DOI: http://dx.doi.org/10.21686/2413-2829-2026-1-72-78

ИНТЕГРАЦИЯ МУЛЬТИМОДАЛЬНЫХ ДАННЫХ В ГЕНЕРАЦИЮ ТЕКСТОВЫХ ОПИСАНИЙ: МЕТОДЫ, ВЫЗОВЫ И ПЕРСПЕКТИВЫ

Н. А. Чиняков
Российский экономический университет имени Г. В. Плеханова,
Москва, Россия

Современные системы искусственного интеллекта все чаще используют мультимодальные данные, комбинируя визуальную, текстовую и аудиальную информацию для решения сложных задач. Одной из ключевых областей применения таких систем является генерация текстовых описаний на основе изображений и видео. Интеграция мультимодальных данных позволяет повысить точность и выразительность создаваемых текстов, обеспечивая более полное и осмысленное представление содержимого. В статье рассматриваются современные методы интеграции мультимодальных данных в генерацию текстовых описаний, анализируются ключевые вызовы, с которыми сталкиваются исследователи, а также обсуждаются перспективные направления развития этой области. Особое внимание уделяется использованию сверточных нейронных сетей (CNN) и трансформеров для обработки визуальной информации, а также механизмов внимания и моделей последовательной генерации текста. Исследуются подходы к фьюжну данных из разных модальностей, включая раннее и позднее объединение признаков, а также мультимодальные модели, обученные на больших корпусах данных. Несмотря на значительный прогресс, интеграция мультимодальных данных сопровождается рядом вызовов, включая проблему синхронизации информации, сложности в интерпретации и контексте, ограничения в обучающих данных и др. Обсуждаются перспективные направления развития. Полученные результаты могут быть полезны для разработчиков систем компьютерного зрения, обработки естественного языка и мультимодального машинного обучения, а также для создания интеллектуальных приложений в области автоматической аннотации изображений, видеосуммаризации и человеко-машинного взаимодействия.

Ключевые слова: задача детекции, сверточные нейронные сети, мультимодальные модели, искусственный интеллект.

INTEGRATION OF MULTI-MODAL DATA INTO GENERATION OF TEXT DESCRIPTIONS: METHODS, CHALLENGES AND PROSPECTS

Nikita A. Chinyakov
Plekhanov Russian University of Economics, Moscow, Russia

Current systems of AI more and more often use multi-modal data by combining visual, text and audio information to resolve complicated problems. One key sphere of such system application is description generation on the basis of images and video. Integration of multi-modal data provides an opportunity to improve accuracy and expressiveness of texts being created, which gives more complete and sensible representation of the content. The article studies current methods of multi-modal data integration into generation of text descriptions, analyzes key challenges that face researchers and discusses promising trends in this field of development. Special attention is paid to application of convolution neuron nets (CNN) and transformers to process visual information, as well as attention mechanisms and models of successive text generation. The article researches approaches to data fusion formed by different modalities, including earlier and later combination of signs and multi-modal models trained on big blocks of data. In spite of serious progress, integration of multi-modal data provokes a number of challenges, including information synchronization, problems in interpretation and context, restrictions in learning data and others. Promising lines in development are being discussed. Obtained results can be used by developers of computer vision systems, in

processing natural language and multi-modal machine learning and for elaboration of intellectual applications in the field of automatic image abstracts, video-summarizing and man-machine interaction.
Keywords: detection tasks, convolution neuron nets, multi-modal models, AI.

Введение

Современные системы искусственного интеллекта все чаще используют мультимодальные данные, объединяя визуальную, текстовую и аудиальную информацию для решения сложных задач. Одной из таких задач является генерация текстовых описаний на основе изображений, видео или других типов данных. Интеграция мультимодальных данных в процесс генерации текстовых описаний позволяет существенно повысить качество и информативность создаваемых текстов, обеспечивая более глубокое понимание содержимого.

Развитие методов обработки мультимодальных данных основано на достижениях в области компьютерного зрения, обработки естественного языка и глубинного обучения. Важную роль в этом процессе играют сверточные нейронные сети (CNN), трансформеры и механизмы внимания, позволяющие эффективно анализировать и объединять данные из разных источников. Однако интеграция таких данных сопровождается рядом вызовов, включая проблему несоответствия между модальностями, неоднородность представлений и необходимость синхронизации информации на разных уровнях абстракции.

