И. А. Крупкин, А. А. Лесив, Г. Б. Барская
Тюменский государственный университет, г. Тюмень
УДК 004.931

ПРИМЕНЕНИЕ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ И ИЗВЛЕЧЕНИЯ ТАБЛИЧНЫХ ДАННЫХ ИЗ PDF-ФАЙЛОВ

Аннотация. Представлено решение задачи с использованием методов компьютерного зрения. Проведено сравнение точности работы алгоритмов и рассмотрены проблемы обработки и извлечения данных из PDF-файлов. Результаты экспериментов показали высокую точность и эффективность алгоритма YOLOv8.

Ключевые слова: компьютерное зрение, обнаружение PDF-таблиц, извлечение табличных данных, распознавание объектов, YOLOv8.

Введение. В 2016 г. L. Hao, L. Gao, X. Yi и Z. Tang для обнаружения таблиц в PDF-документах использовали сверточную нейронную сеть на базе архитектуры LeNet [1]. Достигнут результат в 94% по метрике F1-Score, однако используемый набор довольно качественный и неизвестно, как покажет себя модель на более зашумленных данных.

В 2017 г. Azka Gilani, Shah Rukh Qasim, Imran Malik и Faisal Shafait использовали подход с выделением участков текста трансформацией изображения и применением алгоритма Faster R-CNN для обнаружения таблиц [2]. Метод показал точность выше 80% по метрике F1-Score, но обнаружение таблиц без рамок показала посредственные результаты.

В 2017 г. Akash Gupta, Anand Shankar S. и Manjunath C.R. предложили метод извлечения данных, заключающийся в обнаружении структуры данных и распознавании конкретных объектов в этой структуре. Это повысит точность извлечения важных данных и уменьшит влияние шума [3].

В 2020 г. Jyothi E, K Tejaswini., Lakshmi Chintalapati и Mr. MD. Shafiulla демонстрируют результаты разработки системы для извлечения текста из изображений с использованием Tesseract OCR [4].

Система показывает хорошие результаты извлечения данных, особенно в вырезных областях чтения [4].

В 2021 г. Borra Vineetha, D. N. D. Harini и Ravi Yelesvarupu предлагают для обнаружения и извлечения табличных данных использовать ПО, состоящее из трех модулей: обнаружение таблиц, распознавание структуры таблиц и чтение табличных данных [5]. Итоговые результаты оказались хорошими, но 2 и 3 этапы сильно зависят от имеющегося словаря и структуры таблицы.

Задачей обнаружения и извлечения табличных данных занимаются довольно продолжительное время. Однако присутствует недочеты и открытые места для улучшений. К задаче проявляют интерес такие компании, как Microsoft и Yandex, и множество энтузиастов из сообщества искусственного интеллекта на ежегодной конференции ICDAR.

Проблематика и постановка задачи. Проблема, решаемая в работе — отсутствие у Компании собственного ПО для автоматического обнаружения и извлечения табличных данных из PDF-файлов.

Задача работы — провести эксперименты по обучению моделей для обнаружения PDF-таблиц и извлечению данных из нее, разработать ПО на основе оптимальных моделей.

Материалы и методы. Для разработки архитектуры ПО решено использовать две независимые модели: одна — для обнаружения таблиц, вторая — для обнаружения объектов таблиц. Для второй задачи были выбраны алгоритмы: YOLOv8, Faster R-CNN, SSD и Fast R-CNN.

Для обучения моделей подобрано два набора данных на 6519 и 3400 изображений с применением аугментаций.

Результаты. Для обнаружения таблиц выбран алгоритм YOLOv8 и при обучении в 400 эпох достигнуты результаты (рис. 1) с оценкой точности по метрике mAP в 91.1%.

