УДК 004.932:004.032.26
DOI: 10.31673/2412-9070.2025.042042
Т. М. КИСІЛЬ, ст. викладач;
ORCID:0000-0002-5123-0768
О. В. ЗІНЧЕНКО, доктор техн. наук, професор,
ORCID:0000-0002-3973-7814
Державний університет інформаційно-комунікаційних технологій, Київ

ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ ДЛЯ АНАЛІЗУ РУХОМИХ ОБ'ЄКТІВ У ВІДЕОПОТОЦІ

У статті проведено огляд та аналіз сучасних методів використання згорткових нейронних мереж (CNN) для обробки рухомих об'єктів на відео. Розглянуто особливості застосування CNN у задачах розпізнавання, відстеження та класифікації об'єктів у відеопослідовностях. Окреслено ключові переваги згорткових нейромереж, зокрема їх здатність до автоматичного виділення просторових і часових ознак, а також обробки великих обсягів даних.

Аналізуються основні архітектури CNN, які використовуються для роботи з відеоданими. Особлива увага приділяється технологіям обробки даних у реальному часі, адаптивному масштабуванню моделей та інтеграції нейромереж із апаратними платформами. Висвітлено основні виклики, включаючи проблему масштабованості, оптимізації обчислювальних ресурсів, врахування часових залежностей та забезпечення точності в умовах складного фону.

Отримані результати можуть бути корисними для подальшого розвитку інтелектуальних систем комп'ютерного зору, зокрема у сфері автономних транспортних засобів, систем відеоспостереження, аналізу поведінки та автоматизації промислових процесів. Представлений аналіз сприяє розумінню перспектив і обмежень використання CNN у динамічних сценаріях і окреслює напрями для майбутніх досліджень.

Ключові слова: згорткові нейронні мережі; розпізнавання об'єктів; відстеження руху; комп'ютерний зір; 3D-CNN; гібридні архітектури.

Вступ

Постановка проблеми. Аналіз рухомих об'єктів у відеопотоці є актуальною проблемою в галузях комп'ютерного зору, безпеки, автономного транспорту, відеоспостереження та інтелектуальних систем. Традиційні алгоритми обробки відео, що базуються на класичних методах виявлення руху, часто демонструють низьку стійкість до шумів, змін освітлення, складного фону та частих змін сцени.

Сучасні методи глибокого навчання, зокрема згорткові нейронні мережі (CNN), показали високу ефективність у розв'язанні задач класифікації, сегментації та відстеження об'єктів на зображеннях. Проте застосування CNN до обробки відеопотоку в реальному часі вимагає вирішення низки технічних і методологічних питань, пов'язаних з обробкою послідовностей кадрів, збереженням просторово-часових залежностей та забезпеченням продуктивності системи. Відеодані стали одним із наймасовіших джерел інформації — вони активно застосовуються у сфері безпеки, автономного транспорту, міського моніторингу, медицини, спорту та промисловості. Ключовим завданням постає автоматичний аналіз рухомих об'єктів у відеопотоці, зокрема їх виявлення, класифікація, відстеження та інтерпретація поведінки. Традиційні алгоритми обробки відео — зокрема методи на основі віднімання фону, оптичного потоку або статистичних моделей — мають суттєві обмеження при змінному освітленні, складному фоні, великій кількості об'єктів або частковому їх перекритті. Такі методи не можуть забезпечити необхідного рівня точності та адаптивності в реальних умовах.

Сучасні досягнення у сфері штучного інтелекту, особливо глибокого навчання, відкрива-

ють нові можливості для вирішення вказаних задач. Зокрема, згорткові нейронні мережі (CNN) довели свою ефективність у задачах аналізу зображень, однак їх пряме застосування до відеопотоку не дозволяє повноцінно враховувати часову динаміку об'єктів. Виникає потреба у нових підходах, що поєднують просторове сприйняття CNN з моделями, здатними аналізувати часову послідовність (наприклад, рекурентними мережами або тривимірними згортками).

Постає проблема у розробці методів та архітектур згорткових нейронних мереж, які забезпечать точне, стабільне та швидке виявлення й відстеження рухомих об'єктів у відеопотоці в умовах реального часу. Вона безпосередньо пов’язана з важливими науковими завданнями в галузі комп’ютерного зору, обробки відео та інтелектуальних систем. З практичного боку, вирішення цієї проблеми має велике значення для: створення ефективних систем відеоспостереження та безпеки; підвищення автономності роботизованих систем і транспортних засобів; аналітики в охороні здоров’я (аналіз руху пацієнтів, реабілітація); розвитку "розумних міст" та інфраструктурного моніторингу.

У зв’язку з цим, метою дослідження є розробка та дослідження моделей згорткових нейронних мереж, здатних ефективно і точно виявляти, розпізнавати та відстежувати рухомі об’єкти у відеопотоці з урахуванням часової динаміки сцени. Особлива увага приділяється інтеграції згорткових мереж із рекурентними або тривимірними структурами, а також оптимізації архітектур для використання у реальному часі.