В работе были использованы научные статьи, обзорные работы и исследования ведущих специалистов в области мультимодального ИИ. Анализ источников позволил систематизировать существующие подходы и выявить основные направления дальнейших исследований.

Мультимодальные данные – это информация, представленная в различных форматах или модальностях, таких как текст, изображение, аудио, видео и сенсорные данные. Их объединение позволяет улучшить точность и выразительность моделей искусственного интеллекта [1]. В отличие от унимодальных данных, которые содержат только один тип информации, мультимодальные данные обеспечивают более глубокое понимание контекста и расширяют возможности анализа.

Современные методы генерации текстовых описаний на основе мультимодальных данных опираются на мощные архитектуры глубокого обучения, способные обрабатывать разнородную информацию. Наиболее распространенными подходами являются конкатенация признаков, внимание (attention) и фьюзия (слияние) на различных уровнях [1].

Конкатенация признаков

Конкатенация признаков (feature concatenation) – это метод объединения различных типов признаков (или характеристик) в одну векторную репрезентацию, которая затем может быть использована в машинном обучении или других алгоритмах анализа данных. Этот подход особенно полезен в контексте работы с мультимодальными данными, где разные источники информации (например, текст, изображения и звук) могут быть представлены разными наборами признаков [4].

В мультимодальных системах, таких как системы генерации текстов из изображений или видео, конкатенация признаков используется для объединения информации из различных источников. Например, можно объединить вектор признаков, извлеченный из изображения (с помощью сверточной нейронной сети), с вектором текстовых признаков (извлеченным с помощью рекуррентной нейронной сети или трансформера).

В качестве примера можно привести системы описания изображений, так как в таких системах можно взять вектор, представляющий изображение, и конкатенировать его с вектором, представляющим текстовые метки или контекстные данные,

чтобы создать более полное представление для генерации описания [1].

Наиболее распространенный способ – это простое объединение векторов признаков:

$$\text{Concat}(x_1, x_2) = [x_1; x_2],$$

где x_1 и x_2 – векторы признаков [1].

После конкатенации часто применяются полносвязные слои для уменьшения размерности и улучшения представления:

$$y = W \cdot \text{Concat}(x_1, x_2) + b,$$

где W и b – веса и смещения полносвязного слоя [1].

Механизмы внимания

Механизмы внимания (attention mechanisms) в машинном обучении представляют собой важный инструмент, который позволяет моделям сосредотачиваться на наиболее значимых частях входных данных. Это особенно важно в задачах, где необходимо учитывать контекст или взаимосвязи между элементами данных.

Механизм внимания позволяет модели взвешивать различные части входной информации, фокусируясь на наиболее релевантных элементах при выполнении задачи. Это можно сравнить с тем, как человек обращает внимание на определенные слова или фразы при чтении текста [5].

Виды внимания:

– *сквозное внимание* (Global Attention): модель рассматривает все элементы входной последовательности;

– *локальное внимание* (Local Attention): модель фокусируется только на определенной части последовательности;

– *самовнимание* (Self-Attention): используется для обработки последовательностей, где каждый элемент может взаимодействовать с другими элементами в пределах одной последовательности (например, в трансформерах).

Один из самых известных примеров применения механизмов внимания – это архитектура трансформера, предложенная в статье Attention is All You Need [5]. Трансформеры используют самовнимание для обработки последовательностей данных.

Нейронные сети (RNN, LSTM)

Рекуррентные нейронные сети (RNN) – это класс нейронных сетей, специально разработанный для обработки последовательных данных. Они находят широкое применение в таких задачах, как обработка естественного языка, распознавание речи, анализ временных рядов и др. Рассмотрим их структуру, принципы работы, преимущества и недостатки, а также примеры использования [6].

В отличие от обычных нейронных сетей, которые обрабатывают фиксированные входные данные, RNN могут принимать последовательности переменной длины. Это достигается за счет наличия циклов в их структуре, что позволяет сети сохранять информацию о предыдущих входах:

1. *Входной слой*: принимает последовательные данные, например, слова в предложении.

