DOI: 10.15514/ISPRAS-2022-34(6)-10

Автоматическая разметка данных для сегментации изображений документов с использованием глубоких нейронных сетей

А.А. Михайлов, ORCID: 0000-0003-4057-4511<mikhailov@icc.ru>

Институт динамики систем и теории управления имени В.М. Матросова СО РАН,
664033, Россия, Иркутск, ул. Лермонтова, 134
Институт системного программирования им. В.П. Иванникова РАН,
109004, Россия, г. Москва, ул. А. Солженицына, д. 25

Аннотация. В статье предложен новый метод автоматической аннотации данных для решения задачи сегментации изображений документов с помощью глубоких нейронных сетей обнаружения объектов. В качестве исходных данных для разметки рассматривается формат помеченных файлов PDF. Особенность данного формата заключается в том, что он включает в себя скрытые метки, которые описывают логическую и физическую структуру документа. Для их извлечения разработано инструментальное средство, которое имитирует работу стековой машины вывода на печать согласно спецификации формата PDF. Для каждой страницы документа генерируются изображение, и аннотация в формате PASCAL VOC. Классы и координаты ограничивающих рамок вычисляются в процессе интерпретации помеченного PDF файла на основе меток. Для апробации метода была сформирована коллекция размеченных PDF файлов из которой в автоматическом режиме получены изображения страниц документов и аннотации для трех классов сегментации. На основе этих данных обучена нейронная сеть архитектуры EfficientDet D2. Произведено тестирование модели на данных из того же домена, размеченных вручную, которое подтвердило эффективность применения автоматически сгенерированных данных для решения прикладных задач.

Ключевые слова: сегментация документов; сегментация изображений документов; глубокие нейронные сети; обнаружение объектов

Для цитирования: Михайлов А.А. Автоматическая разметка данных для сегментации изображений документов с использованием глубоких нейронных сетей. Труды ИСП РАН, том 34, вып. 6, 2022 г., стр. 137-146. DOI: 10.15514/ISPRAS-2022-34(6)-10

Automatic data labeling for document image segmentation using deep neural networks

A.A. Mikhailov, ORCID: 0000-0003-4057-4511 <mikhailov@icc.ru>

Matrosov Institute for System Dynamics and Control Theory of the SB RAS
134, Lermontova st., Irkutsk, 664033, Russia.
Ivannikov Institute for System Programming of the RAS,
25 Alexander Solzhenitsyn st., Moscow, 109004, Russia

Abstract. The article proposes a new method for automatic data annotation for solving the problem of document image segmentation using deep object detection neural networks. The format of marked PDF files is considered as the initial data for markup. The peculiarity of this format is that it includes hidden marks that describe the logical and physical structure of the document. To extract them, a tool has been developed that simulates the operation of a stack-based printing machine according to the PDF format specification. For each page of the document, an image and annotation are generated in PASCAL VOC format. The classes and

coordinates of the bounding boxes are calculated during the interpretation of the labeled PDF file based on the labels. To test the method, a collection of marked up PDF files was formed from which images of document pages and annotations for three segmentation classes (text, table, figure) were automatically obtained. Based on these data, a neural network of the EfficientDet D2 architecture was trained. The model was tested on manually labeled data from the same domain, which confirmed the effectiveness of using automatically generated data for solving applied problems.

Keywords: document layout analysis; PDF accessibility; ANN models; artificial intelligence

For citation: Mikhailov A.A. Automatic data labeling for document image segmentation using deep neural networks. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 6, 2022. pp. 137-146 (in Russian). DOI: 10.15514/ISPRAS-2022-34(6)-10

1. Введение

Произвольные документы являются распространённым способом представления информации. Они повсеместно распространены в веб-пространстве. Большой объем и свойства структуры таких документов делают их ценным источником в приложениях науки о данных и бизнес аналитики. Однако, как правило, они не сопровождаются явной семантикой необходимой для машинной интерпретации своего содержания так, как задумано их автором. Накапливаемая в них информация часто является неструктурированной и не стандартизированной. Анализ этих данных нуждается в их предварительной трансформации к структурированному представлению с заданной формальной моделью.

