DOI: https://doi.org/10.26565/2304-6201-2025-66-09

УДК (UDC) 004.93

Ясінський Ярослав Андрійович *Аспірант кафедри комп'ютерних систем та робототехніки Навчально-наукового інституту комп'ютерних наук та штучного інтелекту; Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022; e-mail: yaroslav.yasinskyi@karazin.ua;* **https://orcid.org/0009-0008-0460-5687**

Бакуменко Ніна Станіславівна *доцент кафедри комп'ютерних систем та робототехніки Навчально-наукового інституту комп'ютерних наук та штучного інтелекту; Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022; e-mail: n.bakumenko@karazin.ua;* **https://orcid.org/0000-0003-3496-7167**

Порівняльний аналіз моделей YOLOv5 та MobileNetV3 для розпізнавання зображень в реальному часі

Актуальність: у сучасних умовах зростаючої потреби у швидкому й точному розпізнаванні об'єктів у реальному часі, особливо для мобільних і вбудованих систем, постає питання вибору оптимальних моделей штучного інтелекту. Порівняння легковагових та високоточних архітектур, таких як YOLOv5 і MobileNetV3, є важливим для розробки ефективних комп'ютерних зорових систем та дослідження принципів побудови гібридних моделей.

Мета: порівняння архітектур YOLOv5 і MobileNetV3 з метою аналізу ефективності для застосування у задачах розпізнавання об'єктів у реальному часі, та підтвердження, що гібридні моделі можуть підвищити ефективність виконання цих задач.

Методи дослідження: методи препроцесінгу зображень, методи навчання глибоких нейронних мереж, вимірювання точності, швидкості обробки та використання ресурсів; порівняльний аналіз результатів для оцінки ефективності моделей.

Результати: експериментальне дослідження показало, що YOLOv5 демонструє кращу загальну точність на тестовому наборі COCO, проте вимагає більше обчислювальних ресурсів. MobileNetV3, натомість, забезпечує пришвидшене виведення та ефективне функціонування на пристроях із низькою потужністю, жертвуючи частково точністю. Таким чином, обидві моделі підтвердили свою придатність для реальних застосувань, а вибір між ними залежить від конкретного балансу між швидкістю, точністю та обмеженнями платформи. Поєднання цих моделей дає кращі результати в розпізнаванні об'єктів, хоча це може збільшити розмір самої моделі та споживання ресурсів.

Висновки: у результаті дослідження проведено порівняння моделей YOLOv5, MobileNetV3 та гібридної моделі для задачі розпізнавання об'єктів. Гібридна модель продемонструвала кращу точність та баланс між швидкістю обробки і використанням ресурсів порівняно з окремими моделями. Це свідчить про доцільність використання гібридних підходів для підвищення ефективності систем комп'ютерного зору в реальних умовах. Отже, гібридна модель є перспективним напрямком для подальших досліджень і практичної реалізації.

Ключові слова: *розпізнавання зображень, комп'ютерний зір, гібридна модель, CNN, YOLOv5, MobileNetV3.*

Як цитувати: Ясінський Я. А., Бакуменко Н. С. Порівняльний аналіз моделей YOLOv5 та MobileNetV3 для розпізнавання зображень в реальному часі. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2025. вип. 66. С. 90-98. https://doi.org/10.26565/2304-6201-2025-66-09

How to quote: Yasinskyi Y. A., Bakumenko, N. S., "Comparative analysis of YOLOv5 and MobileNetV3 models for real-time image recognition". *Bulletin of Kharkiv National University named after V. N. Karazin, series Mathematical modeling. Information Technology. Automated control systems,* vol. 66, pp. 90-98. https://doi.org/10.26565/2304-6201-2025-66-09[in Ukrainian]

1 Вступ

Виявлення та класифікація об'єктів залишаються фундаментальними проблемами в галузі комп'ютерного зору, особливо в сценаріях, де швидкість і точність є критично важливими, наприклад, в автономних системах, спостереженні та мобільних додатках. В даній роботі проведено порівняльний аналіз двох широко використовуваних архітектур згорткових нейронних мереж – YOLOv5 і MobileNetV3 з акцентом на їхню застосовність та ефективність у задачах розпізнавання зображень у реальному часі.

Мета роботи – огляд та порівняння архітектур нейронних мереж, розроблених для розпізнавання зображень у реальному часі та дослідження гібридної моделі створеної на базі обраних з метою підвищення ефективності.

