СЕГМЕНТАЦИИ ТЕКСТА ДОРЕВОЛЮЦИОННЫХ ГАЗЕТ С ПРИМЕНЕНИЕМ НЕЙРОСЕТЕВОЙ МОДЕЛИ YOLO

Ломтев И.К. [1], *Кайда А.Ю.* [2]
1 ТПУ, ИШИТР, гр. 8К13, e-mail: ikl4@tpu.ru
2 ТПУ, ОИТ, ст. преподаватель, e-mail: ayk13@tpu.ru

Аннотация

Исследование направлено на поиск инструмента, основанного на нейронных, сетях, для сегментации текста газет.

Ключевые слова: нейронные сети, компьютерное зрение, NLP

Введение

Сегментация текста на изображениях является важной задачей в области компьютерного зрения и обработки изображений. Особенно важно решение этой задачи для газет и журналов, где текст часто представлен в различных форматах и стилях. Сегментация текста на изображениях является важным этапом для различных приложений, таких как оптическое распознавание символов (OCR), анализ структуры документа, извлечение информации и другие. Однако, сегментация текста на изображениях газет представляет собой сложную задачу из-за разнообразия шрифтов, размеров текста, стилей и сложной структуры страниц.

Существующие методы сегментации текста на газетах часто сталкиваются с проблемами, такими как недостаточная точность и непригодность для обработки разнообразных и сложных изображений. В этой статье рассматривается проблема автоматической сегментации текста на страницах газет с использованием технологий машинного зрения.

Использование стандартных инструментов

Начальным этапом исследования подходов к сегментации текста было рассмотрение стандартных инструментов машинного зрения, таких как OpenCV и Tesseract. Эти библиотеки предлагают функционал для обнаружения и выделения контуров текста на изображениях. Однако при использовании этих инструментов были обнаружено, что полученные результаты не полностью удовлетворяют требованиям. В связи с этим возникла необходимость искать более эффективные и точные методы сегментации текста.

```python
import cv2
import pytesseract
from PIL import Image, ImageDraw
```

```python
image = cv2.imread(r"datasets\bw-1.png")

# Применение порогового преобразования для выделения текста
_, thresh = cv2.threshold(image, 150, 255, cv2.THRESH_BINARY_INV)

# Применение OCR с помощью Tesseract с получением ограничивающих рамок
boxes = pytesseract.image_to_boxes(thresh)

# Отображение ограничивающих рамок на изображении
for b in boxes.splitlines():
    b = b.split()
    # Отрисовка прямоугольника вокруг каждой буквы/слова
    x, y, w, h = int(b[1]), int(b[2]), int(b[3]), int(b[4])
    cv2.rectangle(image, (x, y), (w, h), (255, 0, 0), 2)

# Отображение изображения с ограничивающими рамками
image_pil = Image.fromarray(image)
image_pil.show()
```

Рис. 1. Код для выделения области текста

Большинство стандартных инструментов машинного зрения обладают ограниченной способностью к распознаванию текста на изображениях, особенно при работе с текстом в сложных условиях освещения, смазывания или размытия. Это может привести к ошибкам в выделении контуров текста и неправильному определению границ слов или фраз.

Помимо этого, традиционные методы сегментации текста могут оказаться неэффективными при работе с текстом на старых или нестандартных шрифтах, а также при наличии шума или других искажений на изображении.

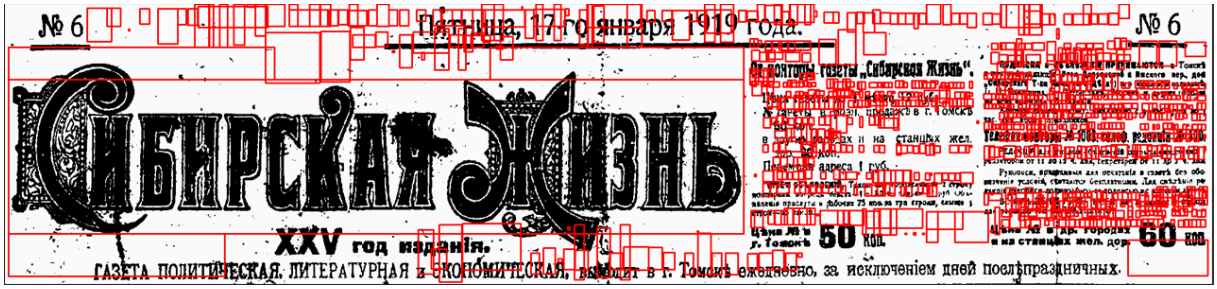

Рис. 2. Текст с выделенными сегментами

В результате было принято решение о необходимости исследования более продвинутых и точных методов сегментации текста, которые могут обеспечить более надежное и точное выделение текста на изображениях, учитывая разнообразные условия и характеристики текста.

Поиск существующих решений

В ходе решения этой проблемы был обнаружен на GitHub репозиторий, содержащий инструмент, способный выделять границы текста на страницах газет XIX-XX веков. Описанный инструмент обладает способностью адаптироваться к особенностям структуры и форматирования текста, характерным для газетных изданий того времени. Это включает в себя учет различных шрифтов, размеров текста, форматов колонок и изображений, а также прочих факторов, характерных для печатных материалов конкретной эпохи.

Использование такого инструмента может существенно упростить процесс анализа и извлечения информации из дореволюционных газет, что, в свою очередь, способствует расширению возможностей исследования в области истории, культурного наследия и социальных исследований.

