DOI: 10.15514/ISPRAS-2025-37(6)-14

Сегментация документов на основе графовых нейронных сетей: от строк к словам

[1,3] *Д.Е. Копылов, ORCID: 0009-0000-6348-4004 <it-daniil@yandex.ru>*
[1,2] *А.А. Михайлов, ORCID: 0000-0003-4057-4511 <mikhailov@icc.ru>*
[1,3] *Р.И. Трифонов, ORCID: 0009-0006-0024-8964 <tr1fonov.roman@yandex.ru>*

[1] *Институт динамики систем и теории управления имени В.М. Матросова СО РАН, Россия, 664033, г. Иркутск, ул. Лермонтова, д. 134.*

[2] *Институт системного программирования РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

[3] *Институт математики и информационных технологий Иркутского государственного университета, Россия, 664003, Иркутск, бульвар Гагарина, д. 20.*

Аннотация. В работе представлен метод анализа макета PDF документов на основе графовых нейронных сетей (GNN), использующий слова в качестве узлов графа для преодоления ограничений современных подходов, опирающихся на строки или локальные области. Предложенная модель WordGLAM, основанная на модифицированных графовых сверточных слоях, демонстрирует возможность построения иерархических структур через агрегацию слов, что обеспечивает баланс между точностью детекции элементов и их семантической связностью. Несмотря на отставание от лидирующих моделей в данной области (например, от модели Vision Grid Transformer) по метрикам точности, исследование выявляет системные проблемы области: дисбаланс данных, неоднозначность кластеризации слов («цепные связи», «мосты» между несвязанными регионами), а также спорные критерии выбора классов при разметке. Ключевым вкладом работы является формулировка новых исследовательских задач, включая оптимизацию векторных представлений слов, учет признаков ребер и разработку методов оценки для сложных иерархий. Результаты подтверждают перспективность подхода для создания адаптируемых моделей, способных обрабатывать разноформатные документы (научные статьи, юридические тексты). Работа фокусирует внимание на необходимости дальнейших исследований в области регуляризации и расширения обучающих данных, открывая пути для улучшения переносимости методов анализа макета на новые домены. Код и модели были опубликованы на GitHub (https://github.com/YRL-AIDA/wordGLAM).

Ключевые слова: графовые нейронные сети; сверточные графовые нейронные сети; сегментация документа; анализ макета документа; регионы документа; блоки документа; сегменты документа.

Для цитирования: Копылов Д.Е., Михайлов А.А., Трифонов Р.И. Сегментация документов на основе графовых нейронных сетей: от строк к словам. Труды ИСП РАН, том 37, вып. 6, часть 1, 2025 г., стр. 219–232. DOI: 10.15514/ISPRAS–2025–37(6)–14.

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

Segmentation of Documents Based on Graph Neural Networks: from Strings to Words

[1, 3] *D.E. Kopylov, ORCID: 0009-0000-6348-4004 <it-daniil@yandex.ru>*
[1, 2] *A.A. Mikhailov ORCID: 0000-0003-4057-4511 <mikhailov@icc.ru>*
[1, 3] *R.I. Trifonov, ORCID: 0009-0006-0024-8964 <tr1fonov.roman@yandex.ru>*

[1] *Matrosov Institute for System Dynamics and Control Theory of the Russian Academy of Sciences, 134, Lermontov st., Irkutsk, 664033, Russia.*

[2] *Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*

[3] *Irkutsk State University Institute of Mathematics and Information Technologies, 20, Gagarin Boulevard, Irkutsk, 664003, Russia.*

Abstract. The paper presents a method for analyzing the layout of PDF documents based on graph neural networks (GNN), which uses words as graph nodes to overcome the limitations of modern approaches based on strings or local areas. The proposed WordGLAM model, based on modified graph convolutional layers, demonstrates the possibility of constructing hierarchical structures through word aggregation, which ensures a balance between the accuracy of element detection and their semantic connectivity. Despite lagging behind state-of-the-art models (for example, Vision Grid Transformer) in accuracy metrics, the study reveals systemic problems of the region: data imbalance, ambiguity in word clustering ("chain links", "bridges" between unrelated regions), as well as controversial criteria selecting classes in the markup. The key contribution of this work is the formulation of new research tasks, including optimization of vector representations of words, consideration of edge embeddings, and development of estimation methods for complex word hierarchies. The results confirm the prospects of the approach for creating adaptable models capable of processing multi-format documents (scientific articles, legal texts). This paper highlights the need for further research in the field of regularization and extension of training data, opening up ways to improve the portability of layout analysis methods to new domains. The code and models were published on GitHub (https://github.com/YRL-AIDA/wordGLAM).

Keywords: graph neural networks; convolutional graph neural networks; document segmentation; document layout analysis; document regions; document blocks; document segments.

For citation: Kopylov D.E., Mikhailov A.A., Trifonov R.I. Segmentation of documents based on graph neural networks: from strings to words. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 6, part1, 2025, pp. 219-232 (in Russian). DOI: 10.15514/ISPRAS-2025-37(6)-14.

Acknowledgements. The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (theme No. 1023110300006-9).

1. Введение

Современные предприятия и организации ежедневно сталкиваются с огромными объемами документов, большая часть которых хранится в формате PDF. Этот формат, разработанный для кроссплатформенного представления данных, существует в двух основных вариантах: 1) в виде изображения (растрового/векторного) без семантической разметки; 2) в виде структурированного файла, содержащего текст, метаданные, векторную графику и инструкции для визуализации, что обеспечивает точное воспроизведение макета на любых устройствах. Несмотря на свою универсальность, PDF документы остаются сложными для автоматизированного извлечения структурированной информации, что порождает необходимость применения методов анализа макета документа (Document Layout Analysis, DLA). Хорошими обзорами указанной проблемы являются [1] и [2].

