

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет прикладної математики

Кафедра прикладної математики

ЗВІТ З ПЕРЕДДИПЛОМНОЇ ПРАКТИКИ

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. _____

Прийнято з оцінкою _____ балів _____

Члени комісії:

ПБ _____

ПБ _____

ПБ _____

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

Студент _____

Київ --- 2024

ВСТУП ДО ЗВІТУ

В сучасному світі технологій розвиваються все швидше і швидше, додаючи до цього збільшення кількості людей отримаємо стрімкий ріст у кількості даних, яку генерує людство на добу. Зображення є одним із підтипів цих даних, воно вимагає певних додаткових описових даних для пошуку якоїсь підмножини серед сотень мільйонів інших зображень. В якості описових даних можуть бути лейбли, саме тому важливо мати метод для автоматичного присвоювання цих лейблів кожному зображенню.

Існуючі рішення зазвичай використовують нейронні мережі, які можна розділити на дві групи:

1) Прості нейронні мережі. Такі рішення, зазвичай, не дуже вибагливі до системни, на якій вони працюють, проводять маркування швидко та досить таки легко натренувати. Основним недоліком таких рішень є низька якість опису зображення.

2) Складні нейронні мережі. Такі рішення, зазвичай, навпаки, - мають високу якість опису зображення, однак вибагливі до системи на якій працюють та потребують значно більше часу на тренування.

Дана робота націлена на вирішення цих трьох проблем, а саме створення системи автоматичного маркування зображення, яка має просту архітектуру, проста в тренуванні та якісно маркує зображення.

В якості предмету дослідження було обрано "шпалери робочого столу", які є підмножиною зображень. Однак, варто зазначити, що предметна область задається насамперед множиною цільових класів (лейблів), тому дане рішення, гіпотетично, можна буде адаптувати до маркування інших видів зображень, зокрема:

рентгенівських знімків, супутникових знімків, зображень зегенерованих ШІ та ін.

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 зображенню лейбли, 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.

Огляд існуючих рішень задачі маркування зображення. В даному розділі було розглянуто описано ключові проблеми, які виникають при створенні системи автоматичного маркування зображень: аналіз зображення як основної модальності даних, аналіз додаткової модальності даних, кількість лейблів у результуючому маркуванні. На основі цих проблем було розглянуто існуючі рішення задачі маркування зображень. В якості існуючих рішень, було розглянуто 11 систем, які

були створені у різних роках (2013-2024). Основним критерієм для відбору саме цих рішень був датасет, так як для порівняння якості маркування необхідно, щоб тренування / тестування моделі відбувалось на одному і тому ж датасеті. На основі даного аналізу в подальшому було проведено моделювання системи маркування зображень.

Моделювання. На основі огляду існуючих рішень було обрано найоптимальніші змодельовано систему для вирішення відповідних проблем та описано їх архітектуру.

Для аналізу зображень було обрано нейронну мережу ResNext101 [resnext], яка є більш сучасною версією Resnet [resnet], яку використовували у більшості інших робіт.

В якості додаткової модальності було вирішено обрати додаткову інформацію у вигляді тегів до зображення. Теги надаються людьми, і містять шумну інформацію. В даній роботі 'шумні теги' - це інформація, яка має опосередковане відношення або взагалі не мати ніякого відношення до зображення. Наприклад: для зображення кота це може бути: ['Канада', 'Левеня', 'Пухнастик']. Використання тегів дає змогу покращити опис зображення, так як, зазвичай, вони описують у собі екстраординарні об'єкти на зображенні, на відміну від тієї ж моделі для аналізу зображень, яка, зазвичай, 'бачить' більш загальні об'єкти ('людина', 'вода', 'небо' і тд.). Для аналізу тегів використовується звичайний багатосаровий перцептрон.

Для обрання кількості лейблів використовується метод top k, де k - це результат роботи регресійної нейронної мережі на основі багатосарового перцептрону.

Також у даному розділі описано особливості тренування / тестування даної системи.

Експерименти. В даному розділі розглянуто 3 пункти: датасет, тестові метрики, процес тренування.

