Построение мультиагентной системы работы с ключевыми графиками научных статей

A Preprint

Maksim M. Mariasov
Department of Mathematical Methods of Forecasting
Moscow State University
Lomonosov Moscow State University
MSU Institute for Artificial Intelligence

Roman V. Ischenko
Department of Mathematical Methods of Forecasting
Moscow State University
Lomonosov Moscow State University
MSU Institute for Artificial Intelligence

Abstract

В работе исследуется способ построения мультиагентной системы автоматического выявления ключевых графиков научных статей, с целью повысить прозрачность и объяснимость процесса выбора графиков, что способствует эффективному анализу публикаций. Поскольку визуальная часть часто концентрирует в себе основные результаты статьи и сопровождает текстовые выводы, система, способная определять ключевые графики и анализировать причины их выбора, позволит исследователям быстрее и точнее выявлять наиболее значимые результаты публикации. Предлагается мультиагентная архитектура, объединяющая работу с текстовыми и визуальными признаками графиков с возможностью генерации объяснений, оценивающих вклад признаков в принятие решения. Проводится исследование отдельных агентов в зависимости от используемых подходов и архитектур и анализируется точность выбора на размеченном наборе статей и его согласованность с человеческими оценками, демонстрируя потенциал подхода.

Keywords Explainable AI · Key Figure Selection · Multimodal Analysis · Multi-agent Systems

1 Введение

Ключевые графики научных статей часто содержат основные результаты исследования и дополняют текстовые выводы, что делает их важным инструментом для быстрого понимания публикации [Lee et al., 2016]. Выделение ключевых графиков требует изучение содержания статьи, что замедляет процесс понимания релевантности её для исследователя. Автоматизация процесса выбора ключевых графиков позволяет эффективнее анализировать литературу и быстрее выявлять значимые результаты. Более того, система с объяснимостью повышает доверие пользователей, делая процесс прозрачным.

Существующие подходы для выявления ключевых графиков опираются на текстовые представления [Beltagy et al., 2019, Reimers and Gurevych, 2019, Герасименко et al., 2024] или на визуальные представления графиков [Radford et al., 2021, Li et al., 2022, 2023]. Мультимодальные модели объединяют информацию из текста и графиков, что позволяет улучшить качество выбора ключевых визуальных элементов. Для упорядочивания графиков по значимости применяются методы ранжирования, включая парный (pairwise) и списковый (listwise) подходы, рассмотренные в работе Liu [2007]. Для интерпретируемости решений используются методы Explainable AI, такие как Grad-CAM [Selvaraju et al., 2017],

Grad-CAM++ [Chattopadhay et al., 2018], Score-CAM [Wang et al., 2019], Text-CAM [Zhang et al., 2025] и LIME [Ribeiro et al., 2016]. Кроме того, применяются подходы рационализации [Zhou et al., 2024, Chen et al., 2025, Hendricks et al., 2018], которые позволяют объяснить, какие признаки текста или изображения оказали наибольшее влияние на решение модели. Использование мультиагентных систем обеспечивает модульность решения, независимость обработки разных модальностей и масштабируемость, облегчая интеграцию различных агентов для работы с текстом, изображениями и ранжированием.

Существующие решения сохраняют ряд существенных ограничений. Во-первых, интеграция текстовой и визуальной модальностей остаётся сложной задачей вследствие различий в их структуры, что затрудняет построение согласованных представлений. Во-вторых, большинство методов автоматического определения ключевых графиков функционируют как «чёрные ящики», не предоставляя интерпретируемых объяснений, что ограничивает их практическую применимость и снижает доверие со стороны исследователей. В-третьих, модели, использующие крупные архитектуры для оценки межмодального сходства, характеризуются высокой вычислительной сложностью, особенно при масштабировании на большие наборы научных публикаций. Наконец, распространённые подходы анализируют графики изолированно, без учёта глобального контекста статьи и взаимосвязей между визуальными элементами, что приводит к снижению точности и устойчивости выбора ключевых графиков.

В данной работе предлагается мультиагентная система, включающая агента обработки текста (Text Agent) и агента обработки изображений (Image Agent), формирующих представления текста и графиков, а также агента ранжирования (Ranking Agent), который объединяет эти представления для окончательного выбора ключевых графиков. Агент объяснений (Explainable Agent) интерпретирует решения системы, демонстрируя вклад отдельных текстовых и визуальных признаков в выбор графиков. Такая архитектура позволяет использовать сильные стороны каждой модальности, при этом централизованно производя ранжирование и обеспечивая объяснимость процесса. Система способна адаптироваться к различным форматам научных статей и обеспечивает воспроизводимость выбора ключевых графиков.

Предложенный подход демонстрирует высокую точность выбора ключевых графиков на размеченном наборе статей, показывая согласованность с оценками экспертов. Мультиагентная архитектура обеспечивает более стабильные и интерпретируемые результаты, позволяя проводить гибкие эксперименты с отдельными агентами. Агент объяснений предоставляет исследователям возможность визуально и текстово анализировать причины выбора графиков, повышая прозрачность и доверие к системе. В целом, предложенная система способствует автоматизации анализа научных публикаций и улучшает инструменты для выявления наиболее значимых результатов исследования.