Аналіз останніх досліджень і публікацій. Проблематика автоматичного аналізу відео та розпізнавання рухомих об’єктів є однією з ключових у галузі комп’ютерного зору. Упродовж останніх років відзначено розвиток методів глибинного навчання, зокрема згорткових нейронних мереж (CNN), які продемонстрували високу ефективність в задачах виявлення та класифікації об’єктів на статичних зображеннях [5,6]. Проте класичні CNN обмежені у здатності враховувати часову динаміку, що критично важливо для обробки відеопотоків. У зв’язку з цим виникли численні модифікації, спрямовані на адаптацію CNN до обробки послідовностей кадрів.

Значним внеском у розвиток цієї галузі стали дослідження, присвячені тривимірним згортковим мережам (3D-CNN), які дозволяють враховувати часові характеристики на рівні об’ємної згортки [12]. Також активно розвиваються гібридні архітектури, що поєднують CNN з рекурентними мережами, зокрема ConvLSTM (Convolutional Long Short-Term Memory) [5], які ефективно моделюють просторово-часові залежності. Серед практичних рішень широко використовуються моделі YOLO (You Only Look Once) [8], Faster R-CNN [9], а також DETR (DEtection TRansformer), що поєднує глибинні згортки з attention-механізмами [1].

Попри успіхи науковців, все ще залишаються невирішеними питання підвищення стійкості моделей до завад, ефективного використання ресурсів у реальному часі та інтеграції просторово-часових контекстів у компактних архітектурах. Саме ці аспекти формують основу актуальності дослідження.

Основна частина

Розробка методів аналізу рухомих об'єктів на відео є важливою складовою сучасних комп’ютерних зорових систем. Завдяки стрімкому розвитку технологій штучного інтелекту та машинного навчання, зокрема Convolutional Neural Networks (CNN), з’явилися інноваційні підходи до обробки візуальної інформації. Ці методи дозволяють ефективно вирішувати задачі, пов’язані з розпізнаванням, класифікацією та відстеженням об’єктів на відеопослідовностях, що має критичне значення для таких галузей, як автономні транспортні системи, відеоспостереження, робототехніка та аналіз поведінки.

Згорткові нейронні мережі (ЗНМ) зарекомендували себе як потужний інструмент для обробки зображень завдяки здатності до автоматичного виділення ознак та адаптації до складних структур даних. Особливий інтерес викликає їх застосування для обробки відеоданих, де виникають додаткові складнощі, пов’язані з часовими змінами, динамікою об'єктів та великими обсягами інформації. Однак ефективність CNN залежить від правильно обраних архітектур, алгоритмів навчання та параметрів моделі, що потребує систематичного аналізу існуючих методів.

Згорткові нейронні мережі набули широкого застосування в обробці відеоданих завдяки їхній здатності аналізувати просторово-часові характеристики. Відео розглядається як послідовність зображень (кадрів), тому для його аналізу необхідно враховувати не лише просторові особливості, а й часову динаміку між кадрами. Існують алгоритми, здатні ідентифікувати символи, розпізнавати об'єкти й виконувати інші завдання, що значно спрощує робочі процеси, підвищуючи їхню точність та надійність шляхом усунення людського фактора. Однак навчити комп'ютер розпізнавати об'єкти доволі складно, адже його сприйняття кардинально відрізняється від людського. На відміну від людини, комп'ютер не володіє досвідом і не здатний інтуїтивно ідентифікувати об'єкти без попередньо отриманих даних. Для вирішення цієї задачі застосовуються технології машинного навчання, що включають формування великих баз даних та створення датасетів. Навчання моделі передбачає виділення специфічних ознак і закономірностей, що дозволяє комп'ютеру ідентифікувати об’єкти на основі наявної інформації [7].

Процес навчання моделі, попри завантаження великих обсягів даних, не гарантує повної точності, тому необхідне подальше "донавчання" на нових наборах. У контексті відеоспостереження ключовим етапом є аналіз, що починається із розпізнавання зображень. Штучний інтелект, використовуючи технології машинного навчання, здатен класифікувати дії та аналізувати поведінку, однак це можливо лише за умови попереднього навчання моделі на релевантних даних. Успіх навчання значною мірою залежить від розміру та якості датасетів [13], які можуть бути створені самостійно або завантажені з відкритих джерел.

Для досягнення високої точності обробки зображень кожне з них розбивається на дрібні ділянки (пікселі), які слугують вхідними нейронами. Сигнали проходять через багатошарові структури і тисячі нейронів з мільйонами параметрів порівнюють отримані дані з раніше обробленими зразками. Наприклад, якщо потрібно розпізнати зображення кішки, то відповідні ділянки порівнюються з мільйонами зразків, що вже є в базі навченої моделі.