2. *Скрытый слой*: содержит нейроны, которые передают информацию друг другу через временные шаги.

3. *Выходной слой*: генерирует выходные данные на каждом временном шаге [7].

LSTM (Long Short-Term Memory) – это тип рекуррентной нейронной сети (RNN), специально разработанный для решения проблемы долгосрочных зависимостей в последовательных данных. Их способность запоминать долгосрочные зависимости делает их подходящими для обработки временных последовательностей и контекста в текстах [8].

Для генерации текстовых описаний мультимодальных данных часто используется комбинированная архитектура, которая включает:

1. *экстракцию признаков*. Для изображений часто используются сверточные нейронные сети (CNN), такие как ResNet или Inception, для извлечения высокоуровневых признаков. Для видео могут исполь-

зоваться 3D-CNN или LSTM для обработки временных аспектов;

2. *интеграцию признаков*. Извлеченные признаки из разных модальностей (например, изображения и текст) комбинируются для создания единого представления. Это может быть сделано с помощью объединения (concatenation) или более сложных методов, таких как внимание (attention);

3. *генерацию текста*. На основе объединенных признаков LSTM используется для генерации текстового описания. Модель принимает на вход вектор признаков и начинает генерировать последовательность слов [4].

Вызовы интеграции мультимодальных данных

Проблемы синхронизации разных модальностей. Одним из основных вызовов при работе с мультимодальными данными является синхронизация различных модальностей. Например, в задачах, связанных с видео и текстом, необходимо точно сопоставить временные метки видео с соответствующими текстовыми описаниями. Неправильная синхронизация может привести к недопониманию контекста и ухудшению качества генерируемого текста. Исследования показывают, что несоответствие во времени может значительно снизить эффективность моделей, работающих с мультимодальными данными [9].

Сложности в интерпретации и контексте. Другой важной проблемой является интерпретация и контекстуализация информации из различных модальностей. Модели могут не всегда правильно интерпретировать информацию, если она представлена в разных форматах. Например, визуальные данные могут содержать нюансы, которые не могут быть адекватно переданы текстом. Это приводит к необходимости разработки более сложных архитектур, способных учитывать контекст и взаимосвязи между различными модальностями [1].

Ограничения в обучающих данных. Качество мультимодальных моделей напрямую зависит от доступных обучающих данных. Однако зачастую такие данные могут быть ограничены или неравномерно распределены между различными модальностями. Например, в случае анализа настроений на основе текста и изображений может быть много текстовых данных, но недостаточно изображений, что приводит к проблемам с обучением моделей [10]. Кроме того, недостаток разнообразия в данных может привести к переобучению модели и ее низкой обобщающей способности.

Этические и социальные вопросы. Эти вопросы также играют важную роль в интеграции мультимодальных данных. Одной из основных проблем является предвзятость в данных. Если обучающие данные содержат предвзятости, модели могут наследовать их воспроизводство в своих выводах. Это может привести к дискриминационным результатам в таких областях, как распознавание лиц или анализ настроений [11]. Исследования показывают, что важно учитывать этические аспекты на всех этапах разработки мультимодальных систем, чтобы минимизировать возможные негативные последствия [12].

Новые подходы и технологии интеграции мультимодальных данных

Современные технологии, такие как глубокое обучение и нейронные сети, открывают новые горизонты для интеграции мультимодальных данных. Например, использование архитектур, таких как Transformer и его модификации, позволяет эффективно обрабатывать информацию из различных источников [5]. Эти подходы способствуют созданию более точных моделей для анализа и генерации мультимодальных данных.

Кроме того, развитие методов обучения с подкреплением генеративных моделей, таких как GAN (Generative Adversarial Networks), также вносит свой вклад в эту область, позволяя создавать более сложные и реалистичные мультимодальные представления [13].

Интеграция мультимодальных данных имеет широкий спектр применения в различных областях, таких как:

1. *Медицина*: в области здравоохранения мультимодальные системы могут использоваться для диагностики заболеваний на основе анализа медицинских изображений, текстовых отчетов и биометрических данных [14]. Например, комбинирование изображений МРТ с текстовыми данными о состоянии пациента может улучшить точность диагностики.

