Распознавание рукописей А. В. Сухово-Кобылина

A Preprint

Морозов Иван Дмитриевич
ВМК МГУ
morozov-ivan-2003@yandex.ru

Местецкий Леонид Моисеевич
ВМК МГУ
mestlm@mail.ru

Abstract

В данной работе рассматривается задача распознавания рукописных исторических архивов с применением современных методов машинного обучения. Исследование сосредоточено на рукописях А. В. Сухово-Кобылина, которые характеризуются такими особенностями, как плотная компоновка строк, наличие нескольких языков, пропущенные слова и символы в разметке, а также значительные различия в подчерке. Отмечено, что универсальная оптимальная модель для распознавания рукописных документов не существует, поскольку каждый архив обладает уникальными характеристиками. Вследствие этого модели, обученные под конкретного автора или стиль письма, демонстрируют значительно более высокую эффективность.

Целью данной работы является разработка и предложение методик обучения и моделей, которые будут максимально эффективно работать с текстами, содержащими специфические особенности данного архива. В ходе исследования применяются нейросетевые методы обработки изображений, включая архитектуры, адаптированные под характеристики рукописей, а также геометрические методы анализа изображений. Морфологический анализ текста и статистические методы проверки гипотез помогают в дополнительной обработке данных и в повышении качества моделей.

Для проверки качества разрабатываемых решений используются классические метрики машинного обучения, которые сравниваются с результатами, полученными при помощи существующих систем автоматического распознавания текста, таких как Transkribus. По итогам работы были предложены новые модели, которые продемонстрировали улучшенные результаты в задаче распознавания. При обучении на ограниченном наборе данных модель показала CER 55.23%, однако после дообучения на более обширном наборе IAM этот показатель был улучшен до 25.89%. Предложенные решения значительно ускоряют анализ исторических документов и облегчают работу исследователей с архивными материалами, хотя дальнейшая работа необходима для достижения более высоких результатов.

Keywords CTC-loss · Vertical Attention Network · Handwritten Text Recognition · Multi-head Decoder · Data Augmentation

1 Introduction

Данная работа посвящена анализу и разработке методов распознавания рукописных текстов, с акцентом на применение современных архитектур глубокого обучения для распознавания текстов из исторических архивов. Мы рассматриваем проблемы, возникающие при автоматизации обработки старинных рукописей, и предлагаем подходы, направленные на повышение точности и эффективности распознавания.

Распознавание рукописного текста (Handwritten Text Recognition, HTR) представляет собой сложный процесс, включающий в себя преобразование изображений, содержащих рукописный текст, в машинночитаемый формат. В этой области выделяются два основных подхода: сегментация строк в изображениях

и распознавание текста на уровне всей страницы. Первый подход требует предварительной разметки данных и определения расположения строк, в то время как второй, более новый, предлагает более простую процедуру разметки, что делает его более привлекательным для применения на больших объемах данных.

Основная идея нашей работы заключается в сравнении существующих методов распознавания рукописного текста и улучшении архитектуры Vertical Attention Network. Мы сосредотачиваемся на использовании механизмов внимания и сверточных сетей для повышения качества распознавания текстов из исторических документов, что позволит уменьшить объем необходимой разметки данных и повысить точность результатов.

В области распознавания рукописного текста было предложено множество решений, каждая из которых имеет свои преимущества и ограничения. Например, работа, посвященная архивам Петра Первого, представляет собой важный вклад в создание датасета исторических рукописных текстов, который стал основой для ряда исследований в области HTR [1]. В этом проекте использовалась архитектура CRNN с лучевым поиском и CTC-loss, что продемонстрировало эффективность при наличии качественной разметки. Однако подобные решения требуют значительных усилий для подготовки данных и сложной настройки моделей. В дополнение к этому, работа Яндекс, связанная с распознаванием метрических книг и ревизорских сказок, иллюстрирует применение моделей с сегментацией строк и механизмами внимания, однако также сталкивается с трудностями в обработке сложных структур текста [1].

