

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КІЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
Факультет прикладної математики
Кафедра прикладної математики

«До захисту допущено»

Завідувач кафедри ПМА

Чертов Олег Романович

«__» _____ 2024

Дипломна робота
на здобуття ступеня бакалавра
за освітньо-професійною програмою
«Наука про дані та математичне моделювання»
спеціальності 113 Прикладна математика
на тему: «Математичне та програмне забезпечення для автоматичного тегування
зображень»

Виконав:
студент IV курсу, групи КМ-01
Скороденко Д. О.

Керівник:
доцент кафедри ПМА
Сирота С. В.

Консультант з нормоконтролю:
доцент кафедри ПМА
Мальчиков В. В.

Засвідчую, що в цьому звіті немає запозичень
із праць інших авторів без відповідних посилань.

Студент _____

Київ --- 2024

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КІЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
Факультет прикладної математики
Кафедра прикладної математики

Рівень вищої освіти - перший (бакалаврський)

Спеціальність - 113 «Прикладна математика»

Освітньо-професійна програма «Наука про дані та математичне моделювання»

«Затверджую»
Завідувач кафедри ПМА
Чертов Олег Романович
«___» _____ 2024

ЗАВДАННЯ
на дипломну роботу
Скороденку Дмитру Олександровичу

1. Тема роботи «Математичне та програмне забезпечення для автоматичного тегування зображень», науковий керівник роботи Сирота Сергій Вікторович, доцент кафедри ПМА, затверджені наказом по університету від «...» 2024 р. №
2. Термін подання студентом роботи «...» червня 2024 р.
3. Вихідні дані до роботи: розроблювана система для маркування зображень повинна забезпечити якість опису зображення краще або на рівні із існуючими рішеннями згідно із тестовими метриками.
4. Зміст роботи: виконати огляд існуючих рішень задачі маркування зображення, провести моделювання системи на основі аналізу існуючих рішень, імплементація системи, тренування / тестування системи, аналіз ефективності компонентів системи, аналіз ефективності системи у порівнянні із існюючими рішеннями.

5. Перелік обов'язкового ілюстративного матеріалу: діаграма архітектури композитної системи, розподіл лейблів у тренувальному/тестовому датасеті, числові характеристики датасету, графіки процесу тренування для кожної із компонент системи, таблиця для порівняння ефективності компонентів системи, таблиця для порівняння ефективносіт маркування у порівнянні з існуючими рішеннями, ілюстративні приклади роботи системи.

6. Дата видачі завдання «...» ... 2024 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1.	Вивчення та збір літератури за темою "маркування зображень"	25.01.2024	+
2.	Проведення аналізу особливостей існуючих систем маркування зображень	15.02.2024	+
3.	Вибір та підготовка датасету	25.02.2024	+
4.	Проектування системи маркування зображення	10.03.2024	+
5.	Проведення тестового тренування системи та відладка програми	25.03.2024	+
6.	Проведення валідації результатів тестового тренування	28.03.2024	+

7.	Проведення тренування фінальної версії системи	14.04.2024	+
8.	Проведення порівняльного аналізу компонентів системи	15.04.2024	+
9.	Проведення порівняння ефективності з існуючими рішеннями	15.04.2024	+
10.	Підготовка першої версії дипломної роботи	08.05.2024	+
11.	Оформлення пояснювальної записки	20.05.2024	+

Студент _____

Дмитро СКОРОДЕНКО

Керівник _____

Сергій СИРОТА

АНОТАЦІЯ

Дипломну роботу виконано на 30 аркушах, вона містить 3 додатки та перелік посилань на використані джерела з 22 найменувань. У роботі наведено 8 рисунків та 3 таблиці.

Актуальність теми: В сучасному світі технології розвиваються надзвичайно швидко. Швидкість цього розвитку можна виміряти об'ємами даних, яким оперують люди. Так для 2000-х років було цілком достатньо мати дискети із максимальним вмістом до 10-15 МБ. Станом на 2024 рік, існують різні накопичувачі від 16 ГБ до декількох десятків ТБ. Чому це важливо? Для того щоб переносити велику кількість даних потрібно, щоб генерувалось ще більше даних. І це дійсно так, якщо наприклад розглянути сучасні хостинги зображень, бази даних медзакладів, бази даних супутникових знімків і тд. то всюди ми побачимо уже від сотень тисяч до сотень мільйонів зображень, при чому швидкість появи нових зображень стрімко зростає. Основними причинами такого росту є: збільшення кількості людей та стрімка цифровізація більшості аспектів людської життєдіяльності. Такий вибуховий ріст у швидкості появи нових зображень створює проблему структуризації зображень. Для вирішення цієї проблеми можна присвоїти кожному зображеню лейблі, які загально описують його вміст. Без таких лейблів будь яка структура, яка містить у собі велику кількість зображень перетвориться у звалище. Саме тому важливо мати швидкий та якісний метод для автоматичного маркування зображень.

Мета дослідження: Метою даної роботи є розробка ПЗ для маркування зображень (шпалерів робочого столу) для покращення системи категоріального пошуку зображень (шпалерів робочого столу).

Завдання дослідження: Створення системи, яка виконуватиме маркування шпалерів робочого столу на основі двох модальностей даних: зображення та шумних тегів.

Для досягнення цієї мети було виконано наступні завдання:

- Проведено аналіз існуючих рішень
- Змодельовано систему
- Проведено тренування системи
- Проведено аналіз ефективності компонентів системи
- Проведено порівняльний аналіз якості і повноти опису розглянутої моделі відносно існуючих рішень на основі тестових метрик
- Проведено аналіз ілюстративних прикладів роботи системи

Об'єкт дослідження: Об'єктом дослідження є маркування зображень, та методи покращення маркування зображення. Для порівняння ефективності маркування серед існуючих рішень було обрано моделі, для яких обраховані тестові метрики для того ж датасету, який обрано в даній роботі. До множини існуючих рішень належать: SR-CNN-RNN [11], Resnet-SRN [23], MS-CMA [22], Query2Label [12], Resnet-CPSD [21] та ін.

Предмет дослідження: Предметом дослідження є множина шпалерів робочого столу в якості основної модальності даних, та додаткова інформація (надані людьми шумні теги) в якості додаткової.

Методи дослідження:

- * Теорія системного аналізу
- * Проектування систем Data Science / Deep learning
- * Проектування інформаційних систем
- * Обробка та аналіз зображень та тегів на основі методів глибинного навчання
- * Теорія алгоритмів
- * Аналіз даних та математична статистика

Кінцевий результат: Кінцевим результатом даної роботи є математичне та програмне забезпечення, архітектура моделі, вагові коефіцієнти натренованої моделі та код програмного забезпечення, в якому реалізовано дану роботу.

Ключові слова: Маркування зображень, глибинне навчання, нейронні мережі, згорткові нейронні мережі, багатошарові персепtronи, композиція нейронних мереж.

ABSTRACT

The thesis presented in 30 pages. It contains 3 appendixes and bibliography of 22 references, 8 figures and 3 tables are given in the thesis.

Topic relevance. In modern world technologies are progressing with very high speed. The speed of this progress can be measured by volume of stored data which people operate with. In this regard for the 2000s, it was quite enough to have diskettes with the maximum volume of up to 10-15 MB. As of 2024, there are various drives from 16GB to several tens of TB. Why is this important? In order to transfer large amounts of data, even more data should be generated. And it really is, let's for example consider modern image hostings, databases of medical institutions, databases of satellite images, etc. then everywhere we will see from hundreds of thousands to hundreds of millions of images, and the rate of appearance of new images is rapidly increasing. The main reasons for this growth are: an increase in the number of people globally and the rapid digitalization of every aspect of our everyday lives. Such explosive growth in speed the appearance of new images creates a problem of image structuring. To solve this problem, you can assign each image with labels that generally describe its content. Without such labels, any structure that contains a large number of images will be like a waste disposal site, where you can't find anything. That's why it's important to have a fast and quality method for automatic image labeling.

Research goal: The purpose of this work is to create deep learning model for automatic image labelling, specifically wallpapers, to improve categorical search using output labels.

Research objectives: Creating system, which will label images (wallpapers), given data of two modalities: image and associated noisy tags.

To reach said objective the following tasks were completed:

- Conducted analysis of existing solutions
- Modeled image labeling system
- Conducted model training

- Conducted analysis of performance gain from each component of composite system
- Conducted comparative analysis of model's labeling effectiveness with existing solutions
- Conducted analysis of illustrative examples of system's result

Research object: The object of research is image labeling and methods to improve it. To conduct comparative analysis to existing solution, only models which are trained/tested on specific dataset were considered as 'existing solutions'. To 'existing solutions' belong following models: SR-CNN-RNN [11], Resnet-SRN [23], MS-CMA [22], Query2Label [12], Resnet-CPSD [21] etc.

Research subject: The subject of this work is subset of images which are called 'wallpapers'. These are the images that you see as background image when you use your computer. The main modality of data is image. Human provided tags would be additional modality.

Research methods:

- * System analysis
- * Data Science / Deep learning
- * Projecting of informational systems
- * Processing and analysis of images and tags using methods of deep learning
- * Theory of algorithms
- * Data analysis and mathematical statistics

End result of research: The end result of research is mathematical and software implementation, architecture of model, weights of trained models and source code of software which implements this work.

Keywords: Image labeling, deep learning, neural networks, convolutional neural networks, multilayer perceptrons, composition of neural networks.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

- * Модель - нейронна мережа
- * Система - композиція нейронних мереж
- * Задача класифікації - це задача, яка вирішує проблему приналежності чогось виключно до одного класу із довільного набору класів.
- * Задача маркування - це задача, яка вирішує проблему приналежності чогось до декількох класів із довільного набору класів.
 - * Тегування та маркування - взаємозамінні поняття
 - * DNN - Глибинна нейронна мережа (Deep Neural Network)
 - * CNN - Згорткова нейронна мережа (Convolutional Neural Network)
 - * RNN - Рекурсивна нейронна мережа (Recursive Neural Network)
 - * ViT - Візуальні трансформери (Visual Transformers)
 - * Тег (Tag) - шумна інформація надана користувачем у формі тексту (наприклад для зображення кота "Cat, Canada, Cola")
 - * Лейбл (Label) - синонім слова ground truth для класифікації

ЗМІСТ

Перелік умовних позначень, скорочень і термінів	10
1 Вступ.....	13
2 Огляд існуючих рішень задачі маркування зображення	15
2.1 Базове рішення	15
2.2 Додаткова модальності даних	18
2.3 Кількість лейблів	20
3 Моделювання	23
3.1 Модель VCNN	23
3.2 Модель MLP	24
3.3 Модель LP	25
3.4 Модель LQP	26
3.5 Процес тренування.....	27
3.5.1 Цільові функції	27
3.5.2 Тренування VCNN	28
3.5.3 Тренування MLP	29
3.5.4 Тренування LP	29
3.5.5 Тренування LQP	29
3.6 Процес тестування	30
3.6.1 Тестування класифікаційних моделей (VCNN, MLP, LP)	30
3.6.2 Тестування композитної моделі.....	30
4 Експерименти	31
4.1 Датасет	31
4.1.1 Множина цільових класів	33
4.1.2 Числові характеристики датасету	34
4.2 Тестові метрики	35
4.3 Тренування.....	37
5 Аналіз результатів	39
5.1 Аналіз компонентів системи	39

	12
5.2 Аналіз впливу додаткової інформації	40
5.3 Порівняння з існуючими рішеннями.....	41
5.4 Демонстративні приклади.....	43
Висновки	45
Перелік посилань.....	46
Додаток А Код лістинг	48
Додаток Б Додаткові приклади.....	49
Додаток В Ілюстративний матеріал	50

1 ВСТУП

Задача класифікації - це одна із основних задач в аналізі зображень, вона полягає у присвоєнні кожному зображенню один із класів. Таким чином дане формулювання накладає обмеження - зображення містить тільки один об'єкт. Поява DNN [3] та її подальший розвитком у CNN [15, 14] разом із створенням великих датасетів як-от ImageNet [4] дало змогу вирішувати задачу класифікації зображень значно швидше і якісніше ніж люди.

Зрозуміло, що зображення - це той тип даних, який у абсолютної більшості випадків містить більше одного об'єкта, і відповідно більше одного класу для класифікації. Для поглиблення опису існує задача маркування зображень (image labeling). На відміну від класифікації, вона полягає у маркуванні зображення більше ніж одним класом. Таким чином повнота опису зображення кратно зростає у порівнянні із звичайною класифікацією, однак привносить декілька складних завдань.

1) Наявність декількох класів у одного зображення створює можливість описувати значно ширший спектр візуальної інформації: різні об'єкти, стилі, дії, і тд. Це створює потребу у розгляді додаткових джерел інформації, так як одного лише зображення вже недостатньо. Поява великих хостингів зображень таких як Imgur, Flickr, та ін., де користувачі можуть як завантажувати різноманітні зображення, так і додавати до них описову інформацію у вигляді тегів / анотацій, дала змогу створити досить різноманітні датасети: ImageNet [4], MS-COCO [10], NUS-WIDE [2], та ін. Також існують і інші види датасетів, наприклад: рентгенівські знімки та додаткова інформація (інші аналізи пацієнта, історія хвороб ...), супутникові знімки та додаткова інформація у вигляді метаданих, геолокацій тощо. Таким чином задача якісного маркування зображення вже охоплює значно більший спектр даних ніж просто зображення.

