Vol. 6, No. 1, 2024

Наталія Мельникова[1], Петро Поберейко[2]

[1] Кафедра систем штучного інтелекту, Національний університет "Львівська політехніка", вул. С. Бандери 12, Львів, Україна, E-mail: melnykovanatalia@gmail.com, ORCID 0000-0002-2114-3436
[2] Кафедра систем штучного інтелекту, Національний університет "Львівська політехніка", вул. С. Бандери 12, Львів, Україна, E-mail: pobereyko.petro26@gmail.com, ORCID 0000-0002-8884-1255

ПОКРАЩЕННЯ МОЖЛИВОСТЕЙ ПОШУКУ ВІДЕО: ІНТЕГРАЦІЯ НЕЙРОННОЇ МЕРЕЖІ ПРЯМОГО ПОШИРЕННЯ ДЛЯ ЕФЕКТИВНОГО ФРАГМЕНТНОГО ПОШУКУ

Отримано: березень 12, 2024 / Переглянуто: березень 28, 2024 / Прийнято: квітень 01, 2024

https://doi.org/

Анотація. В умовах стрімкого збільшення обсягів відеоданих актуалізується проблема їх ефективного пошуку та аналізу. Це дослідження має на меті розробку та апробацію інноваційної системи для покращення швидкості та точності пошуку відео, використовуючи можливості глибоких згорткових нейронних мереж (DCNN) та нейронних мереж прямого поширення (FFNN). У рамках методології, розробленої для цього дослідження, відеодані обробляються через декілька послідовних етапів: від вилучення ознак до ідентифікації ключових кадрів і формування абстрактного векторного представлення. Центральне місце в системі відводиться глибоким згортковим нейронним мережам для аналізу зображень та нейронним мережам прямого пошорення для оптимізації процесу пошуку. Основними результатами дослідження є підвищення ефективності пошуку відео за рахунок зниження часу обробки даних та підвищення точності ідентифікації відповідних фрагментів. Оригінальність роботи полягає в інтеграції двох типів нейронних мереж для структурованого аналізу відеоданих, що є новим кроком у розвитку технологій пошуку відео. Практичне значення дослідження виражається у можливості застосування розробленої системи у різноманітних сферах, де потрібен швидкий та точний пошук відео: від медіаіндустрії до систем безпеки. Масштаби подальших досліджень включають адаптацію системи під специфічні типи відеоконтенту та розширення можливостей штучного інтелекту для глибшого розуміння відеоданих.

Ключові слова: глибокі згорткові нейронні мережі, вилучення візуальних ознак, мережі прямого поширення, машинне навчання.

Вступ

Протягом останніх років спостерігається стрімкий розвиток відеоконтенту та зростання його обсягів, що свідчить про перетворення цього типу візуальних даних на один з найефективніших способів передачі інформації. Така динаміка спонукає дослідників до розробки інформаційних систем, здатних аналізувати візуальні дані з метою задоволення потреб користувачів. Особливо цінними є розробки автоматизованих систем пошуку відео за довільними фрагментами. Академічний і практичний інтерес до цієї області зумовлений потенціалом таких систем у вдосконаленні пошукових механізмів у медіасфері та ефективності у боротьбі з розповсюдженням нелегального контенту. Ефективність пошуку відео на основі окремих фрагментів набуває особливої актуальності в контексті стрімкого збільшення обсягів цифрового контенту. Автоматизовані системи пошуку відео за фрагментами мають надзвичайно важливе значення у захисті інтелектуальної власності. Вони дають змогу своєчасно виявити та заблокувати копії відео, які порушують авторські права. Впровадження цих систем істотно підвищує ефективність

управління великими обсягами відеоданих в сферах медіа-індустрії, освітніх установ та наукових організацій. Окрім цього, технології побудови цих систем істотно впливають на розвиток досліджень в області штучного інтелекту. Оскільки удосконалення існуючих та розробка нових систем пошуку відео за фрагментами є можливими лише за умови удосконалення складних алгоритмів машинного навчання та комп'ютерного зору.

На сьогодні частково спостерігається певна невідповідність між низькорівневими даними відео та вимогами користувачів. Переважна більшість сучасних методів пошуку відео грунтується на парадигмі трансформації низькорівневих характеристик у вищі семантичні концепції. Цей підхід передбачає необхідність попередньої обробки даних, а результати такої трансформації часто виявляються нестабільними. Задача стає ще більш складною без врахування специфіки конкретної предметної області. Окрім цього, в сучасному цифровому медіапросторі спостерігається зростання кількості так званих нечітких дублікатів відео, складовими яких є уривки зі схожими, але не ідентичними змістовими елементами. У зв'язку з цим пошук оригіналу чи подібного відеоматеріалу без знання точного опису чи ключових слів за окремими фрагментами суттєво ускладнюється. Одним із шляхів вирішення цієї задачі є розроблення нових методик та систем пошуку відео, які б базувались на аналізі візуального контенту.

Аналіз сучасних досліджень та результатів з цієї галузі знань свідчить, що для ідентифікації відео та підвищення ефективності пошуку його оригіналу доцільно удосконалювати багатоетапні системи пошуку відео за результатами аналізу візуального контенту. Ці системи повинні забезпечувати оптимальний баланс між швидкістю пошуку та точністю співпадіння, мінімізуючи помилки у результатах. Одним з основних завдань задачі побудови таких систем є завдання виділення ознак зображень (кадрів) для створення метаданих (моделі представлення) зі збереженням сутності відеофрагменту для подальшого порівняння з даними в базі. На цьому етапі засобами технології обробки та аналізу зображень вирішуються завдання перетворення, виділення та аналізу ознак зображень. На цьому етапі, зображення піддаються трансформаціям для зміни їхніх геометричних чи кольорових характеристик з метою підготовки до ефективнішої обробки. Другим завданням є створення векторів числових значень, які ефективно репрезентують зображення для подальшого аналізу. Третє завдання полягає у вилученні та репрезентації семантичної інформації зображення. Аналіз характеристик кадру включає оцінку колірних параметрів (гістограми, моменти, корелограми та моделі на основі гаусових розподілів), а також текстурних особливостей, які незалежні від тону чи насиченості кольору і відображають однорідність зображення (часто вилучаються за допомогою фільтрів Габора). До цього додається аналіз контурних та формових характеристик, включаючи визначення форм об'єктів, які вилучаються на основі контурів або областей цих об'єктів.

