УДК 004.352.242
DOI 10.47367/0021-3497_2026_3_280

ПРИМЕНЕНИЕ OCR-НЕЗАВИСИМОЙ МОДЕЛИ НА ОСНОВЕ ТРАНСФОРМЕРОВ ДЛЯ АВТОМАТИЗАЦИИ ДОКУМЕНТООБОРОТА НА ТЕКСТИЛЬНЫХ ПРЕДПРИЯТИЯХ

APPLICATION OF AN OCR-INDEPENDENT TRANSFORMER-BASED MODEL TO AUTOMATE DOCUMENT FLOW IN TEXTILE ENTERPRISES

А.С. СИРОТКИНА[1], М.А. ЛЫСОВА[1], А.А. АРБУЗОВА[2], Н.А. ГРУЗИНЦЕВА[2]

A.S. SIROTKINA[1], M.A. LYSOVA[1], A.A. ARBUZOVA[2], N.A. GRUZINTSEVA[2]

**([1]Ивановский государственный химико-технологический университет,
[2]Ивановский государственный политехнический университет)**

**([1]Ivanovo State of chemistry and technology University,
[2]Ivanovo State Polytechnic University)**

E-mail: anastasiaeeva@mail.ru, lysova7@yandex.ru, annaarb215@yandex.ru, gruzincevan@mail.ru

В условиях масштабной цифровизации промышленных отраслей предприятия текстильного сектора сталкиваются с необходимостью автоматизированной обработки значительных объемов многоформатной и многоязычной документации. Традиционные подходы, основанные на системах оптического распознавания символов (OCR), демонстрируют недостаточную эффективность при работе с документами низкого качества, вариативными макетами и нестандартными форматами представления данных. Цель настоящего исследования заключается в разработке и комплексной экспериментальной валидации принципиально новой OCR-независимой end-to-end модели, построенной на базе передовых трансформерных архитектур, для прямого преобразования визуальных представлений документов в структурированные данные.

Новизна работы заключается в архитектурной интеграции визуального кодировщика Swin Transformer и текстового декодера BART, что обеспечивает выполнение преобразования изображения документа в структурированный JSON-формат в едином вычислительном контуре без привлечения внешних OCR-систем. Методология обучения модели включает двухэтапный процесс с использованием комбинированного корпуса синтетических и реальных данных. Экспериментальная часть исследования охватывает тестирование модели на публичных наборах данных (RVL-CDIP, CORD) и на специализированном приватном датасете текстильной фабрики (ПДТФ), включающем отраслевые счета-фактуры, спецификации и товаросопроводительные накладные.

Результаты сравнительного анализа продемонстрировали превосходство предложенного решения над классическими OCR-зависимыми аналогами (включая LayoutLM и SPADE) по ключевым метрикам точности (значение F1-меры до 98,7% на ПДТФ) и производительности (время обработки до 0,6 секунды на одно изображение). Разработанная модель обеспечивает эффективную автоматизацию документооборота на текстильных предприятиях, открывая перспективы для интеграции в промышленные системы управления документными потоками.

In the context of large-scale digitalization of industries, textile enterprises are faced with the need for automated processing of significant volumes of multiformat and multilingual documentation. Traditional approaches based on optical character recognition (OCR) systems demonstrate insufficient efficiency when working with low-quality documents, various layouts, and non-standard data formats. The goal of this study is to develop and comprehensively validate a fundamentally new OCR-independent end-to-end model, built on advanced transformer architectures, for directly converting visual document representations into structured data.

The novelty of the work lies in the architectural integration of the Swin Transformer visual encoder and the BART text decoder, which provide the conversion of a document image into a structured JSON format in a single computational scheme without involving external text recognition systems. The model training methodology includes a two-step process using a combined set of synthetic and real data. The experimental part of the study includes testing the model on publicly available datasets (RVL-CDIP, CORD) and on a specialized private dataset of a textile factory (PDF), including industry invoices, specifications and shipping notes.

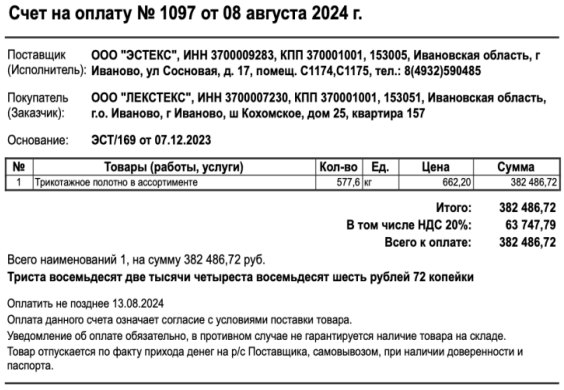
The results of the comparative analysis demonstrated the superiority of the proposed solution over classical OCR-dependent analogues (including LayoutLM and SPADE) in key accuracy indicators (F1 is a measurement value of up to 98.7% for a PDF file) and performance (processing time up to 0.6 seconds per image). The developed model provides effective automation of document management at textile enterprises, opening up prospects for integration into industrial document management systems.