Розпізнавання образів є однією з основних задач комп’ютерного зору [15], яка полягає у виявленні та класифікації візуальних об’єктів у цифрових зображеннях, таких як фотографії або відеокадри. Ця задача виконується двома основними способами:

- *традиційними методами обробки зображень*;
- *сучасними підходами з використанням глибокого навчання*.

Методи обробки зображень є неконтрольованими, не потребують великих обсягів анотованих даних і використовуються в інструментах на кшталт OpenCV [16]. Проте вони обмежені низкою факторів, таких як погане освітлення, оклюзія об'єктів або неоднорідний фон. У свою чергу, методи глибокого навчання базуються на контрольованому навчанні, що дозволяє досягти високої стійкості до складних умов, але вимагає значних обчислювальних ресурсів і великих датасетів.

Об’єктне розпізнавання з використанням глибокого навчання активно застосовується як у наукових дослідженнях, так і у створенні комерційних продуктів, включаючи автономні транспортні засоби, системи відеоспостереження [14] та інші інтелектуальні технології. Завдяки розвитку графічних процесорів та доступності анотованих баз даних (MS COCO, Caltech, KITTI), ці технології стають дедалі ефективнішими та доступнішими для практичного використання.

Відеодані мають унікальні характеристики, що робить їх складними для обробки. Просторова складова представлена кожним кадром відео, який можна розглядати як окреме зображення, тоді як часова складова демонструє зміну сцен або рух об'єктів у послідовності кадрів. Крім того, великий обсяг відеоінформації, що може включати тисячі або мільйони кадрів, вимагає значних обчислювальних ресурсів.

Для обробки відео за допомогою згорткових нейронних мереж використовуються різні архітектури. Двовимірні CNN (2D-CNN) обробляють кожен кадр відео окремо [15], однак втрачають часову інформацію. Тривимірні CNN (3D-CNN) розширюють згортковий оператор у тривимірному просторі, що дозволяє одночасно враховувати просторові та часові залежності. Інтеграція CNN із рекурентними нейронними мережами (RNN), наприклад, LSTM або GRU, допомагає аналізувати довгострокові часові залежності. Двопотокові CNN (Two-Stream

CNN) комбінують просторовий аналіз окремих кадрів із часовим аналізом руху між ними. Крім того, архітектури типу Spatio-Temporal Networks (STN) інтегрують обидва підходи для кращого опрацювання просторово-часових особливостей.

CNN відіграють ключову роль у розпізнаванні відео завдяки своїй здатності автоматично виділяти просторові особливості кожного кадру [14], такі як форми, текстури та об'єкти. Завдяки багатошаровій архітектурі вони аналізують як низькорівневі, так і високорівневі ознаки, одночасно враховуючи часові залежності між кадрами [16]. Наприклад, 3D-CNN дозволяють інтегрувати просторову та часову інформацію, що робить їх ефективними для завдань розпізнавання дій, подій або об'єктів [13].

CNN вирішують низку завдань у відеоаналізі, включаючи класифікацію відео, виявлення та локалізацію об’єктів, їх відстеження, сегментацію відео та розпізнавання дій [8, 15]. Наприклад, вони можуть ідентифікувати категорію відео, виявляти дії, як-от біг чи танці, а також аналізувати динаміку сцени. Такі можливості широко використовуються у спортивній аналітиці, системах безпеки, автономному транспорті та інших галузях.

Серед популярних архітектур для обробки відео вирізняються:

- *2D-CNN*, що аналізують просторові особливості окремих кадрів;
- *3D-CNN*, які обробляють як просторові, так і часові залежності;
- *Two-Stream CNN* - поєднують аналіз руху кожного кадру;
- *I3D (Inflated 3D-CNN)* - базуються на використанні попередньо натренованих 2D-CNN для інтеграції просторово-часових особливостей.

Згорткові нейронні мережі є надзвичайно ефективним інструментом для розпізнавання рухомих об'єктів, забезпечуючи високу точність і адаптивність у багатьох сферах застосування. Наприклад, CNN широко використовуються у відеоспостереженні для автоматичного виявлення та відстеження людей або транспортних засобів, у спортивній аналітиці для ідентифікації рухів спортсменів, а також у робототехніці для навігації автономних пристроїв. Інтеграція CNN із такими архітектурами, як 3D-CNN, Two-Stream CNN або рекурентні нейронні мережі (RNN) дозволяє вирішувати складні задачі, наприклад, класифікацію дій у відео або передбачення поведінки об'єктів.

У контексті завдання виявлення, класифікації та відстеження рухомих об'єктів у відеопотоці розглядається відеопослідовність наступного вигляду:

$$V=\{I_1,I_2,...,I_T\} \ , \tag{1}$$

де I_t — це окремий кадр відео у момент часу t; T — загальна кількість кадрів у відеофрагменті.