Огляд нейромережевих моделей для виявлення об'єктів у реальному часі був проведений з акцентом на їхні принципи побудови, операційну ефективність і застосовність у практичних сценаріях. Також був проведений детальний аналіз моделей YOLOv5 і MobileNetV3, включаючи дослідження їхніх архітектурних варіацій і модифікацій, спрямованих на підвищення продуктивності в різних обчислювальних середовищах.

Оцінка ефективності моделей була проведена на тестовому наборі даних, що забезпечило стандартизовану основу для навчання та тестування моделей. Нейронні мережі навчалися в контрольованих умовах, а їхня продуктивність вимірювалася за допомогою точності, швидкості отримання висновків та ефективності використання ресурсів.

2 Опис нейромережевих архітертур

Згорткові нейронні мережі (ЗНМ) є основною архітектурою для більшості систем розпізнавання зображень, завдяки своїй здатності використовувати просторові ієрархії та інваріантність трансляції у візуальних даних. Згорткові шари, які ідентифікують локальні ознаки за допомогою фільтрів, що навчаються, поєднуються з шаром агрегації (пулінгу), який зменшує просторові розміри та обчислювальну складність. Однак традиційні глибокі архітектури ЗНМ, такі як VGG та ранні варіанти ResNet, часто виявляються занадто обчислювально дорогими для програм реального часу, що вимагає архітектурних інновацій, спеціально розроблених для підвищення ефективності [1].

Згортки, розділені по глибині, є важливою архітектурною модифікацією, яка значно зменшує обчислювальну складність, зберігаючи при цьому репрезентативну здатність. Цей підхід, популяризований архітектурами MobileNet, факторизує стандартні згортки на згортки по глибині, які застосовують фільтри до кожного вхідного каналу незалежно, а потім на точкові згортки, які об'єднують вихідні. Блоки стиснення та збудження забезпечують механізм уваги, який покращує якість представлення об'єктів без значних обчислювальних витрат. Ці блоки обчислюють вагу уваги для кожного каналу за допомогою глобального усередненого пулінгу з подальшим повним з'єднанням шарів, що дозволяє мережі адаптивно підкреслювати інформативні ознаки, пригнічуючи менш релевантні. Обчислювальні витрати залишаються мінімальними, оскільки обчислення уваги оперує стислими представленнями ознак. Методи дистиляції знань дозволяють розробляти компактні учнівські мережі, які навчаються від більших і точніших мереж вчителів. Архітектура учнів розроблена таким чином, щоб бути ефективною за своєю суттю, зберігаючи при цьому здатність фіксувати основні знання, закодовані в мережі вчителя. Такий підхід дозволяє створювати мережі, які досягають майже оптимальної точності при дотриманні обмежень реального часу [2].

Трансформери, такі як Vision Transformer (ViT) та Swin Transformer, продемонстрували високу продуктивність у різних завданнях комп'ютерного зору, ефективно моделюючи залежності на великій відстані. Незважаючи на досягнуту точність, їхні високі обчислювальні та пам'ятні вимоги створюють значну перешкоду для розгортання в малих пристроях, які обмежені розміром, вагою та обчислювальною потужністю [3].

В даній роботі MobileNet та YOLOv5 були обрані для більш детального розгляду через їхні взаємодоповнюючі переваги в задачах розпізнавання об'єктів. MobileNet добре підходить для використання на пристроях з ресурсами, що можуть бути обмеженими, завдяки своїй легкій архітектурі та ефективній продуктивності, що робить його ідеальним для додатків реального часу, де обчислювальна потужність обмежена. YOLOv5, з іншого боку, пропонує баланс між швидкістю і точністю, забезпечуючи високу продуктивність виявлення при відносно короткому часі

висновку. Разом ці моделі представляють практичний компроміс між ефективністю і точністю, що відповідає цілям дослідження.