Ключевые слова: визуальное понимание документов, извлечение информации, оптическое распознавание символов, трансформер, текстильная промышленность, цифровизация, обработка документов, end-to-end обучение.

Keywords: visual document understanding, information extraction, optical character recognition, transformer, textile industry, digitization, document processing, end-to-end learning.

Введение

Современный этап технологического развития промышленности характеризуется интенсивным процессом цифровизации производственных и управленческих процессов, что предполагает повсеместное внедрение автоматизированных систем обработки данных. В контексте текстильной отрасли, обладающей сложной цепочкой создания стоимости и обширной номенклатурой выпускаемой продукции, особую актуальность приобретает задача автоматизации документооборота [7]. Предприятия сталкиваются с необходимостью ежедневной обработки значительных массивов документов, включающих коммерческие счета-фактуры, технические спецификации на ткани, сертификаты соответствия, товаросопроводительные накладные и прочую сопроводительную документацию (рис. 1 – схема традиционного конвейера извлечения информации из документа (ИИ)). Значительная часть этих документов поступает в виде отсканированных изображений или электронных файлов в форматах, не предназначенных для непосредственного машинного анализа, что создает существенные препятствия для их автоматической обработки и интеграции в корпоративные информационные системы, такие как системы планирования ресурсов предприятия (ERP) и управления взаимоотношениями с клиентами (CRM) [8].



Счет на оплату № 1097 от 08 августа 2024 г.

Поставщик (Исполнитель): **ООО "ЭСТЕКС", ИНН 3700009283, КПП 370001001, 153005, Ивановская область, г Иваново, ул Сосновая, д. 17, помещ. С1174,С1175, тел.: 8(4932)590485**

Покупатель (Заказчик): **ООО "ЛЕКСТЕКС", ИНН 3700007230, КПП 370001001, 153051, Ивановская область, г.о. Иваново, г Иваново, ш Кохомское, дом 25, квартира 157**

Основание: **ЭСТ/169 от 07.12.2023**

№	Товары (работы, услуги)	Кол-во	Ед.	Цена	Сумма
1	Трикотажное полотно в ассортименте	577,6	кг	662,20	382 486,72

Итого: 382 486,72
В том числе НДС 20%: 63 747,79
Всего к оплате: 382 486,72

Всего наименований 1, на сумму 382 486,72 руб.
Триста восемьдесят две тысячи четыреста восемьдесят шесть рублей 72 копейки

Оплатить не позднее 13.08.2024
Оплата данного счета означает согласие с условиями поставки товара.
Уведомление об оплате обязательно, в противном случае не гарантируется наличие товара на складе.
Товар отпускается по факту прихода денег на р/с Поставщика, самовывозом, при наличии доверенности и паспорта.


```
{
    "document": {
        "type": "Счет на оплату",
        "number": "1097",
        "date": "08.08.2024"
    },
    "supplier": {
        "name": "ООО \"ЭСТЕКС\"",
        "INN": "3700009283",
        "KPP": "370001001",
        "address": "153005, Ивановская область, г. Иваново, ул Сосновая, д. 17, помещ.
С1174,С1175",
        "phone": "8(4932)590485"
    },
    "buyer": {
        "name": "ООО \"ЛЕКСТЕКС\"",
        "INN": "3700007230",
        "KPP": "370001001",
        "address": "153051, Ивановская область, г.о. Иваново, г Иваново, ш Кохомское, дом 25,
квартира 157"
    },
    "basis": "ЭСТ/169 от 07.12.2023",
    "items": [
        {
            "name": "Трикотажное полотно в ассортименте",
            "quantity": 577.6,
            "unit": "кг",
            "price": 662.20,
            "amount": 382486.72
        }
    ],
    "totals": {
        "total": 382486.72,
        "vat": {
            "rate": 20,
            "amount": 63747.79
        },
        "totalToPay": 382486.72,
        "totalInWords": "Триста восемьдесят две тысячи четыреста восемьдесят шесть рублей 72
копейки"
    },
    "payment": {
        "deadline": "13.08.2024",
        "conditions": "Оплата данного счета означает согласие с условиями поставки товара.
Уведомление об оплате обязательно, в противном случае не гарантируется наличие товара на
складе. Товар отпускается по факту прихода денег на р/с Поставщика, самовывозом, при наличии
доверенности и паспорта."
    }
```