Ефективний вибір та застосування цих ознак впливає на модель зображення, що використовується в алгоритмах аналізу, особливо при виявленні та розпізнаванні об'єктів. Важливим аспектом є також урахування умов функціонування систем, для яких ці алгоритми будуть реалізовані. Сучасні тенденції та результати вказують на кілька значущих підходів до представлення зображень, які ефективно використовуються у машинному навчанні. Ці підходи і методи, оптимальні для систем візуального пошуку, відіграють ключову роль у процесі машинного навчання та розробці програмних продуктів.

Постановка проблеми

У контексті стрімкого розвитку цифрових медіа, ключовим завданням є створення інноваційних систем пошуку відео для ефективного аналізу великих обсягів відеоданих. Цей розділ присвячено аналізу сучасних наукових досліджень та технологічних розробок у галузі пошуку відео за фрагментом. Зокрема, розглянуто та проаналізовано передові методи і техніки, що використовуються у провідних лабораторіях та компаніях, з метою їх використання для створення власної системи пошуку відео.

Більшість сучасних методів аналізу відеоданих для створення ефективного «хеш-контенту» ґрунтуються на вилученні просторових атрибутів із статичних зображень та часових параметрів із

відео. Важливим інструментом у цьому процесі є колірні гістограми, які виступають як потужний механізм у виявленні візуально схожого контенту. Це обумовлено тим, що подібні матеріали часто мають схожу колірну статистику, яка зазвичай зберігається навіть після перекодування або інших видів обробки. Hsu та інші дослідники пропонують методику розділення відео на сегменти відповідно до вмісту кадрів і використання локальних колірних гістограм для кожного сегменту. Однак, існує суттєве обмеження у використанні виключно колірних гістограм: є висока ймовірність «конфліктів» у випадку аналізу матеріалів з різним змістом, але схожими кольоровими характеристиками. Не менш ефективними методами аналізу відеоданих для створення ефективного «хеш-контенту» є методи сутність яких полягає в інтеграції методів хешування зображень з аналізом структури відео. Наприклад, визначивши межі відеоматеріалів можна вибрати ключові кадри для детальнішого аналізу. Також існує можливість скорочення тривалого відео до коротких фрагментів без втрати значущої інформації, що підтверджується низкою алгоритмів [5]. Ці алгоритми можна класифікувати на дві основні категорії: підходи до піксельної області та підходи до стисненої області. Підходи до піксельної області використовують гістограми та аналіз країв для виявлення суттєвих колірних відмінностей між послідовними кадрами. У той час як підходи до стисненого домену базуються на таких параметрах, як значення постійного струму, кількість кодованих макроблоків і вектори руху, що дає змогу уникнути повного розшифрування закодованого відео.

У дослідженні під назвою "Searching surveillance video contents using convolutional neural network", розроблена система пошуку вмісту відеоспостереження, що використовує глибокі згорткові нейронні мережі (CNN). Ця система ґрунтується на застосуванні заздалегідь навченої моделі VGG-16, використовуваної для тренування датасету. Особливістю системи є обробка ключових кадрів відео за допомогою детектора країв Sobel і методу Max-pooling, що дає змогу усувати надлишкові дані та забезпечувати компактність інформації. Модель VGG-16, яка є інтегральною частиною цієї системи, відома як одна з провідних глибоких згорткових нейронних мереж для класифікації зображень, включає в себе 16 шарів, включаючи згорткові шари, активаційні шари ReLU (Rectified Linear Unit), шари Max-pooling та повнозв'язні шари. Детектор країв Sobel ефективно вилучає ключові особливості зображення, такі як краї та контури, що сприяє виділенню структурних характеристик кадрів. Метод Max-pooling зменшує розмірність даних, зберігаючи при цьому важливу інформацію, що допомагає знизити обсяг даних для обробки та зменшити ризик перенавчання моделі. Функція активації ReLU працює за принципом: вхідний сигнал передається без змін у випадку позитивного значення та стає рівним нулю у випадку негативного. Математично це виражається як $f(x)=\max(0,x)$. Головною перевагою ReLU є її здатність прискорювати процес навчання мережі через ефективніший градієнтний спуск, особливо у порівнянні з іншими функціями активації, такими як сигмоїд або тангенс гіперболічний.

Zhuang [6] розробив методику кластеризації для визначення ключових кадрів, яка базується на групуванні схожих знімків та їх подальшому розподілі у кластери відповідно до їх позиціонування у відео. Wolfe [7] запровадив метод використання оптичного потоку для відбору ключових кадрів. Wang та інші [8] рекомендують вибір ключових кадрів із зон високої стисненості, використовуючи критерій високої інтенсивності руху, зокрема, зосереджуючи увагу на високій інтенсивності руху та його концентрації в центральній частині кадру. Liu та інші [10] запропонували інноваційний підхід, заснований на моделі «сприйнятої енергії руху», який дає змогу визначати ключові кадри на основі пікової енергії руху.

У дослідженні «VEDL: a novel Video Event searching technique using Deep Learning» подано триетапний підхід щодо ефективного пошуку подій у відео за допомогою методів глибокого навчання. Цей підхід передбачає:

1. Вилучення ключових кадрів: на цьому етапі з використанням методу решета Ератосфена, вибір ключових кадрів з відео, включає як глобальні так і локальні перспективи кадрів. В основі цього підходу покладено концепцію map-reduce, використання якої істотно зменшує загальний час обробки візуальних даних.

2. Виявлення подій: попередньо навчена модель глибокого навчання, що складається зі згорткових нейронних мереж (CNN) і рекурентних нейронних мереж (RNN), визначає події в ключових кадрах. Модель створена для ефективного опису подій в образах.

3. Визначення межі події та створення індексу: на цьому етапі обчислюються межі події та створюється індекс різних подій у відео. Процес передбачає визначення часу початку та закінчення подій, що сприяє ефективному пошуку відеоподій.

Дослідження зосереджується на ключових кадрах і використовує глибоке навчання, цей підхід значно економить час. Комплексний аналіз, який поєднує глобальні і локальні перспективи забезпечує ретельний аналіз відео. Але є ряд недоліків:

4. Складність: модель глибокого навчання може потребувати значних обчислювальних ресурсів.

