Классификация блоков изображения на основе графового представления выделенных слов с использованием нейронных сетей

Даниил Копылов[1,2], Андрей Михайлов[1,2]

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

[2]*Институт системного программирования им. В.П. Иванникова Российской академии наук, 109004, г. Москва, ул. Александра Солженицына, д. 25, Российская Федерация*

Аннотация

В данной работе предложен метод классификации структурных элементов текста на основе графового представления этих элементов. Метод использует граф, построенный по словам сегмента, и его статистические характеристики. Точность метода составляет 0.793.

Ключевые слова

Классификация сегментов изображения, анализ макета страницы, представление сегментов

1. Введение

В современном мире информация хранится преимущественно в цифровом виде, значительная часть которой представлена сканированными документами. Обработка такой информации требует распознавания текста и восстановления его логической структуры.

Задача детекции структурных блоков текста (заголовки, основной текст, таблицы) не новая. Одной из ранних работ в этой области является [1]. На сегодняшний день наиболее перспективные решения в этой сфере представлены в работах [2] и [3].

В этих работах для решения задачи используются нейронные сети, основанные на архитектуре Transformer. Однако такие модели, как правило, являются достаточно ресурсоемкими и не способны быстро обрабатывать большие объемы документов, что становится проблемой при работе с тысячами или миллионами файлов.

В качестве альтернативы Transformer-архитектурам в задачах обработки документов набирают популярность подходы, основанные на графовых нейронных сетях. Одним из примеров является работа [4]. Графовые сети уже применялись для решения отдельных задач в этой области, например, сегментации [5].

В данном исследовании мы предлагаем использовать графовый подход для классификации структурных блоков текста. Помимо того, что мы не используем стандартные

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

it-daniil@yandex.ru (Копылов); mikhailov@icc.ru (Михайлов)

0009-0000-6348-4004 (Копылов); 0000-0003-4057-4511 (Михайлов)

ICCS-DE 2024 Workshop Proceedings (iccs-de.icc.ru)
DOI: 10.47350/ICCS-DE.2024.29

графовые сети, мы также строим граф по-другому. В отличие от большинства существующих решений, где графы строятся из участков изображения [6] или строк текста [5], мы предлагаем использовать графы, построенные на основе слов.

Такой подход имеет ряд преимуществ. Во-первых, использование слов естественным образом вписывается в процесс обработки документов, так как на выходе всегда требуется получить текст, а значит его в любом случае необходимо детектировать. Во-вторых, не требуется получать дополнительные элементы из документа, что экономит время и ресурсы при обработке. В-третьих, слова, как правило, распознаются более четко, чем строки, и даже если они сливаются в одно слово, это происходит в рамках одного и того же структурного блока.

Предложенный подход позволит повысить эффективность и скорость обработки больших объемов текстовых документов, что имеет большое значение для различных областей, таких как цифровизация архивов, автоматизация обработки юридической документации и т.д.

2. Предлагаемый подход

В качестве набора данных использовались сегменты текста из PubLayNet [7]. От туда было отобрано 10% изображений, которые затем были сбалансированы и разделены на три набора: обучающий (4386 сегментов), проверочный (1096 сегментов) и тестовый (1370 сегментов).

Для распознавания слов в изображениях использовался OCR Tesseract [8]. Было обнаружено, что для повышения качества распознавания необходимо увеличить разрешение изображений (например, разрешение 100 первых изображений из PubLayNet было увеличено в 4 раза, что привело к увеличению среднего количества распознанных слов с 539 до 681).

Граф строится так, чтобы каждое слово имело ссылки на четырех соседей: 2 по строке (слева/справа) и 2 по абзацу (сверху/снизу). В случае границ, слово может ссылаться само на себя (петля). Для ускорения построения: изображение разбивается на секторы, число которых пропорционально числу слов. Поиск соседа начинается в секторе, где находится слово, а затем в соседних секторах (поиск в глубину). Ограничение по глубине поиска позволяет сократить время поиска и избавиться заведомо лишних связей.

После граф обогащается признаками. Условно их можно поделить на признаки ребер и узлов. Для ребер рассчитывается длина и косинус угла наклона ребра к горизонту. Для узлов высота блока слова и оценка "жирности"(отношение площади к периметру области).

Граф напрямую не может быть передан в MLP-модель. Поэтому он преобразуется в вектор, который представляет собой конкатенацию векторов для каждого признака. Вектора признаков - это интервальные ряды распределения значений признаков. Длина векторов подбиралась экспериментально.

Модель представляет собой MLP с 3 слоями:

1. Первый слой: 182 нейрона, функция активации ReLU.
2. Второй слой: 40 нейронов, функция активации ReLU.

3. Выходной слой: 5 нейронов, функция активации Softmax.

Перед скрытым слоем используется dropout с вероятностью отключения 0.1. В качестве оптимизатора использовался Adam. Функция потерь была выбрана categorical_crossentropy. Число эпох обучение 20 с размером batch равным 1000.

3. Заключение

В рамках данного исследования была продемонстрирована работоспособность предлагаемого подхода к классификации структурных элементов текста на основе графового представления из слов. Точность распознавания составила 0.793. Использование статистических характеристик построенного графа позволило добиться приемлемых результатов классификации. Дальнейшее исследование, направленное на применение графовых нейронных сетей для решения данной задачи, имеет потенциал для достижения результатов, сопоставимых с лидерами в этой области.

Благодарности

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 1023110300006-9).

Список литературы

[1] S. Tsujimoto, H. Asada, Major components of a complete text reading system, Proceedings of the IEEE 80 (1992) 1133–1149. doi:10.1109/5.156475.

[2] C. Da, C. Luo, Q. Zheng, C. Yao, Vision grid transformer for document layout analysis, in: ICCV, 2023.

[3] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, Layoutlmv3: Pre-training for document ai with unified text and image masking, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022.

[4] J. Wang, M. Krumdick, B. Tong, H. Halim, M. Sokolov, V. Barda, D. Vendryes, C. Tanner, A graphical approach to document layout analysis, in: International Conference on Document Analysis and Recognition, Springer, 2023, pp. 53–69.

[5] R. Wang, Y. Fujii, A. C. Popat, General-purpose OCR paragraph identification by graph convolution networks, CoRR abs/2101.12741 (2021). URL: https://arxiv.org/abs/2101.12741. arXiv:2101.12741.

[6] A. L. Maia, F. D. Julca-Aguilar, N. S. Hirata, A machine learning approach for graph-based page segmentation, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018, pp. 424–431. doi:10.1109/SIBGRAPI.2018.00061.

[7] X. Zhong, J. Tang, A. J. Yepes, Publaynet: largest dataset ever for document layout analysis, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1015–1022. doi:10.1109/ICDAR.2019.00166.

[8] R. W. Smith, An overview of the tesseract ocr engine, Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) 2 (2007) 629–633. URL: https://api.semanticscholar.org/CorpusID:7038773.