Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження моделей машинного навчання
для автоматичного опису зображень
(тема)

Виконав:
студент 2 курсу, групи ІПЗм-22-3

Голобородько Б.Ю.
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)
Тип програми освітньо-наукова

Керівник проф. Смеляков К.С.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри _____ З.В.Дудар
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Голобородько Богдану Юрійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи «Дослідження моделей машинного навчання для Автоматичного опису зображень»
Затверджена наказом по університету від 29.03. 2024р. № 230 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 17.06.2024
3. Вихідні дані до роботи опис досліджуваних моделей, критерії для оцінки згенерованих описів, вимоги до розробки системи для проведення досліджень за обраною предметною областю, мова програмування Python, бібліотеки Nunpy, Tourch, PIL, Transformers, CLIP, Pandas, Flask
4. Перелік питань, що потрібно опрацювати в роботі
аналіз існуючих моделей машинного навчання, вибір моделей, метрик та датасетів для дослідження, пошук зображень та їх анотація, проектування системи для проведення експериментальних досліджень, написання програмних рішень, проведення експериментів та аналіз отриманих результатів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз предметної галузі та постановка задачі	29.03.2024	*виконано*
2	Аналіз предметної області та вибір моделей для дослідження	31.03.2024	*виконано*
3	Експериментальні дослідження	24.04.2024	*виконано*
4	Аналіз результатів експериментальних досліджень та розробка рекомендацій	24.04.2024 – 26.05.2024	*виконано*
5	Підготовка пояснювальної записки	02.06.2024	*виконано*
6	Підготовка презентації та доповіді	06.06.2024	*виконано*
7	Нормоконтроль	07.06.2024	*виконано*
8	Рецензування	12.06.2024	*виконано*
9	Занесення диплома в електронний архів	13.06.2024	*виконано*
10	Попередній захист	13.06.2024	*виконано*
11	Допуск до захисту у зав. кафедри	14.06.2024	*виконано*
12	Захист магістерського дослідження	17.06.2024	*виконано*

Дата видачі завдання 29 березня 2024р.

Студент (ка) _____ Голобородько Б.Ю.
(підпис)

Керівник роботи _____ проф. Смеляков К.С.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 74 с., 29 рис., 6 табл., 20 джерел.

МАШИННЕ НАВЧАННЯ, НЕЙРОМЕРЕЖІ, ОПИС ЗОБРАЖЕННЯ, ТРАНСФОРМЕР, BLEU, BLIP, BLIP-2, CLIP, NLTK, VIT

Об’єкт дослідження – методи машинного навчання для генерації опису зображень та оцінка ефективності їх роботи.

Мета роботи – дослідження процесу автоматичної генерації опису зображення, моделі машинного навчання та критерії оцінки згенерованих результатів.

Методи розробки базуються на основі роботи нейромереж, їх ефективності та показників ефективності. При виконанні роботи були використані загально-логічні, теоретичні та емпіричні методи дослідження.

У результаті кваліфікаційної роботи було проаналізовано моделі машинного навчання, проведено тестування їх роботи з урахуванням об’єктивних та суб’єктивних критеріїв, розроблено веб-додаток для подальшого дослідження.

MACHINE LEARNING, NEURAL NETWORKS, IMAGE DESCRIPTION, TRANSFORMER, BLEU, BLIP, BLIP-2, CLIP, NLTK, VIT

Object of study – machine learning methods for generating image descriptions and evaluating their performance.

Purpose – to study the process of automatic image description generation, machine learning models and criteria for evaluating the generated results.

The development methods are based on the operation of neural networks, their efficiency and performance indicators. General logical, theoretical, and empirical research methods were used in the course of the work.

As a result of the qualification work, machine learning models were analyzed, their performance was tested taking into account objective and subjective criteria, and a web application was developed for further research.

Я, Голобородько Богдан Юрійович, студент гр. ІПЗм-22-3, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження моделей машинного навчання для автоматичного опису зображень», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

Перелік скорочень 7
Вступ 8
1 Аналіз предметної галузі 10
1.1 Загальний аналіз галузі 10
1.2 Огляд сучасних рішень для опису зображень 13
1.3 Виявленя проблем 18
1.4 Постановка задачі 19
2 Аналіз існуючих методів та алгоритмів 21
2.1 Трансформер та механізм уваги 21
2.2 Огляд архітектур нейронних мереж для опису зображень 24
2.3 Критерії оцінки якості генерованих описів 29
2.4 Інформаційна підготовка прийняття рішень 31
3 Експеримент 36
3.1 Розробка вимог до програмної системи 36
3.2 Підготовка зображень для експерименту 38
3.3 Практична реалізація 42
3.4 Результати експерименту 47
Перелік джерел посилання 58
Додаток А Перелік джерел посилання за науковими напрямами керівника та науковців кафедри програмної інженерії 61
Додаток Б Звіт результатів перевірки на унікальність тексту в базі ХНУРЕ 62
Додаток В Апробація результатів роботи 63
Додаток Г Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008: 2015 66
Додаток Д Слайди презентації 67

ПЕРЕЛІК СКОРОЧЕНЬ

BERT – Bidirectional Encoder Representations from Transformers

BLIP – Bootstrapping Language-Image Pre-training

CLIP – Contrastive Language-Image Pre-training

CNN – Convolutional Neural Network

LSTM – Long Short-Term Memory

RNN – Recurrent Neural Network

ViT – Vision Transformer

ВСТУП

У сучасному світі, де обсяг візуальної інформації зростає експоненціально, автоматичне описування зображень стає важливим завданням для багатьох сфер діяльності. Від створення доступних для читання веб-сторінок для людей з вадами зору до автоматизації систем керування вмістом та покращення пошукових систем, автоматичне генерування описів зображень має широкий спектр застосувань. Проте, ця задача залишається складною і вимагає комплексного підходу до обробки та інтерпретації візуальних даних.

Проблематика досліджень моделей для автоматичного опису зображень включає кілька ключових аспектів. По-перше, це складність побудови моделей, які здатні адекватно розуміти і описувати зображення в різних контекстах. По-друге, існують виклики, пов'язані з оцінкою якості згенерованих описів, оскільки традиційні метрики часто не відображають повною мірою семантичну та стилістичну релевантність тексту. По-третє, ефективність моделей значно залежить від обчислювальних ресурсів, що може бути проблемою для їхнього широкого застосування.

Актуальність досліджень у цій сфері обумовлена постійним зростанням кількості візуальних даних та необхідністю автоматизації їх обробки. Автоматичні системи, здатні генерувати описання для зображень, можуть значно спростити роботу в багатьох галузях, включаючи медіа, освіту, електронну комерцію та інші. Впровадження таких систем сприятиме підвищенню ефективності обробки даних, забезпеченню доступності інформації та покращенню користувацького досвіду.

Об'єктом дослідження є процес автоматичного генерування описів для зображень за допомогою моделей машинного навчання. Предметом дослідження є методи та алгоритми оцінки якості згенерованих описів, а також ефективність різних моделей машинного навчання в контексті автоматичного опису зображень.

У цій роботі розглянуто розробку та впровадження веб-системи для тестування різних моделей машинного навчання для автоматичного опису зображень. Основна увага приділяється створенню зручного інтерфейсу для

взаємодії з моделями, забезпеченню об'єктивної оцінки якості згенерованих описів за допомогою статистичних метрик та експертних оцінок, а також інтеграції та тестуванню моделей для забезпечення їх максимальної ефективності та точності.

Таким чином, дана робота спрямована на вирішення актуальних проблем автоматичного опису зображень, забезпечуючи інструменти для детального аналізу та вдосконалення існуючих моделей, що сприятиме їх подальшому розвитку та широкому застосуванню в різних сферах.

Результати роботи пройшли апробацію на 28-му Молодіжному форуму «Радіоелектроніка та молодь у XXI столітті» за темою «Тенденції розвитку генерації автоматичного опису зображень» у співавторстві з керівником кваліфікаційної роботи, що продемонстровано у відповідному додатку пояснювальної записки (Зб. матеріалів форуму. Т. 6., Харків: ХНУРЕ. 2024.С.408-410. (DOI: https://doi.org/10.30837/IYF.IIS.2024.408)).

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Загальний аналіз галузі

Автоматичний опис зображень, або image captioning, є однією з ключових проблем у сфері комп'ютерного зору та обробки природної мови. Це міждисциплінарна галузь, що поєднує методи аналізу візуальних даних з алгоритмами обробки текстової інформації. Основна ідея полягає у створенні системи, яка здатна автоматично генерувати текстові описи для заданих зображень, що має велике практичне значення для різних застосувань, таких як полегшення доступу до візуальної інформації для людей з порушеннями зору, автоматичне створення метаданих для зображень, покращення результатів пошукових систем, тощо.

Термін "image captioning" почав активно використовуватися в наукових дослідженнях у 2010-х роках, коли з'явилися значні досягнення у поєднанні методів комп'ютерного зору та обробки природної мови для автоматичного генерування текстових описів зображень. Одна з ключових публікацій, що суттєво вплинула на розвиток цієї галузі, була представлена у 2015 році [1]. В статті було запропоновано модель глибоких візуально-семантичних вирівнювань для генерації описів зображень, що стало важливим кроком у розвитку методів image captioning.

З розвитком глибокого навчання та збільшенням обчислювальних потужностей з'явилися нові моделі та архітектури, які значно покращили якість автоматичних описів зображень. Серед них особливої уваги заслуговують конволюційні нейронні мережі для витягування ознак зображення та рекурентні нейронні мережі для генерації тексту. Більш сучасні підходи включають використання трансформерів та гібридних архітектур, які об'єднують переваги обох архітектур для створення більш точних і контекстуально збагачених описів.

В спрощеному вигляді, архітектура для створення підписів до зображень зазвичай включає два основні компоненти: модуль аналізу зображень, відомий як енкодер (encoder), і модуль генерації тексту, що використовує інформацію, закодовану енкодером.

Енкодер застосовує нейронні мережі для обробки вхідного зображення. Його головна задача полягає у витягуванні візуальних ознак (features) із зображення, включаючи ідентифікацію об'єктів, визначення їхніх позицій, розмірів, кольорів і взаємозв'язків між різними об'єктами. Ми будемо зосереджуватися на конволюційних нейронних мережах (таких як VGG, ResNet або Inception), оскільки вони ефективно розпізнають об'єкти та їхні характеристики. Схема взаємодії енкодер – декодер для генерації підписів до зображень має наступний вид:

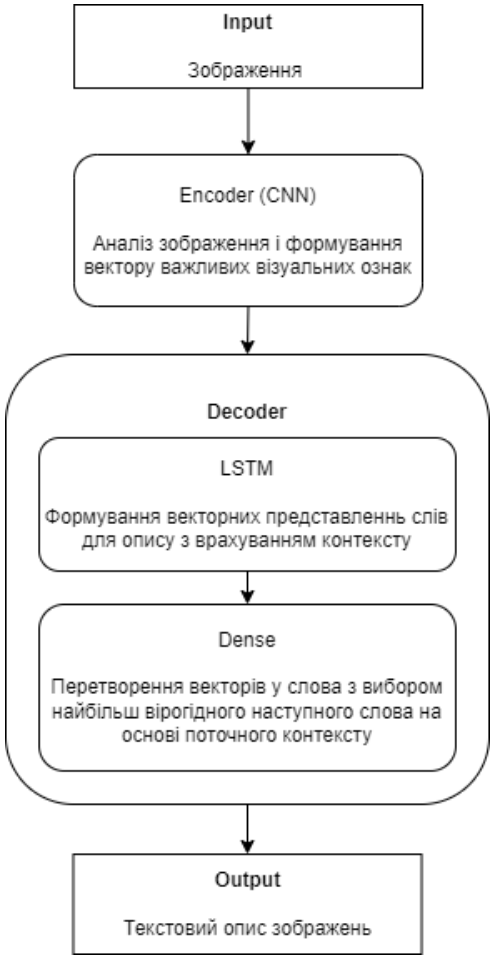


Рисунок 1.1 – Процес генерації текстового опису зображення за допомогою нейронної мережі [Рисунок виконаний самостійно]

Після витягнення візуальних ознак ці дані передаються до модуля генерації тексту. Зазвичай цей модуль базується на рекурентних нейронних мережах (RNN), LSTM або трансформерах. Він відповідає за перетворення візуальних ознак в описовий текст, враховуючи граматичну структуру та контекст.

Ключовою частиною цього процесу є створення зв'язного та релевантного опису, який точно відображає зміст зображення.

Dense - це шар нейронної мережі, який перетворює векторні представлення на конкретні слова, вибираючи найбільш відповідне наступне слово на основі поточного контексту. Це дозволяє системі будувати фрази та речення, що логічно і послідовно підходять одне до одного.

Підхід, показаний на рис. 1, використовується через його ефективність у вилученні ознак із зображень (що забезпечується CNN) та здатність зберігати контекстуальну інформацію в послідовностях даних (що забезпечується рекурентною нейронною мережею з архітектурою LSTM).

Продуктивність такої архітектури, як CNN-LSTM, знижується при обробці зображень із великою кількістю інформації та об'єктів. Це відбувається через відсутність семантичного розуміння сцен, представлених на зображеннях, що впливає на ефективність моделі в геометричній прогресії [2].

Ця пробема була вирішена за рахунок механізму уваги (attention mechanism), який дозволяє моделі динамічно зосереджуватись на важливих частинах вхідних даних. Це імітує людську здатність зосереджувати увагу та обробляти великі обсяги інформації. З появою трансформерів та архітектур на зразок BERT для обробки тексту і сучасних варіантів CNN для аналізу зображень, дослідники та інженери почали використовувати більш ефективні методи. Моделі, які можуть краще розуміти та описувати зміст зображень, стають все більш точними і контекстуально релевантними. Одним з таких прикладів є модель CNN – GRU [3, 4]. У цьому випадку CNN виокремлює просторові особливості зображень, а модель GRU (Gated Recurrent Unit) ефективно вирішує проблему градієнтного вибуху при довгостроковому прогнозуванні. GRU має простішу та швидшу структуру порівняно з LSTM, забезпечуючи вищу точність прогнозування.

Помітним прогресом у розвитку механізмів уваги стало впровадження моделі адаптивної уваги з візуальним вартовим (Adaptive Attention with Visual Sentinel) [5]. Цей метод ефективно вирішує проблему вибору оптимальних

моментів і місць для фокусування уваги на зображенні, що є особливо важливим для слів, які не мають чіткого візуального представлення.

Далі була представлена модель навмисної залишкової уваги (Deliberate Residual Attention Network), яка використовує двоетапний механізм уваги. Перший етап формує початкову версію підпису, а другий уточнює її, що призводить до більш точних та контекстуально відповідних підписів.

Крім того, продовжується розвиток енд-ту-енд моделей, таких як Vision Transformers [5], які здатні обробляти зображення без використання традиційних CNN. Ці моделі можуть інтегруватися з трансформерами для тексту, створюючи ще потужніші системи для генерації описів зображень.

Значний прогрес у цій галузі також пов'язаний із наявністю великих датасетів, таких як MS COCO, що містять тисячі зображень з людськими анотаціями. Це дозволяє тренувати моделі на великому обсязі даних, що покращує їхню здатність узагальнювати та генерувати описи для нових, раніше невідомих зображень. Крім того, впровадження нових методів оцінювання, таких як BLEU, METEOR, ROUGE та CIDEr, дозволяє більш об'єктивно оцінювати якість згенерованих описів, сприяючи розвитку більш досконалих моделей. Більш детально датасети та методи оцінювання ми розглянемо в наступних розділах.

Таким чином, дослідження в галузі автоматичного опису зображень активно розвиваються, зокрема завдяки постійному вдосконаленню алгоритмів машинного навчання та збільшенню доступності обчислювальних ресурсів. Це відкриває нові можливості для застосування таких технологій у різних сферах, від медицини до розваг, та стимулює подальші наукові пошуки в цій області.

1.2 Огляд сучасних рішень для опису зображень

Сучасні рішення включають різноманітні моделі та підходи, які дозволяють генерувати зв'язні та контекстуально релевантні описи для зображень. Для користувача процес генерації опису для зображення має більш-менш уніфікований інтерфейс: треба завантажити зображення або надати посилання.

Розглянемо найбільш розповсюджені рішення та їхні ключові особливості. Наприклад, онлайн-сервіс 10015.io дозволяє завантажити зображення і отримати 4 різних варіанти опису (рис. 1.2).

