ВІСНИК ПРИАЗОВСЬКОГО ДЕРЖАВНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ
2026р. Серія: Технічні науки Вип. 53. Т. 1
p-ISSN: 2225-6733; e-ISSN: 2519-271X

УДК 004.93:004.8:004.032.26 DOI: 10.31498/2225-6733.53.1.2026.359780

РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ ЗА ДОПОМОГОЮ ГІБРИДНОГО ПІДХОДУ ДО НА ОСНОВІ МУЛЬТИМОДАЛЬНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ТА ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

Чичкарьов Є.А. *д-р техн. наук, професор, Державний університет інформаційно-комунікаційних технологій, м. Київ, ORCID: https://orcid.org/0000-0002-4362-5129, e-mail: e.chychkarov@duikt.edu.ua;*

Семенов О.В. *аспірант, Державний університет інформаційно-комунікаційних технологій, м. Київ, ORCID: https://orcid.org/0009-0005-9749-3343, e-mail: sasha.sem.fti@gmail.com*

У роботі розглядається актуальна наукова проблема розпізнавання зображень в умовах обмежень сучасних нейромережевих архітектур. Обґрунтовано доцільність застосування гібридного підходу, що поєднує структурну точність згорткових нейронних мереж (CNN) із семантичною гнучкістю мультимодальних великих мовних моделей (MLLM). Метою дослідження є розробка та апробація методу, що дозволяє усунути обмеженість категорій класичних систем розпізнавання та схильність мультимодальних моделей до формування неіснуючих об'єктів. У ході дослідження застосовано методи ієрархічної класифікації, узгодженого голосування (Hybrid Voting) та формування текстових інструкцій (Prompt Engineering). Для програмної реалізації використано моделі MobileNetV2 та локальні MLLM (серії Gemma 3, LLaVA) у хмарному середовищі Google Colab із застосуванням інструменту Ollama. Валідація підходу здійснювалася на датасетах MNIST, CIFAR-10 та Flowers102. Ключовим результатом роботи є встановлення того, що гібридна архітектура забезпечує вищу достовірність розпізнавання порівняно з автономним використанням моделей. Зокрема, на складному датасеті Flowers102 зафіксовано приріст точності до 92,3%. Визначено, що використання CNN як базового шару дозволяє нівелювати слабкість мовних моделей у роботі з низькорозмірними даними, тоді як MLLM забезпечує глибоку семантичну інтерпретацію в умовах візуальної схожості об'єктів. Наукова новизна полягає у розробці динамічного механізму взаємодії нейромережевих архітектур, що дозволяє оптимізувати обчислювальні ресурси та підвищити надійність систем розпізнавання в умовах невизначеності. Практичне значення роботи підтверджується можливістю впровадження запропонованого конвеєра в системи пакетної обробки даних та інтерактивні застосунки.

Ключові слова: *розпізнавання зображень; згорткові нейронні мережі; мультимодальні великі мовні моделі; гібридний підхід; MobileNetV2; Gemma 3; розпізнавання без попереднього навчання; Google Colab; Ollama.*

Постановка проблеми

Сучасний етап розвитку штучного інтелекту характеризується стрімкою парадигмальною зміною: переходом від вузькоспеціалізованих моделей до універсальних мультимодальних систем. Попри значні успіхи згорткових нейронних мереж (CNN) у задачах класифікації, вони залишаються функціонально обмеженими системами, здатними оперувати лише тими категоріями об’єктів, що були визначені на етапі навчання. Водночас поява мультимодальних великих мовних моделей (MLLM), таких як Gemini, GPT-4V та LLaVA, відкрила нові можливості для розпізнавання без попереднього навчання (Zero-shot). Проте їх впровадження виявило низку специфічних викликів, зокрема схильність до «галюцинацій» та високу обчислювальну складність [1–4].

Порівняльний аналіз можливостей CNN та MLLM вказує на фундаментальну наукову проблему: механізми опрацювання візуальної інформації в цих архітектурах суттєво розрізняються. Моделі типу MobileNetV2 або ResNet [5-6] орієнтовані на прецизійне виділення локальних ознак – текстур, контурів та геометричних патернів. Хоча такі архітектури демонструють високу ефективність у задачах із фіксованою множиною класів, відсутність широкого когнітивного контексту обмежує їхню здатність до семантичної інтерпретації об’єктів, які мають суттєві візуальні відмінності від еталонних навчальних зразків.