2) Маркування зображень передбачає динамічну к-сть промаркованих класів, так для опису зображенням із широким спектром понять необхідно 5-6 класів, для зображення із простим вмістом - 2-3 класи.

3) Маркування зображень потребує оцінки якості проведеного маркування. Оскільки будь який датасет буде містити в собі дизбаланс класів в тій чи іншій мірі, важливо оцінювати маркування із урахуванням цього.

4) Маркування зображень значно складніша задача ніж класифікація і відповідно зростає складність моделей. З однієї сторони складніша модель потенційно здатна покращити якість маркування, з іншої - може сильно збільшити як час на виконання маркування, так і час затрачений на тренування системи. До складних систем належать ті, які використовують трансформери та/або мають велику к-сть параметрів. Отже, важливо обрати певний баланс відносно складності моделі.

Все це робить задачу маркування зображення досить складною.

2 ОГЛЯД ІСНУЮЧИХ РІШЕНЬ ЗАДАЧІ МАРКУВАННЯ ЗОБРАЖЕННЯ

Розглянемо ключові аспекти задачі маркування зображення.

2.1 Базове рішення

Базовим рішенням в задачі маркування зображення є аналіз основної модальності даних - зображення. У абсолютній більшості існуючих робіт використовується CNN (Convolutional Neural Network). Застосовуються різні архітектури даної моделі ResNet [7], AlexNet [1], GoogleNet [17], ResNext [20].

В якості базового рішення також можна використовувати ViT [5]. Дано модель використовує трансформери, і аналізує зображення по частинам (patches).

Спільним між CNN та ViT є те, зазвичай їх рідко тренують з нуля, так як для цього необхідно багато ресурсів. Саме тому використовують вже натреновані (pretrained) моделі на великому датасеті, здебільшого ImageNet [4]. Для адаптації моделі до обраного контексту така модель дотреновується (fine tune), замінюючи існуючий класифікатор. Це працює завдяки тому, що всі архітектури сучасних CNN та ViT моделей містять десятки мільйонів параметрів, даючи широку репрезентацію зображень. Для CNN - це ієрархічне представлення: перші шари репрезентують базові особливості, а останні - більш специфічні особливості. Для ViT - це представлення, яке будується на основі взаємозв'язків між різними частинками зображення за допомогою трансформерів. Такі репрезентації зображень дозволяють адаптувати модель під різні задачі після проведення підгонки (finetune).

CNN (Convolutional Neural Networks)

a) Архітектура:

- Використовують згортки (convolutions) для обробки зображень.
- Складені з шарів згорток, активаційних функцій (наприклад, ReLU), пулінгу (max pooling або average pooling) та повнозв'язних шарів.

б) Принцип роботи:

- Згортки виділяють просторові особливості зображення (наприклад, краї, текстури).
- Кожен згортковий шар витягує ознаки вищого рівня по мірі проходження шарів (наприклад, краї → форми → об'єкти).

в) Особливості:

- Висока ефективність у задачах розпізнавання зображень.
- Використовують локальну інформацію через згортки з невеликими ядрами.
- Сильна локальна інваріантність, тобто можуть розпізнавати об'єкти незалежно від їхнього місця розташування в зображенні.

г) Приклади моделей:

- LeNet, AlexNet, VGG, ResNet, Inception.

ViT (Vision Transformers)

a) Архітектура:

- Базуються на архітектурі трансформерів, які були спочатку розроблені для обробки послідовностей у задачах обробки природної мови (наприклад, BERT, GPT).
- Складаються з шарів самоспрямованої уваги (self-attention) та повнозв'язних шарів.

б) Принцип роботи:

- Розбивають зображення на невеликі патчі (наприклад, 16x16 пікселів).
- Перетворюють кожен патч у вектор ознак за допомогою лінійної проекції.
- Додають позиційні кодування, щоб зберегти інформацію про розташування патчів.
- Використовують механізм самоспрямованої уваги для обробки глобальних залежностей між патчами.

в) Особливості:

- Добре працюють із глобальними залежностями в зображенні завдяки самоспрямованій увазі.
- Менше обмежені локальними взаємозв'язками в порівнянні з CNN.
- Потребують великої кількості даних для ефективного навчання.

г) Приклади моделей:

- Оригінальний Vision Transformer (ViT), DeiT (Data-efficient Image Transformers), Swin Transformer.

Порівняння

- Локальна vs Глобальна обробка:

- CNNs зосереджуються на локальних особливостях через згортки, тоді як ViT використовують глобальний контекст через самоспрямовану увагу.

- Ефективність:

- CNNs можуть бути більш ефективними на невеликих наборах даних, тоді як ViT зазвичай потребують великих наборів даних та більше обчислювальних ресурсів для навчання.

- Простота інтерпретації:

- Архітектура CNN часто вважається більш інтуїтивно зрозумілою через свою структуру шарів і використання згорток.
- ViT можуть бути складнішими для інтерпретації через складність механізму

уваги.

- Гнучкість:

- ViT є більш гнучкими в обробці довільних залежностей між частинами зображення, що може бути перевагою у складних задачах розпізнавання образів.

Обидва підходи мають свої переваги та недоліки, і вибір між ними залежить від конкретної задачі, доступних даних та обчислювальних ресурсів, так як ViT, попри всі свої переваги потребує в середньому більше даних та ресурсів.

2.2 Додаткова модальність даних

Більш нові роботи також розглядають додаткові джерела інформації для підвищення якості та повноти маркування зображень. Існує два основних підходи:

а) *Аналіз додаткової інформації.* Даний підхід аналізує додаткову до зображення інформацію. Це може бути як текстова інформація (теги / анотації) [8, 11], так і метадані зображення [9, 18]. Очевидним недоліком даного методу є потреба у цій додатковій інформації, яку можуть мати не всі зображення, а відсутність даної інформації знижує точність результуючого маркування.

б) *Аналіз цільових класів.*

На відміну від загальної інтерпретації класів для задачі маркування (коли кожен клас - це незалежна сутність), даний підхід аналізує зв'язки між цільовими класами, створюючи нову модальність на основі набору цільових класів [23]. Більш новим та узагальненим підходом є технологія word2vec (або аналогічне рішення), яка дозволяє впорядкувати слова у певному векторному просторі таким чином, що їх просторове значення має прямий зв'язок із їх семантичним значенням. Наприклад модель Resnet-CPSPD [21] використовує для аналізу класів граф, в якому вказуються класи (поняття) які часто знаходяться на одному зображені (co-occurrence; наприклад: риба, вода) та класи які рідко знаходяться на одному зображені (dis-occurrence; наприклад: риба, пустеля). Перевагою даного підходу є те, що йому не потрібні ніякі нові дані

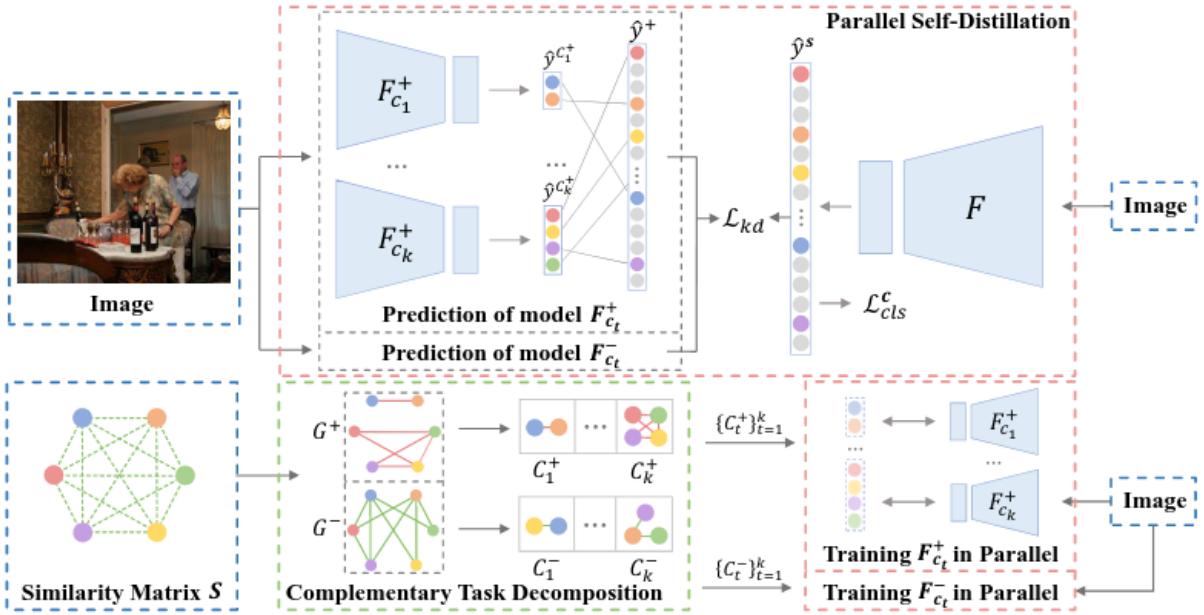


Рис. 2.1 – Приклад архітектури моделі, яка використовує графове представлення цільових класів Resnet-CPSD [21].

крім зображень та відповідних їм цільових класів, а нова модальності, яка репрезентує зв'язок між класами створюється під час тренування системи. Основним недоліком таких систем полягає у високій складності, і як наслідок - довше по часу тренування / розпізнавання.

2.3 Кількість лейблів

Image				
Truth	clouds grass house sky	leaf plants sky	animal grass	person
Top 5 pred	clouds grass house road sky	clouds grass leaf plants sky	animal grass horses plants sky	animal clouds person road sky
Model pred	clouds grass sky	plants sky	animal grass horses	person

Рис. 2.2 – Приклад адаптивної кількості лейблів, на основі роботи моделі на датасеті NUS-WIDE. 'Truth' - правдиве маркування, 'Top 5 pred' - ілюстрація вибору top k , при $k = 5$, 'Model pred' - приклад роботи моделі

Результатом роботи класифікаційної моделі є вектор ймовірностей, який репрезентує приналежність до класів. Для задачі класифікації вибір результату на основі цього вектора очевидний - клас із найбільшою ймовірністю, однак для задачі маркування все складніше. Блільшість наведених вище робоїт розглядаєуть задачу вибору k-сті лейблів як найкращі k (top k) маркувань (Розділ 4.2), де k - наперед задана константа. Очевидно, що такий вибір k-сті класів не є оптимальним, так як більш змістовні зображення будуть містити менше описової інформації і навпаки - менш змістовні будуть містити лишню інформацію, яка до того ж може не мати нічого спільногого із цим зображенням (Рис. 2.2)

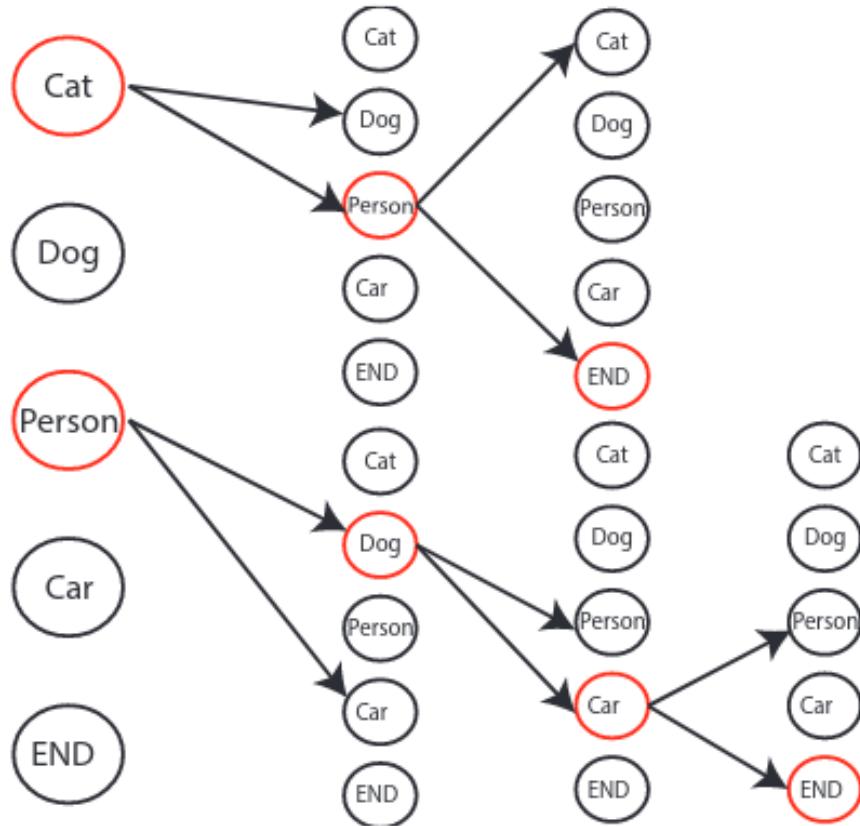


Рис. 2.3 – Приклад застосування алгоритму "beam search", для пошуку найбільш оптимального маркування зображення на основі найвищої ймовірності, який реалізовано у системі CNN-RNN [19].

