DOI: 10.46299/ISG.2024.MONO.TECH.2.6.2

6.2 Інформаційна технологія аналізу медичних зображень на основі моделі пояснювального штучного інтелекту

Вступ

Зі зростанням населення, пропорційно зростає і попит на лікарів, зокрема, на рентгенологів опорно-рухового апарату. Останні статистичні дослідження прогнозують значний дефіцит експертів-рентгенологів, а також інших спеціалістів у галузі медицини. Лише в США нестача лікарів у галузі рентгенології може перевищити 35 000 до 2034 року, згідно з нещодавно опублікованим щорічним аналізом потреби у лікарських спеціальностях, зробленим Асоціацією американських медичних коледжів [1].

Дана статистика підводить до висновку, що існує потреба в надійних комп'ютеризованих системах, які могли б допомогти ренгтенологам у виявленні патологічних станів під час аналізу МРТ пацієнтів. Основним застосуванням таких систем має бути зменшення робочого навантаження лікарів-спеціалістів, яке на даний момент постійно зростає.

В питанні інтерпретації результатів автоматизованого діагностування, також можна відзначити відсутність актуальних досліджень для зображень МРТ колінного суглоба, в першу чергу через суттєву складність побудови моделей для інтерпретації, зважаючи на багатовимірність вхідних даних.

Метою роботи є підвищення точності та інтерпретованості аналізу медичних даних, зокрема, серій знімків магнітно-резонансної томографії, на основі ансамблю методів машинного навчання (МН).

Інтерпретації рішень у автоматизованому діагностуванні. Існуючі алгоритми інтерпретації.

Розуміння та інтерпретація класифікаційних рішень автоматизованих систем діагностування в медицині має велике значення, оскільки це дозволяє перевірити міркування системи та надає додаткову інформацію людині – фахівцю, котрому надаються дані для підтримки прийняття рішень. Незважаючи

на те, що методи машинного навчання дуже успішно вирішують безліч завдань, у більшості випадків вони мають той недолік, що діють як чорна скринька, себто не надають жодної інформації про те, яка частина вхідних даних вплинула на конкретне рішення класифікації [170].

Відсутність інтерпретабельності пов'язана з нелінійністю різних відображень, які обробляють пікселі зображення до його представлення як вектору ознак, і надають це представлення до кінцевої функції класифікатора. Це є значним недоліком у застосуванні методів машинного навчання до класифікації у реальному житті, оскільки це заважає фахівцям ретельно перевіряти рішення щодо класифікації. Проста відповідь «так» або «ні» іноді має обмежену цінність у застосуваннях, де питання розташування або структури є більш доречними, ніж проста бінарна оцінка. Крім цього, є багато інших аргументів на користь інтерпретабельного штучного інтелекту [171], наприклад:

- **Перевірка коректності**: як згадувалося раніше, у багатьох застосуваннях, не можна безумовно довіряти системі типу чорної скриньки. Наприклад, у сфері охорони здоров'я використання моделей, які можуть бути інтерпретовані та перевірені медичними експертами, є абсолютною необхідністю. Існує приклад [172], де система штучного інтелекту, яка була навчена передбачати ризик пневмонії у людини, приходить до абсолютно хибних висновків. У процесі навчання модель дізнається, що хворі на астму з проблемами серця мають набагато менший ризик смерті від пневмонії, ніж здорові люди. Лікар-фахівець би одразу помітив, що це не може бути правдою, оскільки астма та проблеми з серцем є факторами, які негативно впливають на прогноз щодо одужання. Однак модель штучного інтелекту нічого не знає про такі явища як астма або пневмонія, а лише робить висновки з наявних даних. У цьому прикладі дані були систематично зміщені, оскільки на відміну від здорових людей більшість хворих на астму та серцево-судинні захворювання перебували під більш суворим медичним наглядом, ніж інші пацієнти.