3 Аналіз архітертури YOLOv5s

Модель YOLO (You Only Look Once) – це фундаментальна архітектура в галузі виявлення об'єктів у реальному часі, відома своїм балансом швидкості та точності. На відміну від класичних фреймворків виявлення об'єктів, які застосовують класифікатори до різних областей зображення (такі як R-CNN та його похідні), YOLO розглядає виявлення об'єктів як єдину регресійну задачу. Він безпосередньо прогнозує обмежувальні границі та ймовірності класів на базі повних зображень за одну оцінку, що забезпечує швидке виявлення, яке може бути застосоване у реальному часі. Оригінальна архітектура YOLO, представлена у 2016 році, розділяє вхідне зображення на сітку. Кожна комірка сітки відповідає за прогнозування фіксованої кількості обмежувальних рамок та показників достовірності для цих рамок, а також показників достовірності класів. Показник достовірності відображає ймовірність того, що прогнозована рамка містить об'єкт, та точність обмежувальної рамки. Прогнозуючи всі обмежувальні рамки та ймовірності класів за один прохід через нейронну мережу, YOLO досягає високої швидкості виведення, хоча спочатку мала проблеми з точністю локалізації та продуктивністю на малих об'єктах [4].

З метою подолання цих обмежень, було розроблено кілька вдосконалених версій YOLO. Алгоритм YOLOv2 запровадив кілька вдосконалень, таких як використання пакетної нормалізації, класифікатори високої роздільної здатності, опорні рамки для кращих апріорних значень обмежувальних рамок та кластеризацію розмірів для підвищення точності прогнозування рамок. Він також включив ієрархічну систему класифікації, яка дозволила йому виявляти понад 9000 категорій об'єктів, навіть з частково позначеними наборами даних.

YOLOv3 ще більше вдосконалив архітектуру, впровадивши глибшу магістральну мережу під назвою Darknet-53, яка використовує залишкові зв'язки для покращення продуктивності навчання. Ця версія підтримує багатомасштабні прогнози, що дозволяє моделі ефективніше виявляти об'єкти різних розмірів. YOLOv3 прогнозує рамки у трьох різних масштабах, використовуючи карти ознак з різної глибини мережі, значно підвищуючи точність як для малих, так і для великих об'єктів [5].

YOLOv4, базується на YOLOv3, інтегруючи досягнення комп'ютерного зору, такі як зважені залишкові з'єднання (WRC), міжетапні часткові з'єднання (CSP), крос-міні-пакетна нормалізація (CmBN) та самозмагальне навчання (SAT). YOLOv4 наголошує на збалансованому компромісі між швидкістю та точністю, оптимізуючи мережу для використання як з графічними процесорами, так і з традиційними обчислювальними середовищами [6].

YOLOv5, неофіційне продовження, розроблене спільнотою та розміщене на GitHub компанією Ultralytics, що є реалізованим у бібліотеці Python PyTorch, пропонує більш модульну конструкцію, легкість експериментів та постійні оновлення. Він включає такі функції, як автоматичне навчання обмежувальних рамок, еволюція гіперпараметрів та масштабовані розміри моделей (YOLOv5s, YOLOv5m, YOLOv5l та YOLOv5x), щоб врахувати різні компроміси між точністю та швидкістю. Незважаючи на деякі суперечки щодо його правил найменування та походження, ця архітетура є популярною серед фахівців на практиці [7].

Наступні версії, включаючи YOLOv6, YOLOv7 та YOLOv8, продовжують розширювати межі виявлення об'єктів. YOLOv6 зосереджена на продуктивності промислового рівня, оптимізуючи ефективність розгортання на периферійних пристроях. YOLOv7 інтегрує додаткові архітектурні інновації, такі як розширені мережі ефективної агрегації шарів (E-ELAN) та методи повторної параметризації моделі, для підвищення здатності до навчання без шкоди для швидкості. YOLOv8, розроблена Ultralytics, являє собою перехід до уніфікованої моделі для виявлення, сегментації та класифікації [8].

Загалом, сімейство моделей YOLO значно вплинуло на ландшафт розпізнавання зображень у реальному часі.

4 Аналіз архітертури MobileNetV3

Архітектура є MobileNetV3 — це архітектура згорткової нейронної мережі, призначена для ефективного розпізнавання зображень на мобільних та периферійних пристроях. MobileNetV3 базується на попередніх версіях MobileNetV1 та MobileNetV2, поєднуючи автоматизований

пошук нейронної архітектури (NAS) з серією оптимізаторів. Результатом стала модель, яка досягає балансу між точністю та обчислювальною ефективністю, що робить її особливо ефективною для завдань розпізнавання зображень у реальному часі за обмежених апаратних умов [9].

