

Распознавание рукописей А. В. Сухово-Кобылина

Морозов Иван Дмитриевич

ВМК МГУ имени М. В. Ломоносова

22 ноября 2024 г.

Цели

- ▶ Разработать модель для распознавания исторических рукописей.
- ▶ Произвести разбивку страниц рукописей на отдельные строки, выполнить нормализацию
- ▶ Оптимизировать методы для работы с пропусками, зачеркиваниями, многозначными текстами.
- ▶ Снизить Character Error Rate (CER).
- ▶ Создать аугментации, адаптированные для рукописных документов.

Распознавание рукописей А. В. Сухово-Кобылина

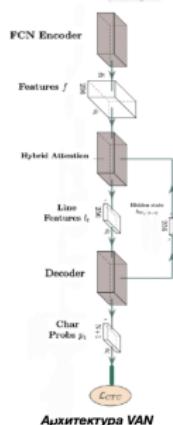
Главная цель:



- ▶ Расшифровка рукописей (≈ 10000 страниц) с Character Error Rate менее 10%

Альтернативная цель:

- ▶ По слабой расшифровке разработать алгоритм выделения ключевых слов



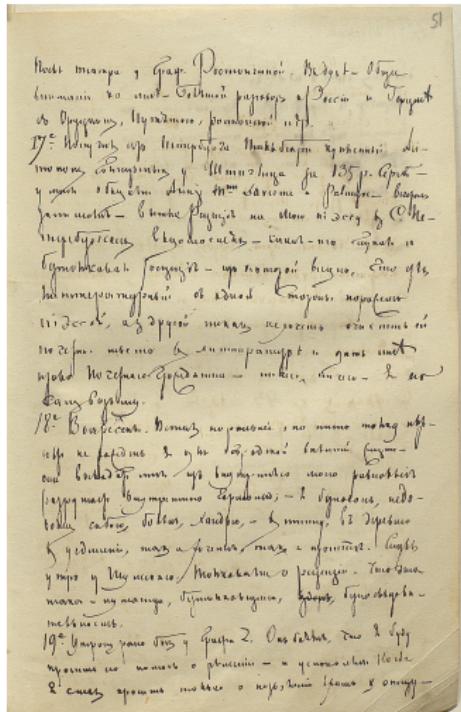
Ключевые особенности решения:

1. **Предобработка данных.** Подготовка и разметка датасета.
2. **Использование СТС-loss.** Использование для работы с пропусками и неравномерностью.
3. **Архитектура Vertical Attention Network.** Анализ рукописных страниц с вертикальным вниманием.

Литература

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 and Tengchao Lv and Jingye Chen and Lei Cui and Yijuan Lu and Dinei Florencio and Cha Zhang and Zhoujun Li and 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

Гипотеза



Гипотеза: возможно распознать тексты Сухово-Кобылина с достаточно высоким качеством, используя архитектуру VAN, предобученную на датасете IAM.

Постановка задачи

Предложить метод обучения и модели, эффективно работающие на архиве дневников Сухова-Кобылина. Главными особенностями данного архива является достаточно неплотная компоновка строчек, наличие нескольких языков, пропущенные слова и символы в разметке. Для каждой особенности будут рассматриваться решения, позволяющие обучить модель и улучшить ее итоговое качество.

Описание алгоритма СТС

- ▶ Рассматривается целевая строка I длины N и матрица вероятностей P ширины $T \geq N$, соответствующей числу предсказаний.
- ▶ Выравнивания I' формируются путем:
 - ▶ вставки символов ϵ между всеми повторяющимися символами,
 - ▶ произвольного повторения любого символа,
 - ▶ добавления ϵ в начало и конец выравнивания.
- ▶ Стока $[\epsilon, I_1, \epsilon, I_2, \dots, I_N, \epsilon]$ имеет длину $2 \cdot N + 1 = K$.
- ▶ Динамическое пересчет выравниваний $\alpha_{k,t}$ для левых подстрок целевой строки:

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

где $\alpha_{0,0} = P(0, \text{ord}(\epsilon))$ для первого символа ϵ в I' .

- ▶ Итоговая вероятность всех выравниваний:

$$P(I|X) = \alpha_{K,T} + \alpha_{K-1,T}.$$

СТС-loss:

$$\mathcal{L}_{CTC} = -\ln P(I|X).$$

Критерий качества

Основные метрики качества:

- ▶ Character Error Rate (CER):

$$CER = \frac{\text{Число ошибок}}{\text{Общее количество символов}} \cdot 100\%$$

где ошибки включают вставки, удаления и замены символов.