В якості датасету для тренування / тестування було обрано NUS-WIDE [nus-wide-civr09]. Даний датасет було сформовано у 2008 році на основі одного із найпопулярніших хостингів зображень на той момент - Flickr.

Для оцінки якості маркування використовуються метрики precision та recall які рахуються для зображень (overall) та для окремих класів (per-class), які були запропоновано у одній із перших робіт [cnn-labeling]. Наступні роботи розраховують метрики F1 на основі precision та recall (C-F1, O-F1). В цій роботі також розглянуто метрику H-F1, яка обчислюється так само як і канонічна метрика F1 (середнє гармонічне), однак вже на основі C-F1 та O-F1.

Також в даному розділі наводяться певні прикладні аспекти тренування: графічний процесор, метод тренування, необхідний на тренування час.

Аналіз результатів. В цьому розділі розглядається 3 пункти: аналіз компонентів, порівняння ефективності системи із існуючими рішеннями, аналіз демонстративного прикладу роботи запропонованої системи.

Аналіз компонентів розглядає чотири конфігурації композитної моделі:

- 1) Аналіз виключно зображень (top 3)
- 2) Аналіз виключно зображень (top k)
- 3) Аналіз зображень та тегів (top 3)
- 4) Аналіз зображень та тегів (top k)

На основі тестових метрик було проаналізовано вплив кожної конфігурації, і, відповідно, кожного із компонентів системи.

Схожим чином було проведено порівняння композитної системи із існуючими рішеннями.

ФОРМАЛІЗАЦІЯ ПОСТАНОВКИ ЗАДАЧІ

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 зображенню лейбли, 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 приклади 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 [**sr-cnn-rnn**], Resnet-SRN [**srn**], MS-CMA [**cma**], Query2Label [**q2l**], Resnet-CPSD [**cpsd**] 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.

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ 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

«КИЇВСЬКИЙ 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 існуючими 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 ефективності 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 зображенню лейбли, 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 приклади 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 [**sr-cnn-rnn**], Resnet-SRN [**srn**], MS-CMA [**cma**], Query2Label [**q2l**], Resnet-CPSD [**cpsd**] 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, мультимодальні дані, багатошарові перцептрони, 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, speccifically 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 [**sr-cnn-rnn**], Resnet-SRN [**srn**], MS-CMA [**cma**], Query2Label [**q2l**], Resnet-CPSD [**cpsd**] 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, multimodal data, multilayer perceptrons, composition of neural networks.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

- * Модель - нейронна мережа
- * Система (система для маркування зображень) - композиція моделей
- * 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)
- * Тег (Tag) - шумна інформація надана користувачем у формі тексту (наприклад для зображення кота "Cat, Canada, Cola")
- * Лейбл (Label) - синонім слова ground truth для класифікації

3MICT

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 [**dnn-cls**] та її подальший розвитком у CNN [**cnn-cls-1**, **cnn-cls-2**] разом із створенням великих датасетів як-от ImageNet [**deng2009imagenet**] дало змогу вирішувати задачу класифікації зображень значно швидше і якісніше ніж люди.

Зрозуміло, що зображення - це той тип даних, який у абсолютній 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 лиш 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 [**deng2009imagenet**], MS-COCO [**cocodataset**], NUS-WIDE [**nus-wide-civr09**], 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 дисбаланс 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 кількість 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.

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 [resnet], AlexNet [alexnet], GoogleNet [googlenet], ResNext [resnext]. Спільним між ними є те що вони вже натреновані на великому датасеті, здебільшого ImageNet [deng2009imagenet]. Для адаптації моделі до обраного контексту така модель дотреновується (fine tune), замінюючи існуючий класифікатор на класифікатор із адаптованою к-стю вихідних класів [cnn-labeling]. Це працює завдяки тому, що всі архітектури сучасних CNN моделей є багат шаровими, і перші шари розпізнають базові особливості (features) зображення, а останні шари вивчають більш глибокі особливості зображення, таким чином дозволяючи адаптовувати модель під різні задачі після проведення підгонки (finetune).

