УДК 004.89

СИСТЕМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТОВ СЛОЖНОЙ СТРУКТУРЫ

Бутенко Екатерина Александровна[1], Задорожный Александр Михайлович[2], Любовинкина Наталья Яковлевна[3], Потемкина Снежана Владиславовна[4]

[1]Студент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: e.bytenko2003@gmail.com.

[2]Кандидат физико-математических наук, доцент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: alex.zador@gmail.com.

[3]Старший преподаватель
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
e-mail: natali@uni-dubna.ru.

[4]Кандидат технических наук, доцент;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, д. 19;
141980, Московская область, г. Дубна, ул. Университетская, д. 19;
e-mail: snezhana@uni-dubna.ru.

В работе представлен метод восстановления логической связности текстов, полученных в после применения методов оптического распознавания текстов (OCR) для классификации отсканированных копий деловой документации. Метод включает два этапа. Вначале проводят предварительную сегментацию интересующих областей средствами сверточной нейронной сети (CNN) глубокого обучения с архитектурой yolo. Полученная в результате информация позволяет восстановить логическую связность текста документа. Затем тот же подход применяется для сопоставления названия атрибута и его значения для одного из распространенных видов их представления в виде 2-х колонок: колонки имен и колонки значений. Метод успешно решает задачи классификации документов и извлечения ключевых атрибутов в контексте системы электронного документооборота.

Ключевые слова: система искусственного интеллекта, сегментация документа, сверточная нейронная сеть глубокого обучения, электронный документооборот.

Для цитирования:

Система искусственного интеллекта для классификации документов сложной структуры / Е. А. Бутенко, А. М. Задорожный, Н. Я. Любовинкина, С. В. Потемкина // Системный анализ в науке и образовании: сетевое научное издание. 2023. № 1. С. 7–12. URL: https://sanse.ru/index.php/sanse/article/view/567.

ARTIFICIAL INTELLIGENCE SYSTEM FOR CLASSIFYING COMPLEX STRUCTURE DOCUMENTS

Butenko Ekaterina A.[1], Zadorozhniy Alexandr M.[2], Lyubovinlina Natalya J.[3], Potemkina Snezhana V.[4]

[1]Student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;

e-mail: e.bytenko2003@gmail.com.

[2]PhD in Physical and Mathematical Sciences, associate professor;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: alex.zador@gmail.com.

[3]Senior teacher;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: natali@uni-dubna.ru.

[4]PhD in Engineering sciences, associate professor;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: snezhana@uni-dubna.ru.

The paper presents method for restoring the logical coherence of texts obtained in a result of using Optical Character Recognition (OCR) methods for classifying scanned copies of business documents. The method includes two stages. First, the preliminary segmentation of the areas of interest is performed by means of a convolutional neural network (CNN) deep learning with yolo architecture. The obtained information allows you to restore the logical coherence of the document text. Then the same approach applies to compare the attribute name and its value for one of the common types of their representation in the form of two columns: a column of names and a column of values. The method successfully solves the issues of document classification and extraction of key attributes in the context of an electronic document management system.

Keywords: artificial intelligence system, document segmentation, deep learning convolutional neural network, electronic document management.

For citation:

Butenko E. A., Zadorozhniy A. M., Lyubovinlina N. J., Potemkina S. V. Artificial intelligence system for classifying complex structure documents. *System analysis in science and education*, 2023;(1):7-12 (in Russ). Available from: https://sanse.ru/index.php/sanse/article/view/567.

Введение

Современные системы электронного документооборота являются неотъемлемым ИТ инструментом крупных компаний. Как правило, такие системы включают эффективных помощников в виде систем искусственного интеллекта (СИИ) для структурирования информации. Документы поступают в систему в виде графических образов бумажных носителей. Средний объем информации для обработки в типовой системе электронного документооборота (СЭД) ~100 тыс. страниц в сутки. При этом каждый документ должен сохраняться в базе данных компании с необходимым набором атрибутов для последующего его извлечения в рамках решаемых компанией задач. Речь идет о типе документа, регистрационном номере, дате, наименовании компании, адресе, сумме к оплате, стоимости договора и пр.

В настоящей статье предложен подход к построению СИИ для решения задачи классификации документов со сложной структурой и извлечения из них ключевых атрибутов. Подход, в известной степени аналогичный тому, что использован в данной статье, применялся в работах [1-3] для задач классификации журнальных статей, выделения формул и пр.