5. Потенційні неточності. Покладення на виділення ключових кадрів і прогнозування моделі може призвести до пропущених або неправильно визначених подій у складних сценаріях відео.

Wu та інші [11] використовували тривалість запису як часовий параметр, застосовуючи швидкий алгоритм зіставлення. Зіставлення здійснювалось за допомогою алгоритму рядків суфіксів для ефективності, зосереджуючись виключно на часових даних. Jalousie та інші [12] впровадили метод вибору ключових кадрів, використовуючи радіальні вектори дизайну та дискретне косинусне перетворення (DCT) для формування хешу. Zargari та інші [14] розробили методику для вилучення стиснених функцій домену в H.264/AVC, використовуючи гістограму просторового передбачення як дескриптор.

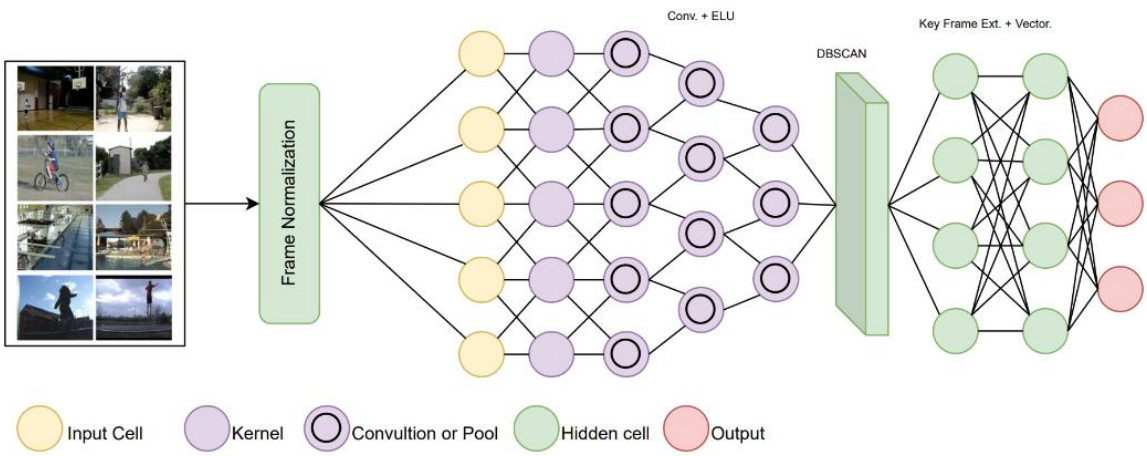
Загалом, автори досліджень вважають, що методи підсумовування та визначення ключових кадрів у відео є найефективнішими лише у випадках використання машинного навчання. Вони зазначають, що завдяки застосуванню алгоритмів машинного навчання системи пошуку відеоданих дають змогу користувачеві більш точніше проаналізувати великі обсяги відеоданих, виявити значущі моменти та автоматично відібрати важливі кадри. Ці технології дозволяють точно аналізувати та індексувати великі обсяги відеоданих, ефективно виявляючи і відновлюючи фрагменти відео на основі складних запитів. Інтеграція цих методів з розширеними алгоритмами обробки зображень та аналізу відео дозволяє створювати системи, які не лише точно відповідають на запити користувачів, але й забезпечують швидке та ефективне пошукове вирішення, адаптуючись до різноманітних форматів та типів відеоконтенту.

Виклад основного матеріалу

У даній роботі, засобами глибокий конволюційних нейронних мереж (DCNN), розроблено іноваційну систему пошуку відео, яка істотно підвищила швидкість опрацювання візуальних даних при збережені точності результатів пошуку відеофайлів. У запропонованій системі процес обробки даних структуровано на декілька послідовних етапів, кожний з яких виконується окремим модулем. Ця архітектура (Рис.1) реалізує послідовний аналіз відеоконтенту.

Процес пошуку відео системою починається з ініціації завантаження відеофрагменту, що є вихідним пунктом для його детальної обробки [15]. Для підвищення точності аналізу відео відеофрагмент розбивається на окремі кадри, після цього система обробляє кожний кадр індивідуально. Дальше для забезпечення однорідності даних проводиться нормалізація кожного кадру, яка здійснюється з метою приведення розмірів кадрів відеофрагменту до єдиного стандарту. Цей процес включає зменшення кількості пікселів у кожному кадрі за допомогою алгоритму білінійної інтерполяції, який використовує лінійну інтерполяцію спочатку в одному напрямку, а потім в іншому, для визначення нових значень пікселів. Спочатку визначаються чотири найближчі пікселі навколо цільової точки у вихідному зображенні. Потім застосовуються формули лінійної інтерполяції: спочатку горизонтально між верхніми та нижніми пікселями, у результаті отримуються проміжні результати, а потім вертикально між цими проміжними результатами для одержання кінцевого значення пікселя [15].

Цей метод забезпечує плавне та природне зображення після зміни розміру, ефективно зменшуючи артефакти та зберігаючи ключові візуальні характеристики оригінального зображення.



Мал. 1. Схематичне зображення архітектури нейронної мережі для системи пошуку відео за фрагментами.

Після процесу нормалізації кадрів, кожний кадр піддається процедурі екстракції особливостей, яка реалізується алгоритмом глибоких згорткових нейронних мереж (DCNN). Вибір DCNN для екстракції особливостей, обгрунтовується здатністю DCNN ефективно розпізнавати та аналізувати візуальні особливості на різних рівнях абстракції. Ці мережі автоматично вчаться визначати значущі характеристики, зокрема, текстури, кольори та форми, що має важливе значення для розрізнення різних типів сцен або об'єктів у відео. Для пошуку ключових кадрів ми обрали інтенсивність кольору як характеристику особливостей. Давайте коротко розглянемо два оптимальних колірних простори RGB і YUV. Якщо завдання передбачає пошук схожих відеокадрів на основі їхнього візуального вмісту, то використання колірного простору RGB може бути більш доцільним, оскільки воно безпосередньо представляє основні кольори, які сприймає людський зір. Це дозволяє точніше зіставляти подібні кадри на основі схожості кольорів. Однак пошук на основі кольору RGB може вимагати більше обчислень. Завдання нашої системи передбачає швидкий пошук подібних відеокадрів у цьому випадку ми обрали характеристики на основі вмісту яскравості та кольоровості, а саме використання колірного простору YUV [16]. Дозволило розділити інформацію про яскравість і кольоровість та забезпечити більш ефективний пошук. Крім того, інформацію про кольоровість ми субдискретизували для зменшення обчислювальних вимог. За допомогою обмеженої палітри кольорів та коефіцієнтів.