Один із підходів як-от CNN-RNN [19], використовує RNN для аналізу візуальних даних (visual features) та автоматично виконує як задачу маркування, так і задачу динамічного вибору кількості лейблів, однак в силу особливості RNN є певні обмеження накладенні на порядок кодування класів. При чому важливо відмітити, що вибір динамічної кількості лейблів за допомогою RNN має суттєвий недолік - залежність від балансу цільових класів у даних.

Найновіші моделі [12, 21, 22], які розглядають зв'язок між цільовими класами, обирають к-сть цільових класів на основі порогового значення threshold (Розділ 4.2). Даний підхід є ефективним рішенням для вибору кількості лейблів, однак він релевантний тільки для цього типу моделей, так як вектор ймовірностей, який отримується на виході даної моделі досить сильно дискретизований, тобто для

позитивного лейблу, який маркується 1 ймовірність буде $\approx 0.7 - 0.9$, а для негативного, тобто 0 ймовірність $\approx 0.1 - 0.3$.

3 МОДЕЛЮВАННЯ

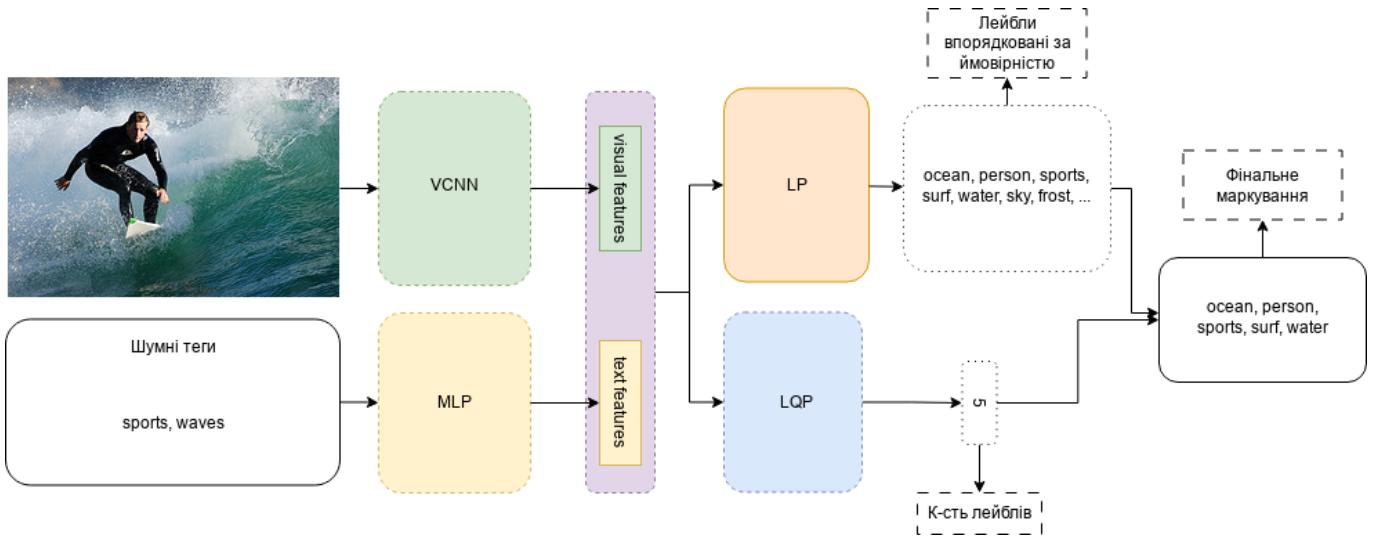


Рис. 3.1 – Структура композитної системи

На основі проведного аналізу альтернатив, дана робота пропонує розглянути модель, яка розглядає дві модальності даних: зображення та текстові теги. Структура даного рішення складається із чотирьох компонентів (Рис. 3.1).

3.1 Модель VCNN

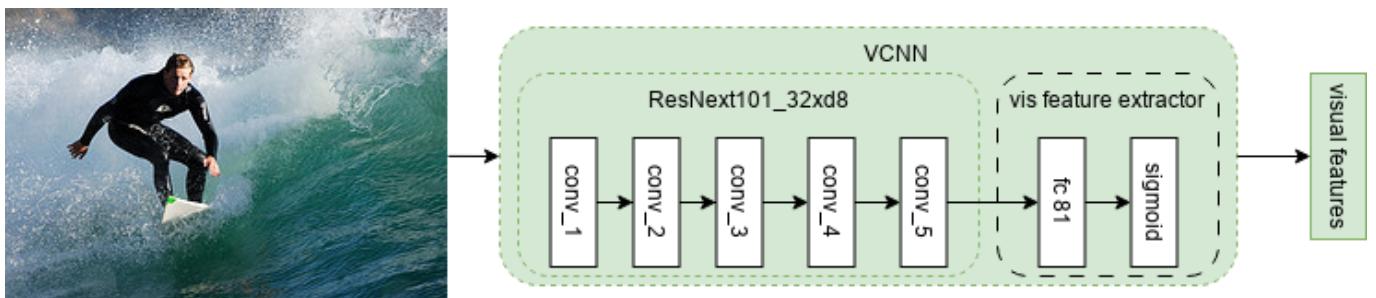


Рис. 3.2 – Архітектура моделі VCNN

Модель VCNN (Рис. 3.2) призначена для вивчення особливостей (features) із зображення. Отримує на вхід пікселі зображення I , у формі матриці розмірності (B,C,W,H) , де B - к-сть зображень у групі для тегування, C - к-сть каналів у зображеннях зазвичай 1 або 3, Grey або RGB відповідно, W,H - розмірність зображень.

За базове рішення використовується ResNext101_32x8d [20] (сучасна версія resnet), так як у порівнянні із широко застосуваною моделлю Resnet [7] вона краще описує зображення, та легше дотреновується при використанні натренованої на ImageNet [4] версії.

На виході даної моделі ми отримуємо вектор вірогідностей vf (visual feature vector), який вказує вірогідність маркування зображення класом j на основі візуальної інформації.

3.2 Модель MLP

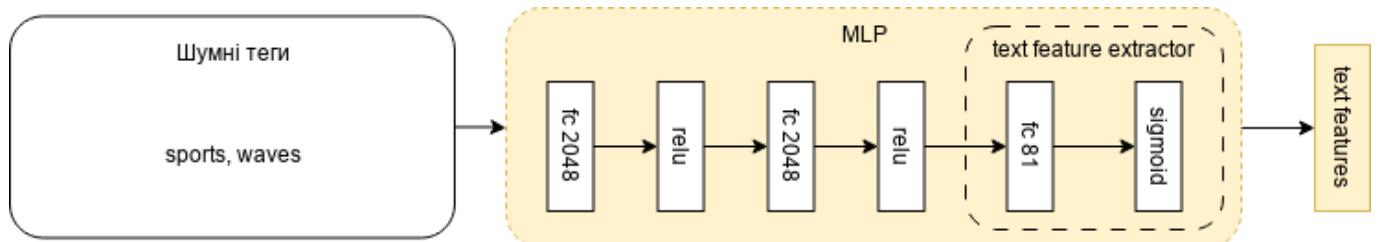


Рис. 3.3 – Архітектура моделі MLP

MLP (Рис. 3.3) - аналізує текстові особливості (text features) тегів до зображення. Теги до зображення i репрезентуються як бінарний вектор $I = [1,0,1,0, \dots, N]$, де 1 - це наявність тегу, а N - к-сть тегів.

Головна причина вибору звичайної MLP моделі для аналізу текстової інформації - це те, що вхідна інформація - це шумні теги (наприклад: для фото кота - теги "Канада", "Кіт"). Важливим правилом щодо ефективності цього рішення є

кількісне співвідношення між множиною цільових класів та тегів, воно має бути приблизно 1 до 10.

На виході даної моделі ми отримуємо вектор tf вірогідностей (text feature vector), який вказує вірогідність маркування зображення класом j на основі текстової інформації.

3.3 Модель LP

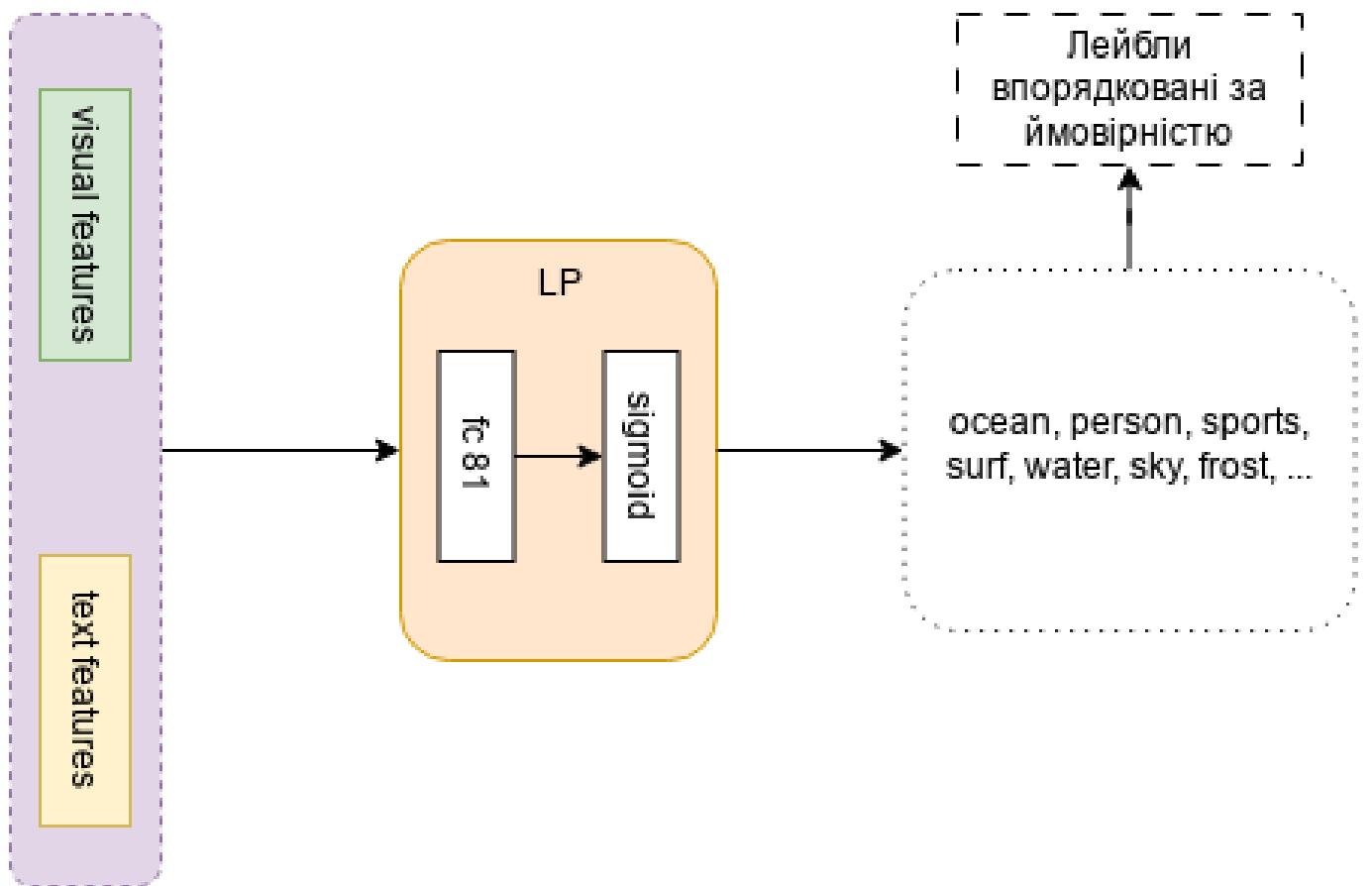


Рис. 3.4 – Архітектура моделі LP

LP (Рис. 3.4) - аналізує вектор вірогідності v , який є композицією векторів vf та tf : $v = [vf, tf]$.

На виході даної моделі ми отримуємо вектор вірогідностей, який комбінує

інформацію отриману як із візуальної так і з текстової інформації.