В литературе по анализу и распознаванию документов такую задачу принято называть анализом компоновки документов (Document layout analysis). В последние годы наряду с классическими методами анализа компоновки документов активно развиваются подходы на основе глубоких нейронных сетей обнаружения и классификации объектов на изображениях. О чем свидетельствуют результаты одной из ведущих научных конференций по анализу документов – ICDAR (International Conference on Document Analysis and Recognition). Методы на основе глубоких нейронных сетей являются эффективными и демонстрируют высокие результаты, но они очень требовательны к размеру и качеству обучающей выборки. Разметка изображений для обучения очень медленный, трудоемкий и дорогостоящий процесс. Одним из методов решения данной проблемы является автоматическая разметка данных.

Таким образом, целью данной работы является разработка метода автоматической разметки данных для обучения глубоких нейронных сетей сегментации изображений документов.

В статье предложен оригинальный метод автоматической аннотации данных из помеченных PDF файлов. Собрана коллекция документов, из которых сгенерирована обучающая выборка емкостью 28 000 изображений страниц, охватывающая типичные структурные элементы для документа: текст, таблица, рисунок. Обучена нейронная сеть архитектуры EfficientDet D2 на полученном наборе данных. Тестирование полученной модели показало, что автоматически сформированных набор данных подходит для решения задачи сегментации изображений документов.

2. Современное состояние исследований

Изображения документов часто генерируются из физических документов путем оцифровки, с использованием сканеров или различных программ генерации (принтеров). Многие документы, такие как газеты, журналы и брошюры, содержат очень сложную компоновку из-за размещения рисунков, заголовков и подписей, сложного фона, художественного форматирования текста и т. д. Человек использует множество дополнительных подсказок, таких как контекст, условные обозначения, информацию о языке. Автоматический анализ произвольного документа со сложной верстkой является чрезвычайно сложной задачей и выходит за рамки возможностей современных систем анализа компоновки документов. В научной литературе предложено большое количество методов анализа компоновки

документов. Согласно статье [1] они могут быть разделены на три группы: методы классификации на основе областей [2, 3]; методы классификации на основе анализа пикселей [4, 5]; анализ связных компонент [6-8].

С ростом эффективности и популярности свёрточных нейронных сетей область их применения постоянно расширяется. Начиная с 2014 года известны первые попытки использования искусственных нейронных сетей для решения задачи анализа компоновки документов [9-12]. Эти работы продемонстрировали свою эффективность по сравнению с классическими подходами, что подтверждается результатами соревнования 2017 года на конференции ICDAR [13]. С другой стороны, соревнования 2019 года показало, что на разнообразных данных, с большим количеством классов (10) комбинирование классических методов [14] наиболее эффективно по сравнению с глубокими нейронными сетями. Это обусловлено отсутствием достаточного количества разнообразных размеченных данных с большим количеством классов. В то время как для частных случаев нейронные сети работают намного эффективнее [15]. Следует отметить, что для решения задачи анализа компоновки документов в этих работах используются либо нейронные сети архитектуры (Fast/Faster) R-CNN, либо авторские разработки. Для обучения нейронных сетей обычно используются открытые наборы размеченных данных, в редких случаях авторы статей указывают, что они разметили собственную обучающую выборку. Вручную размеченные наборы данных редко достигают размера в 20 000 экземпляров и чаще всего не публикуются в открытом доступе.