- **Удосконалення**. Першим кроком до вдосконалення системи ШІ є розуміння її слабких сторін. Очевидно, що важче виконати такий аналіз слабких місць на моделях типу чорної скриньки, ніж на моделях, які можна інтерпретувати. Також легше виявити упередження в моделі або наборі даних (як у прикладі з пневмонією), якщо зрозуміти, що робить модель і чому вона приходить до своїх прогнозів. Можливість інтерпретації моделі може бути корисною при порівнянні різних моделей або архітектур. Деякі моделі можуть бути приблизно однаковими в плані якості класифікації, але їх архітектура передбачає суттєві відмінності між ознаками, що використовуються як основа для прийняття рішень. Це демонструє, що більш повні методи порівняння моделей вимагають інтерпретованості цих моделей. Також можна спробувати прослідкувати тенденцію, що краще розуміння поведінки моделей машинного навчання, дає більше простору до їх удосконалення.
- **Навчання на результатах передбачення**: оскільки сучасні системи штучного інтелекту навчаються на мільйонах прикладів, вони можуть спостерігати закономірності в даних, які раніше не були відомі широкому загалу оскільки люди здатні вчитися на значно мершій кількості прикладів. Використовуючи пояснювані системи штучного інтелекту, можна спробувати використовувати системи штучного інтелекту для отримання нових знань.
- **Відповідність законодавству**: системи ШІ впливають на все більше сфер нашого повсякденного життя. Разом з тим останнім часом підвищена увага приділяється також юридичним аспектам, наприклад, питанню відповідальності за рішення класифікації, коли система приймає неправильне рішення. Оскільки неможливо знайти задовільні відповіді на ці юридичні питання, покладаючись на моделі типу чорної скриньки, майбутні системи ШІ обов’язково повинні стати інтерпретабельним.

На сьогоднішній день існує декілька способів побудови пояснень для систем автоматизованого діагностування. Найпростіші методи опираються на

карту помітності, або карту салієнтності (англ. saliency maps). Карту помітності можна розглядати як приклад сегментації зображення. У комп'ютерному баченні процес сегментації зображення – це поділ цифрового зображення на декілька множин пікселів, відомих як суперпікселі. Мета сегментації полягає в тому, щоб спростити представлення зображення до більш значущого і легшого для аналізу. Зазвичай, сегментація використовується для визначення місцезнаходження об'єктів та їх меж на зображеннях [173].

Згідно із останніми дослідженнями ефективності використання підходів на базі карт помітності для пояснення результатів класифікації зображення глибинними нейронними мережами, дані алгоритми не є достатньо повними, оскільки погано враховують зміни в параметрах чи архітектурі моделей, результати класифікації яких вони намагаються пояснити.

У роботі [173], п'ять та два методи використання карт помітності (з восьми загалом) не пройшли рандомізований тест моделі на сегментаційному наборі даних та наборі виявлення об'єктів відповідно. Було зроблено припущення, що методи карт помітності не є чутливими до змін параметрів самої моделі. Повторюваність і відтворюваність більшості методів із використанням карт помітності були значно гіршими, аніж спеціально навчені підмережі локалізації, що було показано також на сегментаційному наборі даних, та наборі виявлення об'єктів.

У дослідженні [174], автори оцінювали ефективність кількох популярних методів карт помітності на наборі даних RSNA Pneumonia Detection щодо можливостей локалізації, стійкості до параметрів моделі при рандомізації позначок класів, а також повторюваності та відтворюваності з різноманітними архітектурами моделей. Було виявлено, що метод GradCAM [175] показав чудову чутливість до параметрів моделі та рандомізації міток і був агностичним до архітектури моделі, в той час як традиційні методи на основі карт помітності не змогли показати суттєвих результатів у цьому.

Робота описує техніку для більш прозорих передбачень моделей на основі згорткових нейронних мереж, шляхом візуалізації вхідних областей, які є

«важливими» для прогнозів, створюючи візуальні пояснення. Саме цей підхід було названо Gradient-weighted Class Activation Mapping, або ж Grad-CAM. Він використовує інформацію про градієнти класу для локалізації важливих регіонів. Ці локалізації поєднуються з існуючими візуалізаціями піксельного простору, щоб створити нову візуалізацію з високою роздільною здатністю та розрізненням класів під назвою Guided Grad CAM. Ці методи допомагають краще зрозуміти моделі на основі згорткових нейронних мереж, включаючи моделі підписів до зображень і візуальних відповідей на запитання. Grad-CAM пропонує принципово новий спосіб пояснення результатів передбачення. Принцип роботи даного алгоритму зображено на рисунку 1.