3.4 Модель LQP

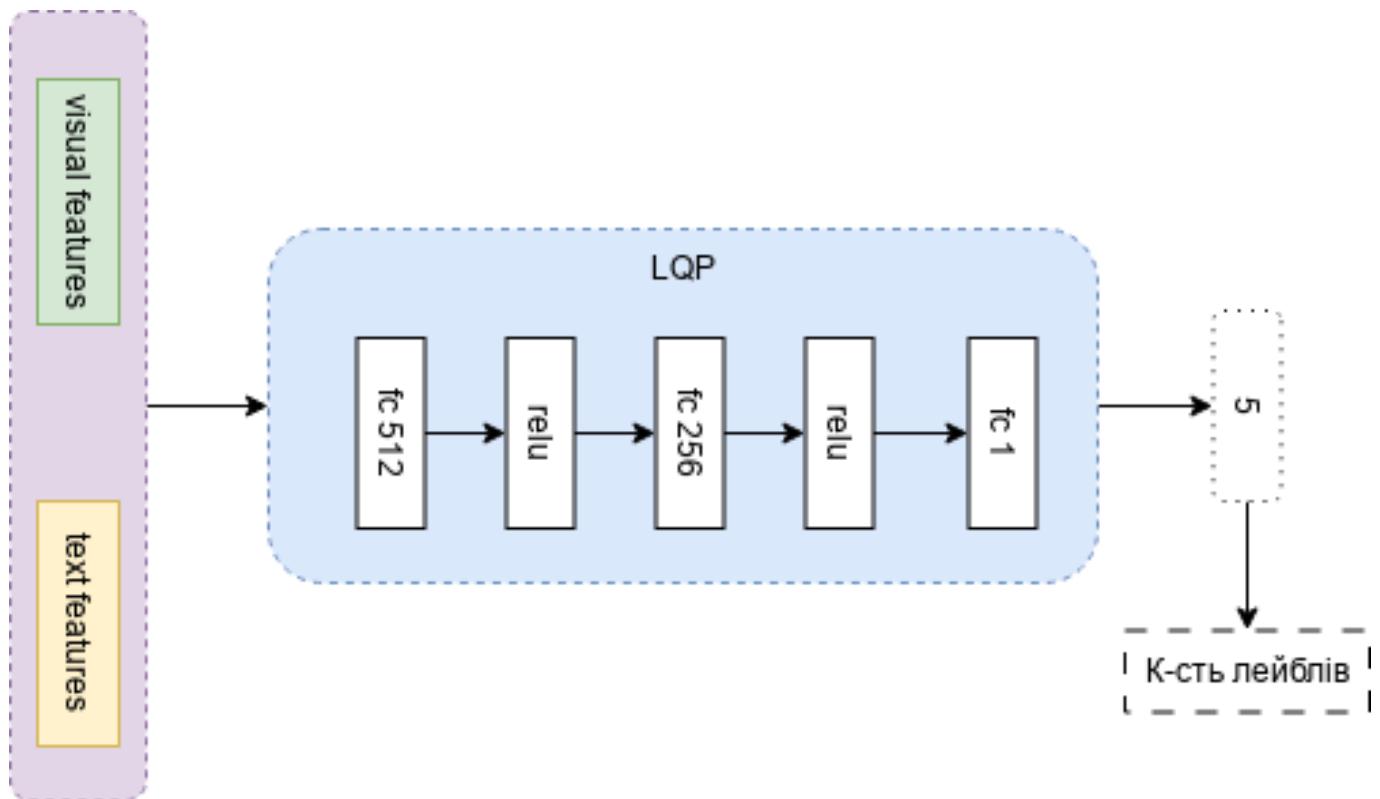


Рис. 3.5 – Архітектура моделі LQP

Модель LQP (Рис. 3.5) аналізує кількість лейблів на основі вектору вірогідностей v , який є композицією векторів vf та tf : $v = [vf, tf]$.

Існує два підходи до визначення к-сті за допомогою нейронних мереж: класифікація та регресія. LQP - регресійна модель.

Оскільки регресійні моделі досить швидко перенавчаються (overfitting), то необхідно задіяти регуляризацію. В даній роботі, в якості регуляризатора задіяні Dropout шари [16], із вірогідністю відкидання (dropout rate) 0.5.

На виході даної моделі є число, яке вказує на кількість лейблів у зображені.

3.5 Процес тренування

Модель складається із декількох компонентів, що створює декілька проблемних місць під час тренування: досить багато параметрів, дві різні цільові функції, проблема затухаючого градієнта, - тому тренування відбувається у декілька стадій, у якому кожна із моделей тренується окремо (деякі з них можна тренувати синхронно).

3.5.1 Цільові функції

Для початку варто розглянути цільові функції (функція втрат, loss function). Дані функції є базовим компонентом глибинного навчання.

В даній роботі використовуються дві функції: BCEWithLogitsLoss та MSELoss.

BCEWithLogitsLoss

Для тренування класифікаційних моделей (VCNN, MLP, LP) вихідні логіти z_{ij} для групи (batch) зображень I_N при $i = 1 \dots N$, $j = 1 \dots C$, де N - кількість зображень в групі, C - кількість цільових класів, цільова функція має вигляд:

$$\mathcal{L}_{cls} = \frac{1}{NC} \sum_i^N \sum_j^C y_{ij} \cdot \ln(\sigma(z_{ij})) + (1 - y_{ij}) \cdot \ln(1 - \sigma(z_{ij})) \quad (3.1)$$

, де $y_{ij} = 1$ якщо зображення i анатоване класом j , інакше - $y_{ij} = 0$, а $\sigma(\cdot)$ - це сигмоїdalна активаційна функція

MSELoss

Для тренування регресійної моделі LQP вихідні логіти z_i для групи (batch)

зображень I_N при $i = 1\dots N$, де N - кількість зображень в групі, цільова функція має вигляд:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_i^N (y_i - z_i)^2 \quad (3.2)$$

, де y_i - це кількість лейблів для зображення I_i .

3.5.2 Тренування VCNN

Тренування моделі ResNext [20] з нуля є досить складною задачою (дана модель має $\approx 80M$ параметрів), адже для цього потрібні значні обчислювальні потужності.

Саме тому використовується натренована модель із адаптованим класифікатором (visual feature extractor) (Рис. 3.1), яка підганяється (finetuned) на обраному датасеті.

Існує два підходи для підгонки:

1) Підгонка всієї моделі (finetuning): всі шари моделі підганяються (дотреновуються) з низькою швидкістю навчання (learning rate). Даний підхід вимагає великої обчислювальної потужності, однак надає високу точність та досить таки швидко тренується (в порівнянні із тренуванням з нуля).

2) Підгонка класифікатора (transfer learning): відбувається тренування тільки класифікатора, фіксуючи всі інші параметри моделі. Даний підхід значно пришвидшує тренування в обмін на певну деградацію точності в порівнянні із 1-им варіантом.

В даній роботі використовується 1 варіант підгонки, так як він надає вищу точність.

3.5.3 Тренування MLP

Дана модель є звичайним багатошаровим персепtronом, тому її можна без проблем натренувати з нуля.

Також цю модель можна тренувати паралельно із VCNN.

3.5.4 Тренування LP

Дана модель призначена для обрахування вірогідностей на основі вектору $f = [vf, tf]$, оскільки вона складається із одного шару то її тренування також очевидне.

Також цю модель можна тренувати паралельно із LQP.

3.5.5 Тренування LQP

Дана модель є регресійним багатошаровим персепtronом, її тренування також є очевидним, однак варто нормалізувати вхідну к-сть лейблів, так як це пришвидшить та/або покращить збіжність моделі.

Під час тренування використовуються шари Dropout, так як дана модель дуже швидко перенавчається. В даній роботі вірогідність відкидання (dropout rate) 0.5.

Також цю модель можна тренувати паралельно із LP.

3.6 Процес тестування

3.6.1 Тестування класифікаційних моделей (VCNN, MLP, LP)

Кожна із даних моделей обраховує вектор ймовірностей P , для тестування необхідно перевести вектор ймовірностей (наприклад: $P = [0.9, 0.6, 0.1, 0.4, 0.6]$) у вигляд маркування (наприклад: $M = [1, 1, 0, 0, 1]$). Дане перетворення називається індикаторною функцією.

Розглянемо два основних види індикаторної функції:

- 1) Порогове значення (threshold): для вектору ймовірностей P та порогу α - вектор маркувань обраховується наступним чином: якщо $y_i > \alpha$, то маркуємо 1, інакше 0. Наприклад при $\alpha = 0.5$: $[0.9, 0.6, 0.1, 0.4, 0.6] \rightarrow [1, 1, 0, 0, 1]$.
- 2) Найкращі k (top k): для вектору ймовірностей P та числа k - вектор маркувань обраховується наступним чином: маркуємо 1 найкращі k ймовірностей, інакше 0. Наприклад при $k = 4$: $[0.9, 0.6, 0.1, 0.4, 0.6] \rightarrow [1, 1, 0, 1, 1]$.

Оскільки дані моделі **не передбачають** передбачення кількості лейблів, то для тренування даних моделей використовується метод top k , при чому $k = 3$.

Тобто для зображення I , із маркуванням $Y = [1, 0, 0, 0, 1]$, і вектором ймовірностей $P = [0.8, 0.9, 0.1, 0.5, 0.2]$ та результиручим вектором маркувань M : $k = 3$, перетворення $P \rightarrow M \equiv [0.8, 0.9, 0.1, 0.3, 0.2] \rightarrow [1, 1, 0, 1, 0]$

3.6.2 Тестування композитної моделі

Для тестування композитної моделі (Рис. 3.1) необхідно обрахувати результиуючі значення для моделей LP та LQP, і обрати top k лейблів LP, де k - це передбачення LQP.

4 ЕКСПЕРИМЕНТИ

4.1 Датасет

Один із найбільш часто використовуваних датасетів для тестування моделей маркування зображень - NUS-WIDE [2], він складається із 269,655 зображень, 81 лейблу, та ≈ 5000 тегів в якості сторонньої текстової інформації. Для проведення тренування/тестування використовується розподіл, надведений разом із датасетом, так як він є збалансований настільки, наскільки це можливо (Рис. 4.2).

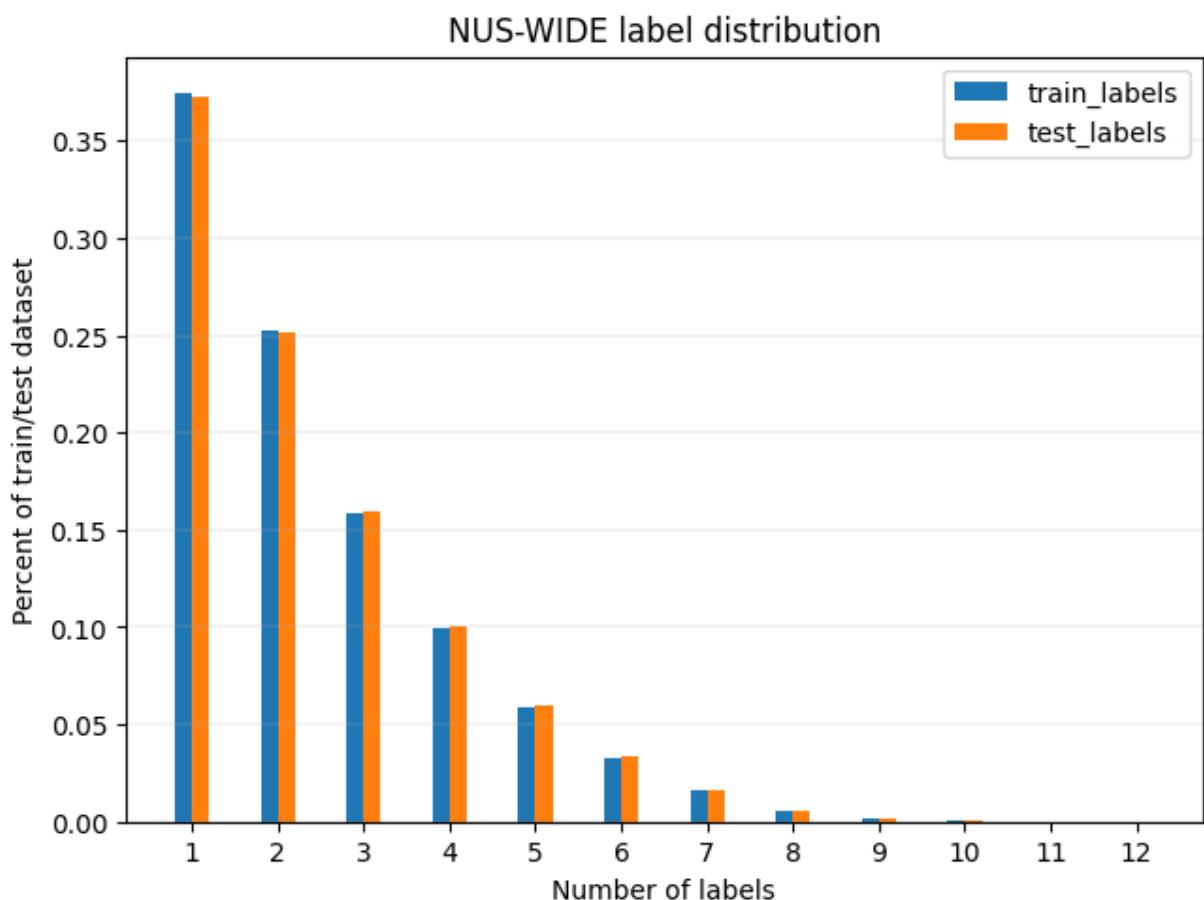


Рис. 4.1 – Розподіл лейблів у тренувальному/тестовому датасеті

З наведеного графіку можна зробити висновок, що він містить значний дизбаланс відносно кількості класів, так як він був зібраний на основі реальних даних, якими користувались люди, із вебсайту 'Flickr'.

Важливо відмітити, що даний датасет містить посилання на зображення на ресурсі 'Flickr', і деякої частина цих зображень вже там немає.

