УДК 004.8 https://doi.org/10.15407/pp2025.03.066

М. Г. Шевченко, М.В. Андрощук

МУЛЬТИМОДАЛЬНИЙ RAG З ВИКОРИСТАННЯМ ТЕКСТОВИХ ТА ВІЗУАЛЬНИХ ДАНИХ

Стаття присвячена розробці та дослідженню мультимодальної системи генерації, доповненої пошуком (Retrieval-Augmented Generation), призначеної для аналізу та інтерпретації медичних зображень. Об'єктом дослідження є рентгенівські знімки грудної клітки та відповідні їм радіологічні звіти. Основна мета роботи полягала у створенні системи, здатної виконувати два ключові завдання: генерувати детальний радіологічний звіт для вхідного зображення та надавати точні відповіді на конкретні запитання щодо нього. Додатковою ціллю було демонстрування того, що застосування мультимодального підходу з пошуком релевантної інформації суттєво покращує якість генерації порівняно з використанням великих мультимодальних моделей без компонента пошуку. Для реалізації системи було використано комбінацію сучасних моделей глибокого навчання. За основу для створення векторних представлень текстових та візуальних даних було взято модель BiomedCLIP, яку було додатково навчено на цільовому наборі даних. Функції генератора виконувала велика мовна модель LLaVA-Med 1.5, адаптована для медичної галузі та квантизована для роботи в умовах обмежених обчислювальних ресурсів. Архітектура системи також включає допоміжні класифікатори на основі DenseNet121 для визначення проєкції знімка та ідентифікації наявних клінічних ознак, що дозволило підвищити точність пошуку. У процесі експериментального дослідження було протестовано шість різних конфігурацій розробленої системи. Оцінювання проводилося з використанням низки метрик, зокрема, точності та F1 для задачі відповіді на питання, а також BLEU, ROUGE, F1-CheXbert та F1-RadGraph для оцінки якості згенерованих звітів. Результати тестування продемонстрували значну перевагу всіх конфігурацій системи над базовою моделлю-генератором. Найкращі результати показала конфігурація, що використовує класифікатори проєкції та клінічних ознак із вимогою точного збігу знайдених патологій. Дослідження підтвердило, що інтеграція механізму пошуку релевантних даних значно підвищує структурну та змістовну якість генерованих текстових описів для медичних зображень.

Ключові слова: генерація доповнена пошуком, мультимодальність, медичні зображення, генерація звітів, глибинне навчання, великі мовні моделі.

M.H. Shevchenko, M.V. Androshchuk

MULTIMODAL RAG USING TEXT AND VISUAL DATA

This paper presents the development and investigation of a multimodal Retrieval-Augmented Generation system designed for the analysis and interpretation of medical images. The research focuses on chest X-ray images and their corresponding radiology reports. The primary goal was to create a system capable of performing two key tasks: generating a detailed radiology report for an input image and providing accurate answers to specific questions about it. A secondary goal was to demonstrate that employing a multimodal retrieval-augmented approach significantly improves generation quality compared to using large multimodal models without a retrieval component. The system's implementation utilizes a combination of state-of-the-art deep learning models. The BiomedCLIP model, fine-tuned on the target dataset, was used to generate vector embeddings for both text and visual data. The generator component is based on the large language model LLaVA-Med 1.5, which is adapted for the medical domain and quantized to operate under limited computational resources. The system architecture also includes auxiliary classifiers based on DenseNet121 to determine the image projection and identify clinical findings, thereby enhancing retrieval accuracy. The experimental evaluation involved testing six different configurations of the developed system. The evaluation was conducted using a range of metrics, including accuracy and F1-score for the question-answering task, as well as BLEU, ROUGE, F1-CheXbert, and F1-RadGraph for assessing the quality of the generated reports. The test results demonstrated a significant advantage of all system configurations over the baseline generator model. The best results were achieved by the configuration that utilizes projection and clinical finding classifiers with an exact match requirement for the identified pathologies. The study confirmed that integrating a relevant data retrieval mechanism significantly enhances both the structural and semantic quality of the generated textual descriptions for medical images.

Keywords: Retrieval-Augmented Generation, multimodality, medical imaging, report generation, deep learning, large language models.

Вступ

Мультимодальний RAG — це RAG, який використовує не один, а кілька типів даних для ретривера(англ. retriever) й генератора одночасно, наприклад: текст, зображення, відео, аудіо тощо. Застосування кількох типів даних часто надає генератору більше інформації, завдяки чому генеруються вичерпніші та релавантніші відповіді на запит користувача [34].

Мультимодальний RAG із використанням текстових та візуальних даних — це найпоширеніший вид мультимодального RAG. Він застосовує тексти, зображення та відео для генерації відповіді. Найчастіше використовують саме зображення, а не відео, оскільки робота з відео є значно складнішою через великий обсяг, необхідність обробки послідовностей кадрів, а також складність виокремлення релевантної інформації з часових даних. Застосування мультимодального RAG із використанням текстових та візуальних даних часто підвищує коректність і релевантність відповідей генераторів [3], оскільки велика кількість інформації зберігається саме у візуальному вигляді й не має текстового відповідника. Наприклад, така інформація, як просторові відношення між об'єктами, емоційний контекст, специфічні деталі зображень (текстури, кольори, форми), а також складні візуальні структури (графіки, діаграми, медичні знімки), часто не мають точного текстового відповідника або потребують значного спрощення у разі опису словами.