MobileNetV3 включає кілька основних архітектурних удосконалень, які відрізняють її від попередніх версій. Одним із найважливіших удосконалень є використання мобільної оберненої згортки вузького місця (MBConv), спочатку представленої в MobileNetV2, яка додатково вдосконалена в MobileNetV3 шляхом додавання легких механізмів уваги, відомих як блоки Squeeze-and-Excitation (SE). Ці блоки SE адаптивно перекалібрують реакції на ознаки каналів, що дозволяє моделі підкреслювати більш інформативні ознаки, одночасно пригнічуючи менш корисні [10].

Ще однією ключовою інновацією є інтеграція функцій активації hard-swish (h-swish) замість традиційних функцій ReLU або swish. Функція h-swish апроксимує активацію swish обчислювально ефективним способом, що дозволяє виконувати кращі нелінійні перетворення без значного збільшення часу обчислень. Крім того, мережа включає нелінійності, адаптовані для апаратної ефективності, що зменшує доступ до пам'яті та споживання енергії, що є важливим для розгортання на мобільних платформах [11].

MobileNetV3 постачається у двох основних варіантах, MobileNetV3-Large та MobileNetV3-Small, кожен з яких розроблений для різних рівнів доступності ресурсів та вимог до продуктивності. MobileNetV3-Large оптимізований для вищої точності та підходить для випадків використання, коли доступно більше обчислювальних ресурсів. Він містить глибшу архітектуру з більшою кількістю параметрів і зазвичай використовується в завданнях, що вимагають високої точності класифікації. І навпаки, MobileNetV3-Small оптимізований для сценаріїв з більш жорсткими обмеженнями затримки або потужності, таких як програми реального часу на мікроконтролерах або смартфонах. Він використовує більш компактну архітектуру, яка жертвує деякою точністю на користь меншого розміру моделі та швидшого виводу [12].

Архітектуру MobileNetV3 було розроблено з використанням платформо-орієнтованої стратегії NAS, що означає, що процес пошуку враховував фактичні апаратні обмеження, такі як затримка мобільних процесорів, під час проектування моделі [13].

5 Огляд датасетів

Розпізнавання об'єктів у реальному часі є критично важливою можливістю в різних програмах комп'ютерного зору, включаючи автономну навігацію, спостереження, робототехніку та доповнену реальність. Ключовим компонентом у розробці точних та ефективних моделей розпізнавання об'єктів у реальному часі є наявність високоякісних наборів даних, які відображають різноманітність, складність та динамічний характер реальних середовищ.

Серед найпоширеніших наборів даних для розпізнавання об'єктів загального призначення є набір даних COCO (Common Objects in Context). Маючи понад 330 000 зображень та понад 1,5 мільйона екземплярів об'єктів у 80 категоріях об'єктів, COCO пропонує багато анотований набір даних у натуралістичних та часто захаращених середовищах. Він підтримує як виявлення об'єктів, так і сегментацію екземплярів, що робить його модельним для розробки та порівняльного аналізу високопродуктивних моделей [14].

Аналогічно, набір даних PASCAL VOC, хоча й менший за масштабом, відрізняється стабільною якістю анотацій та своєю роллю у встановленні фундаментальних орієнтирів у цій галузі [15].

Для програм, що вимагають продуктивності в режимі реального часу на мобільних та вбудованих пристроях, особливо актуальними є набори даних, такі як ImageNet VID та YouTube-BoundingBoxes. Набір даних ImageNet VID, отриманий з більшої колекції ImageNet, містить відеопослідовності з покадровими анотаціями, що дозволяє навчати моделі, які можуть відстежувати та розпізнавати об'єкти з часом. Цей часовий вимір має вирішальне значення для розробки алгоритмів, які повинні надійно працювати в динамічних середовищах. Аналогічно, набір даних YouTube-BoundingBoxes містить мільйони позначених відеокадрів, отриманих з YouTube, що забезпечує реальну мінливість освітлення, руху та оклюзії, що є важливим для створення надійних систем розпізнавання в режимі реального часу [16].

Спеціалізовані набори даних, наприклад, BDD100K, був розроблений для підтримки програм автономного водіння. Ці набори даних пропонують анотовані відеодані, отримані з транспортних

засобів, що працюють у різних дорожніх умовах та середовищах. Вони надають достовірну інформацію для різноманітних завдань, включаючи виявлення об'єктів, відстеження, виявлення смуг руху та семантичну сегментацію. Такі набори даних є важливими для навчальних моделей, які можуть працювати в режимі реального часу з урахуванням обмежень автомобільного обладнання та критичних вимог безпеки [17].