Следует отметить, что задача данной работы сложнее, чем постановка в работах [1-3], в силу разнообразия типов документов и их форматов, а также разнообразия атрибутов, которые подлежат извлечению. Внимание авторов сосредоточено на таких типах документов компаний как: Договоры, Соглашения, Акты, Отчеты, Универсальные передаточные документы, Счета-фактуры и других видах документов производственной, коммерческой и хозяйственной деятельности компаний, структура и форма которых установлена государством и не может быть изменена ни при каких обстоятельствах.

Описание проблемы

Рассмотрим, как происходит процесс распознавания документов компании. Вначале отсканированную копию страницы документа в формате *jpeg* подвергают распознаванию с использованием алгоритма оптического распознавания текста (*OCR*). В результате формируется *json*-документ, содержащий текст и координаты всех распознанных слов. На вход в СИИ поступают отсканированные копии страниц документов в формате *jpeg* и *json*-документы. Задача СИИ – классифицировать по типу документа каждую страницу, определить и извлечь необходимые атрибуты. Классификация страниц и извлечение атрибутов выполняется с применением технологий обработки естественного языка (*NLP*) и регулярных выражений к текстам, полученным в результате *OCR*.

Проблема заключается в том, что визуально понятный текст в результате работы *OCR* теряет логически связную структуру. Фрагмент оригинального скана и полученный в результате применения *OCR* текст представлены на рис. 1 и 2. Потеря логической связности текста затрудняет как классификацию документа по типам, так и поиск и извлечение атрибутов документа.

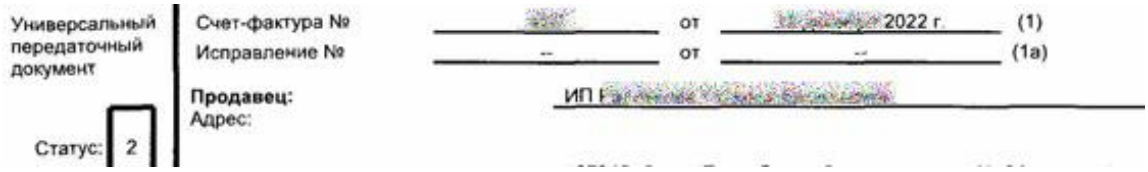

Рис. 1. Фрагмент скана документа

Универсальный Счет—фактура № от 2022 г. (1)
передаточный Исправление № — от —— (1а)
документ
Продавец: ИП (2) Покупатель: (6)
Адрес: Адрес: 117218, город Москва, вн.тер. г. Муниципальный Округ
Статус:й 2 Академический,

Рис. 2. Фрагмент текста в результате OCR

Располагая координатами слов после *OCR* можно алгоритмически восстановить связность текста. Например, предлагается проводить кластеризацию слов по расположению, группируя их в близкие группы. Однако, разнообразие типов и структур документов, помехи, связанные с неточной ориентацией исходных документов при сканировании (отклонения при сканировании документов составляют до 5 градусов) и ошибки при самого *OCR*, делают такой подход сложно прогнозируемым в случае добавления новых типов документов или вариантов их оформления.

Описание предлагаемого решения

В данной статье представлен подход к созданию СИИ для распознавания разных документов со сложной структурой, включающий предварительную сегментацию связных областей образа страницы. Идея заключается в том, чтобы обучить *CNN* выполнять сегментацию документа - распознавать логически связные фрагменты изображения, и полученную информацию использовать для восстановления логической связности текста после *OCR*.

Для сегментации страниц документа использовалась нейронная сеть *CNN* глубокого обучения с архитектурой *yolo* версии 5. Сети этой архитектуры являются на данный момент лидерами в области визуального детектирования объектов. Детали архитектуры описаны в [5], а также достаточно рассмотрены в [1-3]. *Yolo* включает несколько моделей. В работе выбрана модель *medium*, как оптимальная по качеству сегментации и производительности для целей работы.

Набор данных для работы представил из себя комплект из 200 сканов различных страниц документов разных типов.

Первая из задач заключается в выборе системы разметки изображений (электронных копий документов). Виды выделяемых областей или областей интересов зависят от конкретной задачи, решаемой фирмой. Это отдельная область конфигурации СИС. В работе [2], например, выделялись Печать, Логотип, Текст, Подпись и Таблица.

Применительно к области приложения данной работы (коммерческий документооборот) в качестве детектируемых объектов выделены типы областей: Заголовок, Связная прямоугольная область, Таблица, Печать (круглая), Форма, Поле (структурированное поле с данными).

Тексты, которые занимают всю ширину страницы и не теряют связности, в результате применения *OCR*, не размечались. Для единообразной и качественной разметки данных необходимо дать непротиворечивое формализованное описание каждого типа областей, а также привести примеры разметки. Неточности на этом шаге не позволят получить качественно обученную модель *CNN*.