В роботі для реалізації мультимодального RAG було обрано медичну предметну сферу, зокрема, аналізу та інтерпретації рентгенівських знімків грудної клітки та їхніх звітів. Медична предметна сфера є доволі популярним напрямком створення RAG. Вибір такої предметної сфери надає широкі можливості для створення різних RAG систем та дозволить оцінити як використання мультимодального RAG покращує роботу існуючих систем. Іншим фактором вибору такої предметної сфери є достатня кількість доступних наборів даних у вільному доступі, що не завжди властиво іншим предметним сферам. Найчастіше такі набори даних включають певну кількість рентгенівських знімків грудних кліток та звітів у текстовому форматі для кожного знімку.

Мультимодальна RAG-система, створена в роботі, підтримує два сценарії використання. Перший сценарій використання — це генерація звіту на вхідний рентгенівський знімок, щодо наявності паталогій, хвороб тощо. Інший сценарій використання — це генерація відповіді на вхідне питання щодо певного рентгенівського знімку. Ці сценарії є типовими для мультимодального RAG. Також такі сценарії використання є стандартними для використання в медичній сфері. Для реалізації даних сценаріїв використання, ретривер буде знаходити релевантні звіти до вхідного зображення. Потім знайдені звіти будуть разом із вхідним зображенням та питанням надаватися генератору для генерації звіту або відповіді на вхідне питання.

Під час розробки мультимодальної RAG-системи з використанням текстових та візуальних даних для роботи з рентгенівськими знімками грудної клітки та звітами до них були використанні набори даних: MIMIC-CXR [24], MIMIC-CXR-JPG [14], CheXpert [16] та IU-Xray [9].

Методологія дослідження

Рентгенівські знімки грудей бувають двох типів: передні і бічні. Передні — зроблені спереду грудної клітки. Бічні — зроблені збоку грудної клітки. Передні і бічні знімки значно відрізняються один від одного, бо на передніх можна бачити паталогії і клінічні ознаки, яких не видно на бічних і навпаки. Через це було вирішено реалізувати дві моделі, які створюють індекси (далі — моделі індексації): одна модель буде працювати з передніми зображеннями і їхніми звітами, а інша з бічними зображеннями і звітами. Для того, щоб ретривер міг ідентифікувати, яке зображення — переднє, а яке — бічне, було створено додаткову модель для ідентифікації сторони (далі — класифікатор сторін). Також для покращення ретривера, щоб він знаходив

найбільш релевантні звіти, було розроблено модель для створення метаданих, а саме інформацію про наявність певної клінічної ознаки на кожному зображенні (далі — класифікатор хвороб).

Як модель для створення індексів була вибрана модель BiomedClip [33]. BiomedClip — це модель, створена на основі CLIP [25], для задач біомедичного бачення. BiomedClip був натренований на наборі даних PMC-15M, створений дослідниками BiomedClip. PMC-15M складається із 15 мільйонів пар підпис-картинок з різних медичних галузей: офтальмологія, стоматологія, рентгенографія та інші. Також дослідники BiomedClip створили модель PubMedBERT та свій власний Vision Transformer, які використовуються як кодувальник тексту та кодувальник зображень відповідно до архітектури CLIP.

Хоча BiomedClip натренований на медичних даних, він не навчений безпосередньо на тих даних, які планується використовувати в ретривері для релевантного пошуку. Також варто враховувати, що BiomedClip натренований на широкому спектрі медичних галузей, а нам потрібно лише рентгенівські знімки грудей. Враховуючи попередні зауваження, зрозуміло, що просте використання BiomedClip як моделі для створення індексів, може часто призводити до нерелевантних результатів [32].

Для покращення роботи BiomedCLIP застосовують передавальне навчання. У межах цієї роботи BiomedCLIP було донавчено на даних MIMIC-CXR, оскільки саме ці дані надалі використовуються ретривером для релевантного пошуку.

Для донавчання був створений програмний застосунок із використанням бібліотеки PyTorch. Як функція втрат [2] використовується перехрестна втрата ентропії (англ. cross entropy loss) [21] та AdamW [20] як оптимізатор [29].

В результаті було донавчено три окремі моделі BiomedClip.

Перша — для індексації передніх рентгенівських знімків і їхніх звітів (далі — модель індексації передніх знімків). Вона була донавчена на 100000 випадково вибраних пар передніх рентгенгенівських знімків та їхніх звітів з MIMIC-CXR. Модель донавчалася протягом 100 епох [5] на відеокарті Nvidia GeForce RTX 3090 24 ГБ. Для донавчання використовувалися такі гіперпараметри [7]: розмір партії (англ. batch size) [4] — 64, темп навчання (англ. learning rate) [1] — $0.00005\sqrt{8}$. Значення втрат на останній епосі склало: 0.0336.

Друга — для індексації бічних рентгенівських знімків і їхніх звітів (далі — модель індексації бічних знімків). Вона була донавчена на 50000 випадково вибраних пар бічних рентгенгенівських знімків та їхніх звітів з MIMIC-CXR. Модель донавчалася протягом 100 епох на відеокарті Nvidia GeForce RTX 2060 6 ГБ. Для донавчання використовувалися такі гіперпараметри: розмір партії — 8, темп навчання — 0.00005. Значення втрат на останній епосі склало: 0.0306.

Третя — для індексації передніх та бічних рентгенівських знімків і їхніх звітів (далі — загальна модель індексації). Вона була донавчена на 150000 випадково вибраних пар передніх та бічних рентгенгенівських знімків та їхніх звітів з MIMIC-CXR. Модель донавчалася протягом 95 епох на відеокарті Nvidia GeForce RTX 3090 24 ГБ. Для донавчання використовувалися такі гіперпараметри: розмір партії — 64, темп навчання — $0.00005\sqrt{8}$. Значення втрат на останній епосі склало: 0.0337.