Анализ макета документа направлен на автоматическое обнаружение и классификацию семантических элементов, таких как заголовки, абзацы, таблицы, изображения и списки. Как

правило, он предшествует решению других задач, например, поиска информации по документам, пересказа или вопросно-ответных систем для документов.

Традиционные подходы к анализу макета документов можно разделить на методы, основанные на правилах (rule-based), и системы использующие нейронные сети. Пионерские работы, например, работа [3], применяли правила, учитывающие расстояние между блоками, выравнивание и характеристики шрифтов, что требует ручной настройки правил для каждого типа документов и ограничивает адаптивность к разнообразным макетам. С другой стороны, нейронные сети, особенно основанные на архитектуре Трансформер (см., например, [4-7]), обрабатывают документ как изображение, набор текстовых блоков или комбинацию изображения и текста. На текущий момент они демонстрируют высочайший уровень качества в решении задач, связанных с анализом макетов документов.

Насколько известно авторам, лучшим результатом на наборе данных [8] остается модель Vision Grid Transformer (VGT) из работы [6], ее оценка по метрике mAP@IoU[0.5:0.95] составляет 0.962. Однако ключевым недостатком модели является ее плохая переносимость на новые домены данных.

В этом контексте перспективными представляются модели, основанные на графовых нейронных сетях (Graph Neural Networks, GNN) [9-14]. Основное преимущество GNN заключается в их способности моделировать топологические зависимости между элементами, что соответствует природе документов. Использование графа для представления документа не является современным подходом, а отсылает к идеям, ранее высказанным в работах [15-17].

Ключевой проблемой многих современных подходов к сегментации документов с использованием GNN является использование строк [10-11] или локальных областей [9] как узлов графа. Строки в сложных документах, таких как научные статьи или юридические тексты, часто детектируются некорректно из-за разнообразия форматов, шрифтов и структуры. Фрагментарность небольших областей затрудняет установление семантических связей между ними. В отличие от этих подходов, использование слов в качестве узлов графа позволяет достичь баланса между детекцией и связностью. Слова детектируются легче, чем строки, так как они имеют более четкие границы и меньше зависят от форматирования. Кроме того, их можно агрегировать в строки, абзацы или разделы, сохраняя иерархию документа. Таким образом, слова являются «золотой серединой», сочетающей точность детекции с возможностью моделирования структурных зависимостей, что делает их более подходящими для задач сегментации документов.

2. Исследуемая архитектура модели

В данном разделе представлена предлагаемая модель сегментации документов на основе GNN. Первый подраздел описывает структуру входных данных, где узлами графа являются слова. Второй подраздел детализирует архитектуру сети, включая модификации базового подхода из работы [12].

2.1 Графы

Все GNN работают непосредственно с графами, представленными в виде матриц. Граф, как известно, характеризуется парой (V, E), где V – множество узлов, E – множество ребер. В качестве узлов выступают вещественные векторы $v_i \in R^m, i = 1,2,\dots,n$. Эти векторы могут быть получены из участков документа, в частности, регионов, строк, слов. Ребра между узлами, часто кодируются в виде матрицы смежности $A \in \{0,1\}^{n\times n}$, где единицы характеризуют наличие связи, а номера строк и столбцов – номера соответствующих узлов. В качестве альтернативы используется Лапласиан $L = D - A \in Z^{n\times n}$ (целочисленная матрица), где $D = diag\{d_i\}$ – диагональная матрица, где d_i – степень i–го узла. Обе матрицы, как правило, разреженные и хранятся в компактном виде (в данной работе матрица

смежности хранится как массив пар из двух индексов). В добавок к матрице формируется множество векторов $e_i \in R^l, i = 1,2,\dots,k$, содержащим признаки для каждого ребра.

Процесс построения графа начинается с детекции слов. Отметим, что документ может не иметь текстовый слой. Если документ представлен в виде изображения, для детекции слов применяется OCR Tesseract [18], который распознаёт текст и определяет координаты ограничивающих рамок (bounding boxes, bbox) для каждого слова. Для PDF документов с текстовым слоем, текст и сопутствующие метаданные извлекаются при помощи PDF парсера [19]. Вопрос распознавания слов является отдельной задачей и выходит за рамки настоящей работы.

Ключевое отличие от работы [12] заключается в использовании слов вместо строк в качестве узлов. Это обусловлено тем, что строки в документах (например, таблицах или диаграммах) часто содержат ошибки детекции или не несут самостоятельной семантики.

Каждое слово представляется в виде узла графа с тремя группами признаков:

1. Геометрические: координаты левого верхнего угла bbox'а, координаты правого нижнего угла bbox'а, высота и ширина;
2. Текстовые: первые 32 координаты из векторного представления слова (используется базовая многоязычная модель BERT с учетом регистра [4], который возвращает векторы длиной 512), индикаторы ключевых слов (для пяти классов), индикатор маркированного и нумерованного списка, индикатор знаков препинаний («.», «;», «,», «:»);
3. Стилевые: векторное представление шрифта размерности 3, полученное с использованием сверточной модели, обученной на генерированных данных распространенных шрифтов.

Для построения ребер применяется модифицированный метод k-ближайших соседей (k=4). Для каждого узла выбираются четыре соседа: ближайшие слева, справа, сверху и снизу [20]. Расстояние вычисляется от границ ограничивающих рамок слов с учетом направления поиска. Увеличение значения k усугубляет проблему, описанную в разделе 5.1. Альтернативные методы, такие как триангуляция Делоне, не улучшают результаты, но значительно увеличивают вычислительную сложность.