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) [cnn-sinn, sr-cnn-rnn], так і метадані зображення [cnn-neighbors, cnn-location]. 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 [srn]. Більш новим та узагальненим підходом є технологія word2vec (або аналогічне рішення), яка дозволяє, проаналізувавши набір слів, впорядкувати його у векторному просторі за семантичним значенням і побудувати граф, в якому вказуються класи (поняття) які часто знаходяться на одному зображенні (co-occurrence; наприклад: риба, вода) та класи які рідко знаходяться на одному зображенні (dis-occurrence; наприклад: риба, пустеля) [q2l, cpsd, cma]. Перевагою даного підходу є те, що йому не потрібні ніякі нові дані крім зображень та відповідних їм цільових класів, а нова модальність, 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.

Кількість лейблів

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. Більшість наведених вище робіт розглядають задачу вибору k -сті лейблів як найкращі k (top k) маркувань (Розділ ??), 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 спільного із цим зображенням (Рис. ??)





| | | | | |
|------------|--|--|---|--|
| 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. 2.1 – 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

Один із підходів як-от CNN-RNN [cnn-rnn], 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.

Найновіші моделі [**q2l**, **cpsd**, **cma**], які розглядають зв'язок між цільовими класами, обирають к-сть цільових класів на основі порогового значення (Розділ ??). Варто зазначити, що даний підхід є ефективним 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

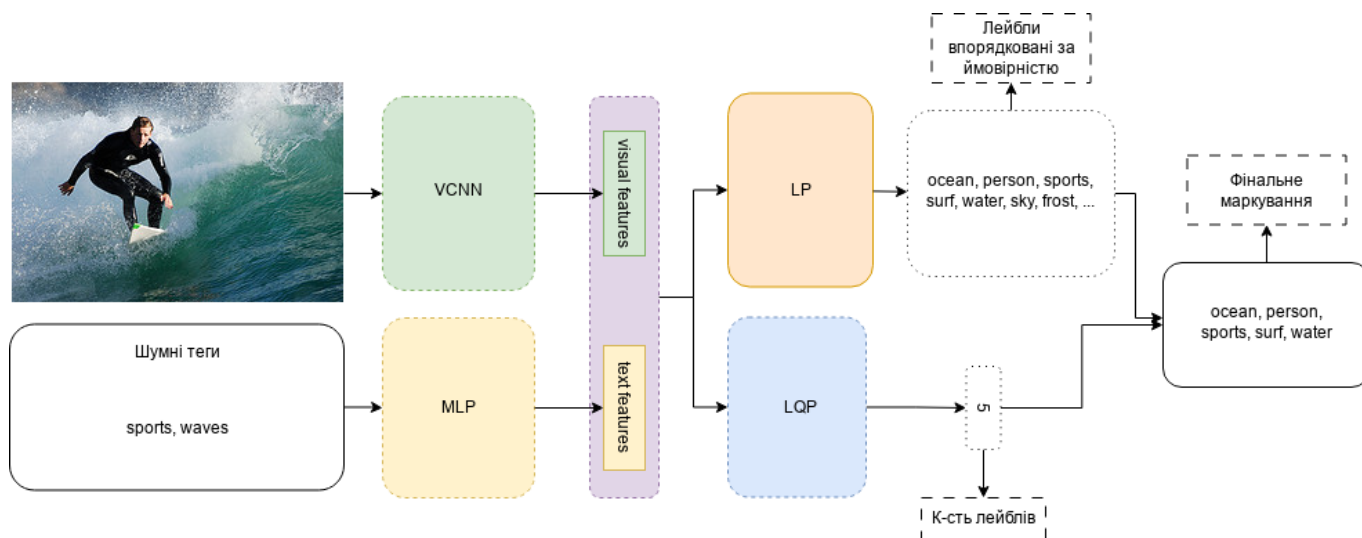


Рис. 3.1 – Архітектура композитної системи

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

3.1 Модель VCNN

Модель VCNN (Рис. ??) призначена для вивчення особливостей (features) із зображення. Отримує на вхід піселі 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 [resnext] (сучасна версія resnet), із адаптованим класифікатором (vis feature extractor), натреновану на датасеті

ImageNet [deng2009imagenet].

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

MLP (Рис. ??) - 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").