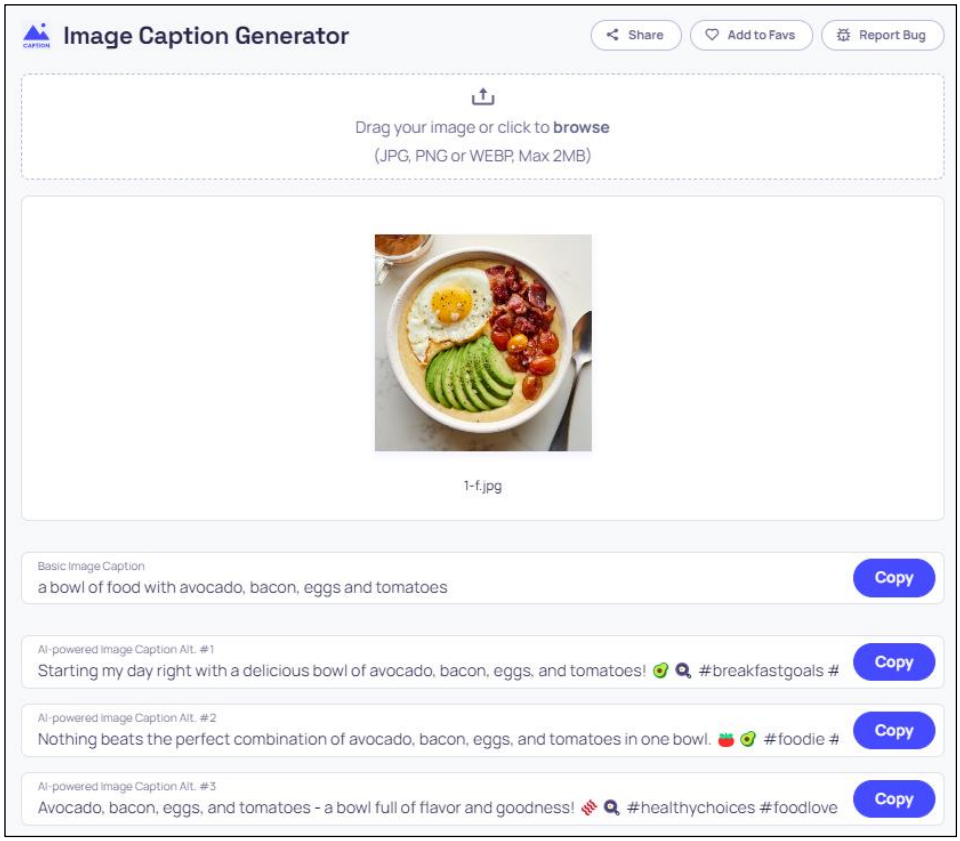


Рисунок 1.2 – Сервіс Image Captioning Generator від 10015.io [Рисунок виконаний самостійно]

Значно більш простим є сервіс для опису зображень від SRDdev (див. Рисунок 1.3), який генерує один опис зображення.

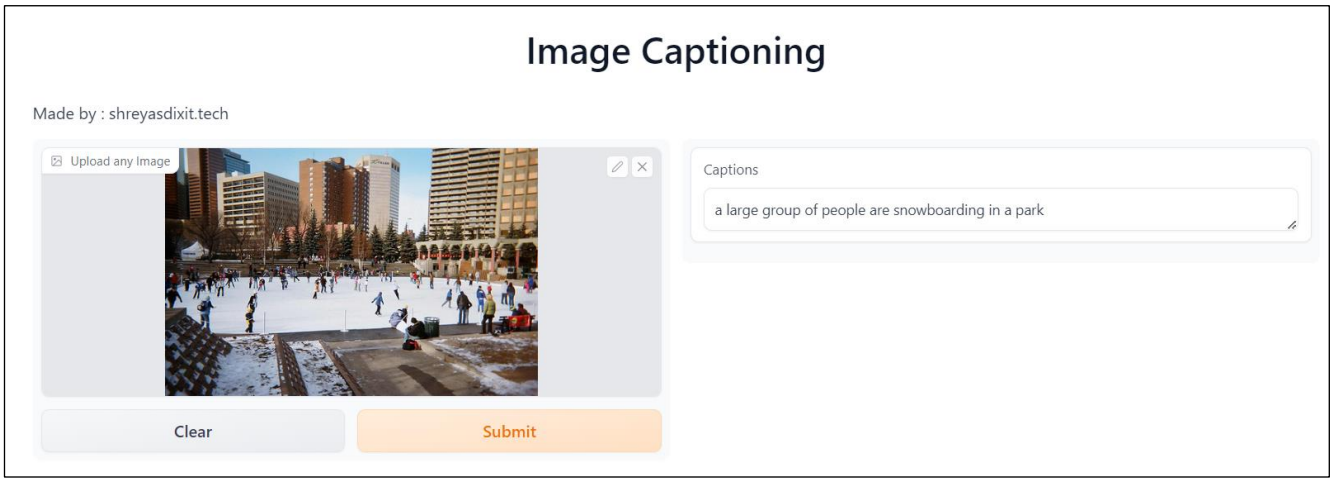


Рисунок 1.3 – Сервіс SRDdev/Image-Caption [Рисунок виконаний самостійно]

Схожим чином працює і сервіс imagecaptiongenerator.com, який окрім зображення пропонує вказати тональність, мову на якій має бути згенерований опис та додаткові деталі, що допоможуть створити більш релевантний опис. Приклад роботи наведено на рис 1.4.

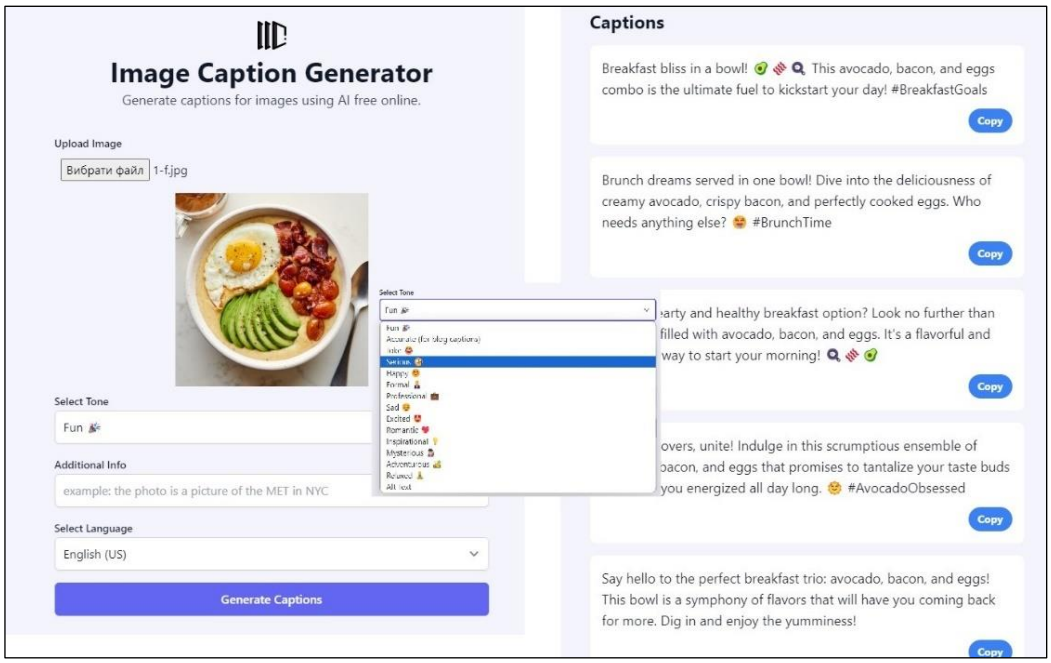


Рисунок 1.4 – Сервіс imagecaptiongenerator.com [Рисунок виконаний самостійно]

Окрім онлайн сервісів, є ще купа хмарних провайдерів, які пропонують свої моделі для опису зображень, які доступні як API. Так, в Google Cloud доступний Imagen for Captioning & VQA [6], який не тільки може генерувати описи до зображення а й відповідати на питання про них. Azure пропонує модель Image Analysis 4.0 [7], яка створює опис з одного речення для всього вмісту зображення (рисунок 1.5).

Рисунок 1.5 – Сервіс Image Analysis 4.0 від Azure [Рисунок виконаний самостійно]

Якщо ж розглядати моделі які доступні у вільному доступі, то вони не мають подібного інтерфейсу і потребують певних технічних навиків та їх результати можуть бути дещо гіршими за комерційні моделі.

Для прикладу, на рисунку 1.6 наведено результат роботи моделі для опису зображень:

Рисунок 1.6 – Результати опису зображень з бібліотеки PyTorch [8]

Згідно з статистикою сайту huggingface.co [9], де користувачі можуть ділитись моделями машинного навчання та наборами даних, найпопулярнішими моделями за кількістю завантажень в розділі Image Captioning є:

– vit-gpt2-image-captioning – 2.17 мільйонів завантажень;
– blip-image-captioning-large – 1.46 мільйонів завантажень;
– blip2-opt-2.7b – 574 тисячі завантажень;
– blip-image-captioning-base – 514 тисячі завантажень;
– GIT (GenerativeImage2Text), base-sized - 457 тисячі завантажень;
– InstructBLIP model – 50,4 тисячь завантажень.

Кількість завантажень зазначена на кінець 2023 року. Найбільш популярною на той момент є модель vit-gpt2, яка використовує архітектуру трансформерів для одночасної обробки зображень та генерування описів. Vision Transformer розділяє зображення на патчі, подібно до слів у реченні, а потім

використовує механізм уваги для виявлення найважливих частин зображення. Генерація тексту здійснюється за допомогою GPT-2, яка використовує велику кількість параметрів для створення зв'язного та релевантного тексту на основі вхідних даних.

Друга (blip-image-captioning-large) та четверта (blip-image-captioning-base) моделі є варіаціями BLIP і відрізняються лише кількістю параметрів. Особливістю BLIP є здатність одночасно обробляти зображення та текст, встановлюючи взаємозв'язки між візуальним і текстовим контентом [10]. Модель представили у 2022 році, поєднавши трансформер для обробки тексту та конволюційну нейронну мережу для роботи з зображенням.

На третьому місці знаходиться blip2-opt-2.7b, яка поєднує BLIP-2 та OPT (Open Pre-trained Transformer). Модель що представлена у 2023 році, є вдосконаленою версією BLIP, включає в себе 3 моделі: CLIP-подібний кодер зображень, трансформатора запитів (Q-Former) та великої мовної моделі (LLM) .

Моделі blip-image-captioning-large/base є варіаціями BLIP, що відрізняються лише кількістю параметрів. GIT використовує трансформерну архітектуру для генерування тексту на основі зображень а InstructBLIP розширює архітектуру BLIP, щоб краще розуміти та виконувати складні інструкції.

Модель CLIP фокусується на нульовому навчанні (Zero-shot Learning), тобто може розуміти і категоризувати зображення, яких не зустрічала під час тренування, що демонструє здатність до загального розуміння і адаптації.

Всі ці моделі об'єднує те, що вони навчені на величезних наборах даних зображень та тексту. Це дозволяє їм навчатися розуміти широкий спектр візуальних та мовних концепцій, генеруючи вражаючі та інформативні описи зображень.

Швидкий розвиток штучного інтелекту гарантує, що в найближчі роки з'являться ще більш потужні та досконалі моделі генерування зображень та опису. Ці моделі матимуть значний вплив на різні галузі, від мистецтва та дизайну до освіти та досліджень.

1.3 Виявленя проблем

Незважаючи на значний прогрес у галузі опису зображень та наявність потужних моделей, все ще існують певні виклики та проблеми, які потребують вирішення.

Моделі машинного навчання навчаються на величезних наборах даних, які можуть містити упередження, пов'язані з культурним контекстом, гендером, расою та іншими факторами. Ці упередження можуть призвести до того, що моделі генеруватимуть описи, які є неточними, упередженими або образливими.

Друга проблема пов'язана з обмеженою здатністю до узагальнення. Моделі машинного навчання добре справляються з завданнями, на яких вони були навчені, але їм може бути важко генерувати описи для зображень, які вони ніколи раніше не бачили. Це може призвести до того, що моделі генеруватимуть неточні або неінформативні описи для нових зображень.

Окрім того, моделі потребують великих обсягів даних для навчання. Забезпечення якості та різноманітності цих даних є складним завданням, оскільки вони повинні охоплювати різні типи зображень та сценарії. Це, в свою чергу впливає на швидкість і необхідні обчислювальні ресурси, так як генерація описів зображень у режимі реального часу вимагає значних обчислювальних ресурсів а з зростанням кількості вхідних даних ефективний процес навчання та використання моделі є ключовим.

Моделі іноді можуть генерувати описи, що базуються на фантазії, а не на реальних об'єктах, зображених на фото. Це може створити хибні уявлення про зображення.

Окрім того, адаптація до нових доменів теж є проблемою. Моделі, що навчені на одних наборах даних, можуть неефективно працювати в інших доменах або умовах. Важливо розробляти методи, що дозволяють адаптувати моделі до нових середовищ та типів зображень. Цей аспект звісно втрачає ствою актуальність по мірі зростання розміру даних, на яких навчена модель, проте все ще є аспекти які потребують врахування.

Зазначені проблеми потребують подальшого дослідження та розробки нових методів, щоб зробити моделі опису зображень більш точними, справедливими, ефективними, доступними та етичними.

Звісно, вирішити всі ці проблеми в рамках однієї роботи неможливо. Проте, важливо зосередитися на найактуальніших та найбільш перспективних напрямках досліджень. Так, доцільно дослідити ефективність використання існуючих моделей для опису зображень на фото, які не використовувались при їх навчанні з врахуванням факторів якості згенерованого опису та використаних ресурсів.

1.4 Постановка задачі

Сучасні технології автоматичного опису зображень широко використовуються у різних сферах, таких як соціальні медіа, електронна комерція, допоміжні технології для людей з обмеженими можливостями, а також в автоматизованих системах відеоспостереження та архівації даних. Проте зниження продуктивності моделей на нових, раніше не бачених зображеннях, обмежує їхню ефективність та надійність. Вирішення цієї проблеми є ключовим для подальшого розвитку та впровадження технологій автоматичного опису зображень у різні практичні додатки.

В попередніх пунктах встановлено, що сучасні моделі автоматичного опису зображень демонструють високу ефективність на навчальних наборах даних, проте їх продуктивність може суттєво знижуватися на реальних зображеннях, які не використовувалися при навчанні. Це створює значні виклики у використанні цих моделей у практичних додатках.

Основна мета дослідження полягає в оцінці та покращенні ефективності існуючих моделей автоматичного опису зображень на нових, раніше не використовуваних зображеннях.

Об'єктом дослідження є моделі машинного навчання для автоматичного опису зображень. Це включає як сучасні моделі глибокого навчання, так і більш традиційні підходи, що використовуються для генерування текстових описів на основі візуальних даних.

Предметом дослідження є ефективність цих моделей на нових, раніше не бачених зображеннях. Особлива увага приділяється якості згенерованих описів та використаним ресурсам, зокрема: впливу невідомих зображень на продуктивність моделей; визначенню метрик оцінки якості описів; аналізу методів обробки та генерації тексту у моделях, таких як ViT-GPT2, BLIP та інші.

Для досягнення мети дослідження треба виконати наступні завдання:

- провести аналіз існуючих моделей автоматичного опису зображень;
- вибрати критерії оцінки якості згенерованих описів;
- виконати експериментальне тестування моделей на нових зображеннях та оцінити їх продуктивність за обраними критеріями;
- проаналізувати отримані результати, виявити сильні та слабкі сторони моделей;
- розробити рекомендації щодо покращення моделей для роботи з новими зображеннями.

Для виконання поставлених завдань будуть використовуватись фреймворки машинного навчання, такі як TensorFlow та PyTorch, для розробки, навчання та тестування моделей. Ці інструменти надають потужні можливості для реалізації нейронних мереж, обробки даних та оцінки продуктивності моделей.

Окрім того, бібліотека Hugging Face суттєво спростить роботу з сучасними моделями, такими як BLIP та ViT. Ця бібліотека забезпечує простий інтерфейс для імплементації та налаштування сучасних моделей глибокого навчання.

Таким чином, об'єкт і предмет дослідження охоплюють не лише аналіз поточних моделей та їх продуктивності, але й пошук шляхів підвищення точності та ефективності автоматичного опису зображень у реальних умовах. А використання Python та сучасних інструментів, таких як TensorFlow, PyTorch, і бібліотеки Hugging Face, дозволить ефективно реалізовувати та тестувати моделі машинного навчання, що значно підвищить якість та швидкість проведення досліджень.

2 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ТА АЛГОРИТМІВ

2.1 Трансформер та механізм уваги

Трансформери та механізм уваги стали одними з найбільш впливових інновацій у галузі машинного навчання та обробки природної мови за останнє десятиліття. Вони суттєво змінили підходи до задач перекладу, текстового генерування та інших завдань, що потребують аналізу послідовностей.

Архітектура трансформера була представлена у 2017 році командою дослідників з Google Brain та Університету Торонто. Основна стаття, що описує цю модель, має назву "Attention is All You Need" [11].