Для того, щоб ретривер міг правильно вибирати модель для класифікації передніх чи бічних знімків, він повинен вміти класифікувати зображення на передні і бічні. Це класична задача класифікації зображень [11]. Для такої задачі найчастіше використовуються нейронні мережі. Тренування класифікатора сторін відбувалося на наборі даних CheXpert, бо він містить інформацію про сторону кожного рентгенівського знімка. За нейронну мережу для дотренування була обрана модель DenseNet121, яка також використовувалася в дослідженні набору даних CheXpert і показала кращі результати, ніж інші моделі в дослідженні. DenseNet121 — це згорткова нейронна мережа [13], яка складається з 121 шару, кожен пов'язаний один з одним.

Для тренування використовувалася передавальне навчання. Для цього бралася модель DenseNet121, яка була попередньо натренована на наборі даних ImageNet [10]. Як функція втрат використовувалася перехрестна втрата ентропії та стохастичний градієнтний спуск [27] як оптимізатор.

Тренування класифікатора сторін відбувалося протягом 10 епох на відеокарті Nvidia GeForce RTX 2060 ГБ. Для тренування використовувалося 20000 випадково вибраних рентенівських знімків з набору даних CheXpert. Тестування моделі після кожної епохи відбувалося на 2000 рентгенівських знімках. Уже після першої епохи тренування, точність визначення сторони зображення на тренувальних та тестових даних склала 100%, тому тренування було закінчене достроково на 5-ій епосі. Для донавчання використовувалися такі гіперпараметри: розмір партії — 32, темп навчання — 0.001.

Для покращення релевантності знайдених результатів ретривером було створено класифікатор хвороб. Класифікатор хвороб визначає, які клінічні ознаки із заданого списку наявні на зображенні, а які ні. Це задача класифікації за багатьма класами (англ. multi-label classification) [26], задача в якій один об’єкт може відповідати декільком класам одночасно. Ця задача схожа на класифікацію зображень, використану в класифікаторі сторін. Для її розв’язання також найчастіше використовують нейронні мережі.

Для створення класифікатора хвороб було використано модель DenseNet121, як в класифікаторі сторін. Для тренування було використано набір даних CheXpert, тому що він містить інформацію про наявність певної ознаки зі списку 14 клінічних ознак для кожного зображення. Кожна ознака для кожного зображення в наборі даних CheXpert може мати одне з 4-ох значень: «наявна», «відсутня», «не зазначена», «не зрозуміло». В тренуванні значення «не зазначена» і «не зрозуміло» будуть сприйматися як один клас, тому що важлива лише наявність або відсутність ознаки. Вихідним значенням тренувальної моделі DenseNet121 є вектор із 14 чисел, кожне з яких відповідає за наявність певної хвороби. Значення чисел вектора знаходяться в межах від нуля включно до одиниці включно, числа більше 0.7 будуть сприйматися як наявність хвороби, числа менше 0.3 будуть сприйматися як її відсутність. Усе, що між 0.3 та 0.7, буде відповідати значенням «не зазначена» або «не відомо». Для тренування, як в класифікаторі сторін, було використане передавальне навчання та модель DenseNet121, яка була попередньо натренована на наборі даних ImageNet. Як функція втрат використовувалася перехрестна втрата ентропії та Adam [17] як оптимізатор.

Тренування класифікатора хвороб відбувалося протягом 50 епох на відеокарті Nvidia GeForce RTX 2060 6 ГБ. Для тренування використовувалося 156553 випадково вибраних рентенівських знімків з набору даних CheXpert. Тестування моделі після кожної епохи відбувалося на 67095 рентгенівських знімках. Після останньої епохи точність визначення наявості та відсутності хвороб на тестових даних склала 86,71%, що майже відповідає результатам у досліджені CheXpert. Для донавчання використовувалися такі гіперпараметри: розмір партії — 32, темп навчання — 0.0001.

Для збереження індексів, звітів, метаданих та виконання пошуку за індексами, як векторна база даних була взята ChromaDb, котра повністю відповідає вимогам для створення системи. В ChromaDb дані зберігаються в колекціях аналогічно до таблиць в реляційних базах даних. Кожен запис в колекції повинен мати індекс, за яким буде відбуватися пошук, та якісь дані. Також запис може містити додаткові дані — метадані, за якими може відбуватися додатковий пошук. Індекси всіх записів в колекції повинні бути однієї розмірності.

У межах реалізації системи було створено три окремі колекції в ChromaDb, кожна з яких відповідала певному типу рентгенівських знімків і використовувала відповідну модель індексації. Для створення колекцій використовувся набір даних MIMIC-CXR. Перша колекція була створена за допомогою моделі індексації передніх знімків на 100000 випадково вибраних пар передніх ретгенівських знімків та їхніх звітів. Друга була створена за допомогою

моделі індексації бічних знімків на 50000 випадково вибраних пар бічних ретгенівських знімків та їхніх звітів. Третя була створена за допомогою загальної моделі індексації на 150000 випадково вибраних пар передніх і бічних ретгенівських знімків та їхніх звітів.