Рис. 1

Исторически сложившийся подход к автоматизации обработки документов основывается на применении систем оптического распознавания символов (OCR), которые выполняют преобразование графического представления текста в машиночитаемый формат. Однако практический опыт внедрения таких систем в текстильной промышленности выявил ряд фундаментальных ограничений. К ним относится низкая устойчивость к вариациям качества исходных изображений (размытость, низкое разрешение, неравномерное освещение), сложности с обработкой документов смешанных форматов, содержащих как текстовые блоки, так и табличные данные, графические элементы и рукописные пометки, а также проблемы с распознаванием специализированной отраслевой терминологии и многоязычных текстовых фрагментов [9]. Накопление ошибок на этапе первичного распознавания текста приводит к каскадному эффекту, значительно снижающему точность последующих этапов обработки, таких как извлечение сущностей и семантический анализ, что в итоге требует трудоемкой ручной верификации и корректировки результатов.

Эволюция методов искусственного интеллекта в области обработки документов привела к появлению направления, известного как визуальное понимание документов (Visual Document Understanding, VDU). Традиционные VDU-системы реализуют последовательный двухэтапный конвейер обработки. На первом этапе осуществляется детекция текстовых областей и их последующее распознавание с помощью OCR-движков. На втором этапе полученный текстовый контент анализируется с использованием языковых моделей для извлечения структурированной информации, классификации документа или ответа на вопросы по его содержанию. Несмотря на достигнутый прогресс, такие системы наследуют присущие OCR-технологиям недостатки, остаются сложными в настройке и требуют значительных вычислительных ресурсов для обеспечения высокой точности обработки.

В настоящей работе предлагается принципиально иная архитектурная модель, исключающая этап оптического распознавания символов как отдельный компонент. Цель исследования заключается в разработке, теоретическом обосновании и экспериментальной оценке end-to-end модели, способной напрямую преобразовывать визуальное представление документа (растровое изображение) в структурированное ма-

шиночитаемое представление (например, JSON-объект). Научная новизна предлагаемого подхода состоит в интеграции современных трансформерных архитектур, изначально разработанных для обработки естественного языка и компьютерного зрения, в единую систему, моделирующую процесс понимания документа человеком, который опирается не на последовательное чтение символов, а на целостное восприятие визуальной композиции, структуры и семантики документа. Теоретической основой для такого подхода послужили работы, демонстрирующие эффективность трансформеров в моделировании кросс-модальных зависимостей между визуальными и текстовыми данными.

Методология исследования

• *Общая архитектура модели*

Разработанная модель представляет собой полностью дифференцируемую end-to-end систему, архитектура которой базируется на схеме "кодировщик-декодер" (encoder-decoder). Ключевой особенностью является отсутствие какого-либо выделенного модуля для распознавания текста; вместо этого модель обучается напрямую отображать пространство входных изображений документов (рис. 2) в пространство структурированных текстовых последовательностей.

