Багатоміткова класифікація стилізованих зображень за допомогою глибоких нейронних мереж

Виконав:
Студент IV курсу Тюкалов Н. С. групи КА-13.
Керівник:
Асистент Древаль М.М.

Актуальність

Комп'ютерний зір (CV) є одним з найбільших напрямків машинного навчання, який знайшов застосування в різноманітності індустрій та сфер.

Незалежно від сфери, головним типом задачі комп'ютерного зору є класифікація та її варіації. Дуже реальною проблемою є необхідність виявити наявність великої кількості класів або об'єктів одночасно. Ця задача називається **багатомітковою класифікацією**.

Багатоміткова класифікація тегів зображень в аніме-стилістиці є проблемою з унікальними труднощами та викликами. Модель з подібною специфікою може бути інтегрована в image-board платформи для сортування, пошуку та модерації контенту.

Об’єкт дослідження

Об’єкт дослідження – великий незбалансований датасет стилізованих аніме-зображень, створений на основі даних з відкритого ресурсу Danbooru, що характеризується детальною комплексною системою тегів.

Предмет дослідження

Предмет дослідження – архітектури глибоких нейронних мереж, зокрема Attention-механізми, методи багатоміткової класифікації, а також стратегії роботи з комплексними та проблемними даними.

Мета дослідження

Мета дослідження – адекватна модель багатоміткової класифікації стилізованих аніме-зображень.

Постановка задачі

Постановка задачі – спроектувати та натренувати модель багатоміткової класифікації аніме-зображень з використанням сучасних архітектур нейронних мереж, яка здатна адекватно класифікувати найпопулярніші теги.

Вхідні дані

В якості набору даних в даній роботі використовується датасет на основі архиву аніме-зображень Danbooru.

Danbooru це image-board платформа яка існує з 2006 року, і спеціалізується на аніме контенті. За оцінкою цей ресурс має більше 7 мільйонів зображень.

Для категоризації зображень використовується комплексна система тегів. Кількість тегів оцінити складно, але кількість “справжніх” тегів вимірюється щонайменше в тисячах.

Безпосередньо для моделі використовувався піднабір з 25к зображень та топ-100 тегів, в зв’язку з обмеженими обчислювальними ресурсами.

Структура даних

Кожне зображення детальне описано та розмічено з використанням тегів, які розділені на 5 головних категорії:

—художник;

—авторські права;

—зображений персонаж;

—загальні описові теги (наприклад “1girl” та “solo”);

—мета теги (наприклад “highres”, якщо зображення високої якості).