Зрештою, моделі, обрані в цій роботі, базується на наборі даних COCO що пропонує великий, різноманітний набір зображень з детальними анотаціями для кількох категорій об'єктів, що дозволяє моделям добре узагальнювати та точно й швидко виявляти широкий спектр об'єктів у складних реальних сценах.

6 Огляд методів створення гібридних моделей

Гібридні моделі для розпізнавання зображень у реальному часі поєднують різні архітектури нейронних мереж та алгоритмічні стратегії, щоб використовувати унікальні сильні сторони кожної з них, прагнучи досягти як високої точності, так і швидкої обробки, придатної для застосувань у реальному часі. Існує кілька методів створення таких гібридних моделей, кожен з яких вирішує конкретні проблеми в задачах розпізнавання зображень.

Одним із поширених підходів є архітектурне об'єднання, де різні типи нейронних мереж інтегруються в одну модель. Наприклад, згорткові нейронні мережі, які чудово виявляють просторові ознаки із зображень, можна поєднувати з трансформаторами, які вміло моделюють довгострокові залежності та механізми уваги. Було показано, що ця гібридна архітектура ЗНМ-Трансформер підвищує як точність, так і швидкість розпізнавання об'єктів у реальному часі, особливо в складних середовищах, таких як сцени в приміщенні зі змінним освітленням та захараченим фоном. Компонент ЗНМ зазвичай обробляє початкове вилучення ознак, тоді як модуль трансформера зосереджується на уточненні цих ознак за допомогою уваги, дозволяючи моделі пріоритезувати критичну інформацію для виявлення та класифікації [18].

Об'єднання ознак – це ще один ключовий метод, де ознаки, отримані з різних моделей або модальностей, об'єднуються перед тим, як робити прогноз. Наприклад, ознаки з попередньо навченої EfficientNet (тип CNN) можна подавати в детекторну головку YOLO, поєднуючи ефективне вилучення ознак EfficientNet з можливостями виявлення в реальному часі YOLO. Цей підхід, як продемонстровано в гібриді E-YOLO (EfficientNet + YOLO), зменшує розмір моделі та обчислювальне навантаження без шкоди для точності виявлення, що робить його добре придатним для застосувань у реальному часі на периферійних пристроях [19].

Також використовується об'єднання рішень, де прогнози з кількох моделей поєднуються за допомогою таких стратегій, як голосування або зважене усереднення. Цей метод підвищує стійкість шляхом агрегування сильних сторін різних архітектур, що особливо корисно в сценаріях з різноманітними або зашумленими даними.

Крім того, гібридні моделі можуть поєднувати CNN з рекурентними нейронними мережами (RNN) для завдань, які потребують як просторового, так і часового розуміння, таких як аналіз відео. Тут CNN аналізують просторові ознаки з кожного кадру, тоді як RNN фіксують часові залежності між кадрами, забезпечуючи надійне розпізнавання в реальному часі в динамічних сценах [20].

Загалом, гібридні моделі для розпізнавання зображень у реальному часі будуються шляхом архітектурного об'єднання (наприклад, CNN-Transformer), об'єднання ознак (наприклад, EfficientNet + YOLO) та об'єднання рішень (ансамблеві методи). Ці стратегії дозволяють розробляти моделі, які є одночасно точними та ефективними, здатними відповідати вимогам програм реального часу в різних середовищах та варіантах використання.

7 Порівняння архітертур MobileNetV3 та YOLOv5

У таблиці представлено порівняльний аналіз різних моделей глибокого навчання. Результати подані у табл. 1:

Таблиця 1. Порівняння результатів продуктивності та ефективності моделей виявлення об'єктів
Table 1. Comparison of performance and efficiency results of object detection models

Модель	Точність	Розмір моделі (MB)	Середнє використання RAM (GB)	Затримка у часі
MobileNetV3-Small (TensorFlow-Lite)	0.6832	9.72	1.91	15.5 ms
MobileNetV3-Large	0.7342	15.43	2.1	21.2 ms
YOLOv5 (original, baseline)	0.8681	92.75	7.91	101.7 ms
YOLOv5 + MobileNetV3-small	0.9041	120.7	7.47	69.7 ms

Оригінальна YOLOv5 показує високу точність (0,8681), але є ресурсозатратною (92,75 МБ, 7,91 ГБ пам'яті, 101,7 мс). Гібридна модель YOLOv5 з MobileNetV3-small покращує точність до 0,9041 при зменшенні ресурсоспоживання (120,7 МБ, 7,47 ГБ, 69,7 мс).

