УДК 004.93
DOI: 10.26102/2310-6018/2025.50.3.006

Разработка легковесной модели автоматической классификации структурированных и неструктурированных данных в потоковых источниках для оптимизации оптического распознавания символов

В.С. Гаврилов✉, С.А. Корчагин, В.И. Долгов, Н.А. Андриянов

Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация

Резюме. В настоящей статье рассмотрена задача предварительной оценки входящего электронного документооборота на основе технологий компьютерного зрения. Авторами был синтезирован датасет изображений со структурированными данными на основе формы счета-фактуры, а также собраны сканы различных документов от страниц научных статей и документации в электронном почтовом ящике научной организации до отчетности Росстата. Таким образом, первая часть датасета относится к структурированным данным, имеющим строгую форму, а вторая часть относится к неструктурированным сканам, поскольку на разных отсканированных документах информация может быть представлена по-разному: только текст, текст и изображения, графики, так как разные источники имеют разные требования и свои стандарты. Первичный анализ данных в потоковых источниках можно делать с помощью моделей компьютерного зрения. Проведенные эксперименты показали высокую точность работы сверточных нейронных сетей. В частности, для нейросети с архитектурой Xception достигается результат с точностью более 99 %. Преимущество по сравнению с более простой моделью MobileNetV2 достигает около 9 %. Предложенный подход позволит проводить первичную фильтрацию документов по отделам без применения больших языковых моделей и моделей распознавания символов, что обеспечит повышение скорости и снижение вычислительных затрат.

Ключевые слова: интеллектуальная обработка документов, компьютерное зрение, сверточные нейронные сети, обработка потоковых данных, машинное обучение.

Благодарности: Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финансового университета при Правительстве РФ.

Для цитирования: Гаврилов В.С., Корчагин С.А., Долгов В.И., Андриянов Н.А. Разработка легковесной модели автоматической классификации структурированных и неструктурированных данных в потоковых источниках для оптимизации оптического распознавания символов. *Моделирование, оптимизация и информационные технологии.* 2025;13(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1918 DOI: 10.26102/2310-6018/2025.50.3.006

Development of a lightweight model for automatic classification of structured and unstructured data in streaming sources to optimize optical character recognition

V.S. Gavrilov✉, S.A. Korchagin, V.I. Dolgov, N.A. Andriyanov

Financial University under the Government of the Russian Federation, Moscow, the Russian Federation

Abstract. This article discusses the task of preliminary assessment of incoming electronic document management based on computer vision technologies. The authors synthesized a dataset of images with

structured data based on the invoice form and also collected scans of various documents from pages of scientific articles and documentation in the electronic mailbox of a scientific organization to Rosstat reports. Thus, the first part of the dataset refers to structured data with a strict form, and the second part refers to unstructured scans, since information can be presented in different ways on different scanned documents: only text, text and images, graphs, since different sources have different requirements and their own standards. The primary analysis of data in streaming sources can be done using computer vision models. The experiments performed have shown high accuracy of convolutional neural networks. In particular, for a neural network with the Xception architecture, the result is achieved with an accuracy of more than 99%. The advantage over the simpler MobileNetV2 model is about 9%. The proposed approach will allow for the primary filtering of documents by department without using large language and character recognition models, which will increase speed and reduce computational costs.

Keywords: intelligent document processing, computer vision, convolutional neural networks, stream data processing, machine learning.

Acknowledgements: The article is based on the results of the research carried out at the expense of budgetary funds under the state assignment of Financial University under the Government of the Russian Federation.

For citation: Gavrilov V.S., Korchagin S.A., Dolgov V.I., Andriyanov N.A. Development of a lightweight model for automatic classification of structured and unstructured data in streaming sources to optimize optical character recognition. *Modeling, Optimization and Information Technology*. 2025;13(3). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1918 DOI: 10.26102/2310-6018/2025.50.3.006

Введение

Оптическое распознавание символов является одним из ключевых направлений в развитии технологий компьютерного зрения. Связано это с широкой прикладной ценностью автоматизации ручной работы по оцифровке материалов изображения. Развитие моделей и методов в области оптического распознавания символов помогает ускорить перевод символов в машиночитаемый формат при одновременном увеличении точности и эффективности поиска необходимой информации [1]. OCR нашло широкое применение во многих областях, где есть большое количество не оцифрованных данных, в том числе рукописный текст, например, в библиотеках и архивах. Однако особенно ценным OCR стало в бизнесе, поскольку внедрение такой технологии помогает сократить издержки на обработку стандартизированных документов, таких как счет-фактура, фискальные чеки и товарные накладные [2, 3]. Другая прикладная ценность задачи OCR лежит в онлайн-распознавании, то есть возможность детектировать и классифицировать объекты в режиме реального времени [4]. Но несмотря на значительные успехи в области оптического распознавания символов, задачи, связанные с низким качеством изображений, сложными шрифтами и нестандартным расположением текста, остаются актуальными [1, 4]. Кроме того, современные модели компьютерного зрения на основе «трансформерных» архитектур, хоть и способны давать высокое качество задачи распознавания символов, требуют достаточно мощных вычислительных ресурсов [5, 6].

Современные решения по обработке документов часто предполагают применение одинаковых моделей OCR ко всем входящим изображениям, что неэффективно с точки зрения вычислений и логики обработки [7]. Между тем, важным предварительным этапом может быть классификация изображений по признаку структурированности [8]. Структурированные документы, такие как счета-фактуры, чеки, накладные, как правило, имеют стабильную структуру и могут быть успешно обработаны базовыми OCR-моделями. В то же время неструктурированные документы – письма, сканы с произвольной компоновкой текста, формы свободного формата – требуют не только

распознавания текста, но и его последующего анализа, что предполагает использование более сложных решений на базе систем интеллектуальной обработки документов (IDP).

Таким образом, задача данной статьи не сводится к задаче разработки OCR-модели, а заключается в создании легковесной модели классификации изображений на изображения со структурированными и неструктурированными данными, что позволяет оптимизировать дальнейший пайплайн обработки изображений из потоковых источников. Классифицированные структурированные документы будут направлены напрямую в OCR-систему с минимальной предобработкой, а неструктурированные – на последующий анализ с привлечением IDP-моделей. Кроме того, статья демонстрирует, как результаты такого подхода можно автоматически переводить в структурированные JSON-форматы, что важно для интеграции в корпоративные системы электронного документооборота.

