МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РЕШЕНИЕ ЗАДАЧИ РАСПОЗНАВАНИЯ РУКОПИСНОГО ТЕКСТА С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование
информационных систем
факультета компьютерных наук и информационных технологий
Анферовой Анастасии Викторовны

Научный руководитель:
зав. кафедрой, к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Несмотря на то, что печатный текст, находясь в постоянной конкуренции с рукописным, пытается вытеснить последний – такое все же вряд ли произойдет полностью.

Таким образом, распознавание рукописного текста – важная прикладная задача, которая, в первую очередь, предоставляет удобство пользователю при переносе информации из рукописного вида в электронный, например, при выделении текста из рукописных конспектов, писем, при автоматической обработке заполненных форм. Информацию в электронном виде не составляет труда обрабатывать программным образом, так данная программа может использоваться для проверки заданий в школе, университетах, что значительно упростит работу сотрудникам образовательных учреждений.

Под оффлайн-распознаванием понимается выделение текста с отсканированного или сфотографированного документа, в то время как онлайн-распознавание – это выделения текста непосредственно во время его написания.

Распознавание рукописного текста берет свое начало от OCR-систем (англ. optical character recognition – оптическое распознавание символов), первая из которых была запатентована Густавом Таушеком еще в 1929 году. Но системы OCR ориентированы, в первую очередь, на печатный текст.

Способность качественного распознавания текста на изображениях во многом зависит от типа текста. Традиционно выделяют распознавание рукописного печатного и рукописного курсивного текста: отличие рукописного печатного от рукописного курсивного текста заключается в том, что в первом случае подразумеваются печатные буквы, написанные от руки раздельно друг от друга, а во втором – написанные от руки буквы с соединительными линиями. И OCR-системы, как было отмечено ранее, главным образом направлены на распознавание печатного текста. Так,

точность распознавания латинских печатных символов при условии сравнительно высокого качества изображений может достигать 99 %.

Актуальность данной работы подтверждается тем, что распознавание рукописного курсивного текста не только является важной прикладной задачей, как было отмечено ранее, но и все еще остается задачей, требующей детального исследования. Системы, которые хорошо справляются с этой задачей, существуют (например, Abbyy FineReader), однако их нет в свободном доступе.

Цель бакалаврской работы – разработка системы автоматизированного распознавания русского рукописного текста в режиме оффлайн.

Поставленная цель определила **следующие задачи** работы:

1. Разобрать основные понятия и определения, связанные с проблемой распознавания текстов.
2. Изучить методы предварительной обработки данных.
3. Изучить и подобрать методы распознавания рукописного текста.
4. Собрать, разметить и разместить набор данных русского рукописного текста на открытой платформе.
5. Выполнить реализацию алгоритмов для решения задачи распознавания рукописного текста.

Для решения задачи распознавания рукописного текста в режиме оффлайн была выбрала архитектура нейронной сети CRNN (англ. convolutional recurrent neural network – сверточная рекуррентная нейронная сеть) с применением CTC (англ. connectionist temporal classification – коннекционистская временная классификация).

Методологические основы распознавания рукописного текста и применения методов машинного обучения для распознавания последовательностей представлены в работах Кеннет М. Сэйр, Дэна Киресана, Кучуганова А. В и Лапинской Г. В, Рахуль Калы, Кунхонг Йу, Гафарова Ф.М, М. Потанина.

Структура и объем работы. Бакалаврская работа состоит из введения, пяти разделов, заключения, списка использованных источников и шести приложений. Общий объем работы – 106 страниц, из них 76 страницы – основное содержание, включая 55 рисунков и список использованных источников информации (27 наименования).

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Базы данных рукописного текста и общие подходы для его распознавания» посвящен предшествующим работам по сбору баз данных рукописного текста и его распознаванию.

В **подразделе 1.1 «Базы данных рукописного текста»** говорится об уже имеющихся базах данных рукописного текста для разных языков мира, в какой коллекции они представлены. И также отмечается, что базы данных русского рукописного текста существуют, но в большинстве случаев они являются закрытыми.