На противагу цьому, мультимодальні моделі (зокрема серії Gemini та Gemma 3) базуються на принципах контекстуального висновку та широкої семантичної генералізації. Однак їхні архітектурні особливості зумовлюють виникнення явища «зміщення до глобального контексту» (Global context bias). У практичних сценаріях це призводить до домінування лінгвістичних імовірностей над безпосереднім візуальним аналізом деталей. Як наслідок, спостерігається зниження точності в задачах дрібнозернистої класифікації (наприклад, у наборах даних Flowers102 або MNIST), де критично важливим є аналіз морфологічних мікроознак, а не загальна семантика сцени [1–4].

Наявність цих розбіжностей створює передумови для розробки гібридного підходу, метою якого є конвергенція структурної точності згорткових мереж та семантичної гнучкості мультимодальних моделей для підвищення загальної надійності систем розпізнавання.

У реальних системах (IoT, мобільні додатки) використання MLLM для кожного вхідного кадру є економічно та технічно недоцільним через величезні витрати пам’яті та енергії. З іншого боку, використання лише CNN не забезпечує необхідної адаптивності в умовах динамічної зміни наборів даних. Виникає

потреба у створенні динамічного механізму перемикання (gating mechanism), який би дозволяв системі самостійно вирішувати, коли достатньо швидкого висновку CNN, а коли необхідне залучення інтелектуального арбітражу MLLM.

Аналіз останніх досліджень та публікацій

Протягом останнього десятиліття згорткові нейронні мережі (CNN) домінували в архітектурах комп'ютерного зору завдяки їхній здатності ефективно виділяти просторові ієрархії ознак. Фундаментальні праці, зокрема дослідження [5, 6] щодо залишкових мереж (ResNet), заклали основу для створення глибоких моделей, здатних до навчання на тисячах категорій об'єктів. Проте деякі роботи [7] продемонстрували суттєве обмеження таких архітектур: процес розпізнавання в них суворо обмежений переліком міток, які були представлені на етапі навчання. Це створює проблему при обробці динамічних або нестандартних візуальних даних, де виникає потреба у залученні зовнішніх знань для інтерпретації об'єктів, що виходять за межі фіксованого набору категорій.

Важливим етапом подолання обмежень класичних CNN стала розробка концепції контрастивного навчання зображень та мови (CLIP). Автори [7] довели, що навчання візуальних моделей на основі природної мови дозволяє досягти високої точності в задачах «Zero-shot» класифікації (класифікація без попереднього навчання на конкретних мітках). Модель CLIP навчилася зіставляти візуальні патчі з текстовими описами в спільному латентному просторі, що дозволило інтерпретувати зображення як семантичні сутності, а не просто як набори пікселів. Це відкрило шлях до створення мультимодальних великих мовних моделей (MLLM), які здатні «міркувати» над візуальним контентом.

Подальший розвиток мультимодальних систем пов’язаний із впровадженням методів візуального налаштування інструкцій (Visual Instruction Tuning). У дослідженні [1-2] була представлена модель LLaVA, яка інтегрує візуальний енкодер CLIP із великою мовною моделлю через проекційну матрицю. Це дозволило перетворити завдання розпізнавання зображень на діалоговий процес. Проте аналіз публікацій останніх років виявив нову проблему: такі моделі схильні до надмірного покладання на лінгвістичний контекст, що іноді призводить до ігнорування дрібних візуальних деталей, які класична згорткова мережа розпізнала б безпомилково.

Ґрунтовний огляд сучасного стану галузі мультимодальних великих мовних моделей наведено в [6]. Автори [8] систематизували архітектурні підходи, методи навчання на інструкціях та стратегії вирівнювання модальностей. Автори [9] підкреслили вразливість мультимодальних систем до явища «галюцинацій». За даними [8, 9] MLLM можуть генерувати впевнені, але фактично невірні описи зображень, особливо якщо візуальна сцена є зашумленою або об'єкти мають низьку роздільну здатність (наприклад, у наборі даних MNIST). Це створює потребу в гібридизації, де CNN виступає як «стабілізатор», що забезпечує структурну точність і перевіряє семантичні припущення мовної моделі.

