Дообучение визуально-языковой модели с помощью метода LoRA для генерации структурированных клинических отчетов по рентгенограммам легких

В. О. Сечин
МФТИ
sechin.vo@phystech.edu

А. Н. Аверкин
ФИЦ ИУ РАН
averkin2003@inbox.ru

Е. Н. Волков
ФИЦ ИУ РАН
envolkoff@gmail.com

Аннотация. **В данной работе описывается методика дообучения визуально-языковой модели для автоматической генерации структурированных клинических отчетов по рентгеновским снимкам легких. С использованием метода LoRA модель адаптируется к специфике задачи, что позволяет формировать ответ в строго заданном формате с полями: метки изображения, пол пациента, возраст и описание изображения с диагностической информацией. Представлено описание архитектуры модели, используемого датасета и особенностей процесса дообучения с применением 4-битной квантизации. Экспериментальные результаты демонстрируют значительное улучшение качества генерируемого текста по ряду метрик, что подтверждает эффективность предложенного подхода для автоматизации составления клинических отчетов.**

Ключевые слова: искусственный интеллект, визуально-языковая модель; дообучение; LoRA; клинический отчет; рентгеновское исследование; квантизация; медицинская диагностика

I. Введение

Автоматизация процесса создания клинических отчетов по медицинским изображениям является актуальной задачей, решение которой способно повысить эффективность работы врачей-рентгенологов и снизить вероятность ошибок, связанных с человеческим фактором. В последние годы значительный прогресс в области обработки естественного языка и компьютерного зрения привел к появлению визуально-языковых моделей (VLM), способных генерировать текстовые описания по изображениям. Однако, прямое применение существующих VLM для генерации клинических отчетов по рентгенограммам легких часто оказывается неэффективным, поскольку требует формирования текста в строго заданном, структурированном формате, содержащем не только описание изображения, но и метаданные пациента. [1]

Настоящая работа посвящена решению этой проблемы путем дообучения визуально-языковой модели с использованием метода Low-Rank Adaptation (LoRA), позволяющего адаптировать модель к специфике задачи при ограниченных вычислительных ресурсах. LoRA вносит небольшие изменения в веса предобученной модели, фокусируясь на адаптации к целевому домену, в данном случае – к генерации структурированных клинических отчетов, включающих поля: «метки изображения», «пол пациента», «возраст» и «описание изображения с диагностической информацией». [2]

Исследование выполнено за счет гранта Российского научного фонда (проект № 24-21-00330), https://rscf.ru/project/24-21-00330/.

II. Методы

A. Постановка задачи

Цель данного исследования заключается в дообучении большой визуально-языковой модели с применением метода LoRA (Low-Rank Adaptation) и для генерации клинических отчетов по рентгеновским снимкам легких. Особенность поставленной задачи состоит в необходимости формирования ответа в строго заданном формате, который представлен в табл. I.

ТАБЛИЦА I. Структура вывода ответа моделью ИИ

Наименование признака	**Значение ответа**
Image Labels	[thorax, radiology, etc.]
Gender	[Male/Female]
Age	[numeric value]
Image Caption	[detailed description of the image with medical diagnosis]

Выходной текст должен содержать именно следующие поля: Image Labels, Gender, Age и Image Caption, разделённые символом "|". На вход же модели подаётся описание клинического случая и изображение рентгеновского снимка лёгких. Основная гипотеза заключается в том, что дообучение с использованием LoRA позволяет адаптировать модель к специфике медицинского языка и визуального анализа, обеспечивая генерацию как структурированного, так и содержательного клинического отчета.

B. Архитектура модели ИИ

В качестве визуально-языковой модели была выбрана модель SmolVLM-Instruct [3], разработана командой Hugging Face. Схема модели представлена на рис. 1.