В **подразделе 1.2 «Общие подходы к распознаванию рукописного текста»** были рассмотрены различные подходы к распознаванию рукописного текста и рассмотрены работы исследователей по распознаванию рукописного текста и последовательностей в режиме оффлайн с помощью нейронных сетей.

Итоги. В первом разделе были рассмотрены существующие базы данных рукописного текста: так как существует проблема нехватки баз данных русского рукописного текста было принято решение собрать собственную базу данных. Также были рассмотрены общие подходы к распознаванию рукописного текста.

Второй раздел «Сбор и подготовка графических изображений» ознакамливает со способами сбора и подготовки данных для составления графических баз данных.

В **подразделе 2.1 «Сбор датасета»** приведены основные понятия, связанные со сбором данных и рассматриваются основные этапы для сбора

датасета.

Датасет – это набор данных. Наличие полностью размеченного датасета означает, что каждому примеру в обучающем наборе соответствует ответ, который алгоритм и должен получить.

В **подразделе 2.2 «Предварительная обработка изображений»** рассмотрены основные этапы для предварительной обработки изображений, что является важным этапом при составлении датасета. Основная цель предварительной обработки изображения состоит в том, чтобы удалить из изображения нерелевантную информацию, восстановить полезную реальную информацию, улучшить обнаруживаемость релевантной информации и максимально упростить данные, тем самым повышая надежность извлечения признаков, сегментации изображения, сопоставления и распознавания.

Итоги. Во втором разделе были рассмотрены основные этапы для сбора датасета и предварительной обработки изображений, что необходимо для составления качественного датасета.

Третий раздел «Теория нейронных сетей» посвящен определению понятий нейронных сетей, описанию алгоритмов работы таких архитектур нейронной сети, как сверточная нейронная сеть (CNN), рекуррентная нейронная сеть (RNN). Описание работы коннекционистская временной классификации (CTC). Также представлено описание работы архитектуры сверточно-рекуррентной нейронной сети с применением коннецкционистской временной классификации (CRNN + CTC).

В **подразделе 3.1 «Основные понятия теории нейронных сетей»** приведены основные понятия теории нейронных сетей.

Нейронная сеть – это математическая модель, которая является системой связанных между собой нейронов, каждый из которых обменивается сигналами с нейронами соседних слоев.

Слой – модуль обработки данных, который фильтрует входную информацию и выдает на выходе сведения, более приближенные к желаемому результату.

Полносвязная нейронная сеть – это сеть, в которой каждый нейрон связан со всеми остальными нейронами, находящимися в соседних слоях.

Батч – часть данных датасета, проходящих через нейронную сеть за один раз. Размер батча – число данных, входящих в состав одного батча. Число батчей называют итерациями. Обработка всех батчей означает завершение одной эпохи.

В **подразделе 3.2 «Сверточная нейронная сеть (CNN)»** описывается алгоритм работы сверточной нейронной сети.

Сверточная нейронная сеть (англ. convolutional neural network, CNN) – архитектура нейронных сетей, нацеленная на эффективное распознавание образов.

В **подразделе 3.3 «Рекуррентная нейронная сеть (RNN)»** рассматривается алгоритм работы рекуррентной нейронной сети.

Рекуррентные нейронные сети (англ. recurrent neural network, RNN) – архитектура нейронных сетей, где связи между элементами образуют направленную последовательность.

В **подразделе 3.4 «Коннекционистская временная классификация (CTC)»** дается объяснение алгоритму работы коннекционистской временной классификации.

Коннекционистская временная классификация (CTC) – это выходной слой нейронной сети, используемый для решения таких задач, которые представлены в виде последовательности.

В **подразделе 3.5 «CRNN + CTC Loss для задач распознавания текста»** представлен подробной разбор алгоритма распознавания рукописного текста с помощью архитектуры CRNN + CTC Loss с примерами.

Итоги. В третьем разделе были рассмотрены основные понятия, связанные с теорией нейронных сетей, были определены алгоритмы работы нейронных сетей, используемые в данной работе. Также был подробно рассмотрен алгоритм работы архитектуры CRNN + CTC Loss.

Четвертый раздел «Инструментарий и технологии» посвящен описанию технологий, с помощью которых была реализована практическая часть работы.