Трансформер є архітектурою, яка виключає використання рекурентних та згорткових нейронних мереж, що були основними підходами до обробки послідовних даних (див. рис. 2.1). Натомість він використовує механізм самоуваги (self-attention), який дозволяє моделі одночасно враховувати всі позиції вхідної послідовності, що значно покращує паралелізацію та ефективність обчислень.

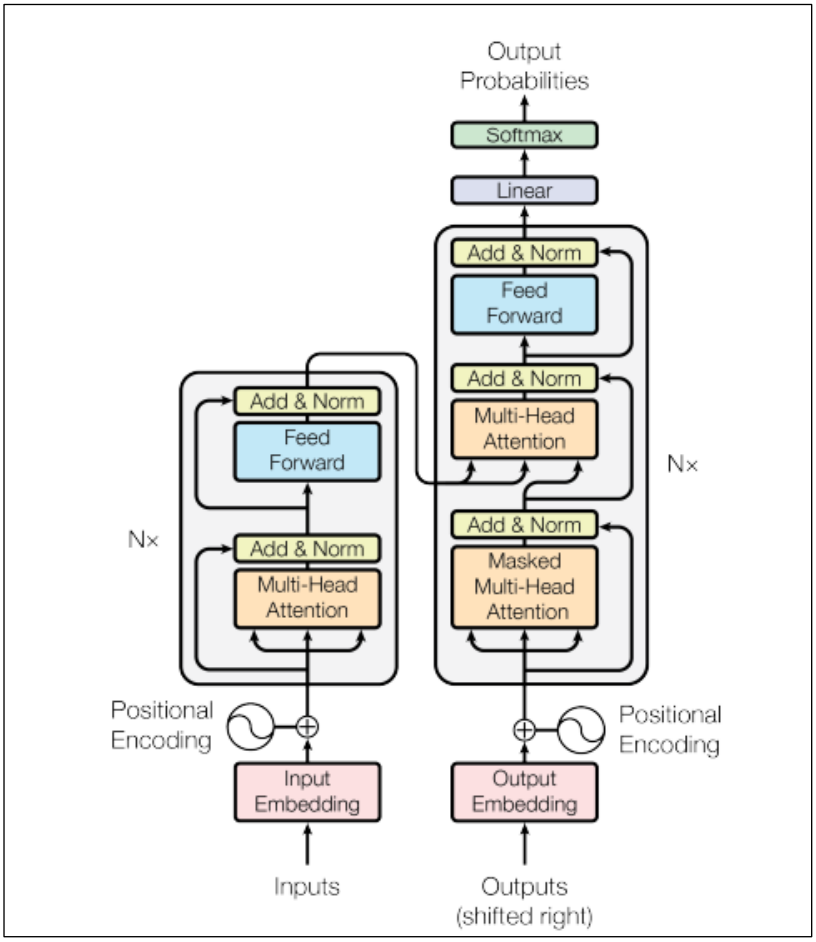


Рисунок 2.1 – Архітектура Трансформер [11]

Трансформер складається з двох компонентів: енкодера та декодера, кожен з яких містить шість ідентичних шарів. Кожен шар енкодера та декодера включає два додаткові шари:

– механізм самоуваги (multi-head self-attention mechanism),

– повнозв'язну мережу (position wise fully connected feed-forward network).

Механізм уваги дозволяє моделі фокусуватися на різних частинах вхідної послідовності під час генерації кожного елемента вихідної послідовності. Самоувага обчислює ваги для кожного елемента вхідної послідовності, визначаючи, наскільки важливий кожен елемент для поточної задачі. Його структура наведена на рисунку 2.2:

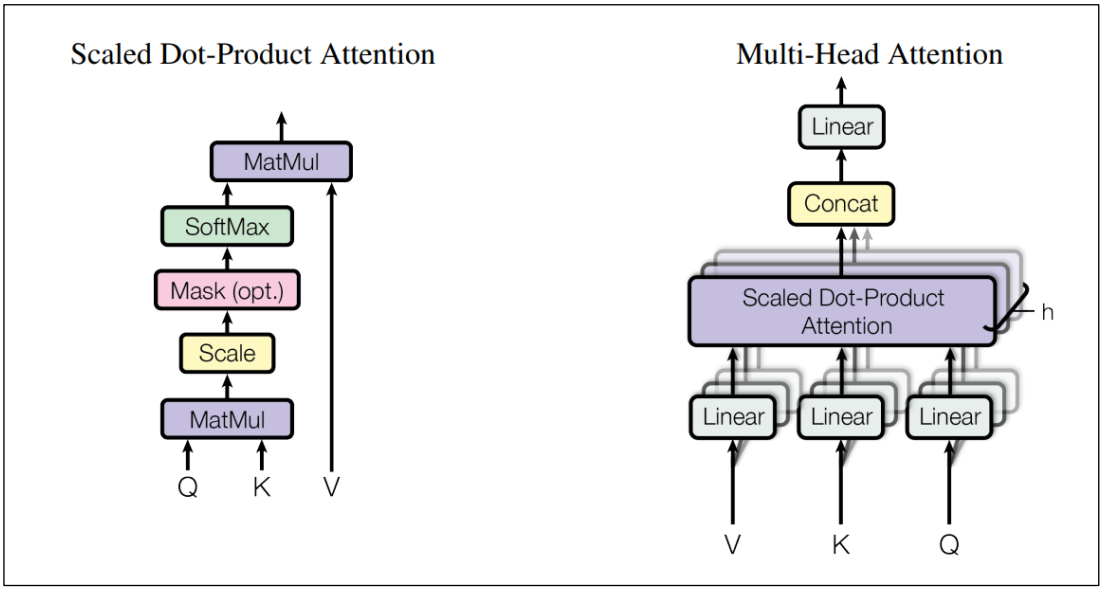


Рисунок 2.2 – Архітектура механізму самоуваги [11]

Query, Key, Value (Q, K, V) є основними компонентами механізму уваги. Вхідна послідовність спочатку перетворюється у три матриці - запити (Q), ключі (K) та значення (V). На основі Q, K та V, модель обчислює ваги за допомогою скалярного добутку (dot-product) між запитами та ключами, що нормалізуються за допомогою функції softmax. Зважені значення (V) комбінуються для формування вихідного представлення.

Використовуючи механізм уваги, модель може зосередитися на обличчі дитини під час опису її емоцій або на конкретному об'єкті, що є центральним у сцені. Візуалізація метанізму уваги наведена на рис. 2.3:

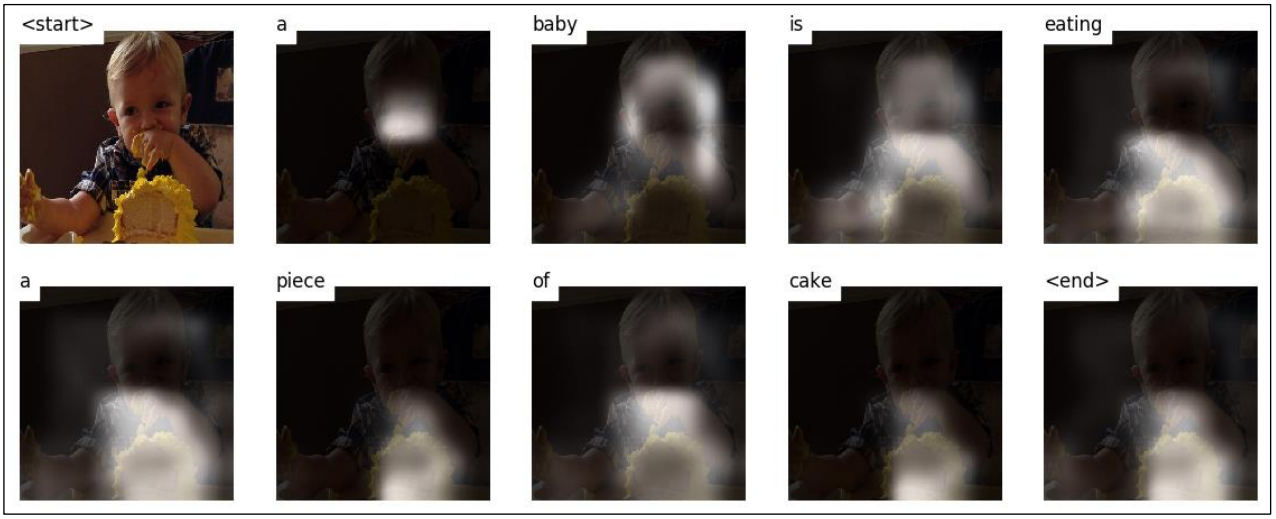


Рисунок 2.3 – Візуалізація механізму уваги для генерації опису [11]

Рисунок, представлений вище, демонструє процес генерування опису зображення за допомогою штучного інтелекту або машинного навчання. Кожен кадр зосереджується на обличчі дитини в центрі, а також на навколишніх предметах. Кожен кадр підписаний окремим словом, які в сукупності утворюють опис зображення: "дитина їсть шматок торта".

Згенерований опис зазвичай починається з токена <start> і закінчується токеном <end>. Це дає змогу моделі розуміти межі генерованого тексту, що особливо важливо під час навчання та генерування вихідних даних, коли модель має вирішити, коли починати та закінчувати текст.

Трансформери пропонують значні переваги для image captioning завдяки своїй здатності генерувати описи кращої якості, гнучкості та масштабованості. Проте, вони також мають певні недоліки, такі як високі обчислювальні витрати, вимоги до даних та складність налаштування [11, 20].

Вибір трансформерів як моделі для image captioning дійсно залежить від конкретних задач та доступних ресурсів. Їхні переваги, такі як гнучкість, масштабованість та паралелізація обчислень, роблять їх привабливими для багатьох завдань.

Однак, високі обчислювальні витрати, потреба в великих наборах даних та складність налаштування можуть стати перешкодою для використання трансформерів у деяких випадках.

2.2 Огляд архітектур нейронних мереж для опису зображень

Автоматичне створення описів для зображень вимагає інтеграції текстових і візуальних даних, що здійснюється за допомогою багатомодальних моделей. Ці моделі поєднують можливості обробки тексту та зображень, що дозволяє їм генерувати зв'язні та релевантні описи для візуальних даних. У цьому розділі ми проаналізуємо архітектури моделей CLIP, BLIP, BLIP-2, InstructBLIP, ViT (vit_bert і vit_gpt2), розглянемо їхні схожості та відмінності, а також переваги та недоліки.

CLIP розроблена компанією OpenAI і представлена 5 січня 2021 року. Це багатомодальна моделль, яка поєднує обробку тексту та зображень у спільному просторі представлення (див. рис. 2.4).

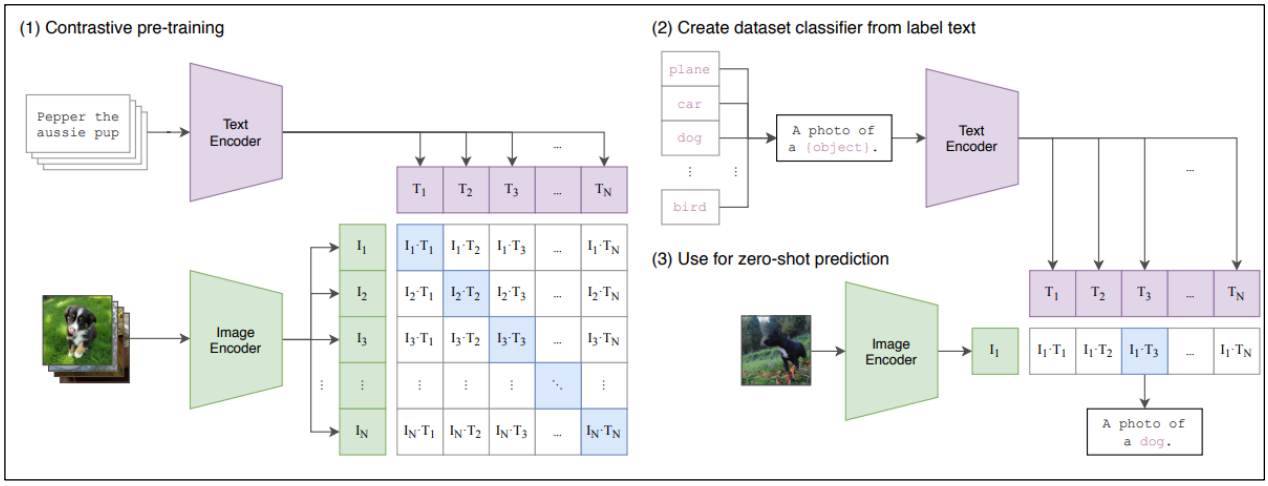


Рисунок 2.4 – Візуалізація процесу навчання моделі CLIP [12]

Основна ідея полягає в тому, що модель навчається відповідати зображенням та текстовим описам через контрастивне навчання. Це дозволяє CLIP досягати високої ефективності в різних задачах комп'ютерного зору без необхідності додаткового навчання на специфічних наборах даних.

Контрастивне переднє навчання (1) пропускає текст "Pepper the aussie pup" через енкодер, який перетворює його на серію векторів (T_1, T_2, ..., T_N). Зображення собаки проходить через енкодер зображень, який перетворює його на серію векторів (I_1, I_2, ..., I_N). Ці вектори комбінуються (I_1T_1, I_1T_2, ..., I_NT_N), які використовуються для навчання моделі визначати схожість між зображеннями та текстом.

Після цього формується набір міток з різними об’єктами, на основі якого створюється класифікатор (2). Ці описи знову проходять через текстовий енкодер для створення векторів, які слугують як класифікатор для різних категорій.

Далі зображення використовується для нульового прогнозування (3). Зображення проходить через енкодер зображень, щоб перетворити його на вектор (L_1). Цей вектор порівнюється з векторами текстового класифікатора, створеними раніше, для ідентифікації зображення без попереднього навчання на конкретних даних зображень.

Однією з головних переваг CLIP є її універсальність та здатність до узагальнення, однак вона може вимагати значних обчислювальних ресурсів для навчання та виконання.

BLIP використовує аналогічний підхід до інтеграції тексту та зображень, як і CLIP, але зосереджується на покращенні продуктивності через більш ефективне використання даних для навчання. BLIP також використовує механізми уваги для фокусування на важливих частинах зображень під час генерації тексту, що підвищує точність і релевантність описів. Однією з відмінностей BLIP є акцент на використанні великих, різноманітних наборів даних для навчання, що дозволяє моделі краще адаптуватися до нових задач. Проте, як і CLIP, BLIP вимагає значних ресурсів для навчання.

На рисунку 2.5 представлена схема, яка ілюструє роботу нейронної мережі BLIP для опису зображень.

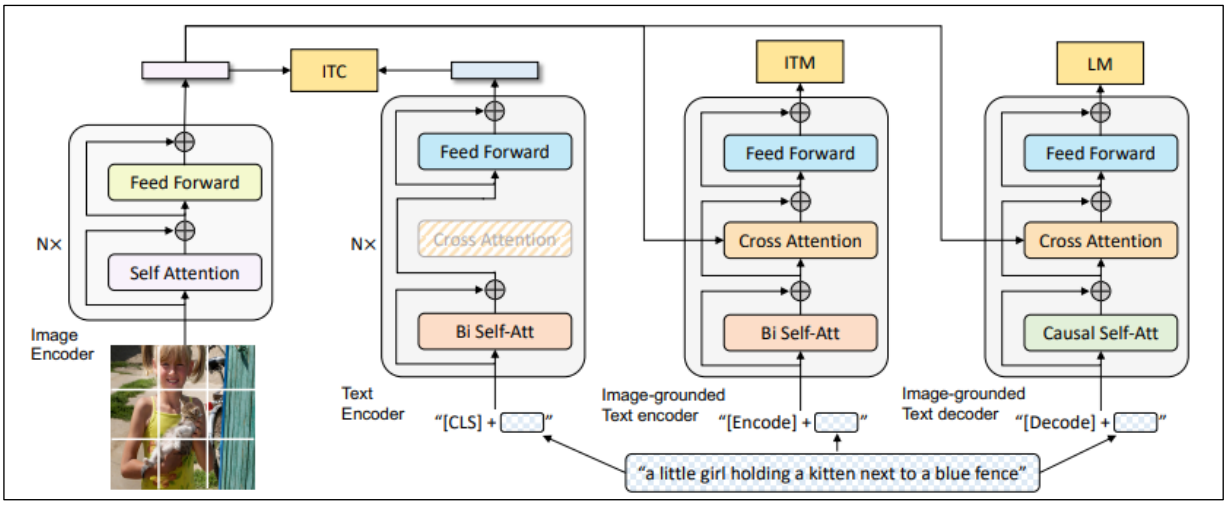


Рисунок 2.5 – Процес роботи моделі BLIP [13]

BLIP складається з трьох основних модулів:

– ITC (Image-Text Contrastive learning) використовує контрастивне навчання, щоб допомогти моделі вивчити спільні візуальні та текстові представлення;
– ITM (Image-Text Matching) відповідає за зіставлення зображень з відповідними описами. Він використовує двонаправлену та перехресну увагу, щоб краще зв'язати зображення та текст;
– LM (Language Modeling) являє собою мовну модель, яка генерує текст, наприклад, описи зображень. Він використовує причинну увагу, щоб забезпечити послідовне генерування тексту на основі попереднього контексту.