Набор данных	
Размер	6598 избр.
Классы	5 классов

Результаты	
mAP	91.1%
precession	86.8%
recall	84.9%

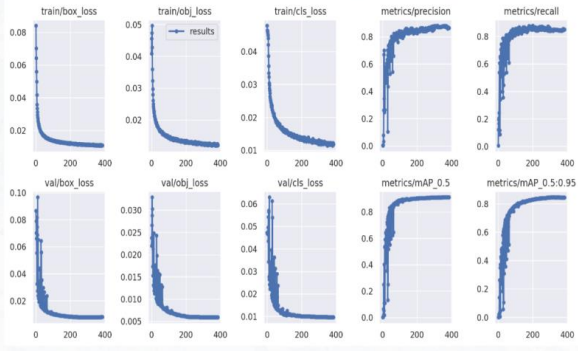

Рис. 1. Результаты обучения модели для обнаружения таблиц

Эти результаты подтверждаются на примере (рис. 2), где границы и классы обеих таблиц были точно определены.

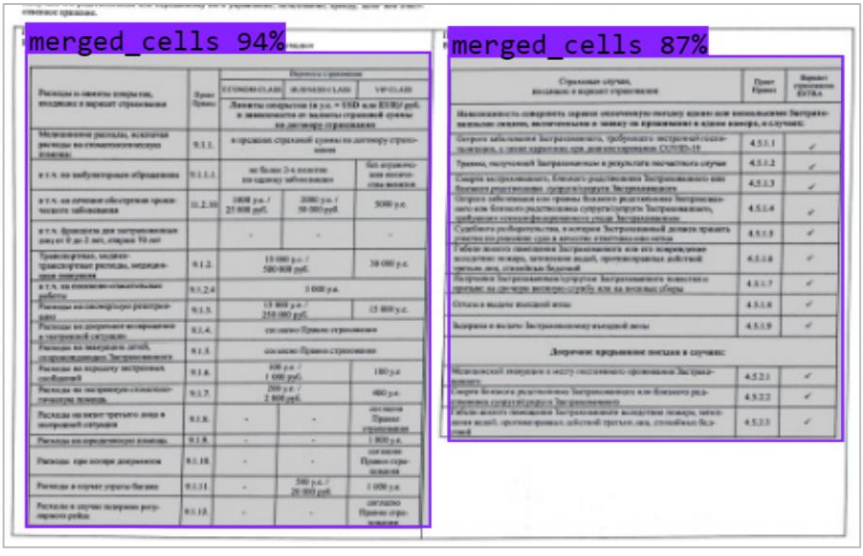

Рис. 2. Пример работы обученной модели для обнаружения таблиц

Для выбора оптимального алгоритма распознавания объектов таблиц, был проведен сравнительный эксперимент между моделями с 4 разными алгоритмами. По полученным результатам лучшую точность по метрике mAP показала модель на основе YOLOv8 и составила 82.2%.

Работа выбранной модели продемонстрирована ниже (рис. 3). Из рисунка видно, что модель отлично определяет строки и ячейки, но посредственно справляется с колонками и ячейками заголовков. Таким образом, принято решение ориентироваться на строки и ячейки при формировании таблицы.

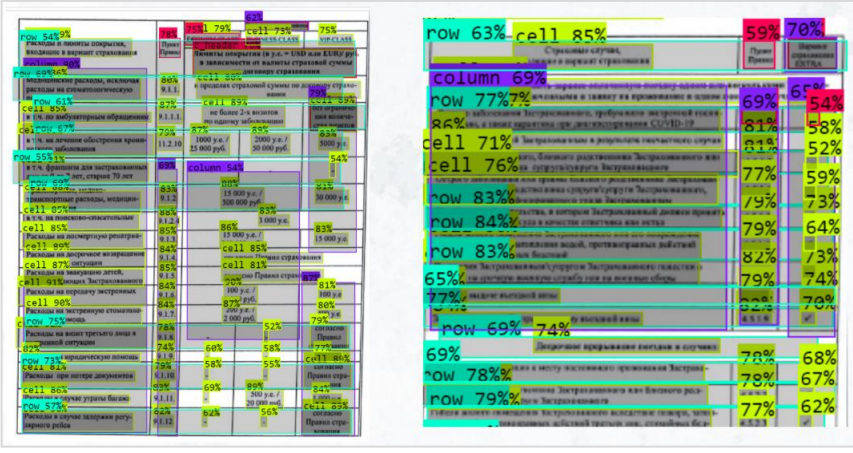

Рис. 3. Пример работы модели YOLOv8
для распознавания объектов таблиц

Для чтения данных таблицы формируется матрица, размерностью по количеству строк в таблице, и максимальному количеству определенных ячеек по горизонтали. Итоговая матрица приводится к формату DataFrame. Результат отображается в тестовом приложении (рис. 4).