Перенесення навчання є ключовим механізмом, що дозволяє використовувати знання, отримані на великих датасетах (ImageNet), для задач із малими вибірками [10, 11]. У гібридному підході CNN виступає в ролі провідника ознак (feature extractor) з замороженими ваговими коефіцієнтами, тоді як LLM забезпечує семантичну інтерпретацію отриманих ознак. Дослідження [10, 11] підтвердили, що поєднання самонавчання та перенесення навчання суттєво підвищує стійкість до розподільного зсуву, що особливо критично для тонкозернистої класифікації.

Сучасні тенденції вказують на ефективність поєднання легких архітектур, таких як MobileNetV2, з потужними хмарними моделями. Наприклад, в роботі [12] наведено модель Flamingo, яка містить наступні ключові архітектурні інновації: (i) поєднання потужних попередньо навчених моделей, що працюють лише з візуалізацією, та моделей, що працюють лише з мовою, (ii) обробки послідовностей довільно чергуваних візуальних та текстових даних, та (iii) безперешкодного завантаження зображень або відео як вхідних даних. На думку авторів [12], модель Flamingo може досягти нового рівня техніки за допомогою навчання за кілька спроб, просто підказуючи моделі приклади, що відповідають конкретним завданням.

Гібридні системи дозволяють використовувати переваги обох підходів: MobileNetV2 забезпечує швидке виділення локальних ознак, а велика мовна модель додає семантичну глибину та здатність до узагальнення. Такий підхід дозволяє оптимізувати обчислювальні ресурси та підвищити загальну надійність систем розпізнавання в умовах невизначеності.

Мета статті

Мета цієї роботи полягає у розробці та апробації гібридного підходу, який забезпечує поєднання структурної точності згорткових нейронних мереж із контекстуальним аналізом мультимодальних великих мовних моделей.

Для досягнення поставленої мети визначено наступні завдання:

- Розробка стратегії узгодженого голосування, за якої результати роботи згорткової мережі піддаються перевірці за допомогою спеціально сформованих текстових інструкцій у мовній моделі.
- Аналіз логічних збоїв у процесах розпізнавання обох типів архітектур та побудова матриці помилок для їх подальшого усунення.
- Оптимізація процесів виконання та взаємодії мультимодальних моделей у хмарних середовищах без втрати достовірності результатів класифікації.

Матеріали та методи

Об'єктом дослідження є процес автоматичної класифікації та семантичної інтерпретації візуальних даних у гетерогенних середовищах.

Предметом дослідження є архітектурні рішення, алгоритми гібридної взаємодії між локальними екстракторами ознак (на основі CNN) та системами контекстуального виводу (на основі MLLM), а також методи оптимізації їхньої спільної роботи в хмарних інфраструктурах.

Для досягнення мети роботи було застосовано комплексний підхід, що базується на наступних методах:

- Метод ієрархічної класифікації: реалізація дворівневої структури прийняття рішень, де перший рівень забезпечує швидку структурну обробку, а другий – поглиблений семантичний аналіз.
- Метод узгодженого голосування (Hybrid Voting): алгоритм порівняння результатів роботи різних архітектур для мінімізації логічних помилок та подолання проблеми «закритого світу» класичних мереж.
- Метод формування текстових інструкцій (Prompt Engineering): використання спеціально розроблених запитів для спрямування уваги мультимодальної моделі на специфічні ознаки зображення (колір, морфологія, просторове розміщення).
- Метод статистичного аналізу: побудова матриць помилок (Confusion Matrices) та розрахунок метрик точності (Accuracy) для оцінки ефективності розроблених алгоритмів.

Запропонована система базується на спільному використанні двох типів моделей:

1. Згортковий рівень (екстракція ознак): Використовується модель MobileNetV2, яка завдяки архітектурі з інвертованими залишковими блоками (Inverted Residuals) забезпечує високу швидкість виконання в умовах обмежених ресурсів. Вона відповідає за первинну ідентифікацію геометричних паттернів.
2. Мультимодальний рівень (арбітраж): Використовується локальна LLM з переліку Gemma3:4b, Gemma3:12b, Llama3.2-Vision. Ця ланка активується для верифікації результатів CNN у складних випадках або для розпізнавання об'єктів поза навчальною вибіркою.