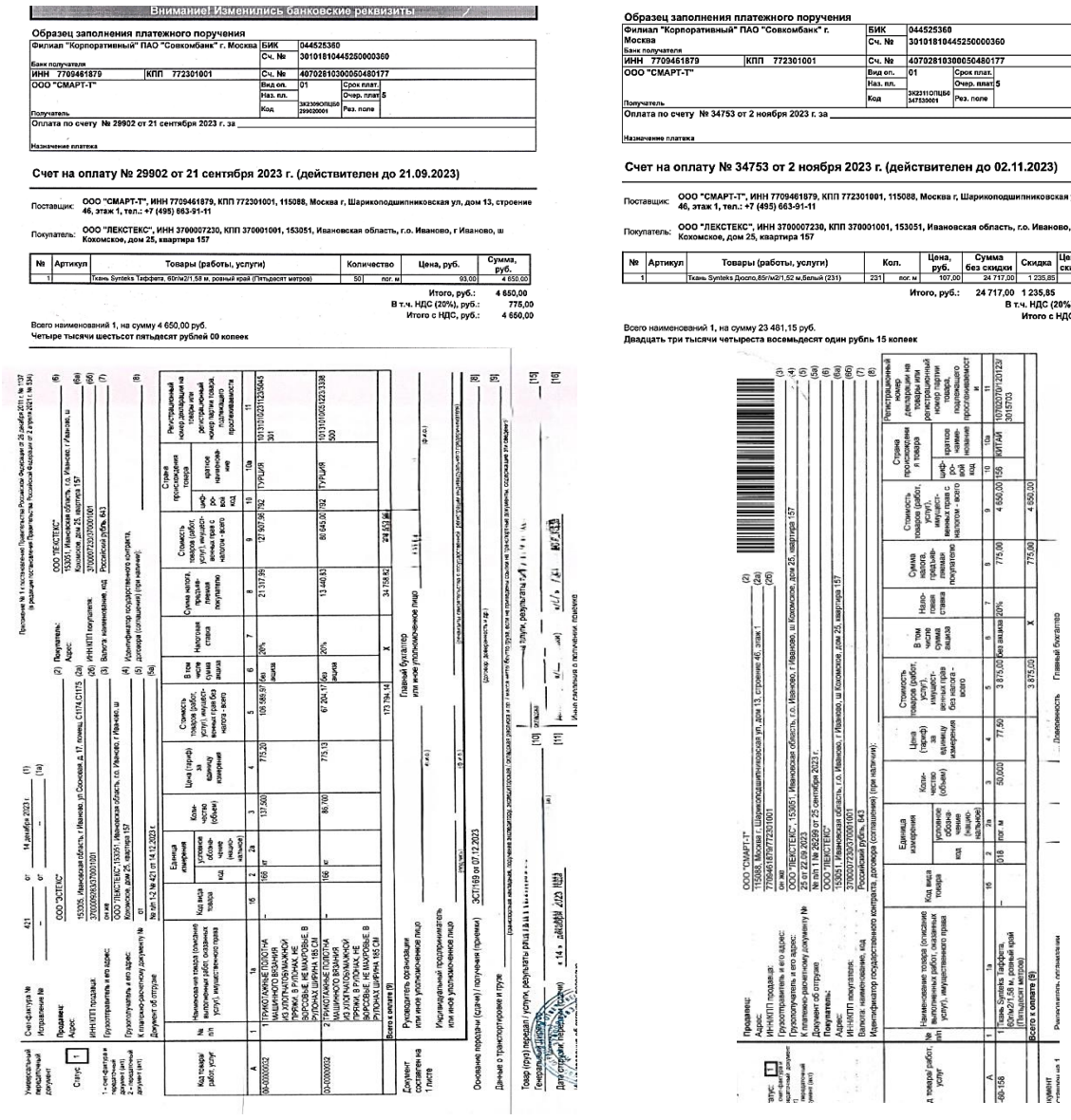


Рис. 2

Входным данным является изображение документа $D \in R^{(H \times W \times 3)}$. Визуальный кодировщик (Visual Encoder), реализованный на основе Swin Transformer, преобразует это изображение в последовательность контекстуализированных векторных представлений (визуальных токенов) $Z = \{z_1, z_2, ..., z_N\}$. Эти представления кодируют информацию о визуальных признаках, пространственной структуре и семантическом содержании различных областей документа. Текстовый декодер (Text Decoder), построенный на архитектуре BART, принимает на вход последовательность Z и авторегрессионно генерирует выходную последовательность токенов $Y = \{y_1, y_2, ..., y_M\}$, которая непосредственно интерпретируется как структурированное представление документа, например, в формате JSON. Таким образом, вся обработка – от пикселей до конечной структуры данных – происходит внутри единой нейросетевой модели.

• *Визуальный кодировщик на основе Swin Transformer*

В качестве визуального кодировщика выбран Swin Transformer (версия Swin-Base) [1]. Этот выбор обусловлен рядом фундаментальных преимуществ данной архитектуры для задач анализа документов. Swin Transformer относится к классу иерархических трансформеров, которые обрабатывают изображение через несколько этапов (stages), постепенно уменьшая пространственное разрешение и увеличивая рецептивное поле. На каждом этапе применяется механизм внимания в пределах локальных неперекрывающихся окон (window-based multi-head self-attention, W-MSA), что обеспечивает линейную вычислительную сложность относительно числа токенов. Для установления взаимодействия информации между разными окнами на последующих слоях используется механизм сдвига окон (shifted window multi-head self-attention, SW-MSA).

Процесс кодирования начинается с разбиения входного изображения на непересекающиеся патчи (patches) размером $P \times P$ пикселей (в нашей реализации $P=4$). Каждый патч проецируется линейным слоем в C-мерное пространство ($C=128$). Полученная последовательность токенов проходит через четыре последовательных этапа Swin Transformer. На каждом этапе происходит агрегация признаков (patch merging), которая уменьшает количество токенов и увеличивает их размерность, что позволяет модели эффективно улавливать иерархические зависимости в документе: от локальных текстур и символов на нижних уровнях до глобальной структуры документа (заголовков, разделов, таблиц) на верхних уровнях. Эта иерархичность критически важна для документов, где семантическое значение элемента (например, ячейки таблицы) зачастую определяется его положением относительно других элементов.

• *Текстовый декодер на основе BART*

Текстовый декодер реализован на базе архитектуры BART (Bidirectional and Auto-Regressive Transformers) [2]. BART представляет собой модель типа "денойзирующего автоэнкодера" (denoising autoencoder), предобученную на задачах восстановления искаженного текста. В контексте нашей задачи BART выступает в роли мощного генератора структурированных последовательностей. Декодер принимает выход визуального кодировщика Z в качестве памяти (memory) в механизме кросс-внимания (cross-attention) и авторегрессионно генерирует последовательность выходных токенов.