Цель исследования: разработка легковесной модели для автоматической классификации документов на структурированные и неструктурированные в потоках изображений, с целью повышения эффективности последующей обработки с применением OCR и IDP.

Задачи исследования:

– Провести обзор существующих методов и моделей оптического распознавания символов.

– Формализовать задачу OCR с математической точки зрения.

– Подобрать оптимальную модель для классификации изображений со структурированными и неструктурированными данными.

– Оценить качество и применимость разработанной модели.

На сегодняшний день существует три основные группы моделей оптического распознавания символов [5].

1. Модели на основе шаблонов.
2. Модели на основе классического машинного обучения.
3. Модели на основе глубокого обучения.

Первая группа методов использует наперед заданные шаблоны символов, которые необходимо распознать. На первом этапе происходит сегментация изображения на определенные символы и слова, а на втором этапе происходит шаблонное сопоставление сегментов с наперед заданными шаблонами. Такие модели не подразумевают никакого обучения. Вторая группа методов основана на извлечении признаков из изображения и обучении на размеченных данных. Далее с помощью обученных моделей классифицируются символы. Данная группа моделей характеризуется относительной легковесностью и высокой эффективностью, однако в некоторых задачах классификации такие модели могут показывать низкие метрики качества из-за особенности поступающих изображений. Третья группа моделей появилась относительно недавно и решает указанную задачу лучше, чем модели из первой и второй группы [1, 9]. Модели на основе глубокого обучения или нейросетевые модели являются флагманскими моделями в задачах оптического распознавания символов, зарекомендовав себя как очень точные и эффективные инструменты для оцифровки информации из изображений [5, 10]. Обратной стороной таких точных и устойчивых к изменению поступающих изображений моделей является их тяжеловесность, большие вычислительные мощности для размещения таких моделей и большие издержки для их поддержания [1].

Наиболее «сильными» нейросетевыми моделями в первом квартале 2025 года являются нейросетевые модели семейства OpenAI, Qwen, Claude 3.x, Gemini, TrOCR, EasyOCR, Florence, Moondream, Idefics [2]. Для сравнения качества было отобрано случайным образом по десять изображений из каждого доменного набора данных из

Roboflow Universe. При выборе изображений использовался критерий «человекочитаемости»: если изображение может было легко прочитано и расшифровано человеком, оно включено в тестовый набор данных. Для более качественного сравнения результатов распознавания моделями, каждое изображение было обработано вручную: добавлены комментарии и непосредственно сам текст на изображении размечен человеком. В качестве основной метрики качества использовалось расстояние Левенштейна, которое позволяет оценить различия между распознанным текстом и эталонным. Это особенно важно для задач OCR, где точность распознавания символов критична. В качестве контрольных метрик качества использовались время ответа и стоимость распознавания. Наилучшие результаты в точности показали модели GPT-4.5 (93,2 %), Gemini 2.0 (92,2 %), GPT-4o (92,1 %). Хуже всего оказалась модель Florence 2 Large (19.2 %). Также эта модель оказалась наиболее медленной (15 секунд), а наиболее быстрыми – Gemini 1.5 Flash-8B (0,57 с), TrOCR (0,22 с), EasyOCR (0,15 с). По третьей метрике (стоимости), модель Florence 2 Large показала себя как самая дорогая в пересчете на единицу инференса. За ней следуют модели OpenAI. Наиболее выгодными моделями являются open source модели, затраты на которые могут измеряться только стоимостью аренды виртуальной машины в Google Cloud. Если используется локальная машина, то стоимость единицы инференса равна нулю. Следует отметить, что поставленные эксперименты выполнялись на англоязычных текстах, содержащихся на изображениях.

Другими эффективными и хорошо зарекомендовавшими себя моделями на основе глубокого обучения считаются модели семейства YOLO (расшифровывается как You Only Look Once) [4–5]. Первая версия модели YOLOv1 вышла в 2015 году и обладала большим преимуществом в скорости детекции, что позволяло использовать ее в режиме реального времени. Она была основана на GoogLeNet и имела 24 сверточных слоя и 2 полносвязных на выходе. Главным новшеством YOLOv1 было то, что разработчики сформулировали и решили задачу детекции как задачу регрессии [4]. На начало 2025-го года самой актуальной версией YOLO является 12-я версия, которая была выпущена в начале 2025 года и претерпела большое количество изменений. Во-первых, в 12-ю версию было добавлено улучшенное извлечение признаков, включая более эффективную обработку больших рецептивных полей, улучшение баланса между вниманием и вычислениями в сети feed forward. Во-вторых, улучшена архитектура, включая уменьшение количества параметров при сохранении точности, исключение позиционного кодирования, оптимизация коэффициентов MLP для эффективного распределения вычислительных ресурсов, а также реализация архитектуры R-ELAN – агрегационные сети с эффективным остаточным уровнем. Имея преимущество в точности на 2,1 процентных пункта при одновременном увеличении скорости на 42 процентных пункта по сравнению с предыдущей 11-й версией и многократное преимущество в метриках качества относительно первой версии модели, данная нейросеть решает широкий спектр задач: от самых важных по обнаружению объектов, сегментации объектов, классификации объектов до специфичных, таких как ориентированное обнаружение объектов и оценка позы объекта [5]. Однако для обучения YOLOv12 на непосредственное распознавание символов требуется достаточно сложная разметка. Фактически необходимо размечать каждый символ на изображении отдельным объектом, а затем еще и разрабатывать алгоритмы анализа положений обнаруженных символов для получения слов и предложений. Поэтому использование YOLO видится, скорее, на предварительном этапе, например, для детектирования текстовых областей. Вместе с тем применять YOLO для задачи первичной фильтрации документов нецелесообразно в силу затрачиваемых вычислительных ресурсов.