Генератором була обрана велика мовна модель LLaVA-Med 1.5 [18]. Ця модель є дотренованою версією LLaVA 1.5 на медичних даних, розробленою дослідниками з Microsoft [18]. У роботі використовувалася LLaVA-Med, яка містить 7 мільярдів параметрів. Оскільки для параметрів в нейронних мережах використовується 32-бітні числа з рухомою комою, то для запуску LLaVA-Med 1.5 на відеокарті треба понад 26 ГБ відеопам'яті. Під час виконання роботи не було доступу до відеокарт з такою кількістю відеопам'яті. Для того, щоб запускати нейронні мережі на відеокартах з меншою кількісю відеопам'яті, використовують техніку квантинізації [22]. Квантинізація дозволяє зменшити кількість пам'яті, яку займає модель, за допомогою представлення параметрів моделі 16-ти, 8-ми або 4-ма бітними числами з рухомою комою. Звісно, використання квантинізації може призвести до деградації моделі, але за дослідженням це або не відбувається взагалі, або зміни не суттєві. Для запуску LLaVA-Med 1.5 на Nvidia GeForce RTX 2060 6 ГБ була використана техніка квантинізації у режимі 4-біт. В результаті модель займає приблизно 5 ГБ відеопам'яті.

Для демонстрації роботи системи у процесі генерування звіту на вхідне зображення було використано випадковий рентгенівський знімок з набору даних MIMIC-CXR. В результаті мультимодальний RAG згенерував такий звіт: «The chest X-ray image shows a large right pleural effusion, which is an abnormal accumulation of fluid in the pleural space surrounding the lungs. This finding is concerning for pneumonia, which is an infection that causes inflammation in the air sacs of the lungs. It is important to consult a healthcare professional for a thorough evaluation and proper diagnosis of the underlying cause of these findings.». Еталонний звіт із набору даних для цього ж самого зображення: «impression: Increased right pleural loculated effusion with chest tube in place. Increasing consolidation in the right lung is concerning for pneumonia. Findings: PA and lateral views of the chest provided. Port-A-Cath is unchanged in position with its tip positioned in the expected location of the mid SVC. A right pleural drain is in place with increased opacity in the right lung and probable increase in size of the loculated right pleural effusion. Findings are concerning for a superimposed consolidation/pneumonia. The left lung remains essentially clear. The heart is difficult to assess given the effacement of the right heart border. The prominence of the mediastinum may reflect in part adjacent loculated pleural fluid. No pneumothorax is seen.».

Для демонстрації роботи системи як оцінювача конкретного питання за вхідне зображення було використано випадковий рентгенівський знімок з набору даних MIMIC-CXR Вхідним питанням до системи було: «Is the cardiomediastinal silhouette within normal limits?». В результаті система згенерувала таку відповідь: «Yes, the cardiomediastinal silhouette appears to be within normal limits in the image.». Еталонна відповідь на це питання до цього ж самого зображення з набору даних: «Yes».

Тестування системи

Тестування створеної системи проводилося на шістьох конфігураціях мультимодального RAG, кожна з яких була налаштована на знаходження 5-ти релевантних звітів до кожного запиту. Як зразок порівняння конфігурацій також була протестована LLaVA-Med 1.5 без використання ретриверу.

В рамках створення конфігурацій системи використовуються поняття «точне співпадіння хвороб» і «неточне співпадіння хвороб», позначаючи точне співпадіння хвороб у сховищі відносно вхідного зображення, чи щоб було хоча б одне співпадіння хвороб відповідно.

Ретривер першої конфігурації мультимодальної RAG-системи (рис.1) включає класифікатор сторін, модель індексації передніх знімків, модель індексації бічних знімків і класифікатор хвороб. Ретривер використовує неточне співпадіння хвороб.

Рис. 1. Конфігурація 1 (з класифікатором сторін, з класифікатором хвороб, неточне співпадіння хвороб)

Ретривер другої кофігурації мультимодальної RAG-системи (Рис.2) майже такий самий, як у конфігурації 1, але використовує точне співпадіння хвороб.

Рис. 2. Конфігурація 2 (з класифікатором сторін, з класифікатором хвороб, точне співпадіння хвороб)

Ретривер третьої кофігурації мультимодальної RAG-системи (рис.2) схожий на ретривер конфігурації 1 і конфігурації 2, але не використовує класифікатор хвороб.

Рис. 3. Конфігурація 3 (з класифікатором сторін, без класифікатора хвороб)

Ретривер четвертої конфігурації мультимодальної RAG-системи (рис.4) включає загальну модель індексації та класифікатор хвороб. Ретривер використовує неточне співпадіння хвороб.

Рис. 4. Конфігурація 4 (без класифікатора сторін, з класифікатором хвороб, неточне співпадіння хвороб)

Ретривер п’ятої кофігурації мультимодальної RAG-системи (рис.5) майже такий самий, як у конфігурації 4, але використовує точне співпадіння хвороб.

Рис. 5. Конфігурація 5 (без класифікатора сторін, з класифікатором хвороб, точне співпадіння хвороб)

Ретривер шостої кофігурації мультимодальної RAG-системи (рис.6) схожий на ретривер конфігурації 4 і конфігурації 5, але не використовує класифікатор хвороб.

Рис. 6. Конфігурація 6 (без класифікатора сторін, без класифікатора хвороб)

Для тестування мультимодальної RAG-системи для відповіді на вхідне запитання і зображення до нього були використані дві вибірки тестових даних, створені дослідниками мультимодального RAG [32]. Кожна вибірка даних містить приблизно 2500 запитань, відповідей і зображень. Одна вибірка містить зображення з набору даних MIMIC-CXR, а інша з набору даних IU-Xray. Дослідники, які створили вибірки даних, використовували велику мовну модель ChatGPT-4 для створення запитань і відповідей до рентгенівських знімків. Потім дослідники власноруч відфільтрували і перевірили кожне питання і відповідь. Усі запитання створені таким чином, щоб вони мали тільки два варіанта відповіді: «так» або «ні».