Данный подход интегрирует локальные атрибуты слов с глобальной структурой документа, учитывая как пространственные, так и семантические связи.

2.2 Архитектура графовой нейронной сети

Представленная в настоящей работе модель обозначается как WordGLAM. Наименование данной модели отсылает к модели GLAM, описанной в источнике [12], архитектура которой была взята за основу. Архитектура WordGLAM представлена на рис. 1.

Для реализации модели использовалась библиотека PyTorch 2.6.0 и расширение PyG 2.6.1.

Основу архитектуры составляют слои TagConv (Topology Adaptive Graph Convolutional Networks) [21], являющиеся обобщением идеи обычного сверточного слоя в GNN [22]. Слои TagConv чередуются с линейными слоями. Слои TagConv выполняют свертку на графе, учитывая топологию и признаки соседних узлов, что позволяет эффективно агрегировать информацию о локальной структуре графа. Математически слой TagConv можно записать в следующем виде:

$$X_{j+1} = \sum_{k=0}^{K} \left(D^{\frac{-1}{2}} A D^{\frac{-1}{2}} \right)^k X_j W_{jk},$$

где X_j – матрица векторных представлений узлов на j–м слое, D – диагональная матрица

степеней узлов (числом связей у каждого элемента), W_{jk} – обучаемые веса. Параметр K – характеризует глубину агрегации признаков (для всех слоев выбран $K = 2$). Число слоев TagConv и значение параметра K подбирались экспериментально (подробнее в подразделе 4.3).

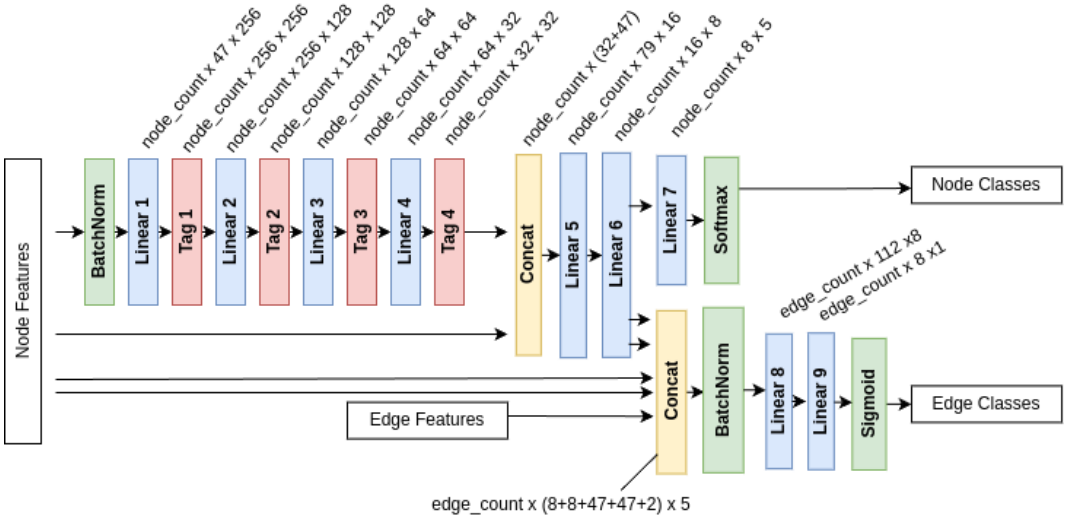

Рис. 1. Архитектура WordGLAM.
Fig. 1. WordGLAM architecture.

На вход первой части модели подаются признаки узлов, сгруппированные в порции данных. Каждая порция подвергается процедуре компонентной нормализации. Данный процесс включает в себя расчет разницы между исходным значением и ее математическим ожиданием, а также последующее деление полученной разницы на ее среднеквадратичное отклонение. Такая нормализация позволяет стабилизировать процесс обучения и улучшить обобщающую способность модели. Выходом первой части являются обогащенные представления узлов, учитывающие их локальный контекст, которые затем передаются на следующий этап обработки.

Вторая часть оценивает значимость ребер. Для каждого ребра входной вектор формируется конкатенацией: исходных признаков связанных узлов, их обогащенных представлений и признаков ребер (длины и наклона). Вторая часть модели на входе также работает с группированными в порцию векторами, которые предварительно проходят компонентную нормализацию. После нормализации векторы проходят через каскад полносвязных слоев, так чтобы на выходе каждый вектор, характеризующий ребро, преобразовывался скалярное значение. Это значение определяет вероятность того, что ребро соединяет два слова внутри региона, а не слова из разных регионов.

Между слоями в обеих частях модели используется нелинейная активация GeLU [23].

Применение такой архитектуры позволяет эффективно выделять компоненты связности в графе, которые соответствуют отдельным блокам текста. Удаление ребер, классифицированных как лишние, приводит к разбиению исходного графа на подграфы, каждый из которых представляет собой отдельный текстовый блок. Это обеспечивает не только сегментацию текста, но и возможность его дальнейшей классификации на основе структуры графа и признаков узлов. Для каждого подграфа строится ограничивающая рамка, которая и является блоком. Если блоки пересекаются, то они объединяются в один (данный вопрос обсуждается в подразделе 5.1).

3. Данные и метрика качества

В работе используется набор данных PubLayNet [8], ставший основным для задачи восстановления макета документа. Для подбора гиперпараметров при обучении использовались 1000 документов из обучающей выборки PubLayNet (из них 10% использовались для валидации модели), при тестировании использовались 50 документов из валидационной выборки PubLayNet. Финальное обучение проводилось на всем наборе данных (также 10% использовалось для валидации модели), а тестирование выполнялось на 200 документах из валидационной выборки.