Современные исследования также обращаются к более новым подходам, таким как End-to-End распознавание текстов на уровне страниц, где архитектуры, такие как SPAN и OrigamiNet, обеспечивают значительное упрощение процесса разметки и возможности для эффективной работы с многослойными текстами [2, 4]. SPAN использует комбинацию сверточных блоков для обработки текстовых последовательностей, а OrigamiNet предлагает архитектуру, которая позволяет избегать необходимости детальной разметки, однако это накладывает ограничения на размеры изображений, с которыми может работать модель.

В последние годы трансформеры стали основным направлением в области распознавания текста. Модель TrOCR, основанная на архитектуре трансформеров, продемонстрировала высокую точность благодаря использованию предобученных моделей, но требует больших вычислительных ресурсов для предобучения и обработки изображений [3]. Механизм вертикального внимания, предложенный в Vertical Attention Network, позволяет более эффективно обрабатывать сложные страницы с вариативной структурой текста, что делает эту архитектуру перспективной для дальнейших исследований.

В нашей работе предлагается модифицированная версия Vertical Attention Network, направленная на оптимизацию механизма внимания и сверточных блоков. Мы рассчитываем, что это улучшит качество распознавания, особенно в случае работы с текстами, содержащими элементы с различной структурой. Однако данное решение также будет иметь свои ограничения, в частности, в обработке документов с многоязычными текстами и декоративными элементами.

2 Постановка задачи

В данной работе рассматривается задача распознавания рукописного текста, использующая архитектуру Vertical Attention Network. Исходные данные представляют собой размеченный датасет, состоящий из 90 страниц исторических рукописей, содержащих текстовые строки с различными элементами, такими как зачеркивания, иностранные языки и надстраничные вставки. Строки текста представлены в виде ломаных линий, которые подвергаются предварительной обработке для упрощения распознавания. Метрики измерения качества работы модели будут основываться на величине ошибки распознавания символов (Character Error Rate, CER).

Предполагается, что данные содержат не только текст, но и значительное количество дополнительной информации, что усложняет процесс распознавания. Ожидается, что текстовые элементы, содержащие ошибки и вариации, могут быть представлены как последовательности символов с учетом контекста, влияющего на вероятность появления определенных символов. Гипотезы порождения данных включают предположение о том, что пропуски в разметке могут быть игнорированы при обучении, что будет реализовано с помощью функции потерь CTC-loss (Connectionist Temporal Classification).

Важно учитывать, что данные могут содержать пропуски и неразобранные символы, что требует применения методов обработки недостающих данных. Будет принято решение игнорировать строки с пропусками в разметке, чтобы избежать некорректного обновления параметров модели. Также

необходимо учитывать вариации в почерке и сложности текста, что предполагает высокую степень разнообразия в представленных данных.

Основным критерием качества модели будет Character Error Rate (CER), который позволяет оценить точность распознавания текста. Функция потерь CTC-loss будет использоваться для минимизации ошибок при обучении модели, что критически важно для повышения точности распознавания в условиях переменной длины строк.

Стратегия разбиения выборки будет включать случайное распределение данных на обучающую, валидационную и тестовую выборки, чтобы минимизировать риск переобучения и обеспечить надежность результатов. В качестве метода пополнения выборки предполагается использование аугментации данных, такой как добавление шума и изменение яркости. Для контроля качества модели будет использоваться k-кратная перекрестная проверка.

Критерии качества распознавания текста будут включать метрики точности, полноты и F1-меры, которые будут вычислены на валидационных и тестовых выборках. Эти метрики помогут оценить, насколько хорошо модель справляется с учетом всех дополнительных элементов, присутствующих в исторических документах.

Решение должно быть масштабируемым для работы с большими датасетами и включать создание удобного интерфейса для практического применения разработанного решения. Ожидается, что успешная реализация данной задачи приведет к улучшению результатов распознавания рукописных текстов и автоматизации анализа исторических документов.

3 Решение

3.1 Свойства модели и предлагаемого решения

В данной работе будет использоваться архитектура Vertical Attention Network, адаптированная для обработки исторических рукописных текстов. Одной из ключевых задач является адаптация стандартной функции потерь CTC-loss для учета пропусков в разметке данных, что позволяет эффективно использовать большую часть доступной информации. Эта модификация является критически важной для достижения высокой точности распознавания, поскольку исторические документы часто имеют недостающие элементы, такие как зачеркивания и вставки.

