УДК 004.9, 519.6

ПРИМЕНЕНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ В ЗАДАЧАХ СЕГМЕНТАЦИИ ТЕКСТОВЫХ ИЗОБРАЖЕНИЙ

Бурикова Анна Георгиевна[1], Ершов Николай Михайлович[2]

[1]Студент;
Московский государственный университет им. М. В. Ломоносова;
Россия, 119991, г. Москва, ул. Ленинские горы, 1;
e-mail: anya.burikova@yandex.ru.

[2]Старший научный сотрудник;
Московский государственный университет им. М. В. Ломоносова;
Россия, 119991, г. Москва, ул. Ленинские горы, 1;
e-mail: ershovnm@gmail.ru.

Работа посвящена решению задачи сегментации текстовых изображений, целью которой является выделение на изображении документа текстовых блоков, соответствующих колонкам, заголовкам, колонтитулам и т.д. Проводится обзор существующих методов сегментации изображений, в том числе предназначенных и для поиска и выделения на изображениях текстовых блоков. Анализируются как классические методы, так и методы, основанные на использовании искусственных нейронных сетей. Для решения поставленной задачи предлагается подход на основе свёрточных нейронных сетей и модели U-Net. Описывается метод автоматической генерации обучающих примеров для обучения нейронной сети. Рассматривается процессы настройки модели, её обучения и тестирования. Приводятся результаты численного исследования обученных моделей на реальных данных.

Ключевые слова: сегментация изображений, распознавание образов, глубокое обучение, свёрточные нейронные сети, модель U-Net.

Для цитирования:
Бурикова А. Г., Ершов Н. М. Применение методов глубокого обучения в задачах сегментации текстовых изображений // Системный анализ в науке и образовании: сетевое научное издание. 2024. № 2. С. 39-46. EDN: HEARBF. URL : https://sanse.ru/index.php/sanse/article/view/618.

APPLICATION OF DEEP LEARNING METHODS IN THE PROBLEMS OF TEXT IMAGE SEGMENTATION

Burikova Anna G.[1], Ershov Nikolay M.[2]

[1]Student;
Lomonosov Moscow State University;
1 Leninskiye Gory, Moscow, 119991, Russia;
e-mail: anya.burikova@yandex.ru.

[2]Senior Researcher;
Lomonosov Moscow State University;
1 Leninskiye Gory, Moscow, 119991, Russia;
e-mail: ershovnm@gmail.ru.

The paper is devoted to solving the problem of text image segmentation, the purpose of which is to select text blocks in the document image that correspond to columns, headers, footers etc. A review of existing image segmentation methods is carried out, including those intended for searching and selecting text blocks in images. Both classical methods and methods based on the use of artificial neural networks are analyzed.

To solve given problem, an approach based on convolutional neural networks and the U-Net model is proposed. A method for automatically generating training examples for training a neural network is described. The processes of setting up a model, training and testing it are considered. The results of a numerical study of trained models on real data are presented.

Keywords: image segmentation, pattern recognition, deep learning, convolutional neural networks, U-Net model.

For citation:
Burikova A. G., Ershov N. M. Application of deep learning methods in the problems of text image segmentation. *System analysis in science and education*, 2024;(2):39-46 (in Russ). EDN: HEARBF. Available from: https://sanse.ru/index.php/sanse/article/view/618.

Введение

Настоящая работа посвящена разработке нейросетевых методов для поиска и выделения текстовых блоков на изображениях отсканированных текстовых документов. Решение такой задачи является необходимым этапом предобработки изображений текстовых документов для их последующего распознавания (*OCR*, оптическое распознавание символов). Данная задача является частным случаем более общей задачи сегментации изображений на отдельные части [1], соответствующие различным объектам, из которых и состоит анализируемое изображение. Оцифровка отсканированных текстовых документов, т. е. распознавание их изображений с последующим преобразованием в текстовый формат, все ещё остаётся актуальной задачей, особенно для работы с историческими документами. Оцифровка таких документов – это путь к сохранению исторического и культурного наследия, а так же возможность быстрого и удобного доступа к обработанным материалам.