Для оценки моделей детекции традиционно применяют метрику mAP@IoU [0.5:0.95]. Качество детекции одного региона оценивается с помощью IoU (Intersection over Union) – отношения площади пересечения верного и предсказанного региона к площади их объединения. Для подсчета общей точности используется mAP (Mean Average Precision) – площадь под кривой на графике с полнотой (Recall) в качестве абсциссы и точностью (Precision) в качестве ординаты. Эта кривая строится путем варьирования порога IoU (от 0.5 до 0.95 с шагом 0.05). Более подробно с метрикой можно ознакомиться, например, в работе [24].

Критерий IoU основан на площади пересечения и объединения. Для регионов текста такой критерий не отражает семантической целостности региона. К ошибкам в семантике можно отнести ложные штрафы за неточные границы, даже если все слова региона попали в него. Другой пример, когда при разделении одной и той же строки в регионах с одним и тем же текстом, но разными межстрочными интервалами оценки будут разными, хотя семантически ошибка одна и та же.

Общий результат подсчитывается с использованием метрики mAP с заданным порогом. В своем расчете метрика учитывает уверенность модели в наличие региона, который она детектировала. На практике уверенность модели заменяется фиксированным порогом, что делает оценки метрик не показательными для практических задач.

Авторами в качестве критерия предлагается рассматривать не отношение площадей, а отношение числа слов WordIoU. В качестве самой оценки, используется классическая F1-мера с двумя порогами: 0.5 и 0.95 (оптимистичный и пессимистичный).

$$WordIoU = \frac{count(W_{true} \cap W_{pred})}{count(W_{true} \cup W_{pred})},$$

где W_{true}, W_{pred} – слова находящиеся в правильно размеченном и предсказанном регионах соответственно. Далее в таблицах указана и классическая метрика mAP@IoU [0.5:0.95], которая по мнению авторов не показывает объективно качество для документов.

4. Обучение модели

В данном разделе, перед тем как перейти непосредственно к обучению модели на всем наборе данных (подраздел 4.4), обсуждается важный для нашей модели вопрос балансировки данных (подраздел 4.1), функции потерь (подраздел 4.2), подбор числа слоев и глубины агрегации (подраздел 4.3). Для подбора гиперпараметров везде использовался оптимизатор Adam с темпом обучения (learning rate), равным 5×10^{-3}. Обучение проходило в течение 30 эпох, веса обновляются через каждые 80 документов.

4.1 Балансировка данных

При сегментации на уровне строк распределение связей внутри и между регионами близко к сбалансированному. В случае выстраивания графа из слов, это становится не так. Однако при использовании слов в качестве узлов доля внутрирегиональных связей снижается в 5 раз по сравнению с межрегиоными, что создает дисбаланс классов. Для компенсации дисбаланса

применен метод взвешивания классов (class weighting). Для положительного класса, когда связь внутри региона (оба слова из одного региона), ставится вес равный 0.15. Аналогичный дисбаланс обнаружен в задаче классификации узлов, где также использовано взвешивание (для изображений вес равен 2.63, для текста 0.015, для заголовка 0.946, для списка 1.268, для таблицы 0.136). Результаты обучения сравнения моделей с балансировкой и без нее представлены в табл. 1.

Табл. 1. Важность балансировки данных для обучения модели.
Table 1. The importance of balancing data for model training.

Балансировка	**F1@WordIoU[0.5]**	**F1@WordIoU[0.95]**	**mAP@IoU[0.5:0.95]**
Нет	0.1103	0.0441	0.0050
Есть	0.5257	0.3457	0.0834

4.2 Функция потерь

Обучение исходной модели GLAM и новой модели WordGLAM осуществлялось с использованием комбинированной функции потерь, которая совмещает задачу классификации ребер и узлов. Функция потерь определяется как:

$$Loss = \alpha Loss_{node} + (1 - \alpha) Loss_{edge},$$

где

- $Loss_{edge}$ – бинарная кросс-энтропия для классификации ребер;
- $Loss_{node}$ – кросс-энтропия для классификации узлов;
- α – весовой коэффициент.

В модели GLAM совмещаются задачи классификации и сегментации регионов. На данном этапе решается только задача сегментации, одна из причин описана в подразделе 5.4. Тем не менее, функция потерь представляет собой композицию функций $Loss_{edge}$ и $Loss_{node}$. При сегментации используются векторные представление узлов, и тем самым классификация выступает некоторой регуляризацией. За счет такого штрафа модель строит векторное представление узлов не только для решения задачи сегментации, а в целом возвращает векторы, которые содержат информацию о словах, в частности, к какому классу (к тексту, списку, заголовку или таблице) они относятся.

Был проведен эксперимент для разных значений α, который показал необходимость наличия данного слагаемого (табл. 2).

Табл. 2. Эксперимент с параметром функции потерь.
Table 2. Experiment with the parameter of the loss function.

α	**F1@WordIoU[0.5]**	**F1@WordIoU[0.95]**	**mAP@IoU[0.5:0.95]**
0.05	0.1536	0.0607	0.0060
0.25	0.1780	0.0803	0.0097
0.50	0.4083	0.2485	0.0526
0.75	**0.5257**	**0.3457**	**0.0835**
0.85	0.4950	0.3137	0.0630
0.95	0.0053	0.0011	0.0000

4.3 Подбор числа слоев и глубины агрегации

В архитектуре WordGLAM ключевую роль играют слои TagConv, эффективность которых критически зависит от глубины агрегации соседей (K) и количества слоев. Оптимальные значения этих параметров зависят от сложной иерархической структуры обрабатываемых документов. Слишком малое K или недостаточное число слоев ограничивает охват контекста,

лишая модель преимуществ графового подхода. Чрезмерно большое K или избыток слоев может привести к «размытию» полезной информации в графе и переобучению. По результатам анализа данных (табл. 3) было принято решение использовать 4 слоя с параметром K=2.

