ИССЛЕДОВАНИЕ И СРАВНЕНИЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ГЕНЕРАЦИИ ЗАКЛЮЧЕНИЯ ДЛЯ РАДИОЛОГИЧЕСКИХ СНИМКОВ

Пан А.Э[1], Аксёнов С.В.[2]
[1] НИ ТГУ, ИПМКН, аспирант, tolikpan0403@gmail.com
[2] НИ ТПУ, ИШИТР, к.т.н., доцент каф. ОИТ, axyonov@tpu.ru

Аннотация

Данный доклад представляет обзор современных методов автоматизированной генерации подписей к медицинским изображениям. Проведен анализ подходов на основе шаблонов, поиска, генеративных и гибридных моделей, оценены их преимущества и недостатки. Представлены результаты сравнительного анализа, выявляющие потенциал для клинического применения.

Ключевые слова: MIC, генерация подписей, медицинские изображения, глубокое обучение, нейронные сети, генеративные модели, гибридные методы, Transfer Learning.

Введение

В современном мире резкое увеличение объема медицинских изображений требует создания высокоэффективных систем автоматизированного анализа. Автоматизация генерации подписей к медицинским изображениям (англ. Medical Image Captioning, MIC) способствует ускоренной диагностике и снижению нагрузки на специалистов [1]. Обзор литературы демонстрирует, что исследования в области генерации подписей активно развиваются благодаря достижениям в глубоких нейронных сетях и трансформерах [2-3]. Однако остаются проблемы, такие как недостаток качественных данных и сложность генерации связного текста, что послужило основанием для постановки цели – провести сравнительный анализ существующих методов и выявить пути повышения их клинической применимости [4].

Методы решения задачи

Шаблонные методы. Шаблонный подход использует заранее подготовленные структуры, куда подставляются ключевые признаки изображения. Например, для рентгенограммы грудной клетки применяется шаблон вида: *«На рентгенограмме грудной клетки в проекции [проекция] выявлены [обнаружения]»* [1].

Преимущества:

- высокая интерпретируемость.
- простота проверки экспертом.

Ограничения:

- ограниченность при обработке сложных и редких патологий.

Методы на основе поиска. Данный метод заключается в сравнении нового изображения с обширной базой данных аннотированных медицинских изображений. Система выбирает наиболее похожие случаи и использует их подписания для формирования нового отчета [2].

Преимущества:

- адаптивность описания.
- снижение зависимости от объема размеченных данных.

Ограничения:

- эффективность зависит от полноты и актуальности базы данных.

Генеративные модели реализуют архитектуру кодер-декодер, где сверточная нейронная сеть преобразует изображение в вектор признаков, а декодер (на базе RNN/LSTM или трансформеров с механизмами внимания) генерирует текстовое описание [3]. Дополнительно, применение diffusion models и self-supervised методов позволяет улучшить

связность генерируемого текста, однако такие модели требуют больших объемов размеченных данных и могут генерировать ошибки [3-4].

Преимущества:

- создание связного и детализированного описания.

Ограничения:

- высокая потребность в размеченных данных;
- риск генерации некорректного текста.

Гибридные методы объединяют методы на основе поиска и генеративные модели, что позволяет компенсировать слабости каждого из методов. Сначала система осуществляет поиск шаблонов, а затем генеративная модель адаптирует найденное описание с учетом специфики нового изображения [5]. Использование Transfer Learning и интеграция экспертных знаний врачей повышают клиническую релевантность результата. Ограничения: высокая вычислительная нагрузка и сложность настройки системы.

Результаты и анализ

Для количественной оценки качества генерируемых подписей используются следующие метрики.