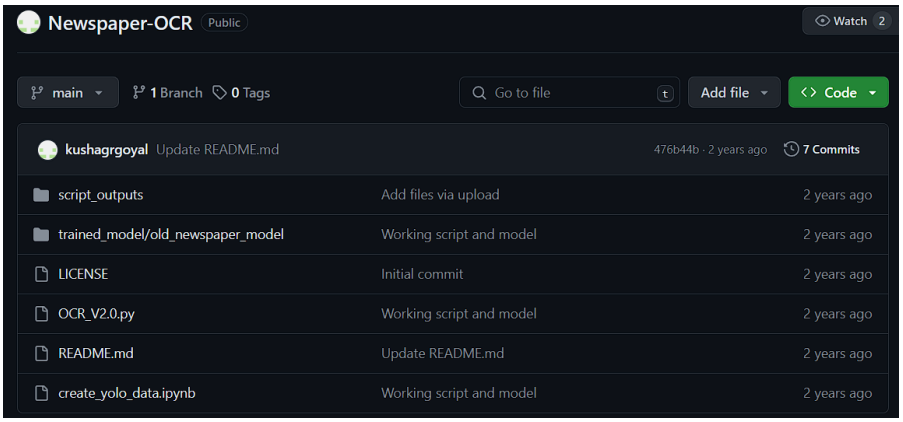

Рис. 3. Страницы репозитория на GitHub

Данный подход основан на применении модели YOLOv5L, которая является сверточной нейронной сетью, предназначенной для обработки матриц изображений. Автор данного инструмента выбрал модель YOLOv5L из-за её простоты в обучении и высокой эффективности. В ходе обучения модели использовался датасет "Newspaper Navigator", содержащий множество изображений газет с размеченными контурами текста и их координатами. Этот датасет обеспечил модель достаточным объемом данных для обучения и значительно улучшил её способность к выявлению текста на изображениях газет.

Для работы с моделью использовалась библиотека PyTorch, которая является одной из самых известных и широко используемых библиотек машинного и глубокого обучения с открытым исходным кодом.

Для улучшения производительности модели также применяются техники предварительной обработки изображений, такие как аугментация данных, нормализация и фильтрация шума. Это помогает улучшить точность и обобщающую способность модели на разнообразных данных.

Результатом этого подхода является инструмент, способный эффективно извлекать текст из газетных изображений, что облегчает дальнейшую обработку и анализ текстовой информации. Это имеет важное значение для исследования исторических данных и культурного наследия, а также для решения практических задач, связанных с архивированием и доступом к информации из старых газет.

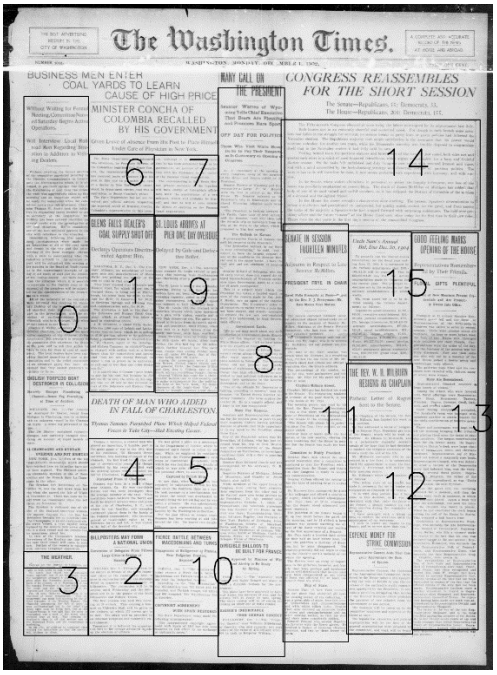

Рис. 4. Пример работы модели

Заключение

В ходе исследования был обнаружен инструмент, способный высокоэффективно сегментировать текст на изображениях газет, что в значительной степени упрощает последующую обработку и анализ текстовой информации, содержащейся на страницах печатных изданий. Это открывает новые перспективы в области автоматизации анализа текста и может стать важным инструментом для библиотек, архивов и исследовательских проектов.

Этот инструмент представляет собой значимое научное достижение, поскольку решает проблему извлечения информации из исторических газет, которые часто содержат ценные данные для исследований в различных областях, включая историю, культуру и социальные науки. Применение данного инструмента может существенно ускорить процесс анализа и обработки огромного объема текстовых данных, что позволяет исследователям и архивариусам сосредоточиться на более глубоком анализе и интерпретации результатов, а также облегчает доступ к информации для широкого круга пользователей.

Таким образом, этот инструмент имеет потенциал стать ценным инструментом в области обработки текста, способствуя улучшению эффективности исследований и расширению возможностей анализа исторических текстовых материалов.

Список литературы

1. Репозиторий kushagrgoyal // github.com [Электронный ресурс]. – URL: https://github.com/kushagrgoyal/Newspaper-OCR/tree/main (дата обращения: 25.11.2023)

2. Васильев Ю.В. «Обработка естественного языка. Python и spaCy на практике» // Питер, – 2021. – С. 22–100. (дата обращения 19.11.2023)

3. Vinotheni Coumar. Deep Learning-Based Text Segmentation in NLP Using Fast Recurrent Neural Network with Bi-LSTM // researchgate.net [Электронный ресурс] – URL: https://www.researchgate.net/publication/355348611_Deep_Learning_Based_Text_Segmentation_in_NLP_Using_Fast_Recurrent_Neural_Network_with_Bi-LSTM(дата обращения: 05.11.2023)

4. Брайан Макмахан, Делип Рао "Знакомство с PyTorch: глубокое обучение при обработке естественного языка" // Справочные издания. – Питер, – 2020. – С. 47–203. (дата обращения 25.10.2023)

5. Репозиторий RepML // github.com [Электронный ресурс]. – URL: mibo8/RepML https://github.com/mibo8/RepML?ysclid=lu8chsdkm6535385277 (github.com)