Кожен модуль складається з мережі прямого зв'язку (Feed Forward) та шарів самоуваги, які повторюються N разів [13]. Для генерування опису зображення пара зображення-текст проходить через візуальний перетворювач один раз і текстовий перетворювач три рази.

Важливо зазначити, що кодер і декодер тексту мають спільні параметри, за винятком шарів самоуваги. Це робиться для ефективного попереднього навчання за допомогою багатозадачного навчання.

BLIP-2 розвиває ідеї BLIP, додаючи більше шарів та покращуючи механізми уваги для більш глибокого аналізу зображень і тексту [14]. Ця модель спрямована на досягнення ще вищої точності та деталізації описів завдяки вдосконаленим алгоритмам обробки даних. Структура моделі наведена на рис. 2.6.

BLIP-2 зберігає переваги BLIP у здатності до узагальнення та адаптації, але ще більше підвищує вимоги до обчислювальних ресурсів.

InstructBLIP (рис. 2.7) є новою варіацією BLIP, яка інтегрує можливості навчання з підкріпленням (reinforcement learning) для покращення генерації тексту на основі інструкцій. Ця модель дозволяє генерувати більш контекстуально відповідні описи, враховуючи специфічні вимоги або інструкції. Основною перевагою InstructBLIP є її гнучкість у генерації тексту під конкретні завдання, однак це також додає складності та потребує додаткових налаштувань.

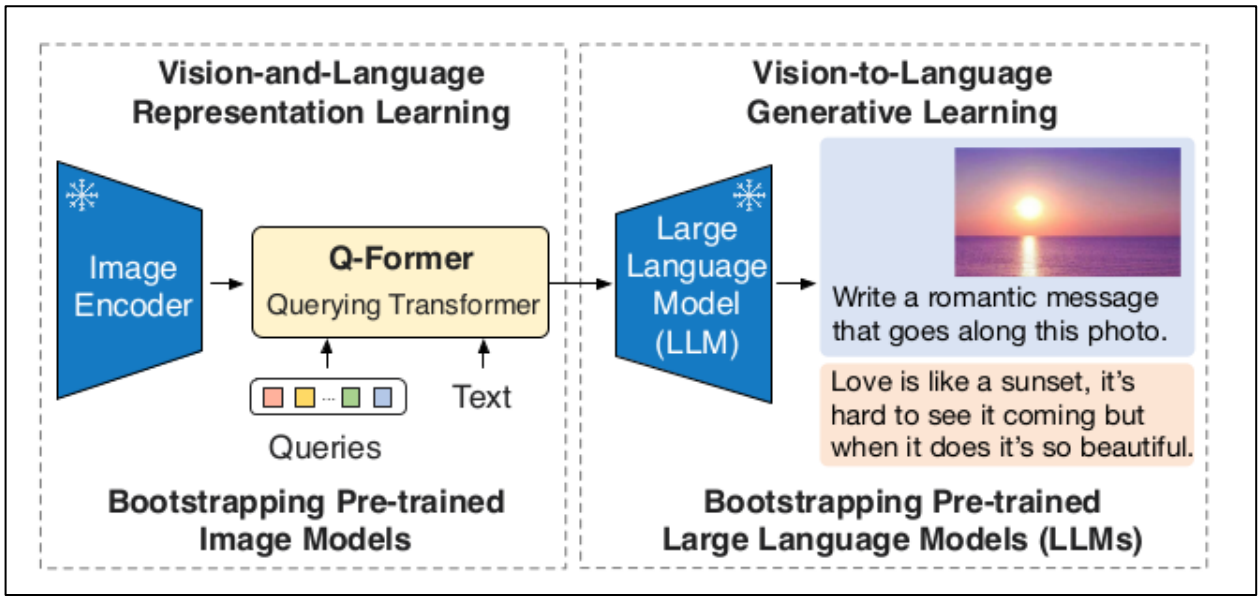


Рисунок 2.6 - Структура BLIP-2 з двоетапною стратегією підготовки [14]

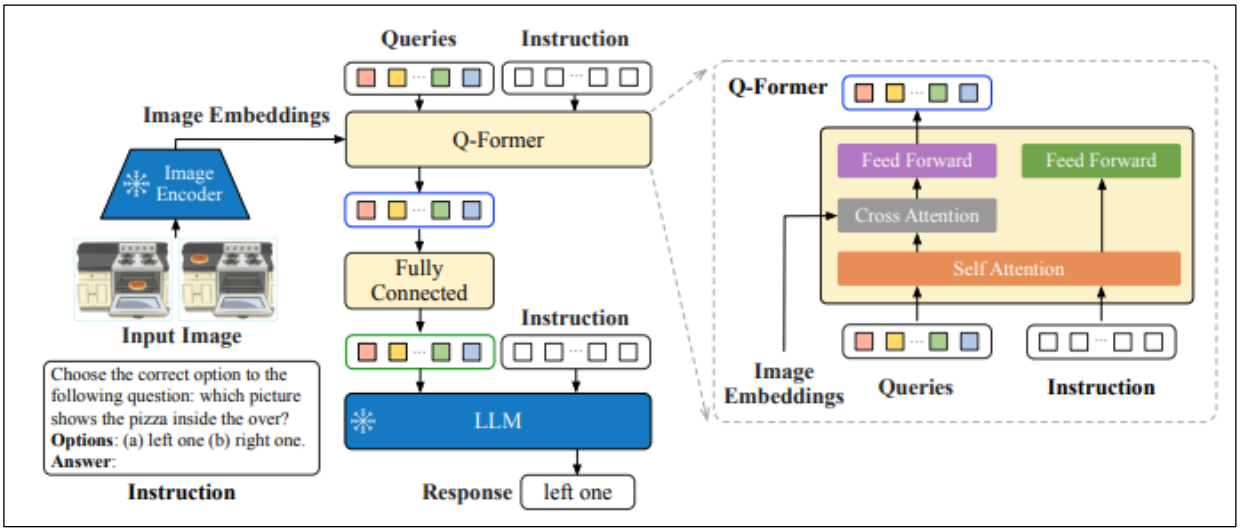


Рисунок 2.7 - Архітектура моделі InstructBLIP [15]

Моделі ViT, такі як vit_bert та vit_gpt2, використовують архітектуру трансформера для обробки візуальних даних. ViT розбиває зображення на патчі та обробляє їх як послідовність, що дозволяє враховувати глобальні залежності між частинами зображення (див. рис. 2.8). Модель ViT-BERT поєднує Vision Transformer з BERT для обробки тексту, тоді як ViT-GPT2 інтегрує GPT-2 для генерації тексту. Це дозволяє створювати детальні та релевантні описи зображень, однак обидві моделі потребують значних обчислювальних ресурсів для навчання та виконання.

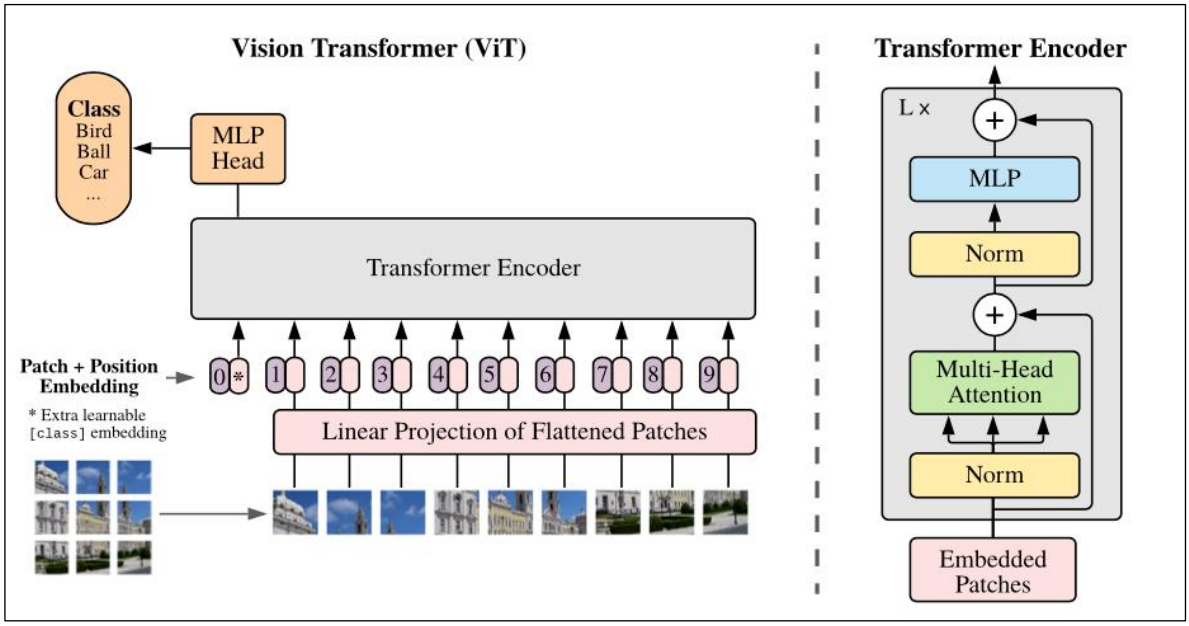


Рисунок 2.8 – Архітектура моделі Vision Transformer [16]

Моделі CLIP, BLIP, BLIP-2, InstructBLIP та ViT мають спільну мету – інтегрувати візуальні та текстові дані для автоматичного створення описів зображень. Вони використовують механізми уваги та трансформери для досягнення високої точності та релевантності описів. Основні відмінності між ними полягають у підходах до навчання та використанні даних. CLIP зосереджується на контрастивному навчанні, BLIP і BLIP-2 використовують великі набори даних та вдосконалені механізми уваги, а InstructBLIP додає можливості навчання з підкріпленням. Моделі ViT обробляють зображення за допомогою трансформерів, що дозволяє враховувати глобальні залежності візуальних ознак.

Переваги всіх цих моделей включають високу точність, здатність до узагальнення та інтеграцію мультимодальних даних. Однак вони також мають недоліки, такі як високі вимоги до обчислювальних ресурсів та складність налаштування [17, 18]. Моделі, що використовують трансформери, як ViT, можуть забезпечувати детальніші описи, але потребують значних ресурсів для ефективної роботи. InstructBLIP додає гнучкість у генерацію тексту, але вимагає додаткових налаштувань для кожного конкретного завдання.

Розглянуті моделі демонструють різні підходи до інтеграції текстових та візуальних даних для автоматичного опису зображень. Вибір конкретної моделі залежить від завдань та доступних ресурсів, але всі вони сприяють покращенню якості та релевантності згенерованих описів. Подальший розвиток цих архітектур може включати оптимізацію обчислювальних витрат та підвищення адаптивності до нових типів даних та задач.

2.3 Критерії оцінки якості генерованих описів

Для оцінки якості автоматично згенерованих описів зображень використовуються різні метрики, що дозволяють кількісно оцінити відповідність згенерованого тексту референтним описам. Нижче наведено основні критерії та метрики, які застосовуються для цього завдання.

BLEU (Bilingual Evaluation Understudy) є однією з найпопулярніших метрик для оцінки якості машинного перекладу і генерування тексту. Вона вимірює відповідність згенерованого тексту до одного або більше референтних текстів за допомогою порівняння n-грам. BLEU має кілька варіантів, залежно від кількості врахованих n-грам (наприклад, BLEU-1, BLEU-2, BLEU-4).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) є набором метрик, що оцінюють якість текстових узагальнень шляхом порівняння з референтними текстами. Найбільш використовуваними варіантами є ROUGE-N (вимірює відповідність n-грам), ROUGE-L (вимірює довжину найдовшої загальної підпослідовності), та ROUGE-S (вимірює відповідність біграм).

CIDEr (Consensus-based Image Description Evaluation) оцінює текстові описи шляхом порівняння їх з кількома референтними описами, враховуючи частоту n-грам та їх значимість у контексті. CIDEr розраховує ваги для n-грам на основі їх частоти в описах, щоб уникнути переоцінки частих фраз.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) враховує відповідність слів, їх синоніми, варіанти форм і порядок слів. Метрика використовує вирівнювання на основі словників та лексичних ресурсів для точнішої оцінки відповідності тексту.

SPICE (Semantic Propositional Image Caption Evaluation) оцінює опис зображень на основі семантичних дерев, що відображають значення тексту. Ця метрика вимірює відповідність семантичної інформації між згенерованим текстом та референтним описом.

Окрім того, ще є 2 показники, що використовують можливості CLIP для об'єднання візуальних і текстових представлень у спільному просторі, що дозволяє оцінювати відповідність між зображенням і текстом.

CLIP-S (Contrastive Language-Image Pre-training Similarity) - метрика, яка оцінює відповідність між зображенням і текстом, використовуючи модель CLIP для обчислення косинусної подібності між вектором зображення та вектором згенерованого текстового опису. Ця метрика є reference-free, тобто не потребує наявності референтних текстів для оцінки якості.

RefCLIP-S (Reference-based CLIP Similarity) - це розширена версія метрики CLIP-S, яка включає референсні описи для більш точного оцінювання. RefCLIP-S обчислює гармонічне середнє між косинусною подібністю зображення та згенерованого опису і косинусною подібністю між референтним і згенерованим описами. Це дозволяє враховувати додаткову інформацію з референтних описів для більш точної оцінки якості згенерованого тексту. Узагальнимо переваги та недоліки кожного з показників у таблиці 2.1.

Таблиця 2.1 – Переваги та недоліки показників оцінки згенерованого тексту (таблиця виконана самостійно)

Показник	Переваги	Недоліки
BLEU	Швидкість обчислення , підходить для порівняння моделей на великих наборах даних.	Не враховує семантичну відповідність та синтаксичні особливості тексту.
ROUGE	Враховує довжину та важливість великих спільних фрагментів.	Не повністю відображає семантичну релевантність.
CIDEr	Враховує значимість n-грам.	Складність обчислення.
METEOR	Враховує семантичну релевантність та гнучкість мови.	Потребує додаткових словників та лексичних баз.
SPICE	Врахування семантичної відповідності та структури.	Складність обчислення.
CLIP-S	Не потребує референсних описів,	Обмежена контекстуальна точність.

Кінець таблиці 2.1

Показник	Переваги	Недоліки
	швидкість роботи.	
RefCLIP-S	Краща точність порівняно з CLIP-S, врахування семантичної відповідності.	Залежність від референсних описів, складність обчислення.

Вибір відповідної метрики для оцінки якості згенерованих описів зображень залежить від конкретних вимог та контексту задачі. Для комплексної оцінки часто використовують комбінацію кількох метрик[19], що дозволяє отримати більш збалансовану оцінку якості тексту з різних аспектів. Кожна з метрик має свої переваги та обмеження, що необхідно враховувати при інтерпретації результатів.

2.4 Інформаційна підготовка прийняття рішень

В цьому розділі ми проведемо порівняльний аналіз різних моделей, що використовуються для опису зображень, з метою забезпечення інформаційної підготовки для прийняття рішень. Сучасні моделі машинного навчання значно відрізняються за своїми характеристиками та іншими параметрами .

Для порівняння моделей, які використовуються для опису зображень, ми розглянемо наступні параметри:

- кількість параметрів моделі, що визначає здатність моделі до навчання та її обчислювальну складність. Моделі з більшою кількістю параметрів мають більшу здатність до навчання та можуть зберігати більше інформації, але вони також потребують більше ресурсів для навчання та використання;
- кількість параметрів токенізатора, що дозволяє обробляти більший словниковий запас, що може бути корисним для роботи з багатшими текстовими даними. Водночас це може збільшити обчислювальні витрати на попередню обробку тексту;
- розмір прихованих шарів, що означає, що модель може зберігати більше інформації в кожному шарі. Це зазвичай покращує здатність моделі до

генералізації та точність, але також збільшує обчислювальні витрати та вимоги до пам'яті;

– кількість прихованих шарів, яка зазвичай покращує здатність моделі вивчати більш складні функції та залежності у даних, що може підвищити точність. Проте це також призводить до збільшення обчислювальної складності та часу навчання;
– ліцензія визначає, чи можна використовувати модель для комерційних цілей, змінювати її або розкривати параметри. Це важливий аспект для вибору моделі, особливо у бізнес-середовищі;
– кількість завантажень моделі на платформі Hugging Face може відображати її популярність та корисність у реальних застосуваннях.