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

LP (Рис. ??) - 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

Модель LQP (Рис. ??) 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 регуляризацию. В даній роботі, в якості регуляризатора задіяні Dropout шари [**dropout**], із вірогідністю відкидання (dropout rate) 0.5.

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

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).

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 $i = 1...N$, $j = 1...C$, 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 анотоване класом j , інакше - $y_{ij} = 0$, а $\sigma(\cdot)$ - це сигмоїдальна 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 .

Тренування VCNN

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

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

Існує два підходи для підгонки:

1) Підгонка всієї моделі: 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: 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.

Тренування MLP

Дана модель є звичайним багатошаровим перцептроном, тому її тренування досить таки прямолінійне.

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.

Тренування LQP

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

3.6 Процес тестування

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 результуючим вектором маркувань M : $k = 3$, перетворення $P \rightarrow M \equiv [0.8, 0.9, 0.1, 0.3, 0.2] \rightarrow [1, 1, 0, 1, 0]$

Тестування композитної моделі

Для тестування композитної моделі (Рис. ??) необхідно обрахувати результуючі 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 [nus-wide-civr09], 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 збалансовний настільки, наскільки це можливо (Рис. ??).

Рис. 4.1 – Розподіл лейблів у тренувальному/тестовому датасеті

Важливо відмітити, що даний датасет містить посилання на зображення на ресурсі Flickr, і деякої частина цих зображень вже там немає. Також буде використано тільки 1000 найбільш частих тегів з ≈ 5000 , при чому зображення, які не містять жодного тега відфільтровано.

Далі наведено деякі числові характеристики тренувальної/тестової вибірок:

| | 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

4.2 Метрики

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

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

,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 [cnn-labeling].

Для того щоб отримати уніфіроване 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 [resnext] натренована на датасеті ImageNet [deng2009imagenet], 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 [deng2009imagenet].

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 навчались із швидкістю навчання (learning rate) 0.001, а LP - зі швидкістю 0.01. Також для навчання 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 l1 регуляризації (weight decay) 0.0003.

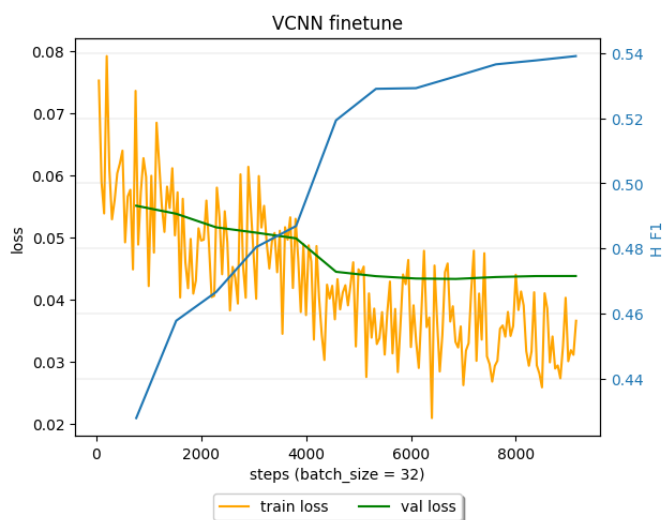
Розмір групи (batch size) 32.