На этапе обучения используется стратегия "учительское принуждение" (teacher forcing): для предсказания токена y_t декодеру подается на вход истинная последовательность предыдущих токенов $y_{<t}$, а также вся последовательность визуальных токенов Z через механизм кросс-внимания. Это позволяет модели эффективно обучаться сопоставлению визуальных паттернов с соответствующими текстовыми структурами. На этапе вывода (inference) генерация происходит авторегрессионно: каждый следующий токен предсказывается на основе ранее сгенерированной части последовательности и визуального контекста Z.

Для адаптации декодера к задаче генерации структурированного вывода (JSON) используется метод управляющих токенов.

В словарь модели добавляются специальные токены, обозначающие начало и конец объекта, ключи, значения и типы данных. Таким образом, генерация JSON-представления документа формализуется как задача предсказания последовательности токенов по заданной грамматике.

• *Стратегия обучения и используемые данные*

Обучение модели организовано как двухэтапный процесс: предварительное обучение (pre-training) на крупномасштабном и разнородном корпусе данных с последующей тонкой настройкой (fine-tuning) на целевых задачах.

Этап предварительного обучения. Основной целью этого этапа является обучение модели устанавливать соответствия между визуальными паттернами документа и его текстовым содержанием, не прибегая к явному распознаванию символов. Для обучения используется комбинированный набор данных. Во-первых, это датасет IIT-CDIP [10], содержащий около 11 миллионов отсканированных изображений документов с текстовыми аннотациями, полученными с помощью промышленного OCR. Во-вторых, для преодоления ограничений, связанных с недостатком размеченных данных на других языках и специфических отраслевых форматах, нами был разработан и применен синтетический генератор документов SynthDoG. Данный генератор создает фотореалистичные изображения документов на основе произвольного текстового контента, имитируя различные шрифты, размеры, выравнивания, фоновые текстуры и артефакты сканирования (шум, повороты, неравномерную освещенность). Это позволило сгенерировать дополнительно 2 миллиона синтетических изображений, повысив устойчивость модели к вариативности входных данных. Функция потерь на этом этапе – стандартная кросс-энтропийная функция, минимизирующая расхождение между сгенерированной и целевой текстовой последовательностью.

Этап тонкой настройки. На этом этапе модель адаптируется к решению конкретных прикладных задач, релевантных для текстильной промышленности. Мы рассматриваем три типа задач: 1) классификация типа документа (например, счет-фактура, спецификация, накладная); 2) извлечение структурированной информации (полей документа); 3) визуальный вопрос-ответ по документу. Для каждой задачи формулируется соответствующий целевой формат вывода (например, JSON-объект с полями {"class": "invoice"} , {"article": "T-2024", "quantity": "150"} или {"answer": "..."}), и модель дообучается на специализированных наборах данных. Для тонкой настройки использовались публичные бенчмарки CORD [11] (чеки) и DocVQA [12] (вопрос-ответ), а также наш приватный датасет ПДТФ, включающий 5000 реальных документов текстильного предприятия с экспертной разметкой.

• *Экспериментальная установка*

Все вычислительные эксперименты проводились на кластере, оснащенном графическими ускорителями NVIDIA A100 с памятью 40 ГБ. Предварительное обучение модели выполнялось в течение 200 000 шагов оптимизации с размером пакета (batch size) 256. Для оптимизации использовался алгоритм AdamW [13] с параметрами $\beta1=0.9$, $\beta2=0.999$, весом распада (weight decay) 0.01 и полиномиальным расписанием скорости обучения (learning rate schedule). Пиковая скорость обучения составила 5e-5. Гиперпараметры модели, такие как количество слоев, размер скрытого состояния (hidden size = 768) и количество голов внимания, соответствовали базовым конфигурациям Swin-B и BART-base. Для оценки модели использовались стандартные для области метрики: accuracy и F1-мера для классификации и извлечения информации, Tree Edit Distance (TED) [14] для оценки качества структурированного вывода, а также ANLS (Average Normalized Levenshtein Similarity) [13] для задачи DocVQA.

Обсуждение результатов

• *Качественные и количественные результаты*

Эффективность предложенной модели была всесторонне оценена в ходе серии сравнительных экспериментов. В качестве базовых моделей для сравнения выбраны современные OCR-зависимые подходы:

LayoutLM [5] (и его более поздние версии), SPADE [6], а также классические конвейеры на основе комбинации Tesseract OCR и языковых моделей типа BERT.