Метою поставленої формалізованої задачі є побудова функції F(V) для кожного кадру I_t та об'єкту *i*, який присутній у кадрі, і визначається за наступною формулою:

$$F(V)\rightarrow\{(B_{t,i},C_{t,i})\} \ , \tag{2}$$

де $B_{t,i}$ — координати граничного прямокутника (bounding box), який окреслює і-й об'єкт на кадрі t; $C_{t,i}$ — клас або ідентифікатор *i*-го об'єкта.

Функція F апроксимується згортковою нейронною мережею з параметрами θ, які навчаються під час оптимізації. Навчання моделі здійснюється шляхом мінімізації сукупної функції втрат:

$$L=L_{det}+\lambda_1 \cdot L_{cls}+\lambda_2 \cdot L_{track} \ , \tag{3}$$

де L_{det} — функція втрат, пов’язана з точністю виявлення об’єктів, що часто базується на метриках IoU, GIoU або L1-помилках координат обмежувальних рамок; L_{cls} — функція втрат класифікації, яка зазвичай реалізується як крос-ентропійна втрата між передбаченим та істинним класами; L_{track} — функція втрат для оцінки коректності відстеження об'єктів між кадрами. Ця складова може включати відстані між ознаками об'єктів, втрати через ID-switch, або похибки асоціації; λ_1, λ_2 — гіперпараметри, що регулюють внесок відповідних компонент у загальну функцію втрат.

Таким чином, задача зводиться до навчання нейронної мережі, яка здатна одночасно вирішувати три взаємопов’язані підзадачі: *детекцію, класифікацію, трекінг* з урахуванням просто-

рово-часової природи відеоданих та специфіки реального середовища функціонування систем комп'ютерного зору.

Розглянемо архітектури CNN, які демонструють різні підходи до роботи з відеоданими (рис. 1). Так, *2D-CNN* обробляють просторові ознаки, витягуючи інформацію з кожного кадру окремо. Це може бути ефективним для задач із мінімальними змінами між кадрами, наприклад, для статичних сцен [3]. Наприклад, у системах безпеки такі моделі можуть виявляти об'єкти у статичному відео. У випадках, коли важливими є як просторові, так і часові залежності, використовуються *3D-CNN*. Ці архітектури працюють із блоками кадрів, обробляючи їх як єдиний тривимірний масив. Вони ефективні, наприклад, у завданнях розпізнавання рухів у спортивних змаганнях, таких як аналіз техніки бігу або плавання [4].

Two-Stream CNN поєднує аналіз статичних кадрів із часовою інформацією, наприклад, оптичним потоком. Цей підхід широко використовується для класифікації складних дій, наприклад, у кіберспорті для розпізнавання маневрів гравців або в кіноіндустрії для аналізу бойових сцен. Інтеграція CNN з рекурентними нейронними мережами (наприклад, LSTM) дозволяє моделювати довгострокові часові залежності, що є корисним для аналізу відео тривалістю кілька хвилин або годин, наприклад, для моніторингу дорожнього трафіку.

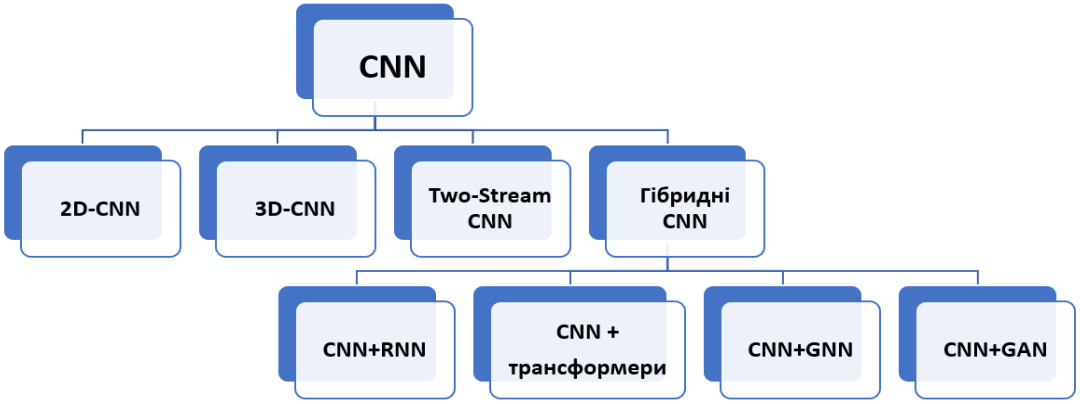

Рис. 1. Архітектури CNN, які використовуються при розпізнаванні рухомих об'єктів відеоконтенту

Гібридні архітектури CNN із іншими методами відкривають ще ширші можливості [2]. Поєднання CNN із трансформерами дозволяє враховувати глобальний контекст, що важливо для аналізу довгих послідовностей, наприклад, у задачах розумного розпізнавання сцен у фільмах. Інтеграція CNN із графовими нейронними мережами (GNN) корисна для моделювання взаємодій між об'єктами, наприклад, у відеоаналітиці, де люди або автомобілі в сцені розглядаються як вузли графа, а їхні зв'язки — як ребра. Подібний підхід може бути використаний у системах розумних міст для управління транспортними потоками.

