Министерство науки и образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПОСТРОЕНИЕ МОДЕЛИ НЕЙРОННОЙ СЕТИ ДЛЯ ЗАДАЧИ РАСПОЗНАВАНИЯ РУКОПИСНОГО ТЕКСТА НА ДОКУМЕНТАХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента (ки) 2 курса 248 группы

направления 09.04.03 – Прикладная информатика
код и наименование направления

механико-математического факультета
наименование факультета

Маловой Ольги Дмитриевны
фамилия, имя, отчество

Научный руководитель:

Доцент, к. ф.-м. н.		О.А. Мыльцина
должность, уч. степень, уч. звание	дата, подпись	инициалы, фамилия

Заведующий кафедрой:

Зав. кафедрой, д. ф.-м. н., доцент		С.П. Сидоров
должность, уч. степень, уч. звание	дата, подпись	инициалы, фамилия

Саратов 2025 год

ВВЕДЕНИЕ

Актуальность темы. В настоящее время с развитием информационных технологий растет уровень автоматизации и роботизации процессов, как в науке и промышленности, так и в повседневности. В следствие этого возникает необходимость эффективных решений задачи обработки информации различных видов. Многие коммерческие компании нуждаются в оптимизации процессов документооборота, учитывая наличие большого объема рукописных данных, включая подписи, исправления и ручное заполнение реквизитов. Подобные документы остаются важными в логистике, бухгалтерском учете и юридической практике, что требует создания специализированных решений для их обработки. Для решения таких задач используются технологии компьютерного зрения.

Технологии компьютерного зрения применяются в различных прикладных областях: от изучения состояния деревьев до управления космическими спутниками. Существует множество работ, посвященных обработке изображений, в том числе с распознаванием текста, однако на данный момент не создано универсального решения данной задачи. Каждая прикладная область задает свои особенные условия на входные данные и требуемые результаты. Поэтому для построения эффективного алгоритма необходимо не только глубокое знание методов компьютерного зрения, но и детальное изучение предметной области.

Целью магистерской работы является подготовка данных и реализация архитектуры нейронной сети для решения задачи распознавания рукописного русскоязычного текста на распространенных сопроводительных документах.

Объект исследования - рынок коммерческих банков России.

Предмет исследования - набор данных, содержащий скан-образы документов форм «ТОРГ-12», «Товарная накладная» и «Универсальный передаточный документ».

Для достижения поставленной цели в работе необходимо выполнить следующие задачи:

— провести обзор существующих методов оптического распознавания сим-

волов;

— изучить этапы алгоритмов оптического распознавания символов;
— рассмотреть примеры существующих инструментов для решения задач компьютерного зрения;
— реализовать разметку документов в наборе данных;
— провести обучение нейронной сети детектированию рукописного текста на документах;
— применить модель CRNN для распознавания;
— построить модель DenseNet-121 и улучшить ее с помощью CRNN;
— реализовать функцию распознавания для наглядного отображения результатов.

Практическая значимость и новизна работы состоят в том, что в настоящее время на рынке отсутствует универсальное решение задачи распознавания рукописного русскоязычного текста, которое возможно интегрировать в существующие системы документооборота и объединить с распознаванием печатного текста.

Структура и содержание магистерской работы. Работа состоит из введения, трех разделов, заключения, списка использованных источников, содержащего 22 наименования, на которые в тексте работы приведены ссылки и приложений, в которых приведены фрагменты кода программы для реализации задачи.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение посвящено обоснованию актуальности выбранной темы работы и набора данных, формулируется цель работы, приводится постановка необходимых задач, отмечается практическая значимость полученных результатов.

В **первом разделе** приводится описание алгоритма оптического распознавания символов (OCR), определения, связанные с ним, а также этапы и различные варианты их выполнения.

Системы OCR состоят из нескольких шагов алгоритма, которые выполняются последовательно, результат каждого шага подается на вход следующего.

1. На этапе предобработки изображение подготавливается к распознаванию, для чего его необходимо обработать и выделить нужную информацию. К изображению применяют операции очистки от шумов, фильтрация, сглаживание, увеличение контрастности. Для точного выделения текста и удаления фона используется бинаризация.

2. Вторым этапом осуществляется сегментация, то есть выделение полезной информации из изображения с ее последующей обработкой. В области распознавания текста сегментация включает этапы сегментации строк, слов и символов.