4.4 Обучение модели на всем наборе данных

В отличие от экспериментов для подбора гиперпараметров, которые осуществлялись на 1000 документах, а веса модели обновлялись каждые 80 документов, в эксперименте по обучению модели на всем наборе данных веса обновлялись после каждых 500 документов. Все остальные параметры в эксперименте остаются неизменными. Поведение функции потерь при обучении представлено на рис. 2. Несколько примеров обработанных документов приведены на рис. 3. Сравнения с моделью VGT [6] (модель возвращает уровень доверия, поэтому берутся только те регионы, которые имеют уровень доверия больше 0.5, и считаем, что они имеют уровень доверия равный 1), приводится в табл. 4. Модель WordGLAM значительно уступает лидирующей модели VGT по точности. При увеличении объема обучающей выборки наблюдается значительное снижение точности по всем метрическим показателям, что превышает двукратное уменьшение. Причины такого поведения требуют дальнейшего исследования.

Табл. 3. Значения метрик mAP@IoU[0.5:0.95] и F1@WordIoU[0.5] в зависимости от числа слоев (строки) и глубины K (столбцы).
Table 3. mAP@IoU[0.5:0.95] and F1@WordIoU[0.5] Metrics vs. Number of Layers (Rows) and Depth (Columns).

		Глубина (K)			
		2	**3**	**4**	**5**
Число слоев	**1**	0.0543/0.4595	0.0546/0.4422	0.0379/0.4211	0.0449/0.3775
	2	0.0504/0.4162	0.0478/0.4332	0.0601/0.4609	0.0478/0.4598
	3	0.0475/0.4401	0.0503/0.4454	0.0488/0.4231	0.0486/0.3662
	4	**0.0770/0.5262**	0.0486/0.4249	0.0465/0.4090	0.0551/0.4464
	5	0.0468/0.4036	0.0314/0.3609	0.0568/0.4370	0.0623/0.4849

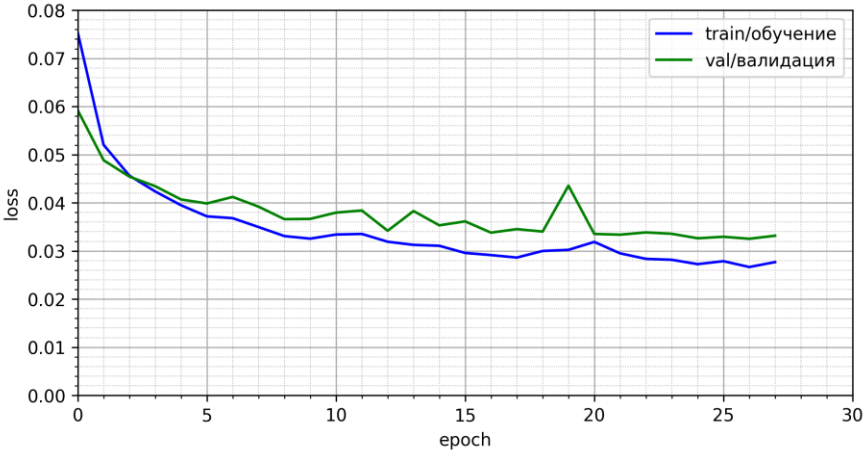

Рис. 2. Обучение WordGLAM на всем наборе данных.
Fig. 2. Learning Word GLAM on the entire dataset.

5. Трудности при работе с документом, представленным в виде графа слов

При переходе с уровня строк на уровень слов возник ряд трудностей, не все из которых удалось решить. Для части из них удалось только сформулировать проблемы. В этом разделе перечислены открытые вопросы, которые возникают из-за того, что происходит переход на более низкий уровень – уровень слов.

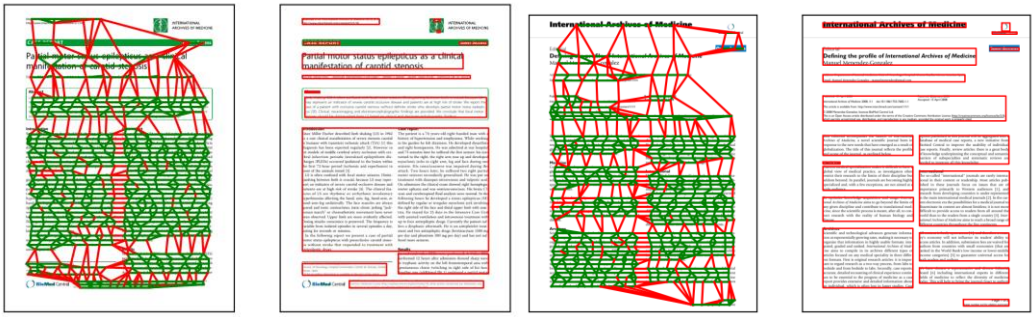

Рис. 3. Пример работы WordGLAM.
Fig. 3. An example of how WordGLAM works.