Генеративні моделі, такі як Generative Adversarial Networks (GAN), у поєднанні з CNN відкривають можливості для створення синтетичних відеоданих, які можуть бути використані для навчання або покращення існуючих моделей. Наприклад, GAN дозволяє створювати реалістичні відео, які використовуються для симуляції руху автомобілів у навчальних системах для автономного транспорту.

Попри переваги, CNN стикаються з певними викликами. Їх висока обчислювальна склад-ність вимагає значних ресурсів, а потреба у великих обсягах навчальних даних може стати бар'єром для їх широкого застосування. Однак завдяки оптимізації моделей, наприклад, вико-ристанню компресії або створенню моделей для пристроїв із низькою потужністю, CNN мо-жуть застосовуватися навіть у пристроях із обмеженими ресурсами, таких як дрони або камери спостереження.

Експериментально дослідимо роботу *CNN* моделей VGG16 / VGG19 при розпізнаванні відеоконтенту (рис. 2, 3). На представлених зображеннях показано порівняння результатів розпізнавання об'єктів на двох кадрах відео у реальному часі. Зображення (рис. 2) демонструє роботу навченої моделі згорткової нейронної мережі (CNN), інше зображення (рис. 3) — ненавченої моделі CNN.

Навчена модель CNN коректно ідентифікує декілька об'єктів, таких як кішка на підлозі та декоративна іграшка біля ялинки. Рамки виділення об'єктів чіткі, ідентифікація виконується з високою точністю — об'єкти розпізнаються як цілісні та мають коректні контури. *Ненавчена модель CNN* демонструє велику кількість хибно розпізнаних або дубльованих об’єктів. На ялинці та навколо неї видно багато випадкових підписів (як наприклад, "dugong", "pillow", "plastic bag"), які не відповідають дійсним об'єктам сцени. Об’єкти визначаються некоректно, з низькою узгодженістю з реальністю, а підписані області виглядають випадковими або неточними.

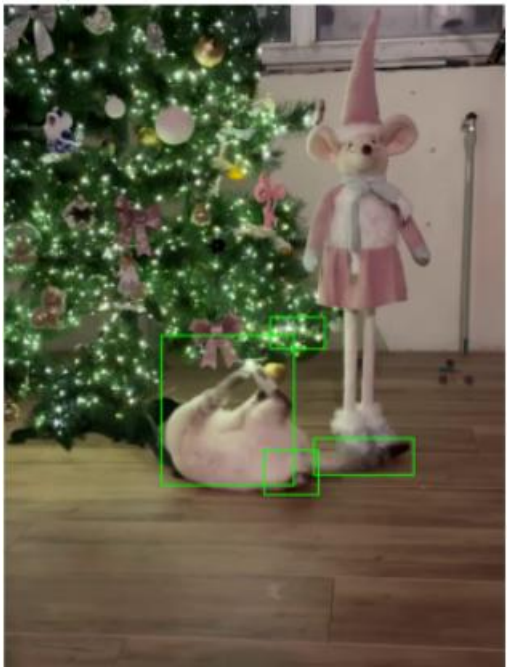

Рис. 2. Результати розпізнавання навченої моделі CNN

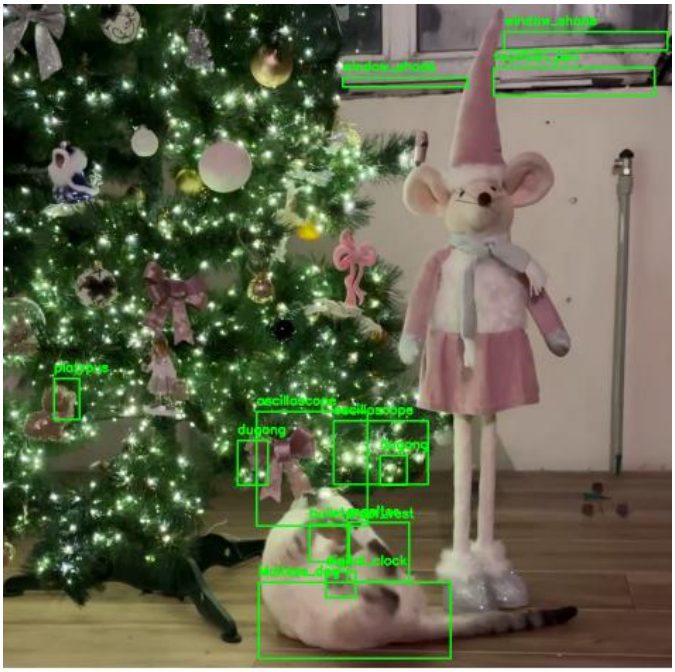

Рис. 3. Результати розпізнавання ненавченої моделі CNN

Порівняння ілюструє важливість процесу навчання моделі CNN. Навчена модель значно краще справляється з точним розпізнаванням об’єктів у кадрі, тоді як ненавчена модель демонструє низьку якість класифікації, хаотичне розпізнавання та численні помилки.