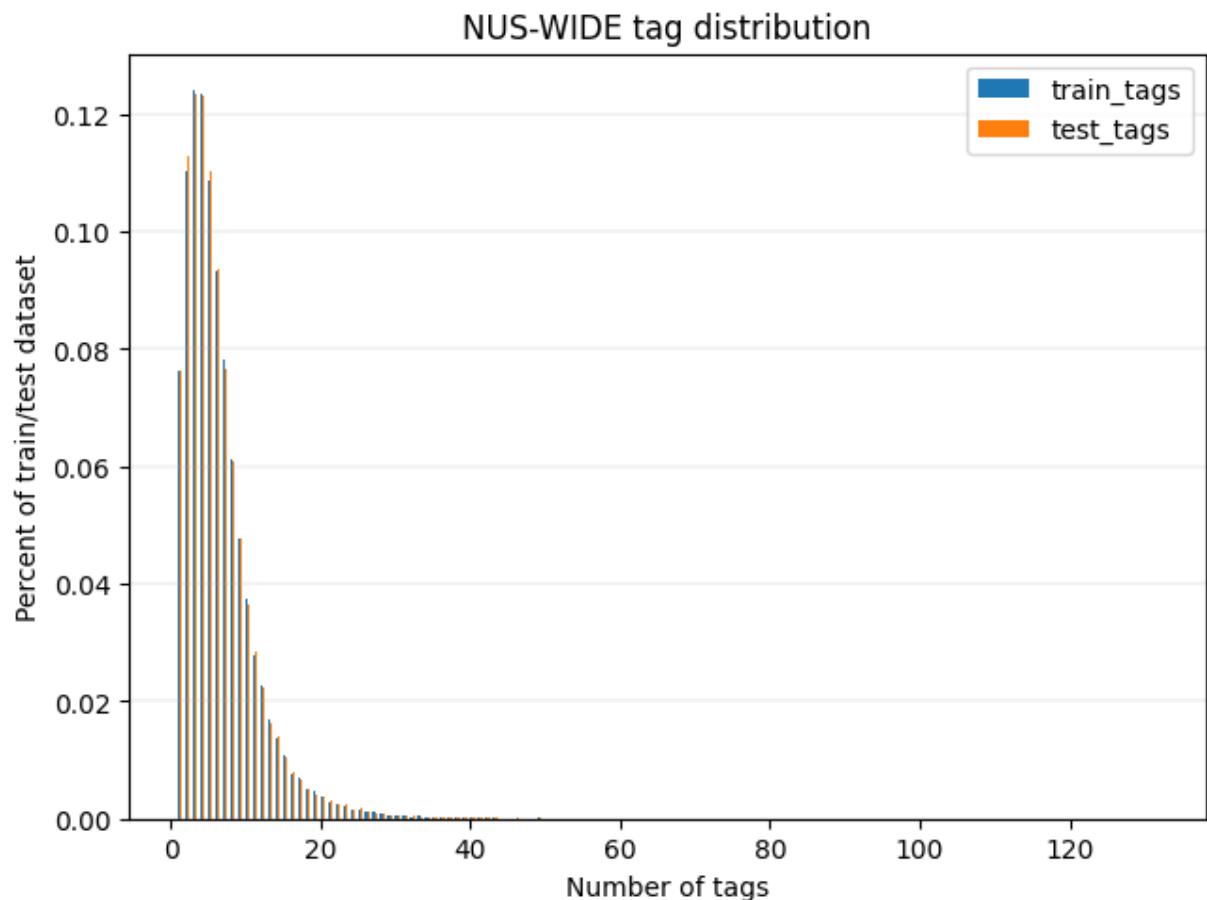


Рис. 4.2 – Розподіл тегів у тренувальному/тестовому датасеті

Також буде використано тільки 1000 найбільш частих тегів з ≈ 5000 , при чому зображення, які не містять жодного тега відфільтровано.

4.1.1 Множина цільових класів

Датасет NUS-WIDE [2] містить 81 цільовий клас.



Рис. 4.3 – Візуалізація цільових класів у датасеті. Класи із більшою частотою мають більший шрифт тексту та темніший колір.

Множина цільових класів містить значний дизбаланс відносно частоти.

4.1.2 Числові характеристики датасету

	Тренування	Тестування
Кількість зображень	121962	81636
Середня к-сть лейблів	2.42	2.43
Медіана к-сть лейблів	2	2
Мінімальна к-сть лейблів	1	1
Максимальна к-сть лейблів	12	13
Середня к-сть тегів	6.27	6.26
Медіана к-сть тегів	5	5
Мінімальна к-сть тегів	1	1
Максимальна к-сть тегів	131	125

Табл. 4.1 – Характеристики тренувальної/тестової вибірок

З наведеної таблиці варто відзначити характеристики, що стосуються тегів. Так як теги - це стороння інформація, враховуючи що медіана = 5, то варто розглянути який вплив несе менша кількість тегів для маркування, так як зазвичай при завантаженні люди додають власноруч приблизно 3 теги. Тому в подальшому буде розглянуто вплив тегів, при виборі фіксованої максимальної кількості тегів.

4.2 Тестові метрики

Для оцінки точності будуть використовуватись метрики, які є загально вживаними для оцінки задачі маркування зображень (multi-label image annotation).

$$\begin{aligned}
 C\text{-P} &= \frac{1}{C} \sum_{j=1}^C \frac{NI_j^c}{NI_j^p} & O\text{-P} &= \frac{\sum_{i=1}^N NL_i^c}{\sum_{i=1}^N NL_i^p} \\
 C\text{-R} &= \frac{1}{C} \sum_{j=1}^C \frac{NI_j^c}{NI_j^g} & O\text{-R} &= \frac{\sum_{i=1}^N NL_i^c}{\sum_{i=1}^N NL_i^g} \\
 C\text{-F1} &= \frac{2 \cdot C\text{-P} \cdot C\text{-R}}{C\text{-P} + C\text{-R}} & O\text{-F1} &= \frac{2 \cdot O\text{-P} \cdot O\text{-R}}{O\text{-P} + O\text{-R}}
 \end{aligned} \tag{4.1}$$

,де

- * C - к-сть класів
- * N - к-сть тестових зображень
- * NI_j^c - к-сть зображень які **коректно** промарковано як клас j
- * NI_j^g - к-сть зображень які мають клас j
- * NI_j^p - к-сть зображень які промарковано як клас j
- * NL_i^c - к-сть **коректно** промаркованих лейблів для зображення i
- * NL_i^g - к-сть лейблів які має зображення i
- * NL_i^p - к-сть промаркованих лейблів для зображення i

Варто відзначити, що дані метрики є зміщеними (biased), при чому по-класові метрики (C) зміщені в сторону рідкісних класів, а загальні метрики (O) - в сторону частих класів [6].

Для того щоб отримати унформене представлення про ефективність моделі буде використовуватись наступна метрика, яка бере до уваги як C-F1 так і O-F1, що полегшує інтерпретацію результатів:

$$H-F1 = \frac{2 \cdot C-F1 \cdot O-F1}{C-F1 + O-F1} \quad (4.2)$$

4.3 Тренування

Для програмної реалізації запропонованої моделі було використано PyTorch.

Оскільки в даній роботі, використовується модель ResNext [20] натренована на датасеті ImageNet [4], то для вхідних зображень потрібно застосувати певне перетворення:

- 1) Зміна розміру (Resize) 232×232 , використовуючи білінійну інтерполяцію
- 2) Центральний кроп (Central crop) 224×224
- 3) Зміна масштабу (Rescale) $[0,1]$
- 4) Нормалізація на основі статистичних величин ImageNet [4]. А саме: mean = $[0.485, 0.456, 0.406]$ та std = $[0.229, 0.224, 0.225]$

Дане перетворення доступно у бібліотеці PyTorch.

Параметри навчання

Класифікаційні моделі VCNN та MLP навчалися із швидкістю навчання (learning rate) 0.001, а LP - зі швидкістю 0.01. Також для начання цих трьох моделей використовувався контроллер швидкості навчання (learning rate scheduler), який множив швидкість навчання на 0.5, досягаючи 5-ої та 10-ої епохи.

Регресійна модель LQP навчалась із сталою швидкістю навчання (learning rate) 0.0005.

Для всіх навчання всіх вище згаданих моделей використовувався оптимізатор AdamW, із параметром 11 регуляризації (weight decay) 0.0003.

Розмір групи (batch size) 32.

Також варто відзначити, що в даній роботі епоха - це 20% від усіх даних, при чому після кожної епохи дані перемішуються (shuffle), отримаючи нові 20% даних.

Процес навчання

Для тренування було використано графічний процесор 'Nvidia L4'. Для тренування всіх елементів моделі знадобилось ≈ 3 години.

Для оптимізації процесу тренування було застосовано техніку mixed precision, яка використовує f16 замість f32, під час певних етапів тренування [13].

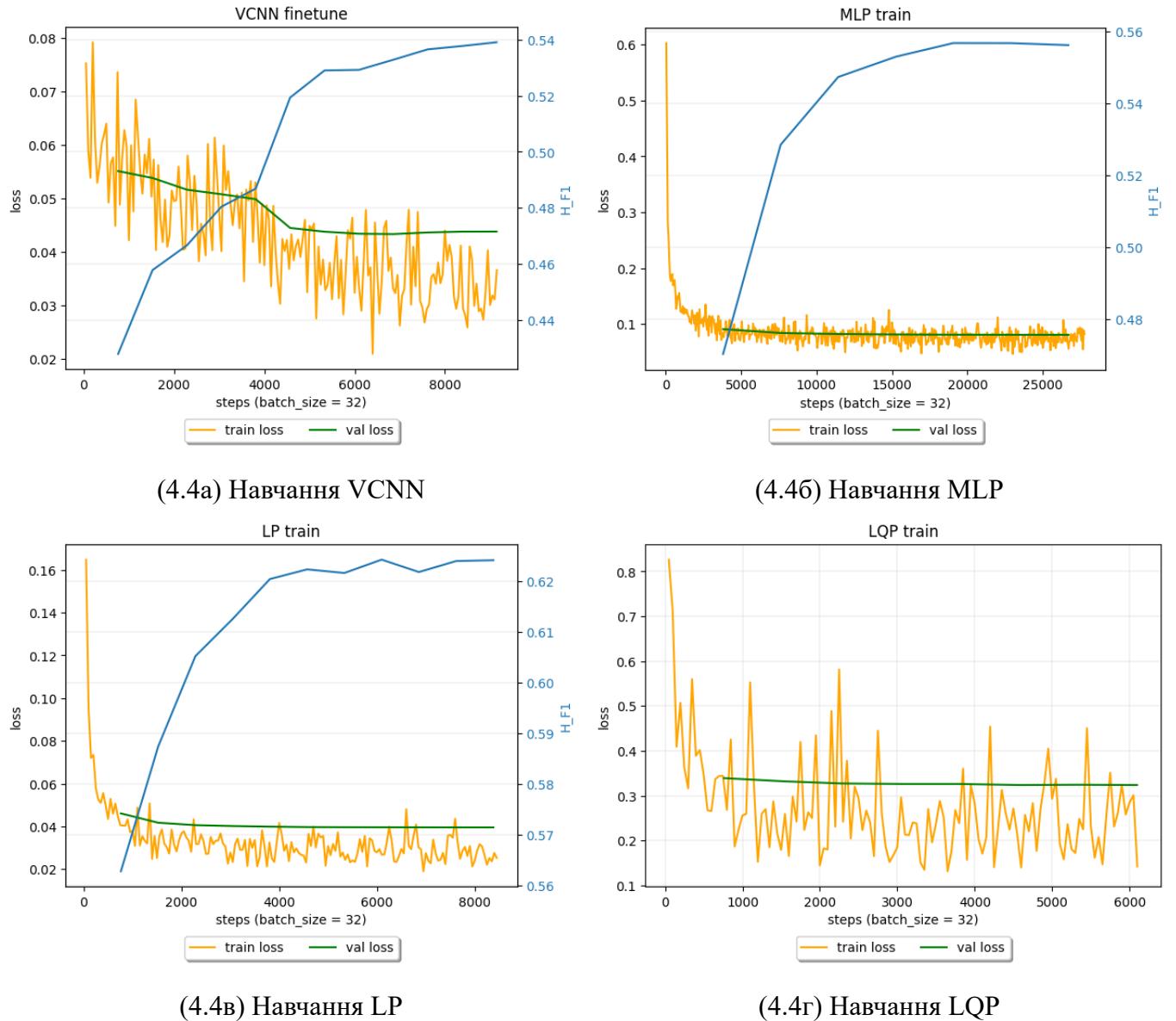


Рис. 4.4 – Процес навчання моделей

5 АНАЛІЗ РЕЗУЛЬТАТІВ

5.1 Аналіз компонентів системи

Модель	Індикаторна ф-ція (3.6.1)	Модальність	C-P	C-R	C-F1	O-P	O-R	O-F1	H-F1
Композитна	top k	Зображення+теги	72.49	59.51	65.36	76.53	74.41	75.46	70.04
VCNN+MLP+LP	top 3	Зображення+теги	60.51	61.28	60.89	67.87	71.52	63.98	62.40
VCNN+LQP	top k	Зображення	64.62	37.61	47.54	77.07	59.01	66.84	55.56
VCNN	top 3	Зображення	44.48	53.32	48.50	55.44	68.52	61.29	54.15

Табл. 5.1 – Порівняння компонентів моделі

Результатуюча композитна модель показала значно кращий результат ніж базове рішення (VCNN).