Додатково можна було б додати такий показник як мультимодальність, проте ми обирали тільки моделі, які її підтримують. Адже це вимога часу та тренд у розвитку машинного навчання. Значення критеріїв та їх вагові коефіцієнти наведені в таблиці 2.2:

Таблиця 2.2 – Значення критеріїв вибору моделей (таблиця виконана самостійно)

Модель	Параметри моделі	Параметри токенізатора	Приховані шари для тексту	Кількість прихованих шарів	Ліцензія	Кількість завантажень
git_base	153147648	30,522	768	6	MIT	551836
git_large	394196026	30,522	768	24	MIT	844
blip_base	247414076	30,522	768	12	bsd-3-clause	1098881
blip_large	469732924	30,522	1024	16	bsd-3-clause	810998
BLIP2	135764352	30,522	768	12	MIT	170011
vit_gpt_2	239195904	50,257	768	12	apache-2.0	2402781
vit_bert	224270394	30,522	768	12	apache-2.0	175
flickr8	239195904	50,257	768	12	apache-2.0	171
CLIP	151277313	30,522	512	12	MIT	50439600
Вагові коефіцієнти	0.25	0.2	0.2	0.15	0.1	0.1

Проаналізуємо значення критеріїв, які були обрані для моделі. Кількісні параметри ми не змінюємо, так як у них є абсолютні значення.

Тільки ліцензії повинні бути перетворені. Враховуючи, що нас цікавлять не тільки дослідницькі цілі, а й потенційне комерційне використання, вони отримують такі бали:

– MIT: Відкритий код, комерційне використання дозволене – 5 балів;
– BSD-3-Clause: Відкритий код, комерційне використання дозволене, з обов'язковим зазначенням авторських прав – 5 балів;
– Apache-2.0: Відкритий код, комерційне використання дозволене, з обов'язковим наданням повідомлень про зміни – 3 бали.

Для вирішення задачі, було обрано лінійну згортку. Так як одні параметри є важливішими за інші і мають різні розмірності, має сенс використовувати лінійну адитивну згортку з ваговими коефіцієнтами.

Формули 2.1 для розрахунку:

$$Z^* = max \sum_{j=1}^{n} \alpha_j \beta_j a_j, \tag{2.1}$$

де α_j – нормуючі множники;

β_j – вагові коефіцієнти, що відображають відносний внесок окремих критеріїв до загального критерію.

Для нормування даних використаємо критерій mini-max (формула 2.2):

$$f = \frac{f_i - f_{min}}{f_{max} - f_{min}}, \tag{2.2}$$

де f – нормоване значення;

f_i – початкове значення, яке підлягає нормуванню;

f_{min} – мінімальне значення в наборі даних;

f_{max} – максимальне значення в наборі даних.

Здійснимо нормування показників у таблиці, враховуючи мінімальні та максимальні значення. Введемо вагові коефіцієнти для наших критеріїв та обчислимо згорткову оптимізаційну модель Z^*, використовуючи лінійну адитивну згортку з ваговими коефіцієнтами.

Результати оцінки наведено в таблиці 2.3.

Таблиця 2.3 – Результати оцінки альтернатив за критерієм корисності (таблиця виконана самостійно)

Модель	Z^*
git_base	0.077887644
git_large	0.126688204
blip_base	0.096583131
blip_large	0.128037517
BLIP2	0.085979187
vit_gpt2	0.105707793
vit_bert	0.087625761
Flickr8	0.101383095
CLIP	0.170298078

Згідно з отриманими результатами, найкращі результати мають бути у моделі CLIP. Дещо гірші але максимально наближені результати також мають git_large та blip_large. Це може свідчити про їхню високу ефективність та здатність успішно виконувати поставлені завдання.

Хоча на папері CLIP виглядає цікавим варіантом, ми не будемо його використовувати як модель для розпізнавання. Ця модель виступить у ролі перевіряючого фактора (розрахунок показників CLIP-S та REF-CLIP-S, про які говорили в пункті 2.3).

Окрім того, доцільно відмовитись від кількості завантажень, тому що дані там досить нерівномірні і не можна виключати того факту, що завантаження за останній місяць не є релевантними, адже більш нові моделі апріорі будуть мати вищі показники. Результати розрахунків наведено в таблиці 2.4.

Таблиця 2.4 – Оцінка альтернатив (таблиця виконана самостійно)

Модель	Z^*
git_base	0.091350035
git_large	0.169960122
blip_base	0.119640943
blip_large	0.168422732
BLIP2	0.102893484
vit_gpt_2	0.126441013
vit_bert_flickr8k	0.110287038
flickr8k	0.126441013

Як видно з таблиці 2.4, large версії моделей все ще найкращі, хоча відрив у між варіантами суттєво зменшився.

Вагові коефіцієнти при цьому були змінені, і кількість параметрів отримала найбільшу вагу – 30%, щодо інших показників: кількість прихованих шарів (+5%), розмір прихованих шарів (-5%).

Для більш детального аналізу та підтвердження цих висновків, перейдемо до експериментальної частини, де на практиці перевіримо результати. Під час експерименту ми порівняємо продуктивність моделей у реальних умовах. Це дозволить нам оцінити їхню ефективність у виконанні конкретних завдань, враховуючи різні фактори та умови використання.

3 ЕКСПЕРИМЕНТ

3.1 Розробка вимог до програмної системи

Створення програмної системи для автоматичного опису зображень вимагає чіткого визначення вимог, які забезпечать її функціональність, надійність та ефективність. Ці вимоги базуються на потребах користувачів, цільових задачах, а також на технологічних можливостях.

Програмна система має бути реалізована у вигляді веб-сайту. Мета розробки веб-сайту:

- надати зручний інтерфейс для завантаження зображень та тестування різних моделей Image Captioning;
- відображати результати роботи моделей, включаючи згенеровані описи та оцінки якості;
- зберігати історію оброблених зображень та описів;
- надати можливість оцінювати якість описів вручну.

Сформуємо функціональні вимоги.

Інтерфейс користувача має бути простим та зрозумілим. Доцільно використати фреймворк Bootstrap, який дозволяє швидко та ефективно створювати прототипи застосунків.

Основний екран має включати форму, де можна завантажити одне або більше зображень з локального диску, обрати модель для розпізнавання та запустити процес генерації описів. Після завершення роботи моделей користувач бачить зображення та згенеровані описи.

Сторінка «History» - відображає всі опрацьовані зображення, що зберігаються в системі. При кліку на зображення має відкритись сторінка зображення.

Сторінка зображення – це конкретне зображення з можливістю перегляду згенерованих описів, додавання референтних текстів для оцінки та запуску генерації описів іншими моделями. Відображення таблиці з результатами оцінки якості описів за метриками BLEU, ROUGE, ROUGE2, ROUGEL, CIDEr, METEOR.

Завантажені зображення та їх описи повинні зберігатися в базі даних (sqlite3).

Веб-сайт повинен зберігати інформацію про використовувані моделі та їх параметри.

Для роботи з моделями Image Captioning веб-сайт повинен використовувати бібліотеки transformers, nltk, torch, Pillow, CLIP та інші. При цьому, моделі машинного навчання повинні завантажуватися з Hugging Face та зберігатись длокально.

В якості веб сервера використовуємо Flask для реалізації серверної логіки та обробки запитів. Flask обрано через його легкість, простоту налаштування та інтеграцію з бібліотеками машинного навчання на Python, такими як transformers, nltk, torch, Pillow та CLIP. Flask також надає гнучкість та розширюваність, що дозволяє швидко розгортати прототипи та невеликі проекти. Django, хоча і є потужнішим фреймворком, може бути надмірним для цього проекту через його складність та важкість налаштування для простих задач.

Окрім функціональних можливостей, веб-сайт для тестування та порівняння моделей Image Captioning повинен відповідати низці нефукціональних вимог, які гарантують його якісну роботу та позитивний досвід користувачів.

Система повинна генерувати описи зображень та розраховувати метрики якості за мінімальний час. Це важливо для забезпечення плавної роботи та зручного користування.

Веб-сайт має бути спроектований таким чином, щоб він міг обробляти велику кількість зображень без значного зниження продуктивності. Це дозволить використовувати його для дослідницьких проектів та комерційних цілей, що потребують обробки великих обсягів даних.

Система повинна працювати стабільно навіть під високим навантаженням, без збоїв та помилок. Це критично важливо для забезпечення безперебійної роботи та довіри користувачів.

У разі збоїв система повинна мати механізми відновлення, які дозволять їй швидко повернутися до нормального функціонування без втрати даних. Це

мінімізує негативний вплив на користувачів та гарантує збереження цінної інформації.

Відповідність веб-сайту всім вищезазначеним нефукціональним вимогам гарантуватиме його високу якість, зручність використання та надійність, що робить його цінним інструментом для досліджень та тестування моделей Image Captioning.

3.2 Підготовка зображень для експерименту

Для проведення експерименту необхідно підготувати набір зображень, які відповідатимуть певним вимогам, щоб забезпечити об'єктивне тестування моделей для автоматичного опису зображень. Кожна група зображень повинна містити мінімум по 10 зображень, що відповідають специфічним критеріям.

Перша група складається з простих зображень, на яких зображений один об'єкт. Ці об'єкти повинні бути добре освітлені, без перекривань іншими елементами, та відповідати тим об'єктам, які зустрічалися у навчальних наборах даних моделей. Зображення мають бути максимально зрозумілими та легкими для обробки, що дозволить моделям генерувати точні та релевантні описи. Зображення з цієї групи наведені на рис. 3.1.

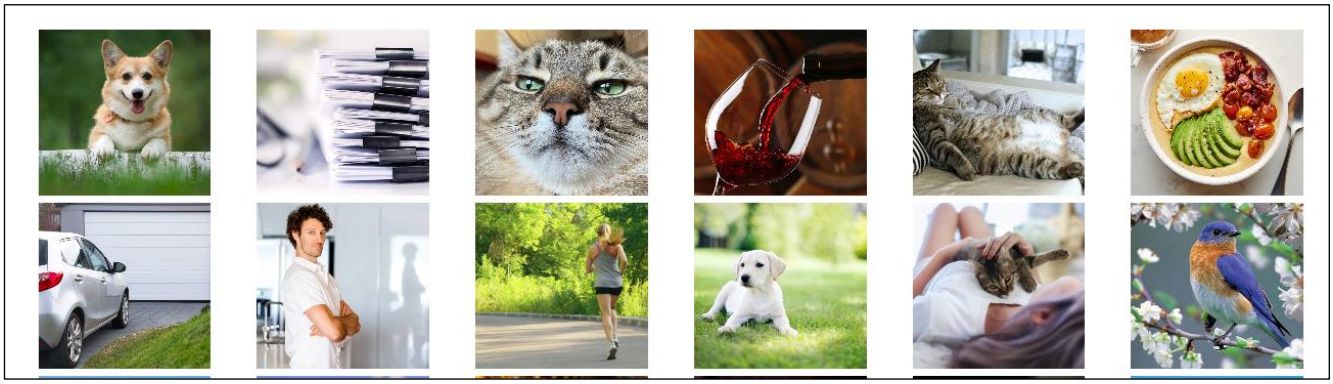


Рисунок 3.1 – Зображення для групи 1 [Рисунок виконаний самостійно]

Друга група включає фотографії (див. рис. 3.2), зроблені вночі або за умов поганого освітлення. На кожному зображенні повинен бути один об'єкт, який оточений світлом або має чіткі контури, щоб його можна було легко розпізнати.

Ці зображення допоможуть оцінити здатність моделей працювати з менш оптимальними умовами освітлення.

Рисунок 3.2 – Зображення для групи 2 [Рисунок виконаний самостійно]

Третя група містить зображення зі складними сценами, де представлено багато однотипних об'єктів. Ці зображення (див. рис. 3.3) створюють додаткові виклики для моделей, оскільки вони повинні не тільки розпізнати окремі об'єкти, але й правильно їх описати в контексті всієї сцени. Ця група дозволяє оцінити, наскільки ефективно моделі можуть працювати зі складними візуальними даними.

Рисунок 3.3 – Зображення для групи 3 [Рисунок виконаний самостійно]

Четверта група складається зі зображень відомих особистостей, знятих у різних умовах. Ці зображення (див. рис. 3.4) допоможуть визначити, наскільки добре моделі можуть розпізнавати та описувати відомі об'єкти, які можуть мати значну кількість варіацій у зовнішньому вигляді.

П'ята група (див. рис. 3.5) включає зображення з фоном, що містить багато деталей, наприклад, міські пейзажі або природні сцени. Такі зображення

допоможуть оцінити здатність моделей фокусуватися на основних об'єктах і не втрачати контекст у складному середовищі.

Рисунок 3.4 – Зображення для групи 4 [Рисунок виконаний самостійно]

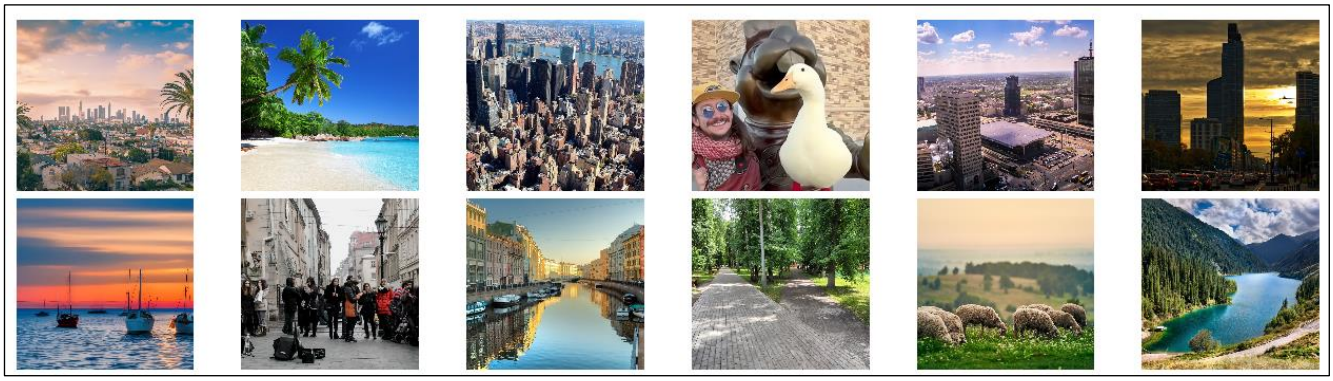


Рисунок 3.5 – Зображення для групи 5 [Рисунок виконаний самостійно]

Шоста група складається з зображень, що містять текстові елементи, такі як дорожні знаки, плакати або книги. Ця група дозволить оцінити здатність моделей розпізнавати та інтерпретувати текст на зображеннях.

Рисунок 3.6 – Зображення для групи 6 [Рисунок виконаний самостійно]

Незважаючи на те, що розпізнавання тексту це окрема задача і для отримання гарних результатів варто використовувати спеціально треновані моделі, спробуємо оцінити універсальність моделі.

Таким чином, підготовка різноманітного набору зображень для експерименту дозволить всебічно оцінити продуктивність та здатність моделей до автоматичного опису зображень в різних умовах та для різних типів сцен.

Всі зображення зібрані з мережі інтернет. Не використовувались картинки з фотостоків та різноманітних датасетів. Всі зображення у форматі jpeg, розміром більше ніж 500*500 пікселів. Кожне зображення має відповідні анотації, які виступатимуть у ролі референсів для прорахунку якості згенерованих описів.

Окрім того, що будуть рахуватись статистичні показники якості згенерованих описів, ми також використаємо метод експертних оцінок, коли люди будуть оцінювати описи та те, як вони підходять до тієї чи іншої картинки. Шкала оцінок варіюватиметься від 0 до 5.

Нуль балів – опис взагалі не відповідає зображенню. Наприклад, на картинці зображена повітряна кулька, а в описі йдеться про людину.

Один бал – опис містить деякі елементи, які можуть бути на зображенні, але в цілому дуже неточний. Наприклад, на зображенні є людина, але опис говорить про тварину.

Два бали – опис частково відповідає зображенню, але містить значні помилки або пропускає важливі деталі. Наприклад, на зображенні дві собаки, а в описі згадується лише одна тварина.

Три бали – опис в цілому відповідає зображенню, але може пропускати деякі деталі або містити незначні неточності. Наприклад, на зображенні людина тримає кота, а в описі згадується лише людина.

Чотири бали – опис майже повністю відповідає зображенню, містить більшість важливих деталей, але може не враховувати деякі стилістичні аспекти або відносини між об'єктами. Наприклад, на зображенні двоє дітей грають з м'ячем, а в описі згадується тільки гра дітей.

П’ять балів – гарний опис, окрім правильно визначених об'єктів, включає стилістичні деталі або відносини між ними. Наприклад, на зображенні два брати грають у футбол на трав'яному полі під вечірнє сонце, і в описі враховані всі ці аспекти.

Будь-яка експертна оцінка є суб'єктивною, проте буде цікаво порівняти оцінки експертів із результатами статистичних показників. Це дозволить не лише отримати кількісні дані про точність моделей, але й оцінити їх з точки зору людського сприйняття, що може виявити важливі аспекти, не відображені в статистичних метриках. Порівняння експертних оцінок із автоматичними показниками може також допомогти виявити, де моделі особливо добре або погано справляються з певними типами зображень, надаючи можливість для подальшого вдосконалення алгоритмів.

