

УДК 004

МЕТОДЫ РАЗМЕТКИ ДАННЫХ ДЛЯ ЗАДАЧ КОМПЬЮТЕРНОГО ЗРЕНИЯ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Прохоров К.В., студент гр. ИИБ-231, III курс

Научный руководитель: Сыркин И.С., к.т.н., доцент

Кузбасский государственный технический университет имени Т.Ф. Горбачева,
г. Кемерово

Разметка данных (data labeling) является одним из ключевых этапов подготовки обучающих выборок для нейросетевых моделей компьютерного зрения. Качество аннотирования напрямую определяет точность и надежность итоговой системы обнаружения объектов. В статье рассматриваются основные подходы к разметке данных: ручной, автоматизированный и гибридный. Производится сравнительный анализ их преимуществ и недостатков, а также обзор современной системы аннотирования CVAT. Особое внимание уделяется вопросам экономической эффективности, качества разметки и применимости различных методов в зависимости от масштаба проекта и специфики предметной области.

Процесс создания высококачественных размеченных данных остается одной из наиболее трудоемких и дорогостоящих стадий разработки систем искусственного интеллекта. В зависимости от масштаба проекта, требования к точности и доступных ресурсов применяются различные методы аннотирования. На данный момент популярностью пользуются три основных метода разметки данных: ручная разметка, автоматизированная разметка и гибридных подход.

Ручная разметка предполагает выполнение аннотирования исключительно человеческими усилиями с использованием специализированных инструментов. Данный метод считается золотым стандартом в подготовке обучающих выборок.

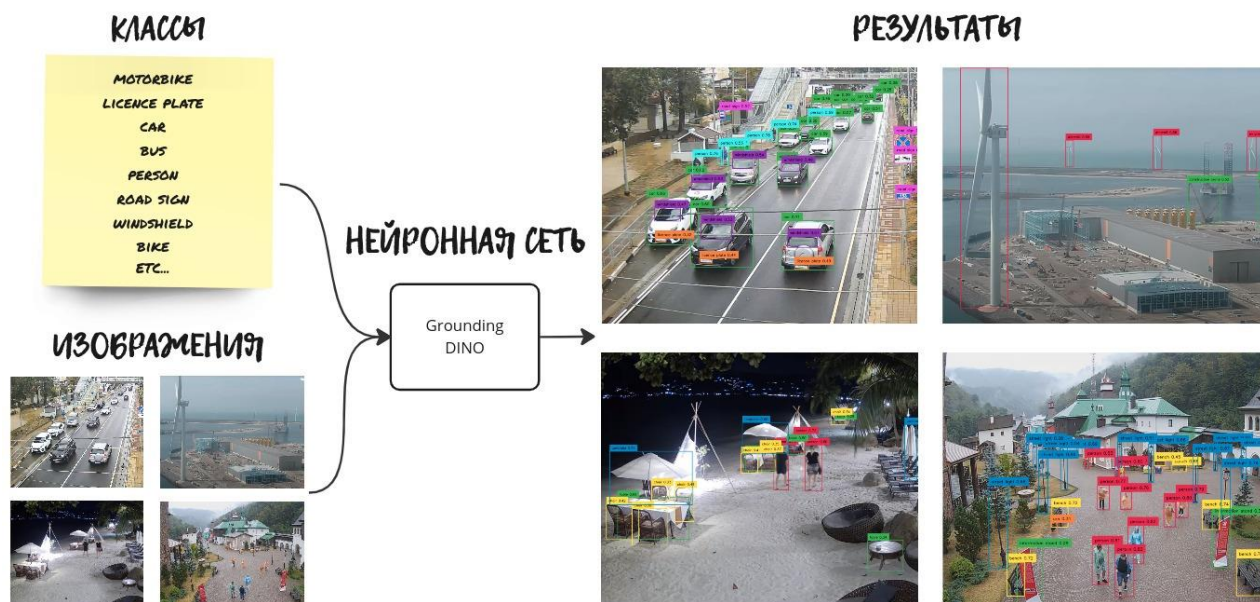
Преимущества ручной разметки:

1. Высокая точность как с простыми, так и со сложными объектами.
2. Конфиденциальность данных при использовании штатных сотрудников.

Недостатки:

1. Высокие временные и финансовые затраты
2. Снижение качества работы при большом объеме данных из-за человеческого фактора.
3. Сложность масштабирования при увеличении объема данных.

Автоматизированная разметка использует алгоритмы и предобучение модели для генерации аннотаций. Среди методов автоматизации выделяют предразметку (pre-labeling), когда модель создает черновые аннотации для последующей проверки человеком, и активное обучение (active learning), при котором модель самостоятельно выбирает наиболее информативные образцы для разметки



Преимущества автоматизированной разметки:

1. Высокая скорость обработки
2. Низкая стоимость

Недостатки:

1. Зависимость от качества модели
2. Сложность обработки редких классов
3. Невозможность полной автоматизации для высокоточных приложений