Додаткова модальність (MLP+LP)

Додавання додаткової модальності у вигляді тегів внесло значний вклад у підвищення якості маркування. Порівнюючи відповідні метрики H-F1 для моделей VCNN та VCNN+MLP+LP, можна побачити приріст на 8.25%. При чому варто відзначити, що цей ріст в основному забезпечений приростом метрики C-F1, яка є зміщеною в сторону більш рідкісних класів. Варто відзначити, що це працює завдяки тому що зазвичай теги надані користувачами відмічають досить рідкісні поняття, які на відміну від частих тегів (небо, сонце, людина, вода і тд.) складно розпізнати маючи одне лише зображення.

Передбачення кількості лейблів (LQP)

При використанні компоненту LQP кількість лейблів обирається за принципом 'top k', а не 'top 3' (3.6.1). Це очевидним чином підвищує точність фінального маркування, адже деякі зображення можуть мати більше трьох лейблів, інші - менше трьох. Порівнюючи вплив компоненти LQP для базового рішення (VCNN) та композитної

моделі (VCNN+MLP+LP+LQP) можна зробити припущення, що VCNN аналізує загальні поняття на зображенні, а враховуючи дизбаланс класів у датасеті, використання принципу 'top k' збільшує точність (precision), сильно жертвуючи по-класовим охопленням (C-R), і, як наслідок, не сильно збільшує величину головної метрики H-F1. На практиці це виливалось у те, абсолютна більшість зображень маркувалась частими лейблами (людина, вода, небо і тд.), а рідкісні теги - ігнорувались. Натомість у композитній моделі вищезгаданий принцип чудово проявив себе. Згідно із тестовими метриками покращення становить 7.64%.

5.2 Аналіз впливу додаткової інформації

Максимальна кількість тегів	C-P	C-R	C-F1	O-P	O-R	O-F1	H-F1
1	72.49	59.51	65.36	76.53	74.41	75.46	70.04

Табл. 5.2 – Порівняння впливу кількості тегів на маркування

5.3 Порівняння з існуючими рішеннями

Модель	Індикаторна ф-ція (3.6.1)	Модальність	C-P	C-R	C-F1	O-P	O-R	O-F1	H-F1
Композитна	top k	Зображення+теги	72.49	59.51	65.36	76.53	74.41	75.56	70.04
Query2Label [12]	threshold α	Зображення (+аналіз класів)	-	-	67.60	-	-	76.3	71.69
SR-CNN-RNN [11]	top 3	Зображення+теги	71.73	61.73	66.36	77.41	76.88	77.15	71.35
Resnet-CPSD [21]	threshold α	Зображення (+аналіз класів)	-	-	64.00	-	-	75.30	69.19
MS-CMA [22]	threshold α	Зображення (+аналіз класів)	-	-	60.50	-	-	73.80	66.49
Resnet-SRN [23]	threshold 0.5	Зображення (+аналіз класів)	65.20	55.80	58.50	75.50	71.50	73.40	65.10
SINN [8]	top 3	Зображення+теги	58.30	60.63	59.44	57.05	79.12	66.29	62.68
TagNeighbour [9]	top 3	Зображення+метадані	54.74	57.30	55.99	53.46	75.10	62.46	59.05
CNN+Logistic [8]	top 3	Зображення	45.60	45.03	45.31	51.32	70.77	59.50	51.44
CNN-RNN [19]	top 3	Зображення	40.50	30.40	34.70	49.9	61.70	55.20	42.61
CNN+WARP [6]	top 3	Зображення	31.65	35.60	33.51	48.59	60.49	53.89	41.32
CNN+Softmax [6]	top 3	Зображення	31.68	31.22	31.45	47.82	59.52	53.03	39.48

Табл. 5.3 – Порівняння результатуючих метрик для різних моделей на датасеті
NUS-WIDE

Точність запропонованого рішення Композитна модель (VCNN+MLP+LP+LQP) продемонструвала високу якість маркування на тестових метриках у порівнянні із розглянутими альтернативними рішеннями. Згідно із метрикою H-F1 запропоноване рішення є третім.

Модальність даних Задача маркування зображень розглядає зображення як основну модальність, однак додавання модальності, очікувано, покращує результати маркування. Це підтверджують метрики наведені в Табл. 5.3. Серед розглянутих рішень є 3 варіанти модальності даних з якимим працюють нейронні мережі.

Найменш ефективним, як і очікувалось, виявились моделі які аналізують виключно зображення. Введення інших двох видів додаткових модальностей: теги, аналіз класів, - надають значно кращі результати. Варто відзначити, - аналіз класів (найкраще імплементовано в: Query2Label [12], Resnet-CPSC [21] та MS-CMA [22]) не потребує ніяких додаткових даних окрім зображення, що є вагомою конкурентною перевагою, враховуючи незначну відміність в точності моделей.

Індикаторна функція Для оцінки ефективності запропонованої індикаторної функції 'top k', варто ізолювати вплив саме цієї функції. Для цього розглянемо існуючі моделі, які працюють із тією ж модальністю даних. Найкращою із таких моделей є SINN [8]. Використання під-системи LQP, яка передбачає роботу із динамічною кількістю лейблів при маркуванні (top k) значено підвищує якість. Згідно із наведеними метриками покращення складає 7.36% (Табл. 5.3), що є вагомим приростом.

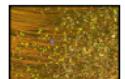
5.4 Демонстративні приклади

З тестового датасету випадковим чином обрано декілька зображень, для демонстрації роботи моделі:

Image	Truth	Model pred	Image	Truth	Model pred
	clouds sky	clouds sky		snow water	glacier lake water
	flowers plants	flowers plants		clouds mountain rocks sky	mountain rocks sky
	cityscape clouds sky	buildings clouds nighttime sky		buildings clouds	flowers
	clouds	clouds sky		clouds person sky	buildings person sky

(5.1a)

(5.1б)

Image	Truth	Model pred	Image	Truth	Model pred
	clouds ocean water	clouds military sky		person wedding	person wedding
	person	person sky		bear	animal
	clouds moon ocean sky water	lake moon ocean sky water		clouds nighttime sky window flowers	buildings clouds nighttime sky
	flowers grass water	grass plants		garden grass plants water	flowers garden grass plants

(5.1в)

(5.1г)

Рис. 5.1 – Демонстративні приклади

Задамо умовне позначення "a::1", що означає демонстративний приклад "a", перше зображення зверху, "в::2" - приклад "в", друге зображення зверху і тд.

Серед наведених прикладів можна розглянути кілька цікавих моментів, які не

відображають тестові метрики:

* Іноді модель передбачає маркування, якого немає у датасеті, однак присутнє на зображені. Наприклад: [a::3,a::4,б::1,б::3,б::4,в::1,в::2,г::2,г::3].

* Існують випадки коли модель відмічає поняття, які не можуть бути присутніми на одному зображені. Це є прямим наслідком того, що наша модель розглядає цільові класи як незалежі сутності. Наприклад: lake та ocean як-от в прикладі 'в::3'.

* Іноді передбачення моделі відсікають неіснуючі поняття та маркують зображення краще ніж це було зроблено в датасеті. Так для зображення 'б::3' на якому зображено якусь рослину датасет вказує що це: 'buildings, clouds'; а модель - 'flowers'.

Окрім наведених вище особливих випадків, іноді модель, звичайно, помиляється. Однак у наведених прикладах немає значних помилок у маркуванні.

ВИСНОВКИ

В даній роботі було розглянуто композитну модель, для маркування зображень (шпалерів робочого столу) проведено оцінювання її точності на тестових метриках.

Розглянута модель показала хороший результат у порівнянні із існуючими альтернативними рішеннями.

До переваг розглянутого рішення належать:

- + Висока точність
- + Невелика кількість параметрів ($\approx 95\%$ параметрів має модель для аналізу зображень)
- + Висока швидкість тренування

Недоліками є:

- Неможливість тренування моделі в один етап (end-to-end)
- Необхідність використання додаткових даних (тегів) для отримання високої якості опису зображення

В подальшому варто розглянути ефективність даної моделі на інших датасетах (наприклад MSCOCO [10]). також варто розглянути і методи для аналізу цільових класів, так як незважаючи на значне ускладнення фінальної моделі це надає досить високий приріст до точності маркування.

Результатуюча композитна модель збережена у форматі safetensors.

Після проведеного дослідження було висунуто гіпотезу, що дане рішення можна застосувати і у інших схожих предметних областях. Так для аналізу рентгенівських знімків - це може бути історія хвороб пацієнта, для супутниковых знімків - різні метадані, геолокація тощо, а для аналізу звичайних фотографій - теги, анотації, метадані, тощо.

ПЕРЕЛІК ПОСИЛАНЬ

- [1] Krizhevsky Alex, Sutskever Ilya та Hinton Geoffrey. „ImageNet Classification with Deep Convolutional Neural Networks“. В: *Advances in Neural Information Processing Systems*. За ред. F. Pereira та ін. Т. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- [2] Tat-Seng Chua та ін. „NUS-WIDE: A Real-World Web Image Database from National University of Singapore“. В: *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*. Santorini, Greece., July 8-10, 2009.
- [3] Dan Cireşan, Ueli Meier та Juergen Schmidhuber. „Multi-column deep neural networks for image classification“. В: (лют. 2012). arXiv: 1202.2745 [cs.CV].
- [4] Li Deng та ін. „Imagenet: A large-scale hierarchical image database“. В: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, С. 248—255.
- [5] Alexey Dosovitskiy та ін. „An image is worth 16x16 words: Transformers for image recognition at scale“. В: (жовт. 2020). arXiv: 2010.11929 [cs.CV].
- [6] Yunchao Gong та ін. *Deep Convolutional Ranking for Multilabel Image Annotation*. 2013. eprint: arXiv:1312.4894.
- [7] Kaiming He та ін. „Deep residual learning for image recognition“. В: (груд. 2015). arXiv: 1512.03385 [cs.CV].
- [8] Hexiang Hu та ін. „Learning structured inference neural networks with label relations“. В: (листоп. 2015). arXiv: 1511.05616 [cs.CV].

- [9] Justin Johnson, Lamberto Ballan та Fei-Fei Li. „Love thy neighbors: Image annotation by exploiting image metadata“. B: (серп. 2015). arXiv: 1508.07647 [cs.CV].
- [10] Tsung-Yi Lin та ін. „Microsoft COCO: Common Objects in Context“. B: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: <http://arxiv.org/abs/1405.0312>.
- [11] Feng Liu та ін. „Semantic Regularisation for Recurrent Image Annotation“. B: (листоп. 2016). arXiv: 1611.05490 [cs.CV].
- [12] Shilong Liu та ін. „Query2Label: A Simple Transformer Way to Multi-Label Classification“. B: (лип. 2021). arXiv: 2107.10834 [cs.CV].
- [13] Paulius Micikevicius та ін. „Mixed Precision Training“. B: (жовт. 2017). arXiv: 1710.03740 [cs.AI].
- [14] Keiron O'Shea та Ryan Nash. „An Introduction to Convolutional Neural Networks“. B: (листоп. 2015). arXiv: 1511.08458 [cs.NE].
- [15] Karen Simonyan та Andrew Zisserman. „Very deep convolutional networks for large-scale image recognition“. B: (вер. 2014). arXiv: 1409.1556 [cs.CV].
- [16] Nitish Srivastava та ін. „Dropout: A Simple Way to Prevent Neural Networks from Overfitting“. B: *Journal of Machine Learning Research* 15.56 (2014), C. 1929—1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html>.
- [17] Christian Szegedy та ін. *Going Deeper with Convolutions*. 2014. eprint: arXiv: 1409.4842.
- [18] Kevin Tang та ін. „Improving image classification with location context“. B: (трав. 2015). arXiv: 1505.03873 [cs.CV].
- [19] Jiang Wang та ін. „CNN-RNN: A unified framework for multi-label image classification“. B: (квіт. 2016). arXiv: 1604.04573 [cs.CV].

- [20] Saining Xie та ін. „Aggregated residual transformations for deep neural networks“. В: (листоп. 2016). arXiv: 1611.05431 [cs.CV].
- [21] Jiazhi Xu та ін. „Boosting multi-Label Image Classification with complementary Parallel Self-distillation“. В: (трав. 2022). arXiv: 2205.10986 [cs.CV].
- [22] Renchun You та ін. „Cross-modality attention with semantic graph embedding for multi-label classification“. В: (груд. 2019). arXiv: 1912.07872 [cs.CV].
- [23] Feng Zhu та ін. „Learning spatial regularization with image-level supervisions for multi-label image classification“. В: (лют. 2017). arXiv: 1702.05891 [cs.CV].

Додаток А

Код лістинг

Код міститься у публічному github репозиторії [посилання](#)

Далі наведено загальний опис елементів проекту:

Дані:

Файли в директорії scripts 'nuswide2ndjson.py' та '1ktags.py' призначені для обробки сиріх даних із датасету в формат ndjson для подальшого тренування.

Файл data.py містить адаптери та визначення датасету для тренування.

Тренування:

Скрипти для тренування для всіх моделей (VCNN, MLP, LP, LQP) знаходяться в директорії 'scripts/train'.

Тестування:

Скрипт для тестування моделей: 'scripts/test.py'. Даний скрипт передбачає тестування як фінальної моделі, так і деяких конфігурацій її компонентів.

Інше:

Скрипт 'scripts/compose2safe.py' призначений для конвертації вагів моделей (VCNN, MLP, LP, LQP) формату .ckpt у композитну модель формату .safetensors.

Ноутбук 'testing.ipynb' призначений для наглядного тестування моделі.