BLEU (BiLingual Evaluation Understudy) – метрика качества машинного перевода, которая основана на подсчете n-грамм слов (n варьируется от 1, до заданного значения, например, 4); чувствительна к вариациям в формулировках, не учитывает семантические связи между словами.

$$BLEU = BP \times \exp\left(\sum_n w_n \log p_n\right);$$

где BP – коэффициент краткости, p_n – точность n-грамм, w_n – веса.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) основан на оценке длины наибольшей общей подпоследовательности (LCS) между сгенерированным и эталонным текстом, учитывает структурную схожесть предложений и способна оценить сохранение смысловых блоков даже при изменении порядка слов. Для ROUGE-L сначала определяются показатели Recall и Precision на основе LCS:

$$R_{LCS} = \frac{LCS(P,R)}{\text{len}(R)}, \qquad P_{LCS} = \frac{LCS(P,R)}{\text{len}(P)}.$$

Затем вычисляется F-мера:

$$ROUGE\text{-}L = F_{LCS} = \frac{(1+\beta^2) \times R_{LCS} \times P_{LCS}}{R_{LCS} + \beta^2 \times P_{LCS}}.$$

Обычно β устанавливается равным 1.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) объединяет точность и полноту с использованием гармонического среднего, а также включает механизм сопоставления синонимов и вариаций в формулировках, что позволяет более гибко оценивать семантическое сходство между сгенерированным текстом и эталоном. Это делает METEOR особенно полезной при анализе качества сложных медицинских описаний.

$$METEOR = F_{\text{mean}} \times (1 - Penalty).$$

где F_{mean} – гармоническое среднее точности и полноты, $Penalty$ – штраф за несовпадения.

CIDEr (Consensus-based Image Description Evaluation) использует взвешивание n-грамм по их TF-IDF значимости и сравнивает косинусное сходство между сгенерированным и эталонным текстами. Эта метрика специально разработана для оценки описаний изображений. Большое значение придается клинически важным терминам.

$$CIDEr = \sum_n (TF\text{-}IDF_n \times \cos(\theta)),$$

где $TF\text{-}IDF$ отражает важность n-грамм, а $\cos(\theta)$ – косинусное сходство между сгенерированным и эталонным текстами.

Хочется отметить, что наличие разнообразных и качественных медицинских наборов данных является ключевым фактором для тренировки и оценки MIC-систем. Качество данных напрямую влияет на стабильность и надежность результатов работы методов (табл. 1).

Таблица 1. Основные медицинские наборы данных

Набор данных	Модальность/Описание	Объем	Комментарий
IU XRay	Рентгенограммы грудной клетки	7 468 изображений	Экспертная разметка, широко используется
MIMIC-CXR	Рентгенограммы грудной клетки	370 110 изображений	Автоматическая разметка (CheXpert labeler)
CheXpert	Рентгенограммы грудной клетки	224 316 изображений	Аннотации проверены экспертами
PadChest	Рентгенограммы грудной клетки	>160 000 изображений	Высокое качество разметки
ChestX-ray8	Фронтальные рентгенограммы грудной клетки	108 948 изображений	Дополнение к существующим коллекциям

Ниже приведена табл. 2, демонстрирующая сравнительный анализ некоторых работ в данной области.

Таблица 2. Сравнительный анализ методов MIC

Метод	Набор данных	B1	B2	B3	B4	M	R-L	CIDEr
Седа-Махмуд (2020)	Собственный	0.56	0.51	0.50	0.49	0.55	0.58	–
Ван (2019)	IU X-Ray	0.34	0.22	0.15	0.10	0.14	0.30	0.32
Шин (2016)	IU X-Ray	0.78	0.40	0.00	0.00	–	–	–
Хуан Синь (2019)	IU X-Ray	0.48	0.34	0.24	0.17	–	0.35	0.30
Цзен (2018)	Собственный	0.30	0.22	0.18	–	0.19	0.29	0.99
Се (2019)	IU X-Ray	0.44	0.34	0.24	0.18	–	0.35	0.47
Ван (2020)	IU X-Ray	0.50	0.33	0.24	0.18	–	0.36	0.33
	CX-CHR	0.71	0.64	0.59	0.55	–	0.68	0.33
Ли (2018)	IU X-Ray	0.48	0.33	0.23	0.16	–	0.34	0.28
	ChexPert	0.67	0.59	0.53	0.47	–	0.62	0.29

Методы, использующие собственные датасеты, демонстрируют умеренные показатели BLEU и ROUGE-L, однако результаты CIDEr варьируются. Наборы, такие как IU X-Ray, дают более низкие значения, возможно, из-за ограничений в качестве разметки. Использование специализированных датасетов (CX-CHR, ChexPert) приводит к существенно более высоким CIDEr, что свидетельствует о лучшей клинической релевантности.