Набор обучения включает 150 изображений, 20 сканов для валидации в процессе обучения и 30 – для контрольного тестирования. После проведения 500 эпох обучения *CNN* интегральный результат составил: *Precision* (точность) – 0.87, *Recall* (воспроизведение) – 0.79 и *mAP* (средняя точность при пересечениях, *IoU* > 0.5) – 0.76.

Пример сегментации связных областей обученной *CNN* приведен на рис. 3.

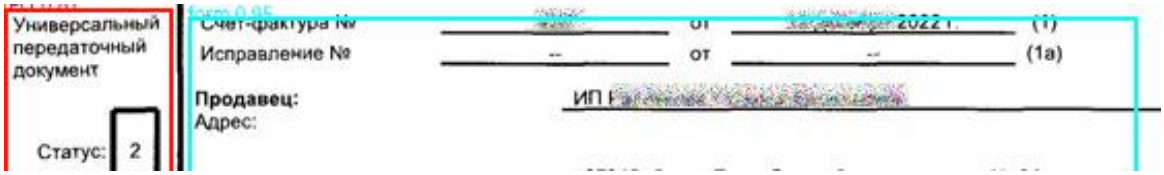

Рис. 3. Сегментация скана при помощи CNN

На рис. 3 слово «Счет-фактура» частично выходит за сегмент «Форма». Для компенсации ошибок *CNN* (низкий *IoU*) на практике достаточно включать слово в ту область, в которую попадает большая его часть.

В процессе анализа страницы *OCR*-текст на основе результатов сегментации (см. рис. 3) преобразовывается в список связных фрагментов с сохранением их порядка в скане.

Для того же примера полученный список приведен на рис. 4.

Рис. 4. Фрагмент текста полученным после обработки OCR и результатов сегментации оригинального изображения

Из примера видно, что теперь один из ключевых атрибутов документа – фраза: «Универсальный передаточный документ» и его статус – представлены одним блоком. Исчезли и нерелевантные слова из поля адреса.

Время обработки одной страницы без использования графического процессора составило ~0.5 сек. Что является вполне приемлемым результатом, особенно с учетом возможностей аппаратного ускорения.

Особенности работы с сегментом типа «Форма»

На практике, при извлечении атрибутов документа, возникла еще одна проблема – с областями типа «Форма». Под формой в работе понимается область структурированного текста, которая логически содержит названия атрибутов и их значения. Часто форма имеет структуру двух колонок – колонку названий атрибутов и колонку значений. Причем, в отличие от таблицы, название и значение не выделены в отдельную строку.

Извлечение большинства атрибутов документа таких как Наименование компании, Адрес или Имя выполняется по ключевым текстовым признакам. При этом, однозначное сопоставление имени и значения без привлечения *NLP*–технологий анализа текста оказывается затруднительным. Даже в пределах одного документа значение поля формы выравнивается относительно имени как по верхнему, так и по нижнему краю (см. рис. 5). Встречаются случаи выравнивания и по центру.

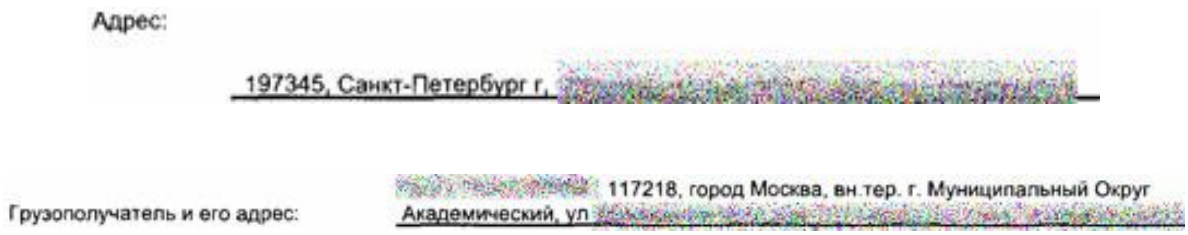

Рис. 5. Значение поля формы выровнено относительно имени по верхнему и по нижнему краю

Дополнительно задача усложняется в тех случаях, когда значения атрибутов занимают несколько строк, которые практически не разделены по вертикали.

Для решения этой проблемы в работе также применен метод сегментации пар Имя-Значение. Если в процессе обработки документа обнаруживается область типа «Форма», то для ее анализа, т.е. извлечения атрибутов, привлекается вторая *CNN*, обученная для сегментации пар Имя-Значение именно для образцов «Форма».