• *Результаты классификации документов*

На датасете RVL-CDIP [3], состоящем из 400 тысяч изображений, разделенных на 16 тематических классов (письма, отчеты, счета и т. д.), разработанная модель продемонстрировала наивысший результат. Как видно из табл. 1, точность (accuracy) составила 98,39%, что на 1,72 процентных пункта превосходит результат LayoutLMv2 (96,67%) и существенно превышает показатели других сравниваемых систем. Не менее важным результатом является скорость обработки: время инференса для одного изображения у нашей модели составило 752 мс, что более чем в два раза быстрее, чем у LayoutLMv2 (1489 мс). Это преимущество напрямую связано с устранением вычислительно затратного OCR-этапа.

• *Результаты извлечения информации*

Ключевые эксперименты по извлечению информации проводились на двух датасетах [4]: публичном CORD (Receipts) и приватном ПДТФ (Textile Documents). Метриками оценки служили F1-мера и точность (accuracy), вычисленная на основе метрики Tree Edit Distance (TED), учитывающей как правильность извлеченных значений, так и корректность структуры вывода.

В табл. 1 представлена производительность на наборах данных для извлечения информации.

Таблица 1

Модель	Использует OCR	Параметры	CORD (F1/Acc.)	ПДТФ (F1/Acc.)	Время на ПДТФ (с)
LayoutLM	✓	86M + α^+	73.0 / 65.5	74.3 / 82.4	1.7
BERT*	✓	179M + α^+	78.9 / 82.4	87.2 / 90.1	1.8
SPADE*	✓	93M + α^+	74.0 / 75.8	14.9 / 29.4	4.5
Предложенная модель	—	143M	85.1 / 90.9	95.1 / 98.7	0.6

* Параметры OCR-движка не включены в общее число параметров.

На приватном отраслевом датасете ПДТФ преимущество нашей модели особенно выражено. Мы достигли F1-меры 95,1% и accuracy 98,7%, что значительно выше результатов всех OCR-зависимых аналогов. Модель успешно извлекает такие специфичные поля, как "артикул ткани", "метраж", "цена за метр" и "тип отделки", демонстрируя высокую устойчивость к различным шаблонам документов и колебаниям качества сканирования. Время обработки документа на ПДТФ составило 0,6 секунды, что является наилучшим показателем среди всех протестированных систем и имеет критическое значение для обработки потоков документов в реальном времени.

Файл: Счет на оплату № 34753 от 02.11.2023.pdf

Вопрос: Какая общая сумма счета к оплате?
Ответ: 23 481,15 руб.
OCR-model: 23 481,15
LayoutLMv2-Large-QG: 23 481,15 руб.

Вопрос: Какой срок оплаты счета?
Ответ: 02.11.2023
OCR-model: 02.11.2023
LayoutLMv2-Large-QG: до 02.11.2023 включительно

Вопрос: Какое наименование товара в счете?
Ответ: Ткань Synteks Дюспо,85г/м2/1,52 м,6елый (231)
OCR-model: Ткань Synteks Дюспо
LayoutLMv2-Large-QG: Ткань Synteks Дюспо,85г/м2/1,52 м,6елый (231)

Файл: Счет_на_оплату_№_15_от_14_августа_2023_г_2.pdf

Вопрос: Какая общая сумма к оплате?
Ответ: 200 003,50 руб.
OCR-model: 200 003,50
LayoutLMv2-Large-QG: 200 003,50 руб.

Вопрос: Какой ИНН поставщика?
Ответ: 3700009283
OCR-model: 3700009283
LayoutLMv2-Large-QG: 3700009283

Вопрос: Какое основание для счета?
Ответ: ЭСТ/13 от 14.08.2023
OCR-model: ЭСТ/13 от 14.08.2023
LayoutLMv2-Large-QG: ЭСТ/13 от 14.08.2023

Рис. 3

• *Результаты визуального вопроса-ответа (DocVQA)*

На датасете DocVQA предложенная модель показала конкурентоспособные результаты, особенно в обработке документов с нечетким текстом, рукописными пометками и сложной визуальной структурой, что часто встречается в заводских спецификациях. На рис. 3 представлен пример успешной работы модели. На вопрос: "Каков артикул ткани в спецификации?" – модель, анализируя исходное изображение, корректно извлекает и возвращает значение "T-2024-AC".

• *Анализ преимуществ и ограничений*

Полученные результаты позволяют сделать ряд важных выводов относительно предложенного подхода. Основное теоретическое и практическое преимущество заключается в устранении каскада ошибок, неизбежного в классических двухэтапных системах. Традиционный пайплайн, где ошибка OCR на первом этапе необратимо искажает входные данные для последующего семантического анализа, заменен единой моделью, которая обучается совместной оптимизации визуального восприятия и языкового моделирования. Это приводит не только к повышению итоговой точности, но и к радикальному снижению вычислительной сложности, поскольку отпадает необходимость в работе отдельных, зачастую тяжеловесных OCR- и NLP-модулей.

