Распознавание рукописных архивов А. В. Сухово-Кобылина

A Preprint

Зыков Валерий Павлович
ВМК МГУ
valera_zykov_2003@mail.ru

Местецкий Леонид Моисеевич
ВМК МГУ
mestlm@mail.ru

2024

Abstract

В данной работе будет решаться задача распознавания архива исторических рукописных архивов. Нет универсальной оптимальной модели для распознавания рукописных исторических документов, потому что каждый из них содержит свои особенности. Кроме того, из-за большого отличия в почерках, модели, обучающиеся под конкретного автора, оказываются эффективнее. Цель данной работы – предложить метод обучения и модели, эффективно работающие на архиве дневников Александра Васильевича Сухово-Кобылина. Главными особенностями данного архива является плотная 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.

1 Введение

1.1 Подходы к распознаванию

Существует два основных подхода к распознаванию страниц текста:

- Сегментация строк в изображении и распознавание текста в строках.

  Отдельная модель или процедура находит строки, которые затем поступают в модель распознавания. Для обучения нужна построчная разметка и подготовленные изображения. Для решения задачи детекции также нужна информация о расположении строк в странице.

  В близких работах с распознаванием русскоязычных исторических архивов рассматриваются именно такие подходы.

  Архивы Петра Первого (1). Основным вкладом данной работы является создание датасета на основе архива исторических рукописных текстов Петра Первого. Помимо датасета вокруг решения задачи распознавания было проведено соревнование, главным треком в котором было распознавание строк. Лучшая архитектура была описана в статье и представляет собой CRNN (Convolutional Recurrent Neural Network) с лучевым поиском с N-граммной моделью, обучается с CTC-loss.

  Яндекс: Поиск по архивам. В данной работе также были подготовлены массивы исторических документов и решалась задача их расшифровки, такими документами являлись метрические книги, ревизорские сказки, которые позволяют проводить генеалогические исследования. Для страниц была составлена разметка на строчки, а также на смысловые блоки. Сама модель состояла из шага сегментации строк в виде Gaussian heatmap, а для распознавания использовался сверточный кодировщик и RNN-декодировщик с механизмом внимания, поэтому она обучалась с использованием кросс-энтропии. Блоки строк находились с моделью Instance Segmentation.

  Наиболее перспективным архитектурным решением оказалось использование трансформерной кодировщик-декодировщик архитектуры в статье TrOCR (3). Авторы отказались от сверточных слоев и подают в кодировщик уменьшенное исходное изображение, нарезанное на патчи.

Декодировщик генерирует bpe-токены. Модель обучается с кросс-энтропией и использует для инициализации предобученные ViT и RoBERTa.

- Распознавание текста на уровне страниц. Недавно появившийся и развивающийся подход- End-to-End распознавание страниц с текстом. Преимущество заключается в упрощенной процедуре разметки и конечного использования модели. Рассмотрим несколько примеров архитектур. Они предполагают моноколоночную структуру текста.

  SPAN (2) – наиболее простая архитектура. Концептуально представляется как комбинация энкодера – последовательности сверточных блоков с пулингом и depthwise сверточных блоков с residual-связями, и декодера – одного сверточного слоя. Декодер моделирует 2d матрицу символов, с векторами вероятностей символов в каждой клетке. Затем эта матрица с помощью последовательной горизонтальной конкатенации приводится в одну строку- цепочку векторов вероятностей символов. Далее эта цепочка преобразовывается по правилам CTC.

  Vertical Attention Network (5) также содержит сверточный энкодер, но кроме того для нее был предложен механизм вертикального внимания. Этот механизм позволяет рекуррентно собирать представления для строк в странице, которые затем приводятся к матрицам вероятностей сверточным и lstm-декодером. Эта модель показывает лучшее качество на открытых датасетах и будет использоваться в качестве базовой.

  В OrigamiNet (4) статье предложена другая архитектура для распознавания. Модель представляет полностью сверточную сеть, которая концептуально представляется как комбинация из двух частей – сверточный backbone и модуль OrigamiNet. OrigamiNet модуль получает на вход внутреннее представление изображения, трехмерный тензор, и при помощи последовательных операций вертикального расширения, горизонтального сужения и сверток приводит это представление к вертикальной цепочке вероятностей символов. Горизонтальная размерность схлопывается, а вертикальная вытягивается в высоту всего возможного числа символов. Далее эта цепочка обрабатывается с CTC. OrigamiNet обучается напрямую на страницах и не требует дополнительной разметки. Однако эта модель применима только к изображениям заданных размеров- чтобы после всех сверточных преобразований получилось схлопнуть горизонтальную размерность.

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 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] MohamedYousef, 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