

Сегментация изображений документов при помощи глубоких нейронных сетей

Андрющенко В.С.¹, Михайлов А.А.^{1,2}

¹Институт динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук, 664033, Иркутск, ул. Лермонтова, д. 134, Российская Федерация

²Институт математики и информационных технологий ИГУ, 664003, Иркутск, бульвар Гагарина, 20, Российская Федерация

Аннотация

В статье рассматривается задача автоматического извлечения заголовков из электронных документов. Для решения этой задачи предлагается подход на основе моделей сегментации документов. В ходе исследования была проверена гипотеза о том, что преобразование изображений поможет улучшить качество модели, были собраны и подготовлены данные, а также произведен дообучение моделей нейронной сети.

Keywords

Обработка неструктурированных данных, Анализ электронных документов, Обнаружение заголовков, Глубокое обучение, Методы обработки изображений

1. Введение

Анализ электронных документов является важной задачей в области обработки неструктурированных данных. Он включает в себя распознавание текста, выявление ключевых данных, классификацию, оценку достоверности и соответствия нормативным требованиям. Анализ может использоваться для автоматизации документооборота, повышения эффективности обработки информации и поддержки принятия решений на основе полученных данных. [1]

Одной из задач анализа электронных документов является обнаружение заголовков. Обнаружение заголовков позволяет разбить текст на отдельные физические блоки, что упрощает дальнейший анализ документа. [2] Цель работы — автоматическое извлечение заголовков из электронных документов.

2. Методы и материалы

Для достижения поставленной цели использовались методы на основе глубокого обучения. Используя библиотеку TensorFlow [3], производилось дообучение модели ней-

6th International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2024), July 01–05, 2024, Irkutsk, Russia

✉ vitya.andr.02@mail.ru (B.C.); mikhailov@icc.ru (A.A.)

ORCID 0009-0001-3404-2044 (B.C.)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ICCS-DE 2024 Workshop Proceedings (iccs-de.icc.ru)

DOI: 10.47350/ICCS-DE.2024.27

ронной сети ResNet 50. [4] Для того, чтобы собрать новый набор данных, использовался инструмент Tagged PDF¹.

3. Результаты

На данном этапе работы гипотеза о том, что преобразования изображений документов поможет улучшить точность модели не подтвердилась, преобразования не дали положительного результата. Был создан новый набор данных, с использованием инструмента: Tagged PDF и обучена модель, которая показала средние результаты при тестировании.

4. Выводы

Получены промежуточные результаты: создан новый набор данных и обучена модель, однако результаты модели при тестировании были лишь средними. Это подчёркивает важность продолжения исследований и оптимизации подходов. Неподтверждение гипотезы о положительном влиянии преобразования изображений на точность модели свидетельствует о необходимости дальнейшего изучения и экспериментирования с различными методами предварительной обработки данных.

Дальнейшие исследования должны сосредоточиться на улучшении качества данных, расширении набора данных, оптимизации архитектуры моделей и применении более сложных методов обработки изображений.

5. Заключение

Полученные результаты подчёркивают сложность задачи автоматического извлечения заголовков из электронных документов и необходимость использования комплексного подхода, который включает в себя не только глубокое обучение, но и тщательное предобучение и подготовку данных.

Список литературы

- [1] Томашевская, Яковлев, Способы обработки неструктурированных данных, Russian Technological Journal 9 (2021) 7–17.
- [2] H. Kim, S. Park, S. Yang, Y. Jung, Cilab@ fintoc-2021 shared task: Title detection and table of content extraction for financial document, in: Proceedings of the 3rd Financial Narrative Processing Workshop, 2021, pp. 106–110.
- [3] D. TensorFlow, Tensorflow, Site oficial (2018).
- [4] B. Koonce, B. Koonce, Resnet 50, Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization (2021) 63–72.

¹https://github.com/sunveil/tagged_pdf/tree/master