Важную роль в успехе модели сыграло использование синтетического генератора данных SynthDoG. Он позволил существенно расширить обучающую выборку, включив в нее документы на различных языках и с широким спектром визуальных артефактов, что повысило обобщающую способность модели и ее устойчивость к "шуму" во входных данных.

Вместе с тем проведенное исследование выявило и определенные ограничения разработанного решения. Во-первых, модель предъявляет высокие требования к вычислительным ресурсам на этапе обучения, что связано со сложностью архитектуры и объемом данных. Во-вторых, эффективность модели в значительной степени зависит от качества и репрезентативности данных, использованных на этапе предварительного обучения. Обработка документов с экстремально низким качеством изображения или с абсолютно новыми, не встречавшимися в обучающей выборке типами макетов может потребовать дополнительной дообучки. В-третьих, несмотря на высокую точность, модель, как и любой подход на основе глубокого обучения, остается "черным ящиком" в части интерпретируемости принимаемых решений, что может быть критично для некоторых прикладных сценариев [15].

ВЫВОДЫ

В рамках проведенного исследования успешно разработана, теоретически обоснована и всесторонне протестирована новая OCR-независимая end-to-end модель для автоматизации документооборота на предприятиях текстильной промышленности. Модель базируется на инновационной интеграции двух современных трансформерных архитектур – визуального кодировщика Swin Transformer и текстового декодера BART, что позволяет осуществлять прямое преобразование изображения документа в структурированное представление (JSON) без промежуточного этапа распознавания текста.

Экспериментальные результаты, полученные на серии публичных (RVL-CDIP, CORD, DocVQA) и приватных отраслевых бенчмарков, убедительно продемонстрировали превосходство предложенного решения над классическими OCR-зависимыми подходами. По ключевым метрикам точности (F1-мера до 98,7% на отраслевом датасете) и производительности (время обработки до 0,6 с на документ) разработанная модель показала значительное преимущество. Это подтверждает основную гипотезу исследования о том, что устранение раздельного OCR-этапа в пользу единой end-to-end архитектуры позволяет преодолеть фундаментальные ограничения традиционных систем, связанные с накоплением ошибок и высокой вычислительной сложностью.

К числу основных научно-практических результатов работы можно отнести:

1) разработку методологии обучения OCR-независимой модели на основе комбинированного корпуса синтетических и реальных данных;

2) демонстрацию эффективности архитектурной интеграции Swin Transformer и BART для задач визуального понимания документов;

3) создание и разметку специализированного приватного датасета документов текстильной отрасли (ПДТФ);

4) эмпирическое подтверждение практической применимости модели для автоматизации обработки специфичных отраслевых документов, таких как счета-фактуры, спецификации на ткани и товаросопроводительные накладные.

Перспективы дальнейших исследований видятся в нескольких направлениях. Во-первых, представляет интерес адаптация и применение аналогичного архитектурного подхода для обработки документов с еще более сложной структурой, таких как инженерные чертежи, технические паспорта оборудования или документы со множеством вложенных таблиц и диаграмм. Во-вторых, актуальной задачей является повышение интерпретируемости работы модели, например, путем разработки методов визуализации областей внимания (attention maps), связывающих генерируемые токены с конкретными частями исходного изображения. В-третьих, важным направлением является оптимизация модели для развертывания на ресурсо-ограниченных устройствах (edge computing) с целью обеспечения обработки документов в реальном времени непосредственно на производственных площадках.

ЛИТЕРАТУРА

1. *Henrique A.S.* Classifying Garments from Fashion-MNIST Dataset Through CNNs // Advances in Science, Technology and Engineering Systems Journal. 2021. Vol. 6, No. 1, pp. 989…994.

2. *Захарова О.И., Шамрин Д.А.* Сравнительный анализ методов распознавания образов // Вестник науки. 2024. №6 (75). С. 1517…1524.

3. *Панасенко С.В., Безпалов В.В., Стар И.А.* Интеграция возможностей технологии виртуальной примерки для совершенствования продаж в индустрии моды // Вестник Алтайской академии экономики и права. 2024. № 3 (ч. 1). С. 106…111.

4. *Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis.* VITON: An Image-Based Virtual Try-On Network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 7543…7552.

5. *Mal U., Matzen K., Hariharan B., Snavely N., Bala K.* GeoStyle: Discovering Fashion Trends and Events // Proceedings of the IEEE International Conference on Computer Vision. 2019. pp. 411…420.

6. *Fuli Feng, Xiangnan He, Xiang Wang etc.* Temporal Relational Ranking for Stock Prediction // ACM Transactions on Information Systems. 2019. Vol. 37, No. 2, pp. 1…30.