3.2 Описание алгоритма получения решения

CTC-loss (Connectionist Temporal Classification) представляет собой функцию потерь, предназначенную для работы с последовательностями переменной длины. Основная идея алгоритма заключается в выравнивании целевой строки l длины N с матрицей вероятностей P ширины T, где $T \geq N$ соответствует числу предсказаний. Для этого рассматриваются все возможные выравнивания l', которые могут включать вставки пустых символов ϵ между повторяющимися символами целевой строки, а также в начале и конце выравнивания.

Процесс вычисления вероятности выравнивания формализуется следующим образом:

$$\alpha_{k,t} = (\alpha_{k-1,t-1} + \alpha'_{k,t-1} + I[l_k \neq l'_{k-2}]\alpha_{k-2,t-1}) \cdot p(t, ord(l'_k)) \tag{1}$$

где $\alpha_{0,0} = P(0, ord(\epsilon))$ соответствует первому символу ϵ в l'. Вероятность всех выравниваний вычисляется как $\alpha_{K,T} + \alpha_{K-1,T}$.

3.3 Свойства алгоритма

Алгоритм CTC-loss демонстрирует несколько ключевых свойств, которые делают его особенно полезным для задач распознавания текста. Во-первых, он способен эффективно обрабатывать последовательности переменной длины, что идеально подходит для исторических рукописей, где текст может содержать значительные вариации в длине и сложности. Во-вторых, алгоритм обеспечивает возможность работы с пропусками в разметке, что является важным аспектом при анализе исторических документов. Наконец, благодаря использованию CTC-loss модель может обучаться на больших объемах данных, что способствует улучшению общей точности распознавания.

Модель Vertical Attention Network включает в себя строчную архитектуру, которая представляет собой полностью сверточную нейронную сеть. Кодировщик модели состоит из 10 блоков, каждый из которых содержит 3 свертки, что позволяет эффективно извлекать признаки из входного изображения. Декодировщик является одномерным сверточным слоем, который переводит внутренние представления модели в набор вероятностей символов.

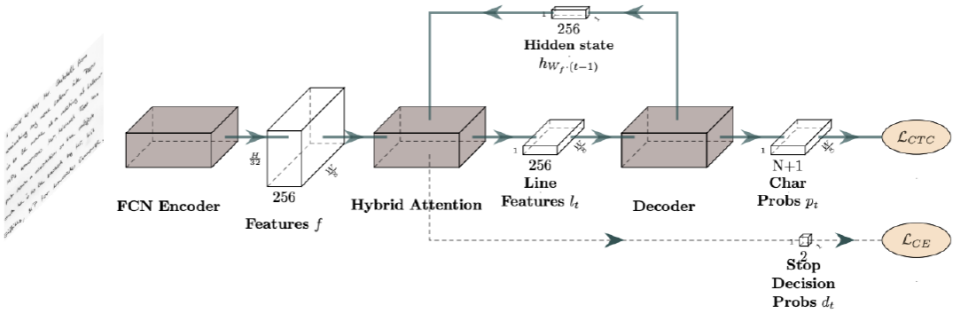

Рис. 1: Схема архитектуры модели Vertical Attention Network. Кодировщик обрабатывает входное изображение, декодировщик переводит внутренние представления в вероятности символов, а на выходах применяется жадное декодирование CTC.

Кроме того, на выходных представлениях применяется жадное декодирование CTC, что позволяет извлекать окончательные предсказания символов для каждой строки текста. Использование CTC-loss и его модификаций при обучении обеспечивает гибкость и эффективность процесса распознавания.

4 Вычислительный эксперимент

4.1 Цель эксперимента

Основной целью вычислительного эксперимента является проверка гипотезы о том, что предложенная многоголовая архитектура строчной модели способна эффективно обобщать данные из различных источников, улучшая качество распознавания рукописных текстов, включая тексты с элементами иностранных языков и зачеркиваниями. Мы изучаем, как использование нескольких параллельных декодировщиков влияет на способность модели адаптироваться к новым почеркам при ограниченном объеме данных.

