УДК 51 (075.8)

РАСПОЗНАВАНИЕ ТАБЛИЧНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

А. А. Дроздов[1)], Л. Л. Голубева[2)]

[1)] *Белорусский государственный университет, Беларусь, Минск, alexejdrosdov@gmail.com*
[2)] *Белорусский государственный университет, Беларусь, Минск, goloubeva@bsu.by*

Рассматриваются задачи, связанные с извлечением информации из таблиц, содержащихся в документах разных форматов. Как смежная рассматривается задача подготовки данных для корректного обучения и тестирования моделей компьютерного зрения, обусловленная высокой чувствительностью методов к входным данным.

Ключевые слова: распознавание табличных данных; искусственные нейронные сети; разметка обучающих данных.

RECOGNITION OF TABLE DATA WITH USING ARTIFICIAL NEURAL NETWORKS

A. A. Drosdov[1)], L. L. Goloubeva[2)]

[1)] *Belarussian state university, Belarus, Minsk, alexejdrosdov@gmail.com*
[2)] *Belarussian state university, Belarus, Minsk, goloubeva@bsu.by*

Problems related to extracting information from tables contained in documents of different formats are considered. The problem of preparing data for correct training and testing of computer vision models, due to the high sensitivity of methods to input data, is considered as a related problem.

Keywords: tabular data recognition; artificial neural networks; training data labeling.

Введение

Современный мир генерирует огромное количество данных каждый день. В 2018 году аналитики компании IDC в партнерстве с Seagate Technology провели исследование глобальной инфосферы, в ходе которого выполнили количественную оценку всей совокупности созданных, собранных и воспроизведенных данных по всему миру, детально рассмотрели ключевые тренды развития глобального рынка данных и сформировали прогноз. В докладе акцентируется внимание на том, что рост

объема мировых данных будет стремительно продолжаться, и прогнозируется, что к 2025 году совокупный объем информации на нашей планете достигнет отметки в 175 зеттабайт по сравнению с 33 зеттабайтами в 2018 году [1]. Аналитики отмечают, что в последнее время данные все чаще создаются на автоматизированной основе, хранятся в цифровых форматах, при этом они постоянно анализируются и обрабатываются.

Данные существуют во множестве различных форм и размеров, но большинство из них могут быть представлены в виде структурированных и неструктурированных данных. Структурированные данные имеют стандартизированный формат, что обеспечивает эффективный доступ к ним ПО и человеку. Обычно они представлены в форме таблиц со строками и столбцами, четко определяющими атрибуты данных. Примерами структурированных данных являются файлы Excel, базы данных SQL и др. Компьютеры могут эффективно обрабатывать структурированные данные. Однако их обработка и анализ может оказаться трудоемкой задачей, если обрабатываемые таблицы являются либо рукописными, либо не имеют четкого формата, например, представляют собой фотографию, сканированное изображение или файл PDF. Встает задача удобного извлечения данных из таблиц с их последующей цифровизацией или преобразования в подходящий для автоматического анализа цифровой формат.

Задача распознавания таблиц

Задача распознавания таблиц в документах является нетривиальной и достаточно сложной по ряду причин. Сам процесс извлечения информации из таблиц можно разделить на пять задач [2, 3].

Локализация, или *обнаружение таблицы* (table location, table detection): поиск в документе областей, являющихся таблицами, и определение их границ.

Сегментация таблицы (table segmentation): определение структуры таблицы, выделение столбцов, строк, простых и охватывающих ячеек.

Функциональный анализ (functional analysis): определение типа ячеек (ячейки со значениями, ячейки с описателями – заголовками, подзаголовками).

Структурный анализ (structural analysis): выявление связей между ячейками, т.е. зависимостей между заголовками, подзаголовками и значениями; определение ячеек, которые должны быть прочитаны совместно.

Интерпретация (interpretation): преобразование полученного на предыдущих этапах описания табличных данных к целевому представлению.

Далее на этапе постобработки решаются следующие задачи: заполнения цифровой копии таблицы извлеченными данными; сохранения цифровой копии таблицы в удобном формате (csv, excel, и др.).

Кроме того, принимая во внимание специфику работы методов компьютерного зрения, качество обучающих данных играет крайне важную роль для получения хорошей точности обученной модели. Для распознавания табличных данных с использованием методов машинного обучения необходимо иметь большой набор аннотированных обучающих данных. Аннотированные данные – это данные, которые содержат информацию о том, какие объекты находятся в таблице и как они связаны друг с другом. Они играют важную роль в обучении моделей машинного обучения, таких как нейронные сети, и позволяют им распознавать и анализировать данные.