Автономні MobileNetV3 моделі мають нижчу точність (0,6832 для Small, 0,7342 для Large), але є значно легшими та швидшими, особливо MobileNetV3-Small (1,91 ГБ, 15,5 мс). Вони підходять для застосувань з обмеженими ресурсами, де критична швидкість обробки.

MobileNetV3, відомий своєю спрощеною конструкцією та використанням згорток, що розділяються за глибиною, слугує ефективним екстрактором ознак, коли використовується як основа YOLOv5. Ця інтеграція зменшує загальний розмір моделі та обчислювальне навантаження без суттєвого погіршення продуктивності. Наприклад, заміна стандартного ядра YOLOv5 на MobileNetV3 призвела до того, що моделі стали значно меншими і швидшими, що полегшило обробку даних в реальному часі на пристроях з обмеженими ресурсами.

Загалом, результати показують, що хоча моделі MobileNetV3 є високоефективними, вони жертвують точністю. Оригінальна модель YOLOv5 пропонує високу продуктивність, але за рахунок вимог до ресурсів. Комбінація YOLOv5 з MobileNetV3-small досягає найкращого компромісу, покращуючи точність порівняно з базовою моделлю YOLOv5 і водночас підвищуючи обчислювальну ефективність, що робить її ефективним рішенням для сценаріїв, які вимагають як високої точності, так і ефективного використання ресурсів.

8 Висновки

У проведеному дослідженні здійснено комплексний порівняльний аналіз двох провідних архітектур нейронних мереж для розпізнавання об'єктів у реальному часі – YOLOv5 та MobileNetV3, а також досліджено ефективність їх гібридного поєднання. Результати експериментального дослідження дозволяють сформулювати наступні висновки.

На основі отриманих результатів можна сформулювати практичні рекомендації для різних сценаріїв застосування. Для застосувань, де критичним є мінімальне споживання ресурсів (IoT-пристрої, мобільні додатки з жорсткими обмеженнями), оптимальним вибором залишається MobileNetV3-Small. Для систем, що вимагають високої точності при помірних обмеженнях на ресурси, рекомендується використання гібридної моделі. Оригінальна YOLOv5 доцільна у випадках, коли обчислювальні ресурси не є критичним обмеженням, а пріоритетом є баланс точності та часу розробки.

Дослідження підтверджує ефективність архітектурного об'єднання як методу створення гібридних моделей для комп'ютерного зору. Використання MobileNetV3 як backbone-мережі для YOLOv5 демонструє можливість збереження переваг обох архітектур при мітигації їх недоліків. Це відкриває перспективи для подальших досліджень у напрямку покращення гібридних архітектур.

Слід зазначити певні обмеження проведеного дослідження. Дослідження проводилося на наборі даних COCO, що може обмежувати узагальнюваність результатів на інші домени. Також гібридна модель, незважаючи на покращення ефективності, все ще характеризується збільшеним розміром порівняно з окремими компонентами, що може бути критичним для деяких застосувань.

Проведене дослідження демонструє, що гібридні моделі представляють перспективний напрямок розвитку архітектур нейронних мереж для визначення об'єктів у реальному часі.

Поєднання YOLOv5 та MobileNetV3 не лише забезпечує покращення точності, але й оптимізує використання обчислювальних ресурсів, що має важливе значення для практичного впровадження систем комп'ютерного зору в реальних умовах.

Отримані результати підтверджують гіпотезу про те, що архітектурне об'єднання різних типів нейронних мереж може ефективно використовувати унікальні переваги кожної архітектури, створюючи рішення, що переважають окремі компоненти за ключовими показниками продуктивності.

Вибір архітектур нейронних мереж для розпізнавання зображень у реальному часі визначається компромісом між обчислювальною ефективністю, точністю та швидкістю.