Для перетворення використано такі співвідношення та формули:

• Y (яскравість) визначається як зважена сума RGB значень:

$$Y = 0.299R + 0.587G + 0.114B \tag{1}$$

• U and V (color components) define chrominance relative to gray:

$$U = -0.14713R - 0.28886G + 0.436B \tag{2}$$

$$\mathrm{V} = 0{,}615\mathrm{R} - 0{,}51499\mathrm{G} - 0{,}10001\mathrm{B} \tag{3}$$

Числа у формулі перетворення RGB у YUV для яскравості (Y) випливають із способу, яким людське око сприймає світло. Людське око є більш чутливим до зеленого кольору, тому вага для зеленого каналу (G) є найбільшою. Червоний (R) також важливий, але менш чутливий, ніж зелений, тому має другу за величиною вагу. Найменша вага призначена блакитному каналу (B), оскільки людське око найменш чутливе до блакитного кольору. Ці коефіцієнти були визначені на основі стандартів PAL і NTSC, які походять від BT.470 System M та використовувались в SMPTE RP 177 [17]. Вони були встановлені для створення сигналу Y′UV із RGB джерела, використовуючи вагові значення для R, G та B, щоб отримати міру загальної яскравості або світності (Y′).

Отже у системі вхідними даними для DCNN є кадри у форматі YUV. Кожен кадр представлений як матриця з трьома каналами (Y, U, V), де кожен канал є двовимірною матрицею,

що відображає інтенсивність відповідного компоненту. Перший рівень DCNN містить конвульційні шари, із використанням фільтрів з розмірами 5x5. Даний розмір був обраний з ряду причин: більше параметрів (25 ваг), ніж фільтри менших розмірів, мають більше "рецептивне поле" – це означає, що вони більше уважно "дивляться" на ділянку вхідних даних, що допомагає виявляти більше глобальних ознак та використали для зменшення розмірності (downsampling) зображення. Результатом є нова матриця (так звана "feature map"), яка відображає виявлені фільтром особливості. В якості активаційної функції було обрано Exponential Linear Unit (ELU). ELU є розширенням традиційної функції активації Rectified Linear Unit (ReLU) [18], значення якої визначаються за формулою:

$$ELU(x) = \begin{cases} x, x > 0 \\ a(e^x - 1), x \le 0 \end{cases}, \quad (4)$$

де a – позитивна константа, що визначає насиченість до негативних чистих входів.

Основними причинами вибору ELU для нашої системи є:

1. Поліпшення Збереження Інформації: запропонована система вимагає точного витягу та збереження інформації про кольорові відтінки та текстурні особливості відеокадрів. ELU, з її властивістю м'якої насиченості для негативних вхідних значень, дозволяє ефективніше зберігати цю важливу інформацію, уникнути її деформації або втрати в глибоких шарах мережі.

2. Зменшення проблеми зникаючих градієнтів: глибокі нейронні мережі схильні до проблеми зникаючих градієнтів, особливо у випадках коли градієнти, необхідні для навчання, стають дуже малими у нижніх шарах. ELU зменшує цю проблему завдяки своїй нелінійній характеристиці на негативних значеннях, що значно покращує збереження градієнтів в процесі навчання [19].

3. Швидкість збіжності навчання: ефективність передачі градієнтів у ELU сприяє швидшій збіжності під час навчання мережі. Для нашої системи, де часто потрібно обробляти великі обсяги відеоданих, здатність швидко навчитися важливим особливостям є ключовою для забезпечення високої продуктивності.

4. Стабільність та надійність: В порівнянні з іншими активаційними функціями, такими як ReLU, ELU забезпечує більш стабільну та надійну обробку інформації, що є критично важливим для точного відеоаналізу.

У сукупності, вибір ELU як активаційної функції для нашої CNN, орієнтованої на аналіз YUV-відеокадрів, забезпечує оптимальне поєднання точності, швидкості та надійності, що є ключовими для успішного виконання завдань нашої системи.

Після вилучення особливостей прийнято рішення використовувати алгоритми кластеризації для подальшого удосконалення методик вибору ключових кадрів [18].

У якості оптимального алгоритму для цієї мети був обраний DBSCAN (Density-Based Spatial Clustering of Applications with Noise), який є ефективним щодо поставлених завдань [20]. Він виявляє кластери за результатами оцінки густини даних та спроможний автоматично адаптувати кількість та розміри даних кластерів відповідно до складності оброблюваних відеоданих, що є особливо важливим для аналізу великих відеопотоків зі змінною кількістю об'єктів на екрані. Окрім цього він дає змогу виділити окремі шумові точки, які не входять до жодного кластера. Це корисно для ідентифікації аномалій або відхилень у відеоданих, що можуть виникнути через помилки або шум. Крім того, однією з найбільш важливих переваг DBSCAN є його здатність до виконання кластеризації без попередньо визначеної кількості кластерів, завдяки чому забезпечується висока гнучкість і адаптивність обробки відеоданих. Для ефективного застосування DBSCAN, важливим є визначення оптимальних параметрів 'радіусу' (eps) та 'мінімальної кількості сусідів' (min_samples). Ці параметри були встановлені на основі низки числових експериментів, результати яких подано у наступних розділах роботи. Варто зазначити, що оптимальні значення цих параметрів можуть змінюватися в залежності від характеристик та обсягів оброблюваних даних. Процес визначення цих параметрів у ході експериментального аналізу був заснований на такому підході.

Початковий етап експериментального налаштування параметра 'радіус' (eps) для алгоритму DBSCAN полягав у визначенні низки його можливих значень. Для цього нами проведено серію кластеризацій, використовуючи різні значення радіусу, які варіювалися від менших до більших, для оцінки впливу цього параметра на результати кластеризації. Для кожного окремого значення радіусу ми аналізували якість кластеризації, використовуючи стандартизовані метрики оцінки. Цей аналіз дозволив нам ідентифікувати залежності значеннь радіусу, до структури кластерів. Таким чином, було обрано проведено дослідження, і вподальшому введений окрему підсистему для коригування цих значень.