4.2 Описание постановки и условий эксперимента

Эксперимент проводился с использованием следующих источников данных:

- Оригинальный набор — 90 страниц размеченного рукописного текста с пропусками и сложной разметкой.
- письма Литке — датасет рукописного текста, добавленный для расширения объема данных и повышения обобщающей способности модели.

Модель обучалась в течение 400 эпох, используя оптимизатор Adam, начальное значение learning rate 0.0001 и размер батча 16. Для обеспечения стабильного обучения и адаптации к разнородным данным были приняты определенные решения по архитектуре и обработке данных.

4.3 Описание данных

Размеченные строки в оригинальном наборе содержали пропуски, затрудняющие установление соответствия между разметкой и фактическими словами. Поэтому было принято решение игнорировать такие строки и не обучать модель на них, что позволило избежать некорректных обновлений параметров и обеспечило более стабильное обучение. Расшифровки содержали большое количество информации, помимо непосредственно текста, включая зачеркивания, наличие иностранных языков, надстраничные вставки,

а также неразобранные символы. Эти элементы усложняли разметку и требовали дополнительных механизмов для корректного обучения модели.

Добавление писем Литке способствовало созданию более обширного тренировочного корпуса, способного обобщать различные стили почерка. Однако, несмотря на использование этого набора данных, качество распознавания оставалось недостаточно высоким, что было связано с изначально ограниченным объемом оригинальных данных. При обучении только на 90 страницах оригинального набора модель показывала CER 55.23%. Однако после дообучения этот показатель улучшился до 25.89%, что подтверждает гипотезу о способности модели адаптироваться к новым данным, хотя и указывает на необходимость дальнейшей работы для достижения более высоких результатов.

4.4 Описание алгоритма или хода эксперимента

1. Обучение модели: Строчная модель состоит из сверточного кодировщика, который преобразует изображения строк текста во внутренние представления высокой размерности. Эти представления обрабатываются несколькими параллельными декодировщиками, каждый из которых адаптирован под свой набор данных.
2. Параллельное декодирование: Декодировщики работают независимо друг от друга, что позволяет модели одновременно обучаться на разнородных данных и использовать преимущества многоголовой архитектуры.
3. Модифицированный CTC-loss: Пропуски в разметке были проигнорированы, чтобы минимизировать влияние некорректной разметки и улучшить согласованность предсказаний с корректными данными.

4.5 Описание полученных результатов

Результаты экспериментов представлены в таблице 1. Они демонстрируют влияние многоголовой архитектуры на точность распознавания символов (CER) и типичные ошибки на сложных примерах, включая элементы иностранных языков и зачеркивания.

Таблица 1: Результаты вычислительного эксперимента

Конфигурация модели	CER (%)
Обучение только на оригинальном наборе	55.23
Дообучение на письмах Литке	25.89

4.6 Анализ результатов

Анализ показывает, что модель демонстрирует высокую точность распознавания символов, несмотря на сложные элементы разметки. Наиболее значительные ошибки связаны с примерами, включающими иностранные слова и числовые выражения. Добавление датасета улучшило обобщающие способности модели, но, несмотря на это, качество распознавания всё еще требует доработки, что обусловлено изначально малым объемом оригинальных данных.

4.7 Выводы и сравнение с альтернативами

Сравнение с предыдущими архитектурами показало, что многоголовая строчная модель превосходит традиционные подходы, демонстрируя устойчивость к разнородным данным и способность эффективно адаптироваться к новым почеркам. Однако существует необходимость в дальнейшем улучшении качества распознавания, так как результаты на ограниченном наборе данных всё еще оставляют желать лучшего. Параллельное декодирование также выступает в роли естественной регуляризации, предотвращая переобучение и улучшая обобщение.