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

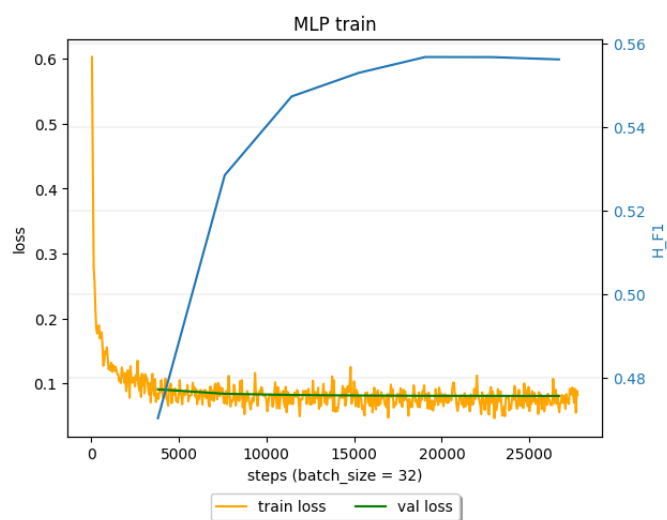
Процес навчання

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

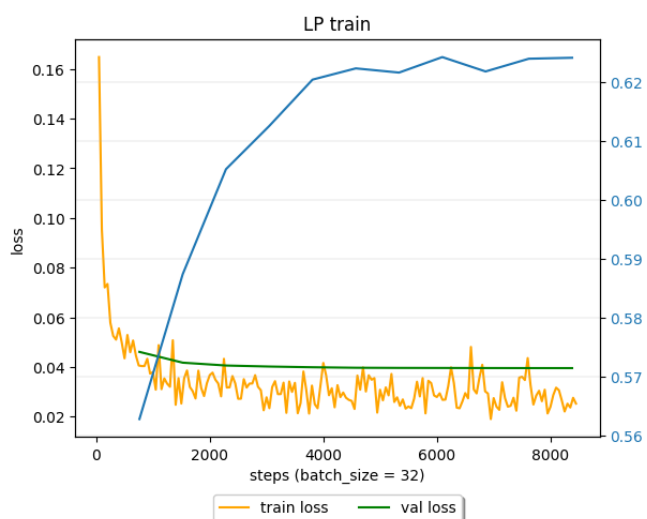
Для оптимізації процесу тренування було застосовано техніку mixed precision, яка використовує f16 замість f32, під час певних етапів тренування [**mixed-precision**].



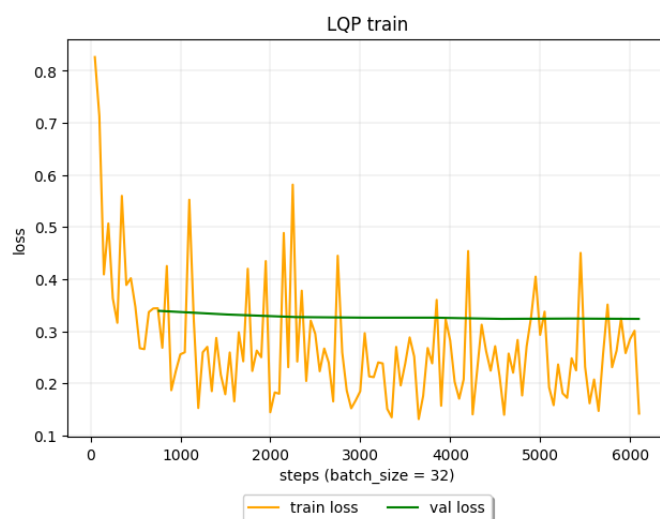
(4.2a) Навчання VCNN



(4.2б) Навчання MLP



(4.2в) Навчання LP



(4.2г) Навчання LQP

Рис. 4.2 – Процес навчання моделей

5 АНАЛІЗ РЕЗУЛЬТАТІВ

5.1 Аналіз компонентів системи

| Модель | Індикаторна ф-ція (??) | 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

Результуюча 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 лиш зображення.

Передбачення кількості лейблів (LQP)

При використанні компоненту LQP кількість лейблів обирається за принципом 'top k', а не 'top 3' (??). 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 дисбаланс класів у датасеті, використання принципу 'top k' збільшує точність (precision), сильно жертвуючи по-класовим охопленням (C-R), і, як наслідок, не сильно збільшуючи 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 Порівняння з існуючими рішеннями