Отличительной особенностью рассматриваемой в работе задачи сегментации текстовых блоков оказывается то, что такие блоки характеризуются не столько формой, сколько своей внутренней текстурой, что делает данную задачу схожей с задачами сегментации медицинских изображений [2]. Поэтому следует ожидать, что применение методов анализа медицинских изображений может оказаться эффективным и при сегментации текстовых изображений.

1. Обзор методов сегментации изображений

Сегментация – это процесс разделения изображения на несколько множеств пикселей (сегментов), то есть присвоения таких меток каждому пикселю, что пиксели с одинаковыми метками имеют общие визуальные характеристики, например, принадлежат изображению одного и того же объекта. К классическим методам сегментации изображений относятся: пороговые методы [3]; методы, основанные на кластеризации [4]; методы разрезания графов [5]; методы выделения границ [6] и т. д. Применение классических методов для сегментации текстовых блоков, однако, выглядит мало перспективным, так как практически все эти методы направлены прежде всего на поиск и выделение структурной информации на изображениях, т. е. в контексте рассматриваемой задачи могут применяться, например, для выделения отдельных символов текста. Текстовые блоки имеют, вообще говоря, слабо выраженную структуру с размытыми границами и характеризуются больше своей текстурой, что позволяет нам легко визуально выделять такого рода блоки даже на документах на незнакомых нам языках.

Современным и более универсальным подходом к решению задачи сегментации изображений является применение нейросетевых методов с использованием свёрточных нейронных сетей. Например, нейронная сеть *SegNet* является автокодировщиком [7], то есть имеет энкодер (уменьшающий размер изображения) и декодер (восстанавливающий исходный размер и выдающий карту сегментации). Архитектура этой сети представлена на рис. 1. Сеть состоит из блоков, в каждом блоке присутствуют свёртки и пулинги – слои, понижающие размер изображения и апсэмплинг слои, повышающие размер, а также активационные слои *ReLU* и слои нормализации *BatchNorm*. Архитектура полностью симметрична, кроме слоя *Softmax* на конце декодера. Этот слой выдаёт вероятность принадлежности каждому классу каждого пикселя. Главное отличие *SegNet* от обычного свёрточного автокодировщика в том, что апсэмплинг слои его декодера информационно соединены с соответствующими пулинг слоями энкодера. То есть, апсэмплинг слои сети не обучаются, а получают необходимую ин-

формацию о том, как повысить размерность и как восстановить сжатую (утерянную) информацию от соответствующих пулинг слоёв, которые хранят индексы пикселей с наибольшим значением в окне.

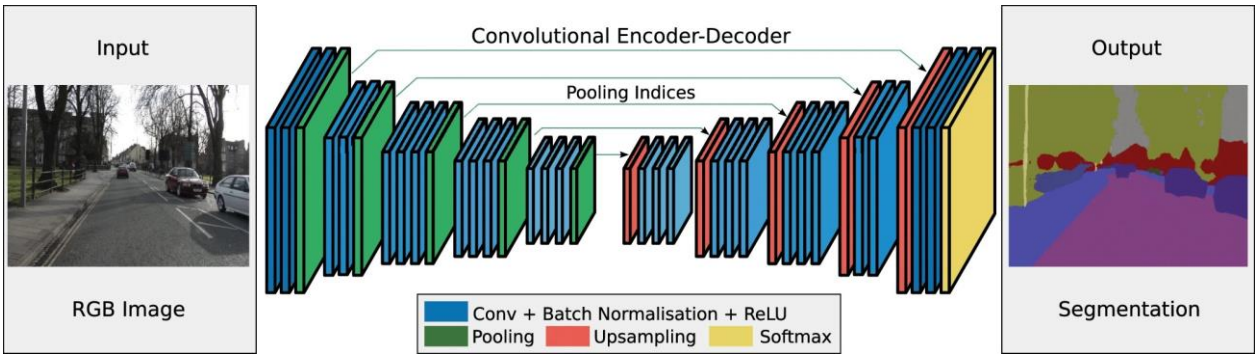

Рис. 1. Архитектура нейронной сети SegNet [7]