Отдельно необходимо отметить открытые репозитории с обученными моделями оптического распознавания на GitHub. Наиболее популярными по метрике количества

добавлений в избранное среди разработчиков являются keras-ocr (1,4 тыс.), yolov12 (1,5 тыс.), EasyOCR (26,3 тыс.), PaddleOCR (48,3 тыс.), tesseract (63,1 тыс.). В первом репозитории представлен исходный код сверточной нейросети CRAFT, которая была разработана в 2019 году в рамках исследования [11] и была наиболее эффективной в свое время. Второй репозиторий связан с моделями семейства YOLO, которые были описаны выше. В третьем же репозитории представлен библиотечный код, фреймворк которого подразумевает несколько этапов: на первом этапе детекции используется та же нейросеть CRAFT (опционально другая аналогичная модель детекции), далее на этапе распознавания используются сети ResNet, LSTM, CTC (опционально другие аналогичные модели распознавания), после этого используется декодер. Таким образом, тут нет уникальной архитектуры и обученной нейросети, а вместо этого представлен фреймворк из ансамбля известных нейросетей. В четвертом репозитории PaddleOCR представлена нейросеть SLANet-LCNetV2 опубликованная в июле 2024 и разработанная командой распознавания OCR из исследовательской группы машинного обучения и когнитивных вычислений Пекинского университета Цзяотун. Особенностью нейросети SLANet-LCNetV2, которая поддерживает более 80-ти языков, является ее легкость, относительно небольшое количество параметров и возможность ее обучения и развертывания на мобильных устройствах. В пятом репозитории tesseract, принадлежащим компанией Google и имеющим историю разработки начиная с 1985 года (тогда в компании Hewlett Packard) представлена нейросеть Tesseract 5, которая основана на архитектуре LSTM – рекуррентной нейронной сети. В основе библиотеки не лежат передовые технологии, но есть большая прикладная ценность за счет сильного функционала библиотеки, включающего разнообразие входных и выходных форматов, более 100 языков, простоту использования и универсальность. Исследования показывают, что модель не имеет высокой точности в сложных задачах, однако же в некоторых задачах с шаблонизированными изображениями, показывает высокую точность и скорость, из-за чего широко используется для решения задач оптического распознавания символов, что и делает ее одной из самых популярных в мире.

Рассмотренные выше LLM/VLM модели являются «большими» моделями и не могут быть развернуты на локальной машине, что делает невозможным их поддержку и эксплуатацию на независимых серверах. Как было сказано ранее, модели на основе классического машинного обучения являются легковесными, при этом, как правило, имеют более низкие метрики качества по сравнению с тяжеловесными моделями на основе глубокого обучения. Существуют также компромиссные модели на основе глубокого обучения, которые не имеют сложной архитектуры по сравнению с большими моделями, но показывают лучшие метрики качества по сравнению с классическими моделями. Такими моделями являются нейросети с архитектурами VGG16, ResNet, Xception. Архитектуру VGG16 создали ученые из Оксфордского университета в 2014 году [12]. Архитектура ResNet была разработана в 2015 году и дала лучшие результаты на классификации ImageNet. Идея архитектуры в последствии была усовершенствована и ее идеи используются в тяжелых языковых моделях семейства OpenAI о которых упоминалось ранее [13]. Модель Xception была разработана в 2016 году и сохранила принцип легковесности с одновременным повышением уровня качества на том же ImageNet [14]. Основная идея архитектуры заключается в том, что вместо того, чтобы выбирать размер ядра, берется сразу несколько вариантов, которые используются все одновременно, а затем результаты конкатенируются. Чтобы существенно не увеличивать количество операций, перед каждым сверточным блоком делается свертка с размером ядра 1x1, которая снижает размерность сигнала, подающегося на вход сверткам с большими размерами ядер [14]. Исследования по возможности увеличения качества таких моделей за счет аугментации и технологий drop out [15] делает возможным их качественное обучение без переобучения, несмотря на легковесность.

Таким образом, современные модели оптического распознавания символов представляют собой мощный инструмент для автоматизации обработки текстовой информации, однако их эффективность во многом зависит от качества входных данных и сложности задачи. Разработка легковесной модели, способной эффективно разделять структурированные и неструктурированные данные в потоковых источниках, является актуальной задачей, которая может значительно оптимизировать процессы обработки поступающей информации. Это позволит более тонко настраивать дальнейшие модели OCR. Например, для изображений со структурированными данными не включать в пайплайн обработки YOLO детектор, а брать фиксированные области изображения для OCR.

Материалы и методы

Математическая формализация. Математическая постановка задачи оптического распознавания символов включает несколько этапов. На первом этапе формулируется постановка задачи, а на втором этапе формулируется математическая модель, которая, в свою очередь, включает в себя подэтапы детекции текста, распознавания символов и введения функции потерь.

Первый этап включает постановку задачи. Дано изображение I:

$$I \in \mathbb{R}^{H \times W \times C}, \tag{1}$$

где H – высота изображения; W – ширина изображения; C – равно 1, если изображение черно-белое, равно 3, если изображение RGB.

Требуется найти последовательность символов S:

$$S = (s_1, s_2, \dots, s_N), s_i \in \mathrm{A}, \tag{2}$$

где N – длина последовательности символов; A – алфавит символов.

Второй этап включает в себя математическую постановку задачи.
Детекция текста:

$$B = Detect(I);\ B = \{b_i\}_{i=1}^{M};\ b_i = (x, y, w, h), \tag{3}$$

где B – ограничивающие рамки текстовых областей.
Распознавание символов: для каждой области b_i

$$P\left(s \middle| I_{b_i}\right) = Recognize(I_{b_i}), s \in \mathrm{A}. \tag{4}$$

Оптимальная последовательность:

$$S^* = \arg\max_S P(S|I). \tag{5}$$

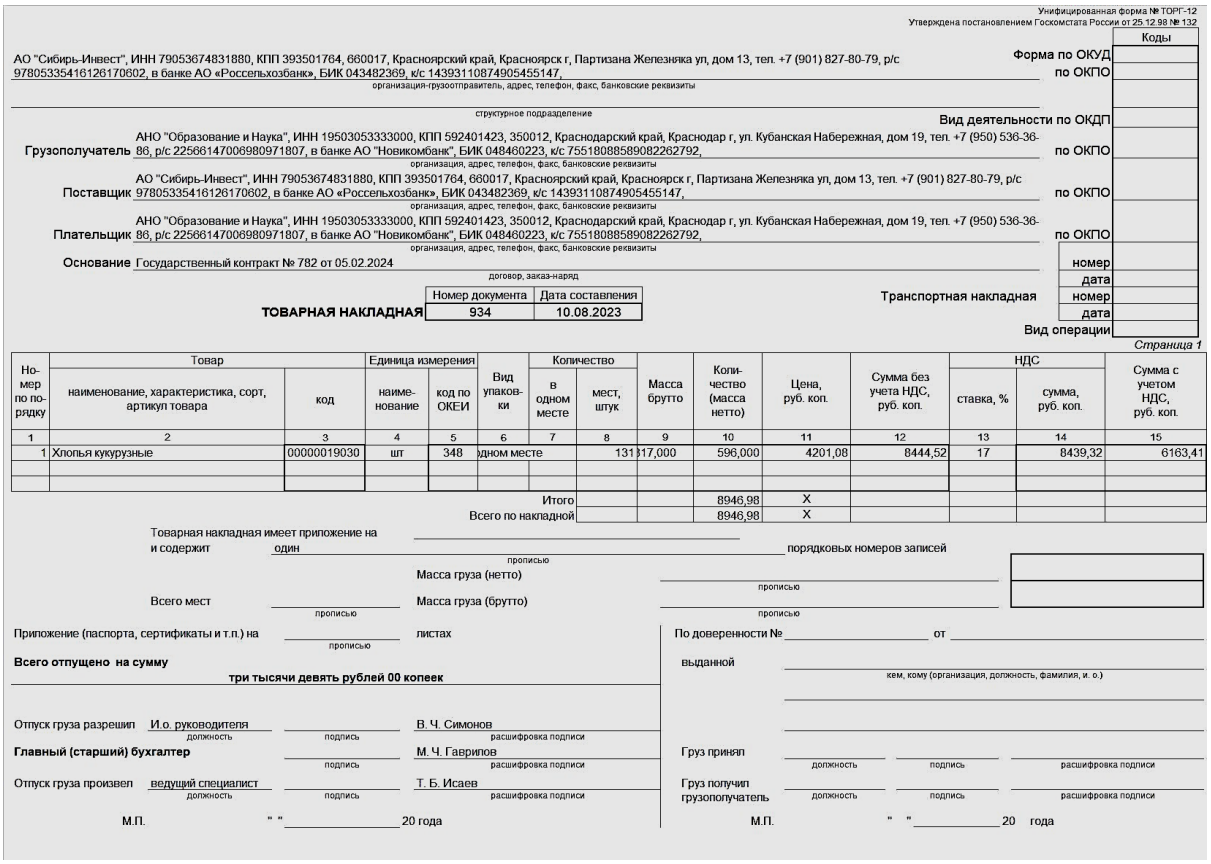
Для обучения модели используется L – функция потерь (loss). Оптимизационная задача выглядит следующим образом:

$$L = -\ln P(S|I) \to \min(S). \tag{6}$$

Используемые данные. Процесс генерации изображений со структурированными данными содержит подготовку Excel-файла, содержащего данные для каждого счета-фактуры, такие как наименование продавца и покупателя, ИНН и КПП, описание товаров или услуг, суммы, ставки НДС и другие обязательные реквизиты. Этот файл служит источником информации для генератора, который извлекает данные из таблицы и автоматически вставляет их в заранее подготовленный шаблон счета-фактуры. С помощью библиотеки *openpyxl*, данные из Excel-файла обрабатываются и заполняются в соответствующие ячейки шаблона. После того как все поля счета-фактуры заполнились необходимой информацией, система генерирует изображение в формате JPEG. Для этого используется библиотека *Pillow*, которая позволяет преобразовать подготовленный

документ в изображение. Это изображение затем используется для тестирования и верификации работы OCR-систем, имитируя реальное изображение счета-фактуры, которое может быть получено в процессе работы с реальными документами. Важную роль играют также документы типа «товарная накладная», которые можно генерировать в программе так что получаются изображения, не содержащие реальной информации о товарах и организациях. На Рисунке 1 показан пример сгенерированного изображения товарной накладной. Общее количество таких изображений со структурированными данными составляет 3060.

Процесс сбора второй части датасета включает в себя поиск научных статей в системе elibrary, электронных писем научной организации, а также документов и отчетностей с официального сайта Росстата. Собранные данные обладают вариативностью: страницы содержат произвольные комбинации текста, графиков, таблиц, формул, что необходимо для обеспечения репрезентативности данных Критерием отбора является стратификация по типам контента: 50 % изображений состоит из табличных страниц, 20 % изображений содержит текст с графиками/таблицами, 30 % изображений содержит смешанный контент, включающий любую комбинацию текста, графиков, формул и таблиц. Для охвата разных стилей оформления было выбрано 76 статей, к которым есть открытый бесплатный доступ из системы elibrary. Из почтового ящика научной организации было выбрано 785 сканов-документов. Из отчетности на электронном ресурсе Росстата выбрано 9 крупных отчетов. На последнем этапе проводилась ручная проверка человеком случайных изображений из выборки на соответствие критериям. Общее количество отобранных изображений с неструктурированными данными составило 3316.



Унифицированная форма № ТОРГ-12
Утверждена постановлением Госкомстата России от 25.12.98 № 132

Коды

АО "Сибирь-Инвест", ИНН 79053674831880, КПП 393501764, 660017, Красноярский край, Красноярск г, Партизана Железняка ул, дом 13, тел. +7 (901) 827-80-79, р/с 97805335416126170602, в банке АО «Россельхозбанк», БИК 043482369, к/с 14393110874905455147,
организация-грузоотправитель, адрес, телефон, факс, банковские реквизиты
Форма по ОКУД
по ОКПО

структурное подразделение
Вид деятельности по ОКДП

Грузополучатель АНО "Образование и Наука", ИНН 19503053333000, КПП 592401423, 350012, Краснодарский край, Краснодар г, ул. Кубанская Набережная, дом 19, тел. +7 (950) 536-36-86, р/с 22586147006980971807, в банке АО "Новикомбанк", БИК 048460223, к/с 75518088589082262792,
организация, адрес, телефон, факс, банковские реквизиты
по ОКПО

Поставщик АО "Сибирь-Инвест", ИНН 79053674831880, КПП 393501764, 660017, Красноярский край, Красноярск г, Партизана Железняка ул, дом 13, тел. +7 (901) 827-80-79, р/с 97805335416126170602, в банке АО «Россельхозбанк», БИК 043482369, к/с 14393110874905455147,
организация, адрес, телефон, факс, банковские реквизиты
по ОКПО

Плательщик АНО "Образование и Наука", ИНН 19503053333000, КПП 592401423, 350012, Краснодарский край, Краснодар г, ул. Кубанская Набережная, дом 19, тел. +7 (950) 536-36-86, р/с 22586147006980971807, в банке АО "Новикомбанк", БИК 048460223, к/с 75518088589082262792,
организация, адрес, телефон, факс, банковские реквизиты
по ОКПО