В сентябре 2021 года состоялась одна из ведущих конференция по анализу документов ICDAR2021[1]. В рамках этой конференции прошли соревнования как по сегментации изображений документов [16], так и по отдельным задачам сегментации, таких как обнаружение формул в изображениях документов [17], обнаружение списка литературы в научных статьях [18]. Большинство участников в своих решениях применяли сложные архитектурные решения, использующих комбинации нейронных сетей и моделей машинного обучения. Результаты победителей данных соревнований варьируются от 82% до 99% по точности и полноте, что позволяет использовать их для решения некоторых практических задач. Но, использование большого количества нейронных сетей накладывает ограничение на ресурсы, что может стать барьером при их внедрении. Использование таких систем на практике затруднено в данный момент и тем, что они очень сильно зависят от предметной области, для которой проводится анализ документов.

Это подтверждают работы 2022 года [19, 20] по обнаружению и извлечению таблиц из документов. Обнаружение таблиц в данном случае является подзадачей сегментации изображений документов только на один класс - таблица. Авторы этих работ для обнаружения таблиц использовали доступные в открытом доступе модели и системы, которые позволяют находить область таблицы. Результаты тестирования на данных из новой предметной области показали, что результаты по качеству и полноте сильно уступают (от 20% до 50%) результатам тестирования из оригинальных статей.

В начале 2020 – конце 2019 года появилось сразу несколько датасетов большого размера: PubTabNet[2] – 568 000 примеров из PubMed, который содержит в себе только таблицы; SciTSR[3] – 15 000 примеров из LaTeX источников, размеченных вручную; TIES-2.0[4] – автоматически сгенерированный датасет на основе набора данных UNLV, который содержит 500 000 примеров. Все эти датасеты содержат всего один класс (таблица).

В статье [1] описан подход генерации синтетических данных для тонкой настройки моделей сегментации документов, размеров в 18000 обучающих примеров. Отдельно следует

[1] https://icdar2021.org.
[2] https://github.com/ibm-aur-nlp/PubTabNet.
[3] https://github.com/Academic-Hammer/SciTSR
[4] https://github.com/shahrukhqasim/TIES-2.0

отметить набор данных PubLayNet[5], который включает в себя около 360 000 размеченных медицинских документов. Авторы этого набора использовали коллекцию документов из базы PubMed[6]. Каждый документ из этой коллекции сопровождается XML аннотацией с разметкой на 5 классов. Для формирования размеченной коллекции использовались методы нечеткого сравнения строк (расстояние Левенштейна) с последующей коррекцией ошибок.

Такой подход обладает рядом недостатков. Во-первых, из-за нечеткого сравнения строк достаточно высока вероятность ошибок разметки. Во-вторых, данный набор содержит небольшое количество классов (5 классов). Такой набор не подходит в полной мере для создания на его основе компьютерной модели. Это подтверждается в статье, опубликованной по результатам соревнования RDCL2019 [14]. В ней отмечается, что методы, основанные на нейронных сетях, сильно подвержены влиянию сравнительно небольшой обучающей выборки. Подходы на основе глубоких нейронных сетей, которые использовали сторонние наборы и специальные техники для увеличения обучающей выборки показали хорошие результаты, но не смогли составить конкуренцию решениям на основе классических методов. Одной из причин этого является чувствительность нейронных сетей к объему и разнообразию размеченных данных. В статье отмечается, что подходы на основе нейронных сетей для отдельных классов значительно превосходят классические методы (например, задача обнаружения таблиц на изображениях документов).

3. Описание метода

В сети Интернет содержится большое количество оригинальных документов LaTex. Одним из самых известных ресурсов является arXiv[7] – бесплатный электронный архив, который содержит более 1,7 миллиона научных статей и препринтов по физике, математике, астрономии, информатике, биологии, электротехнике, статистике, финансовой математике и экономике. Практически все статьи и препринты распространяются по одной из разновидностей свободной лицензии CC BY, что позволяется публиковать полученные на их основе результаты в открытом доступе. Помимо этого, в сети интернет содержится большое количество свободно распространяемых документов в формате MS Word, Open Office.