Целью исследовательской работы является создание инструментария для распознавания табличных данных, обладающего следующим функционалом:

- разметка изображения для подготовки обучающего датасета, что включает в себя: отображение объекта изображения со всеми примененными к нему преобразованиями (поворотом, обрезкой, затемнением и др.); расширяемый набор инструментов для преобразования изображения;
- моделирование архитектуры искусственной нейронной сети (ИНС), подходящей для решения задачи распознавания таблиц; обучение и тестирование спроектированной ИНС на собственных наборах данных; применение обученной ИНС; анализ полученных результатов;
- удобный доступ к обученной модели для распознавания пользовательских изображений.

Так как задача распознавания таблиц является широко исследуемой, то существует большое количество методов, которые используются для ее решения. Из наиболее известных можно выделить: систематический фреймворк с визуальным контролем для обнаружения объединенных таблиц и распознавания структуры ячеек Global Table Extractor (GTE), нейронная сеть YOLO, предназначенная для детекции объектов на изображении, сверточная нейронная сеть Retina Net.

В исследовании рассматриваются задачи локализации, сегментации и функционального анализа. Задачи извлечения и интерпретации данных из ячеек таблицы являются, по большому счету, отдельным направлением исследования, в силу своей нетривиальности относительно содержания ячеек, и требуют отдельной глубокой концентрации на данном направлении.

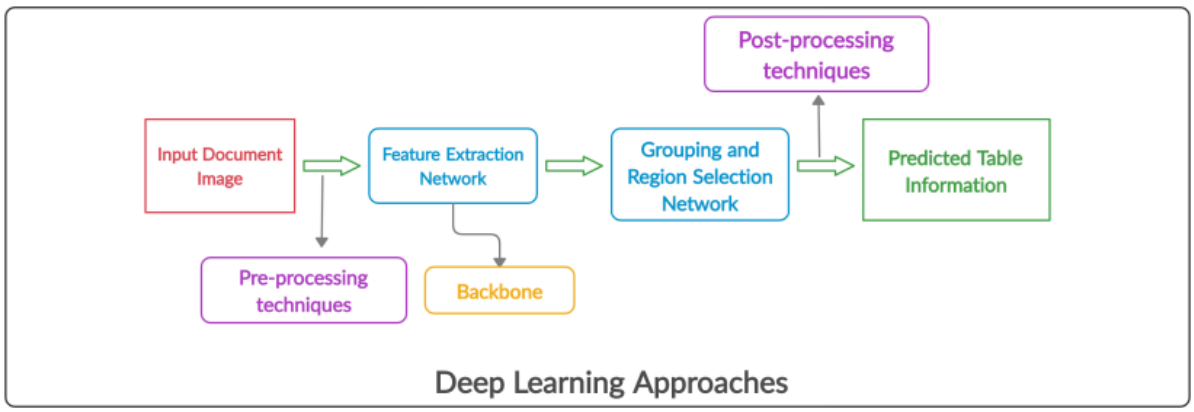

Рис. 1. Использование методов глубокого обучения для анализа таблиц [4].

На рис. 1 отображена последовательность “pipeline” использования ИНС для решения задачи анализа таблиц. Данный подход подразумевает использование ИНС для решения задач обнаружения, сегментации и извлечения данных и, в отличие от традиционных подходов, методы глубокого обучения для понимания таблиц не зависят от данных и обладают лучшими возможностями обобщения.

Предлагаемый метод. Главной целью, поставленной при проектировании метода, являлось создание удобного для использования и последующей модификации метода, который бы решал задачу как распознавания таблицы и ее ячеек, а также позволял производить действия за минимальное время. Для этого выбрана система модульности всех компонентов, основными из которых являются: backbone (тело), neck (шея), head (голова) (dense head, ROI head, mask head), ROI extractor. Для задачи распознавания таблиц были выбраны следующие компоненты: RPNHead, CascadeRoIHead, Shared2FCBBoxHead, SingleRoIExtractor.

Архитектура сети выглядит следующим образом, как представлено на рис. 2: В качестве «тела» выбрана сеть HRNetV2p_W32 [5]. RPNHead (Dense Head) предсказывает предварительные предложения объектов для этих карт объектов. BboxHead принимают функции RoI в качестве входных данных и делают прогнозы с учетом RoI. Mask Head предсказывает маски для объектов. Блок I обозначает входное изображение, блок S – результат сегментации.