Додаток В

Додаткові приклади

Image	Truth	Model pred
	animal birds sky	animal birds sky
	person wedding	person wedding
	buildings	sky
	animal	animal birds

Рис. В.1

Image	Truth	Model pred
	animal clouds	animal
	animal	animal
	flowers	flowers
	animal elk snow	animal snow

Рис. В.2

Image	Truth	Model pred
	sky	animal clouds horses sky
	train	railroad train
	buildings clouds sky window	buildings clouds sky town window
	clouds grass sky vehicle	clouds sky vehicle window

Рис. В.3

Image	Truth	Model pred
	clouds sky	buildings clouds grass sky
	buildings plants	buildings grass sky
	person	person
	animal dog	animal dog

Рис. В.4

Image	Truth	Model pred
	person	person window
	grass road vehicle	road vehicle
	lake ocean water	beach lake ocean water
	flowers plants	flowers

Рис. В.5

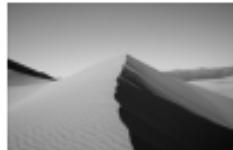
Image	Truth	Model pred
	snow	snow
	animal cat	animal cat
	animal elk lake water	animal lake water
	road sand	sand sky

Рис. В.6

Image	Truth	Model pred
	plane	plane sky
	clouds flowers garden grass mountain plants sky	clouds grass plants road sky
	animal beach	beach
	grass house	grass plants sky tree

Рис. В.7

Image	Truth	Model pred
	flowers plants	flowers plants
	animal	animal
	buildings clouds sky snow tower	buildings clouds sky
	animal bear snow	animal

Рис. В.8

Image	Truth	Model pred
	buildings	buildings town water
	clouds sky sunset	sky sunset
	person	person
	sky	clouds sky

Рис. В.9

Image	Truth	Model pred
	sky	grass sky
	road train window	sky train
	animal grass	animal grass
	water	person

Рис. В.10

Image	Truth	Model pred
	animal whales	animal fish
	beach clouds sky	clouds sky
	buildings	window
	grass house sky	buildings grass house

Рис. В.11

Image	Truth	Model pred
	beach clouds lake sky sunset beach water clouds	clouds ocean sky sunset water clouds lake
	ocean sky sunset water clouds sky sunset	ocean sky sunset water clouds fire sky sunset
		
	flowers sky	flowers sky

Рис. B.12

Image	Truth	Model pred
	animal	animal
	person	person
	airport grass plane	airport clouds plane sky
	buildings clouds tower	sky tower

Рис. B.13

Додаток Г

Ілюстративний матеріал

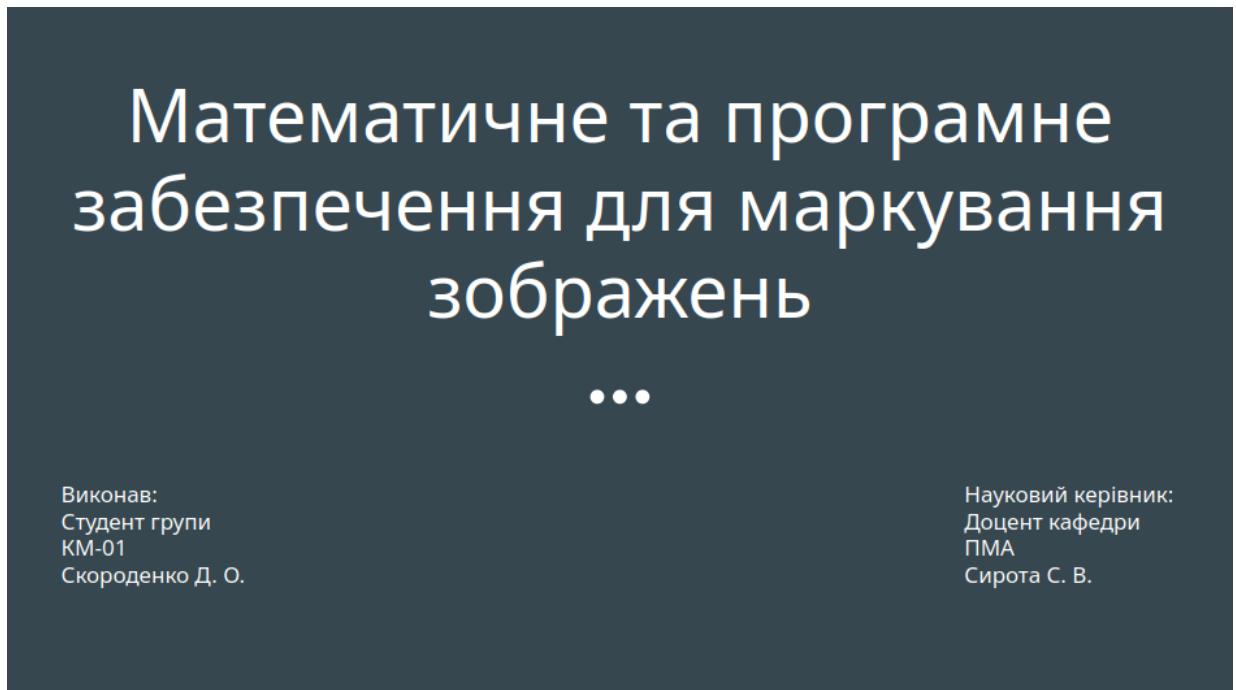


Рис. Г.1 – Слайд 1

Вибір теми. Актуальність.

В сучасному світі, коли людство виробляє все більше і більше даних на добу через стрімку цифровізацію, виникає необхідність структурувати ці дані для того щоб в подальшому мати змогу здійснювати якісний пошук по певним категоріям. Це, зокрема, справедливо для зображень. Для їх структуризації можна використовувати лейблі, а для автоматизації процесу присвоєння лейблів зображеню потрібно використовувати систему для маркування.

Існуючі рішення задачі маркування зображення поділяються на дві категорії:

- Прості нейронні мережі. Такі рішення не дуже виагливі до обчислювальних потужностей, однак не надають високої якості маркування.
- Складні нейронні мережі (композиція нейронних мереж). Такі рішення виагливі до обчислювальних потужностей, потребують значно більше часу на тренування, але значно якісніше маркують зображення.

Дана робота націлена на вирішення цих проблем, а саме створення системи для маркування зображення, яка має просту архітектуру, швидко тренується та якісно маркує зображення.

Рис. Г.2 – Слайд 2

Постановка задачі

Об'єктом дослідження є маркування зображень, та методи покращення маркування зображення. Для порівняння ефективності маркування серед існуючих рішень було обрано моделі, для яких яких обраховані тестові метрики для того ж датасету, який обрано в даній роботі.

Предметом дослідження є множина шпалерів робочого столу в якості основної модальності даних, та додаткова інформація (надані людьми шумні теги) в якості додаткової.

Рис. Г.3 – Слайд 3

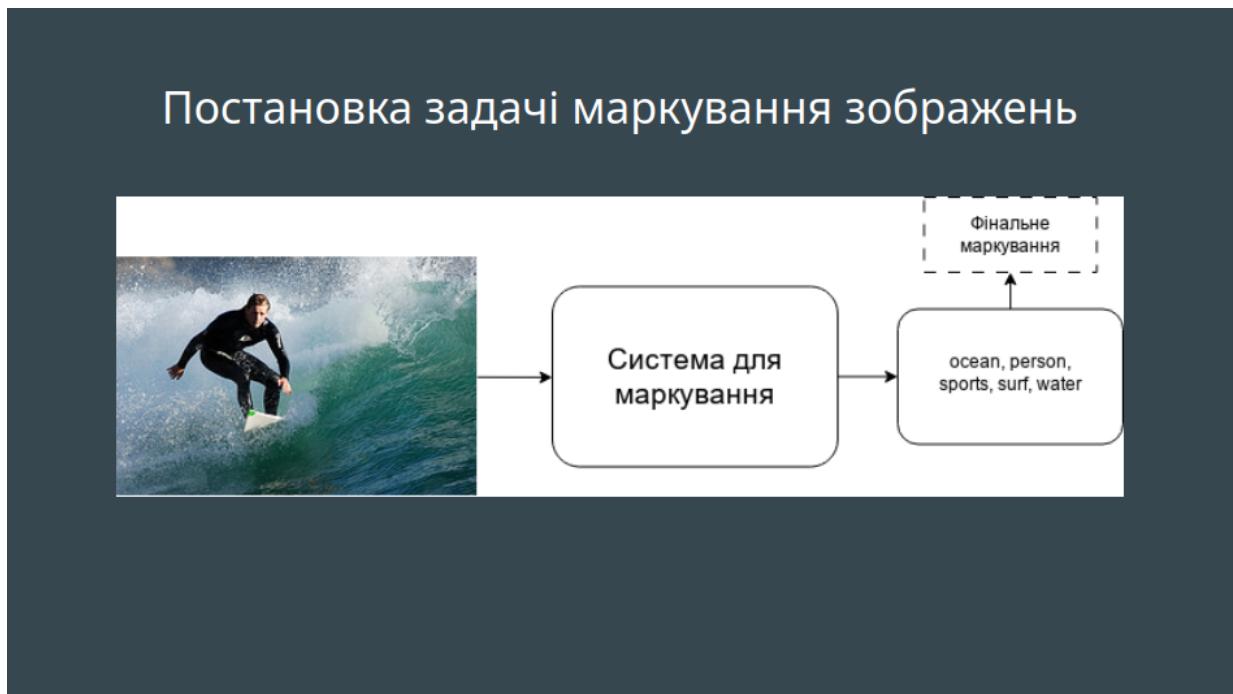


Рис. Г.4 – Слайд 4

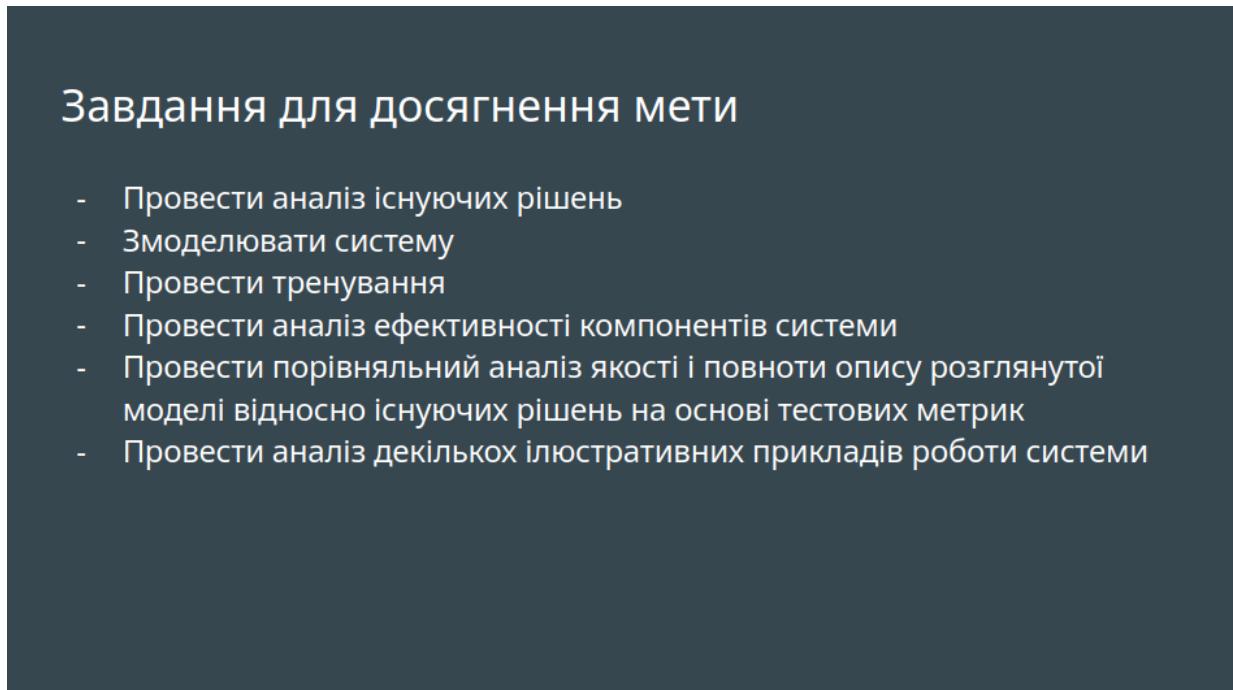


Рис. Г.5 – Слайд 5

Огляд існуючих рішень

Рис. Г.6 – Слайд 6

Базове рішення (аналіз зображення)

Ключовим аспектом задачі маркування є аналіз основної модальності даних - зображення. У абсолютній більшості робіт використовується модель CNN (Convolutional Neural Network). Щодо конкретних архітектур, то найчастіше використовується Resnet, більш рідкісним є використання ResNext, AlexNet та GoogleNet.

Найновіші підходи використовують ViT (Visual Transformer), які надають кращі результати, однак вимагають значно більше часу для тренування і потребують значного об'єму тренувальних даних.