3. Классификация позволяет распознать символ из изображения и перевести его в машиночитаемый формат. Существуют разные виды алгоритмов распознавания, самыми популярными являются: шаблонные алгоритмы, признаковые алгоритмы и нейросетевые алгоритмы.

4. Во многих случаях результат работы системы OCR после классификации не считается конечным. Для обнаружения и исправления ошибок необходимо использовать контекстную информацию. Существуют такие методы постобработки, как глобальные и локальные позиционные диаграммы, триграммы, n-граммы, словари, а также различные сочетания всех этих методов

Второй раздел содержит подробное описание используемых в работе технологий: инструменты для детектирования объектов на изображениях и разметки, а также проанализированы принцип работы и структура моделей нейронных сетей для распознавания текста.

Label Studio была использована для решения задачи разметки. Данный инструмент широко применяется в задачах машинного обучения, включая распознавание рукописного текста. Программа позволяет эффективно составлять аннотации к областям изображений, текстов и других типов данных, что крайне важно для построения качественных моделей машинного обучения.

При использовании Label Studio для задач распознавания текста процесс включает следующие этапы:

1. Подготовка данных - документы загружаются в систему в виде изображений. При необходимости пользователь имеет возможность предварител-

ной обработки изображений (например, нормализация яркости, устранение шума).

2. Разметка данных - выделение области с текстом и вводят их транскрипцию. В некоторых случаях может быть полезно размечать дополнительные характеристики, такие как язык, стиль письма или качество изображения.

3. Обучение и улучшение модели или экспорт данных для использования в других программах.

В данной работе Label Studio использовалась только для составления разметки изображений, затем данные экспортировались в формате YOLO для дальнейшего применения в нейронной сети.

YOLO v5 ultralytics. Нейронная сеть YOLO v5 (You Only Look Once, версия 5) представляет собой одно из современных решений для детектирования объектов. Архитектура этой сети подходит для разбиения сложных изображений на локализованные текстовые области, которые затем могут быть обработаны различными методами. Сеть YOLO v5 имеет следующую структуру:

1. *Backbone* — это базовая сеть, которая отвечает за извлечение признаков из изображения. Она представляет собой сверточную нейронную сеть. Включает такие элементы, как:

— *Focus*: первичный слой, который разделяет входное изображение на участки и применяет свёртки;

— *CSP Bottleneck*: модуль для улучшения обучения глубоких сетей и сокращения избыточных вычислений;

— *SPPF* (Spatial Pyramid Pooling - Fast): модуль, который увеличивает рецептивное поле сети, позволяя учитывать глобальный контекст изображения.

2. *Neck* — это промежуточный слой, который связывает Backbone и Head. Его задача — объединение признаков с различных уровней глубины сети, что необходимо, так как объекты на изображениях могут иметь разный размер.

3. *Head* — выходной слой сети, который отвечает за предсказание координат ограничивающих рамок (bounding boxes), вероятностей классов и

степени уверенности в детекции. В YOLO v5 используется:

— *Anchor-based детекция*: модель использует предварительно определённые якорные рамки (anchors) для прогнозирования позиций объектов;

— *Objectness Score*: метрика, которая определяет вероятность того, что в рамке находится объект нужного класса;

— *Сигмоидальная активация*: используется для нормализации предсказаний и интерпретации вероятностей.

Модель CRNN. В ней сверточная нейронная сеть используется для извлечения локальных признаков из изображения, а рекуррентная сеть, представленная двумя слоями двунаправленных LSTM, занимается обработкой последовательности извлеченных символов.

Сверточный слой сети устроен таким образом, что исходное изображение разбивается на k (гиперпараметр) вертикальных сегментов вдоль горизонтальной оси, каждому из которых на выходе из свертки соответствует вектор признаков размерности (1, n). За счет уменьшения одного из измерений до единицы можно без потери информации понизить размерность карты признаков ((k, 1, n) $\rightarrow$ (k, n)).

Субдискретизационные слои располагаются между сверточными слоями для уменьшения объема информации, получаемой в результате применения фильтров к изображению. Метод Max Pooling выбирает максимальное значение из всего массива входных данных и суммирует эти значения в карту признаков. Этот метод сохраняет наиболее значимые характеристики входных данных, уменьшая их размеры. Его математическая формула имеет вид:

$$\text{MaxPooling}(X)_{i,j,k} = \max_{m,n} X_{i \cdot s_x + m, j \cdot s_y + n, k},$$

где X - вход, (i,j) - индексы выходов, k - индекс канала, s_x и s_y - значения страйдов в горизонтальном и вертикальном направлениях соответственно, а окно объединения определяется размером фильтра f_x и f_y с центром в выходном индексе (i,j).