2. *Образование*: в образовательных технологиях мультимодальные данные могут быть использованы для создания адаптивных учебных материалов, которые учитывают предпочтения и стиль обучения каждого студента [15]. Это может включать анализ видеоуроков, текстовых материалов и взаимодействия студентов в реальном времени.

3. *Искусство*: в сфере искусства мультимодальные технологии могут способствовать созданию интерактивных инсталляций и произведений, которые объединяют различные формы искусства – музыку, живопись и литературу [16]. Это создает новые возможности для взаимодействия зрителей с искусством.

Заключение

Интеграция мультимодальных данных в генерацию текстовых описаний представляет собой динамичную и многообещающую область исследований, которая открывает новые горизонты для взаимодействия человека с технологиями. В процессе работы были рассмотрены ключевые методы, используемые для объединения различных форм данных, таких как текст, изображения и звук, а также основные вызовы, с которыми сталкиваются исследователи и разработчики в этой сфере.

Успешная интеграция мультимодальных данных требует не только продвинутых алгоритмов и моделей, но и глубокого понимания контекста, в котором эти данные используются. Проблемы синхронизации, качества данных и их интерпретации остаются актуальными и требуют дальнейшего изучения.

Перспективы данного направления выглядят многообещающе: с развитием технологий глубокого обучения и нейронных сетей мы можем ожидать появления более совершенных систем, способных генерировать текстовые описания, которые будут не только точными, но и эмоционально окрашенными. Это может значительно улучшить пользовательский опыт в таких областях, как образование, медицина и искусство.

Таким образом, интеграция мультимодальных данных в генерацию текстовых описаний открывает новые возможности для создания умных приложений и сервисов, способных адаптироваться к потребностям пользователей и обеспечивать более глубокое понимание окружающего мира.

Список литературы

1. *Baltrusaitis T., Ahuja C., Morency L.-P.* Multimodal Machine Learning: A Survey and Taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 41 (2). – P. 423–443.

2. *Buolamwini J., Gebru T.* Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification // Proceedings of the 1st Conference on Fairness, Accountability and Transparency. – URL: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf?utm_source=chatgpt.com

3. *Esteva A., Kuprel B., Novoa R. A.* Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. – URL: https://www.researchgate.net/publication/312890808_Dermatologist-level_classification_of_skin_cancer_with_deep_neural_networks

4. *Goodfellow I., Pouget-Abadie J., Mirza M.* Generative Adversarial Nets. – URL: https://www.researchgate.net/publication/263012109_Generative_Adversarial_Networks

5. *Hochreiter S., Schmidhuber J.* Long Short-Term Memory // Neural Computation. – 1997. – Vol. 9 (8). – P. 1735–1780.

6. *Jobin A., Ienca M., Andorno R.* The Global Landscape of AI Ethics Guidelines // Nature Machine Intelligence. – URL: https://www.nature.com/articles/s42256-019-0088-2?utm_source=chatgpt.com

7. *Kizilcec R. F., Piech C., Schneider E. F.* Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. – URL: https://www.researchgate.net/publication/260265661_Deconstructing_Disengagement_Analyzing_Learner_Subpopulations_in_Massive_Open_Online_Courses

8. *McCormack J., Gifford T., Hutchings P.* Autonomy, Authenticity and the Role of the Artist in the Age of AI. – URL: https://www.researchgate.net/publication/331562062_Autonomy_Authenticity_Authorship_and_Intention_in_computer_generated_art

9. *Nguyen H., Wang Y., Zhang J.* Multimodal Sentiment Analysis: A Survey on Methods and Applications // IEEE Transactions on Affective Computing. – URL: https://arxiv.org/abs/2305.07611?utm_source=chatgpt.com

10. *Rihem F.* Image Captioning Using Multimodal Deep Learning Approach // Computers, Materials & Continua. – 2024. – Vol. 81 (3). – P. 3951–3968.