Для обраного значення радіусу, провели експерименти з різними значеннями параметра min_samples. Експериментальні дослідження проводили в діапазоні зміни значень параметра min_samples від найменшого до найбільшого значень, оцінюючи результати кластеризації для кожного значення параметра min_samples. Зокрема, оцінювали, як зміна min_samples впливає на кількість та розмірність кластерів, а також на якість кластеризації.

Отже для початкових даних радіус було встановлено значення 0.5, що для нас є середнім значенням для роздільної здатності відео, швидкості зміни об'єктів у кадрах, кількості шуму у даних і загальної структури відео. Для min_samples значення 25 оскільки ми плануємо аналізувати середньо-завантажені відео і потребуємо високу стабільність та впевненість у виявленні кластерів у процесі навчання і подальшої автоматичної корекції цих значень. Встановлення значення min_samples = 25 робить кластеризацію менш чутливою до випадкових змін у відеоданих, що може бути корисним при аналізі середніх відеопотоків з помірним рівнем динаміки. Результатом кластеризації відеоданих є формування груп відеокадрів, де кожна із сформованих груп відображає виразну візуальну подібність між відеокадрами, враховуючи їх репрезентацію у вигляді компонентів YUV.

Наступний етапом є другий рівень нейронної мережі, який приймає групи кадрів та для кожної групи кадрів шукає такий, який має найбільшу вагу, таким чином будемо вважати його ключовим у групі. Для кожного кадру у форматі YUV, вираховуємо середнє значення яскравості (Y-канал) за допомогою формули:

$$Avg_Y = \sum Y[i, j]/(height * width), \tag{5}$$

де $Y[i, j]$ – значення в Y-каналі для області на позиції (i, j) у матриці, а (*height * width*) – загальна кількість блоків (групи пікселів) у кадрі.

Вихідним результатом другого рівня мережі буде група тривимірних матриць, де кожна матриця представляє ключовий кадр із найвищим середнім значенням яскравості у відповідній групі.

Третій рівень в нейронній мережі відповідатиме за перетворення групи ключових кадрів у вигляді тривимірних матриць, що містять інформацію про яскравість (Y), колір Chroma (U) та колір Chroma (V), в абстрактне векторне представлення. Даний процес є критичним для подальшого ефективного пошуку та порівняння відеофрагментів у базі даних. Кожений ключовий кадр у вихідній групі піддається інтенсивному обробленню на цьому рівні. Нехай масив YUV має розмірність (H, W, 3), де H – висота зображення, W – ширина зображення, і 3 відповідає компонентам Y, U і V.

Процес розгортання масиву полягає у перетворенні тривимірного масиву пікселів YUV у двовимірний масив, де кожен рядок відповідає пікселю з його компонентами Y (яскравість), U і V (кольорові компоненти Chroma). Цей процес дозволяє створити вектор ознак для кожного пікселя, спрощуючи подальшу обробку та аналіз даних.

У подальшому, вектори ознак пікселів об'єднуються у плоске представлення, формуючи векторну матрицю для всього зображення. Таке представлення забезпечує ефективне використання у системах пошуку відеофрагментів, дозволяючи легко проводити пошук та аналіз схожих векторів.

Для пошуку подібних векторів у великих датасетах відео використовується бібліотека FAISS, яка дозволяє швидко знаходити найближчі сусіди для заданих векторів. Використання індексу, створеного на основі YUV векторів, сприяє ефективному та точному пошуку.

Завершальним етапом є аналіз отриманих результатів, вибір та ранжування найбільш релевантних відеофрагментів на основі схожості та кількості співпадінь з ключовими кадрами. Такий підхід дозволяє оптимізувати пошук та виведення найбільш відповідних відео за запитом.

Для оптимізації процесу відеопошуку за допомогою DCNN, підвищення швидкості та ефективності пошуку відео за довільним фрагментом вирішено використати підсистему на базі мережі прямого поширення (Feed-forward Neural Network, FFNN) [21]. FFNN підсистема аналізує вихідні дані та метрики продуктивності DCNN, використовуючи алгоритми машинного навчання та математичні моделі для налаштування її гіперпараметрів. Мережа прямого поширення в цій ролі функціонує як аналітичний інструмент, що використовує набір даних про продуктивність DCNN (таких як точність, втрати, час обробки) для визначення оптимальних налаштувань. FFNN буде мати кілька шарів: вхідний шар, який приймає метрики DCNN, прихований шар для обробки цих даних, та вихідний шар, що генерує рекомендації щодо гіперпараметрів (Рис. 2).

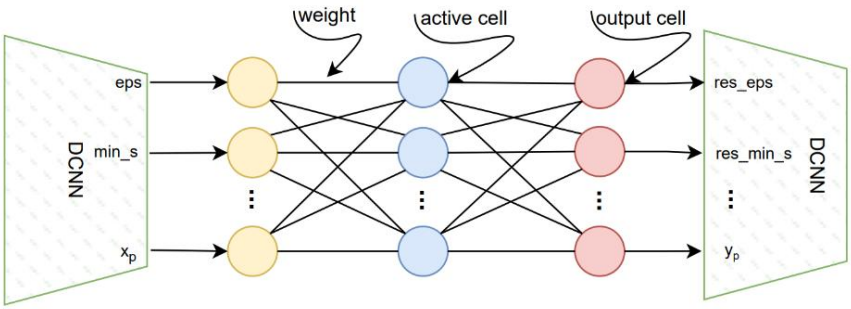


Рис. 2. Схематичне зображення архітектури нейронної мережі для оптимізації системи пошуку

Нейрон у мережі прямого поширення (FFNN) використовує зсув та активаційну функцію ReLU (Rectified Linear Unit) [22]. Працює за допомогою декількох ключових математичних операцій. Ось детальний опис цього процесу:

Лінійна Комбінація – кожен нейрон отримує вхідні сигнали x_1,x_2…,x_n, де n – кількість входів. Кожен вхідний сигнал множиться на відповідну вагу w_1,w_2…,w_n,. Сума цих зважених входів визначається за формулою:

$$z = w_1x_1 + w_2x_2 + ... + w_nx_n + b, \tag{6}$$

де b – зсув (bias), який додається до суми зважених входів.