Як результат, при створенні системи розпізнавання відео на базі ненавченої згорткової нейронної мережі технічно можливе, однак за замовчуванням вона буде працювати неточно й часто видавати помилкові результати розпізнавання. Якщо експериментувати з модифікованою моделлю, то варто взяти одну із готових архітектур CNN (YOLO, MobileNet, ResNet) і протестувати її реакцію на відеопотоках без попередніх вагових коефіцієнтів. Надалі виникає необхідність під’єднати безнаглядові методи кластеризації K-means, сегментації контурів Canny, Sobel для приблизного виокремлення об’єктів. Варто додати ручні евристичні фільтри: наприклад, вважати об’єктом ділянку, яка суттєво відрізняється від фону за кольором, розміром або рухом у кількох послідовних кадрах. Ефективність можна підсилити класичними прийомами комп’ютерного зору — оптичним потоком Лукаса-Канаде для виявлення руху або відніманням фону для фіксації змін у статичних сценах. Оскільки ненавчена модель не розпізнає зображення, усі порогові значення й правила необхідно налаштовувати вручну. В підсумку така CNN залишається моделлю без знань: вона здатна лише реагувати на зміни та/або виділяти підозрілі області, не усвідомлюючи їхньої природи. Цього достатньо для дослідних демонстрацій, порівняння з навченою версією або пояснення переваг глибокого навчання, але для практичних задач краще хоча б мінімально натренувати модель на власному наборі даних та/або застосувати transfer learning.

Оцінювання точності згорткових нейронних мереж у задачах розпізнавання рухомих об'єктів у відео показує, що результат значною мірою залежить від якості навчання, обсягу та специфіки даних, умов зйомки та вибору архітектури мережі. Якщо модель була попередньо навчена на відповідному наборі відео, її ефективність у виявленні об'єктів у русі значно зростає.

У протилежному випадку, коли мережа не проходила донавчання, її результати значно гірші, особливо в умовах реального часу, де критичним є збереження просторово-часової стабі-льності.

Порівняльний аналіз точності таких моделей, як VGG16, VGG19, YOLO, RNN, (таблиця) показує істотну різницю між адаптованими й універсальними варіантами. Наприклад, YOLOv5 виявився найпридатнішим для обробки відео у реальному часі завдяки поєднанню високої швидкості, точності та стабільності. Базові архітектури, VGG16 або VGG19 менш ефективні самостійно, адже не враховують часову динаміку — тому часто використовуються лише як екстрактори ознак у зв'язці з рекурентними мережами або трекінговими алгоритмами. У задачах, де ключовим є розпізнавання руху, особливо добре себе зарекомендували 3D-CNN і ConvLSTM, які враховують динамічні зміни в кадрах. Загалом, моделі без адаптації значно поступаються за точністю, що підтверджує важливість цільового навчання при відеоаналізі.

Порівняння точності моделей CNN/RNN для відеоаналізу

Модель	Тип	Навчена на відео (mAP @0.5)	Не навчена на відео (mAP @0.5)	IoU (навчена)	IDF1 (трекінг)	Примітки
YOLOv5	CNN (1-stage)	85–90%	55–60%	>0.5	75–80%	Висока швидкість, добра точність
VGG16	CNN (2D)	70–75%	45–50%	~0.45	50–60%	Потребує адаптації, не включає рух
VGG19	CNN (2D)	72–78%	48–52%	~0.48	52–63%	Глибша, але повільна
Faster R-CNN	CNN (2-stage)	80–85%	50–60%	>0.5	65–75%	Висока точність, повільна
3D-CNN	CNN (3D)	82–88%	~60%	>0.5	70–78%	Враховує просторово-часову інформацію
ConvLSTM	CNN + RNN	83–89%	~62%	>0.5	75–82%	Враховує рух і пам'ять
RNN (LSTM)	RNN	65–72%	<50%	<0.4	~60%	Потребує CNN для виділення ознак

На графіку (рис. 4) показано порівняння точності різних моделей згорткових і рекурентних