11. *Stojkoska B. R., Avramova A. P., Chatzimisios P.* Application of Wireless Sensor Networks for Indoor Temperature Regulation. – URL: https://arxiv.org/abs/1606.07386

12. *Sutskever I., Vinyals O., Le Q. V.* Sequence to Sequence Learning with Neural Networks. – URL: https://arxiv.org/abs/1409.3215

13. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Gomez A. N., Kaiser L., Polosukhin I.* Attention is All You Need. – URL: https://arxiv.org/abs/1706.03762

14. *Vinyals O., Toshev A., Bengio S., Erhan D.* Show and Tell: A Neural Image Caption Generator. – URL: https://arxiv.org/abs/1411.4555

15. *Zadeh A. B., Liang P. P., Poria S., Cambria E., Morency L-P.* Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. – URL: https://aclanthology.org/P18-1208/

16. *Zhang Y., Liu F., Wang H., Hu Z.* Multimodal Learning for Medical Image Analysis: A Survey // Medical Image Analysis. – 2023. – N 85. – P. 102759.

References

1. Baltrusaitis T., Ahuja C., Morency L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, Vol. 41 (2), pp. 423–443.

2. Buolamwini J., Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Available at: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf?utm_source=chatgpt.com

3. Esteva A., Kuprel B., Novoa R. A. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Available at: https://www.researchgate.net/publication/312890808_Dermatologist-level_classification_of_skin_cancer_with_deep_neural_networks

4. Goodfellow I., Pouget-Abadie J., Mirza M. Generative Adversarial Nets. Available at: https://www.researchgate.net/publication/263012109_Generative_Adversarial_Networks

5. Hochreiter S., Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 1997, Vol. 9 (8), pp. 1735–1780.

6. Jobin A., Ienca M., Andorno R. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*. Available at: https://www.nature.com/articles/s42256-019-0088-2?utm_source=chatgpt.com

7. Kizilcec R. F., Piech C., Schneider E. F. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. Available at: https://www.researchgate.net/publication/260265661_Deconstructing_Disengagement_Analyzing_Learner_Subpopulations_in_Massive_Open_Online_Courses

8. *McCormack J., Gifford T., Hutchings P.* Autonomy, Authenticity and the Role of the Artist in the Age of AI. Available at: https://www.researchgate.net/publication/331562062_Autonomy_Authenticity_Authorship_and_Intention_in_computer_generated_art

9. Nguyen H., Wang Y., Zhang J. Multimodal Sentiment Analysis: A Survey on Methods and Applications. *IEEE Transactions on Affective Computing*. Available at: https://arxiv.org/abs/2305.07611?utm_source=chatgpt.com

10. Rihem F. Image Captioning Using Multimodal Deep Learning Approach. *Computers, Materials & Continua*, 2024, Vol. 81 (3), pp. 3951–3968.

11. Stojkoska B. R., Avramova A. P., Chatzimisios P. Application of Wireless Sensor Networks for Indoor Temperature Regulation. Available at: https://arxiv.org/abs/1606.07386

12. Sutskever I., Vinyals O., Le Q. V. Sequence to Sequence Learning with Neural Networks. Available at: https://arxiv.org/abs/1409.3215

13. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Gomez A. N., Kaiser L., Polosukhin I. Attention is All You Need. Available at: https://arxiv.org/abs/1706.03762

14. Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: A Neural Image Caption Generator. Available at: https://arxiv.org/as/1411.4555

15. Zadeh A. B., Liang P. P., Poria S., Cambria E., Morency L-P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Available at: https://aclanthology.org/P18-1208/

16. Zhang Y., Liu F., Wang H., Hu Z. Multimodal Learning for Medical Image Analysis: A Survey. *Medical Image Analysis,* 2023, No. 85, p. 102759.

Поступила: 01.04.2025 Принята к печати: 23.07.2025

Сведения об авторе

Никита Александрович Чиняков
аспирант кафедры информатики
РЭУ им. Г. В. Плеханова.
Адрес: ФГБОУ ВО «Российский экономический университет имени Г. В. Плеханова», 109992, Москва, Стремянный пер., д. 36.
E-mail: arinachinyakova99@mail.ru

Information about the author

Nikita A. Chinyakov
Post-Graduate Student of the Department for Informatics of the PRUE.
Address: Plekhanov Russian University of Economics, 36 Stremyanny Lane, Moscow, 109992, Russian Federation.
E-mail: arinachinyakova99@mail.ru