УДК 004.896

ГЕНЕРАЦИЯ И ОЦЕНКА АННОТАЦИЙ С ИСПОЛЬЗОВАНИЕМ МУЛЬТИМОДАЛЬНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ И ОПТИМИЗАЦИЯ ТЕКСТОВЫХ ИНСТРУКЦИЙ

Герасимчук М.Ю. (ИТМО), **Сметанин А.А.** (ИТМО),
Научный руководитель – доктор технических наук, доцент Духанов А.В.
(ИТМО)

Введение. В условиях стремительного роста объемов данных и разнообразия визуальной информации автоматизация процессов генерации аннотаций изображений становится ключевым этапом в подготовке данных для задач компьютерного зрения. Современные системы обработки изображений требуют высокой точности разметки, а также эффективности при работе с большими наборами данных. В данной работе исследуется потенциал использования Vision-Language Large Models (VLLM) для автоматической генерации аннотаций, пригодных для обучения моделей YOLO. Особое внимание уделяется сравнительному анализу различных моделей и оптимизации текстовых инструкций, которые задают формат и требования к аннотациям.

Основная часть. В рамках исследования была проведена серия экспериментов по оценке эффективности современных моделей VLLM, таких как GLIP, yolo-world и других аналогичных решений [1,2]. Эксперименты проводились на различных датасетах, охватывающих широкий спектр сценариев и условий съемки. В частности, использовались наборы данных, содержащие изображения транспортных средств, включая съемку с дронов, что добавляло сложности за счет ракурсов и высоты съемки. Другие датасеты включали изображения животных, автомобильных номеров и растений, каждое из которых имело свои уникальные особенности: разные условия освещения, ракурсы, масштаб объектов и их расположение относительно фона. Такое разнообразие данных позволило всесторонне оценить производительность моделей в зависимости от типа изображений и специфики задачи.

Входные данные:

- Аннотированные изображения с экспертной разметкой объектов;
- Текстовые инструкции, задающие формат аннотаций и особенности обработки изображений.

Для оценки эффективности моделей использовались следующие метрики:

- Показатели моделей:
  1. Время отклика модели на CPU и GPU;
  2. Объем используемой памяти на CPU и GPU;
  3. Размер модели.
- Оценка ответов:
  1. Коэффициент совпадения предсказанных и истинных границ объектов (IOU);
  2. Точность обнаружения объектов (Precision);
  3. Полнота обнаружения объектов (Recall);
  4. Гармоническое среднее между точностью и полнотой (F1-score) [3];

Примеры текстовых инструкций варьировались от простых запросов на генерацию аннотаций до более сложных, включающих нормализацию координат, указание классов объектов и примеры формата вывода.

Выводы. Проведенный сравнительный анализ показал, что эффективность моделей VLLM зависит от множества факторов, включая точность детекции, скорость работы, вычислительные затраты и способность адаптироваться к различным текстовым инструкциям. Результаты исследования подтвердили важность оптимизации промтов для повышения

качества генерируемых аннотаций. Предложенный подход может быть успешно применен для препроцессинга данных при обучении моделей компьютерного зрения, что позволит сократить временные и трудозатраты на ручную разметку данных.

Список использованных источников:

1. Zhang, H., Zhang, P., Hu, X., Chen, Y., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.-N., Gao, J. GLIP: Grounded Language-Image Pre-training // arXiv:2206.05836. – 2022.
2. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection // arXiv:2401.17270v3. – 2024.
3. Everingham M., Van Gool L., Williams C.K.I., Winn J., Zisserman A. The Pascal Visual Object Classes (VOC) Challenge // International Journal of Computer Vision. – 2010. – Vol. 88. – № 2. – P. 303-338.