовершенствование модели CLIP, которое добавляет обнаружение объектов на отдельных регионах изображения, а не на всем изображении. В отличие от своих предшественников, RegionCLIP использует самообучение для улучшения сопоставления текстовых запросов с локализованными регионами на изображениях [4]. Особенность этой модели в том, что она связывает текстовые метки с определёнными областями изображения, а не с целыми картинками. Такая технология делает её особенно эффективной для точного обнаружения объектов, что очень важно в таких сферах, как медицинская визуализация и автономная навигация.

Методы обнаружения с открытым словарем (OVD) изменяют подход к обнаружению объектов. Они позволяют моделям адаптироваться к новым категориям объектов, которые не были заранее размечены. Это позволяет моделям выходить за пределы заранее заданных категорий. Развитие таких моделей, как GroundingDINO, OWL-ViT и RegionCLIP, делает эту технологию более гибкой и универсальной для реальных применений, будь то автономные транспортные средства, робототехника или поисковые системы.

Список использованных источников

1. Edozie, E. Comprehensive review of recent developments in visual object detection based on deep learning / Edozie, E., Shuaibu, A.N., John, U.K. // Artificial Intelligence Review – Vol. 58, N. 227 – 2025.

2. Liu, S. Grounding dino: Marrying dino with grounded pre-training for open-set object detection / Liu, S. // European conference on computer vision – 2024 – Pp. 38-55.

3. Minderer, M. Simple open-vocabulary object detection / Minderer M. // European conference on computer vision – 2022 – Pp. 728–755.

4. Zhong, Y. Regionclip: region-based language-image pretraining / Zhong, Y. // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition – 2022 – Pp. 16793-16803.