Слой *Batch Normalization* используется в качестве вспомогательного параметра для оптимизации, позволяет уменьшить число эпох обучения. Включает этапы:

1. Вычисление среднего значения и стандартного отклонения для каждого признака:

$$\mu_d = \frac{1}{N}\sum_{i=1}^{N} x_{i,d},$$

$$\sigma_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{i,d} - \mu_d)^2.$$

2. Нормализация входных данных:

$$z_{i,d} = \frac{(x_{i,d} - \mu_d)}{\sqrt{\sigma_d^2 + \epsilon}},$$

где ϵ — небольшое число, которое добавляется для предотвращения деления на ноль.

3. Масштабирование и сдвиг нормализованных данных: $y_{i,d} = \gamma_d z_{i,d} + \beta_d$, где γ_d и β_d — параметры масштаба и сдвига для каждого признака.

4. Применение функции активации к нормализованным данным: $h(x) = f(y)$, где f — функция активации, например, ReLU или сигмоидная

Рекурентный слой. После сверточных слоев и слоев оптимизации полученные данные преобразуются в последовательность векторов признаков и передаются в двунаправленную рекуррентную сеть LSTM. Входной узел сети обрабатывает входные данные для текущего временного шага. Далее вычисляется забывающий узел, после чего определяются значение-кандидат на изменение ячейки состояния и состояние на данный момент. В выходном узле вычисляется скрытое состояние в текущий момент времени и значение выходного узла.

Функция потерь CTC. Основная идея метода CTC заключается в том, чтобы преобразовать последовательность наблюдений в последовательность меток, а затем использовать эту последовательность для обучения модели. Для этого используется специальный вид функции потерь, называемый CTC loss. Для вычисления вероятности целевой последовательности y вводится множество всех возможных меток $L = \{1, 2, \ldots, L\}$, включая символ $\varnothing$. Обозначим через $p(l_t|x_t)$ вероятность метки $l_t \in \mathcal{L}$ в момент времени t. Тогда вероятность полного выравнивания $\pi = (\pi_1, \pi_2, \ldots, \pi_T)$, где $\pi_t \in L$, задается

как:

$$p(\pi|x) = \prod_{t=1}^{T} p(\pi_t|x_t).$$

Модель DenseNet. Данная архитектурная модель DenseNet вводит концепцию плотных блоков, соединяя каждый слой нейронной сети непосредственно друг с другом. Чтобы обеспечить максимальный поток информации между слоями сети, все слои с соответствующими размерами карты признаков соединяются напрямую друг с другом. Сеть содержит следующие группы слоев:

1. Входной слой: стандартный сверточный слой с фильтрами размера 7×7, за которым следует слой максимального объединения (Max Pooling).

2. Плотные блоки: каждый плотный блок включает множество сверточных слоев, каждый из которых получает входные данные от всех предыдущих слоев блока. Такая структура позволяет минимизировать переобучение и улучшить обобщающую способность модели.

3. Транзитные слои: эти слои, состоящие из сверточных слоев 1×1 и объединяющих слоев 2×2, уменьшают размерность данных и число каналов, что способствует уменьшению вычислительной нагрузки.

4. Выходные слои: глобальный средний пуллинг и полностью связанный слой с функцией активации softmax для классификации.

Третий раздел включает в себя описание наборов данных для обучения и реализации модели распознавания, реализацию детектирования и извлечения рукописного текста на документе, а также построение модели распознавания этого текста.

Описание наборов данных.Тестовым набором данных и объектом исследования является архив отсканированных документов таких форм, как «ТОРГ-12», «УПД» (Универсальный передаточный документ), «Товарная накладная». В качестве тренировочного набора данных использовался Cyrillic Handwriting Dataset, размещенный на платформе Kaggle. Данный набор включает около 70 тысяч изображений рукописных слов и фраз на русском языке, сгенерированных с участием различных авторов, что обеспечивает разнообразие почерков и условий написания. Изображения представлены в градаци-

ях серого с разрешением 256×64 пикселей и содержат аннотации в формате текстовых меток. Кроме того, он применялся для дообучения сетей CRNN и DenseNet.