2 Обзор литературы

2.1 Представление текста

Текстовые данные научных статей содержат ключевую информацию о результатах исследований. Аннотации, подписи к графикам и упоминания графиков в тексте могут содержать в себе признаки их значимости. Модели для извлечения текстовых эмбеддингов, такие как SciBERT [Beltagy et al., 2019], Sentence-BERT [Reimers and Gurevych, 2019] и SciRus [Герасименко et al., 2024], позволяют формировать векторные представления текста, которые применяются для оценки значимости графиков и их связи с основными результатами статьи.

2.2 Представление изображений

Графики и диаграммы отражают результаты статьи, поэтому методы, опирающиеся только на текстовые данные не способны быть полными. Для извлечения признаков изображений применяются модели CLIP [Radford et al., 2021], BLIP [Li et al., 2022] и BLIP-2 [Li et al., 2023], которые создают компактные представления графиков и позволяют сравнивать их с текстовыми описаниями.

2.3 Мультимодальные представления

Объединение текстовой и визуальной информации повышает точность выявления ключевых графиков. Мультимодальные модели, включая BLIP [Li et al., 2022], BLIP-2 [Li et al., 2023], LamRA [Liu et al., 2025] и PaLI-3 [Chen et al., 2023], объединяют представления текста и изображений в единое пространство, учитывая взаимосвязь между подписью и визуальными особенностями графика, что позволяет эффективно анализировать связи между двумя модальностями.

2.4 Ранжирование

Для упорядочивания графиков по значимости применяются методы Learning-to-Rank: парный (pairwise) и списковый (listwise) подходы [Liu, 2007]. Также рассматриваются методы на основе обучения с подкреплением (RL) [Wang et al., 2024] и нейросетевые ранжирующие модели (MLP, Poly-encoders [Humeau et al., 2019]), которые учитывают сложные взаимосвязи между текстовыми и визуальными признаками.

2.5 Explainable AI и рационализации

Интерпретируемость решений моделей критична для доверия пользователей. Для визуальных моделей применяются методы Grad-CAM [Selvaraju et al., 2017], Grad-CAM++ [Chattopadhay et al., 2018], Score-CAM [Wang et al., 2019] и Text-CAM [Zhang et al., 2025]. Для текста и мультимодальных решений используются LIME [Ribeiro et al., 2016] и подходы рационализации [Zhou et al., 2024, Chen et al., 2025, Hendricks et al., 2018], позволяющие показать, какие признаки текста или изображения оказали наибольшее влияние на выбор ключевых графиков.

3 Постановка задачи

3.1 Данные

Рассматриваем набор данных

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},$$

где x_i — данные о i-ой статье, y_i — индексы ключевых графиков статьи.

3.1.1 Структура объекта

Каждый объект x_i представляет собой кортеж

$$x_i = (t_i, a_i, c_i, G_i),$$

где:

- t_i — название i-ой статьи,
- a_i — аннотация статьи,
- c_i — заключение статьи,
- $G_i = (g_{i1}, \ldots, g_{ik_i})$ — набор графиков статьи.

Для каждого графика g_{ij}:

$$g_{ij} = \Big(T_{ij}, \{(t_{ijk}, I_{ijk})\}_{k=1}^{N_{ij}}\Big),$$

где:

- T_{ij} — главное описание j-го графика ($T_{ij} \neq \emptyset$),
- t_{ijk} — описание k-го подграфика j-го графика,
- I_{ijk} — изображение k-го подграфика j-го графика.

Статьи можно разделить на два типа графиков:

- $t_{ijk} \neq \emptyset$ для всех $k = 1, \ldots, N_{ij}$,
- $t_{ijk} = \emptyset$ для всех $k = 1, \ldots, N_{ij}$.

В большинстве случаев графики имеют вид: $N_{ij} = 1$ и $t_{ijk} = \emptyset$ для всех k.

3.2 Структура таргета

Таргет задается множеством индексов ключевых графиков:

$$y_i \subseteq \{1, \ldots, k_i\}.$$

В данной работе ограничиваем $|y_i| \leq 2$, однако в дальнейшем можно изучить влияние выбора количества ключевых графиков. То есть $y_i \in \{1, \ldots, k_i\} \cup \{1, \ldots, k_i\}^2$.

3.3 Отображение

Модель $f_\theta : \mathcal{X} \to \mathcal{Y}$, где $\mathcal{X}$ - множество всех кортежей (t, a, c, G), заданного ранее, которая является представлением статьи, и $\mathcal{Y}$ - множество всех подмножеств $\{1, .., k\}$, где $|G| = k$.