Форматы файлов MS Word, Open Office, LaTex содержат в себе всю необходимую информацию о компоновке документа, такую как заголовки, подзаголовки, таблицы, изображения, диаграммы, формулы и т.д.

Для того, что разработать методы автоматической размети данных на основе этих форматов необходимо исследовать алгоритмы процессинга приведенных форматов и модифицировать их для генерации растровых изображений и сопутствующих файлов с координатами ограничивающих прямоугольников, и классов. Это очень трудоемкий процесс, который связан с рядом трудностей. Например, для форматов MS Word и Open Office сгенерировать изображение документа и правильно извлечь координаты можно только для заранее известного драйвера принтера. В общем случае это сделать затруднительно.

Начиная с 2006 года, спецификация формата PDF включает в себя специальные метки, которые позволяют с помощью скрытых тегов описывать логическую структуру документа (листинг 1).

Теги позволяют пометить в PDF документе такие объекты как таблицы, списки и заголовки, формулы, рисунки, строки, ячейки, сноски, ссылки, параграфы и т.д. Кроме того, PDF – это универсальный формат, который позволяет отображать документы в любой операционной системе, программ е или устройстве в том виде, в котором он был создан.

[5] https://github.com/ibm-aur-nlp/PubLayNet
[6] https://pubmed.ncbi.nlm.nih.gov
[7] https://arxiv.org

```
<ispAnotation index=30>
Ключевые слова:
список ключевых слов, разделенных точкой с
запятой.
(стиль ispAnotation)
</ispAnotation>
<ispAnotation index=31>
Для цитирования:
Иванов И.И., Петров П.П. Заголовок статьи. Труды
ИСП РАН, том 1, вып. 2, 2019
г., стр. 15-19. DOI: 10.15514/ISPRAS-2019-1(2)-1
</ispAnotation>
<ispAnotation index=32>
<ispAnotation2 Знак index=33>
Благодарности:
</ispAnotation2 Знак>
<ispAnotation2 Знак index=34>
В этом блоке перечисляются организации,
поддерживающие исследование, описанное в статье,
гранты и т.д.
</ispAnotation2 Знак>
</ispAnotation>
```

Листинг 1. Результат интерпретации помеченного PDF файла
Listing 1. Tagged PDF parsing result

Основная идея метода автоматической разметки данных (рис 1) заключается в использовании помеченных PDF файлов в качестве промежуточного этапа между слабоструктурированными форматами и аннотированными данными. Такой подход гарантирует полное соответствие координат тегов в PDF с их координатами в сгенерированном изображении страниц, из-за их полной идентичности.

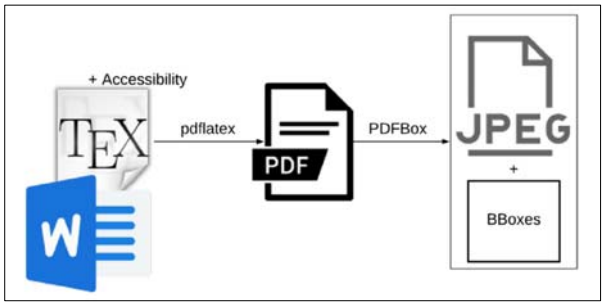

Рис. 1. Метод автоматической разметки данных для сегментации изображений документов с использованием нейронных сетей
Fig. 1. Automatic data labeling method for document image segmentation using neural networks

3.1 Формирование коллекции исходных данных

Коллекция исходных данных в количестве 1200 документов была сформирована из опубликованных в отрытом доступе технических заданий и нормативно-правовых актов[8].

[8] https://zakupki.gov.ru

3.2. Генерация помеченных PDF документов

Начиная с июля 2020 года, в основной дистрибутив LaTex был включен пакет Accessibility. Этот пакет предназначен для создания помеченных PDF документов с тегами из исходного кода LaTex. Для того, чтобы воспользоваться данной функций достаточно добавить в преамбулу исходного LaTex документа пакет Accessibility. Кроме того, современные версии Microsoft Office, Open Office, Libre Office и др. позволяют по исходному формату производить экспорт помеченных PDF (Tagged PDF) документов.