У гібридній системі MobileNetV2 використовувалась без фінального класифікаційного шару (global average pooling + dense). Вихідний вектор ознак розмірності 1280 передавався через softmax-проекцію у простір попередньо навчених класів ImageNet (1000 класів) для отримання вектора ймовірностей. Для формування вхідного контексту LLM використовувались п'ять класів, для яких отримано найвищі значення ймовірності належності.

Покращення поведінки гібридної системи було досягнуто завдяки застосуванню трансферного навчання. Архітектура трансферного навчання для MobileNetV2 базується на використанні попередньо навченої на ImageNet базової мережі як потужного екстрактора візуальних ознак, де всі шари, крім вихідного, залишаються статичними для збереження універсальних знань про форми та текстури. Замість оригінального верхнього блоку класифікатора додається шар Global Average Pooling для перетворення просторових ознак у вектор, після чого слідують нові повнозв'язні шари (Dense) з активацією ReLU та регуляризацією Dropout. Фінальний шар з активацією Softmax адаптується під кількість класів конкретного датасету (наприклад, 10 для MNIST або для Cifar10), що дозволяє моделі ефективно навчатися навіть на невеликих вибірках даних, мінімізуючи обчислювальні витрати.

Дослідження та програмна реалізація проводилися з використанням наступних інструментів:

- Середовище розробки: Хмарна платформа Google Colab, що надає доступ до прискорювачів (GPU T4) для виконання нейронних мереж.
- Бібліотеки машинного навчання: TensorFlow 2.x та Keras – для реалізації та навчання згорткових шарів.
- TensorFlow Datasets (TFDS) – для завантаження та стандартизації наборів даних.
- Засоби роботи з мультимодальними моделями (MLLM): Ollama – для локального розгортання та тестування моделей серії Gemma та LLaVA.
- Обробка даних: Бібліотеки NumPy, Pandas та Matplotlib для статистичної обробки та візуалізації результатів.

Ollama [13-14] – це інструмент із відкритим кодом для локального розгортання великих мовних моделей. Він надає REST API, сумісний з протоколом OpenAI, що спрощує інтеграцію з існуючими застосунками. Ollama підтримує квантизацію моделей (4-bit, 8-bit GGUF), яка критично знижує вимоги до оперативної пам'яті GPU.

Google Colab [15] надає до 12 ГБ RAM та до 15 ГБ відеопам'яті GPU (Tesla T4, A100) у безкоштовній версії. Це забезпечує достатній ресурс для розгортання моделей розміром до 12b параметрів у 4-bit квантизації. Ollama встановлюється однією командою у bash-клітинці Colab і запускається у фоновому режимі як daemon-процес.

Типовий порядок розгортання LLM в Google Colab за допомогою Ollama включає такі кроки:

1. Встановлення Ollama:

```
!curl -fsSL https://ollama.com/install.sh | sh
```

2. Запуск сервера у фоні:

```
subprocess.Popen(["ollama", "serve"])
```

3. Завантаження моделі:

```
!ollama pull gemma3:12b
# або mistral:7b, llama3.2, llava:7b
```

Для валідації гібридного підходу використано три типи датасетів:

1. MNIST: для оцінки здатності розпізнавати прості рукописні символи за умов низької роздільної здатності.

p-ISSN: 2225-6733; e-ISSN: 2519-271X

2. CIFAR-10: для тестування розпізнавання об’єктів реального світу в умовах обмеженої кількості пікселів.

3. Flowers102: для аналізу точності дрібнозернистої класифікації природничих об’єктів з високою міжкласовою схожістю.

Датасет Flowers102 є значно складнішим за CIFAR-10 через велику кількість класів та їх схожість. Семантична компетентність LLM у ботанічних знаннях дає найбільший ефект саме на цьому датасеті.

Ключова перевага LLM полягає у здатності до багатокрокового міркування та використання фонових знань, недоступних CNN. Зокрема, LLM може врахувати сезонність, географічні особливості виду та типові поєднання кольорів, що підвищує точність тонкозернистої класифікації порівняно з ізольованим використанням CNN.