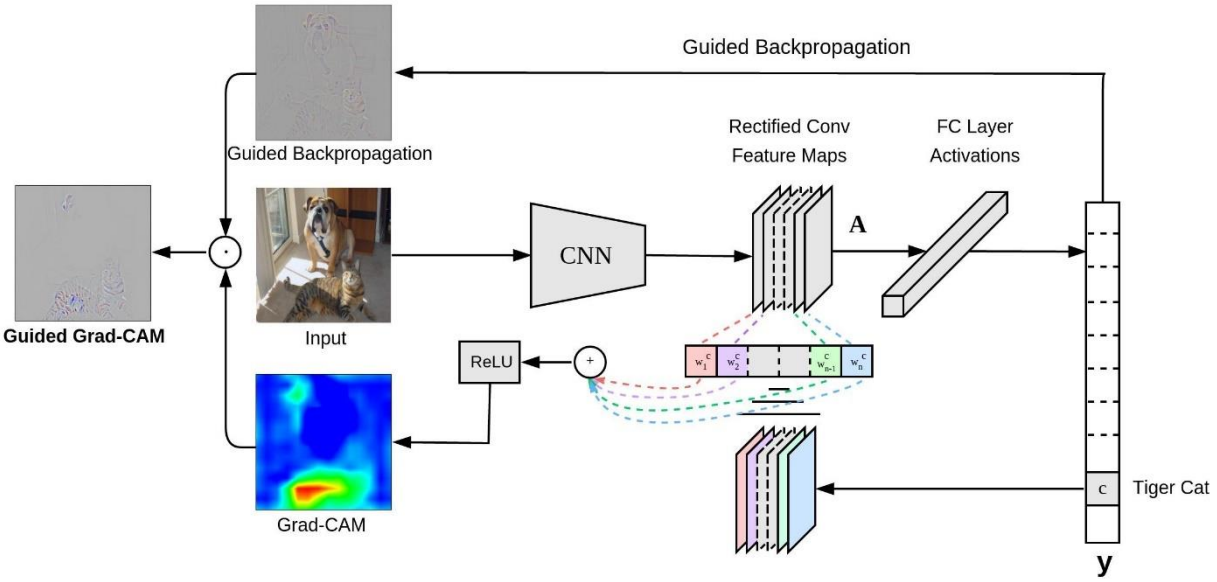


Рисунок 1. Огляд та візуалізація роботи алгоритму пояснення Grad-CAM

Алгоритм працює наступним чином – маючи вхідне зображення та мітку категорії («тигровий кіт») як вхідні дані, зображення проганяється через згорткову нейронну мережу, щоб отримати необроблені оцінки кожного класу перед задіювання функції активації softmax. Для градієнту встановлюється початкове значення як нуль для всіх класів, за винятком мітки класу із вхідних даних, яка має значення 1.

Цей сигнал потім поширюється у зворотному напрямку на випрямлену карту ознак згорткової нейронної мережі, де обчислюється груба локалізацію Grad-CAM теплова карта на рисунку 1. Нарешті, відбувається поточкове

множення теплової карти за допомогою керованого зворотного поширення, щоб отримати візуалізацію Grad-CAM, яка має високу роздільну здатність, і також розрізняє класи.

Підхід латентного зсуву

Нещодавно ідея візуалізації прогнозів через перебільшення ознак, які впливають на передбачення моделі, була запропонована у роботі [176]. Це перебільшення є результатом здатності нейронної мережі створювати, або "галюцинувати" нові ознаки, і цей процес, як відомо, може бути контрольованим.

Замість того, щоб просто генерувати зображення певного класу, згадані методи перебільшення можуть пояснити конкретні особливості, які використовує класифікатор для кожного прогнозу. Це важливо для прогнозування, коли модель передбачає використання неправильних хибних корелятів, щоб переконатися, що вона правильна з правильних причин. У той час як більшість моделей прогнозування патології на зображенні передбачають причинно-наслідкові зв'язки, де конкретні області зображення явно призводять до застосування певної класифікаційної мітки (наприклад, збільшене серце → кардіомегалія), моделі, що прогнозують ризик у майбутньому (наприклад, прогноз сметрності протягом 5 років), не мають заздалегідь відомого причинно-наслідкового зв'язку. У таких сценаріях можна дізнатися, які саме ознаки можуть використовуються в моделях, що здійснюють такі передбачення через перегляд згенерованого зображення із перебільшеними ознаками.

У роботі [177] описано підхід латентного зсуву, або модель латентної змінної, такої як простий автокодувальник $D\big(E(x)\big)$, де E— кодувальник, а D— декодувальник, і також присутній класифікатор f, який робить передбачення цільового показника y наступним чином: $y = f(x)$. Модель латентної змінної та класифікатор навчаються незалежно без будь-яких особливих обмежень, окрім того, що вони повинні бути диференційовними. Автокодувальник використовується з метою простоти навчання та програмної реалізації.

Коли моделі кодувальника та декодувальника є навчені, пояснення рішення класифікатора можна обчислити наступним чином:

Вхідне зображення x кодується за допомогою $E(x)$, створюючи його представлення у просторі латентної змінної z. Збурення латентного простору обчислюються для класифікатора f у рівнянні 1, який потім використовується для отримання λ-зміщених вибірок зображення, показаних у рівнянні 2.

$$z_\lambda = z + \lambda \frac{\partial f(D(z))}{\partial z} \tag{1}$$

$$x'_\lambda = D(z_\lambda) \tag{2}$$

Очікується, що зображення x'_λ дасть вищий прогноз, такий, що $f(x'_\lambda) > f(x)$. Звідси ми можемо генерувати кілька зображень x'_λ, щоб здійснити перебільшення або видалення ознак на зображенні, які призводять до певного прогнозу. Ці зображення можна з'єднати в короткі послідовності, які допомагають пояснити, чому було зроблено конкретне передбачення та які саме ознаки і в якій мірі використовувалися даним класифікатором.