Для тестування системи відповідати на вхідне запитання і зображення до нього було використано дві метрики: метрика точності (англ. accuracy) [12] і метрика F1 (англ. F1-score) [30].

Метрика точності — одна з найпростіших і найпоширеніших оцінок якості

класифікаційних моделей. Вона показує, яка частка всіх передбачень моделі виявилася правильною. В нашому випадку вона покаже яка частка відповідей була правильно визначена. Точність набуває значень від нуля включно до одиниці включно, де нуль означає, що ніяка відповідь не була визначена правильно, а одиниця означає, що всі відповіді були визначені правильно.

Для обчислення метрики F1 треба поділити всі відповіді на позитивні і негативні. Позитивні відповіді — це відповіді, які ми оцінюємо, а негативні — всі інші. F1 показує, наскільки добре модель одночасно уникає хибних спрацьовувань (помилкових позитивних відповідей) та пропусків (хибних негативних відповідей). Вона обчислюється як гармонійне середнє [6] між влучністю [31] (англ. precision, доля правильних позитивних передбачень серед усіх позитивних передбачень) і повнотою [31] (англ. recall, доля правильних позитивних передбачень серед усіх позитивних відповідей), тому низьке значення однієї з цих метрик суттєво знижує F1. F1, як і точність, набуває значень від нуля включно до одиниці включно. Нуль означає, що модель не змогла правильно передбачити жодного позитивного випадку, а одиниця означає ідеальну відповідність. Для обчислення метрики F1 на вибірках даних буде використовуватися зважена F1 (англ. F1-weighted). Зважена F1 обчислюється окремо для кожного типу відповіді, а потім об'єднується, використовуючи кількісне значення кожного типу відповіді.

Тестування відбувалося на 1000 випадково вибраних запитань з кожної вибірки даних. Результати тестування наведені у таблиці 1.

Таблиця 1.

Результати тестування системи відповідати на питання за зображенням

Система	MIMIC-CXR		IU-Xray	
	Точність	Зважена F1	Точність	Зважена F1
LLaVA-Med 1.5	0.717	0.668	0.411	0.418
Конфігурація 1	**0.840**	**0.838**	0.921	0.924
Конфігурація 2	0.844	0.842	**0.909**	**0.913**
Конфігурація 3	0.831	0.830	0.929	0.931
Конфігурація 4	0.766	0.766	0.917	0.919
Конфігурація 5	0.786	0.784	0.918	0.921
Конфігурація 6	0.773	0.773	0.921	0.923

Для тестування можливості мультимодального RAG генерувати звіт до вхідного зображення були використані дві вибірки тестових даних, створені дослідниками мультимодального RAG [32]. Перша вибірка містить 700 зображень та звітів з набору даних MIMIC-CXR, а друга — 1180 зображень та звітів з набору даних IU-Xray. Кожна вибірка сформована так, аби вона містила якнайрізноманітніші звіти. Тестування системи за різними метриками проводилось на 700 зображень та звітів з набору даних MIMIC-CXR (з першої вибірки) та 1000 зображень та звітів з набору даних IU-Xray (з другої вибірки).

Для тестування системи генерувати звіт до вхідного зображення було використано всього чотири метрики: BLEU [23],

ROUGE [19], F1-CheXbert [28], F1-RadGraph [8].

BLEU (Bilingual Evaluation Understudy) — це метрика для порівняння згенерованого чи перекладеного тексту з одним або кількома еталонними варіантами. Для оцінки за метрикою BLEU підраховують кількість однакових послідовностей із *n* слів (*n*-грам), що зустрічаються як у згенерованому, так і в еталонному тексті. Потім обчислюється влучність як відношення числа збігів до загальної кількості n-грам у згенерованому тексті. Зазвичай влучність обраховують для *n*-грам від одного до чотирьох слів. Далі ці влучності об'єднують за допомогою геометричного середнього. Оскільки короткий (порівняно з еталонним) згенерований текст часто має вищу частку збігів, під час розрахунку BLEU застосовують штраф за короткість (англ. brevity penalty). BLEU набуває значень від нуля включно до одиниці включно. Чим більше значення BLEU, тим більше згенерований текст схожий на еталонний. Варто зазначити, що ця метрика оцінює лише збіги слів, а не зміст чи стиль тексту.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) — це набір метрик для порівняння згенерованого чи перекладеного тексту з одним або кількома еталонними варіантами. Набір ROUGE складається з трьох метрик: ROUGE-1, ROUGE-2 і ROUGE-L. Для оцінки за ROUGE-1 підраховують збіги окремих слів між згенерованим і еталонним текстом. Потім обчислюється значення метрики F1 на основі кількості таких збігів. ROUGE-2 обчислюється аналогічно до ROUGE-1, але замість окремих слів рахують двослівні послідовності. Ця метрика дозволяє оцінити, наскільки модель правильно відтворює не лише окремі слова, а й стійкі словосполучення, порядок слів і короткі фрази. Для оцінки за ROUGE-L обраховують довжину найдовшої спільної підпослідовності між згенерованим і еталонним текстом. Ця метрика відображає не лише точні збіги слів, а й загальну послідовність, у якій вони з'являються. Усі метрики ROUGE набувають значень від нуля включно до одиниці включно. Чим більше значення метрики, тим більше згенерований текст схожий на еталонний. Як і BLEU, набір ROUGE не оцінює зміст чи стиль тексту, а лише лексичну подібність. Результати наведені у таблиці 2 і таблиці 3.

Таблиця 2.