В **подразделе 4.1 «Инструментарий и технологии автоматизированной генерации и считывания бланков»** говорится об инструментариях и технологиях, выбранных для написания приложения автоматизированной генерации и считывания бланков. Был выбран язык программирования C# 7.3. Платформа – .NET Framework 4.7.2. При составлении форм бланка использовался Microsoft Word 18. Для генерации и считывания бланков использовалась библиотека Microsoft.Office.Interop.Word, которая нужна для работы с API Microsoft Word. Библиотеки newtonsoft.json, pdfsharp.

В **подразделе 4.2 «Инструментарий и технологии предобработки изображений»** описывается инструментарий и технологии, выбранных для предварительной обработки изображений. Язык программирования Python 3.10. В работе использовалась библиотека OpenCV, которая представляет собой библиотеку компьютерного зрения с открытым исходным кодом. Пакет Imutils, основанный на OpenCV, который может реализовать серию операций, таких как перевод изображения, вращение, масштабирование и скелет. Для изменения размера изображений потребовалось использование библиотеки изображений Python – PIL (Python Imaging Library), которая необходима для обработки графики в Python. Также был необходим математический модуль statistics – для сложных вычислений, и библиотека NumPy – это библиотека языка Python, добавляющая поддержку больших многомерных массивов и матриц.

В **подразделе 4.3 «Инструментарий и технологии обучения нейронной сети CRNN + CTC»** говорится о выбранном инструментарии для построения архитектуры нейронной сети и ее последующего обучения. Был выбран язык программирования Python 3.10. Также была выбрана среда разработки Google Colab – облачный сервис на основе Jupyter Notebook.

Главным образом была использована библиотека глубокого обучения PyTorch — фреймворк машинного обучения для языка Python с открытым исходным кодом. Для визуализации данных – библиотека Matplotlib. Также понадобились такие библиотеки, как numpy, OpenCV, json, os.

Итоги. Изученные в рамках четвертого раздела инструменты и технологии позволяют реализовать приложение для автоматизированной генерации и считывания бланков и в полной мере решить задачу распознавания рукописного текста в режиме оффлайн.

Пятый раздел «Реализация системы автоматизированного распознавания рукописного текста» посвящен реализации системы автоматизированного распознавания рукописного текста в режиме оффлайн. Таким образом, в данном разделе описана разработка приложения для автоматизированной генерации и считывания бланков, предобработка собранных изображений и построение архитектуры нейронной сети с ее последующим обучением.

В **подразделе 5.1 «Сбор данных и реализация приложения для автоматизированной генерации бланков и их последующего считывания»** описывается процесс разработки приложения для автоматизированной генерации и считывания бланков. Задается шаблон, наполняемый необходимым контентом. И далее происходит автоматизированная разметка заполненных бланков. С помощью данного приложения было собрано 279 бланков. В общей сложности это 3802 слова (из них 625 различных) и 438 предложения.

В **подразделе 5.2 «Предварительная обработка изображений»** последовательно рассмотрен процесс предварительной обработки изображений. На данном этапе были обработаны и приведены к единому размеру все имеющиеся изображения. После этапа предобработки из 3802 слов осталось 3415 в результате потерь из-за незаполненных или зачеркнутых полей. Таким образом, было собрано 3415 изображения формата jpeg, размером 256 × 64 пикселей (Рисунок 43) и опубликовано на публичной

веб-платформе исследования и сбора данных kaggle по URL адресу: https://www.kaggle.com/datasets/lelalolkaaa/russian-handwriting-text.

В **подразделе 5.3 «Построение нейронной сети CRNN + CTC»** описывается процесс построения нейронной сети CRNN + CTC.

В **подразделе 5.4 «Обучение нейронной сети на собранных данных»** дается описание процесса обучения нейронной сети на самостоятельно собранных данных в количестве 3415 изображений. В результате полученная модель сильно ошибается в распознавании слов, но некоторые буквосочетания определяет верно. После обработки уже полученных прогнозов модели функцией исправлений орфографических ошибок прогнозы немного улучшились, но некоторые метрики стали хуже из-за специфики выбранного модуля.