Настройка детектирования рукописного текста. Первым шагов в достижении поставленной цели работы было составление разметки. Был сформирован тренировочный набор данных, состоящий из 300 изображений для начального этапа. Было решено выделять три класса рукописных пометок на документах: handwriting (текст), date (даты) и signature (подписи).

В ходе выполнения детектирования модулем *predict* валидационных изображений были получены результаты, частично представленные на рисунке. Можно отметить, что на изображениях были обнаружены все области рукописных фигур. С высокой уверенностью детектируется и классифицируется рукописный текст, который является основным объектом работы.

Результаты работы модели детекции представляют собой исходные изображения скан-образов документов и текстовые файлы с координатами обнаруженных классов на каждом изображении. Эти данные являются входными для последующей программы для автоматизированного извлечения фрагментов изображений, соответствующих классу Handwriting, на основе предсказанных координат.

Метрики оценки качества модели распознавания. Метрика CER (Character Error Rate) является одной из основных метрик для оценки качества систем распознавания текста, таких как OCR (Optical Character Recognition) и системы распознавания речи. CER измеряет количество ошибок при распознавании символов в строке по сравнению с эталонным текстом.

Метрика WER (Word Error Rate), или частота ошибок в словах, также используется в качестве ключевой для оценки производительности моделей распознавания текста. Эта метрика аналогична CER и совместно с ней используется в задачах оптического распознавания символов (OCR) и автоматического распознавания речи (ASR), где важно измерять, насколько точно модель может преобразовывать входные данные в текстовую форму.

Распознавание рукописного текста. **Предобработка изображений** является важным этапом перед применением к ним моделей распознавания текста. Алгоритм предобработки включает последовательность опе-

раций, разработанных с учетом особенностей рукописных документов: вариативности освещения, геометрических искажений и неоднородности фона. Этапы, входящие в функцию предобработки:

1. Загрузка изображения в градациях серого с использованием метода cv2.IMREAD_GRAYSCALE.

2. Этап шумоподавления включает в себя применение медианного фильтра с ядром 1×1 (cv2.medianBlur).

3. Бинаризация методом фиксированного порога cv2.THRESH_BINARY.

4. Морфологическое открытие (cv2.MORPH_OPEN) с ядром $2times2$.

5. Дилатация (расширение) реализуется после операции морфологического открытия методом cv2.dilate.

6. Эрозия (сужение), выполняемая функцией cv2.erode, применяется после дилатации для восстановления исходного масштаба символов.

В результате работы данной функции получены изображения, которые представляют собой бинарную маску, где текстовые элементы класса детекции Handwriting выделены с увеличенной четкостью и сниженным уровнем шума.

Модель CRNN включает три части: сверточную сеть CNN, рекуррентную сеть RNN и блок преобразования слоев между ними. В результате отработки модели был получен график потерь. Наблюдаются быстрое обучение на тренировочном наборе данных и большое снижение потерь. Поведение функции потерь на тестовой выборке существенно отличается. После спада потерь на ранних эпохах кривая становится нестабильной: наблюдаются резкие колебания, значения остаются на достаточно высоком уровне, и отсутствует явная тенденция к снижению. Такая динамика может быть признаком недостаточной способности модели к обобщению, что усиливается особенностью тестового набора. Также о необходимости улучшения модели говорят графики CER и WER, соответственно. Метрика CER демонстрирует большое количество ошибок, сильно колеблясь около уровня 1. Ее значениям соответвует и WER, которая на всем протяжении обучения равна 1 или крайне близка к ней.

Данная модель нуждается в изменении структуры и усложнении, поэтому ее составные слои будут включены в итоговую архитектуру нейронной

сети для распознавания, так как выполняют важные преобразования для распознавания последовательностей символов.

Модель DenseNet-121 Для данной работы в архитектуре нейронной сети DenseNet-121 используется как базовая модель, к которой добавляются и настраиваются необходимые слои обработки и трансформации данных. В структуру сети также была включена модель CRNN с оптимальными параметрами и необходимыми слоями, которые были определены в предыдущем подразделе. Особенностью DenseNet-121 является плотное соединение слоев, при котором каждый слой получает на вход не только выход предыдущего, но и все выходы от более ранних слоев.