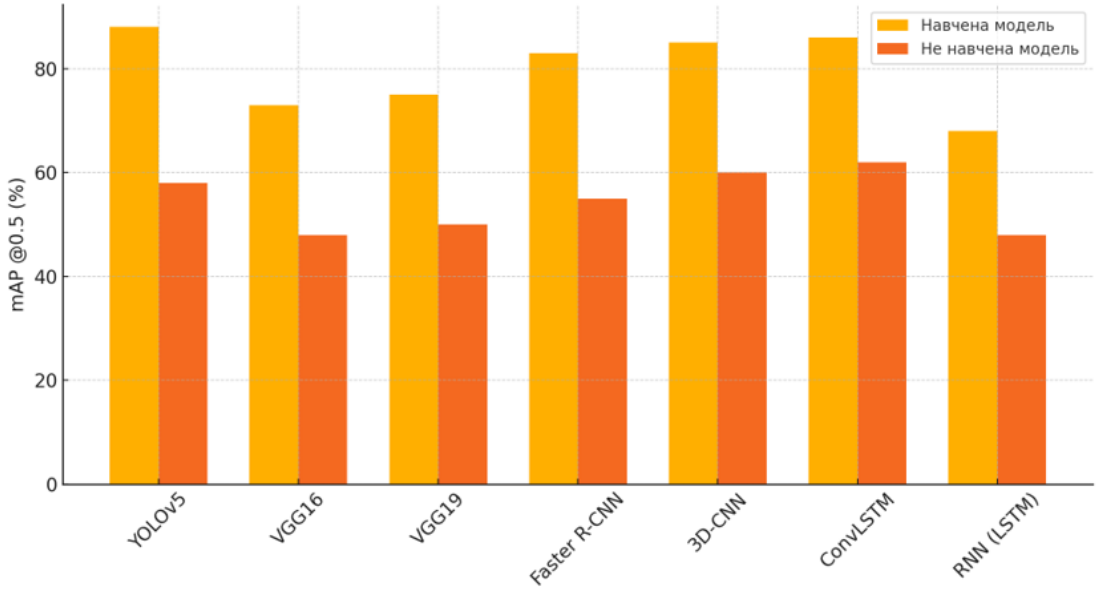

Рис. 4. Графік порівняння точності моделей CNN/RNN

нейронних мереж (CNN і RNN) у задачах відеоаналізу, зокрема розпізнавання об'єктів у динамічних сценах. По вертикальній осі відкладено значення середньої точності (mAP) при порого-

вому значенні IoU 0.5, яка є ключовим показником якості детекції. Горизонтальна вісь відображає назви моделей, серед яких присутні YOLOv5, VGG16, VGG19, Faster R-CNN, 3D-CNN, ConvLSTM і RNN (LSTM).

Як видно, що всі навчено-адаптовані моделі суттєво перевершують ненавчені варіанти, що свідчить про важливість спеціалізованого навчання для досягнення високої точності у відеоаналізі. Найвищі показники демонструють моделі YOLOv5, 3D-CNN і ConvLSTM у навченому варіанті, які перевищують позначку 85%. У той же час, ті самі архітектури без навчання мають точність, що коливається в межах 55–62%. Базові CNN-моделі, такі як VGG16 та VGG19, без навчання демонструють найнижчу ефективність — близько 45–50%. Це вказує на те, що без адаптації до відео даних їхня здатність до аналізу рухомих об'єктів суттєво обмежена. Модель RNN (LSTM) також показує невисоку точність без навчання, але в навченому варіанті її ефективність помітно зростає. Таким чином, графік наочно демонструє ключову залежність результативності відеоаналізу від ступеня підготовки моделі, а також від вибору її архітектури.

Проведене порівняння точності моделей CNN та RNN у задачах відеоаналізу показало суттєву перевагу попередньо навчених моделей над ненавченими. Найвищу ефективність у реальному часі продемонструвала модель YOLOv5, яка поєднує швидкість та високу точність. Моделі, здатні враховувати часові залежності та просторову динаміку (зокрема 3D-CNN і ConvLSTM), також досягають високих результатів за умови адаптації до відеоданих. У той же час, ненавчені варіанти всіх архітектур значно поступаються у точності, що підтверджує критичну роль навчання для реального застосування.

Висновки та перспективи

Для побудови ефективних CNN-моделей у режимі неконтрольованого навчання слід звертати увагу на використання методів самонавчання (self-supervised learning), які дають змогу моделі витягувати патерни з неанотованих відеоданих. Доцільно впроваджувати підходи, що базуються на контрастивному навчанні, кластеризації ознак, автоенкодерах, а також на архітектурах типу Vision Transformer (ViT) у поєднанні з CNN для підсилення узагальнення. Для стабільності детекції об'єктів доцільно додавати часові механізми, як-от ConvLSTM або оптичний потік, які дозволяють моделі вловлювати контекст руху. Крім того, важливо забезпечити постобробку результатів з використанням евристичних або фільтраційних підходів для зменшення шуму та помилкових детекцій, які часто виникають у ненавчених системах. У перспективі, поєднання неконтрольованого навчання з обмеженим донавчанням на спеціалізованих вибірках може забезпечити прийнятний компроміс між автономністю системи та точністю її роботи.