Візуалізацію методу, описаного в роботі [177] показано на рисунку 2.

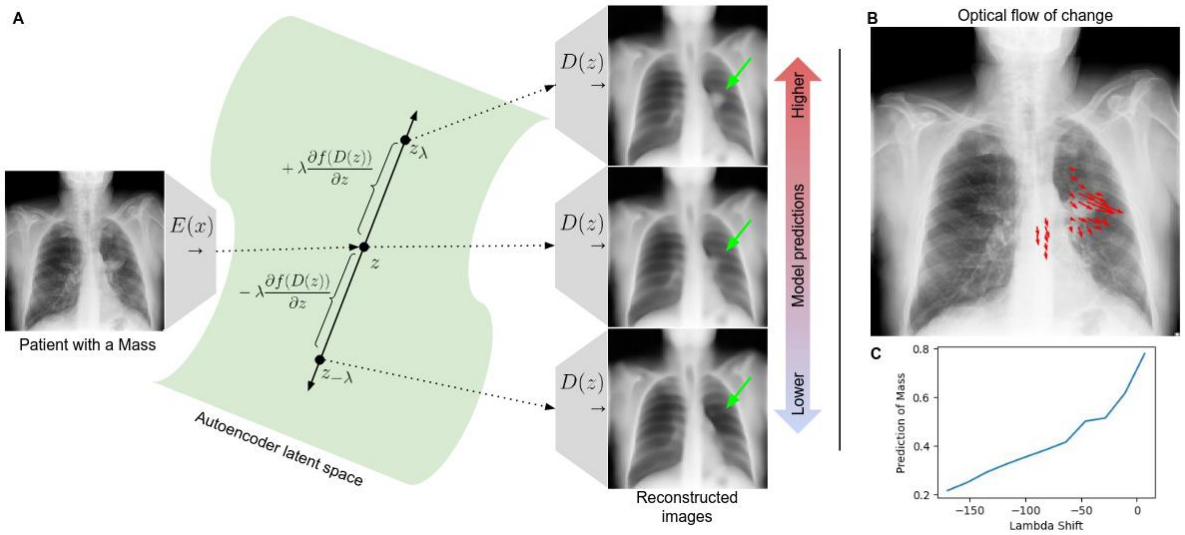


Рисунок 2. Візуалізація методу латентної змінної для побудови зображень із перебільшеними ознаками

При цьому підході також дуже важливо мати на увазі, що метод є обмежений простором латентного представлення ознак автокодувальника. Якщо декодувальник не є достатньо виразний, він не зможе правильно представити ознаки, які використовує класифікатор. Проте цей підхід дозволяє порівнювати декілька класифікаторів із фіксованим автоматичним кодувальником (або вибором моделі латентної змінної), що дозволяє чітко зрозуміти використання різноманітних ознак із зображення у моделях.

По суті, цей метод можна вважати протилежним до методу adversarial attack у GAN. Якби здійснювалася проста модифікація зображення за допомогою градієнта $\frac{\partial f(x)}{\partial x}$, що є традиційним підходом у adversarial attack, модифікація ознак була б непомітною та спотворювала б зображення шляхом вибору помилкових пікселів, які мали б вплив на цільову змінну.

У підході латентної змінної, процес є упорядкованим за допомогою фіксованого декодувальника, щоб запобігти зміні незначущих пікселів зображення. Загалом, у даному підході ставиться завдання змінити лише найбільш семантично значущі пікселі ознак, які призводять до конкретного результату класифікації.

Побудова інтерпретаційної моделі

Які проблеми було виявлено при побудові інтерпретаційної моделі?

МРТ можна вважати багатовимірними зображеннями, оскільки кожен розріз складається із набору вхідних зображень. Також, ці зображення розташовуються в строгій послідовності, інакше втрачається сенс обстеження з точки зору просторових координат. Тому, власне генерація таких складних вхідних даних за допомогою GAN цілком може вписатися в рамки окремого дисертаційного дослідження, оскільки є дуже складною як в плані алгоритмізації, так і в плані наявності ресурсів для обчислень.

Чому? Навіть не враховуючи той факт, що контрафактні зображення повинні абсолютно не відрізнятися від оригіналів у більшості частин за винятком зон із ознаками, додатково потрібно забезпечити їх цілісність в розрізі – тобто,

згенероване зображення повинне лише мінімально відрізнятися від попереднього та наступного у розрізах.