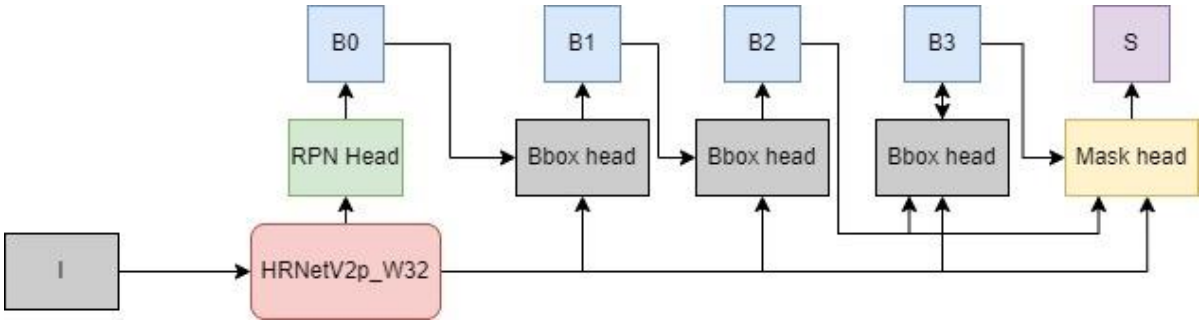

Рис. 2. Архитектура предлагаемого метода

Задачи подготовки данных

Крайне важным для обучения ИНС и методов компьютерного зрения кроме самого метода и архитектуры, является подготовка обучающих данных. Для задачи распознавания табличных данных помимо релевантно выбранных изображений так же необходимо использовать файлы аннотаций, содержащие изначально верные (ground truth) данные об объектах изображения. Существует большое число различных форматов файлов аннотаций, основными из которых являются в настоящее время COCO (Common Objects in Context, Общие объекты в контексте), и VOC (Visual Object Classes, Визуальные классы объектов). Для решения поставленной задачи было спроектировано и реализовано модульное приложение с использованием технологий DirectX и C# .Net WPF для пред- и пост- обработки изображений, а также их разметки с перспективой интеграции метода распознавания пользовательских изображений с таблицами.

Заключение

В результате проведенных исследований:

- выполнен сравнительный анализ существующих методов с выделением их преимуществ и недостатков и обоснована необходимость дальнейшего развития методов для распознавания табличных данных; обоснована необходимость дальнейших исследований в области распознавания и структурной сегментации таблиц;
- разработана легковесная система для распознавания табличных данных с возможностью удобного расширения функционала, основанная на модульности;
- обоснована необходимость корректной подготовки данных для обучения методов машинного зрения; разработано и реализовано приложение, реализующее функционал для разметки обучающих данных и со-

здания файлов аннотации с использованием модульной структуры, с возможностью дальнейшего расширения функционала путем интеграции автоматического применения ИНС сети для разметки данных;

- предложен вектор дальнейшего развития.

Дальнейшим развитием темы исследования является улучшение частей модели, таких как backbone, neck, head, а также создание сервиса для удобного доступа и обмена данными между сервисом для распознавания таблиц и пользовательскими приложениями. Кроме того, перспективным является доработка метода для решения задачи излечения данных.

Библиографические ссылки

1. *Reinsel D., Gantz J., Rydning J.* Data Age 2025. The Digitization of the World From Edge to Core. An IDC White Paper – #US44413318, November 2018. Доступ: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

2. *Бычков И. В., Ружников Г. М., Хмельнов А. Е. и др.* Эвристический метод обнаружения таблиц в разноформатных документах. / Вычислительные технологии. Том 14, № 2, 2009, с.58-73.

3. *e Silva A. C., Jorge A. M., Torgo L.* Automatic Selection of Table Areas in Documents for Information Extraction. // Progress in Artificial Intelligence, 11th Portuguese Conference on Artificial Intelligence, EPIA 2003, Beja, Portugal, December 4-7, 2003, Proceedings. Fernando Moura Pires, Salvador Abreu (Eds.): EPIA 2003, LNAI 2902, pp. 460–465. Springer-Verlag Berlin Heidelberg 2003. [Online]. URL: https://www.researchgate.net/publication/220773906_Automatic_Selection_of_Table_Areas_in_Documents_for_Information_Extraction.

4. *Hashmi K. A., Liwicki M., Stricker D.* Current Status and Performance Analysis of Table Recognition in Document Images With Deep Neural Networks, 2021. DOI:10.1109/ACCESS.2021.3087865. [Online]. URL: https://www.researchgate.net/publication/352270970_Current_Status_and_Performance_Analysis_of_Table_Recognition_in_Document_Images_With_Deep_Neural_Networks.

5. *Jingdong Wang, Ke Sun, Tianheng Cheng*: Deep high-resolution representation learning for visual recognition, 2019. arXiv: 1908.07919. [Online]. URL: https://arxiv.org/abs/1908.07919.