Основним напрямком розвитку є створення програмного забезпечення для обробки відео у реальному часі. Такий додаток повинен використовувати оптимізовані CNN-моделі, що здатні швидко аналізувати відеопотік, ідентифікувати об'єкти, відстежувати їх рух і класифікувати події. Інтеграція з апаратним прискоренням (GPU, TPU) та використання розподілених обчислень у хмарних середовищах забезпечать високу продуктивність навіть на мобільних пристроях. Крім того, адаптивні алгоритми з можливістю оновлення моделей на основі зворотного зв’язку сприятимуть підвищенню точності обробки у змінних умовах. В результаті, такий додаток повинен бути швидким, енергоефективним і стійким до змін освітлення та динаміки сцени. Використання згорткових нейронних мереж (CNN) у поєднанні з трансформерами або рекурентними мережами дозволить точно враховувати просторово-часові залежності. Підтримка апаратного прискорення, інтеграція з хмарними платформами для обробки великих масивів даних і застосування методів стиснення моделей (квантування, прунінг) підвищать продуктивність. У результаті створений додаток матиме розширені можливості налаштування параметрів розпізнавання та візуалізації результатів у режимі реального часу.

Список літератури

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. End-to-End Object Detection with Transformers // Proceedings of the European Conference on Computer Vision (ECCV). – 2020.

2. Donahue J., Hendricks L. A., Guadarrama S., et al. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-14.

3. Feichtenhofer C., Pinz A., Zisserman A. "Convolutional Two-Stream Network Fusion for Video Action Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.1-9.

4. Girshick R., Donahue J., Darrell T., Malik J. "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 3, 2016, pp. 642–657.

5. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016.

6. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems (NeurIPS). – 2012.

7. LeCun Y., Bengio Y., Hinton G. "Deep Learning." Nature, vol. 521, 2015, pp. 436–444.

8. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016.

9. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks // Advances in Neural Information Processing Systems (NeurIPS). – 2015.

10. Shi, X., Chen, Z., Wang, H. та ін. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting // Advances in Neural Information Processing Systems (NeurIPS). – 2015.

11. Szeliski R. Computer Vision: Algorithms and Applications. Springer, 2021, p. 812.

12. Tran D., Bourdev L., Fergus R., Torresani L., Paluri M. "Learning Spatiotemporal Features with 3D Convolutional Networks." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 31-47.

13. Zhou B., Andonian A., Torralba A., Oliva A. "Temporal Relational Reasoning in Videos." Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp.51-67.

14. Зінченко О. В., Звенігородський О. С., Кисіль Т. М. Згорткові нейронні мережі: особливості розпізнавання відео-контенту // XII Міжнародна науково-практична конференція «Математика. Інформаційні технології. Освіта» (2-4 червня 2023 р.), — Луцьк: ВНУ, 2022.- С.86-88

15. Зінченко О.В., Звенігородський О.С., Т.М. Кисіль. Згорткові нейронні мережі для вирішення задач комп'ютерного зору. Телекомунікаційні та інформаційні технології. 2022. № 2 С. 4-12.

16. Кисіль Т.М., Зінченко О.В. OpenCV: розпізнавання та ідентифікація об'єктів відеоконтенту // Всеукраїнська науково-технічна конференція "Застосування програмного забезпечення в інформаційно-комунікаційних технологіях" (24 квітня 2024р.), - Київ: ДУІКТ, 2024 р., с. 328-330.

T. Kysil, O. Zinchenko

CONVOLUTIONAL NEURAL NETWORKS FOR MOVING OBJECT ANALYSIS IN VIDEO STREAMS

This article presents a comprehensive review and in-depth analysis of methods for applying Convolutional Neural Networks (CNNs) to the processing of moving objects in video streams. Modern CNN-based approaches enable effective solutions to tasks such as object recognition, tracking, and classification in video sequences, which form the foundation of many areas in computer vision. The

study examines architectures including 2D-CNNs for processing individual frames, 3D-CNNs for capturing spatiotemporal dependencies, and hybrid models (RNN+CNN and ConvLSTM) that integrate the advantages of different video data processing techniques.

The advantages of CNNs are highlighted, including automatic feature extraction, high adaptability to varying shooting conditions (e.g., changes in lighting, presence of noise), and the ability to handle large-scale data. Particular attention is paid to addressing scalability challenges and optimizing computational resources to ensure real-time data processing, which is critical for fields such as autonomous transportation systems, intelligent video surveillance, and human/object behavior analysis.

A significant portion of the article discusses challenges associated with large-scale video data processing, such as modeling temporal dependencies, preventing the loss of important spatial features, and reducing the energy consumption of models for deployment on mobile and embedded platforms. The article outlines future prospects for CNN development, including the creation of lightweight models, the adoption of transfer learning, and the integration of CNNs with transformers to handle dynamic video environments. The obtained results are of significant importance for the further advancement of computer vision algorithms. The proposed recommendations on the use of CNNs will contribute to the development of high-performance intelligent video processing systems and enhance their efficiency across a wide range of applications.

Keywords: convolutional neural networks; object recognition; motion tracking; computer vision; 3D-CNN; hybrid architectures.