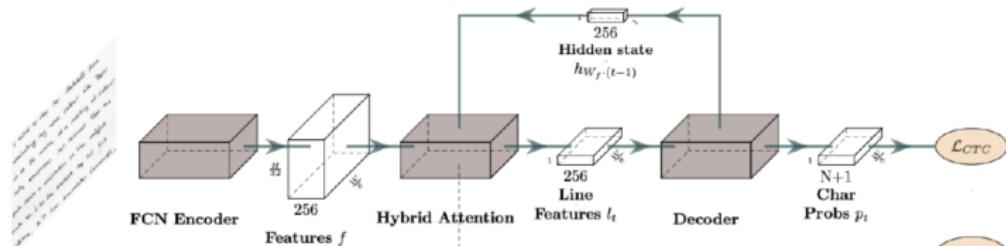
- ▶ Word Error Rate (WER):

$$WER = \frac{S + D + I}{N} \cdot 100\%$$

где:

- ▶ S — количество замен слов,
- ▶ D — количество удалений слов,
- ▶ I — количество вставок слов,
- ▶ N — общее количество слов в эталонной разметке.

Описание модели



Будет использоваться модель Vertical Attention Network в строчном варианте. Строчная архитектура представлена полностью сверточной нейронной сетью, состоящей из кодировщика, содержащего 10 блоков с 3 свертками в каждом, и одномерного сверточного декодировщика, который переводит внутреннее представление модели в набор вероятностей. Модель выдаёт пространственное предсказание символов в строке, поэтому к выходам применяется жадное декодирование СТС, а при обучении используется СТС-loss.

Нарезка страницы на строки

А.В.Сурово Кобицин. Личный дневник. Том 438-1-219

Дневник Александра Васильевича Сухово-

Кобыльского

Маскот Ремонт страниц Нарезка строк

Нарезка строк в файле Том_438-1-219_Страница_706.html

1 Апрель Это. Уехать в Петербург, ибо дядя принял ду
2 обратить в Им. Юст. Настало тихое время. Саки съ
3 грибами. Красивые грибы. Саки съелись. Саки были до
4 срока. Ходить довольно рано гуашь на подору - и
5 пахнет из Невы с пологими апельс.

6 Сентябрь В Петербург. Где-то запнулся и, ф-ф-о
7 предупредил я Ю. России. Отправился на Выху и
8 заключил условие с Ник. Шепелем. Привез обратно
9 из Петербурга 130 небольш. Соловьев Саки и Наненцы в
10 Петербург. Жил в гостинице Демидова. Поставлен
11 на конвой для дачи по утрам. Секунд 200 и 300 акт. \$
12. Затем в Государственную Торговую инспекцию.
13 корабля. Кильев и Шипилов в Петербург. Дней.
14 Возвращение из Москвы. \$
15 1854 год. Новый год на желтой дороге. Быть в Ворон
16 сенокосе. 15 часов выехал из Выху. Апрель Гол. \$
17 Тёплая читалась с утра. В ферзати
18 должно быть выпущено из Москвы по поводу начавшегося
19 съездства. Быть на «кошкою» дне в Петербург. \$
20 Покупать скотину вещи, чтобы скотин.

Задать строку Вставить выше Вставить ниже Редактировать

Копировать страницу в Clipboard Сохранить изменения страницы

Назад Создать Виджеты Просмотр Отмена

Том 438-1-219. Страница 706 Просмотр Отмена

Имена записаны Сухово-Кобыльским. Петербург

Результаты обработки данных и обучения модели

- ▶ В разметке историков присутствовали нераспознанные элементы строк, обозначенные специальными символами.
- ▶ Принято решение удалить такие строки, так как алгоритм локализации пропусков пока не разработан.
- ▶ Используется предобученная на англоязычном IAM модель (модель умеет выделять буквы в строке)
- ▶ Итоговое разбиение данных:
 - ▶ **Обучение (train)**: 1505 строк.
 - ▶ **Валидация (validation)**: 119 строк.
 - ▶ **Тест (test)**: 74 строки.
- ▶ Результаты дообучения (CER, %):
 - ▶ **Обучение (train)**: 0.37%.
 - ▶ **Валидация (validation)**: 18.08%.
 - ▶ **Тест (test)**: 19.03%.

Анализ ошибок

Pred: Рябикова
gt: Рябников,
seg: 0.250



Pred: .Вечеромъ литаль піссу - ядъ чревыоймо
gt: Вечеромъ читаль піссу - Дядъ чревыайин,
seg: 0.200

Pred: понравилась. – бекаеру также.
gt: понравилась – Беккеру также,
seg: 0.148

Pred: 10-г. утромъ выѣхалъ въ дѣксеевскoe. Почваль в
gt: 10-го утромъ выѣхалъ въ Алексеевское. Ночеваль в,
seg: 0.104

Результаты и заключение

- ▶ Снижение CER до 19%.
- ▶ Нормализация строк улучшает результаты.
- ▶ Подход подходит для сложных рукописных текстов (модель распознает текст, который обычному человеку распознать затруднительно).

Будущая работа:

- ▶ Расширение датасета.
- ▶ Бинаризация изображений строк.
- ▶ Удаление свисающих элементов от других строк.
- ▶ Центрирование строк на выделенных фрагментах.