Основание Государственный контракт № 782 от 05.02.2024
договор, заказ-наряд
номер
дата

ТОВАРНАЯ НАКЛАДНАЯ

Номер документа	Дата составления
934	10.08.2023

Транспортная накладная
номер
дата
Вид операции

Страница 1

Номер по порядку	Товар: наименование, характеристика, сорт, артикул товара	Товар: код	Единица измерения: наименование	Единица измерения: код по ОКЕИ	Вид упаковки	Количество: в одном месте	Количество: мест, штук	Масса брутто	Количество (масса нетто)	Цена, руб. коп.	Сумма без учета НДС, руб. коп.	НДС: ставка, %	НДС: сумма, руб. коп.	Сумма с учетом НДС, руб. коп.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Хлопья кукурузные	00000019030	шт	348	одном месте		131	17,000	596,000	4201,08	8444,52	17	8439,32	6163,41
					Итого				8946,98	X				
					Всего по накладной				8946,98	X				

Товарная накладная имеет приложение на
и содержит один порядковых номеров записей
прописью

Масса груза (нетто)
прописью

Всего мест
прописью

Масса груза (брутто)
прописью

Приложение (паспорта, сертификаты и т.п.) на
прописью
листах

По доверенности № от

Всего отпущено на сумму
три тысячи девять рублей 00 копеек

выданной
кем, кому (организация, должность, фамилия, и. о.)

Отпуск груза разрешил И.о. руководителя / подпись / В. Ч. Симонов
должность / подпись / расшифровка подписи

Главный (старший) бухгалтер / подпись / М. Ч. Гаврилов
подпись / расшифровка подписи

Отпуск груза произвел ведущий специалист / подпись / Т. Б. Исаев
должность / подпись / расшифровка подписи

М.П. " " 20 года

Груз принял / должность / подпись / расшифровка подписи

Груз получил грузополучатель / должность / подпись / расшифровка подписи

М.П. " " 20 года

Рисунок 1 – Пример изображения со структурированными данными
Figure 1 – An example of image with structured data

Следующие модели, которые были использованы в рамках этого исследования, принадлежат третьей группе. В силу условия на возможность обучения модели на локальном ПК (легковесность), были выбраны модели глубокого обучения с архитектурами на базе VGG16, ResNet50 и Xception, собой хорошо известные и проверенные архитектуры глубокого обучения, которые широко используются в задачах компьютерного зрения. Указанные архитектуры показали очень высокие метрики качества классификации на ImageNet при своей легковесности и возможности развертывания на низкомощных серверах или устройствах. VGG16 была выбрана благодаря своей простоте и высокой производительности на задачах классификации изображений, ResNet50 из-за использования остаточных связей, которые позволяют эффективно обучать сети с большим количеством слоев, а Xception благодаря своей способности к разделению глубинных признаков [14, 15]. В работе применяется метод трансферного обучения, суть которого заключается в том, что нейросеть сначала обучается на большом объеме данных, затем – на целевом наборе [18].

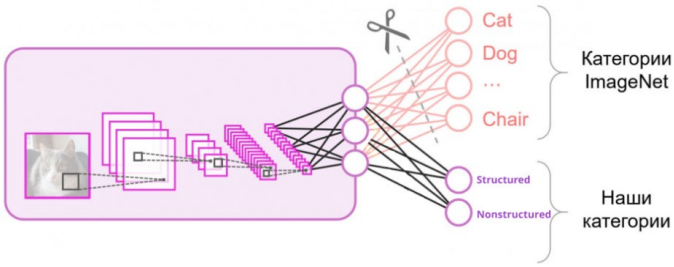


Рисунок 2 – Идея файнтюнинга трансферного обучения
Figure 2 – An idea of transfer learning finetuning

Это позволяет повторно использовать обученную нейросеть для решения нашей задачи классификации или аналогичной проблемы. В распознавании изображений со структурированными и неструктурированными данными, классы отличаются от тех, которые есть в ImageNet, на котором были обучены базовые архитектуры VGG16, ResNet50 и Xception. В связи с этим необходимо отказаться от последнего слоя, заменив его на требуемый, с нужным количеством выходных нейронов, равным двум, как это показано на Рисунке 2.

Результаты

В Таблице 1 представлены метрики качества обученных моделей. Нейронные сети методом трансферного обучения были дообучены на целевом наборе данных со сжатием изображений, соответствующим исходным архитектурам нейросетей. В случае с MobileNetV2, EfficientNetB0, VGG16, ResNet50 – 224×224, а для Xception требуется размер 299×299.

Таблица 1 – Метрики качества обученных моделей
Table 1 – The performance metric of the trained models

Название модели (число параметров)	Площадь под кривой рабочей характеристики приемника (AUC-ROC)	Точность (Precision)	Полнота (Recall)	F1-мера
MobileNetV2 (3,4 млн)	0,91	0,91	0,92	0,91
EfficientNetB0 (5,3 млн)	0,97	1,00	0,95	0,97
ResNet50 VGG16 (138 млн)	0,98	0,98	0,99	0,98

Таблица 1 (продолжение)
Table 1 (continued)

VGG16 (25,6 млн)	0,99	0,99	0,99	0,99
Xception (22,9 млн)	0,99	0,99	1,00	0,99

В Таблице 1 в порядке возрастания метрик качества для обучающего набора данных указаны обученные модели. Всего было обучено 5 моделей классификации. Предобученные на ImageNet модели MobileNetV2 и EfficientNetB0 использовались в качестве базовых моделей, при этом их верхние слои были заменены новыми, специально обученными для решения поставленной задачи. Для нейросети VGG16 на основе предобученной модели с замороженными слоями была создана архитектура для бинарной классификации, включающая глобальный пулинг, полносвязные слои с функцией активации ReLU и случайным удалением нейронов с вероятностью 0,5, скомпилированная с оптимизатором Adam, функцией потерь бинарной кроссэнтропии, метрика качества – точность. Те же гиперпараметры были выбраны для ResNet и Xception. Во всех пяти нейросетях в качестве функции активации на выходе использована сигмоида, все модели обучались на 10 эпохах с размером батча 32.