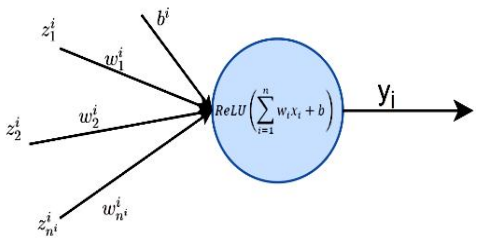


Рис. 3. Схематичне зображення нейрона із функцією активації ReLU

Зсув дозволяє нейрону зміщувати активаційну функцію вліво або вправо на графіку, що дає додаткову гнучкість мережі. Після розрахунку лінійної комбінації, результат z подається на активаційну функцію. Це означає, що якщо z є позитивним, функція повертає значення z, а якщо z є негативним, функція повертає 0 (Рис. 3). Таким чином, вихідний сигнал нейрона у буде:

$$y = \mathrm{Re}\,LU(z) = \mathrm{Re}\,LU\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{7}$$

де z – це лінійна комбінація входів, w_i – вага, пов'язана з і-м входом, x_i – і-й вхідний сигнал, n – загальна кількість входів, b – зсув (bias).

Вхідні дані для FFNN:

1. Час роботи системи: тривалість кожної операції у системі.
2. Параметри DBScan: eps (радіус сусідства) та min_samples (мінімальна кількість точок для формування кластера).

3. Додаткові метрики: включають точність кластеризації, кількість виявлених кластерів, відсоток шуму. Для точності застосовується Adjusted Rand Index (ARI) – міра схожості між двома кластеризаціями, що враховує випадковість. Приймає значення від -1 до 1, де 1 вказує на ідеальну кластеризацію (при експерментах із тим самим фрагментом). Для розрахунку шумових точок в алгоритмі DBScan точки, що не належать жодному кластеру, вважаються шумом. Відсоток шумів розраховуємо як відношення кількості шумофвих точок до загальної кількості точок.

FNN аналізуватиме вплив змін значень параметрів eps та min_samples на ефективність кластеризації. Навчатиметься прогнозувати оптимальні значення для цих параметрів залежно від вхідних даних (розмір і складність датасету).

Завдяки введенню підсистеми DCNN може динамічно коригувати параметри DBScan за результатами поточного аналізу точності та шуму. Регулярний моніторинг цих метрик допомогає в ідентифікації областей для вдосконалення алгоритмів та процесів обробки даних. Використання цих методів дозволить не тільки оцінювати ефективність кластеризації в вашій системі, але й надаватиме інформацію для її оптимізації, і тим самим підвищить якість та продуктивність обробки даних.

Результати та обговорення

У дослідженні використано набір даних для тренування та аналізу ефективності запропонованої системи пошуку відео за довільним фрагментом. Набір даних UCF-101 містить 13320 відеофайлів, які розділені на 101 неперекриваючу категорію [23]. Роздільна здатність кожного файлу складає 240 x 360 пікселів, а також мають фіксовану кадрову швидкість 25 кадрів на секунду (FPS). Тривалість відео варіюється від 1,06 секунди до 71,04 секунди.

Ці дані використовуються для тренування та валідації системи пошуку відео за довільним фрагментом. Вони дозволяють оцінити ефективність пошуку відеофрагментів розробленою системою. Для тренування та валідації роботи розробленої системи рекомендуємо використовувати метод "відкладеного" розділення тренувальних даних на підгрупи, а для визначення точності її тестування – тестовий набір. Дані були розділені на набори для навчання, тестування, які складалися з 70% даних навчання та 30% даних тестування.

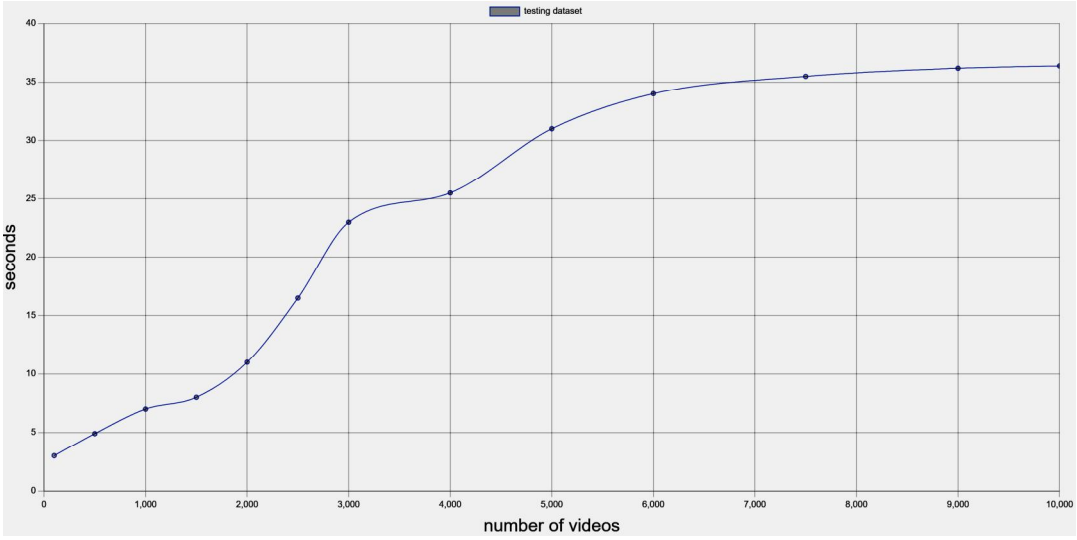


Рис. 4. Залежність швидкості пошуку від кількості відео у системі

Для навчання системи і наповнення даними відеофайли було переведоно у формат YUV. У системі шар пошуку модифіковано для можливості збереження файлу із ключовими кадрами та векторами із привязкою до відповідного відео. Для ідентифікації були обрані відеофрагменти з максимальною тривалістю до 20 секунд, причому частина цих фрагментів не входила до тестової

бази оригіналів. Високий рівень продуктивності було досягнуто за рахунок репрезентації відеоматеріалу у вигляді компактного, мінімізованого набору характеристик. Крім того, для підвищення ефективності обробки даних було застосовано методику кластеризації з індексацією векторного простору. Вході тренування та пошуку задопомогою FFNN [21] ми спостерігали за вихідними параметрами мережі, сформувавши графік залежності часу пошуку та кількості відео у базі даних (Рис. 4). Це дозволяє оцінити швидкість нормалізації та обробки даних.

