УДК 004.048

РАЗРАБОТКА СЕРВИСА РАЗМЕТКИ ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ

Ермаков Т.С. (ИТМО), **Лапин А.А.** (ИТМО), **Ри А.Р.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Кугаевских А.В.
(ИТМО)

Введение. В современной практике машинного обучения с учителем одной из ключевых задач является сбор и подготовка больших объемов данных [1]. Качество и корректность разметки напрямую влияют на результаты обучения нейронных сетей и уровень окончательной точности [2]. В связи с этим актуальным становится создание удобной масштабируемой платформы для исследователей и бизнеса, позволяющей асессорам эффективно производить разметку наборов данных между собой, например, товаров, включающих в себя описания, изображения, ссылки и другую информацию. В зарубежной и отечественной практике представлены различные сервисы для разметки, такие как покадровая разметка видео, выделение области на изображении и другие, но многие из них закрыты, специализированы под нужды отдельной компании, а также не предусматривают сопоставления данных друг с другом, гибкой настройки, масштабирования на большое количество пользователей.

Основная часть. В рамках данного исследования разработан сервис разметки данных о товарах, опирающийся на спецификацию, предоставленную в рамках образовательного интенсива AI Learning Lab. Архитектура решения предполагает модульную структуру со следующими возможностями:

1) Загрузка пакетов данных. Предусмотрен импорт через JSON или CSV для различных типов разметки (поиск, сопоставление). Для каждого батча можно задавать приоритет, критерий перекрытия (overlaps) и иные настройки.
2) Гибкая система ролей и прав. Реализована многоуровневая модель пользователей (асессор, админ), где каждый видит только соответствующие разделы и статистику.
3) Распределение заданий. Сервис автоматически выдаёт задания асессорам с подходящими навыками. Поддерживается параллельная работа большого числа пользователей с помощью архитектуры на Java, Spring, PostgreSQL с использованием Docker.
4) Контроль качества. Включает ханипоты (honeypots), позволяющие оценивать точность разметки. Система ведёт сводную статистику правильных/неправильных ответов и учитывает время, затрачиваемое на разметку, что позволяет выявить проблемные места.
5) Статистика и аналитика. Подробные показатели для каждого асессора (процент ошибок, скорость, количество обработанных заданий) и по пакетам в целом.
6) Модуль обучения. Предусмотрен отдельный режим, в котором каждое разметанное задание сопровождается эталонным результатом и пояснениями, что даёт возможность быстрой подготовки новых асессоров.

Выводы. Разработанный сервис автоматизирует процесс формирования обучающих выборок для нейронных сетей и предоставляет инструменты обучения персонала, контроля качества и детальной аналитики. Внедрение подобной системы позволит компаниям и исследовательским коллективам масштабировать проекты по сбору и аннотированию данных, повысив качество итоговых моделей за счёт точной и оперативной разметки. В дальнейшем планируется расширение функционала за счёт интеграции дополнительных типов данных (видео, аудио), а также внедрение агентов авторазметки и расширенной статистики.

Список использованных источников:

1. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A largescale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
2. He C. et al. Opendatalab: Empowering general artificial intelligence with open datasets //arXiv preprint arXiv:2407.13773. – 2024.