| Модель | Індикаторна ф-ція (??) | 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 [q2l] | threshold α | Зображення (+аналіз класів) | - | - | 67.60 | - | - | 76.3 | 71.69 |
| SR-CNN-RNN [sr-cnn-rnn] | top 3 | Зображення+теги | 71.73 | 61.73 | 66.36 | 77.41 | 76.88 | 77.15 | 71.35 |
| Resnet-CPSD [cpsd] | threshold α | Зображення (+аналіз класів) | - | - | 64.00 | - | - | 75.30 | 69.19 |
| MS-CMA [cma] | threshold α | Зображення (+аналіз класів) | - | - | 60.50 | - | - | 73.80 | 66.49 |
| Resnet-SRN [srn] | threshold 0.5 | Зображення (+аналіз класів) | 65.20 | 55.80 | 58.50 | 75.50 | 71.50 | 73.40 | 65.10 |
| SINN [cnn-sinn] | top 3 | Зображення+теги | 58.30 | 60.63 | 59.44 | 57.05 | 79.12 | 66.29 | 62.68 |
| TagNeighbour [cnn-neighbors] | top 3 | Зображення+метадані | 54.74 | 57.30 | 55.99 | 53.46 | 75.10 | 62.46 | 59.05 |
| CNN+Logistic [cnn-sinn] | top 3 | Зображення | 45.60 | 45.03 | 45.31 | 51.32 | 70.77 | 59.50 | 51.44 |
| CNN-RNN [cnn-rnn] | top 3 | Зображення | 40.50 | 30.40 | 34.70 | 49.9 | 61.70 | 55.20 | 42.61 |
| CNN+WARP [cnn-labeling] | top 3 | Зображення | 31.65 | 35.60 | 33.51 | 48.59 | 60.49 | 53.89 | 41.32 |
| CNN+Softmax [cnn-labeling] | top 3 | Зображення | 31.68 | 31.22 | 31.45 | 47.82 | 59.52 | 53.03 | 39.48 |

Табл. 5.2 – Порівняння результируючих 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. ???. 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 [q2l], Resnet-CPSD [cpsd] та MS-CMA [cma]) не потребує ніяких додакових даних окрім зображення, що є вагомою конкурентною перевагою, враховуючи незначну відмінність в точності моделей.

Індикаторна функція Для оцінки ефективності запропонованої індикаторної функції 'top k', варто ізолювати вплив саме цієї фнкції. Для цього розглянемо існуючі моделі, які працюють із тією ж модальністю даних. Найкращою із таких моделей є SINN [cnn-sinn]. 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. ??), що є вагомим приростом.

5.3 Демонстративні приклади

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















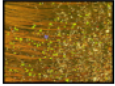

| Image | Truth | Model pred | Image | Truth | Model pred |
|---|---|---|--|---|---|
|  | clouds sky | clouds sky |  | snow water | glacier lake water |
|  | flowers plants | flowers plants |  | clouds mountain rocks sky | mountain rocks sky |
|  | cityscape clouds sky | buildings clouds nighttime sky |  | 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 | buildings clouds nighttime sky |
|  | flowers grass water | grass plants |  | flowers garden grass plants water | flowers garden grass plants |
| (5.1в) | | | (5.1г) | | |

Рис. 5.1 – Демонстративні приклади

Задамо умовне позначення "а::1", що означає демонстративний приклад "а", 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 зображенні. Наприклад: [а::3,а::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 зображенні. 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 [cocodataset]). 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.

Результуюча 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 супутникових знімків - різні метадані, геолокація тощо, а для аналізу звичайних фотографій - теги, анотації, метадані, тощо.

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 'lktags.py' призначенні для обробки сирих 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




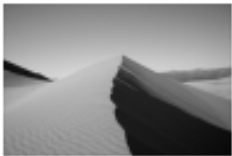
| Image | Truth | Model pred |
|---|--------------------------------|-------------------------|
|  | snow | snow |
|  | animal cat | animal cat |
|  | animal elk lake water | animal lake water |
|  | road sand | sand sky |

Рис. В.6





| Image | Truth | Model pred |
|---|---|--|
|  | plane clouds | plane sky |
|  | 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



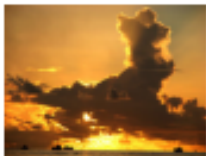
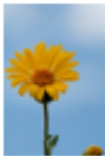
| Image | Truth | Model pred |
|---|--|---|
|  | beach clouds lake sky sunset beach water | clouds ocean sky sunset water clouds |
|  | clouds ocean sky sunset water | lake ocean sky sunset water |
|  | clouds sky sunset | clouds fire sky sunset |
|  | flowers sky | flowers sky |