В качестве второй модели использовалась сеть той же архитектуры. Набор данных для обучения включает 50 фрагментов сканов типа «Форма» (более 250 пар имя-значение).

Обучение составило 400 эпох. Итоговый результат: *Precision* (точность) – 0.96, *Recall* (воспроизведение) – 0.95 и *mAP* (средняя точность при пересечениях, *IoU*) > 0.5) – 0.73.

Пример сегментирования фрагмента формы приведен на рис. 6.

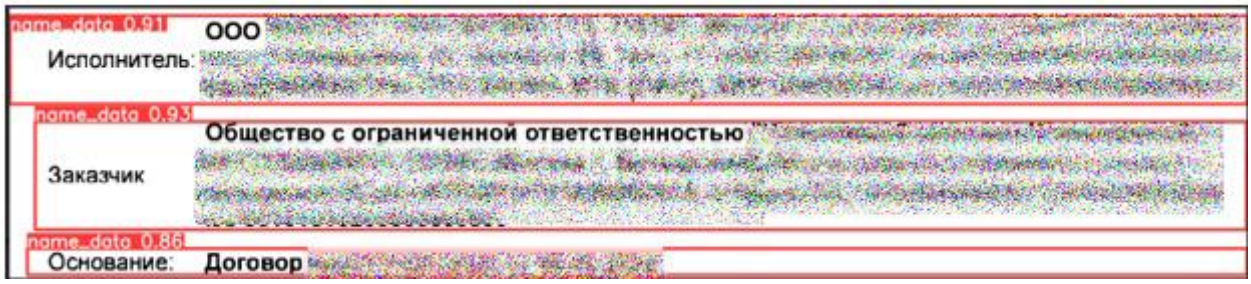

Рис. 6. Сегментирование фрагмента формы с именованными полями данных

Представленный подход принес вполне удовлетворительные результаты. Важно также, что понятен путь улучшения точности и качества сегментации, который заключается в расширении набора данных обучения обеих моделей *CNN*.

Заключение

Совместное применение *NLP*-технологий и технологий сегментации изображений показывает высокую эффективность для задач классификации документов со сложными структурами, распознавания и извлечения их атрибутов.

На первом этапе построения СИИ сегментация графической копии документа при помощи *CNN*-глубокого обучения применяется для реконструкции связности текстов оригинального документа. На втором этапе при детектировании областей типа «Форма», используется та же технология, но с другой моделью *CNN* для сопоставления пар Имя-Значение.

В отличие от алгоритмических и иных подходов к решению задач такого сорта, технологии компьютерного зрения на основе *CNN* имеют контролируемую сложность добавления новых типов документов и образцов их оформления.

Сегментация сканов документов для выявления их логической структуры дополнительно позволяет извлекать атрибуты не текстового характера – подписи, печати, диаграммы, изображения, формулы, области рукописного текста и пр.

Список источников

1. Visual document understanding // John Snow Labs | NLP & AI in Heathcare / John Snow Labs Inc. - [Lewes], 2023. – URL: https://nlp.johnsnowlabs.com/docs/en/ocr_visual_document_understanding (дата обращения 02.02.2023).
2. Yu M. YOLOv4 transfer learning for scanned document structure recognition// Medium : [сайт]. - Дата публикации: 27.07.2020. – URL: https://mingzhi2.medium.com/yolov4-transfer-learning-for-scanned-document-structure-recognition-dc3fc8bfe426/.
3. Segmentation of Scanned Documents Using Deep-Learning Approach / P. Forczmański , A. Smoliński, A. Nowosielski, K. Małecki // Advances in Intelligent Systems and Computing. – 2020. – Vol. 977. – Pp. 141-152. – DOI: https://doi.org/10.1007/978-3-030-19738-4_15
4. Deep learning for table detection and structure recognition: A survey / M. Kasem, A. Abdallah, A. Berendeyev, E. Elkady, [et al.]. // arXiv.org : [open archive]. – 2022. – arXiv:2211.08469 [cs.CV]. – URL: https://arxiv.org/abs/2211.08469. – Submitted on 15 Nov 2022.
5. Automated hierarchical classification of scanned documents using convolutional neural network and regular expression / R. Arief, A. B. Mutiara, T. M. Kusuma, H. Hustinawati // International Journal of Electrical and Computer Engineering. – 2022. – Vol. 12(1). – Pp. 1018-1029. – DOI:10.11591/ijece.v12i1.pp1018-1029.
6. yolov5 : [project] / Ultralytics // GitHub : [web platform]. – GitHub, Inc., 2023. – URL: https://github.com/ultralytics/yolov5?ysclid=lg9f94vy90429921472 (дата обращения: 27.02.2022).