У рамках проведеного дослідження здійснювалась оцінка ефективності кластеризації відеоданих, з особливим фокусом на точності ідентифікації ключових кадрів. Для цього були використані відеофрагменти з максимальною тривалістю до 20 секунд, з яких деякі не входили до бази оригіналів.

Основні параметри для алгоритму кластеризації DBSCAN (Density-Based Spatial Clustering of Applications with Noise) включали радіус (єпсилон) і мінімальну кількість зразків (min_samples). Радіус коливався від 0.3 до 0.6, тоді як min_samples залишався незмінним, оскільки кількість кадрів у датасеті була сталою.

У результаті дослідження було отримано наступні показники точності: для навчального набору даних точність склала 93,36%, а для тестового набору– 86,36%. Для порівняння, у контексті архітектури без використання підсистеми FFNN (Feedforward Neural Network), точність для навчального набору склала 74,36%, а для тестового– 66,04%. Ці результати підкреслюють значення додаткової підсистеми у процесі оптимізації параметрів кластеризації.

Важливо відзначити, що відсутність цієї підсистеми приводить до суттєвої розбіжності в точності, що свідчить про високу чутливість параметрів кластеризації до кінцевої репрезентації кадрів. Таким чином, інтеграція FFNN як частини системи забезпечує значне покращення загальної продуктивності та точності ідентифікації важливих кадрів у відеопотоці.

Остаточні результати цього дослідження вказують на значну вагомість архітектурної інновації, що була запропонована та впроваджена, і її вплив на ефективність виявлення ключових кадрів у відеопотоці.

Дослідження зосереджено на вдосконаленні можливостей пошуку відео, зокрема за допомогою нейронних мереж прямого зв’язку (FFNN) і глибоких згорткових нейронних мереж (DCNN). За результатами досліджень при роботі з візуальними даними можна врахувати такі фактори обмеженнях сучасних методів пошуку відео, зокрема в обробці великих обсягів відеоданих і складності відеовмісту. Усе це потребує вдосконалення алгоритмів розв’язування задач оцінювання та прогнозування. Дослідження висвітлює такі фактори, як виділення ознак, ідентифікація ключових кадрів і абстрактне векторне представлення як вирішальні для покращення можливостей пошуку відео. У ньому обговорюються труднощі перетворення відеоданих низького рівня в семантичні концепції високого рівня та необхідність попередньої обробки цих даних. Запропонована багатоетапна системи відеопошуку, яка збалансовує швидкість і точність пошуку, мінімізуючи помилки в результатах. Це включає складні алгоритми машинного навчання та методи комп’ютерного зору. Система використовує колірний простір YUV для ефективного представлення функцій і використовує алгоритм кластеризації (DBSCAN) для вибору ключових кадрів. Моделі глибокого навчання мають вирішальне значення для розпізнавання та аналізу візуальних особливостей на різних рівнях абстракції.

Оптимізація та виклики: архітектура системи розроблена для оптимізації процесу пошуку відео, адаптації до різних форматів і типів відеоконтенту. Однак складність моделей глибокого навчання та потенційні неточності у виборі ключових кадрів і прогнозуванні подій відзначаються як проблеми. Для розв'язання проблеми складності можна розглянути використання більш ефективних алгоритмів глибокого навчання, які вимагають менших обчислювальних ресурсів. Також можливе використання хмарних обчислень для забезпечення потрібної обчислювальної потужності без необхідності інвестицій у дороге обладнання.

Майбутні наслідки. Дослідження показує, що інтеграція вдосконалених алгоритмів обробки зображень і методів аналізу відео з машинним навчанням може призвести до систем, які не тільки точно відповідають на запити користувачів, але й надають швидкі та ефективні рішення для пошуку.

Висновки

На основі проведеного аналізу, дослідження робить висновок, що поточний інтерес до пошуку та аналізу відео є дуже актуальним. Розроблена система, яка використовує архітектуру Deep Convolutional Neural Networks (DCNN) і Feedforward Neural Networks (FFNN). Система ефективно обробляє великі обсяги відеоданих із структурованим підходом до обробки даних через послідовні етапи та модульні компоненти. Основна увага приділяється виділенню ознак, ідентифікації ключових кадрів і абстрактному векторному представленню для підвищення точності пошуку. Дослідження підкреслює потенціал нейронних мереж у вдосконаленні технологій пошуку відео, пропонуючи постійне вдосконалення для більш складного аналізу відео. Дослідження є значним внеском у сферу пошуку та обробки відеоданих, демонструючи важливість розвитку технологій у цій галузі, що швидко розвивається. Експериментальні результати підтверджують практичність і застосовність запропонованої системи в реальних сценаріях.

Перелік використаних джерел

[1] Indriyani and P. Dewanti, “Analysis of the Effect of Social Media on the Marketing Process in a Store or Business Entity ‘Social Media Store’,” Budapest International Research and Critics Institute-Journal, vol. 4, no. 4, pp. 9804-9814, 2021

[2] A. W. Bridges, “Skills, content knowledge, and tools needed in a 21st century university-level graphic design program,” Visual Communications Journal, vol. 52, no. 2, pp. 1–12, 2016.

[3] M. Y. Saragih and A. I. Harahap, “The Challenges of Print Media Journalism in the Digital Era. Budapest International Research and Critics Institute,” BIRCI-Journal, vol. 3, no. 1, pp. 540-548, 2020. https://doi.org/10.33258/birci.v3i1.805

[4] Konrad J, Wang M, Ishwar P, Wu C, Mukherjee D. LearningBased, Automatic 2D-to-3D Image and Video Conversion. IEEE Transactions on Image Processing. 2013; 22(9):3485–96. https://doi.org/10.1109/TIP.2013.2270375

[5] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In Advances in Neural Information Processing Systems, pages 2069–2077, 2014.