Табл. 4. Сравнение WordGLAM с VGT [6].
Table 4. Comparing WordGLAM with VGT [6].

	mAP@IoU [0.5:0.95]	**F1@WordIoU [0.95]**	**F1@WordIoU [0.5]**
VGT (весь PubLayNet)	0.8713	0.8613	0.9985
WordGLAM (1000 документов из PubLayNet)	0.0835	0.3457	0.5257
WordGLAM (весь PubLayNet)	0.0182	0.1520	0.2645

5.1 «Мосты» и «цепная связь» между регионами

Несмотря на высокую точность обнаружения связей (рис. 4), алгоритм демонстрирует значительные ошибки в сегментации регионов. Основные проблемы связаны с двумя явлениями: «мостами» и «цепными связями».

«Мост» возникает, когда алгоритм ошибочно идентифицирует связь между двумя независимыми регионами, объединяя их в один компонент. Например, на рис. 4 два нижних региона (заголовок и текст) сливаются в единый регион из-за единственной ложной связи. «Цепная связь» – каскадное распространение ошибки объединения на соседние регионы. На том же примере слияние нижних блоков запускает цепную реакцию: алгоритм последовательно объединяет все пересекающиеся регионы, начиная с неверно найденного региона, отмеченного фиолетовой рамкой, который объединяется с регионом в синей рамке в один общий регион. В результате 5 корректно выделенных изначально регионов из-за «цепной связи» объединились в один.

Попытка исправить проблему через запрет объединения регионов потребует введения множества эвристик (например, правил для конкретных типов пересечений), что снизит универсальность обработчика. Альтернативой может стать регуляризация, учитывающая: положение центров блоков (минимизация расстояний между центрами связей) и геометрическую форму (штраф за отклонение от «прямоугольности»). Однако авторам не удалось найти способ для такой регуляризации, что открывает направление для новых исследований.

5.2 Текст или абзацы, таблицы или колонки

Основу геометрического графа составляют пространственные признаки элементов: их координаты, визуальная схожесть (шрифт) и расстояние между ними. Эти параметры позволяют группировать элементы в регионы. Однако ключевая сложность заключается в том, что пространственные параметры границы (например, интервалов между абзацами) нарушает корректность сегментации.

Главная проблема при работе с текстом – ложное объединение абзацев. Если между ними нет выраженных интервалов, геометрический граф интерпретирует их как единый блок, несмотря на смысловую разрозненность (рис. 5). Например, два независимых абзаца, оформленных одинаково, будут слиты в один регион. Такие ошибки мало влияют на понимание содержания, но критичны для метрик качества сегментации (например, IoU), которые не учитывают семантику. Это демонстрирует необходимость разработки метрик, оценивающих не только геометрию, но и контекст.

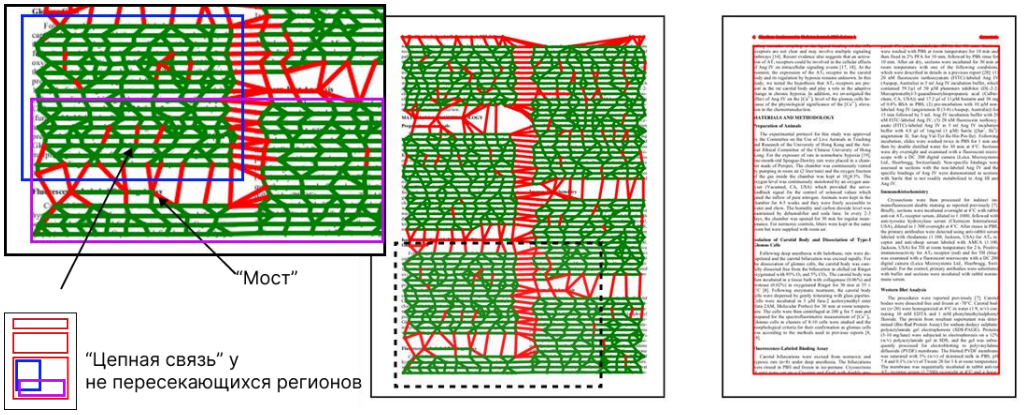

Рис. 4. «Мосты» и «цепные связи».
Fig. 4. «Bridges» and «Chain links».

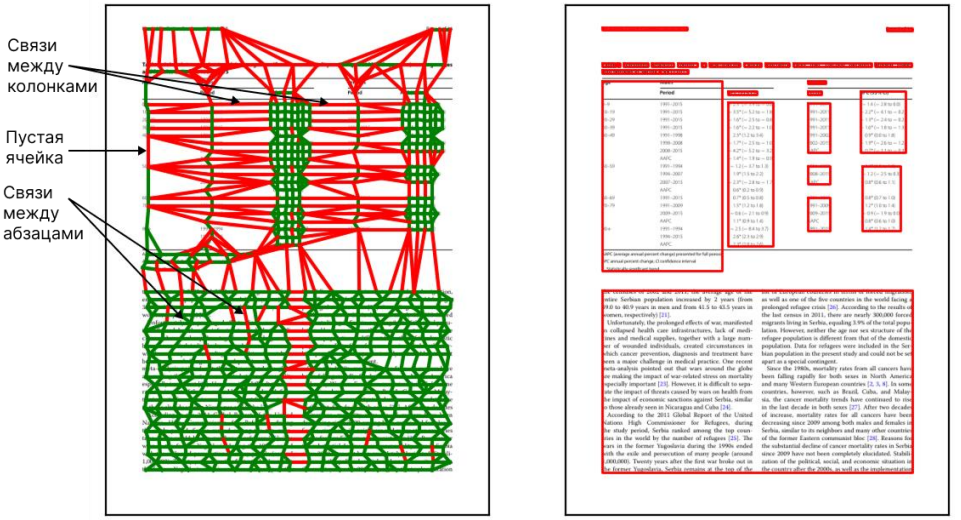

Рис. 5. Связи между колонками и абзацами.
Fig. 5. Edges between columns and paragraphs.