Вторая нейросетевая модель, популярная в области сегментации изображений — модель *UNet*, описанная и реализованная в 2015 году [2] и изначально предназначенная для сегментации медицинских изображений. Архитектура сети показана на рис. 2. Она состоит из энкодера (слева) и декодера (справа). Каждый шаг энкодера содержит два свёрточных слоя 3×3, за каждым из которых идёт слой *ReLU*, и после макс-пуллинг 2×2 с шагом 2. Каждый шаг декодера содержит обратный пуллинг, который расширяет карту признаков, после которого следует свёртка 2×2 (эти две операции вместе — *up-convolution*), которая уменьшает количество каналов. После идёт конкатенация с соответствующей картой признаков из энкодера. Это повышает качество модели, так как *up-convolution* плохо восстанавливает информацию, а конкатенация со слоем из энкодера помогает сети восстановить общую пространственную информацию. И две свёртки 3×3, за каждой из которой идет *ReLU*. На последнем слое свёртка 1×1 используется для приведения каждого 64-компонентного вектора признаков до требуемого количества классов.

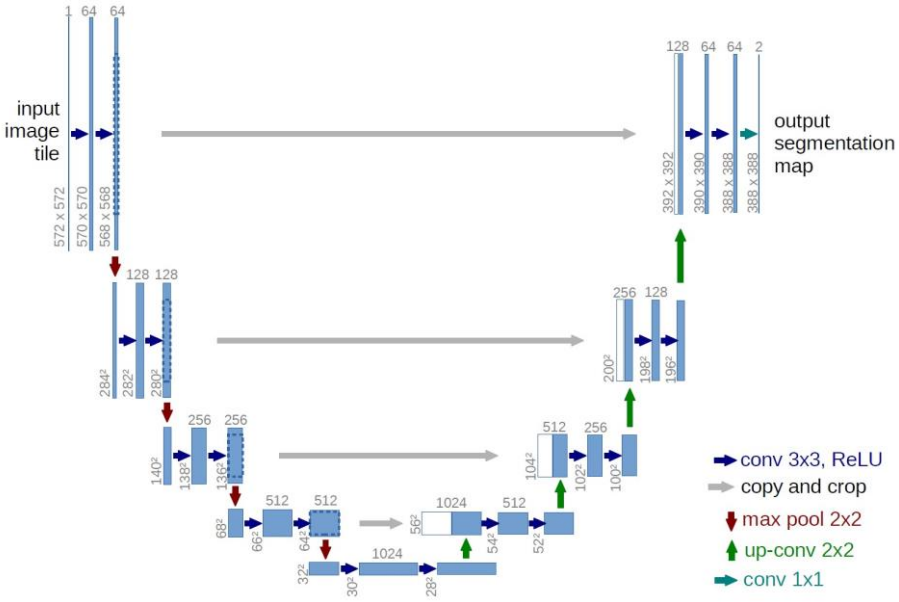

Рис. 2. Архитектура нейронной сети U-Net [2]

В настоящее время также доступно некоторое количество готовых программных продуктов, предназначенных для выделения текстовых блоков на отсканированных изображениях документов: *Google Cloud Vision API* [8] – инструмент для различных задач компьютерного зрения, который использует модель, предварительно обученную компанией *Google*; *CCS Layout Analysis Benchmark* [9] – средство анализа структуры газет; *Book Scan Processing Print Press Edition* [10] – распределенная система для обработки сканов газетных страниц; *Newspaper Navigator* [11] – проект для извлечения ви-

зуального контента с отсканированных документов с открытым кодом. К сожалению, практически все представленные продукты являются платными и не имеют описания методов, лежащих в основе их работы. Таким образом, создание собственного решения с открытым кодом является актуальной исследовательской и практической задачей.

2. Подготовка обучающего набора данных

Целью настоящей работы является применение нейросетевых моделей для решения задачи сегментации текстовых блоков. Для настройки таких моделей требуется наличие обучающего набора данных (примеров). В нашем случае обучающим примером будет служить пара изображений одинакового размера, первое из них является изображением исходного текстового документа, второе — результатом сегментации (маской), на котором белым цветом на чёрном фоне выделены искомые текстовые блоки. Пример такой пары показан на рис. 3.