Для того, аби мати змогу застосувати підхід генерації контрафактних зображень у даному дослідженні, при побудові моделі пояснення було використано наступні спрощення:

1. Модель пояснення може працювати виключно із одним зображенням на одному розрізі МРТ. Зображення повинне знаходитися на відстані не більше ніж F розрізів від умовного центру серії зображень МРТ (у розробленій інформаційній системі передбачається перевірка цієї умови для вибору зображень для запуску моделі пояснення).
2. В якості вхідних даних для класифікатора, використовуються оригінальні дані, за винятком середини серії, де обране зображення викривлюється у просторі ознак дискримінатора із відхиленням λ, яке повторюється $2 \cdot F$ разів у центральній зоні серії.

Тут введено параметр F, котрий означає максимальну відстань від умовного центру серії зображень МРТ, на якій можна вибрати зображення для запуску моделі передбачення. За допомогою валідації моделі передбачення на вхідному наборі даних MRNet, було експериментально підібрано значення параметра $F = 3$. Схематично, принцип підстановки контрафактних зображень на вхід моделі зображено на рисунку 3.

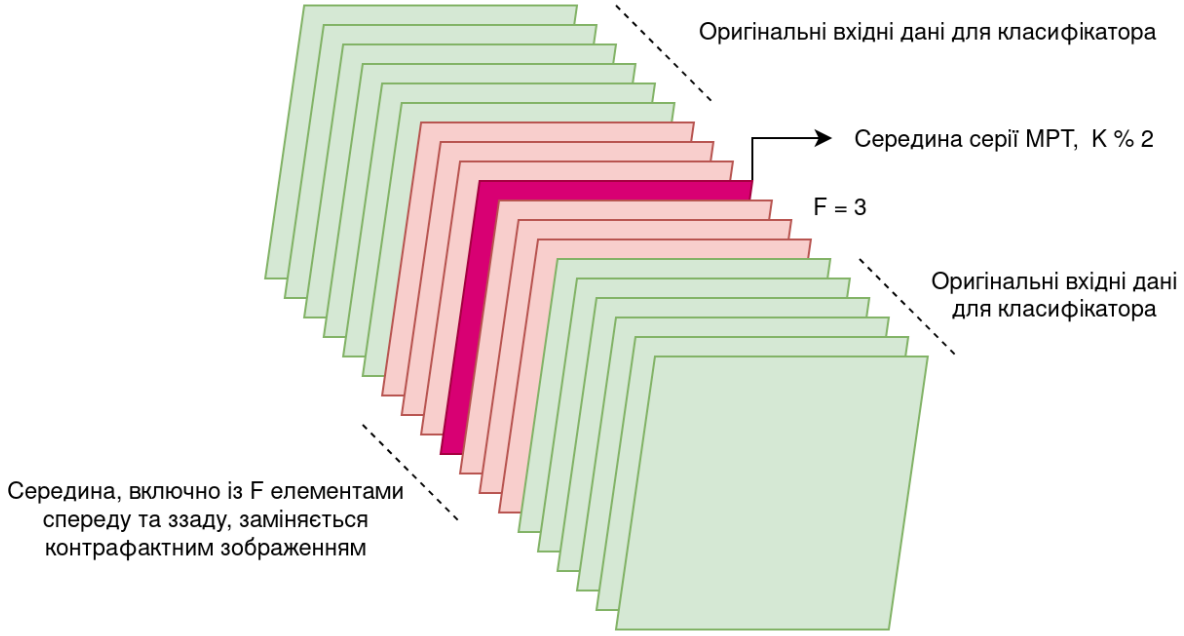


Рисунок 3. Схематика підстановки контрафактних зображень на вхід класифікатора для пояснювальної моделі, де K – кількість зображень у серії, F – параметризована максимальна відстань від центру серії до обраного зображення, на основі якого будується пояснення

Отож, оскільки на вхід до класифікатора подається середня зона серії зображень, яка дещо відрізнається від оригінальної, з метою спростити обчислення. Через це також відбувається і внесення похибок в результат передбачення.

Тому, пояснення не будуть ідеальними (якими вони могли би бути, враховуючи всю повноту вхідних даних). Проте у цьому випадку точністю доводиться в деякій мірі пожертвувати, задля можливості побудувати модель пояснення в цілому.

Яким чином можна визначити відхилення точності передбачення при процесах класифікації та пояснення? У даній роботі проведено експеримент, використовуючи в якості вхідних даних валідаційну частину набору даних MRNet. При фіксації випадкових початкових значень, було обчислено

ймовірність діагнозу на завданні класифікації для оригінальних вхідних даних, а також для даних із видозміненою середньою частиною серії зображень, використовуючи центральне зображення замість $F = 3$ суміжних з кожного боку, порівнюючи результат передбачення.