На Рисунке 3 представлены матрицы ошибок каждой из обученный моделей при прогнозировании на тестовых (отложенных) данных, которые не использовались при обучении нейросетей. MobileNetV2 в классе изображений со структурированными данными ошиблась в 77 случаях, определив их как неструктурированные, и верно предсказала в 700 случаях, а также классифицировала 77 изображений с неструктурированными данными как изображения со структурированными данными. EfficientNetB0 ошиблась в 43 случаях, все ошибки ложноположительные. ResNet50 ошиблась в 25 изображениях, из них – 11 ложноположительные. VGG16 имеет 7 ложноотрицательных и 10 ложноположительных ответа. Нейросеть Xception не ошиблась в классификации ни в одном изображении в «структурированном» классе и имеет 13 ложноположительных ответов, определив изображения с неструктурированными данными как изображения со структурированными данными.

Далее проверялась работа алгоритмов оптического распознавания символов на структурированных документах. Это было проще сделать, поскольку для разметки фактически всегда требовались одни и те же координаты полей или очень близкие. Это помогло использовать мало эпох для обучения (10–15), поскольку далее модели бы переобучались под единственную структуру документов и не могли бы генерализировать данные с табличными вставками. Результаты сравнения по метрике доли верно распознанных символов после обучения моделей YOLOv10–12 размера large приведены в Таблице 2. Также сравниваются разные предобученные OCR модели.

Таблица 2 – Метрики качества OCR моделей
Table 2 – The performance metric of OCR models

Комбинация моделей	**Accuracy**
YOLOv10+Tesseract	0,84
YOLOv10+EasyOCR	0,82
YOLOv11+Tesseract	0,86
YOLOv11+EasyOCR	0,82
YOLOv12+Tesseract	0,88
YOLOv12+EasyOCR	0,84

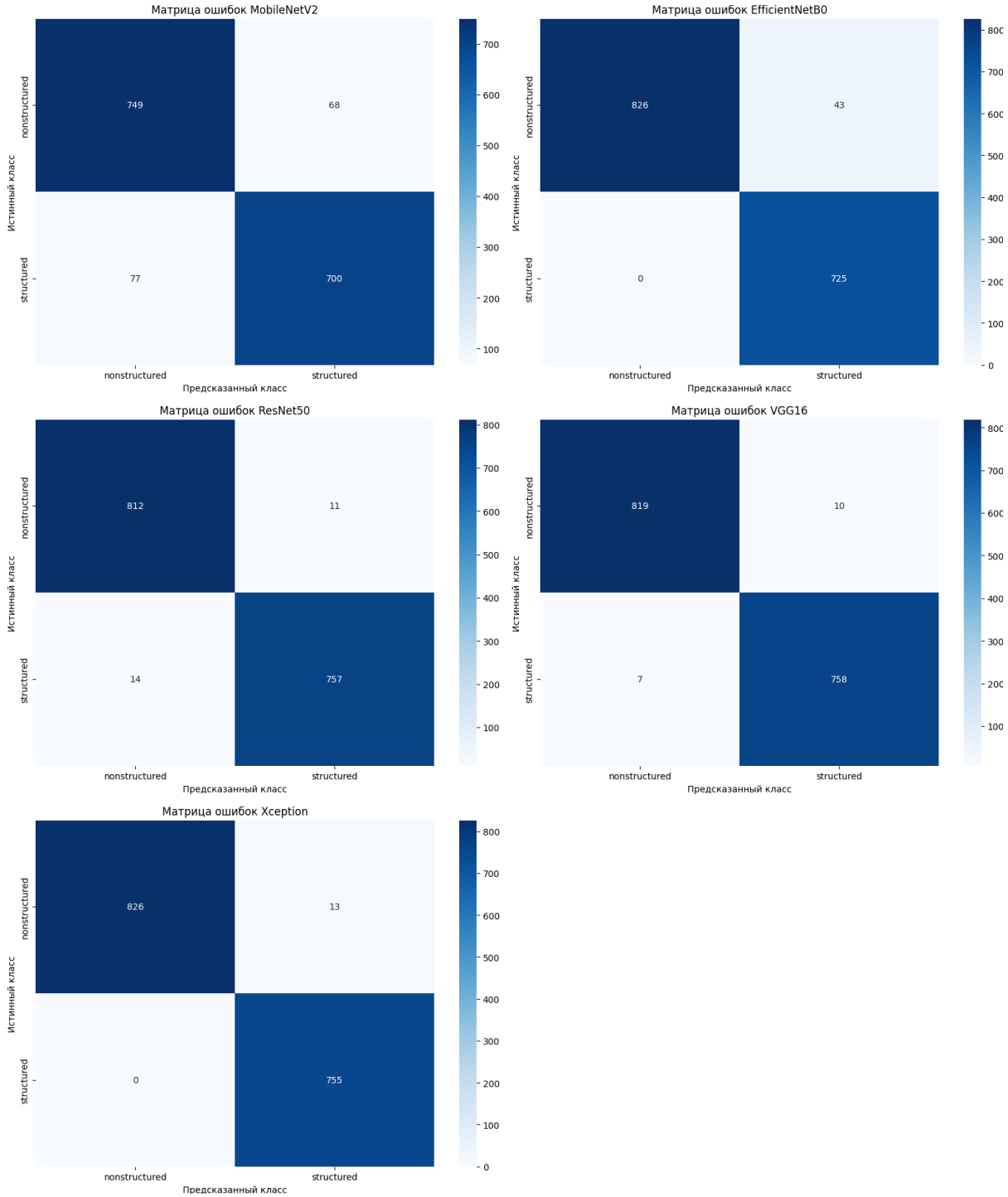


Рисунок 3 – Матрица ошибок каждой из обученных моделей
Figure 3 – Confusion matrix for all trained models

Из Таблицы 2 видно, что лучшее качество обеспечивает комбинация YOLOv12 и EasyOCR. Это связано с тем, что модель YOLOv12 наилучшим образом детектирует области текста на изображении, а версии 10 и 11 иногда могут обрезать часть букв или захватить символы границ ячейки таблицы. Само качество OCR оказалось выше у модели EasyOCR.

Обсуждение

Как видно из Таблицы 1, модель MobileNetV2, выбранная в качестве базовой, показала себя хорошо, значение среднего гармонического точности и полноты составляет 91 %, площади под кривой ROC 91 %, что является очень хорошим результатом для наиболее легковесной модели. Модель EfficientNetB0 показала себя несколько лучше с результатами 97 % по тем же метрикам AUC и F1-мере. Как видно далее, нейронные сети ResNet50, VGG16 дообученные на целевых данных, очень высокий результат со среднегармоническим точности и полноты 98 % и 99 % соответственно. Xception имеет практически эталонный результат со 100 % полнотой и 99 % точностью.