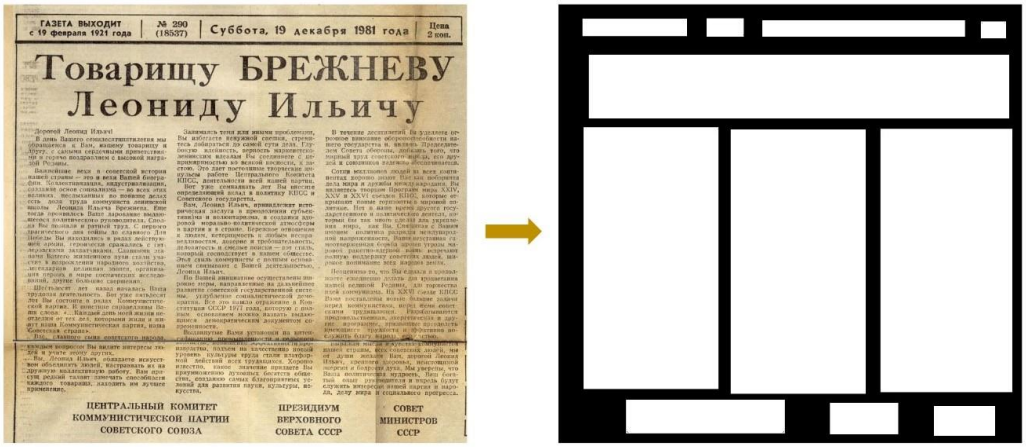

Рис. 3. Пример входного изображения и выходной маски

Для этого в Интернете были найдены *RGB*-изображения газетных страниц (сканы или фотографии газет преимущественно XX века, содержащие колонки текста, картинки, заголовки и т.д.), и для каждого изображения с помощью графического редактора создавалась искомая черно-белая маска. Таким образом был создан набор из 20 пар таких изображений и масок для них. Однако, создание даже такого небольшого набора данных, которого не хватило бы для обучения нейросетевого решения, занимает очень много времени. Поэтому появилась необходимость создания средства для автоматической генерации данных.

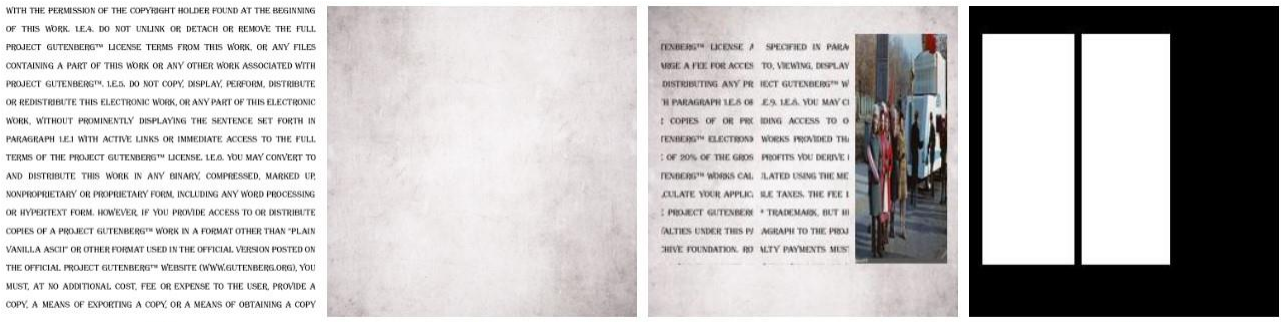

Рис. 4. Пример входного изображения и соответствующей ему выходной маски

Для создания автоматической генерации обучающих примеров были взяты: тексты (из генератора случайного текста или литературных источников), из которых были созданы изображения, полностью заполненные данными текстами различных шрифтов и размеров (рис. 4, а); фон, с текстурой газетной бумаги (рис. 4, б) и различные картинки, представляющие иллюстрации. Был написан скрипт на языке *Python*, который вырезал текстовые блоки прямоугольной формы из изображений с текстом и накладывал их на фоновое изображение с бумажной текстурой. Дополнительно на тот же

фон накладывались и подготовленные картинки. В результате получались входные для модели изображения (рис. 4, в). Одновременно с этим создавались и целевые маски (рис. 4, г).

