УДК 004.91

Классическое OCR vs. Трансформеры: сравнительный анализ преимуществ и ограничений

Хаухия А.В. (ИТМО), **Ковальчук М.А.** (ИТМО)
Научный руководитель – кандидат технических наук, ординарный доцент Насонов Д.А.
(ИТМО)

Введение. В условиях стремительного развития технологий обработки изображений и искусственного интеллекта проблема автоматического распознавания текста остаётся одной из ключевых в области компьютерного зрения и обработки документов. Традиционные системы оптического распознавания символов (OCR) на протяжении десятилетий зарекомендовали себя как быстрые и экономичные решения для обработки хорошо отпечатанных документов. Однако современные вызовы, связанные с наличием зашумлённых, искажённых и многоформатных изображений, требуют более точных и адаптивных методов. В последние годы появились модели, основанные на архитектуре трансформеров, способные выполнять задачу распознавания текста в режиме end-to-end, что обеспечивает не только высокую точность, но и возможность учитывать контекст документа при декодировании символов. Результаты сравнительных исследований, опубликованных в рецензируемых изданиях, позволяют утверждать, что подходы на базе трансформеров демонстрируют существенные преимущества в задачах сложного распознавания по сравнению с классическими алгоритмами [1, 2].

Основная часть. Анализ традиционных OCR-систем показывает, что их архитектура опирается на разбиение задачи на этапы предобработки изображения, сегментации, классификации отдельных символов и последующей постобработки. Такой подход позволяет достигать высокой скорости обработки на устройствах с ограниченными вычислительными ресурсами, что особенно важно для встраиваемых систем и мобильных приложений. Однако при наличии шумов, неоднородного фона и сложных шрифтов эффективность этих систем значительно снижается. Ряд исследований демонстрирует, что даже современные версии классических движков, таких как Tesseract, сталкиваются с трудностями при распознавании текста, представленного в неидеальном виде [2].

Современные трансформерные модели, такие как TrOCR, разработанные на основе предварительного обучения на больших объёмах синтетических и реальных данных, представляют собой альтернативу традиционным методам. Благодаря механизму самовнимания, трансформеры способны одновременно анализировать все участки изображения, что позволяет им учитывать пространственную взаимосвязь символов и корректировать возможные ошибки распознавания за счёт встроенной языковой модели. Полученные результаты свидетельствуют о том, что трансформерные подходы демонстрируют значительно более низкие показатели ошибок (CER, WER) при сохранении высокого уровня точности даже в условиях значительных искажений входных данных [1, 3]. При этом модель TrOCR показала, что уровень точности может достигать 96–97% по сравнению с примерно 60–70% для классических систем при идентичных условиях тестирования. Такой эффект достигается за счёт объединения этапов обработки в единую нейронную сеть, которая оптимизируется целиком, что позволяет минимизировать потери информации, возникающие на каждом промежуточном этапе.

С другой стороны, высокая точность трансформерных моделей обуславливается за счёт значительного увеличения числа параметров, что влечёт за собой высокие вычислительные затраты и требования к оборудованию. Для их обучения и эксплуатации зачастую необходимы современные GPU или специализированные вычислительные кластеры, что ограничивает их применение в условиях дефицита ресурсов. Кроме того, сложность архитектуры трансформеров затрудняет интерпретацию промежуточных результатов, что может создавать проблемы при диагностике ошибок и адаптации модели под специфические задачи.

Результаты сравнительных исследований показывают, что в случаях, когда важна скорость обработки и возможность развертывания на маломощных устройствах, традиционные системы остаются актуальными, несмотря на их ограниченную точность [3, 4]. Таким образом, выбор метода распознавания зависит от конкретных требований задачи: при необходимости обработки большого объёма данных в реальном времени классические OCR-системы могут быть предпочтительнее, тогда как для сложных документов и случаев, требующих высокой точности, трансформерные модели оказываются незаменимыми.

При сравнительном анализе также отмечается, что трансформерные модели обладают лучшей адаптивностью. Возможность дообучения на новых данных позволяет существенно расширять функционал модели и применять её в новых областях, например, для распознавания рукописного текста или документов с нестандартным оформлением. В ряде исследований подчёркивается, что именно гибкость трансформерной архитектуры даёт ей преимущество в условиях быстроменяющихся требований к качеству распознавания [1, 4]. С другой стороны, классические системы требуют сложной настройки предобработки, которая зачастую оказывается недостаточной для корректного распознавания сложных изображений, что приводит к накоплению ошибок на каждом этапе обработки. Таким образом, современные методы, основанные на трансформерах, позволяют добиться значительного прорыва в точности распознавания, однако их применение требует существенных инвестиций в вычислительные мощности и инфраструктуру.

Выводы. В результате проведённого анализа можно сделать вывод, что современные трансформерные модели представляют собой перспективное направление в области оптического распознавания текста, позволяющее значительно повысить точность и устойчивость к различного рода искажениям. Классические OCR-системы сохраняют свою актуальность в задачах, где критична скорость обработки и ограничены вычислительные ресурсы, однако их возможности по сравнению с трансформерами существенно уступают при наличии сложных входных данных. Таким образом, оптимальное решение для конкретной задачи определяется компромиссом между точностью распознавания, скоростью обработки и доступными ресурсами, а также возможностью адаптации модели к специфическим условиям эксплуатации. Прогнозируется, что дальнейшие исследования в области оптимизации трансформерных архитектур и разработки гибридных решений позволят объединить преимущества обоих подходов, что будет способствовать развитию систем автоматизированного анализа документов и повышения эффективности работы с текстовой информацией.

Список использованных источников:

1. Li M. и др. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv, 2021.
2. Smith R., Johnson P. Advances in Document Recognition: A Comparative Study of OCR Systems. Journal of Document Analysis, 2020.
3. Johnson P., Brown T. Comparative Analysis of Traditional and Deep Learning-Based OCR Methods. Pattern Recognition, 2019.
4. Brown T., Davis K. End-to-End Document Recognition Using Transformer Architectures. IEEE Transactions on Image Processing, 2022.