3.4 Внешний критерий качества

Так как мы используем размеченный набор данных с известным таргетом $\hat{y}_i$, то будем использовать классификационные критерии качества:

- Accuracy — доля правильно выбранных графиков относительно всех графиков статьи:
$$\text{Accuracy}_i = \frac{|\hat{y}_i \cap y_i|}{k_i}.$$
- Precision — доля выбранных системой графиков, которые действительно являются ключевыми:
$$\text{Precision}_i = \frac{|\hat{y}_i \cap y_i|}{|\hat{y}_i|}.$$
- Recall — доля ключевых графиков, которые система успешно выбрала:
$$\text{Recall}_i = \frac{|\hat{y}_i \cap y_i|}{|y_i|}.$$
- F1-score — гармоническое среднее Precision и Recall:
$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}.$$

3.5 Оптимизационная задача

Модель $f_\theta : \mathcal{X} \to \mathcal{Y}$ может использоваться двумя способами: она либо присваивает графикам оценки важности для ранжирования, либо предсказывает вероятность того, что график является ключевым (бинарная классификация).

3.5.1 Задача ранжирования

Пусть $f_\theta(x_i, g_{ij}) = s_{ij}$ — оценка важности графика g_{ij} статьи x_i. Требуется, чтобы ключевые графики имели более высокий скор, чем остальные:
$$s_{ij^+} > s_{ij^-}, \quad j^+ \in y_i,\ j^- \notin y_i.$$

Финальное предсказание формируется как выбор двух графиков с наибольшими оценками:
$$\hat{y}_i = \text{argmax}^{(2)}_{j \in \{1,\dots,k_i\}} s_{ij}.$$

Функция потерь для обучения задаётся парной логистической:
$$\mathcal{L}_{rank}(\theta) = \sum_{i=1}^{N} \sum_{j^+ \in y_i} \sum_{j^- \notin y_i} \log\left(1 + e^{-(s_{ij^+} - s_{ij^-})}\right),$$
минимизация которой обеспечивает корректное ранжирование графиков.

3.5.2 Бинарная классификация с выбором двух объектов

Если $f_\theta(x_i, g_{ij}) = p_\theta(g_{ij})$ предсказывает вероятность того, что график является ключевым, финальное предсказание выбирает два графика с наибольшими вероятностями:
$$\hat{y}_i = \text{argmax}^{(2)}_{j \in \{1,\dots,k_i\}} p_\theta(g_{ij}).$$

Функция потерь задаётся через кросс-энтропию с учётом парных ключевых и неключевых графиков:
$$\mathcal{L}_{cls}(\theta) = -\sum_{i=1}^{N} \sum_{j^+ \in y_i} \sum_{j^- \notin y_i} \Big[\log p_\theta(g_{ij^+}) + \log(1 - p_\theta(g_{ij^-}))\Big].$$

Такое формулирование обеспечивает корректный выбор двух наиболее значимых графиков и позволяет напрямую оценивать результаты по внешним метрикам.

4 Предложенный метод

5 Эксперименты

6 Заключение

Список литературы

Po-shen Lee, Jevin D. West, and Bill Howe. Viziometrics: Analyzing visual information in the scientific literature, 2016. arXiv preprint.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620. Association for Computational Linguistics, 2019. doi:10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.pdf.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992. Association for Computational Linguistics, 2019. doi:10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.pdf.

Н.А. Герасименко, А.С. Ватолин, А.О. Янина, and К.В. Воронцов. Scirus: Легкий и мощный мультиязычный энкодер для научных текстов. https://elibrary.ru/item.asp?id=80287449, 2024. eLIBRARY.ru.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. arXiv preprint.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. arXiv preprint.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023), pages 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Tie-Yan Liu. Learning to rank: From pairwise approach to listwise approach. Technical Report TR-2007-40, Microsoft Research, 2007. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2007-40.pdf. Technical Report.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth Balasubramanian. Grad-cam++: Improved visual explanations for deep convolutional networks, 2018. arXiv preprint.

Haofan Wang, Ziyu Wang, Chen Li, Jun Zhang, Zhifeng Li, and Yuxin Pan. Score-cam: Score-weighted visual explanations for convolutional neural networks, 2019. arXiv preprint.

Xinyi Zhang, Yuchen Li, Haoran Wang, et al. Textcam: Explaining class activation map with text, 2025. arXiv preprint.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv preprint.

Xinyun Zhou, Yixin Li, et al. If clip could talk: Understanding vision-language model representations through their preferred concept descriptions. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 547–560, 2024. URL https://aclanthology.org/2024.emnlp-main.547.pdf.

Yiming Chen et al. Explainable saliency: Articulating reasoning with contextual prioritization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025), 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Chen_Explainable_Saliency_Articulating_Reasoning_with_Contextual_Prioritization_CVPR_2025_paper.pdf.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Multimodal explanations: Justifying decisions and pointing to the evidence, 2018. arXiv preprint.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4015–4025, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Liu_LamRA_Large_Multimodal_Model_as_Your_Advanced_Retrieval_Assistant_CVPR_2025_paper.pdf.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger, 2023. arXiv preprint.

Yuchen Wang, Yujie Zhang, Xin Li, Zhiwei Zhang, Fei Wang, Jie Song, and Xiang Chen. Multimodal label relevance ranking via reinforcement learning. In European Conference on Computer Vision (ECCV 2024), pages 833–850, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/08369.pdf.

Samuel Humeau, Jesse Dodge, Julien Gauthier, and Jason Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring, 2019. arXiv preprint.