Результати тестування системи генерувати звіт на вхідне зображення за допомогою метрики BLEU

Система	MIMIC-CXR	IU-Xray
LLaVA-Med 1.5	0.006204	0.014087
Конфігурація 1	0.025932	0.035426
Конфігурація 2	**0.042499**	0.036582
Конфігурація 3	0.027740	0.034322
Конфігурація 4	0.018719	0.035311
Конфігурація 5	0.022738	**0.037315**
Конфігурація 6	0.018848	0.035051

Таблиця 3.

Результати тестування системи генерувати звіт на вхідне зображення за допомогою метрик ROUGE-1, ROUGE-2, ROUGE-L

Система	MIMIC-CXR			IU-Xray		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
LLaVA-Med 1.5	0.177572	0.017352	0.111342	0.178954	0.024311	0.122759
Конфігурація 1	0.255128	0.063660	0.164584	0.258481	0.054633	0.175839
Конфігурація 2	**0.269676**	**0.079149**	**0.178946**	0.257788	0.054402	0.176659
Конфігурація 3	0.256807	0.067752	0.165909	0.258677	0.053155	0.176821
Конфігурація 4	0.233677	0.052150	0.147039	0.267725	**0.059491**	0.174326
Конфігурація 5	0.244741	0.055802	0.157708	**0.275352**	0.058866	**0.182624**
Конфігурація 6	0.234838	0.053292	0.148318	0.262273	0.055589	0.171794

F1-CheXbert — це метрика, яка оцінює наскільки точно модель відтворює набір клінічних ознак з набору даних CheXpert [14] у радіологічних звітах порівняно з еталонними звітами. F1-CheXbert використовує спеціальну модель CheXbert, яка визначає наявність кожної клінічної ознаки в конкретному звіті. Ця модель натренована на наборі даних CheXpert. Після цього CheXbert застосовують для виявлення клінічних ознак як у згенерованих, так і в еталонних звітах. Для кожної ознаки обчислюється значення метрики F1. Значення F1-CheXbert — це середнє арифметичне значення F1 для всіх ознак, щоб рідкісні, але важливі ознаки мали такий самий вплив на фінальний результат, як і найпоширеніші. F1-CheXbert набуває значень від нуля включно до одиниці включно. Чим більше значення F1-CheXbert, тим більше клінічних ознак із еталонного звіту модель правильно відтворила в згенерованому тексті.

F1-RadGraph — це метрика, яка оцінює, наскільки вдало згенерований радіологічний звіт відтворює структуровану інформацію про клінічні ознаки порівняно з еталонним. F1-RadGraph використовує спеціальну модель RadGraph [15], яка виявляє у звіті клінічні ознаки та зв'язки між ними, формуючи граф. RadGraph застосовують для виявлення клінічних ознак та зв'язків між ними, як у згенерованих, так і в еталонних звітах. Значення F1-RadGraph — це значення F1, а саме F1-мікро (англ. F1-micro), що обчислюється для всіх сутностей і зв'язків разом. F1-RadGraph набуває значень від нуля включно до одиниці включно. Чим більше значення F1-RadGraph, тим більше клінічних ознак і зв'язків між ними з еталонного звіту модель правильно відтворила в згенерованому тексті. Результати наведені у таблиці 4.

Таблиця 4.

Результати тестування системи генерувати звіт на вхідне зображення за допомогою метрики F1-RadGraph і F1-CheXpert

Система	MIMIC-CXR		IU-Xray	
	F1-RadGraph	F1-CheXpert	F1-RadGraph	F1-CheXpert
LLaVA-Med 1.5	0.020411	0.011978	0.059309	0.120694
Конфігурація 1	0.095536	0.266570	0.120658	0.217418
Конфігурація 2	**0.101452**	**0.289483**	0.119638	0.219341
Конфігурація 3	0.097419	0.263669	0.122248	**0.226831**
Конфігурація 4	0.076431	0.221617	0.118745	0.182259
Конфігурація 5	0.079204	0.235491	**0.122624**	0.180509
Конфігурація 6	0.078109	0.214062	0.112216	0.179557

За результатами тестування відповідати на запитання до вхідного зображення конфігурація 2 (з класифікатором сторін, з класифікатором хвороб, точне співпадіння хвороб) найкраще впоралася на наборі даних MIMIC-CXR, а конфігурація 3 (з класифікатором сторін, без класифікатора хвороб) найкраще впоралася на наборі даних IU-Xray. На наборі даних MIMIC-CXR система показала приблизно на 26% кращий результат, ніж LLaVA-Med 1.5 без ретриверу відповідно за зваженою F1 метрикою, а на наборі даних IU-Xray приблизно на 123% краще. Такий великий розрив у результатах можна пояснити тим, що розробники LLaVA-Med 1.5 використовували набір даних MIMIC-CXR для тренування, тому результати LLaVA-Med 1.5 на MIMIC-CXR вищі, ніж на IU-Xray. Низькі результати LLaVA-Med на IU-Xray можна пояснити її використанням у режимі 4-біт. Водночас розроблена система показала найкращі результати саме на IU-Xray.

Тестування розробленої системи генерувати звіт за допомогою метрик, які оцінюють збіги слів, а саме BLEU, ROUGE-1, ROUGE-L показало, що конфігурація 2 (з класифікатором сторін, з класифікатором хвороб, точне співпадіння хвороб) найкраще впоралася на наборі даних MIMIC-CXR, а конфігурація 5 (без класифікатора сторін, з класифікатором хвороб, точне співпадіння хвороб) найкраще впоралася на наборі даних IU-Xray. Результати метрики ROUGE-2 відрізняються тим, що конфігурація 4 (без класифікатора сторін, з класифікатором хвороб, неточне співпадіння хвороб) найкраще управилася на наборі даних IU-Xray. З результатів цих метрик зрозуміло, що система генерує більш схожі звіти на еталонні, ніж просто LLaVA-Med 1.5. Ймовірно, це зумовлено тим, що система надає релевантні звіти генератору, і він намагається створити звіти схожої структури, тоді як використання LLaVA-Med 1.5 без генератора генерує доволі короткі і недеталізовані звіти. Конфігурації без використання класифікатора сторін, а з використанням загальної моделі індексації, показали кращі результати на наборі даних IU-Xray ймовірніше за все через те, що в

цьому наборі даних використовується однаковий звіт до передніх і бічних знімків на відміну від MIMIC-CXR.