В ходе работы с данной моделью было осуществлено несколько этапов обучения и доработки архитектуры модели и функции обучения. Основные дополнения к исходной архитектуре модели за три этапа улучшений:

— добавлены слои нормализации *Batch Normalization* после сверточных, плотных и рекуррентных;
— внедрены Dropout-слои с коэффициентом 0.3;
— реализована стратегия экспоненциального затухания *learning rate*;
— добавлен *градиентный клиппинг*, ограничение значений градиентов на этапе обратного распространения;
— внедрена регуляризация L2, которая добавляет к функции потерь штраф за слишком большие значения весов модели;
— включение в регулирование обучения *Learning Rate Scheduler*, для большей восприимчивости показателей по эпохам

В результате внесенных изменений были улучшены все показатели качества модели. Потери на тестовом наборе упали до 9%, что является хорошим результатом для рассматриваемых данных в силу особенностей изображений и русскоязычного рукописного текста. Метрики CER и WER также показали хорошую обучаемость модели и сниженное количество ошибок до 0.25 и 0.4, соответственно.

Примеры распознавания представлены на рисунке 1. По этим примерам можно сделать вывод о хорошей работе модели. Несмотря на все еще присутствующие ошибки в словах, они являются незначительными и гораздо более редко встречаемыми, а смысловое содержание сохраняется лучше.

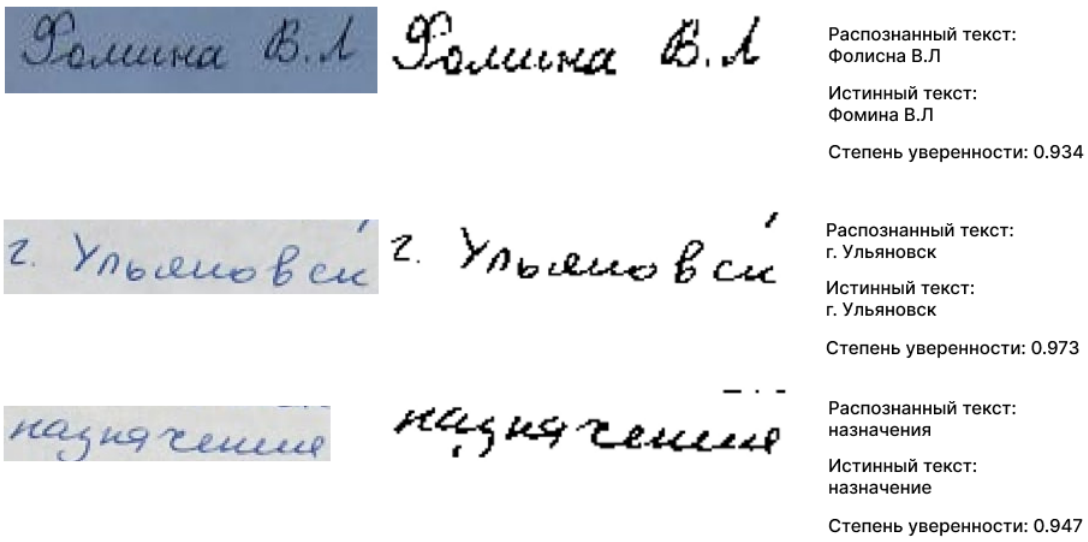

Рисунок 1 – Примеры распознавания текста на тестовом наборе

Например, ошибка в распознавании слова «назначения» как «назначение» не несет больших потерь информации.

Таким образом, на основе полученных показателей качества модели можно сделать вывод о том, что она была достаточно оптимизирована. Внесенные изменения показали свою эффективность в улучшении работы сети.

В **заключении** приведены результаты магистерской работы.

Основные результаты

1. Определены основные показатели финансово-хозяйственной деятельности банков. Рассмотрены основные современные тенденции их изменения.

1. Изучены этапы алгоритмов оптического распознавания символов, а также проведен обзор существующих методов.

2. Рассмотрены основные понятия и характеристики различных видов нейронных сетей и их реализаций для рассматриваемой задачи.

3. Проведен анализ архива документов для распознавания, выявлены ключевые особенности изображений.

4. Реализована разметка, детектирование и извлечение рукописного текста на документе с помошью модели нейронной сети YOLO v5.

5. Построена модель нейронной сети для распознавания рукописного с помощью DenseNet-121 в качестве основы, CRNN как улучшения и внесенных вручную изменений архитектуры для наибольшего соответствия данным.

6. Осуществлена интеграция между модулем детекции рукописного текста и модулем его распознавания.

7. Проверено качество выполнения модели с помощью сетрик CER и WER, а также с помощью анализа результатов работы функции предсказания.