Рис. В.12

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

Рис. В.13

Додаток Г

Ілюстративний матеріал

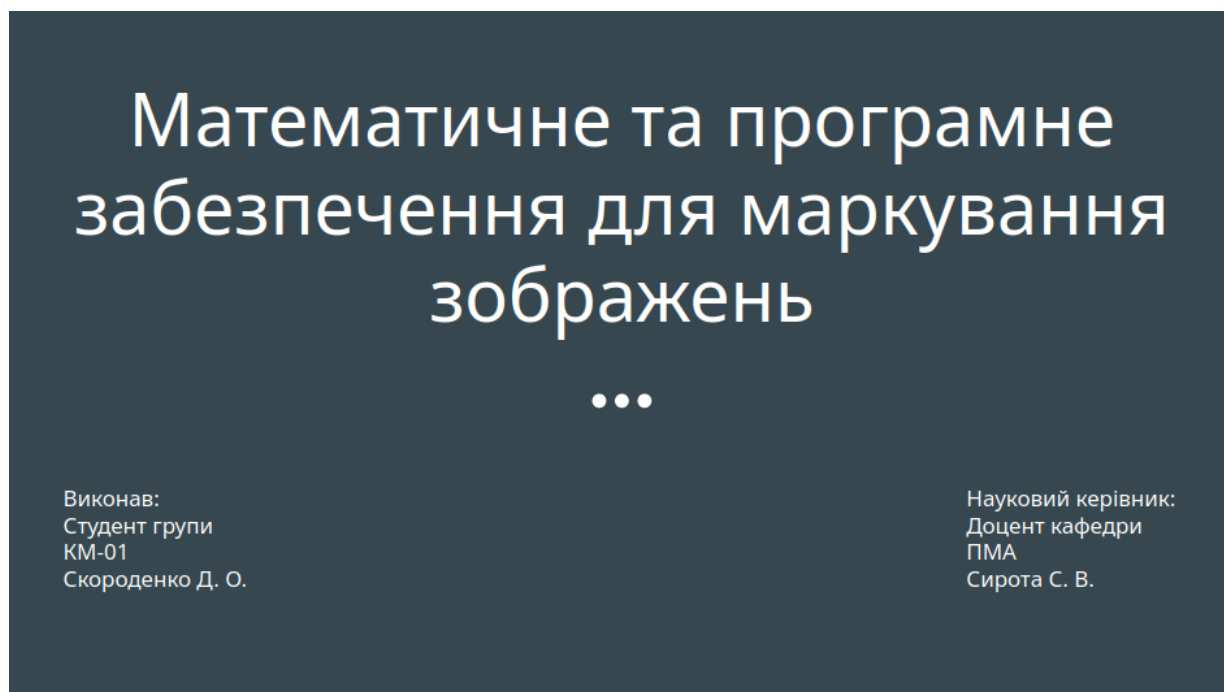


Рис. Г.1 – Слайд 1

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 лейбли, а для автоматизації процесу присвоєння лейблів зображенню потрібно використовувати систему для маркування.

Існуючі рішення задачі маркування зображення поділяються на дві категорії:

- Прості нейронні мережі. Такі рішення не дуже вибагливі до обчислювальних потужностей, однак не надають високої якості маркування.
- Складні нейронні мережі (композиція нейронних мереж). Такі рішення вибагливі 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 обрані 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

Постановка задачі маркування зображень



Рис. Г.4 – Слайд 4

Завдання для досягнення мети

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

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 потребує 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 результуючого 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 вибір результату 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 α .

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 |

Тор 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 тор 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 α

Threshold α , або порогове значення α , - це індикаторна функція яка маркує зображення за пороговим значенням, так для класу який має ймовірність більше ніж α 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