Для потоковой генерации помеченных PDF документов из файлов формата Microsoft Word использовался скрипт экспорта с помощью LibreOffice Community с включенной опцией «Tagged PDF (add document structure)». В результате было получено 1200 помеченных PDF файлов.

3.3. Описание классов сегментации

Название тегов в помеченных PDF документах, полученных из файлов Microsoft Word, задается стилевых файлом, который был использован при их создании. Из этого формируется одно из ограничений метода – необходимость сопоставить названия тегов с ограниченным набором заранее определенных классов.

В данной работе было принято решение использовать три класса:

- Text – текстовые блоки, содержащие однородный текст с единым форматированием (размером, жирностью, шрифтом, отступами между строк);
- Table – класс таблиц с границами, которые могут содержать объединенные ячейки по вертикали или по горизонтали; заголовок таблиц может отличаться от тела другим форматированием;
- Picture – класс, содержащий печати, подписи, изображения в документах.

Этот выбор обусловлен наличием тестового набора данных из того же домена с аналогичным набором классов для проведения сравнительной оценки разработанного метода.

3.4 Извлечение координат и классов

PDF - это универсальный межплатформенный формат, который позволяет отображать документы в любой операционной системе, программе или устройстве в том виде, в котором он был создан. Сам формат задается спецификацией, а его интерпретация производится с помощью стековой виртуальной машины. Фактически, содержимое PDF можно рассматривать как набор инструкций.

Все инструкции выполняются последовательно, каждая извлекает параметры со стека, производит операцию и помещает результат своей работы на стек (если это требуется) и изменяет контекст.

Для извлечения теговой информации разработано и опубликовано в открытом доступе инструментальное средство TaggedPDF[9], которое имитирует стековую машину вывода на печать согласно спецификации формата PDF. Это позволило правильным образом извлечь название класса и вычислить координаты начала и конца тега. Для разработки инструментального средства извлечения теговой информации из помеченных PDF документов использовалась библиотека Apache PDFBox. Данная библиотека предназначена для генерации и извлечения текстовой и графической информации из PDF документов на достаточном уровне для реализации поставленной задачи.

[9] https://github.com/sunveil/tagged_pdf/tree/develop

3.5. Генерация обучающей выборки

С помощью разработанного инструмента в автоматическом режиме из 1200 документов было получено 28000 изображений и аннотаций к ним в формате PASCAL VOC.

3.6. Обучение модели

Для тестирования применимости автоматически сгенерированного обучающего набора данных была выбрана нейронная сеть архитектуры EfficientDet D2 [22]. Данная модель имеет одну из лучших оценок при высокой производительности на наборе данных COCO, что стало определяющим фактором при выборе архитектуры модели.

Обучение модели производилось с помощью фреймворка Object Detection API TensorFlow 2 на двух видеокартах NVIDIA GeForce RTX 2060 Super.

Сгенерированный набор данных был разбит на обучающий (24000) и валидационный (4000). Предобученная модель EfficientDet D2 на наборе данных COCO обучалась 20 эпох (120000 шагов). Каждую эпоху производилась валидация модели и сохранялось ее состояние. В результате для тестирования была выбрана модель, полученная после 18 эпохи.

3.7. Тестирование модели

Для тестирования модели использовался набор реальных данных мощностью 277 изображений опубликованный в работе [21]. Для сравнения в табл. 1 приведены результаты тестирования (с применением постобрабоки (**Post**)) четырех моделей, обученных на разных наборах данных из [21]: **PLN** – набор данных научных статей PupLayNet, 125 тысяч изображений, **GEN** – набор сгенерированных данных, 18 тысяч изображений, **NPA-small** – набор реальных данных, 100 изображений, **NPA-big** – набор реальных данных, 500 изображений. Для оценки качества использовались метрика PASCAL VOC – (AP) = 0.5.