Тестування розробленої системи генерувати звіт за допомогою метрики F1-CheXbert, яка оцінює зміст тексту, показало, що конфігурація 2 (з класифікатором сторін, з класифікатором хвороб, точне співпадіння хвороб) найкраще впоралася на наборі даних MIMIC-CXR, а конфігурація 3 (з класифікатором сторін, без класифікатора хвороб) найкраще впоралася на наборі даних IU-Xray. Конфігурація 2 показала більше, ніж вдвічі кращий результат, аніж LLaVA-Med 1.5 без ретриверу на наборі даних MIMIC-CXR, а конфігурація 3 більше, ніж в 1.8 рази кращий результат на наборі даних IU-Xray.

Тестування розробленої системи щодо генерування звіту за допомогою метрики F1-RadGraph, яка також оцінює зміст тексту, показало, що конфігурація 2 (з класифікатором сторін, з класифікатором хвороб, точне співпадіння хвороб) найкраще впоралася на наборі даних MIMIC-CXR, а конфігурація 5 (без класифікатора сторін, з класифікатором хвороб, точне співпадіння хвороб) найкраще управилася на наборі даних IU-Xray. Конфігурація 2 показала приблизно в 5 разів кращий результат, аніж LlaVA-Med без ретриверу на наборі даних MIMIC-CXR, а конфігурація 5 більше ніж вдвічі кращий результат на наборі даних IU-Xray.

Висновки

Загалом результати тестування демонструють, що використання техніки мультимодального RAG суттєво покращує генерацію відповідей і звітів LLaVA-Med 1.5, навіть попри обмеження 4-бітного режиму, за якого без ретривера генерує короткі і недеталізовані відповіді. Релевантні звіти, отримані через ретривер, дозволяють генератору створювати кращі за структурою й змістом відповіді. В більшості випадків кофігурації з використанням класифікатора хвороб, а саме точним співпадінням хвороб, показали найкращі результати. Випадки, коли такі конфігурації давали гірші результати, ймовірно пов'язані з неідеальною точністю класифікатора хвороб. Класифікатор сторін також у більшості сценаріїв покращував результати системи порівняно з конфігураціями без нього. Враховуючи вище сказане, можна зробити висновок, що конфігурація 2 (з класифікатором сторін, з класифікатором хвороб, точне співпадіння хвороб) є найкращою серед інших протестованих.

Під час роботи було створено мультимодальну RAG-систему для аналізу та інтерпретації рентгенівських знімків грудної клітки та їхніх звітів. Система використовує новітні технології та програмні засоби, а саме: PyTorch, LLaVA-Med 1.5, BioMedCLIP, DenseNet121, ChromaDB. Розроблена система здатна відповідати на запитання до рентгенівського знімку та генерувати радіологічний звіт до знімку. Архітектура системи складається з декількох підсистем: різні моделі індексації, класифікатор сторін, класифікатор хвороб, генератор, сховище даних. Було запропоновано декілька конфігурацій системи для їхнього тестування.

Також проведене тестування шести конфігурацій створеної мультимодальної RAG-системи. Для порівняння була протестована LLaVA-Med 1.5 без ретриверу. Генерацію відповідей на запитання до зображення оцінювали за метриками точності та F1, а генерацію звітів — за метриками BLEU, ROUGE, F1-CheXbert, F1-RadGraph. Результати тестування були детально проаналізовані та визначено найкращу конфігурацію системи. За результатами встановлено, що використання техніки мультимодального RAG значно покращує можливості LLaVA-Med 1.5 у роботі з рентгенівськими знімками і їхніми звітами, навіть за умов обмежених ресурсів, зокрема, у використанні моделі в 4-бітному режимі.

Майбутні перспективи розвитку можливі в напрямку використання не тільки тексту та зображень як даних, а й інших типів даних, наприклад: аудіо, відео тощо. Перспективним є також створення мультимодальних RAG-систем для інших предметних сфер та автоматизація процесу їхньої розробки.