Приклади оброблених даних з датасету

Tags: 1girl, black_hair, collarbone, green_eyes, long_hair, solo

Tags: black_hair, blonde_hair, blue_eyes, gloves, jewelry, multiple_boys

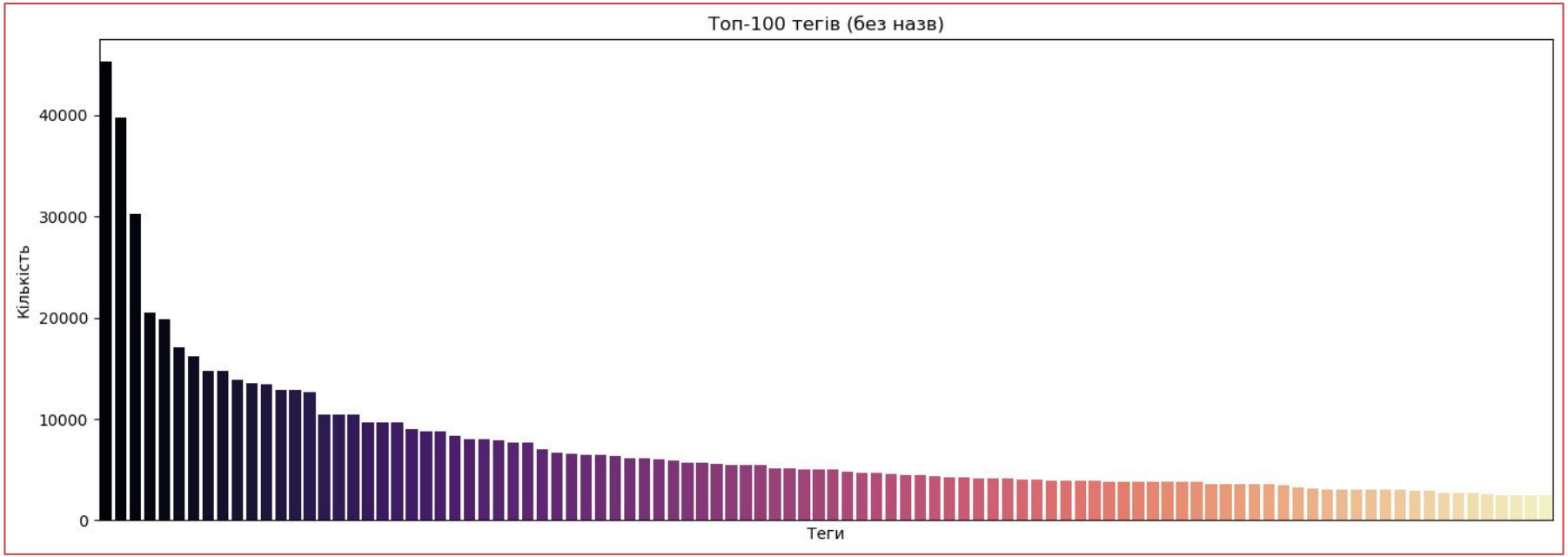
Проблемність даних

Дані характеризуються наступними проблемами:

1) перетини тегів;
2) абстрактність тегів;
3) псевдоієрархічна структура;
4) критичний дисбаланс класів;
5) відсутність тегів.

Також існує ряд проблем, який пов’язаний з самою стилістикою, наприклад невиражений гендерний диморфізм.

Ілюстрація дисбалансу популярності тегів



Архітектура моделі

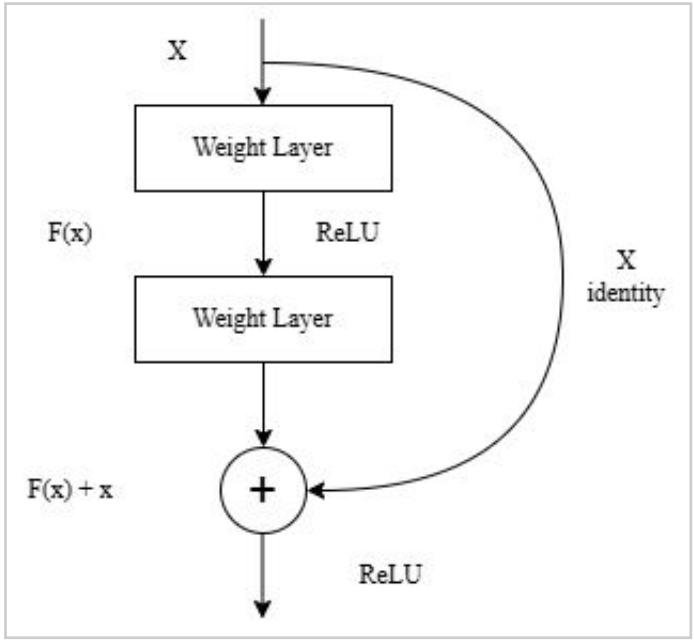
1. Вхідний шар (Input): 256x256, 3 канали.
2. Початковий CNN блок: згортковий шар, batch normalization, ReLu.
3. MaxPooling2D.
4. ResNet: Чотири ResNet блока з послідовним збільшенням кількості фільтрів. В кожному ResNet блоці також присутній SE блок.
5. GlobalAveragePooling2D
6. Dropout
7. Вихідний шар: Dense на 100 нейронів, Sigmoid.

ResNet

ResNet це тип архітектури NN.

ResNet складається з блоків залишку (Residual block).