СПИСОК ЛІТЕРАТУРИ

1. Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., & Henkel, J. (2024). A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. IEEE Access, 12, 41180-41218. https://doi.org/10.48550/arXiv.2402.15490
2. Tu, C. H., Lee, J. H., Chan, Y. M., & Chen, C. S. (2020, July). Pruning depthwise separable convolutions for mobilenet compression. In 2020 *international joint conference on neural networks* (IJCNN) (pp. 1-8). IEEE.
3. Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., ... & Shen, C. (2022). Topformer: Token pyramid transformer for mobile semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12083-12093). https://doi.org/10.48550/arXiv.2204.05525
4. YOLO Object Detection Explained: Evolution, Algorithm, and Applications. URL: https://encord.com/blog/yolo-object-detection-guide/ (дата звернення: 22.03.2025).
5. Реалізація Keras (TF backend) виявлення об'єктів, YoloV3. URL: https://github.com/xiaochus/YOLOv3/tree/master (дата звернення: 22.03.2025).
6. YOLOv4: Високошвидкісне та точне виявлення об'єктів. URL: https://docs.ultralytics.com/models/yolov4/ (дата звернення: 22.03.2025).
7. Офіційний репозиторій YOLOv5. URL: https://github.com/ultralytics/yolov5 (дата звернення: 22.03.2025).
8. Занурення в виявлення об'єктів, YOLO. URL: https://www.picsellia.com/post/a-dive-into-yolo-object-detection (дата звернення: 22.03.2025).
9. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324). https://doi.org/10.48550/arXiv.1905.02244
10. Channel Attention and Squeeze-and-Excitation Networks (SENet). URL: https://www.digitalocean.com/community/tutorials/channel-attention-squeeze-and-excitation-networks (дата звернення: 22.03.2025).
11. Функція Hardswish. URL: https://www.paddlepaddle.org.cn/documentation/docs/en/api/paddle/nn/functional/hardswish_en.html (дата звернення: 22.03.2025).
12. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861
13. MobileNet V3. URL: https://mmclassification.readthedocs.io/en/dev-1.x/papers/mobilenet_v3.html (дата звернення: 22.03.2025).
14. COCO, Набір даних для виявлення, сегментації та субтитрування великомасштабних об'єктів. URL: https://cocodataset.org/#overview (дата звернення: 22.03.2025).
15. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88, 303-338. https://doi.org/10.1007/s11263-009-0275-4

16. YouTube-Bounding Boxes Dataset. URL: https://research.google.com/youtube-bb/ (дата звернення: 22.03.2025).
17. BDD100K: A Large-scale Diverse Driving Video Database. URL: https://bair.berkeley.edu/blog/2018/05/30/bdd/ (дата звернення: 22.03.2025).
18. Agga, A., Abbou, A., Labbadi, M., El Houm, Y., & Ali, I. H. O. (2022). CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. Electric Power Systems Research, 208, 107908. DOI:10.1016/j.epsr.2022.107908
19. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., & Barnard, K. (2021). Attentional feature fusion. In Proceedings of the IEEE/CVF winter conference on applications of computer vision https://doi.org/10.48550/arXiv.2009.14082 (pp. 3560-3569).
20. Dhruv, P., & Naskar, S. (2020). Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review. Machine learning and information processing: proceedings of ICMLIP 2019, 367-381. DOI:10.1007/978-981-15-1884-3_34

REFERENCES

1. Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., & Henkel, J. (2024). A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. IEEE Access, 12, 41180-41218. https://doi.org/10.48550/arXiv.2402.15490
2. Tu, C. H., Lee, J. H., Chan, Y. M., & Chen, C. S. (2020, July). Pruning depthwise separable convolutions for mobilenet compression. In 2020 *international joint conference on neural networks* (IJCNN) (pp. 1-8). IEEE.
3. Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., ... & Shen, C. (2022). Topformer: Token pyramid transformer for mobile semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12083-12093). https://doi.org/10.48550/arXiv.2204.05525
4. YOLO Object Detection Explained: Evolution, Algorithm, and Applications. URL: https://encord.com/blog/yolo-object-detection-guide/ (date of last access: 22.03.2025).
5. Keras(TF backend) implementation of YoloV3 objects detection. URL: https://github.com/xiaochus/YOLOv3/tree/master (date of last access: 22.03.2025). [in Ukrainian]
6. YOLOv4: High-Speed and Precise Object Detection. URL: https://docs.ultralytics.com/models/yolov4/ (date of last access: 22.03.2025).
7. YOLOv5 Official Repository. URL: https://github.com/ultralytics/yolov5 (date of last access: 22.03.2025). [in Ukrainian]
8. A dive into YOLO object detection. URL: https://www.picsellia.com/post/a-dive-into-yolo-object-detection (date of last access: 22.03.2025).
9. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324). https://doi.org/10.48550/arXiv.1905.02244
10. Channel Attention and Squeeze-and-Excitation Networks (SENet). URL: https://www.digitalocean.com/community/tutorials/channel-attention-squeeze-and-excitation-networks (date of last access: 22.03.2025).
11. Function Hardswish. URL: https://www.paddlepaddle.org.cn/documentation/docs/en/api/paddle/nn/functional/hardswish_en.html (date of last access: 22.03.2025). [in Ukrainian]
12. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861
13. MobileNet V3. URL: https://mmclassification.readthedocs.io/en/dev-1.x/papers/mobilenet_v3.html (date of last access: 22.03.2025).
14. COCO, large-scale object detection, segmentation, and captioning dataset. URL: https://cocodataset.org/#overview (date of last access: 22.03.2025).

15. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88, 303-338. https://doi.org/10.1007/s11263-009-0275-4
16. YouTube-Bounding Boxes Dataset:. URL: https://research.google.com/youtube-bb/ (date of last access: 22.03.2025).
17. BDD100K: A Large-scale Diverse Driving Video Database. URL: https://bair.berkeley.edu/blog/2018/05/30/bdd/ (date of last access: 22.03.2025).
18. Agga, A., Abbou, A., Labbadi, M., El Houm, Y., & Ali, I. H. O. (2022). CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. Electric Power Systems Research, 208, 107908. DOI:10.1016/j.epsr.2022.107908
19. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., & Barnard, K. (2021). Attentional feature fusion. In Proceedings of the IEEE/CVF winter conference on applications of computer vision https://doi.org/10.48550/arXiv.2009.14082 (pp. 3560-3569).
20. Dhruv, P., & Naskar, S. (2020). Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review. Machine learning and information processing: proceedings of ICMLIP 2019, 367-381. DOI:10.1007/978-981-15-1884-3_34

Yasinskyi Yaroslav
Ph.D student;
V.N. Karazin Kharkiv National University
Svobody Sq 4, Kharkiv, Ukraine, 61022
e-mail: yaroslav.yasinskyi@karazin.ua;
https://orcid.org/0009-0008-0460-5687

Bakumenko Nina
Candidate of Technical Sciences; Associate Professor of Computer Systems and Robotics Department, Education and Research Institute of Computer Sciences and Artificial Intelligence;
V.N. Karazin Kharkiv National University
Svobody Sq 4, Kharkiv, Ukraine, 61022
e-mail: n.bakumenko@karazin.ua;
https://orcid.org/0000-0003-3496-7167

Comparative analysis of YOLOv5 and MobileNetV3 models for real-time image recognition

Relevance: With the growing need for fast and accurate real-time object recognition, especially for mobile and embedded systems, the question of choosing the optimal AI models arises. Comparisons of lightweight and high-precision architectures such as YOLOv5 and MobileNetV3 are important for developing efficient computer vision systems and exploring the principles of hybrid model construction.
Purpose: Comparison of the YOLOv5 and MobileNetV3 architectures to analyze the efficiency for real-time object recognition applications, and to confirm that hybrid models can improve the efficiency of these tasks.
Research methods: image preprocessing methods, deep neural network training methods, measurement of accuracy, processing speed, and resource usage; comparative analysis of results to assess model effectiveness.
Results: An experimental study showed that YOLOv5 demonstrates better overall accuracy on the COCO test suite, but requires more computing resources. MobileNetV3, on the other hand, provides faster output and efficient functioning on low-power devices, sacrificing accuracy in part. As such, both models have proven their suitability for real-world applications, and the choice between them depends on the specific balance between speed, accuracy, and platform limitations. Combining these models gives better results in object recognition, although this may increase the size of the model itself and resource consumption.
Conclusions: As a result of the study, the YOLOv5, MobileNetV3 and hybrid models for the object recognition problem were compared. The hybrid model demonstrated better accuracy and balance between processing speed and resource utilization than individual models. This indicates the feasibility of using hybrid approaches to improve the efficiency of computer vision systems in real conditions. Therefore, the hybrid model is a promising direction for further research and practical implementation.
Keywords: *image recognition, computer vision, hybrid model, CNN, YOLOv5, MobileNetV3*

.

Надійшла у першій редакції 21.03.2025, в останній - 17.05.2025.
The first version has been received on 21.03.2025, the final version - on 17.05.2025.