References

1. Belcic, I. and Stryker, C. (n.d.) *What is Learning Rate in Machine Learning?*. IBM. [online] Available at: https://www.ibm.com/think/topics/learning-rate [Accessed 25 May 2025].
2. Bergmann, D. and Stryker, C. (n.d.) *What is Loss Function?*. IBM. [online] Available at: https://www.ibm.com/think/topics/loss-function [Accessed 25 May 2025].
3. Chen, W. et al. (2022) *MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text*. [online] Available at: https://arxiv.org/abs/2210.02928 [Accessed 10 October 2025].
4. Coursera (n.d.) *What Does Batch Size Mean in Deep Learning? An In-Depth Guide*. [online] Available at: https://www.coursera.org/articles/what-does-batch-size-mean-in-deep-learning [Accessed 25 May 2025].
5. Coursera (n.d.) *What Is an Epoch in Machine Learning?*. [online] Available at: https://www.coursera.org/articles/epoch-in-machine-learning [Accessed 25 May 2025].
6. DeepAI (n.d.) *Harmonic Mean*. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/harmonic-mean [Accessed 25 May 2025].
7. DeepAI (n.d.) *Hyperparameter*. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/hyperparameter [Accessed 25 May 2025].
8. Delbrouck, J.-B. et al. (2022) *Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards*. [online] Available at: https://arxiv.org/abs/2210.12186 [Accessed 10 October 2025].
9. Demner-Fushman, D. et al. (2015) 'Preparing a collection of radiology examinations for distribution and retrieval', *Journal of the American Medical Informatics Association*, 23(2), pp. 304–310. doi: 10.1093/jamia/ocv080.
10. Deng, J. et al. (2009) 'ImageNet: A large-scale hierarchical image database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, Florida, 20-25 June. IEEE, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
11. Hugging Face (n.d.) *What is Image Classification?*. [online] Available at: https://huggingface.co/tasks/image-classification [Accessed 25 May 2025].
12. IBM (n.d.) *IBM Watson Studio and Knowledge Catalog*. [online] Available at: https://www.ibm.com/docs/en/ws-and-kc?topic=metrics-accuracy [Accessed 25 May 2025].
13. IBM (n.d.) *What are Convolutional Neural Networks?*. [online] Available at: https://www.ibm.com/think/topics/convolutional-neural-networks [Accessed 25 May 2025].
14. Irvin, J. et al. (2019) *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. [online] Available at: https://arxiv.org/abs/1901.07031 [Accessed 10 October 2025].
15. Jain, S. et al. (2021) *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports*. [online] Available at: https://arxiv.org/abs/2106.14463 [Accessed 10 October 2025].
16. Johnson, A. E. W. et al. (2019) *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. [online] Available at: https://arxiv.org/abs/1901.07042 [Accessed 10 October 2025].
17. Kingma, D. P. and Ba, J. (2014) *Adam: A Method for Stochastic Optimization*. [online] Available at: https://arxiv.org/abs/1412.6980 [Accessed 10 October 2025].
18. Li, C. et al. (2023) *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day*. [online] Available at: https://arxiv.org/abs/2306.00890 [Accessed 10 October 2025].
19. Lin, C.-Y. (2004) 'ROUGE: A Package for Automatic Evaluation of Summaries', in *Text Summarization Branches Out*. Barcelona, Spain, 25-26 July. ACL, pp. 74–81. Available at: https://aclanthology.org/W04-1013/ [Accessed 10 October 2025].
20. Loshchilov, I. and Hutter, F. (2017) *Decoupled Weight Decay Regularization*. [online] Available at: https://arxiv.org/abs/1711.05101 [Accessed 10 October 2025].
21. Mao, A., Mohri, M. and Zhong, Y. (2023) *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. [online] Available at: https://arxiv.org/abs/2304.07288 [Accessed 10 October 2025].
22. Nagel, M. et al. (2021) *A White Paper on Neural Network Quantization*. [online] Available

at: https://arxiv.org/abs/2106.08295 [Accessed 10 October 2025].
23. Papineni, K. et al. (2002) 'Bleu: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, 7-12 July. ACL, pp. 311–318. doi: 10.3115/1073083.1073135.
24. PhysioNet (2023) *MIMIC-CXR Database*. [online] Available at: https://physionet.org/content/mimic-cxr/2.1.0/ [Accessed 25 May 2025].
25. Radford, A. et al. (2021) *Learning Transferable Visual Models From Natural Language Supervision*. [online] Available at: https://arxiv.org/abs/2103.00020 [Accessed 10 October 2025].
26. Read, J. and Perez-Cruz, F. (2015) *Deep Learning for Multi-label Classification*. [online] Available at: https://arxiv.org/abs/1502.05988 [Accessed 10 October 2025].
27. Ruder, S. (2016) *An overview of gradient descent optimization algorithms*. [online] Available at: https://arxiv.org/abs/1609.04747 [Accessed 10 October 2025].
28. Smit, A. et al. (2020) *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. [online] Available at: https://arxiv.org/abs/2004.09167 [Accessed 10 October 2025].
29. Sun, S. et al. (2019) *A Survey of Optimization Methods from a Machine Learning Perspective*. [online] Available at: https://arxiv.org/abs/1906.06821 [Accessed 10 October 2025].
30. Wood, T. (n.d.) *F-Score*. DeepAI. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/f-score [Accessed 25 May 2025].
31. Wood, T. (n.d.) *Precision and Recall*. DeepAI. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall [Accessed 25 May 2025].
32. Xia, P. et al. (2024) *MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models*. [online] Available at: https://arxiv.org/abs/2410.13085 [Accessed 10 October 2025].
33. Zhang, S. et al. (2023) *BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs*. [online] Available at: https://arxiv.org/abs/2303.00915 [Accessed 10 October 2025].
34. Zhao, R. et al. (2023) *Retrieving Multimodal Information for Augmented Generation: A Survey*. [online] Available at: https://arxiv.org/abs/2303.10868 [Accessed 10 October 2025].

Одержано: 11.10.2025
Внутрішня рецензія отримана: 20.10.2025
Зовнішня рецензія отримана: 22.10.2025

Про авторів

[1]*Шевченко Михайло Григорович*
Бакалавр
https://orcid.org/0009-0004-5933-2349
[1]*Андрощук Максим Віталійович*
Здобувач ступеня доктора філософії
https://orcid.org/0000-0001-6183-6950

Місце роботи авторів:

[1]Національний університет
«Києво-Могилянська академія»
тел. +38-044-425-60-59
E-mail: vkd@ukma.edu.ua
https://www.ukma.edu.ua/