У результаті експерименту, на валідаційному наборі даних MRNet при значенні параметра $F = 3$ було отримано середнє відхилення результату передбачення 0.095623 що вважається допустимим для можливості побудови інтерпретаційної моделі. Параметр F було підібрано таким чином, щоб забезпечити найменшу похибку у разі використання підходу генерації контрафактних зображень, але при цьому маючи змогу виділити бажане центральне зображення із серії МРТ для більшості МРТ із валідаційного набору даних.

Врахувавши та прийнявши можливі неточності у передбаченні при побудові інтерпретаційної моделі, у даній роботі було здійснено тренування моделі на базі cGAN [175]. Аналогічно, як у згаданій роботі, маємо функцію пояснення, яка виглядає наступним чином:

$$L_f(x, \lambda): (X, R) \to X \tag{3}$$

Тут, x — це вхідне зображення, для якого потрібно згенерувати пояснення, а λ — це додатній або від'ємний зсув, призначений для класифікатора f. Функція $L_f(x, \lambda)$ діє з декартового добутку простору зображень X та простору дійсних чисел R на простір вхідних зображень X.

Функція пояснення складається із кількох компонентів, включаючи згадані вище елементи автокодувальника — кодувальник $E(x)$ та декодувальник $G(z)$. У даному дисертаційному дослідженні, автокодувальник побудовано на базі ResNet. Автокодувальник було навчено з нуля на базі середніх зображень МРТ із сагітальних розрізів тренувального набору даних MRNet, де найбільш помітно передню хрестоподібну зв'язку (ACL).

Функція втрат для автокодувальника складається із декількох частин, і покликана задовільнити деякі обмеження на контрафактне зображення. По-

перше, контрафактне λ — зміщене зображення повинно належати тому ж простору вхідних зображень X, що й оригінальне зображення. Для цього, зображення повинні мінімально відрізнятися одне від одного. Тому, функція втрат для декодувальника та дискримінатора виглядає наступним чином:

$$L_{cGAN} = E_{x,\lambda P(x,\lambda)} \quad (4)$$

Тут в якості P позначаємо розподіл параметрів для λ, x, а також z. Розподіл zP_z – це шум, взятий із нормального розподілу у просторі латентної змінної z.

Саме z – це латентне представлення вхідного зображення X, яке утворюється в результаті роботи кодувальника $E(x)$. E – це математичне сподівання по заданому розподілу.

На додачу до цього, семантично виділені регіони згенерованих зображень повинні бути співставними з оригіналом. Однієї лише функції втрат на основі подібності двох зображень є недостатньо для генерації досить подібних зображень, тому додатково натреновано модель семантичного виділення хрестоподібної зв'зки $S(x)$ на зображеннях МРТ.

Результат роботи моделі семантичного видіення враховується дискримінатором $D(\cdot)$ при кінцевому відборі кандидатів на зображення x_λ. Результат виконання моделі виділення хрестоподібної зв'язки показано на рисунку 4.

Тому, додатково задана функція втрат, що враховує співпадіння регіонів для оригінального та контрафактного зображення. Простими словами, ця частина загальної функції втрат «штрафує» генератор зображень, якщо у заданих зображеннях не співпадають регіони передньої хрестоподібної зв'язки, що виділені моделлю виділення $S(x)$.

Рисунок 4. Приклад роботи екстракції зони, де знаходиться хрестоподібна зв'язка із сагітального розрізу зображення МРТ.

$$L_{rec} = \left|p_{ij} - p_{ij}^{\lambda}\right|, (i,j) \in S(x), \tag{5}$$

де p_{ij} – це значення пікселя в двовимірному чорно-білому зображенні x, а p_{ij}^{λ} – у зображенні x_{λ}, відповідно. $S(x)$ повертає множину точок $(i,j$, котрі включають в себе передню хрестоподібну зв’язку. Таким чином, маємо функцію втрат, що рахує абсолютні значення різниці пікселів на частині зображення, яка прокласифікована як передня хрестоподібна зв’язка. Інтуїтивно, цей регіон повинен бути розташований у тому самому місці на обох зображеннях.