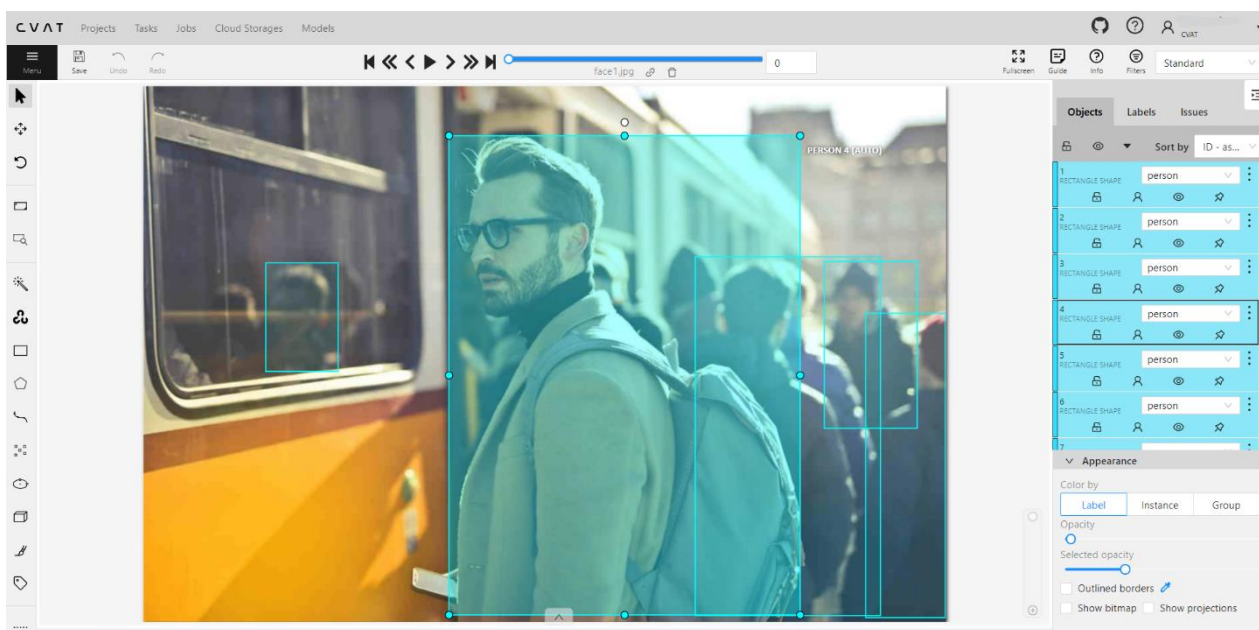
Гибридный подход объединяет преимущества автоматизации и человеческого контроля, создавая итеративный цикл улучшения качества. Процесс включает в себя несколько этапов:

1. Обучение начальной модели на небольшом размеченном вручную датасете.
2. Генерация разметки для новых данных.
3. Верификация и исправление аннотаций человеком.
4. Дообучение модели на исправленных данных.

Данный подход позволяет сократить трудоемкость до 80% по сравнению с чистой ручной разметкой и обеспечивает адаптивность к предметной области.

Существует широкий спектр инструментов для аннотирования данных, но на данный момент лидерам в области разметки данных является такой инструмент как CVAT.

CVAT (Computer Vision Annotation Tool) – инструмент с открытым исходным кодом, разработанный компанией Intel. Платформа поддерживает различные типы аннотаций: ограничивающие рамки, полигоны, полилинии, ключевые точки, а также 3D-кубоиды.



Ключевой особенностью CVAT является возможность AI-assisted labeling с интеграцией моделей YOLO и SAM. Инструмент поддерживает экспорт в форматы Pascal VOC, YOLO и может быть развернут как на сервере, так и локально на личном компьютере.

Преимущества CVAT:

1. Высокая степень автоматизации.
2. Поддержка различных типов данных.
3. Мощные инструменты для командной работы.

Недостатки:

1. Необходимость настройки и обслуживание сервера.

Выбор метода и инструмента разметки определяется масштабом проекта и требованиями к качеству. Для малых проектов (до 1000 изображений) целесообразна ручная разметка с использованием легковесных инструментов (LabelImg, Make Sense). Для средних и крупных проектов оптимальным является гибридный подход с использованием платформ, поддерживающих AI-assisted labeling (CVAT, Label Studio). Для проектов с высокими требованиями к конфиденциальности предпочтительны инструменты с локальным развертыванием и штатные сотрудники. Для проектов со специфическими объектами (например, карьерная техника) критически важна возможность встраивания кастомных моделей, что обеспечивает CVAT.

Можно сделать вывод, что ручная разметка обеспечивает максимальное качество, но характеризуется высокими временными и финансовыми затратами. Автоматизированные методы позволяют существенно ускорить процесс, но требуют доработки человеком. Гибридный подход, объединяющий предразметку на основе предобученных моделей и последующую верификацию, позволяет достичь оптимального баланса между качеством, скоростью и стоимостью. Выбор конкретной платформы определяется масштабом проекта, типами данных, требованиями к конфиденциальности и необходимостью интеграции с пайплайнами обучения.

Список литературы.

1. Бешапошников Н.О., Кузьменко М.А., Леонов А.Г., Матюшин М.А. Автоматизация разметки набора данных для нейронных сетей [Электронный ресурс] // URL: <https://www.vestcyber.ru/jour/article/view/50>
2. Cacciarelli D., Kulahci M. Active learning for data streams: a survey [Электронный ресурс] // URL: <https://arxiv.org/pdf/2302.08893>