Рис. Г.7 – Слайд 7

Додаткова модальність даних

Більш нові роботи також розглядають додаткові джерела інформації для підвищення якості маркування. Існує два основних підходи:

- 1) **Аналіз додаткової інформації.** Даний підхід аналізує додаткову до зображення інформацію. Це може бути як текстова інформація (теги/анотації), так і метадані, геолокації, тощо. Очевидним недоліком даного методу є потреба у цій додатковій інформації, яку можуть мати не всі зображення, а відсутність даної інформації знижує точність результатуючого маркування.
- 2) **Аналіз цільових класів.** На відміну від загальної інтерпретації класів для задачі маркування (коли кожен клас - незалежна сутність), даний підхід аналізує зв'язки між цільовими класами, створюючи нову модальність на основі набору цільових класів. Для цього зазвичай використовуються embeddings (та трансформери). Перевагою даного підходу є те, що йому не потрібні ніякі додаткові дані окрім зображень. Основним недоліком таких систем є висока складність, і як наслідок - значно довший процес тренування / розпізнавання.

Рис. Г.8 – Слайд 8

Кількість лейблів

Результатом роботи будь якої класифікаційної моделі є вектор ймовірностей, який репрезентує приналежність об'єкту до певних цільових класів. Для перетворення вектору в результат класифікації використовується індикаторна функція. Для задачі класифікації вибр результата на основі цього вектора очевидний - клас із найбільшою ймовірністю, однак для задачі маркування все складніше. Існує як мінімум дві основні індикаторні функції: top k та threshold a.

Рис. Г.9 – Слайд 9

Індикаторна функція top k				
Image				
Truth	clouds grass house sky	leaf plants sky	animal grass	person
Top 5 pred	clouds grass house road sky	clouds grass leaf plants sky	animal grass horses plants sky	animal clouds person road sky
Model pred	clouds grass sky	plants sky	animal grass horses	person

Топ k, або "найкращих k", обирає результат як k найбільш ймовірних класів у векторі ймовірностей.

Більшість робіт, які віднесено до множини "існуючих рішень" використовує індикаторну функцію top 3. Очевидно що така кількість лейблів не є оптимальною, так як деякі зображення містять більше 3 тегів, а деякі - менше.

Рис. Г.10 – Слайд 10

Індикаторна функція threshold a	
Threshold a, або порогове значення a, - це індикаторна функція яка маркує зображення за пороговим значенням, так для класу який має ймовірність більше ніж a маркування - позитивне, і навпаки. Перевагою даного методу є "вбудована" адаптивність до динамічної кількості лейблів, однак суттєвим недоліком є те, що вектор ймовірностей має бути досить сильно дискретизованим. Тобто для позитивних класів ймовірність має бути високою (0.6 - 1.0), а для негативних класів низькою (0.0 - 0.4).	

Рис. Г.11 – Слайд 11

Моделювання

Рис. Г.12 – Слайд 12

Архітектура композитної моделі

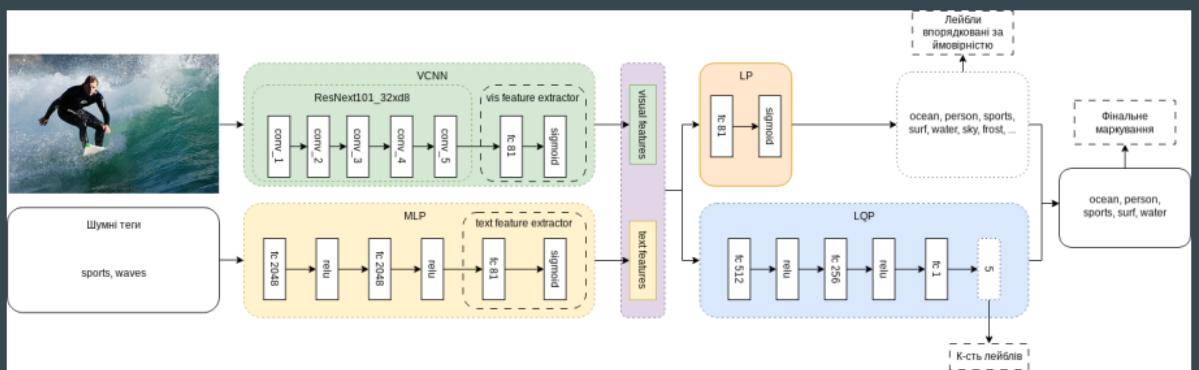


Рис. Г.13 – Слайд 13

Експерименти

Рис. Г.14 – Слайд 14

Датасет

Один із найбільш часто використовуваних датасетів для тестування моделей маркування зображень - NUS-WIDE, він складається із 269,655 зображень, 81 цільового класу (лейблу), та ≥ 5000 тегів. Для проведення тренування/тестування використовується розподіл, наведений авторами датасету, так як він є збалансованим відносно кількості лейблів.

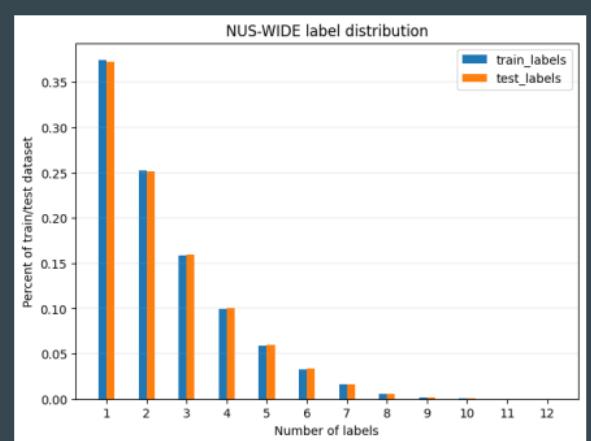


Рис. Г.15 – Слайд 15

Датасет. Підготовка

Обраний датасет містить посилання на зображення на ресурсі Flickr, і деякої частини цих зображень вже не існує. Додатково буде використано тільки 1000 найбільш частих тегів з 5000, при чому зображення які не містять жодного тега - відфільтровано.

	Тренування	Тестування
Кількість зображень	121962	81636
Середня к-сть лейблів	2.42	2.43
Медіана к-сть лейблів	2	2
Мінімальна к-сть лейблів	1	1
Максимальна к-сть лейблів	12	13

Табл. 4.1 – Характеристики тренувальної/тестової вибірок

Рис. Г.16 – Слайд 16

Датасет. Цільові класи



Рис. Г.17 – Слайд 17

Тестові метрики

$$\begin{aligned}
 \text{C-P} &= \frac{1}{C} \sum_{j=1}^C \frac{NI_j^c}{NI_j^p} & \text{O-P} &= \frac{\sum_{i=1}^N NL_i^c}{\sum_{i=1}^N NL_i^p} \\
 \text{C-R} &= \frac{1}{C} \sum_{j=1}^C \frac{NI_j^c}{NI_j^g} & \text{O-R} &= \frac{\sum_{i=1}^N NL_i^c}{\sum_{i=1}^N NL_i^g} \\
 \text{C-F1} &= \frac{2 \cdot \text{C-P} \cdot \text{C-R}}{\text{C-P} + \text{C-R}} & \text{O-F1} &= \frac{2 \cdot \text{O-P} \cdot \text{O-R}}{\text{O-P} + \text{O-R}}
 \end{aligned}$$

Рис. Г.18 – Слайд 18

Процес треування

Для тренування було використано графічний процесор "Nvidia L4".

Приблизний час тренування - 3 години. Для оптимізації тренування було використано техніку mixed precision.

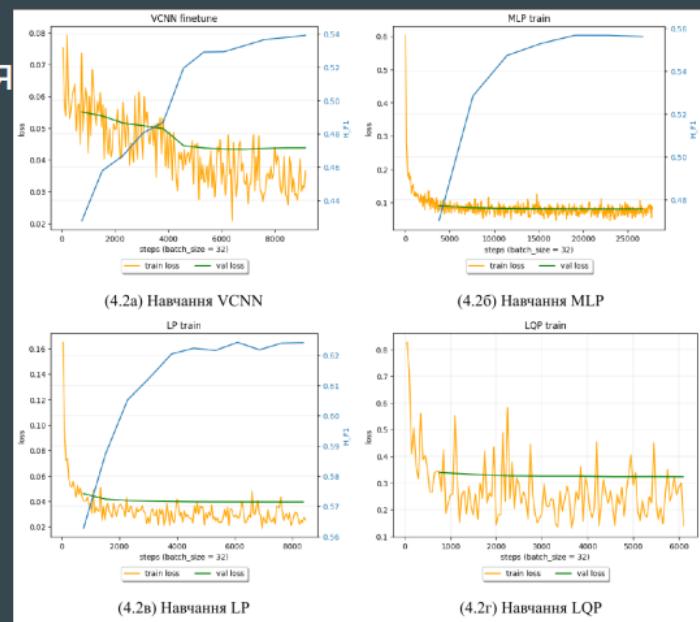


Рис. Г.19 – Слайд 19

Аналіз результатів

Рис. Г.20 – Слайд 20

Аналіз компонентів системи

Модель	Індикаторна ф-ція (3.6)	Модальність	C-P	C-R	C-F1	O-P	O-R	O-F1	H-F1
Композитна	top k	Зображення+теги	72.49	59.51	65.36	76.53	74.41	75.46	70.04
VCNN+MLP+LP	top 3	Зображення+теги	60.51	61.28	60.89	67.87	71.52	63.98	62.40
VCNN+LQP	top k	Зображення	64.62	37.61	47.54	77.07	59.01	66.84	55.56
VCNN	top 3	Зображення	44.48	53.32	48.50	55.44	68.52	61.29	54.15

Табл. 5.1 – Порівняння компонентів моделі

Рис. Г.21 – Слайд 21

Порівняння з існуючими рішеннями

Модель	Індикаторна ф-ція (3.6)	Модальност	C-P	C-R	C-F1	O-P	O-R	O-F1	H-F1
Композитна	top k	Зображення+теги	72.49	59.51	65.36	76.53	74.41	75.56	70.04
Query2Label [11]	threshold α	Зображення (+аналіз класів)	-	-	67.60	-	-	76.3	71.69
SR-CNN-RNN [10]	top 3	Зображення+теги	71.73	61.73	66.36	77.41	76.88	77.15	71.35
Resnet-CPSD [20]	threshold α	Зображення (+аналіз класів)	-	-	64.00	-	-	75.30	69.19
MS-CMA [21]	threshold α	Зображення (+аналіз класів)	-	-	60.50	-	-	73.80	66.49
Resnet-SRN [22]	threshold 0.5	Зображення (+аналіз класів)	65.20	55.80	58.50	75.50	71.50	73.40	65.10
SINN [7]	top 3	Зображення+теги	58.30	60.63	59.44	57.05	79.12	66.29	62.68
TagNeighbour [8]	top 3	Зображення+метадані	54.74	57.30	55.99	53.46	75.10	62.46	59.05
CNN+Logistic [7]	top 3	Зображення	45.60	45.03	45.31	51.32	70.77	59.50	51.44
CNN-RNN [18]	top 3	Зображення	40.50	30.40	34.70	49.9	61.70	55.20	42.61
CNN+WARP [5]	top 3	Зображення	31.65	35.60	33.51	48.59	60.49	53.89	41.32
CNN+Softmax [5]	top 3	Зображення	31.68	31.22	31.45	47.82	59.52	53.03	39.48

Табл. 5.2 – Порівняння результатуючих метрик для різних моделей на датасеті
NUS-WIDE

Рис. Г.22 – Слайд 22

Демонстаривний приклад №1

Image	Truth	Model pred
	animal birds sky	animal birds sky
	person wedding	person wedding
	buildings	sky
	animal	animal birds

Рис. Г.23 – Слайд 23

Демонстаривний приклад №2

Image	Truth	Model pred
	sky	animal clouds horses sky
	train	railroad train
	buildings clouds sky window	buildings clouds sky town window
	clouds grass sky vehicle	clouds sky vehicle window

Рис. Г.24 – Слайд 24

Демонстаривний приклад №3

Image	Truth	Model pred
	sky	grass sky
	road train window	sky train
	animal grass	animal grass
	water	person

Рис. Г.25 – Слайд 25

Демонстаривний приклад №4

Image	Truth	Model pred
	person	person window
	grass road vehicle	road vehicle
	lake ocean water	beach lake ocean water
	flowers plants	flowers

Рис. Г.26 – Слайд 26

Висновки

Рис. Г.27 – Слайд 27

Результатом проходження переддипломної практики є сформований шаблон БАР за заданою темою.

Досягнуто мету - розроблено ПЗ для маркування зображень (шпалерів робочого столу), а саме спроектовано та натреновано композитну нейронну мережу для автоматичного маркування зображень на основі даних із двома модальностями (зображення і теги), що дозволить покращити категоріальний пошук при впровадженні даної системи.

Після завершення процесу тренування було досягнуто точності співставної із існуючими рішеннями по якості маркування, при меншій складності результируючої системи, що внаслідок зменшує вимоги як до ресурсів необхідних, щоб натренувати модель, так і до ресурсів необхідних для запуску моделі.

Проведено порівняльний аналіз компонентів системи, та доведено ефективність кожного з її компонентів. Також проведено порівняння із існуючими рішеннями, що дає можливість зробити висновок, що розроблена система є ефективним рішенням для маркування зображень.

Бакалаврська атестаційна робота готова на 90-95% до захисту та потребує деяких покращень. Основною масою правок буде розширення змісту у деяких розділах, доповнення опису датасету за допомогою додаткових графіків та інші незначні правки.

Рис. Г.28 – Слайд 28