3.3 Практична реалізація

Розроблено веб-сайт, який реалізовано на Flask, що дозволяє спростити користувачам роботу з різними моделями автоматичного опису зображень.

Головна сторінка (рис. 3.7) має простий інтерфейс, що дозволяє завантажити від 1 до 10 зображень, обрати модель якою ми хочемо згенерувати опис та натиснути кнопку «Process».

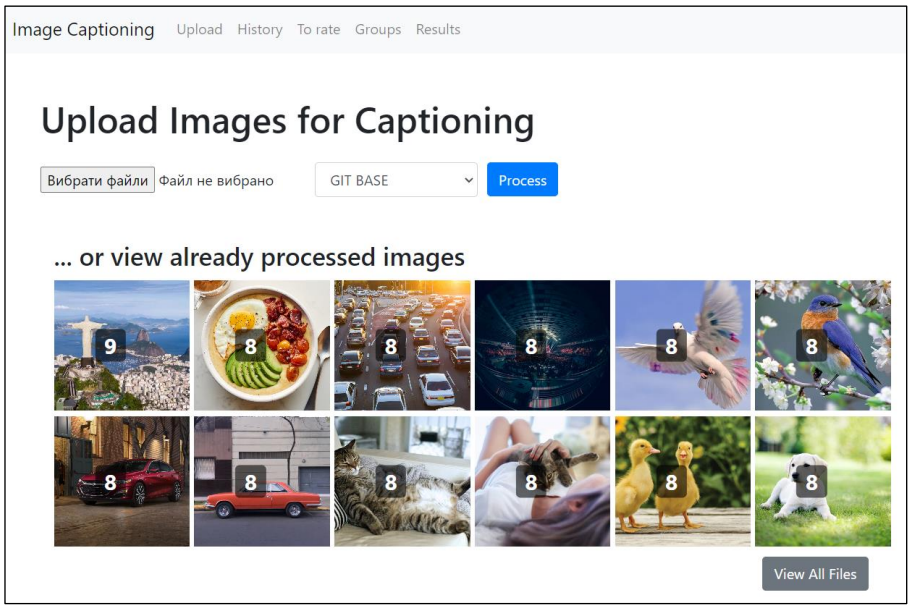


Рисунок 3.7 – Головна сторінка [Рисунок виконаний самостійно]

Також на головній сторінці показано вже опрацьовані зображення та ті зображення, в яких недостатньо інформації для оцінки зображення.

Якщо відкрити сторінку зображення (див. рис. 3.8), то ми побачимо збільшене зображення, групу в яку входе зображення, згенеровані описи, оцінки експертів (якщо вже оцінено) з можливістю їх змінити.

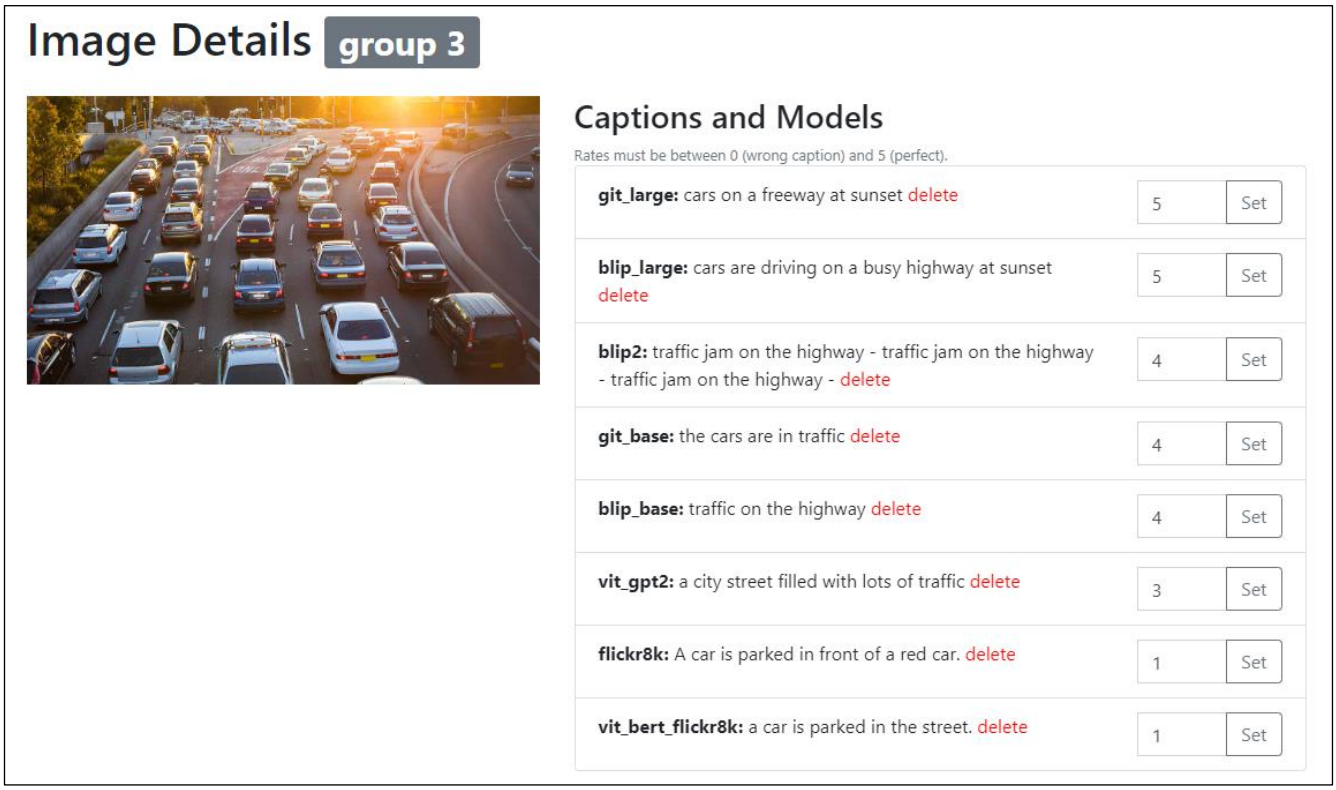
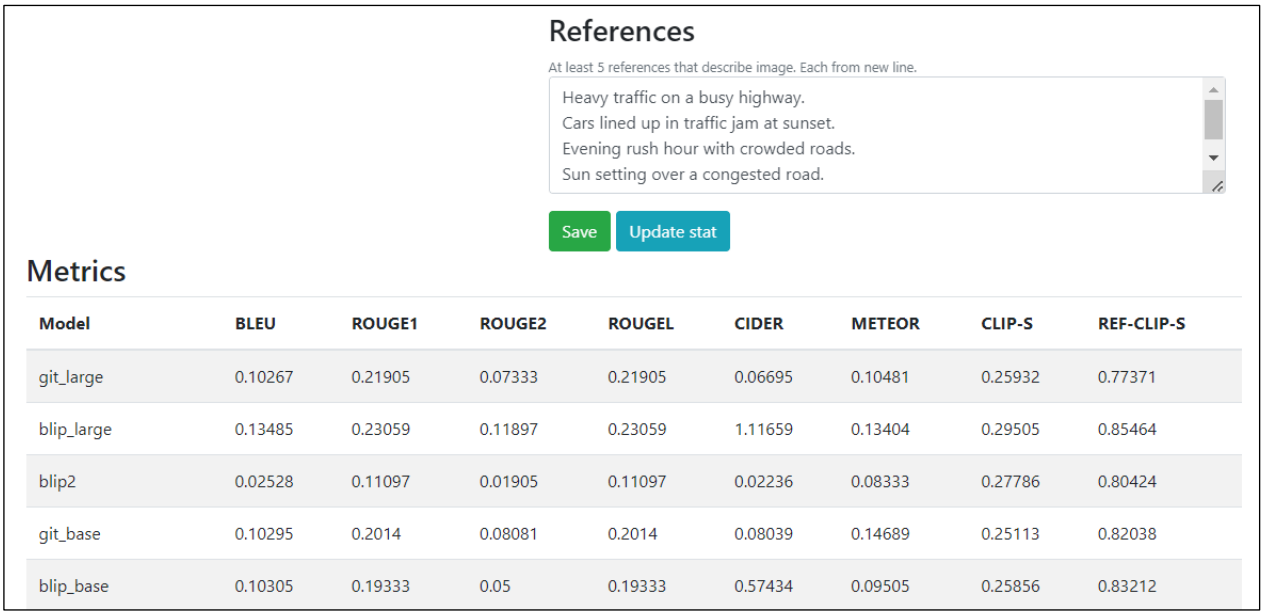


Рисунок 3.8 – Сторінка зображення (екран 1) [Рисунок виконаний самостійно]

На другому екрані (див. рис. 3.9) буде відображено поле з референсами (з можливістю редагувати), метрики для згенерованих описів, запустити перерахунок метрик та моделей для генерації, якщо є такі, що ще не опрацювали зображення.

Окрім головної сторінки та картки зображення передбачено ще кілька функціональних сторінок. На сторінці «History» відображені всі зображення з можливістю фільтрувати по групі. Сторінка «To Rate» виводить список зображень які ще не оцінені експертами та має функціонал оцінки. На сторінці «Groups» можна переглянути існуючі та створити нові групи для зображень. «Results» відображає статистичні показники щодо моделей та зображень.

Щодо моделей, то інтегровані наступні моделі: git_base, git_large, blip_base, blip_large, blip2, flickr8, vit_bert_flickr8, vit_gpt2.



Model	BLEU	ROUGE1	ROUGE2	ROUGEL	CIDER	METEOR	CLIP-S	REF-CLIP-S
git_large	0.10267	0.21905	0.07333	0.21905	0.06695	0.10481	0.25932	0.77371
blip_large	0.13485	0.23059	0.11897	0.23059	1.11659	0.13404	0.29505	0.85464
blip2	0.02528	0.11097	0.01905	0.11097	0.02236	0.08333	0.27786	0.80424
git_base	0.10295	0.2014	0.08081	0.2014	0.08039	0.14689	0.25113	0.82038
blip_base	0.10305	0.19333	0.05	0.19333	0.57434	0.09505	0.25856	0.83212

Рисунок 3.9 – Сторінка зображення (екран 2) [Рисунок виконаний самостійно]

Також інтегрована модель CLIP, проте вона використовується для оцінки згенерованих описів.

Перейдемо безперечно до технічних рішень. Для роботи з різними моделями реалізовано 2 функції: load_model_and_processors та analyze_image. Завантаження моделей реалізовано за шаблоном "фабричного методу", який дозволяє динамічно завантажувати різні моделі та їх процесори в залежності від заданого імені моделі (див. рис. 3.10). Це забезпечує модульність та гнучкість системи, дозволяючи легко додавати нові моделі або змінювати існуючі без значних змін в основному коді.

По завершенню роботи, реалізовано очистку пам’яті, використану для моделей та процесорів, і викликає збірку сміття, що допомагає зменшити використання ресурсів і уникнути витоків пам'яті. Приклад коду:

```
            model = VisionEncoderDecoderModel.from_pretrained("nlpconnect/vit-gpt2-image-captioning", cache_dir=cache_dir)
            feature_extractor = ViTFeatureExtractor.from_pretrained("nlpconnect/vit-gpt2-image-captioning", cache_dir=cache_dir)
```

```
            tokenizer = AutoTokenizer.from_pretrained("nlpconnect/vit-
gpt2-image-captioning", cache_dir=cache_dir)
            return model, feature_extractor, tokenizer
```

Функція analyze_image відповідає за аналіз зображення, використовуючи вказану модель, і генерує опис для цього зображення. Вона інтегрує кілька ключових компонентів, таких як завантаження зображення, підготовка даних для моделі, генерація опису та вимірювання часу обробки.

Самі моделі ми зберігаємо локально на диску. Якщо модель ще не завантажена, вона завантажується і зберігається в кеші. Це забезпечує швидкий доступ до моделей при наступних запитах. Окрім того, ми вимірюємо час на обробку зображення, який потім зберігається в базі даних разом з згенерованим описом. Таке рішення дозволяє легко адаптувати систему до нових вимог і забезпечує високу продуктивність при роботі з різними моделями та великими обсягами даних.

Код використовує бібліотеки transformers, torch, nltk, PIL та інші для завантаження моделей, обробки зображень та генерації описів. Завантажені файли зберігаються на диску, назва файлу є його ідентифікатором в системі. В подальшому це може бути недоліком, адже створює певні ризики що будуть завантажені кілька файлів з однаковими іменами.

Щодо розрахунку метрик, що характеризують якість згенерованого тексту, то для кожного показника реалізовано власну функцію.

Наприклад, calculate_bleu_score(references, candidate) обчислює значення BLEU між кандидатським описом та референсними описами, використовуючи метод згладжування. Фрагмент коду:

```
    def calculate_bleu_score(references, candidate):
        reference_tokens_list = [ref.lower().split() for ref in
references]
        candidate_tokens = candidate.lower().split()
        smoothing_function = SmoothingFunction().method1
        score = sentence_bleu(reference_tokens_list, candidate_tokens,
smoothing_function=smoothing_function)
        return score

    def calculate_rouge_score(references, candidate):
```

```
        scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2', 'rougeL'],
use_stemmer=True)
        aggregated_scores = {'rouge1': 0, 'rouge2': 0, 'rougeL': 0}
        for ref in references:
            scores = scorer.score(ref, candidate)
            for key in aggregated_scores:
                aggregated_scores[key] += scores[key].fmeasure

        for key in aggregated_scores:
            aggregated_scores[key] /= len(references)
        return aggregated_scores
```

Що стосується показників, які засновані на моделі CLIP, то їх розрахунок трохи відрізняється від інших. Це пов’язано з тим, що CLIP є моделлю, яка була натренована на великій кількості пар зображення-текст і здатна порівнювати зображення з текстом у спільному векторному просторі. Приклад коду:

```
    def calculate_clip_s(file, captions, model, preprocess):
        filename = file.split('/')[-1]
        relative_path = os.path.join("static", "uploads", filename)
        image_path = os.path.join(CURRENT_DIR, relative_path)
        image = preprocess(Image.open(image_path)).unsqueeze(0).to("cpu")

        texts = [c['caption'] for c in captions]
        text = clip.tokenize(texts).to("cpu")

        with torch.no_grad():
            image_features = model.encode_image(image)
            text_features = model.encode_text(text)

        image_features /= image_features.norm(dim=-1, keepdim=True)
        text_features /= text_features.norm(dim=-1, keepdim=True)

        similarity = (image_features @ text_features.T).cpu().numpy()

        return similarity[0]
```

Зображення кодується у векторний простір за допомогою методу model.encode_image, а тексти – за допомогою model.encode_text. Векторні представлення нормалізуються (зводяться до одиничної довжини).

Схожість між векторними представленнями зображення і текстів обчислюється за допомогою скалярного добутку нормалізованих векторів. Це дозволяє отримати числове значення, яке відображає схожість між зображенням та кожним з описів.

Показники на основі CLIP дозволяють не лише оцінювати текстові описи зображень, але й робити це з урахуванням як візуальної, так і текстової інформації, що додає додатковий рівень розуміння до традиційних текстових метрик.

3.4 Результати експерименту

В експерименті приймали участь 8 моделей, кожна з яких генерувала описи на основі фото. Ці описи були оцінені як експертною групою, так і за допомогою різних метрик. Зображення були поділені на групи, тож ми можемо проаналізувати результати для кожної групи окремо, а також узагальнити висновки по всіх групах.

Окрім якості згенерованих описів, ми оцінювали також їх продуктивність. Всі генерації відбувались в однакових умовах за наявності ідентичних ресурсів. Тож ми оцінимо також швидкість роботи моделей.

Моделі на основі FLICKR8 були виключені з експерименту після початку, адже стало зрозуміло що вони мають недостатньо високу якість. Враховуючи розмір датасету, на якому вони тренувались (8000 зображень), очікувати чогось суттєвого важко. Для прикладу, доволі просте зображення собаки (див. рис. 3.10).