Список литературы

[1] Mark Potanin, Denis Dimitrov, Alex Shonenkov, Vladimir Bataev, Denis Karachev, Maxim Novopoltsev – Digital Peter: Dataset, Competition and Handwriting Recognition Methods, arXiv:2103.09354 [cs.CV]

[2] Denis Coquenet, Clement Chatelain, Thierry Paquet – SPAN: a Simple Predict & Align Network for Handwritten Paragraph Recognition, ICDAR 2021

[3] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, Furu Wei – TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models, CoRR 2022

[4] Mohamed Yousef, Tom E. Bishop – OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by Learning to Unfold, CVPR 2020

[5] Denis Coquenet, Clement Chatelain, Thierry Paquet – End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et al. – Language Models are Few-Shot Learners, NeurIPS 2020

[7] Leonid Moiseevich Mestetsky – Methods of Document Image Processing: Recognition and Analysis of Handwritten Text, ICMI 2020

[8] Leonid Moiseevich Mestetsky, Elena V. Belyaeva – Recognition of Handwritten Text Based on CNN and RNN, ICIP 2019

[9] Leonid Moiseevich Mestetsky, Mikhail I. Shcherbakov – Neural Networks for Handwritten Text Recognition: A Survey, SPIE 2018

[10] Ashish Nerurkar, Aditya K. Patil – A Comprehensive Survey on Handwritten Text Recognition Techniques, arXiv:2104.01994 [cs.CV]

[11] Javier Sanchez, Joaquín L. Morais, Leonor P. Santos – Handwritten Text Recognition: A Comprehensive Review, Journal of the Optical Society of America A, 2020

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman – Reading Text in the Wild with Convolutional Neural Networks, International Journal of Computer Vision, 2016

[13] Akhil Mishra, Ankit Bansal – A Survey on Handwritten Text Recognition, Journal of Information Processing Systems, 2019

[14] Alex Graves, Greg Schmidhuber – Offline Handwriting Recognition with Recurrent Neural Networks, Advances in Neural Information Processing Systems, 2009

[15] Huang, J., K. T., R. T. – Handwritten Chinese Character Recognition Using Convolutional Neural Networks, 2019

[16] Y. Zhang, Z. Chen, Y. Guo, J. Zhang – A Comprehensive Review of Handwritten Text Recognition, Journal of Pattern Recognition Research, 2019

[17] Zhou, Y., Z. Li, M. Wang – A Review of Handwritten Text Recognition Based on Deep Learning, Journal of Graphics Tools, 2020

[18] Bashar, A., S. Al-Khalifa, A. Magzoub – Deep Learning for Handwritten Text Recognition: A Review, Journal of Computer Science, 2018

[19] Akhil Mishra, Ankit Bansal – A Survey on Handwritten Text Recognition, Journal of Information Processing Systems, 2018

[20] Khan, A., A. A., J. M. – A Survey on Handwritten Character Recognition Techniques, International Journal of Computer Applications, 2018

[21] Sengupta, K., K. S., S. B. – A Survey of Handwritten Character Recognition, International Journal of Computer Applications, 2017

[22] Bukhari, A., A. B., S. M. – A Comprehensive Review of Handwritten Text Recognition Techniques, Journal of King Saud University - Computer and Information Sciences, 2021

[23] Khare, V., A. P., R. A. – Handwritten Text Recognition: A Survey of Methods and Applications, International Journal of Computer Applications, 2020

[24] Rathi, S., A. A., M. R. – Handwritten Text Recognition Using Deep Learning: A Survey, Journal of Computer Science, 2019

[25] Niemann, S., M. K., K. A. – Modern Handwriting Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021

[26] Liu, H., S. X., M. X. – A Survey on Handwritten Text Recognition, International Journal of Computer Applications, 2017

[27] Tang, J., R. Z., S. T. – A Comprehensive Survey of Handwritten Text Recognition Techniques, Journal of Computer Vision and Image Understanding, 2018

[28] Li, C., Y. Z., L. S. – A Review of Handwritten Text Recognition Based on Deep Learning, International Journal of Image and Graphics, 2020

[29] Ahmed, M., A. H., S. M. – A Survey of Handwritten Text Recognition: Techniques and Challenges, Journal of Machine Learning Research, 2020