УДК 004.89

ДЕТЕКЦИЯ И РАСПОЗНАВАНИЕ В LATEX ПРЕДСТАВЛЕНИЕ ФОРМУЛ ИЗ НЕСТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ

Автор: Баталенков С.С. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель: аспирант, Филатова А.А. (Университет ИТМО, г. Санкт-Петербург)
Научный консультант: аспирант, Ковальчук М.А. (Университет ИТМО, г. Санкт-Петербург)

Введение. Задача парсинга сканов PDF документов заключается в извлечении текстовой и графической информации из изображений, полученных путем сканирования документов. В отличие от PDF с текстовым слоем, где информация о тексте документа сохраняется в виде машиночитаемых данных, сканы представляют собой изображения, что усложняет задачу извлечения содержания. Парсинг сканов включает в себя использование методов компьютерного зрения и оптического распознавания символов (OCR) для преобразования изображений в текстовые данные, а также анализ структуры документа (например, заголовков, абзацев, таблиц, изображений) для дальнейшего представления в удобном формате. Важными задачами являются обеспечение высокой точности распознавания при различных качествах сканов, обработка различных шрифтов и стилей, а также правильная интерпретация структуры документа [2].

Задача детекции математических формул на сканах PDF документов включает в себя извлечение и распознавание формул, представленных в виде изображений, из документов, содержащих текстовые или графические данные. Этот процесс важен для автоматизации обработки научных, технических и образовательных материалов, где формулы играют ключевую роль. При отсутствии текстового слоя в PDF (что характерно для сканов документов, а особенно учебников) задача становится значительно сложнее, поскольку необходимо извлекать математическое содержание из графических данных плохого типографского качества и качества самого скана. Важно учитывать, что математические выражения могут быть представлены как сложные изображения с различными шрифтами, стилистикой и ориентацией, что требует применения методов компьютерного зрения, анализа изображений и машинного обучения для их точного распознавания и преобразования в структурированные форматы.

Проблемы в существующих решениях для парсинга документов [1]:

1. **Недостаточная точность распознавания на сканах с низким качеством**: Существующие методы часто ошибаются на сканах с низким качеством изображений или нестандартными шрифтами, что ведет к ошибкам в распознавании формул. Шум на изображениях, плохая контрастность и пикселизация могут существенно ухудшить качество результата.
2. **Отсутствие универсальности и адаптивности**: Существующие решения не всегда хорошо работают с различными стилями и форматами математических выражений. Например, различные способы записи формул (например, встраивание формул в текст или представление в виде отдельных блоков) могут требовать адаптации решений под каждую конкретную задачу.
3. **Проблемы с определением структуры формулы**: Даже если формула распознана, её структура (например, правильность скобок, вложенность операций) может быть трудно интерпретировать, что приводит к ошибкам в дальнейшей обработке или конвертации в стандартизированные форматы, такие как LaTeX.

4. **Неэффективность при работе с многокомпонентными формулами**: В сложных выражениях, состоящих из нескольких элементов (например, индексов, дробей, интегралов), текущие системы часто не могут корректно разделить эти элементы и отобразить их в нужной последовательности.
5. **Проблемы с интеграцией в реальных сценариях**: Многие решения для парсинга математических формул не интегрируются должным образом с другими инструментами и системами (например, базы данных или редакторы научных работ), что ограничивает их практическую применимость в реальных задачах автоматизации.

Основная часть. В данном исследовании предлагается использовать комбинацию алгоритмы на основе архитектуры трансформеров:

- специализированные модели (CNN + language decoder)
- мультимодальные большие языковые модели (БЯМ)

Решено использовать трансформеры, потому что они:

- объединяют два типа информации – графическую и текстовую, благодаря чему повышается точность распознавания;
- меньше реагируют на шумы и артефакты изображения за счёт способности к обобщению получаемой и выдаваемой информации;
- способны к экстракции контента и игнорированию стиля, что важно при последующем использовании полученной информации;
- позволяют донастраивать модель посредством простого дообучения.

Выводы. В результате данной работы было подготовлено описание архитектуры системы, обоснование выбора концепций и алгоритмов. В рамках работы также произведено обсуждение перспектив развития данной концепции и разработана система, реализующая некоторые из них.

Список использованных источников:

1. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model / Wei H., Liu C., Chen J. et al. // ArXiv 2024
2. PP-OCRv2: Bag of Tricks for Ultra Lightweight OCR System / Du Y., Li C., Guo R. // 2021

Баталенков С.С. (Автор)

Филатова А.А. (Научный Руководитель)

Ковальчук М.А. (Научный Руководитель)