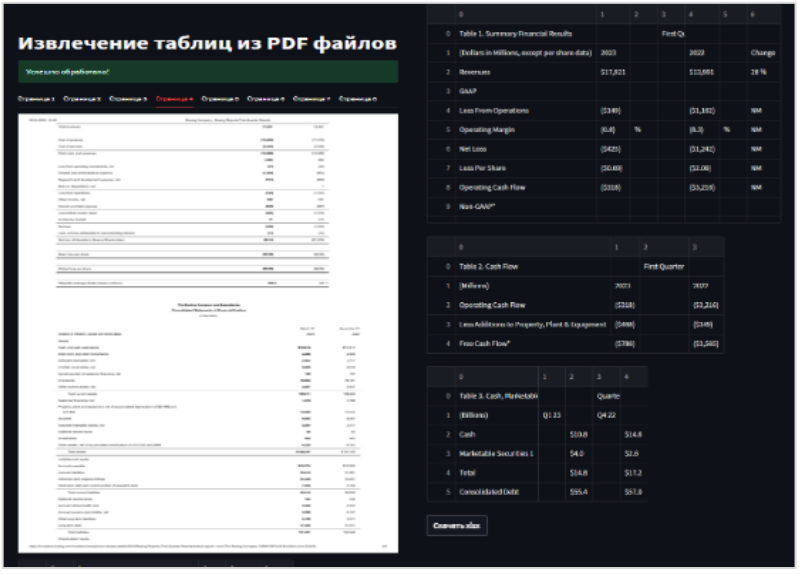

Рис. 4. Пример работы тестового приложения

Заключение. Разработана архитектура ПО, использующая две модели для последовательного обнаружения таблиц и ее объектов.

YOLOv8 показала точность по метрике F1-Score в 91.1% на обнаружении таблицы и 82.2% на обнаружении объектов таблицы. Модель имеет проблемы с обнаружением колонок и заголовков, но отлично определяет строки и ячейки.

Создано тестовое ПО с функциями просмотра извлеченных табличных данных и выгрузкой их в формате xlsx.

СПИСОК ЛИТЕРАТУРЫ

1. A Table Detection Method for PDF Documents Based on Convolutional Neural Networks / L. Hao, L. Gao, X. Yi, Z. Tang. — Текст: непосредственный // 12th IAPR Workshop on Document Analysis Systems (DAS), 2016. — С. 287-292.

2. Table Detection using Deep Learning / Azka Gilani, Shah Rukh Qasim, Imran Malik, Faisal Shafait. — Текст: непосредственный // 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017. — С. 771-776.
3. Akash Gupta. A Comparative Study on Data Extraction and Its Processes / Akash Gupta, Anand Shankar S., Manjunath C.R. — Текст: непосредственный // International Journal of Applied Engineering Research. — 2017. — Vol. 12. — С. 7194-7201.
4. Text Extraction from Image Using OCR / E. Jyothi, K. Tejaswini, Lakshmi Chintalapati, Mr. MD. Shafiulla. — Текст: непосредственный // International Journal of Applied Engineering Research. — 2017. — Vol. 9. — С. 1805-1810.
5. Borra Vineetha. Automated Table Detection and Extraction from PDF Documents using Computer Vision Techniques / Borra Vineetha, D. N. D. Harini, Ravi Yelesvarupu. — Текст: непосредственный // International Journal of Innovative Technology and Exploring Engineering. — 2021. — Vol. 10. — С. 73-79.