УДК 004.93

АНАЛІЗ МОДЕЛІ OPENAI CLIP ДЛЯ ЗАДАЧІ ПОШУКУ ЗОБРАЖЕНЬ НА ПЕРСОНАЛЬНОМУ КОМП'ЮТЕРІ ЗА ТЕКСТОВИМ ЗАПИТОМ

Кошель В.О.
e-mail: vladyslav.koshel@nure.ua
Науковий керівник – к.т.н., доц. Яковлева О.В.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

This study examines the use of the OpenAI CLIP model for image search on personal computers, addressing the inefficiencies of traditional metadata-based methods. By utilizing a custom dataset of 11 thematic, the research explores CLIP's ability to map images and text into a shared feature space using embeddings. The model's effectiveness is demonstrated through its capacity to discern semantic relationships and differentiate between similar and distinct classes. The results indicate that integrating CLIP can significantly enhance media content management and search efficiency. Future work involves developing an application that combines metadata and semantic search capabilities to improve the quality and convenience of image search on personal computers.

В сучасному світі обсяги медіаконтенту стрімко зростають, що створює нові виклики для його ефективної організації та пошуку на пристроях. Традиційні методи пошуку, засновані на метаданих, або методи пошуку за зразком-зображення, які основані на ознаках зображень [1,2], стають менш ефективними через великі витрати ресурсів. У попередній роботі [3] було зазначено, що інтеграція моделей ШІ, зокрема моделей, що використовують ембеддинги, векторні представлення даних, для зіставлення тексту та зображень, може значно покращити функціональність пошукових систем.

Метою даної роботи є аналіз моделі OpenAI CLIP для задачі пошуку зображень на персональному комп'ютері. Для проведення аналізу було сформовано власний датасет, що складається з 11 класів по 10 зображень у кожному класі. Обрані класи представляють як суміжні, так і віддалені тематичні категорії: «Акваріуми», «Коти», «Різдво», «Різдвяний декор», «Одяг», «Комп'ютери», «Собаки», «Квіти», «Їжа», «Хмарочоси», «Заходи сонця» (рис. 1).

Рисунок 1 – Приклад зображень з датасету

Включення суміжних класів дозволяє оцінити здатність моделі розрізняти тонкі семантичні відмінності. Віддалені ж за змістом класи допомагають перевірити загальну класифікаційну потужність моделі.

Для кожного зображення та відповідного текстового опису було обчислено ембеддинги за допомогою моделі OpenAI CLIP. Вона складається з двох окремих енкодерів: текстового та візуального. Текстовий енкодер, зазвичай побудований на основі трансформерів, таких як GPT, перетворює текстові повідомлення у вектори ознак. Візуальний енкодер, базований на Vision Transformer (ViT), виконує аналогічне перетворення для зображень.

Ключовою особливістю CLIP є навчання у спільному просторі для тексту і зображень. Під час навчання модель використовує контрастивний підхід: вона отримує пари «текст-зображення» та навчається максимізувати косинусну схожість між ембеддингами відповідних пар, одночасно мінімізуючи схожість між невідповідними парами. В результаті, текстові та візуальні дані проектуються в єдиний простір ознак, де семантично близькі тексти та зображення розташовуються поруч. Приклад пари «зображення-текст» показано в рис. 2.

Рисунок 2 – Приклад пари «зображення-текст»

Нижче наведено діаграму типу «heatmap» (рис. 3), на якій показано результати порівняння ембеддингів для підмножини повного датасету, в якій знаходяться 5 елементів в одному з 5 класів (всього 25 пар). Було порівняно ембеддинги зображень (за стовпцями) та текстів-описів (за рядками), де за діагоналлю – порівняння зображення та його ж опису. За результатами можна відслідкувати, що модель розуміє семантичні зв’язки. Ембеддингове представлення наших класів та порівняння пар дало змогу вияснити, що векторний простір моделі має властивість розуміння класів, адже на діаграмі помітні квадрати більш інтенсивного забарвлення. Це означає, що в рамках цих квадратів косинусна, а від того і семантична, схожість є більшою.