Таким образом, модели MobileNetV2, EfficientNetB0 являются хорошим выбором для задач, где важна простота модели, а также при крайне ограниченных вычислительных ресурсах. VGG16 и ResNet50 показывают практически эталонные результаты, что делает их подходящими для задач, где требуется высокая точность и полнота. Xception является лидером среди всех моделей по всем метрикам качества, демонстрируя самые высокие показатели. Ее применение может быть наиболее обосновано в задачах, когда критически важно минимизировать ошибки классификации. Еще одним преимуществом является то, Xception имеет относительно небольшое количество параметров, что делает ее инференс относительно быстрым.

Как можно видеть на Рисунке 3, MobileNetV2 имеет минимальную точность среди остальных моделей, а ошибки распределены в равных долях между ложноположительными и ложноотрицательными ответами. В среднем MobileNetV2 имеет 9 % ошибок, что может стать ограничением на ее применение. EfficientNetB0 не имеет ложноотрицательных ответов, что значит, что в случае ее применения в предварительной обработке, ни один документ не будет пропущен, но при этом модель имеет 3 % ложноположительных ответов, что несколько потенциально несколько зашумляет отфильтрованный поток документов. Модель ResNet50 имеет очень высокие точность и полноту, почти идеальное предсказание для класса изображений со структурированными данными. У модели ResNet50 на тестовом датасете всего 11 ошибок из 768 ответов для класса изображений со структурированными данными, а в классификации изображений с неструктурированными данными модель сделала 14 ошибок, что демонстрирует очень высокое качество ее инференса и это является третьим результатом из рассматриваемых моделей. Практически идентичный уровень качества результаты демонстрирует модель VGG16. Модель Xception предсказывает верно все 755 изображений со структурированными данными, из 839 изображений с неструктурированными данными, имеет всего 13 ложноположительных ответ. Учитывая, что модель не видела тестировочный набор данных при обучении, можно заключить, что модель является наилучшей для автоматической классификации изображений со структурированными и неструктурированными данными в потоковых источниках. Сравнивая результаты моделей между собой, можно заключить, что EfficientNetB0, MobileNetV2 могут быть применимы при условии низкой цены ошибки и необходимости в самом быстром инференсе при очень ограниченных ресурсах. ResNet50 и VGG16 хорошо подходят для двух классов. Xception является наилучшей нейросетью для поставленной задачи с точки зрения заявленных метрик качества бинарной классификации. Особое преимущество Xception имеет в случае, когда цена ошибки пропуска структурированного документа (ложнонегативного ответа) может быть очень высокой.

Более того, интересные результаты продемонстрировало сравнение использования комбинации детекторов и OCR моделей в задаче обработки

структурированных документов. Было установлено, что комбинация модели YOLOv12 и EasyOCR обеспечивает долю верных распознаваний на 2–5 % выше, чем использование более ранних версий YOLO и, например, распознавателя символов Tesseract. При этом с выделением структурированных данных значительно упрощается процесс разметки и обучения моделей детекции текстовых блоков. А разработанный алгоритм классификации может быть полезен при подготовке датасетов структурированных данных.

Заключение

В работе были рассмотрены современные подходы к оптическому распознаванию символов (OCR) и разработана легковесная модель, способная эффективно классифицировать изображения со структурированными и неструктурированными данными в потоковых источниках для оптимизации оптического распознавания символов. Достигнутые результаты подтвердили высокую эффективность выбранных моделей глубокого обучения, таких как VGG16, ResNet50 и Xception, которые продемонстрировали практически эталонные метрики качества на тестовых наборах данных. В частности, модель Xception показала наилучшие результаты, допустив минимальное количество в классификации изображений среди остальных рассмотренных моделей, что подчеркивает ее надежность и точность в задачах, требующих минимизации ошибок.

Было также показано, что непосредственно распознавание текста в структурированных документах требует малого количества эпох обучения за счет соответствующей структуризации, и модели с точностью, близкой к 90 %, получаются на базе обучения YOLO в течение 15 эпох и стандартных OCR моделей.

Представленная работа имеет большое количество возможных продолжений для изучения. В зависимости от постановки задачи, пропускать изображения с важной и структурированной информацией может быть критично, в то время как рассмотреть вручную лишнее изображение не столь ресурсозатратно, поэтому можно учитывать при бинаризации это условие. По умолчанию в исследовании использовался порог бинаризации равный 0.5, однако его можно варьировать, отказываясь от ложноотрицательных ошибок в пользу ложноположительных. Также дальнейшим направлением исследования может являться оптимизация гиперпараметров для специфических данных, в случае если потоковые данные поменяют свою структуру. Другим направлением исследования может быть исследование дисбаланса классов. Например, когда неструктурированной информации поступает все больше. Это может быть актуально для таких потоков источников как электронная почта или сайт. Отдельно можно исследовать методы интерпретации нейросетей (например, LIME или SHAP), чтобы лучше понимать инференс и анализировать стабильность во времени.

Наиболее перспективным дальнейшим развитием исследования является интеграция легковесной модели в большую систему, которая может решать сложные задачи оптического распознавания символов и оркестрировать большое количество ML-моделей и технологий разработки и инженерии данных. Выбор модели зависит от задачи и доступных ресурсов, но результаты показывают, что современные архитектуры способны справляться с задачами классификации изображений с высокой эффективностью. В большой системе данные с изображений со структурированной формой (счета-фактуры, накладные) могут направляться в визуально-языковые модели (VLM) для извлечения табличных данных и последующего экспорта в формат Excel (XLSX) или JSON. Благодаря этому данные из стандартных шаблонов могут быть автоматически оцифрованы и включены в бухгалтерский учет или ERP-систему. При

этом неструктурированные не обрабатываются как таблицы, а могут быть направлены на ручную или отдельную интеллектуальную обработку. После фильтрации документы обрабатываются непосредственно в VLM и на выходе получаются данные в нужном формате. Такая система позволит оптимизировать процесс обработки входящего потока и повысить его эффективность, а точность будет превышать точность систем OCR EasyOCR, Pytesseract на 20–30 %.