Табл. 1. Результаты сравнительной оценки
Table 1. Comparation results

Метод	Текст	Таблица	Изображение	Итого
PLN+GEN+NPA-big	0.810	0.937	0.696	0.820
PLN+GEN+NPA-big + Post	**0.840**	**0.968**	0.755	**0.855**
PLN+GEN+NPA-small	0.502	0.824	0.336	0.559
PLN+GEN+NPA-small + Post	0.652	0.888	0.448	0.663
PLN+NPA-small	0.489	0.846	0.346	0.565
PLN	0.045	0.065	0.004	0.039
TaggedPDF	*0.032*	*0.876*	***0.763***	*0.557*

Результаты тестирования, приведенные в Табл. 1. показывают, что модель, обученная только на автоматически аннотированных данных, показывает лучший результат на классе *изображение*. Сопоставимый результат с моделью (PLN+GEN+NPA-small + Post), дообученной на трех разных наборах данных с использованием постобработки достигается для класса *таблица*.

Неудовлетворительная оценка получена для класса *текст*. Это объясняется тем, что Microsoft Word предоставляет большой набор инструментов для редактирования текста. Он может быть отредактирован различными способами при полной визуальной идентичности. При этом семантически единый текстовый блок для эксперта во внутреннем представлении документа будет разбит на множество частей с разными тегами. Это порождает несоответствие экспертной разметки с автоматической, из-за чего данный класс распознается плохо.

С другой стороны, для редактирования таблиц и вставки рисунков в документ представлен ограниченный набор инструментов, что порождает более точное соответствие внутреннего представления с визуальным. Из-за этого автоматическая разметка в основном соответствует экспертной.

4. Заключение

Предложенный метод автоматической разметки данных для решения задачи сегментации изображений документов можно считать эффективным только для классов с высокой согласованностью экспертов. К таким классам относятся таблица и изображение, а из не рассмотренных в данной работе – формула, колонтитул, номер страницы. В других случаях из-за неоднозначности внутреннего представления форматирования документа с его визуальным восприятием данные размечаются некорректно.

Основываясь на полученных результатах, можно сделать вывод, что предложенный метод имеет ряд ограничений, связанных с разнообразием инструментов форматирования Microsoft Word. Для устранения указанных ограничений возможно использование в качестве исходных данных файлы в формате LaTex. Он имеет более строгий набор инструментов форматирования документов, что должно позитивно повлиять на устранение неоднозначности между внутренним представлением документа и его визуальным представлением. В качестве альтернативы можно использовать синтетически сгенерированные документы в качестве исходных данных для аннотирования обучающей выборки.

Список литературы / References

[1] Lee E., Park J. et al. Deep-learning and graph-based approach to table structure recognition. Multimedia Tools and Applications, vol. 81, issue 4, 2022, pp. 5827-5848.
[2] Le V.P., Nayef N. Text and non-text segmentation based on connected component features. In Proc. of the 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1096-1100.
[3] Wong K.Y., Casey R.G., Wahl F.M. Document analysis system. IBM Journal of Research and Development, vol. 26, issue 6, 1982, pp. 647-656.
[4] Okun O., Doermann D., Pietikainen M. Page segmentation and zone classification: The state of the art. Technical Report LAMP-TR-036, CAR-TR-927, CS-TR-4079. University of Maryland, 1999, 38 p.
[5] Moll M.A., Baird H.S., An C. Truthing for pixel-accurate segmentation. In Proc. of the Eighth IAPR International Workshop on Document Analysis Systems, 2008, pp. 379-385.
[6] Moll M.A., Baird H.S. Segmentation-based retrieval of document images from diverse collections. In Proc. of the IS&T/SPIE 20th Annual Symposium on Electronic Imaging, 2008, 8 p.
[7] Fletcher L.A., Kasturi R. A robust algorithm for text string separation from mixed text/graphics images. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 10, issue 6, 1988, pp. 910-918, 1988.
[8] Tombre K., Tabbone S., et al. Text/graphics separation revisited. Lecture Notes in Computer Science, vol. 2423, 2002, pp. 200–211.
[9] Bukhari S.S., Al Azawi M.I.A. et al. Document image segmentation using discriminative learning over connected components. In Proc. of the 9th IAPR International Workshop on Document Analysis Systems, 2010, pp. 183-190.
[10] Kang L., Kumar J. et al. Convolutional neural networks for document image classification. In Proc. of the 22nd International Conference on Pattern Recognition, 2014, pp. 3168-3172.
[11] Harley A.W., Ufkes A., Derpanis K.G. Evaluation of deep convolutional nets for document image

