ОБРАБОТКА ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ГРАФОВЫХ НЕЙРОННЫХ СЕТЕЙ

Трифонов Р. И., Михайлов А. А.

Ключевые слова: логическая структура, сегментация, классификация, графовые нейронные сети, обработка документов

В условиях широкого разнообразия форматов и макетов документов (PDF, сканы со сложным расположением блоков) традиционные OCR-парсеры и rule-based системы требуют ручной настройки и часто теряют семантику структуры. Цель исследования – повысить эффективность существующей системы за счёт переработки архитектуры на основе выводов предыдущих работ [2].

Процесс восстановления логической структуры документа подразделяется на две ключевые задачи: сегментацию фрагментов текста и их классификацию по типу (заголовок, основной текст, список, таблица, изображение). Среди существующих методов, Fast CNN [1] демонстрирует высокую скорость обработки, однако обладает ограниченным учётом контекста. В свою очередь, модель LayoutLM [4] учитывает как текстовую информацию, так и пространственное расположение элементов, но характеризуется значительной ресурсоёмкостью. Модель GLAM [3], основанная на графовых нейронных сетях, объединяет преимущества предыдущих подходов, обеспечивая баланс между скоростью и точностью анализа контекста. Данная методология лежит в основе улучшаемой системы.

Графовые нейронные сети (GNN) — это мощный инструмент для анализа графовых данных. В них узлы представляют слова или строки документа, а рёбра показывают их связи, что позволяет учитывать сложные зависимости и агрегировать информацию от соседей. В отличие от GLAM, в которой граф строится на основе строк, исследуемая система строит граф на базе слов. Этот отличие вызвало уникальную проблему классификаци. Анализ экспериментов [2] показал, что свойства рёбер, такие как длина и угол наклона между словами, важны для точности классификации в графовых представлениях блоков, основанных на словах.

Построение графа осуществляется на основе извлечённых OCR Tesseract слов. Рёбра устанавливаются между ближайшими соседями в четырёх направлениях (влево, вправо, вверх, вниз). Для узлов извлекаются векторные эмбеддинги текста и стилистические признаки

(размер шрифта, жирность, курсив и пр.), а для рёбер – длина и наклон. Такой граф учитывает и визуальные, и текстовые признаки документа.

В качестве архитектуры GNN используется сверточная графовая сеть (GCN), которая агрегирует признаки узлов и их соседей, обучаясь на связях между ними. Сверточный слой GCN описывается формулой (1).

$$H^{(l+1)} = \sigma(\widetilde{D}^{(-1/2)}(A + I)\widetilde{D}^{(-1/2)}H^{(l)}W^{(l)}), \quad (1)$$

где $H^{(l+1)}$ – матрица признаков узлов на l–том слое, A – разреженная матрица смежности графа, I – единичная матрица, $\widetilde{D}$ – диагональная матрица (каждый элемент – это сумма соответствующей строки матрицы A), $W^{(l)}$ – обучаемая матрица весов, σ - функция активации (GELU).

В данной работе предлагается модернизировать архитектуру модели, добавив многоклассовую классификацию типов информации (заголовок, текст, список, таблица, изображение) к бинарной классификации связей, в отличии от классификации узлов как реализовано сейчас. Это повысит точность идентификации информации.

Первоначальные эксперименты с моделью на уровне слов выявили ошибки сегментации и классификации. Это указывает на необходимость настройки модели и дополнительных исследований для оптимизации производительности.

Список литературы

1. Girshick R. Fast R-CNN. https://arxiv.org/abs/1504.08083.
2. Kopylov D., Mikhaylov A. How To Classify Document Segments Using Graph Based Representation and Neural Networks // Ivannikov Memorial Workshop (IVMEM). 2024. P. 36–41. DOI: 10.1109/IVMEM63006.2024.10659393.
3. Wang J., Krumdick M., Tong B. et al. A Graphical Approach to Document Layout Analysis. https://arxiv.org/abs/2308.02051
4. Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. LayoutLM: Pre-training of Text and Layout for Document Image Understanding // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2020. P. 1192–1200. DOI: 10.1145/3394486.3403172.