Для таблиц характерна обратная проблема: из-за разнородных отступов, пустых ячеек и отсутствия визуальных разделителей выделяют изолированные блоки вместо единой структуры (рис. 5). Например, пустая ячейка может трактоваться как граница между столбцами, что разбивает таблицу на несвязанные фрагменты. Это усложняет обработку документов со сложной сеточной структурой, где геометрические признаки не всегда отражают логические связи.

Решение проблемы возможно через расширение спектра классифицируемых объектов. В исходной обучающей выборке было представлено пять типов меток. Для достижения этой цели можно использовать наборы данных с большим количеством типов меток, такие как DocLayNet [25] (11 типов) или M6Doc [26] (74 типа).

5.3 Классификация узлов без признаков ребер

Если документ разбивается на строки, то узлы сами по себе плохо характеризуют блок, чего нельзя сказать о ребрах между словами [20]. Даже визуально таблицы имеют вид прямоугольной сетки; текст имеет горизонтальные линии строк и случайные связи между строками; список будет иметь горизонтальную линию объединяя цифры и маркеры, а остальное как у текста; заголовок в большинстве случаев состоит из одной линии.

В данной архитектуре при классификации не используется информация о ребрах. Возможным улучшением будет являться передача признаков для удаления ребер для классификации.

Другой вариант – обучение модели, которая после сегментации классифицирует граф региона целиком. Такой подход, с одной стороны, разбивает задачу на две, но, с другой стороны, обучение происходит не в общем контексте документов.

5.4 Векторное представление узлов

В модели TAGConv ключевые атрибуты передаются в виде векторного представления узлов. На каждом уровне осуществляется процесс агрегации. В случае признаков стиля, координат и индикаторов их семантика понятна. Однако возникает вопрос о сущности агрегированного вектора слова. Если граф отражает семантические связи между словами, то такой вектор имеет обоснованное значение. Однако в данной модели связи формируются на основе позиции элементов в структуре, и слова не обладают логической взаимосвязью (за исключением случаев, когда они являются элементами таблицы или текстовой строки).

Текстовые признаки формируются с использованием модели и BERT [4]. Для оптимизации размера признакового вектора вместо полных 512 компонентов используются только первые 32, что не превышает удвоенный размер остальных признаков. Эксперимент с полным вектором (512 компонентов) показали значение метрики mAP@IoU[0.5:0.95] = 0.0223. Вопрос об оптимальном способе интеграции текстовой информации требует дальнейшего изучения.

6. Заключение

В данной работе предложен альтернативный подход к сегментации документов на основе графовых нейронных сетей (GNN), где в качестве узлов графа используются слова. Несмотря на то, что текущие результаты модели уступают по метрикам современным подходам, опирающимся на строки или локальные области, метод демонстрирует значительный исследовательский потенциал. Использование слов как базовых элементов позволило выявить ранее незаметные проблемы, связанные с балансировкой данных, регуляризацией функции потерь и оценкой качества моделей в области восстановления логической структуры документов.

Архитектура модели, основанная на модификациях графовых сверточных слоев из работы

[12], хотя и не превзошла существующие аналоги, подтвердила возможность построения иерархических структур документа через объединение слов. Модель требует дальнейшей настройки гиперпараметров и расширения обучающей выборки для улучшения обобщающей способности.

Еще одним вкладом работы стало обнаружение новых исследовательских вызовов, связанных с переходом на уровень слов: например, проблема «мостов» и «цепных связей» при кластеризации; вопрос правильного выбора классов при разметке данных; вопрос о том, что более точно характеризует регион: слова как множество или их взаимное расположение. Эти проблемы, ранее не акцентированные в литературе, открывают направления для будущих исследований.

Несмотря на полученные результаты, предложенная модель WordGLAM закладывает основу для более гибкого анализа документов с возможностью переносимости модели на новые домены данных. Ее развитие, включая улучшение векторного представления слов и учета векторного представления ребер при классификации, может привести к прорыву в задачах обработки документов.

Список литературы / References

[1]. Kise K. Page Segmentation Techniques in Document Analysis. In: Doermann, D., Tombre, K. (eds) Handbook of Document Image Processing and Recognition, 2014, Springer, London, pp. 135-175. DOI: 10.1007/978-0-85729-859-1_5.

[2]. BinMakhashen G. M., Mahmoud S. A. Document Layout Analysis: A Comprehensive Survey. ACM Computing Surveys (CSUR), vol. 52, issue 6, pp. 1-36. DOI:10.1145/3355610.

[3]. Tsujimoto S., Asada H. Major components of a complete text reading system. In Proc. of the IEEE, 1992, 80(7), pp. 1133-1149. DOI: 10.1109/5.156475.

[4]. Koroteev M. V. BERT: a review of applications in natural language processing and understanding. CoRR, vol. abs/2103.11943, 2021 [Online]. Available at: https://arxiv.org/abs/1810.04805.

[5]. Huang Y., Lv T., Cui L., Lu Y., Wei F. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. Proc. of the 30th ACM International Conference on Multimedia, 2022, pp. 4083-4091.DOI: 10.1145/3503161.3548112.

[6]. Da C., Luo C., Zheng Q., Yao C. Vision Grid Transformer for Document Layout Analysis. IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 19405-19415, DOI: 10.1109/ICCV51070.2023.01783.

[7]. Sun T., Cui C., Du Y., Liu Y. PP-DocLayout: A Unified Document Layout Detection Model to Accelerate Large-Scale Data Construction. CoRR, vol. abs/2503.17213, 2025 [Online]. Available at: https://arxiv.org/abs/2503.17213.