Методы с высокими BLEU-значениями (например, Шин (2016) [6]) могут иметь проблемы с генерацией связного текста, если более высокоуровневые n-граммы отсутствуют (BLEU-3, BLEU-4 равны нулю). Гибридные методы (например, Ван (2020) [7] на наборе CX-CHR) демонстрируют значительные улучшения по метрике CIDEr, что указывает на их способность генерировать текст, максимально приближенный к клиническим стандартам.

Высокий результат по метрике CIDEr (0.99) у Цзен (2018) можно объяснить рядом особенностей их подхода и используемого датасета. В его работе использовался собственный набор данных ультразвуковых изображений, масштабированный до примерно 3 042 образцов за счёт элементарных аугментаций (флип, вращение) и feature-wise processing. При этом

описания к изображениям создавались экспертами по единой схеме, что приводило к ограниченному лексическому разнообразию и высокой консистентности эталонных текстов.

Важно отметить, что разнообразие результатов указывает на то, что эффективность каждой модели зависит не только от архитектуры, но и от качества используемого набора, что подчеркивает необходимость комплексного подхода к оценке MIC-систем. Несмотря на неплохие показатели, следует отметить, что каждая из методик имеет свои ограничения. Например, шаблонные и поисковые подходы обладают высокой интерпретируемостью, но страдают от жесткости и неадаптивности. Генеративные модели, напротив, могут создавать более связный текст, однако их успешность во многом зависит от объема качественных данных, что остается серьезным препятствием [4]. Таким образом, комбинирование методов позволяет компенсировать слабости каждого отдельного подхода и приводит к наиболее стабильным результатам.

Заключение

В работе проведен детальный обзор современных методов генерации подписей к медицинским изображениям. Гибридные архитектуры, сочетающие подходы на основе поиска и генеративные модели, демонстрируют высокий потенциал для клинического применения за счет повышения точности и адаптивности описаний. Основные вызовы, такие как ограниченность качественных данных и сложность генерации связного текста, требуют дальнейших исследований, включая развитие self-supervised методов и интеграцию экспертных знаний [4, 5]. Дальнейшие разработки в этой области способствуют созданию устойчивых и адаптивных MIC-систем, способных эффективно поддерживать клиническую диагностику.

Список использованных источников

1. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering / P. Anderson, X. He, C. Buehler [и др.] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – Salt Lake City, UT: IEEE, 2018. – С. 6077-6086. – URL: ieeexplore.ieee.org/document/8578734/ (date of access: 03.01.2025).

2. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Show, Attend and Tell / K. Xu, J. Ba, R. Kiros, [и др.] arXiv:1502.03044 [cs]. – arXiv, 2016. – URL: arxiv.org/abs/1502.03044 (дата обращения: 23.12.2024).

3. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. Knowing When to Look / J. Lu, C. Xiong, D. Parikh, R. Socher arXiv:1612.01887 [cs]. – arXiv, 2017. – URL: arxiv.org/abs/1612.01887 (date of access: 23.12.2024).

4. Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains / P. Chambon, C. Bluethgen, C. P. Langlotz, A. Chaudhari arXiv:2210.04133 [cs]. – arXiv, 2022. – URL: http://arxiv.org/abs/2210.04133 (date of access: 02.04.2025).

5. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation / C.Y. Li, X. Liang, Z. Hu, E. P. Xing arXiv:1805.08298 [cs]. – arXiv, 2018. – URL: http://arxiv.org/abs/1805.08298 (date of access: 10.11.2024).

6. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning / H.-C. Shin, H. R. Roth, M. Gao [и др.] // IEEE Transactions on Medical Imaging. – 2016. – Т. 35. – Deep Convolutional Neural Networks for Computer-Aided Detection. – № 5. – С. 1285-1298. – URL: https://ieeexplore.ieee.org/document/7404017/ (дата обращения: 10.11.2024).

7. Unifying Relational Sentence Generation and Retrieval for Medical Image Report Composition / F. Wang, X. Liang, L. Xu, L. Lin arXiv:2101.03287 [cs]. – arXiv, 2021. – URL: http://arxiv.org/abs/2101.03287 (date of access: 03.01.2025).