Нарешті, класифікатор $f(x)$ повинен давати приблизно λ - зміщену ймовірність передбачення для зображень x та x_{λ}. Для цього, використано функцію втрат на основі розходження Кульбака — Лейблера.

$$L_f(D,G) := r(f(x) + \lambda|x) + D_{KL}(f(x_{\lambda})|f(x) + \lambda).$$

Отож, функція пояснення в процесі свого навчання, маючи заданий автокодувальник $E(G)$, а також дискримінатор D, повинна мінімізувати наступну функцію втрат:

$$L = c_{cGAN} \cdot L_{cGAN}(D,G) + c_f \cdot L_f(D,G) + c_{rec} L_{rec}(E,G) \tag{6}$$

Загальна архітектура функції пояснення показана на рисунку 5.

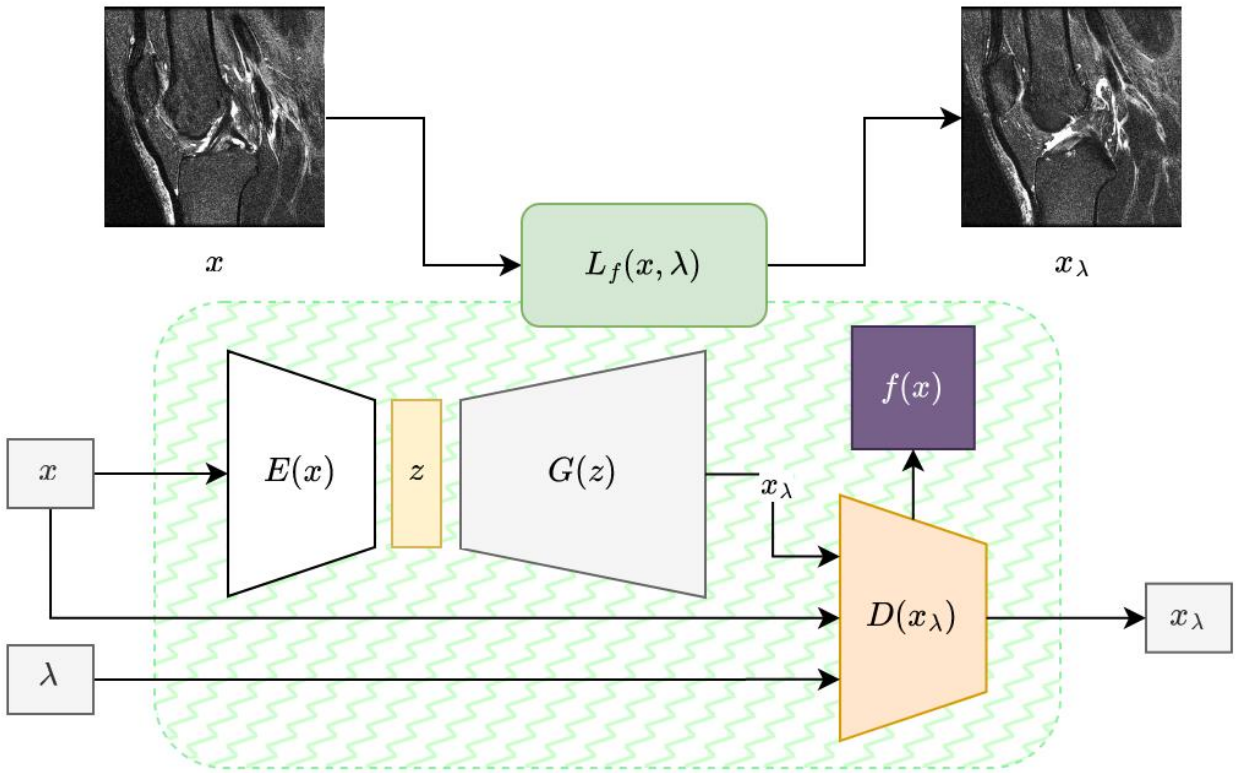


Рисунок 5. Архітектура пояснювальної функції $L_f(x, \lambda)$. Вхідне зображення x проходить через кодувальник E(x), і формується його представлення у просторі латентної змінної z. Далі, декодувальник G(z) здійснює реконструкцію вхідного зображення, та подає його на дискримінатор $D(\cdot)$, який вирішує, чи задовільняє згенероване зображення критеріям функцій втрат, використовуючи як саме зображення, так і функцію-класифікатор.

На відміну від мережі для автоматизованого діагностування, тут не розглядалися різноманітні варіанти каркасних мереж для пояснюваного навчання, оскільки це потребує значних затрат часу та обчислювальних ресурсів. Натомість, для елементів інтерпретаційної моделі було задіяно найбільш ефективні та прості в імплементації мережі з наявних.

У таблиці 1 показано роботу частини інтерпретаційної моделі, котра займається генерацією зображень на основі оригінального, використовуючи підхід латентної змінної із заданими значеннями зсуву λ.