3. Выбор, настройка и обучение моделей

Для решения задачи было решено реализовать, настроить и исследовать три нейросетевые модели: *SegNet, UNet* и *UNet*, предобученную для сегментации аномалий на наборе данных МРТ головного мозга [12]. Выбор нейронной сети *UNet*, а также такой её предобученной версии обусловлен схожестью задач сегментации текстовых и некоторых медицинских изображений. В обоих случаях надо находить разные по тексту объекты — фон и текст; межклеточное пространство и клетки; мягкую ткань и уплотнения.

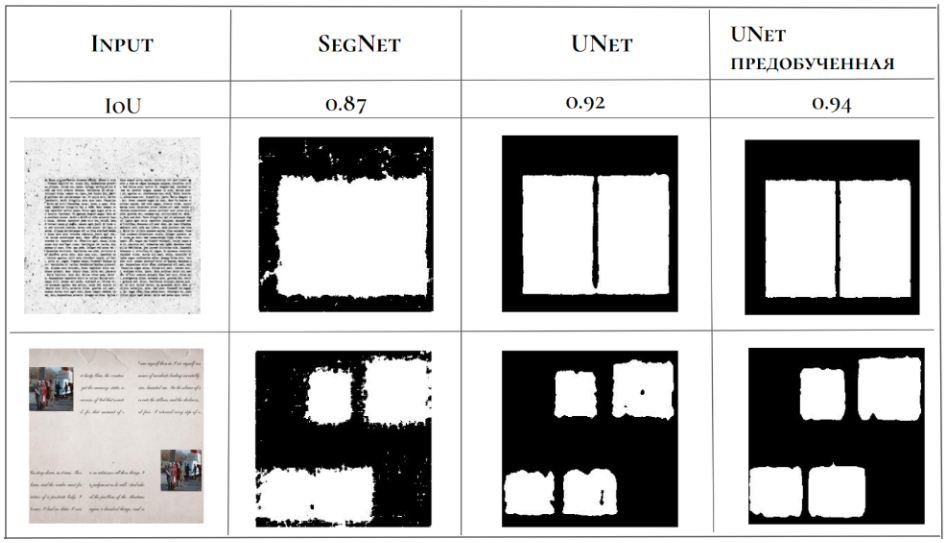

Рис. 5. Сравнение результатов обучения трёх выбранных моделей

Для обучения моделей был сгенерирован обучающий набор из 250 входных изображений: 200 в тренировочной части, 25 в валидационной и 25 в тестовой. Изображения имели размер 256 на 256 пикселей. Все модели обучались с использованием фпеймворка *pytorch* в *Google Colab* на *T4 GPU*. Для обучения использовались: функция потерь — бинарная кросс-энтропия, метрика – *Intersection over Union* (*IoU*).

Результаты работы трёх моделей приведены на рис. 5. Модель *SegNet* выделяет очертания текстовых блоков и не выделяет картинки. Границы блоков выделены очень нечетко, а тонкие границы между блоками в большинстве случаев не выделяются, то есть несколько блоков (колонок) сливаются в один. Модель *UNet*, обученная нуля, выделяет эти границы между блоками и практически точно сегментирует изображение. Её качество внешне практически такое же, как предобученной *UNet*, но у предобученной модели в ряде примеров можно заметить более ровные границы и менее частое слияние границ между блоками. По качеству на тестовой выборке лучше всего работает предобученная *UNet*, поэтому она и была выбрана для дальнейшей более точной настройки.

Рис. 6. Примеры работы предобученной модели UNet на реальных данных

Тестирование третьей модели (предобученная *UNet*) на изображения реальных газет по метрике *IoU* составило 0.60. На рис. 6 представлены примеры изображений и масок, которые выдала сеть. Видим, что на изображениях, где блоки крупные и их немного, получились хорошие маски. На изображениях, где много маленьких блоков и границы между ними очень тонкие, не очень хорошо получилось отделить блоки друг от друга и выделить эти границы. Можно заметить, что так как при обучении использовалась стандартная функция потерь бинарная кросс-энтропия, то все пиксели были одинаково важны для модели. Поэтому сеть напрямую не обучалась на выделение тонких границ между блоками. Поэтому, было решено модифицировать функцию потерь, добавив каждому пикселю вес: у всех пикселей, кроме граничных пикселей блоков, будет вес 1, а у граничных — вес 2. Такой приём должен заставить модель более внимательно относиться к распознаванию границ между блоками.