classification and retrieval. In Proc. of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 991-995.
[12] Oliveira D.A.B., Viana M.P. Fast CNN-based document layout analysis. In Proc. of the IEEE International Conference on Computer Vision, 2017, pp. 1173-1180.
[13] Vincent N., Ogier J.M. Shall deep learning be the mandatory future of document analysis problems? Pattern Recognition, vol. 86, 2019, pp. 281-289.
[14] Clausner C., Antonacopoulos A., Pletschacher S. ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017. In Proc. of the 14th International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1404-1410.
[15] Clausner C., Antonacopoulos A., Pletschacher S. ICDAR2019 Competition on Recognition of Documents with Complex Layouts - RDCL2019. In Proc. of the 15th International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1521-1526.
[16] Gao L., Huang Y. et al. ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). In Proc. of the 15th International Conference on Document Analysis and Recognition (ICDAR), 2021, pp. 1510-1515.
[17] Lopes C.A.M. Junior, das Neves R.B. Junior et al. ICDAR 2021 Competition on Components Segmentation Task of Document Photos. Lecture Notes in Computer Science, vol. 12824, 2021, pp. 678-692.
[18] Anitei D., Sánchez J.A. et al. ICDAR 2021 Competition on Mathematical Formula Detection. International Conference on Document Analysis and Recognition. Lecture Notes in Computer Science, vol. 12824, 2021, pp. 783-795.
[19] Yepes A. J., Zhong P., Burdick D.. ICDAR 2021 Competition on Scientific Literature Parsing. Lecture Notes in Computer Science, vol. 12824, 2021, pp. 605-617.
[20] Adams T., Namysl M. et al, Benchmarking table recognition performance on biomedical literature on neurological disorders. Bioinformatics, vol. 38, issue 6, 2022, pp. 1624-1630,
[21] Беляева О.В., Перминов А.И., Козлов И.С. Использование синтетических данных для тонкой настройки моделей сегментации документов. Труды ИСП РАН, том 32, вып. 4, 2020 г., стр. 189–202 / Belyaeva O.V., Perminov A.I., Kozlov I.S. Synthetic data usage for document segmentation models fine-tuning. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 4, 2020. pp. 189–202 (in Russian). DOI: 10.15514/ISPRAS–2020–32(4)–14.
[22] Tan M., Pang R., Le Q.V. EfficientDet: Scalable and efficient object detection. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. pp. 10778-10787..

Информация об авторах / Information about authors

Андрей Анатольевич МИХАЙЛОВ – кандидат технических наук, старший научный сотрудник лаборатории комплексных информационных систем ИДСТУ СО РАН, научный сотрудник ИСП РАН. Сфера научных интересов: искусственный интеллект, большие данные, анализ документов.

Andrey Anatolievitch MIKHAYLOV – Candidate of Technical Sciences, Senior Researcher in the Laborotory of Complex Information Systems at IDSTU SB RAS, Researcher at ISP RAS. Research interests: artificial intelligence, big data, document analisys.