Таким чином, інтеграція моделей на зразок OpenAI CLIP має значний потенціал для революціонізації способів пошуку та управління медіаконтентом на персональних пристроях.

Для подальшої роботи планується створення застосунку, який реалізує

комбінований метод пошуку – за метаданими та контекстуальний. Використовуючи наявні бази даних векторних ембеддингів, такі як FAISS або подібні, можна забезпечити ефективну та швидку обробку великого обсягу даних.

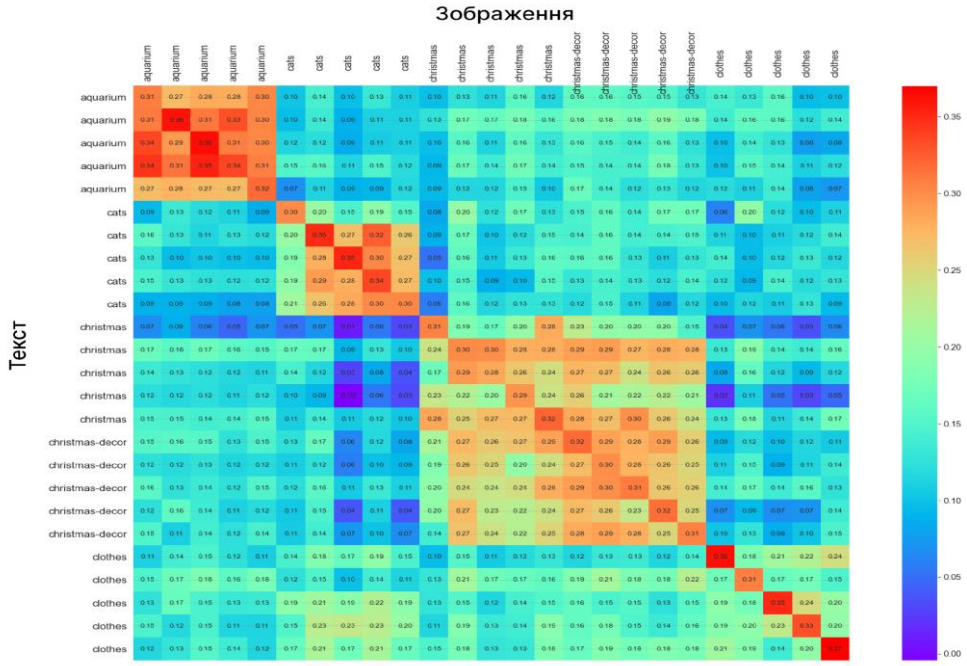

Рисунок 3 – Діаграма порівняльного аналізу пар ембеддингів

Базуючись на результатах проведеного аналізу, цей застосунок зможе поєднувати традиційні методи пошуку з можливостями семантичного пошуку. Це значно підвищить ефективність та зручність користувацького досвіду при управлінні та пошуку зображень на персональному комп'ютері.

Список використаних джерел:

1. Application a Committee of Kohonen Neural Networks to Training of Image Classifier Based on Description of Descriptors Set / V. Gorokhovatskyi et al. *IEEE Access*. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3404371.

2. Gorokhovatskyi O., Yakovleva O. Medoids as a packing of ORB image descriptors. *Advanced Information Systems*. 2024. Vol. 8, no. 2. P. 5–11. URL: https://doi.org/10.20998/2522-9052.2024.2.01.

3. Yakovleva O., Matúšová S., Koshel V. Implementation of AI approaches in current tools for managing image collections to improve the search capabilities. *Proceedings of the IV Correspondence International Scientific and Practical Conference «Science in motion: classic and modern tools and methods in scientific investigations» in Periodical International scientific journal «Grail of science».* (February 21, 2025). Vinnytsia, Ukraine – Vienna, Austria. 2025. Vol. 49. P. 752–755. https://doi.org/10.36074/grail-of-science.21.02.2025.096.