Рис. 7. Входные изображения и выходные маски модели для взвешенной кросс-энтропии

С использованием модифицированной функции потерь была выполнена повторная настройка предобученной модели *Unet*. На сгенерированном обучающем наборе качество по метрике *IoU* выросло до 0.96. На реальных изображениях метрика увеличилась до величины 0.635. На рис. 7 приведены примеры выходных изображений и масок, на выходе этой модели. Видим, что границы стали более чёткими, как на изображениях с небольшим количеством крупных блоков, так и с большим количеством мелких блоков.

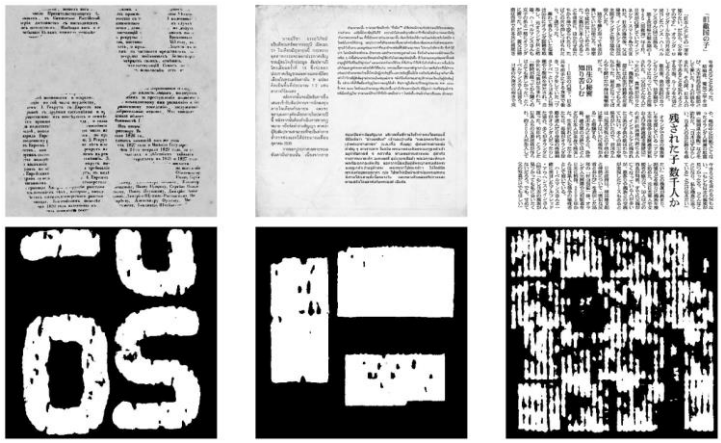

Рис. 8. Обработка изображений с нестандартными текстовыми блоками и языками

Для исследования работы модели с различными языками и блоками нестандартной формы были сгенерированы соответствующие изображения, а затем пропущены через нейронную сеть. Качество на семи изображениях по метрике *IoU* составило 0.75. Пример входных изображений и масок на выходе из нейросети представлены на рис. 8. Видим, что модель может выделять блоки вне зависимости от языка и формы текстовых блоков.

4. Программная реализация

Для того чтобы пользователям было удобно использовать разработанный подход к выделению текстовых блоков, было разработано приложение с графическим интерфейсом с помощью библиотеки *Tkinter* на языке программирования *Python*. Интерфейс приложения представлен на рис. 9. Пользователю необходимо ввести путь к изображению или путь к папке для обработки нескольких изображений. Изображения должны быть представлены в форматах *JPG* или *PNG*. После обработки в папке с исходными изображениями появится папка *processed*, в которой для каждого входного изображения сохранится выходная маска, входное изображение с построенными на нём ограничивающими прямоугольниками (*bounding boxes*) и координаты этих прямоугольников в текстовом файле. Также после обработки на экране приложения появятся изображения, которые были обработаны. Изображения можно листать с помощью кнопки *Next*. Код приложения со всеми необходимыми для его работы файлами доступен на сервисе *GitHub* [13].