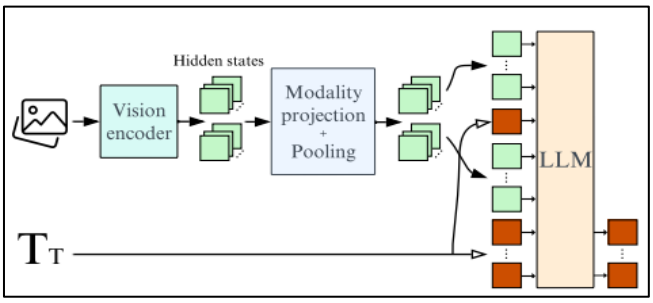

Рис. 1. Архитектура модели SmolVLM-Instruct [3]

Компоненты модели включают Vision Encoder, предназначенный для извлечения высокоуровневых визуальных признаков из входных изображений, и Text Decoder, языковой модуль, основанный на трансформерной архитектуре, который принимает текстовый ввод и интегрирует визуальные токены. Механизм интеграции визуальной и текстовой информации в SmolVLM-Instruct основан на преобразовании визуальных признаков в последовательность токенов посредством специальных «modality projection» слоев, существующих для каждой модальности данных (текст, изображение) для преобразования представлений данных в единое, общее пространство представлений:

- Visual Projection: Визуальные признаки, полученные от ViT, передаются через слой проекции визуальной модальности. Этот слой (обычно линейный) преобразует высокоразмерные визуальные признаки в векторы, которые по размерности и семантике соответствуют пространству представлений языковой модели.
- Text Projection в SmolVLM-Instruct акцент делается на визуальной проекции, часто и текстовые эмбеддинги (перед подачей в трансформер) могут проходить через неявное проецирование, например, за счет самих эмбеддингов или первого слоя трансформера. Это помогает адаптировать предобученные текстовые представления к совместной работе с визуальными данными.

C. *Набор данных*

Для проведения экспериментов по дообучению визуально-языковой модели был использован датасет MultiCaRe [4], представляющий собой многоязычный набор данных для генерации структурированных клинических отчетов по рентгенограммам грудной клетки. MultiCaRe является расширением набора данных MIMIC-CXR, одного из крупнейших публично доступных наборов данных, содержащих рентгеновские снимки грудной клетки и соответствующие им радиологические отчеты.

Помимо оригинальных англоязычных отчетов из MIMIC-CXR, MultiCaRe включает в себя переводы этих отчетов на испанский и португальский языки, выполненные с использованием машинного перевода. Однако, в рамках данного исследования использовалась только англоязычная часть датасета. Общий объем англоязычной части составляет более 370 тысяч пар «изображение-отчет». Для обучения и валидации модели данные были разделены на обучающую, валидационную и тестовую выборки в соотношении 70/10/20, с учетом стратификации по идентификаторам пациентов, чтобы избежать утечки данных.

D. *Метрики оценки качества*

Для оценки качества генерируемых структурированных клинических отчетов использовался комплекс метрик, широко применяемых в задачах машинного перевода и генерации текста. Эти метрики позволяют оценить степень соответствия сгенерированного текста эталонным отчетам, как с точки зрения лексического перекрытия, так и семантической близости элементов текста.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) – это семейство метрик, оценивающих качество текста путем сравнения n-грамм сгенерированного текста и эталонных текстов. Использовались следующие метрики из этого семейства: ROUGE-1 – измеряет перекрытие униграмм (отдельных слов) между сгенерированным и эталонным текстами; ROUGE-2 – измеряет перекрытие биграмм (последовательностей из двух слов); ROUGE-L – основана на самой длинной общей подпоследовательности (Longest Common Subsequence, LCS) между сгенерированным и эталонным текстами. Учитывает структуру предложений и не требует точного совпадения n-грамм. [5]

BLEU (Bilingual Evaluation Understudy) – метрика, оценивающая точность совпадения n-грамм (от 1 до 4) сгенерированного текста с эталонными текстами. Использует модифицированную точность n-грамм, чтобы избежать завышения оценки для текстов, содержащих повторяющиеся слова. [6]