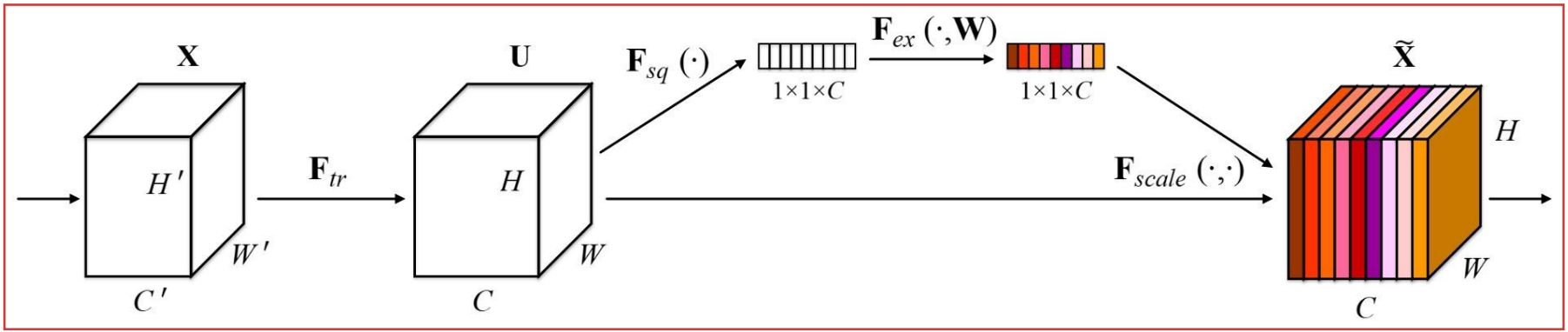
Основна ідея ResNet архітектури в тому, щоб замість тренування вирішення поставленої задачі напряму, мережа вчиться “доповненню до задачи”, що є простішим завданням.



Squeeze-and-Excitation

Squeeze-and-Excitation (SE) блок є простим механізмом уваги, який динамічно зважує важливість каналів у згорткових шарах.

SE складається з двох етапів: **Squeeze** (видобування ознак з GAP) та **Excitation** (ранжування ознак Dense шарами та Sigmoid виходом).



Focal Loss

Focal Loss – це функція втрат, яка фокусується на складних прикладах зменшенням ваги для прикладів, які класифікуються успішно, та збільшенням ваги для прикладів, які класифікуються менш успішно.

$$L_{Focal}(y, \hat{y}) = -\alpha_t(1 - \hat{p}_t)^{\gamma} log(\hat{p}_t)$$

Приклади результату роботи

Image: 1367003.jpg

True tags: 1girl, comic, greyscale, monochrome
Predicted tags: 1girl, comic, greyscale, long_hair, monochrome

Image: 306008.jpg

True tags: 1girl, bare_shoulders, blonde_hair, blue_hair, blush, closed_eyes, dress, short_hair, solo
Predicted tags: 1girl, blush, long_hair, short_hair, smile, solo

Вихідні метрики

Для порівняння взята незалежна модель, яка натренована, по суті, на цьому ж датасеті.

Модель	Кількість зображень	Кількість тегів	Кількість параметрів	F1-micro	F1-macro	Розмір
Власна	25к	100	17м	0.3778	0.1865	200 мб
Camie Tagger	7м	70к	214м	0.576	0.204	850 мб

Висновки

У рамках даної дипломної роботи було успішно досліджено проблему багатоміткової класифікації стилізованих зображень із застосуванням глибоких нейронних мереж.

Була спроектована, побудована та натренована модель багатоміткової класифікації на основі ResNet-архітектури, посиленої блоками Squeeze-and-Excitation.

Було досягнуто адекватних метрик оцінки моделі.

Подальші дослідження

В рамках подальшого дослідження пропонується:

1) використання більшої кількості даних;
2) розширення кількості тегів;
3) пошук оптимального методу обробки даних;
4) використати новітні архітектури, таких як ConvNext;
5) застосування візуальних трансформерів;
6) імплементація оптимізатора GSAM;
7) продовжити пошук оптимальної функції втрат.

Використані технології та інструменти

1. Середовище Google Collab .
2. Фреймворк TensorFlow.
3. Стандартні Python бібліотеки.

Дякую за увагу!