[6] Zhang HJ, Wu J, Zhong D, Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Pattern Recognit 30(4):643–658 https://doi.org/10.1016/S0031-3203(96)00109-4

[7] C. Cotsaces, N. Nikolaidis, and I. Pitas. Shot detection and condensed representation – a review. IEEE Signal Processing Magazine, 23:28–37, 2006. https://doi.org/10.1109/MSP.2006.1621446

[8] D. Mohammad, I. Aljarrah, and M. Jarrah. Searching surveillance video contents using convolutional neural network. IJECE, vol. 11, no. 2, pp. 1656-1665, 2021 https://doi.org/10.11591/ijece.v11i2.pp1656-1665

[9] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction", 2019 IEEE International Conference on Electrical Computer and Communication Technologies (ICECCT), pp. 1-7, 2019. https://doi.org/10.1109/ICECCT.2019.8869364

[10] G. Huang, Z. Liu, L. V. D. Maaten and K. Q. Weinberger, "Densely connected convolutional networks", CVPR, pp. 2261-2269, July 2017. https://doi.org/10.1109/CVPR.2017.243

[11]. L. Shao, F. Zhu, X. Li, Transfer learning for visual categorization: A survey, IEEE transactions on neural networks and learning systems vol. 26, pp. 1019–1034, 2014. https://doi.org/10.1109/TNNLS.2014.2330900

[12]. P. Naveen and B. Diwan, "Relative Analysis of ML Algorithm QDA LR and SVM for Credit Card Fraud Detection Dataset", 2020 Fourth International Conference on I-SMAC (IoT in Social Mobile Analytics and Cloud) (I-SMAC), pp. 976-981, 2020. https://doi.org/10.1109/I-SMAC49090.2020.9243602

[13]. Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra. Adaptive key frame extracting using unsupervised clustering. Proc. of IEEE Int Conf on Image Processing, pages 866–870, 1998.

[14]. Wolf. Key frame selection by motion analysis. IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1228–1231, 1996.

[15]. Li, D.; Wang, R.; Xie, C.; Liu, L.; Zhang, J.; Li, R.; Wang, F.; Zhou, M.; Liu, W. A Recognition Method for Rice Plant Diseases and Pests Video Detection Based on Deep Convolutional Neural Network. Sensors 2020, 20, 578. https://doi.org/10.3390/s20030578

[16]. Liu, Z.G., Zhang, X.Y., Wu, C.C.: A flame detection algorithm based on bag-of-features in the YUV color space. In: Proceedings on International Conference on Intelligent Computing and Internet of Things, Harbin, pp. 64–67 (2015).

[17]. Divya Shree, Chander Kant. Building Efficient Neural Networks For Brain Tumor Detection. Journal of Positive School Psychology, vol. 6, no. 11, 2022.

[18]. Z. Qiumei, T. Dan and W. Fenghua, "Improved Convolutional Neural Network Based on Fast Exponentially Linear Unit Activation Function," in IEEE Access, vol. 7, pp. 151359-151367, 2019, https://doi.org/10.1109/ACCESS.2019.2948112

[19]. L. Li, M. Doroslovački and M. H. Loew, "Approximating the Gradient of Cross-Entropy Loss Function," in IEEE Access, vol. 8, pp. 111626-111635, 2020, https://doi.org/10.1109/ACCESS.2020.3001531

[20]. N. Ohadi, A. Kamandi, M. Shabankhah, S. M. Fatemi, S. M. Hosseini and A. Mahmoudi, "SW-DBSCAN: A Grid-based DBSCAN Algorithm for Large Datasets," 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, 2020, pp. 139-145, https://doi.org/10.1109/ICWR49608.2020.9122313

[21]. T. Kwon, "Average Data Rate Analysis for Hierachical Cell Structure under Nakagami-m Fading Channel with a Two-layer Feed-Forward Neural Network," 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Barcelona, Spain, 2019, pp. 1-4, https://doi.org/10.1109/WiMOB.2019.8923280

[22]. L. D. Medus, T. Iakymchuk, J. V. Frances-Villora, M. Bataller-Mompeán and A. Rosado-Muñoz, "A Novel Systolic Parallel Hardware Architecture for the FPGA Acceleration of Feedforward Neural Networks," in IEEE Access, vol. 7, pp. 76084-76103, 2019, https://doi.org/10.1109/ACCESS.2019.2920885

[23]. X. Luo, O. Ye and B. Zhou, "An Modified Video Stream Classification Method Which Fuses Three-Dimensional Convolutional Neural Network," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2019, pp. 105-108, https://doi.org/10.1109/MLBDBI48998.2019.00026

Nataliia Melnykova[1], Petro Pobereiko[2]

[1] Department of Artificial Intelligence Systems, Lviv Polytechnic National University, 12, St. Bandera str., Lviv, Ukraine, E-mail: melnykovanatalia@gmail.com, ORCID 0000-0002-2114-3436

[2] Department of Artificial Intelligence Systems, Lviv Polytechnic National University, 12, St. Bandera str., Lviv, Ukraine, E-mail: pobereyko.petro26@gmail.com, ORCID 0000-0002-8884-1255

ADVANCING VIDEO SEARCH CAPABILITIES: INTEGRATING FEEDFORWARD NEURAL NETWORKS FOR EFFICIENT FRAGMENT-BASED RETRIEVAL

Received: March 12, 2024 / Revised: March 28, 2024 / Accepted: April 01, 2024

Abstract. In the context of rapidly increasing volumes of video data, the problem of their efficient search and analysis becomes more acute. This research aims to develop and test an innovative system to improve the speed and accuracy of video search, utilizing the capabilities of Deep Convolutional Neural Networks (DCNN) and Feedforward Neural Networks (FFNN). Within the methodology developed for this study, video data are processed through several sequential stages: from feature extraction to key frame identification and the formation of an abstract vector representation. Deep Convolutional Neural Networks are central to the system for image analysis and Feedforward Neural Networks for optimizing the search process. The main results of the study include an increase in video search efficiency by reducing data processing time and increasing the accuracy of identifying relevant fragments. The originality of the work lies in the integration of two types of neural networks for structured analysis of video data, which is a new step in the development of video search technologies. The practical significance of the research is expressed in the possibility of applying the developed system in various areas where fast and accurate video search is needed: from the media industry to security systems. The scope of further research includes adapting the system to specific types of video content and expanding the capabilities of artificial intelligence for a deeper understanding of video data.

Keywords: deep convolutional neural networks, video search, data processing, feature extraction, feedforward neural networks