Для того, аби побудувати карту помітності на основі цих даних, у цій роботі використано алгоритм піксельної відмінності між двома зображеннями у чорно-білому спектрі.

Беруться два вхідних зображення (оригінальне зображення x , та контрафактне із найбільшим значенням λ - $x_{\lambda=0.8}$ конвертується у матрицю інтенсивності пікселів розміром 256x256, і для кожного пікселя обчислюється абсолютне значення різниці інтенсивності I_D між даними зображеннями:

$$x = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,256} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots & p_{2,256} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{256,1} & p_{256,2} & p_{256,3} & \cdots & p_{256,256} \end{bmatrix}$$

$$x_\lambda = \begin{bmatrix} p^\lambda_{1,1} & p^\lambda_{1,2} & p^\lambda_{1,3} & \cdots & p^\lambda_{1,256} \\ p^\lambda_{2,1} & p^\lambda_{2,2} & p^\lambda_{2,3} & \cdots & p^\lambda_{2,256} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p^\lambda_{256,1} & p^\lambda_{256,2} & p^\lambda_{256,3} & \cdots & p^\lambda_{256,256} \end{bmatrix}$$

$$I_D = \left(d_{ij}\right), d_{ij} = \left|p^\lambda_{ij} - p_{ij}\right|, i = \overline{1{,}256}, j = \overline{1{,}256}$$

Абсолютне значення різниці інтенсивності потім кодується у нове зображення розміром 256x256, і використовується в якості пояснювального зображення.

Приклад даного алгоритму візуалізації відмінностей між оригінальним та контрафактним зображеннями можна бачити на рисунку 6.

Таблиця 1. Контрафактні зображення для інтерпретаційної моделі, згенеровані на сагітальному розрізі для діагностування розриву ACL.

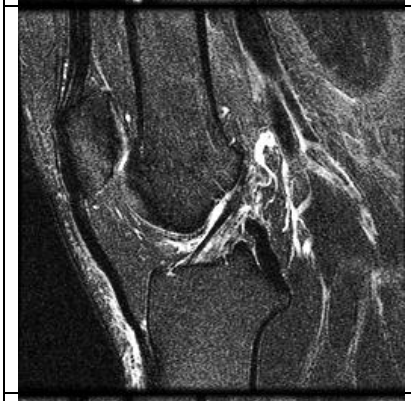
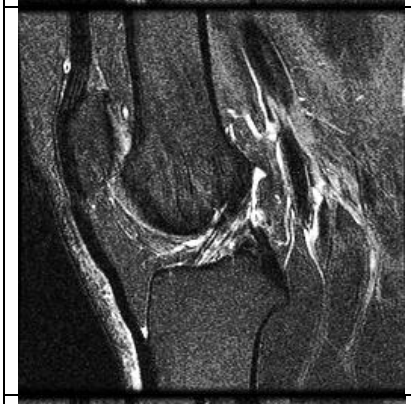
Зображення	λ	$f(x_\lambda)$	$\lvert F(x_\lambda) - F(x) - \lambda \rvert$
	-	0.17803	-
	Оригінальне зображення. Середній розріз МРТ, який прокласифіковано моделлю автоматизованого діагностування як малоймовірний з точки зору розриву ACL. На малюнку видно візуально неушкоджену передню хрестоподібну зв'язку.		
	0.2	0.35100	0.02704
	Згенероване контрафактне зображення із заданим $\lambda = 0.2$. Видно незначне потоншення лівої частини хрестоподібної зв'язки, на що класифікатор відреагував збільшенням ймовірності розриву.		
	0.5	0.59225	0.08577
	Згенероване контрафактне зображення із заданим $\lambda = 0.5$. Видно суттєве видозмінення лівої частини зв'язки, а також незначне видозмінення правої частини. Модель класифікації відреагувала відповідно, хоча із значною похибкою.		
	0.8	0.94476	0.03327
	Згенероване контрафактне зображення із заданим $\lambda = 0.8$. Видно, що на зображенні практично відсутня ліва частина хрестоподібної зв'язки.		

Рисунок 6. Візуалізація алгоритму створення пояснювального зображення(праворуч) для вхідного малюнку(ліворуч), на основі максимально контрафактного малюнку(по центру)

Висновки

У дослідженні розроблено інтерпретаційну модель для задачі діагностування ушкоджень МРТ коліна, на основі методу зсуву латентної змінної. Методологія інтерпретації полягає у генеруванні контрафактних зображень, з подальшою їх параметризованою підстановкою у оригінальну серію зображень МРТ.

Було підібрано оптимальні параметри для вставки з метою максимально зберегти вхідну точність класифікатора, із максимальною похибкою передбачення між оригінальними та контрафактними серіями в межах 0.096.