METEOR (Metric for Evaluation of Translation with Explicit ORdering) – метрика, основанная на взвешенном среднем гармоническом точности и полноты униграмм. Учитывает синонимию и стемминг, а также использует штраф за неправильный порядок слов, что делает ее более устойчивой к перефразированию, чем BLEU. [7]

BERTScore – это метрика, использующая предобученную модель BERT (Bidirectional Encoder Representations from Transformers) для вычисления косинусного сходства между векторными представлениями слов сгенерированного и эталонного текстов. BERTScore учитывает контекст и семантику слов, что позволяет более точно оценить смысловую близость текстов. [8]

В работе используются такой разнообразный набор метрик для того, чтобы иметь возможность изучить с разных сторон качество полученного результата. Это очень важно в силу того, что постановка задачи требует как получения структурированного текста, так и текста, верно отражающего диагноз.

E. *Дообучение*

В работе были использованы данные, полученные после обработки части исходного датасета, где для каждой записи заданы ключевые поля: clinical_case

(подробное описание клинического случая), image_path (путь к файлу с рентгеновским снимком) и expected_output - эталонный ответ в заданном формате, сформированный на основании необходимых полей из датасета

Формирование входного текста (промта) включало указание инструкции с описанием ожидаемой структуры ответа, текста клинического случая clinical_case. Входной текст был ограничен по длине 500 символами для ускорения процесса обучения модели.

Базовая инструкция определяла, что итоговый ответ должен содержать четыре обязательных поля: Image Labels, Gender, Age и Image Caption, а итоговый промт представлял собой конкатенацию инструкции и текста клинического случая, направляя модель на генерацию ответа, строго соответствующего заданному шаблону.

Вместе с промтом моделе подавалось изображение клинического случая. В рамках исследования был применен подход дообучения модели, сочетающий технику LoRA с 4-битной квантизацией.

В свою очередь, применение 4-битной квантизации было реализовано через BitsAndBytesConfig с следующими техническими параметрами, показанными в табл. II.

ТАБЛИЦА II. ПАРАМЕТРЫ КВАНТИЗАЦИИ

```
bnb_config = BitsAndBytesConfig(
  load_in_4bit=True,
  bnb_4bit_use_double_quant=True,
  bnb_4bit_quant_type="nf4",
  bnb_4bit_compute_dtype=torch.bfloat16)
```

Данная конфигурация снизила требования к памяти GPU примерно в 8 раз по сравнению с полной FP32-моделью, что позволило разместить модель на одной NVIDIA A100 (40GB).

Далее была выбрана конфигурация LoRA для дообучения модели. Использовались следующие компоненты:

- q_proj, k_proj, v_proj: проекции запросов, ключей и значений в механизме self-attention, ответственные за извлечение релевантных признаков из входных данных;
- o_proj: выходная проекция после механизма внимания, интегрирующая контекстуальную информацию;
- gate_proj: модуль, управляющий потоком информации через активационные функции в MLP-блоках;
- up_proj, down_proj: проекции, контролирующие сжатие и восстановление размерности в Feed-Forward Network (FFN) компонентах.

Эта конфигурация привела к тому, что из 2.26 млрд параметров базовой модели обучению подверглись только 10,536,960 параметров (0.47 %), что существенно сократило вычислительные затраты и объем необходимой памяти.

III. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Из результатов, представленных в табл. III, следует, что модель после обучения показывает улучшения по всем ключевым метрикам.

ТАБЛИЦА III. РЕЗУЛЬТАТЫ МОДЕЛИ SMOLVLM-INSTRUCT

Метрика	**Базовая модель**	**Дообученная модель**	**Относительное изменение, %**
ROUGE-1	0,15492	0,58234	275,9
ROUGE-2	0,04685	0,54041	1053,41
ROUGE-L	0,12845	0,58047	351,91
BLEU	0,01942	0,37854	1849,44
METEOR	0,14199	0,45004	216,96
BERTScore Precision	0,82858	0,97421	17,58
BERTScore Recall	0,81384	0,90436	11,12
BERTScore F1	0,82036	0,9375	14,28

Резкий рост BLEU указывает на то, что модель после дообучения способна генерировать текст, максимально приближенный по последовательности слов к эталонному образцу, что особенно ценно при автоматизации составления отчетов, где важна единообразность. [9]

Улучшение METEOR и BERTScore подтверждает, что модель не просто запоминает шаблон, а действительно понимает контекст клинического случая и генерирует текст, который не только соответствует требуемой структуре, но и обладает правильной семантикой, критичной для медицинской интерпретации.