В **подразделе 5.5 «Обучение нейронной сети на собранных данных»** описывается процесс обучения нейронной сети на данных сообщества ODS в количестве 66599 изображений. Полученная модель имеет неплохие прогностические способности и достаточно хорошо прогнозирует слова, когда не имеется явных дефектов почерка и помарок в написании. Также полученная модель показала неплохие результаты на данных, собранных самостоятельно, несмотря на то что предобработка изображений значительно отличается.

В **подразделе 5.6 «Сравнение результатов»** были проанализированы результаты работы моделей. Модели, полученные в результате обучения одной нейронной сети, но на разных количествах данных демонстрируют довольно сильно отличающиеся прогностические способности: модель, обученная на большом количестве данных, предсказывает слова в 4 раза точнее. А модель, построенная на малом количестве данных, предсказывает полностью верные слова крайне редко, но способна определять буквосочетания. Применение орфографического модуля способно не на много, но все же улучшить даже достаточно плохо обученную модель для распознавания рукописного текста. А модель, которая была обучена на

данных сообщества ODS, неплохо распознает собранные данные, несмотря на тот факт, что предобработка данных обоих датасетов достаточно отличается.

Итоги. В рамках пятого раздела была разработана система автоматизированного распознавания рукописного текста в режиме оффлайн: разработано приложение для автоматизированной генерации и считывания бланков, произведена предварительная обработка полученных изображений, построена и далее обучена архитектура нейронной сети CRNN + CTC. И в результате был произведен анализ получившихся результатов.

ЗАКЛЮЧЕНИЕ

Распознавание рукописного текста до сих пор является одной из самых исследуемых задач, которая способна не только значительно упрощать деятельность человека, но и, к примеру, применяться в задачах лингвистики по расшифровке рукописей или других целей. Для создания такой модели, в первую очередь, необходимо большое количество данных рукописного текста. В данной работе был собран датасет русского рукописного текста и опубликован на публичную веб-платформу kaggle. Любой желающий сможет воспользоваться собранными данными в исследовательских целях и также добавить свои, в результате чего может быть собран полноценный датасет. Были изучены и опробованы на практике методы предварительной обработки изображений, очищающие изображение от лишних шумов, помех и приводящие данные к единому виду.

Также в данной работе был создан инструментарий автоматизированной генерации бланков и их последующего считывания. Данные приложения являются незаменимыми помощниками при сборе подобных баз данных, в разы уменьшая время составления и обработки бланков.

Была построена архитектура нейронной сети CRNN + CTC Loss, обученная модель которой показала неплохие результаты на данных

сообщества ODS. Полученная модель имеет достаточные прогностические способности, ошибки возникают только при явных дефектах в написании слов. Модель на собранных данных показывает не такие хорошие результаты, но тем не менее способна угадывать определенные символы, что говорит только о недостаточном объеме данных.

Таким образом, в результате данной работы была разработана система автоматизированного распознавания русского рукописного текста в режиме оффлайн.

Основные источники информации:

1. Кеннет М. Сэйр. Машинное распознавание рукописных слов: отчет о проекте. Распознавание образов // Pergamon Press. – 1973. – Издание 5. – С. 213-228

2. Ciresan, D. Multi-column deep neural networks for image classification. Technical report / Meier, U., and Schmidhuber, J // Conference on Computer Vision and Pattern Recognition. – 2012. – Vol. 3. – P. 3642-3649

3. Кучуганов А.В. Распознавание рукописных текстов / Лапинская Г.В. – Ижевск:.Мир, 2006

4. Rahul Kala. Offline handwriting recognition using genetic algorithm / Harsh Vazirani, Anupam Shukla, Ritu Tiwari // International Journal of Computer Science Issues. – 2010. – Vol. 7. – Is. 2

5. Kunhong Yu. Digging Deeper into CRNN Model in Chinese Text Images Recognition / Yuze Zhang. – ArXiv:2011.08505, 2020

6. Гафаров Ф.М. Искусственные нейронные сети и приложения: учеб. пособие / А.Ф. Галимянов – Казань: Изд-во Казан. ун-та, 2018

7. Mark Potanin. Digital Peter: Dataset, Competition and Handwriting Recognition Methods / Vladimir Bataev, Denis Karachev, Maxim Novopoltsev. – ArXiv:2103.09354, 2021