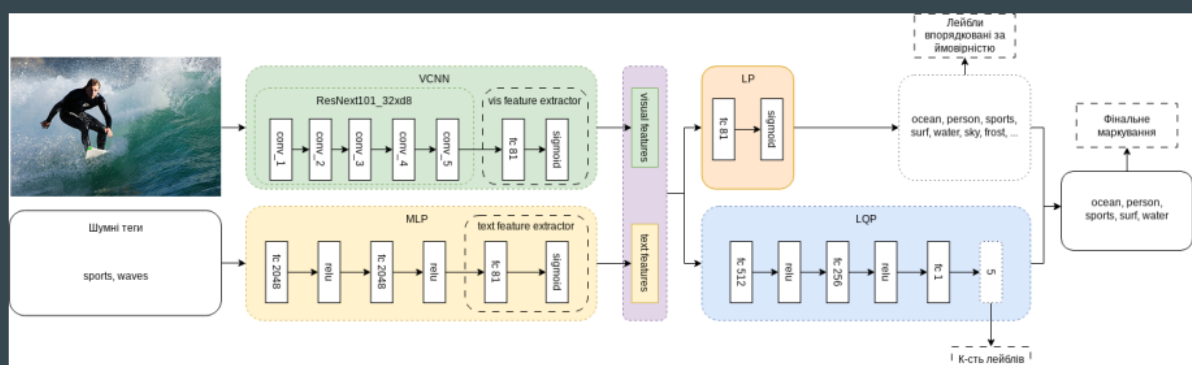


Рис. Г.13 – Слайд 13

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.

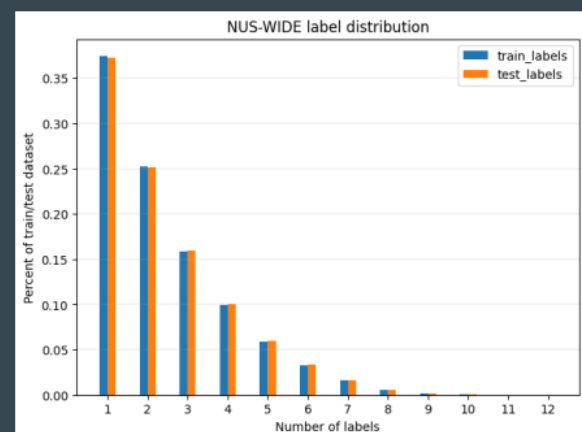


Рис. Г.15 – Слайд 15

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 тренування

Для тренування було використано графічний процесор "Nvidia L4". Приблизний час тренування - 3 години. Для оптимізації тренування було використано техніку mixed precision.

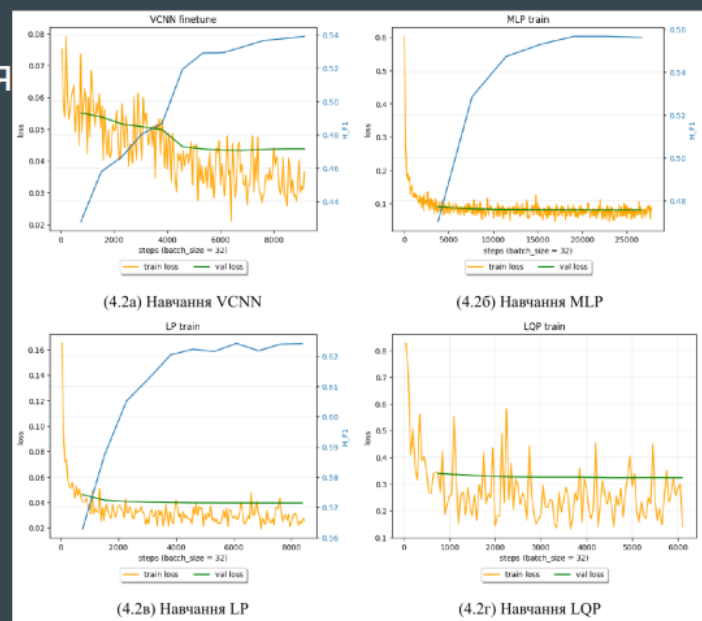


Рис. Г.19 – Слайд 19

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) | 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 [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 результуючої 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

Висновки до звіту

Результатом проходження переддипломної практики є сформований шаблон бакалаврської атестаційної роботи на тему «Автоматичне маркування зображень за допомогою нейронних мереж».

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 зображень при впровадженні даної системи.

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

Дана композитна нейронна мережа складається із однієї моделі типу CNN, двох класифікаційних перцептронів та одного регресійного перцептрону.

Після завершення процесу тренування було досягнуто точності співставної із існуючими рішеннями по якості маркування, при меншій складності результуючої 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.