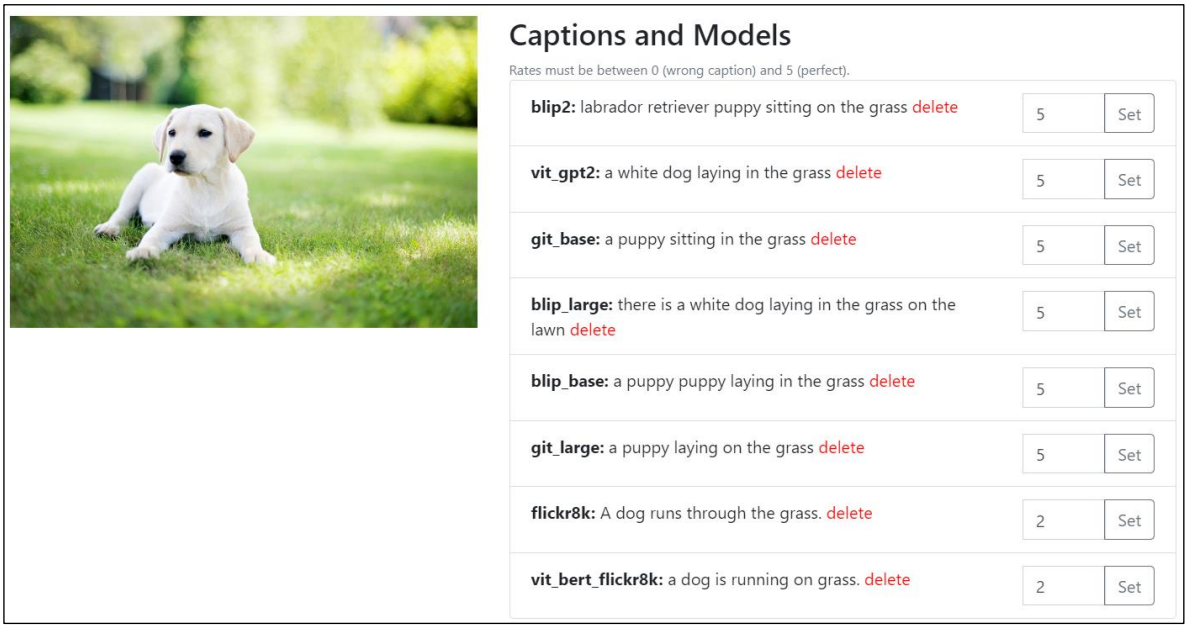


Рисунок 3.10 – Порівняння згенерованих підписів, приклад 1 [Рисунок виконаний самостійно]

Всі моделі отримали макисмальні показники, за винятком FLICKR8. Незважаючи на те, що всі визначили що на фото собака, некоректно зазначено що собака біжить. Розглянемо гірший випадок (рис. 3.11). На фото зображено тарілку з їжею, в той час як FLICKR8 бачить там чоловіка, який або гуляє з собакою або показує фокуси. Тому ці моделі були виключені з подальшої експертної оцінки.

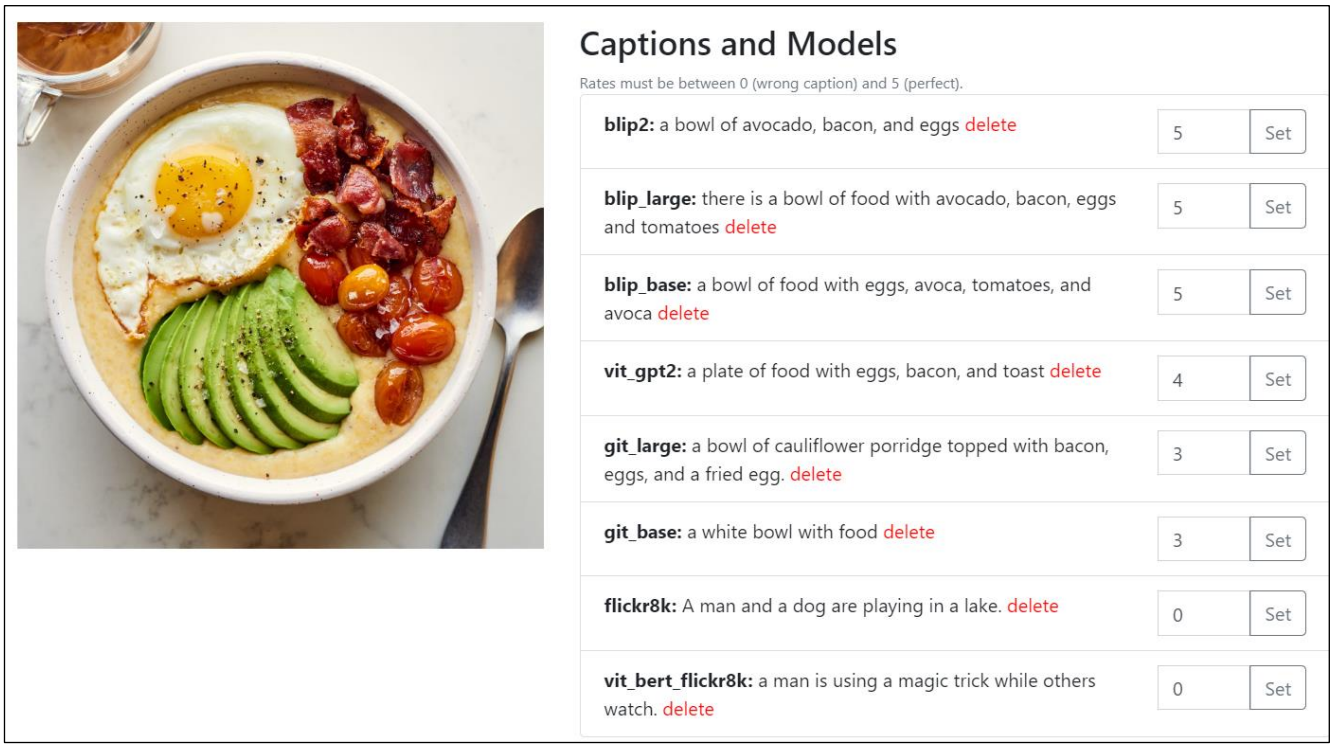


Рисунок 3.11 – Порівняння згенерованих підписів, приклад 2 [Рисунок виконаний самостійно]

Метрики для цього зображення показують, що моделі BLIP (особливо blip_large) демонструють найкращу продуктивність серед усіх розглянутих моделей за більшістю метрик, включаючи BLEU, ROUGE1, ROUGE2, ROUGEL, METEOR, і Ref-CLIP-S, що вказує на їхню високу здатність генерувати точні та змістовні описи для зображень. Модель blip_large досягла найвищих значень у BLEU (0.10025) та ROUGE1 (0.57956), що свідчить про високу точність і відповідність згенерованих описів референсним.

На противагу цьому, моделі Flickr8k та ViT-BERT-Flickr8k показали значно гірші результати: наприклад, BLEU для flickr8k (0.04469) та vit_bert_flickr8k (0.02271) є в 2.2 та 4.5 разів нижчими відповідно порівняно з blip_large.

Аналогічно, показник ROUGE1 для flickr8k (0.19896) та vit_bert_flickr8k (0.11617) є в 2.91 та 4.99 разів нижчим відповідно. Це вказує на суттєві відставання у якості генерованих описів для цих моделей, що потребує їх подальшого вдосконалення для досягнення кращої якості результатів.

Для спрощення порівняння результатів введемо інтегральний показник, який включатиме в себе всі середні значення метрик. Для BLEU, ROUGE1, ROUGE2, ROUGEL, CIDEr, METEOR, CLIP-S, REF-CLIP-S: більші значення означають кращу якість. Тому їх слід максимізувати. Розрахунок такого показника буде відбуватись на рівні бази даних наступним запитом:

```
SELECT
    model_name,
    ROUND((AVG(bleu) + AVG(rouge1) + AVG(rouge2) + AVG(rougeL) +
AVG(cider) + AVG(meteor) + AVG(clip_s) + AVG(ref_clip_s)) / 8, 5) AS
overall_score
FROM
    history
GROUP BY
    model_name
ORDER BY
    overall_score DESC;
```

Порівняємо значення інтегрованого показника з середнім значенням оцінок, які були поставлені експертами. Результати зображено на рис. 3.12:

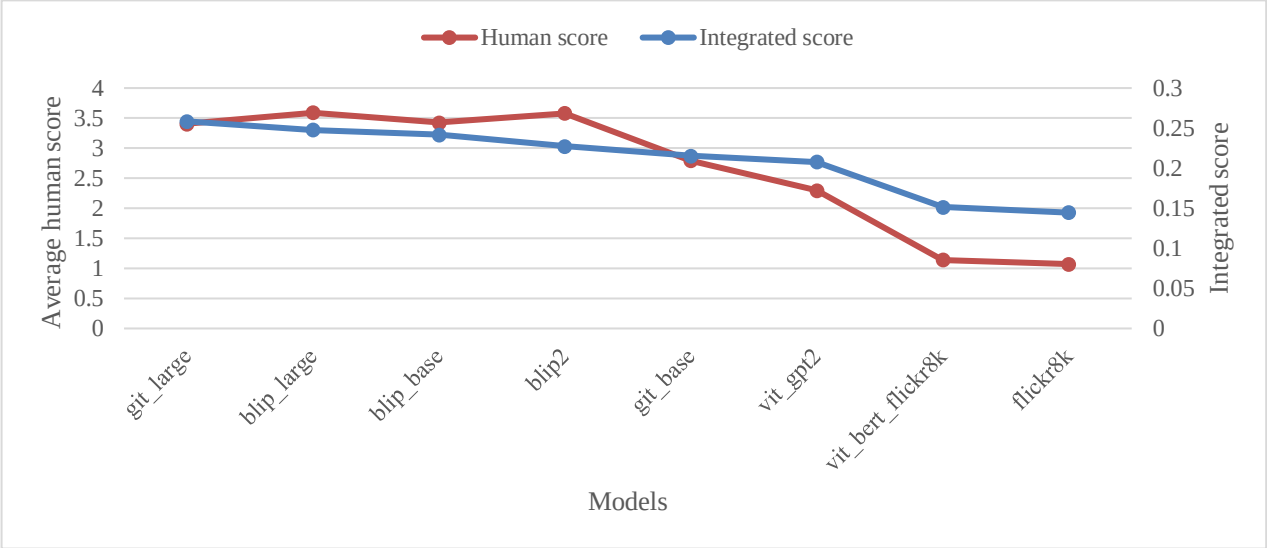


Рисунок 3.12 – Середнє значення показників за моделями [Рисунок виконаний самостійно]

Щодо даних на графіку, то беззаперечними Лідери за інтегральним показником є git_large та blip_large. У першої моделі найвищий інтегральний показник (0.25847) при високому середньому оцінюванні людиною (3.40441). Це свідчить про те, що модель демонструє хороші результати як за автоматичними метриками, так і за оцінками експертів.

BLIP large має другий найвищий інтегральний показник (0.24793) і найвища середня оцінка людиною (3.58824), що робить її найкращою моделлю за оцінками експертів.

Розбіжності між інтегральними показниками та оцінками людиною присутні. Так, git_base та vit_gpt2 мають нижчі інтегральні показники (0.21551 і 0.20778), проте середні оцінки людиною (2.79412 і 2.29412 відповідно) демонструють значну різницю у сприйнятті якості моделями та експертами.

Моделі vit_bert_flickr8k та flickr8k потребують значного покращення для досягнення конкурентоспроможного рівня продуктивності.

Порівняємо аналогічні показники за кожною з груп (див. табл. 3.1).

Таблиця 3.1 – Середне значення оцінок за моделями (таблиця виконана самостійно)

Група	Інтегральний показник	Середня оцінка екпертів
Group 1 - Simple, large objects	0.22884	3.54673
Group 2 - Dark, Night	0.20975	3.15686
Group 3 - Many similar objects	0.22285	3.42667
Group 4 - Famous people	0.20604	2.06383
Group 5 - Views	0.21231	3.25694
Group 6 - Text elements	0.16054	2.05263

Порівняємо результати кожної з груп. Так, у першій групі найвище інтегральне значення і середня оцінку експертів, що вказує на те, що моделі демонструють найкращу якість генерації описів для простих, великих об'єктів. Це очікуваний результат, адже ця група є найлегшою. Проте, й тут були заплутані випадки (див. рис. 3.13).

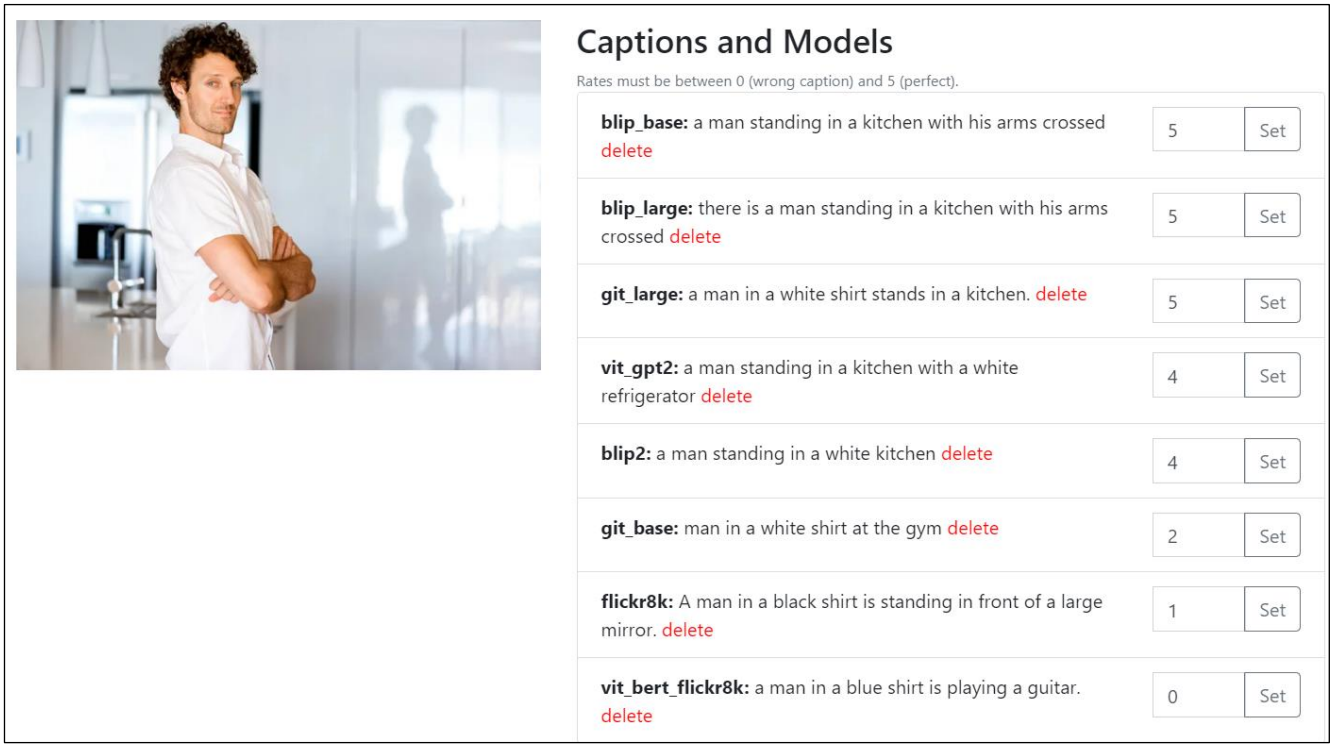


Рисунок 3.13 – Група 1, приклад розпізнавання [Рисунок виконаний самостійно]

Незважаючи на те, що на фото стоїть один чоловік, його відображення на білій поверхні не завадило коректно сформувати опис зображення. Помилки були як з кольором футболки, так і з тим, що робить чоловік. blip_base, blip_large та git_largeчудово впорались з таким фото.

Група темних та нічних зображень має друге за величиною інтегральне значення, але середня оцінка людиною трохи нижча, що може вказувати на деякі труднощі моделей у генерації якісних описів для цього типу зображень.

Третя група показує високу якість за обома показниками, що вказує на те, що моделі здатні ефективно справлятися з генерацією описів для зображень, які містять багато схожих об'єктів. Для прикладу, подивимось на фото 3.14.

Як бачимо, найкраще впоралась модель blip_large, яка не намагалась перелічити прапори на фото, а максимально детально описала зображення. Моделі blip_large і blip2 демонструють найвищу продуктивність за показниками повноти (ROUGE) та лексичної відповідності (BLEU, METEOR), що робить їх найкращими серед розглянутих моделей. Вони ефективно включають важливі текстові елементи та використовують відповідну лексику для генерації описів.