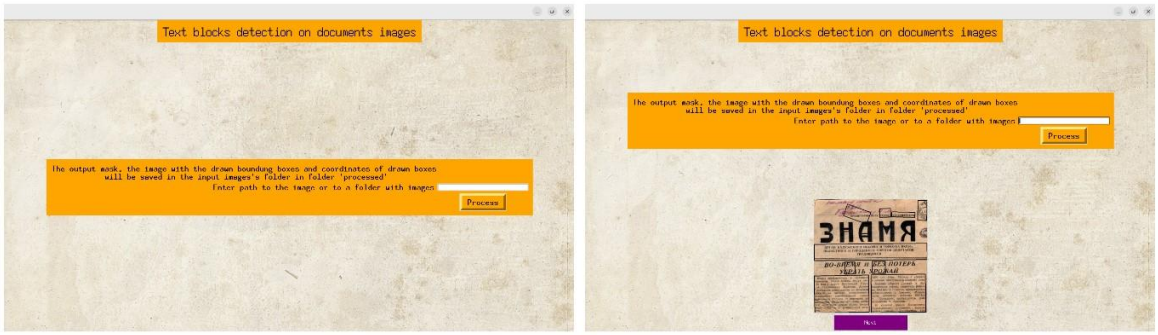

Рис. 9. Графический интерфейс приложения для выделения текстовых блоков

Заключение

В настоящей работе был рассмотрен нейросетевой подход к решению задачи сегментации текстовых блоков на изображениях текстовых документов. Выполнен обзор различных методов общей задачи сегментации изображений, рассмотрены как классические, так и современные нейросетевые подходы решения данной задачи. Были рассмотрены существующие программные продукты для выделения текстовых блоков на изображениях документов и газет. В большинстве случаев разработчи-

ки продуктов не описывают деталей реализации, кроме того, практически все эти программы являются платными в использовании.

Для применения подхода с использованием нейронных сетей к решению задачи сегментации текстовых блоков был создан набор обучающих данных на основе размеченных изображений реальных газет. Также была разработана и программно реализована автоматическая генерация обучающих данных и с её помощью сгенерирован набор размеченных искусственных текстовых изображений.

Были построены и обучены на сгенерированных данных три нейросетевых модели — сети *SegNet* и *UNet*, а также предобученная на медицинских изображениях сеть *UNet*. Проведённое тестирование моделей показало, что лучшей для решения поставленной задачи является предобученная модель *UNet*. Для улучшения выделения границ блоков и тонких границ между блоками предложена собственная взвешенная функция потерь на основе стандартной бинарной кросс-энтропии. Обученная модель была протестирована на изображениях реальных газет, рукописных документов, а также на изображениях с текстом на различных языках и с блоками различной формы. Для удобства использования разработанного решения задачи выделения текстовых блоков с текстовых изображений, было разработано приложение с графическим интерфейсом с помощью библиотеки *Tkinter* на языке программирования *Python*.

Список источников

1. Shapiro L. G. Computer Vision / L. G. Shapiro, G. C. Stockman. Prentice Hall, 2001.
2. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015. Vol. 935. P. 234–241. DOI: 10.1007/978-3-319-24574-4_28.
3. Sezgin M., Sankur B. Survey over image thresholding techniques and quantitative performance evaluation // Journal of Electronic Imaging. 2004. Vol. 13 (1). P. 146-168. DOI: 10.1117/1.1631315.
4. Comaniciu D., Meer P. Mean Shift: A Robust Approach Toward Feature Space Analysis // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. Vol. 24, No. 5. P. 603–619. DOI: 10.1109/34.1000236.
5. Shi Jianbo, Malik J. Normalized Cuts and Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000. Vol. 22, No. 8. P. 888–905. DOI: 10.1109/34.868688
6. Barghout L. Visual Taxometric approach Image Segmentation using Fuzzy-Spatial Taxon Cut Yields Contextually Relevant Regions // Communications in Computer and Information Science (CCIS). Springer-Verlag. 2014.
7. Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017. Vol. 39, No. 12. P. 2481-2495. DOI: 10.1109/TPAMI.2016.2644615.
8. Vision AI: Image & Visual AI Tools | Google Cloud. URL: https://cloud.google.com/vision (дата обращения: 10.06.2024).
9. docWizz | CCS. CCS Content Conversion Specialists Gmb, [2024]. URL: https://content-conversion.com/software/docwizz/ (дата обращения: 10.06.2024).
10. Book Scan Processing Print Press Edition | АЛАНИС Софтвер. URL: https://alanissoftware.wordpress.com/bsp-ppe-book-scan-processing-print-press-edition/ (дата обращения: 10.06.2024).
11. The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America / B. Lee, J. Mears, E. Jakeway [et al.] // arXive.org e-Print archive. DOI: 10.48550/arXiv.2005.01583.
12. U-NET for brain MRI | PyTorch. The Linux Foundation, [2024]. URL: https://pytorch.org/hub/mateuszbuda_brain-segmentation-pytorch_unet/ (дата обращения: 10.06.2024).
13. GitHub - AnnaBurikova / TextBlocksDetection. GitHub, Inc., 2024. URL: https://github.com/AnnaBurikova/TextBlocksDetection (дата обращения: 10.06.2024).