[8]. Zhong X., Tang J., Jimeno-Yepes A. PubLayNet: Largest Dataset Ever for Document Layout Analysis. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1015-1022. DOI: 10.1109/ICDAR.2019.00166.

[9]. Maia A. L. L. M., Julca-Aguilar F. D. Hirata N. S. T. A Machine Learning Approach for Graph-Based Page Segmentation. In Proc. 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 2018, pp. 424-431. DOI: 10.1109/SIBGRAPI.2018.00061.

[10]. Wang R., Fujii Y., Popat A.C. General-Purpose OCR Paragraph Identification by Graph Convolution Networks. CoRR, vol. abs/2101.12741, 2021 [Online]. Available at: https://arxiv.org/abs/2101.12741.

[11]. Wei S., Xu N. PARAGRAPH2GRAPH: A GNN-based framework for layout paragraph analysis. CoRR, vol. abs/2304.11810, 2023 [Online]. Available at: https://arxiv.org/abs/2304.11810.

[12]. Wang, J. et al. (2023). A Graphical Approach to Document Layout Analysis. Proc. of the 17th ICDAR, 2023, vol. 14191, pp. 53-69. DOI:10.1007/978-3-031-41734-4_4.

[13]. Dai HS., Li XH., Yin, F., Yan, X., Mei, S., Liu, CL. (2024). GraphMLLM: A Graph-Based Multi-level Layout Language-Independent Model for Document Understanding. Proc. of the 18th ICDAR, 2024, vol 14804, pp. 227-243. DOI: 10.1007/978-3-031-70533-5_14.

[14]. Chen Y. et al. Graph-based Document Structure Analysis. CoRR, vol. abs/2502.02501, 2025 [Online]. Available at: https://arxiv.org/abs/2502.02501.

[15]. O'Gorman L. The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15 (11), pp. 1162-1173, 1993. DOI: 10.1109/34.244677.

[16]. Kise K., Sato A. Iwata M. Segmentation of Page Images Using the Area Voronoi Diagram. Comput. Vis. Image Underst, 1998, vol. 70, pp. 370-382. DOI:10.1006/cviu.1998.0684.
[17]. Yi Xiao and Hong Yan. Text region extraction in a document image based on the Delaunay. Pattern Recognit, 2003, vol. 36, pp. 799-809. DOI: 10.1016/S0031-3203(02)00082-1.
[18]. Tesseract User Manual, Available at: https://tesseract-ocr.github.io/tessdoc, accessed 5.08.2025.
[19]. PrecisionPDF, Available at: https://github.com/YRL-AIDA/PrecisionPDF, accessed 5.08.2025.
[20]. Kopylov D., Mikhaylov A. How To Classify Document Segments Using Graph Based Representation and Neural Networks. Ivannikov Memorial Workshop (IVMEM), 2024, pp. 36-41. DOI: 10.1109/IVMEM63006.2024.10659393.
[21]. Du J., Zhang S., Wu G. Moura J. M. F., Kar S. Topology Adaptive Graph Convolutional Networks. CoRR, vol. abs/1710.10370, 2018 [Online]. Available at: https://arxiv.org/abs/1710.10370.
[22]. Kipf T. N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks. CoRR, vol. abs/1609.02907, 2017 [Online]. Available at: https://arxiv.org/abs/1609.02907.
[23]. Hendrycks D., Gimpel K. Gaussian error linear units (GELUs). CoRR, vol. abs/1606.08415, 2016 [Online]. Available at: https://arxiv.org/abs/1606.08415.
[24]. Everingham M., Van Gool L., Williams C.K.I. et al. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 2010, vol. 88, pp. 303–338. DOI: 10.1007/s11263-009-0275-4.
[25]. Pfitzmann B., Auer C., Dolfi M., Nassar A. S., Staar P. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In Proc. of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), 2022, pp. 3743-3751. DOI: 10.1145/3534678.3539043.
[26]. Cheng H. et al. M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15138-15147. DOI: 10.1109/CVPR52729.2023.01453.

Информация об авторах / Information about authors

Даниил Евгеньевич КОПЫЛОВ – магистрант направления подготовки «Прикладная математика и информатика» Иркутского государственного университета, сотрудник Института динамики систем и теории управления имени В.М. Матросова Сибирского отделения Российской академии наук. Сфера научных интересов: прикладная математика, анализ данных.

Daniil Evgenievich KOPYLOV is a master student of Irkutsk State University, employee of Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences. Research interests: applied mathematics, data analysis.

Андрей Анатольевич МИХАЙЛОВ является заведующим Молодежной лаборатории Искусственного интеллекта, обработки и анализа данных Института динамики систем и теории управления имени В.М. Матросова. Сфера научных интересов: анализ электронных документов, распознавание образов.

Andrey Anatolievitch MIKHAYLOV is the head of the Youth laboratory of AI, Data Processing and Analysis of Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences. His research interests include document analysis, image recognition.

Роман Игоревич ТРИФОНОВ – студент направления подготовки «Фундаментальная информатика и информационные технологии» Иркутского государственного университета, сотрудник Института динамики систем и теории управления имени В.М. Матросова Сибирского отделения Российской академии наук. Сфера научных интересов: прикладная информатика, анализ данных, нейронные сети.

Roman Igorevich TRIFONOV is a student of Irkutsk State University, employee of Matrosov Institute for Systems Dynamics of and Control Theory of Siberian Branch of Russian Academy of Sciences. Research interests: applied informatics, data analysis, neural networks.