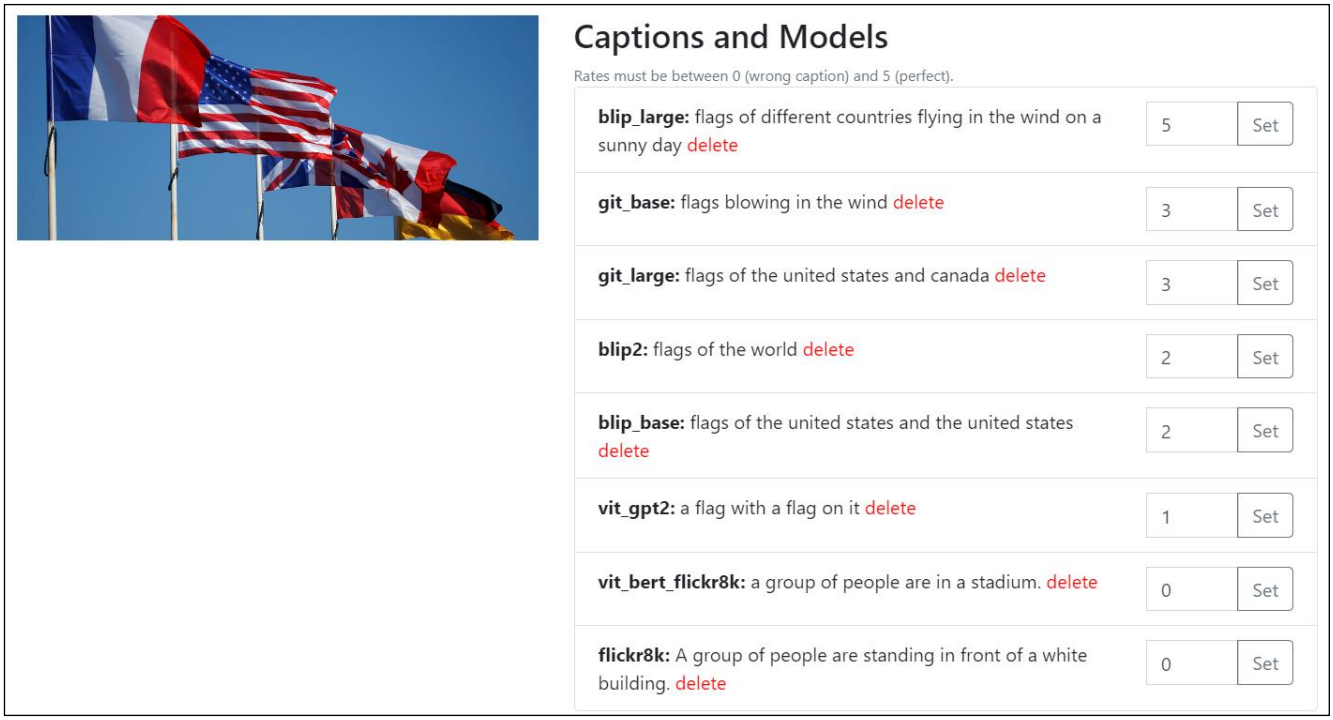


Рисунок 3.14 – Група 3, приклад розпізнавання [Рисунок виконаний самостійно]

Інші моделі, такі як git_large і blip_base, також показують пристойні результати, але трохи відстають від лідерів. Моделі vit_gpt2, vit_bert_flickr8k, та flickr8k потребують значного вдосконалення для покращення якості своїх генерацій, особливо в аспектах повноти та лексичної відповідності.

Незважаючи на середнє інтегральне значення, група з відомими людьми (4) має низьку середню оцінку людиною, що свідчить про труднощі у розпізнаванні та описі зображень відомих людей. Так, моделі не мали проблем з позами, статтю чи одягом людини. Недоліки були у тому, що часто замість однієї особи моделі вказували зовсім інших. Так, Зеленського назвали президентом росії, замість Олександра Усика на фото модель побачила іншого боксера. Щодо інших осіб, то краща ситуація була з президентами США, Ейнштейном і Бонапартом.

Зображення з краєвидами отримали високі оцінки як за інтегральним значенням, так і за середньою оцінкою людиною, що свідчить про хорошу здатність моделей генерувати якісні описи для цього типу зображень.

Шоста група має найнижчі інтегральне значення та низьку оцінку людиною, що вказує на те, що моделі зазнають значних труднощів у генерації якісних описів для зображень з текстовими елементами. Це цілком очікувано, адже в першу

чергу вони намагаються описати зображення. Зауважимо, що були випадки успішного розпізнавання окремими моделями.

Результати показують, що моделі генерації текстових описів найкраще працюють для простих і великих об'єктів, а також для зображень з багатьма схожими об'єктами та краєвидами. Однак, є значні проблеми з генерацією описів для зображень, що містять відомих людей (група 4) та текстові елементи (група 6). Ці групи потребують додаткового вдосконалення моделей для досягнення кращої якості генерації описів.

Одним з основних недоліків розглянутих моделей генерації описів є їхня нездатність коректно обробляти зображення зі складними сценами або великою кількістю об'єктів.

Часто такі моделі не в змозі врахувати всі важливі деталі або правильно зрозуміти контекст сцени, що призводить до неповних або неточних описів. Наприклад, для деяких фото опис взагалі не генерувався, і результат був або пустим, або виглядав як "--------". Крім того, моделі можуть невірно рахувати кількість об'єктів, коли намагаються вказати точну кількість.

Однією з виявлених проблем є те, що на темних фото моделі не завжди коректно визначають колір одягу, що може призводити до помилкових описів. Це пов'язано з труднощами у розпізнаванні кольорів в умовах недостатнього освітлення.

Крім того, моделі інколи намагаються вказувати конкретне місце або країну, але часто помиляються, що може вводити користувачів в оману та знижувати довіру до автоматично згенерованих описів. Ці проблеми вказують на необхідність вдосконалення моделей для більш точної інтерпретації зображень у різних умовах.

Іншою важливою проблемою є час обробки зображень. Деякі моделі вимагають значних обчислювальних ресурсів та часу для генерації описів, що робить їх непридатними для реальних застосувань з обмеженими ресурсами або вимогами до швидкодії. Крім того, моделі не завжди коректно вгадують колір одягу на темних фото і інколи помиляються, вказуючи конкретне місце або

країну. Недостатня узгодженість між результатами автоматичних метрик та оцінками експертів також виявляє потребу в подальшому вдосконаленні моделей для забезпечення більш точних і релевантних описів.

Порівняємо час на генерацію опису до одного зображення. На відміну від попередніх критеріїв, час має бути мінімальний, адже чим швидче модель генерує якісний опис – тим краще. Результати по швидкості можна побачити на зображенні:

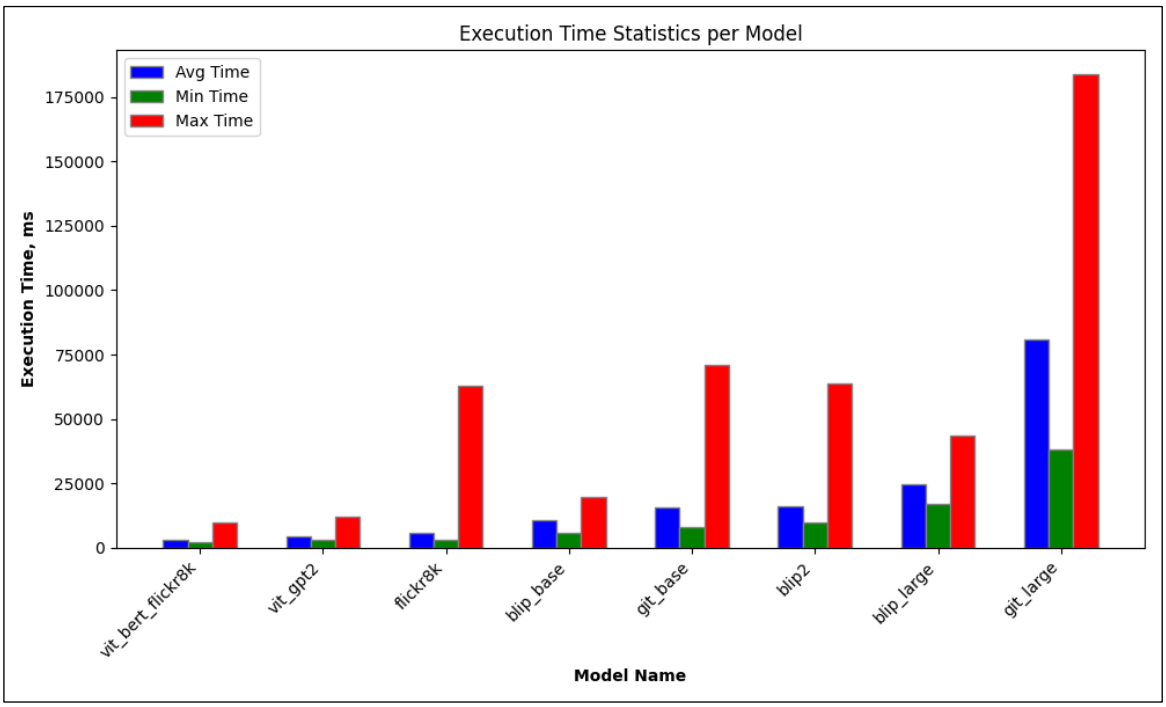


Рисунок 3.15 – Швидкість опису зображення [Рисунок виконаний самостійно]

Час обробки зображень суттєво різниться між моделями. Найшвидшою є vit_gpt2 (середній час 4492.52 мс), тоді як blip_base (10904.27 мс) і git_base (15588.38 мс) показують помірний час обробки. blip2 (16191.11 мс) та blip_large (24512.03 мс) мають вищі середні показники, а git_large є найповільнішою з середнім часом 80945.65 мс. Це свідчить про значну ресурсомісткість останньої моделі та потребу враховувати обчислювальні ресурси при виборі моделі.

Проранжуємо результати за трьома показниками – інтегральним показником, середньою оцінкою експерта та середньою швидкістю обробки одного зображення та визначимо загальну оцінку (див. табл. 3.2).

Згідно з отриманими результатами, моделі blip_base та blip_large набрали однакову кількість балів (10), що робить їх найкращими моделями в контексті збалансованої продуктивності та якості описів.

Таблиця 3.2 – Загальні результати по експерименту (таблиця виконана самостійно)

Модель	Середній час для опису зображення	Середня інтегральна оцінка	Середня оцінка експертів	Загальний бал
blip_base	4	3	3	10
blip_large	7	2	1	10
blip2	6	4	2	12
vit_gpt2	2	6	5	13
git_large	8	1	4	13
vit_bert_flickr8k	1	7	7	15
git_base	5	5	5	15
flickr8k	3	8	8	19

Хоча blip_large та git_large мають схожу якість згенерованих описів, перша модель суттєво швидша, що робить її більш ефективною у використанні. Крім того, blip_large отримала високу оцінку від експертів, що підкреслює її здатність генерувати точні та релевантні описи. Оптимальна якість опису з ефективністю роботи може бути досягнута за допомогою моделей, таких як blip_base та blip_large, які демонструють хороший баланс між швидкістю та якістю генерації описів. Це свідчить про те, що ці моделі є найбільш підходящими для практичного застосування, де важливі як точність, так і швидкодія.

Для подальших досліджень пропонується розширити навчальні датасети, зокрема включити більше зображень у складних умовах освітлення та з різними сценами. Крім того, використання ансамблів моделей, удосконалення алгоритмів обробки природної мови можуть значно покращити якість генерації описів. Інтерактивні підходи, що залучають людину в процес генерації, також можуть бути корисними для виявлення та корекції помилок у режимі реального часу.

ВИСНОВКИ

В ході виконання кваліфікаційної роботи було проведено аналіз проблемної області дослідження, описано методи та інструменти, які були використані для вирішення поставлених завдань.

Атоматичний опис зображень це комплексна задача. За сотанні 10 років є значний прогрес, проте навіть найсучасніші моделі все ще не можуть 100% описати деякі види зображень.

Було визначено основні проблеми, з якими стикаються сучасні моделі генерації описів зображень, такі як невірна інтерпретація кольорів на темних фото, помилки при вказівці кількості об'єктів, та некоректне визначення конкретних місць або країн. Було виявлено, що деякі моделі взагалі не генерують описів для певних зображень, або генерують некоректні результати.

Під час аналізу були використані різні метрики для оцінки якості згенерованих описів, зокрема BLEU, ROUGE, METEOR, CLIP-S, та Ref-CLIP-S. Це дозволило комплексно оцінити ефективність моделей, враховуючи як автоматичні показники, так і оцінки експертів.

Особлива увага була приділена інтегральному показнику, який враховує усі ці метрики, що дозволило порівняти моделі між собою за сукупною оцінкою. Окрім того, встановлено відмінності між оцінками експертів та автоматичними показниками. Дані проаналізовано як в цілому, так і по окремим групам зображень.

В ході дослідження було також оцінено швидкість обробки зображень кожною моделлю. Виявилося, що моделі з кращою якістю генерації описів, такі як blip_base та blip_large, також демонструють високу продуктивність за середнім часом обробки зображень. Проте, деякі моделі, хоча і мали хороші оцінки за якістю описів, значно поступалися в швидкості обробки, що обмежує їх використання у реальних умовах.

На основі отриманих результатів були запропоновані рекомендації щодо вдосконалення існуючих моделей генерації описів зображень. Зокрема, було запропоновано розширити та балансувати навчальні датасети, використовувати

методи аугментації даних, вдосконалювати алгоритми обробки природної мови та застосовувати інтерактивні підходи, що залучають людину в процес генерації. Це дозволить не лише покращити якість генерованих описів, але й підвищити швидкість обробки та зменшити кількість помилок у кінцевих результатах.

В подальшому, дослідження можна розширити використовуючі інші групи зображень та інакше натреновані моделі. Окрім того, досить цікавою була б оцінка запропонованих моделей порівняно з хмарними моделями, які використовує Google. Amazon, Microsoft та інші.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Show and Tell: A Neural Image Caption Generator / O. Vinyals, A. Toshev, S. Bengio, D. Erhan // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 2015. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html (дата звернення: 03.04.2024).

2. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge / Vinyals O., Toshev A., Bengio S., Erhan D. // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016. Т. 39, № 4. С. 652-663. URL: https://arxiv.org/abs/1609.06647 (дата звернення: 11.03.2024).

3. An Image Captioning Algorithm Based on the Hybrid Deep Learning Technique (CNN+GRU) / Ahmad R. A., Azhar M., Sattar H. // arXiv:2301.02440. 2023. URL: https://arxiv.org/abs/2301.02440 (дата звернення: 12.03.2024).

4. Neural machine translation with reconstruction / Tu Z., Liu Y., Shang L., Liu X., Li H. // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Лютий 2017. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10950 (дата звернення: 12.03.2024).

5. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers // arXiv preprint arXiv:2106.10270. 2021. URL: https://arxiv.org/abs/2106.10270 (дата звернення: 03.05.2024).

6. Google Model Garden // Vertex AI. URL: https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/imagetext (дата звернення: 03.05.2024).

7. Microsoft Learn. Describe Images in Computer Vision // Azure AI Services. URL: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/concept-describe-images-40?tabs=image (дата звернення: 03.05.2024).

8. A PyTorch Tutorial to Image Captioning // GitHub. URL: https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning (дата звернення: 03.06.2024).

9. Models - Hugging Face. Hugging Face – The AI community building the future. URL: https://huggingface.co/models?other=image-captioning&sort=downloads (дата звернення: 03.04.2024).

10. An Overview of Image Caption Generation Methods / Wang H., Zhang Y., Yu X. // Computational Intelligence and Neuroscience. 2020. URL: https://www.hindawi.com/journals/cin/2020/3062706/ (дата звернення: 17.03.2024).

11. Attention Is All You Need / Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. // arXiv preprint arXiv:1706.03762. 2023. URL: https://arxiv.org/abs/1706.03762 (дата звернення: 03.06.2024).

12. Learning Transferable Visual Models From Natural Language Supervision / A. Radford et al. 2021. URL: https://arxiv.org/abs/2103.00020 (date of access: 14.03.2024).

13. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation / J. Liat al. 2022. URL: https://arxiv.org/abs/2201.12086 (дата звернення: 10.04.2024).

14. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models / J. Li et al. 2023. URL: https://arxiv.org/abs/2301.12597 (дата звернення: 12.04.2024).

15. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning / Dai W., Li J., Li D., Tiong A. M. H., Zhao J., Wang W., Li B., Fung P., Hoi S. // arXiv preprint arXiv:2305.06500. 2023. URL: https://arxiv.org/abs/2305.06500 (дата звернення: 03.06.2024).

16. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale / Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. // arXiv preprint arXiv:2010.11929. 2021. URL: https://arxiv.org/abs/2010.11929 (дата звернення: 03.06.2024).

17. Neural Attention for Image Captioning: Review of Outstanding Methods / Zohourianshahzadi Z., Kalita J. K. // arXiv preprint arXiv:2111.15015. 2021. URL: https://arxiv.org/abs/2111.15015 (дата звернення: 11.03.2024).

18. Efficiency of image convolution / Smelyakov K., Shupyliuk M., Martovytskyi V., Tovchyrechko D., Ponomarenko O. // 2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL). 2019. С. 578-583. DOI: 10.1109/CAOL46282.2019.9019450.

19. Gradational Correction Models Efficiency Analysis of Low-Light Digital Image / Smelyakov K., Chupryna A., Hvozdiev M., Sandrkin D. // 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream). 2019. С. 1-6. DOI: 10.1109/eStream.2019.8732174.

20. End-to-End Transformer Based Model for Image Captioning / Wang Y., Xu J., Sun Y. // arXiv preprint arXiv:2203.15350. 2022. URL: https://arxiv.org/abs/2203.15350 (дата звернення: 12.03.2024).