Вариант вывода автоматического структурированного описания рентгенограммы грудной клетки представлен на рис. 2.

Клинический случай:

A 29 years old female, physician by profession presented to the emergency department with a history of aggressive vomiting five weeks back followed by left upper abdominal, a single episode of loose motion, subcostal pain radiating to left shoulder associated with shortness of breath (SOB) and was unable to take full inspiration. The patient has a history of heartburn, early satiety, indigestion, and food regurgitation six years ago and diagnosed and managed as gastroesophageal reflux disease in her native country. The primary evaluation shows a toxic looking afebrile patient with vitals as; respiratory rate-27/min, pulse 87/min. The patient did well in her follow-up period. To the best of our knowledge, it is the first reported case of Bochdalek hernia associated with the retrocardiac spleen in an adult female in the published literature.

Ожидаемый ответ:

Image Labels: thorax, radiology, frontal, x_ray |
Gender: Female |
Age: 29.0 |
Image Caption: Pre-op x-ray chest with Chilaiditis Sign.

Полученный ответ:

Image Labels: thorax, radiology, frontal, x_ray |
Gender: Female |
Age: 29.0 |
Image Caption: Chest X-ray showing right upper lobe consolidation.

Рис. 2. Вывод результата

IV. Заключение

В данной работе была предложена и реализована методика дообучения визуально-языковой модели для генерации структурированных клинических отчетов по рентгеновским снимкам легких с применением метода LoRA. Разработанный подход позволил адаптировать базовую модель к специфике медицинского языка и визуального анализа, обеспечивая вывод текста в строго заданном формате. Применение 4-битной квантизации существенно снизило требования к вычислительным ресурсам, что делает методику привлекательной для практического использования в условиях ограниченной вычислительной мощности. Экспериментальные результаты подтвердили значительное улучшение качества генерируемых отчетов по ряду метрик, что свидетельствует о высокой эффективности предложенного решения. Полученные выводы открывают перспективы для дальнейшей интеграции данной методики в клиническую практику, что может способствовать оптимизации процесса диагностики и повышению точности медицинской документации.

Список литературы

[1] Busch F., Hoffmann L., Dos Santos D.P. et al. Large language models for structured reporting in radiology: past, present, and future //European Radiology. 2024. P. 1-14. DOI: 10.1007/s00330-024-11107-6.

[2] Mao Y., Ge Y., Fan Y. et al. A survey on lora of large language models //Frontiers of Computer Science. 2025. Vol. 19. No. 7. P. 197605. DOI: 10.1007/s11704-024-40663-9.

[3] Laurençon H., Marafioti A., Sanh V. et al. Building and better understanding vision-language models: insights and future directions //Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models. 2024.

[4] Offidani M. A. N., Delrieux C. A. Dataset of clinical cases, images, image labels and captions from open access case reports from PubMed Central (1990–2023) //Data in Brief. 2024. Vol. 52. P. 110008. DOI: 10.1016/j.dib.2023.110008.

[5] Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. 2004. P. 74-81.

[6] Papineni K., Roukos S., Ward T. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. P. 311-318.

[7] Banerjee S., Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments //Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. P. 65-72.

[8] Zhang T., Kishore V., Wu F. et al. Bertscore: Evaluating text generation with bert //arXiv preprint arXiv:1904.09675. 2019.

[9] Mallio C. A., Sertorio A. C., Bernetti C. et al. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing //La radiologia medica. 2023. Vol. 128. No. 7. P. 808-812. DOI: 10.1007/s11547-023-01651-4.