Таким образом, результаты работы подчеркивают важность и перспективность дальнейших исследований в области оптического распознавания символов, поскольку это открывает новые горизонты для применения технологий компьютерного зрения в различных сферах, включая как бизнес, так и цифровую трансформацию в целом, а разработка эффективных легковесных моделей может значительно упростить и ускорить последующие процессы обработки данных с применением технологий OCR и IDP.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Shi B., Bai X., Yao C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2017;39(11):2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371
2. Inoue K. Context-Independent OCR with Multimodal LLMs: Effects of Image Resolution and Visual Complexity. arXiv. URL: https://arxiv.org/abs/2503.23667 [Accessed 12th March 2025].
3. Fujitake M. DTrOCR: Decoder-Only Transformer for Optical Character Recognition. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 03–08 January 2024, Waikoloa, HI, USA.* IEEE; 2024. P. 8010–8020. https://doi.org/10.1109/WACV57701.2024.00784
4. Tian Yu., Ye Q., Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv. URL: https://arxiv.org/abs/2502.12524 [Accessed 12th March 2025].
5. Alif M.A.R., Hussain M. YOLOv12: A Breakdown of the Key Architectural Features. arXiv. URL: https://arxiv.org/abs/2502.14740 [Accessed 12th March 2025].
6. Wang X., Li Ye., Liu J., et al. Intelligent Micron Optical Character Recognition of DFB Chip Using Deep Convolutional Neural Network. *IEEE Transactions on Instrumentation and Measurement.* 2022;71. https://doi.org/10.1109/TIM.2022.3154831
7. Li M., Lv T., Chen J., et al. TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. In: *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, 07–14 February 2023, Washington, DC, USA.* AAAI Press; 2023. P. 13094–13102.
8. Dosovitskiy A., Beyer L., Kolesnikov A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *9th International Conference on Learning Representations, ICLR 2021, 03–07 May 2021, Virtual Event, Austria.* 2021. https://doi.org/10.48550/arXiv.2010.11929
9. Xiong S., Chen X., Zhang H. Deep Learning-Based Multifunctional End-to-End Model for Optical Character Classification and Denoising. *Journal of Computational Methods in Engineering Applications.* 2023;3(1):1–13. https://doi.org/10.62836/jcmea.v3i1.030103
10. Kasem M.S.E., Mahmoud M., Kang H.-S. Advancements and Challenges in Arabic Optical Character Recognition: A Comprehensive Survey. [Preprint]. arXiv. URL: https://arxiv.org/abs/2312.11812 [Accessed 14th March 2025].

11. Baek Yo., Lee B., Han D., Yun S., Lee H. Character Region Awareness for Text Detection. arXiv. URL: https://doi.org/10.48550/arXiv.1904.01941 [Accessed 14th March 2025].
12. Zhang Ya., Ye Yu-L., Guo D.-J., Huang T. PCA-VGG16 Model for Classification of Rock Types. *Earth Science Informatics.* 2024;17(2):1553–1567. https://doi.org/10.1007/s12145-023-01217-y
13. Sarwinda D., Paradisa R.H., Bustamam A., Anggia P. Deep Learning in Image Classification Using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. *Procedia Computer Science.* 2021;179:423–431. https://doi.org/10.1016/j.procs.2021.01.025
14. Morani K., Ayana E.K., Kollias D., Unay D. COVID-19 Detection from Computed Tomography Images Using Slice Processing Techniques and a Modified Xception Classifier. *International Journal of Biomedical Imaging.* 2024;2024. https://doi.org/10.1155/2024/9962839
15. Andriyanov N., Andriyanov D. Pattern Recognition on Radar Images Using Augmentation. In: *2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), 14–15 May 2020, Yekaterinburg, Russia.* IEEE; 2020. P. 0289–0291. https://doi.org/10.1109/USBEREIT48449.2020.9117669
16. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-Ch. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18–23 June 2018, Salt Lake City, UT, USA.* IEEE; 2018. P. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
17. Tan M., Le Q.V. EfficientNetV2: Smaller Models and Faster Training. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event.* 2021. P. 10096–10106.
18. Ахмад А., Андриянов Н.А., Соловьев В.И., Соломатин Д.А. Применение глубокого обучения для аугментации и генерации подводного набора данных с промышленными объектами. *Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника.* 2023;23(2):5–16. https://doi.org/10.14529/ctcr230201
Ahmad A., Andriyanov N.A., Soloviev V.I., Solomatin D.A. Application of Deep Learning for Augmentation and Generation of an Underwater Data Set with Industrial Facilities. *Bulletin of the South Ural State University. Series: Computer Technologies, Automatic Control, Radioelectronics.* 2023;23(2):5–16. (In Russ.). https://doi.org/10.14529/ctcr230201

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Гаврилов Вадим Сергеевич, аспирант кафедры искусственного интеллекта, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация.
e-mail: vad093@mail.ru
ORCID: 0009-0006-0849-4561

Vadim S. Gavrilov, Postgraduate at the Department of Artificial Intelligence, Financial University under the Government of the Russian Federation, Moscow, the Russian Federation.

Корчагин Сергей Алексеевич, кандидат физико-математических наук, доцент, доцент кафедры искусственного интеллекта, Финансовый университет при Правительстве

Sergei A. Korchagin, Candidate of Physical and Mathematical Sciences, Docent, Associate Professor at the Department of Artificial Intelligence, Financial University under the

Российской Федерации, Москва, Российская Федерация.
e-mail: sakorchagin@fa.ru
ORCID: 0000-0001-8042-4089

Government of the Russian Federation, Moscow, the Russian Federation.

Долгов Виталий Игоревич, кандидат физико-математических наук, доцент, доцент кафедры информационных технологий, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация.
e-mail: vidolgov@fa.ru
ORCID: 0000-0003-2413-7880

Vitaly I. Dolgov, Candidate of Physical and Mathematical Sciences, Docent, Associate Professor at the Department of Information Technology, Financial University under the Government of the Russian Federation, Moscow, the Russian Federation.

Андриянов Никита Андреевич, кандидат технических наук, доцент, доцент кафедры искусственного интеллекта, Финансовый университет при Правительстве Российской Федерации, Москва, Российская Федерация.
e-mail: naandriyanov@fa.ru
ORCID: 0000-0003-0735-7697

Nikita A. Andriyanov, Candidate of Engineering Sciences, Docent, Associate Professor at the Department of Artificial Intelligence, Financial University under the Government of the Russian Federation, Moscow, the Russian Federation.

Статья поступила в редакцию 26.04.2025; одобрена после рецензирования 15.06.2025; принята к публикации 24.06.2025.

The article was submitted 26.04.2025; approved after reviewing 15.06.2025; accepted for publication 24.06.2025.