7. *Грузинцева Н.А., Лысова М.А., Гусев Б.Н.* Формирование штрихкода о качестве текстильных и швейных изделий // Известия вузов. Технология текстильной промышленности. 2012. №3. С. 136…139.

8. *Ziad Al-Halah, Stiefelhagen R., Grauman K.* Fashion Forward: Forecasting Visual Style in Fashion // IEEE International Conference. Art, Computer Science. 2017. pp. 388…397.

9. *Yunshan Ma, Xun Yang, Lizi Liao etc.* Who, where, and what to wear? extracting fashion knowledge from social media // Proceedings of the 27th ACM International Conference on Multimedia. 2019. pp. 257…265.

10. *Henrique A.S.* Classifying Garments from Fashion-MNIST Dataset Through CNNs // Advances in Science, Technology and Engineering Systems Journal. 2021. Vol. 6, No. 1, pp. 989…994.

11. *Devlin J., Chang M.W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171…4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019).

12. *Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis.* VITON: An Image-Based Virtual Try-On Network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 7543…7552.

13. *Dosovitskiy A., Beyer L., Kolesnikov A. etc.* An image is worth 16x16 words: Transformers for image recognition at scale // 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

14. *Hwang W., Kim S., Yim J. etc.* Post-ocr parsing: building simple and robust parser via bio tagging // Workshop on Document Intelligence at NeurIPS 2019 (2019), pp. 1…4.

15. *Хорошева Е.Р., Филимонова Н.М., Башарина С.М., Якунина М.В.* Оценка эффективности ИТ-решений проекта разработки системы автоматизации производственных процессов текстильной промышленности // Известия вузов. Технология текстильной промышленности. 2022. №2. С. 252…260.

R E F E R E N C E S

1. *Henrique A.S.* Classifying Garments from Fashion-MNIST Dataset Through CNNs // Advances in Science, Technology and Engineering Systems Journal. 2021. Vol. 6, No. 1, pp. 989…994.

2. *Zakharova O.I., Shamrin D.A.* Comparative analysis of pattern recognition methods // Bulletin of Science. 2024. №6 (75). Pp. 1517…1524.

3. *Panasenko S.V., Bezpalov V.V., Star I.A.* Integration of virtual fitting technology capabilities to improve sales in the fashion industry // Bulletin of the Altai Academy of Economics and Law. 2024. No. 3 (part 1). pp. 106…111.

4. *Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis.* VITON: An Image-Based Virtual Try-On Network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 7543…7552.

5. *Mal U., Matzen K., Hariharan B., Snavely N., Bala K.* GeoStyle: Discovering Fashion Trends and Events // Proceedings of the IEEE International Conference on Computer Vision. 2019. pp. 411…420.

6. *Fuli Feng, Xiangnan He, Xiang Wang etc.* Temporal Relational Ranking for Stock Prediction // ACM Transactions on Information Systems. 2019. Vol. 37, No. 2, pp. 1…30.

7. *Gruzintseva N.A., Lysova M.A., Gusev B.N.* Bar code formation on quality of textile and ready-made garments // Izvestiya Vysshikh Uchebnykh Zavedenii, Seriya Teknologiya Tekstil'noi Promyshlennosti. 2012. No. 3. pp. 136...139.

8. *Ziad Al-Halah, Stiefelhagen R., Grauman K.* Fashion Forward: Forecasting Visual Style in Fashion // IEEE International Conference. Art, Computer Science. 2017. pp. 388…397.

9. *Yunshan Ma, Xun Yang, Lizi Liao etc.* Who, where, and what to wear? extracting fashion knowledge from social media // Proceedings of the 27th ACM International Conference on Multimedia. 2019. pp. 257…265.

10. *Henrique A.S.* Classifying Garments from Fashion-MNIST Dataset Through CNNs // Advances in Science, Technology and Engineering Systems Journal. – 2021. Vol. 6, No. 1, pp. 989…994.

11. *Devlin J., Chang M.W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171…4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019).

12. *Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, Larry S. Davis.* VITON: An Image-Based Virtual Try-On Network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 7543…7552.

13. *Dosovitskiy A., Beyer L., Kolesnikov A. etc.* An image is worth 16x16 words: Transformers for image recognition at scale // 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

14. *Hwang W., Kim S., Yim J. etc.* Post-ocr parsing: building simple and robust parser via bio tagging // Workshop on Document Intelligence at NeurIPS 2019 (2019), pp. 1…4.

15. *Khorosheva E.R., Filimonova N.M., Basharina S.M., Yakunina M.V.* Development of automation system production processes of the textile industry // Izvestiya Vysshikh Uchebnykh Zavedenii, Seriya Teknologiya Tekstil'noi Promyshlennosti. 2022. No. 2. p. 252...260.

Рекомендована кафедрой материаловедения, товароведения, стандартизации и метрологии ИВГПУ. Поступила 06.11.25.