Виклад основного матеріалу

Дослідження проводилося шляхом послідовного тестування трьох конфігурацій систем: класичної згорткової нейронної мережі на базі MobileNetV2, автономної мультимодальної моделі (MLLM) та запропонованого гібридного підходу.

Результати класифікації на контрольних вибірках (MNIST, CIFAR-10, Flowers102) наведено у таблиці 1.

Таблиця 1

Порівняльна ефективність розпізнавання

Набір даних	Показник Accuracy, %		
	MobileNetV2 (CNN)	Gemma3:12b (MLLM)	Гібридний підхід
MNIST	98,4	65,2	98,7
CIFAR-10	89,2	78,5	90,4
Flowers102	74,5	87,1	92,3

Ілюстрацію розпізнавання зображення з набору даних Flowers102 наведено на рис. 1.

Рис. 1 – Приклад розпізнавання зображення з набору даних Flowers102 за допомогою гібридного підходу

Результати на рис. 1 і в таблиці 1 було отримано з використанням попереднього трансферного навчання і динамічного промпту, який містив перелік назв класів. Для скорочення часу виконання класифікації тестових зображень MLLM не викликалась для кожного зображення. Вона використовувалася як арбітр лише тоді, коли впевненість CNN падала нижче 75%.

Аналіз результатів на низькорозмірних даних (MNIST, CIFAR-10)

Процес обробки даних MNIST виявив суттєву закономірність: мультимодальні моделі демонструють нижчу точність на монохромних символах низької роздільної здатності. Це обґрунтовується специфікою візуальних енкодерів (ViT), які при апсемплінгу зображень розміром 28x28 пікселів схильні до втрати ключових геометричних ознак.

Натомість згорткова мережа MobileNetV2 демонструє стабільно високу точність, ефективно розпізнаючи структурні примітиви (лінії, дуги). У гібридному підході використання CNN як базового шару дозволило нівелювати слабкість мовної моделі у роботі з абстрактними символами, що підтверджується зростанням загальної точності до 98,7%.

Інтерпретація результатів дрібнозернистої класифікації (Flowers102)

Найбільш виражена синергія спостерігається на наборі даних Flowers102. У процесі аналізу виявлено, що класична CNN часто припускається помилок другого роду (хибне сприйняття схожих текстур), тоді як мультимодальна модель Gemini використовує контекстуальні знання для диференціації об'єктів.

Ключові тенденції, виявлені під час аналізу:

- Контекстуальна корекція: У випадках, коли CNN ідентифікувала об'єкт із низькою впевненістю, залучення мовної моделі через механізм формування текстових інструкцій дозволило уточнити класифікацію за непрямими ознаками (наприклад, форма листя або середовище зростання).
- Зниження рівня галюцинацій: Гібридна система використовує результати CNN для обмеження простору пошуку відповідей мовної моделі. Це значно зменшує ймовірність генерації назв класів, які відсутні в поточному контексті дослідження.

Взаємозв’язок між складністю моделі та точністю виводу

У ході дослідження було зафіксовано кореляцію між кількістю параметрів мультимодальної моделі та якістю семантичного арбітражу. Використання моделі Gemma 3:12b порівняно з 4b версією підвищує точність на складному датасеті Flowers на 8,5%, проте збільшує час обробки одного запиту в хмарному середовищі Google Colab приблизно в 2,4 раза.

p-ISSN: 2225-6733; e-ISSN: 2519-271X

Обґрунтування гібридного підходу через матрицю помилок

Побудовані матриці помилок (Confusion Matrices) свідчать про те, що гібридний підхід ефективно перекриває зони вразливості обох типів архітектур. Там, де CNN помиляється через візуальну схожість (наприклад, плутаючи класи «кіт» та «собака» у CIFAR-10), мультимодальна модель додає логічний фільтр. Там, де мовна модель припускається галюцинацій через низьку роздільну здатність пікселів, згорткова мережа забезпечує жорстку структурну верифікацію.

Результати та їх обговорення

Експериментальні дослідження підтвердили гіпотезу про те, що поєднання структурного аналізу згорткових нейронних мереж (CNN) із семантичними можливостями мультимодальних моделей (MLLM) дозволяє досягти стабільно високих показників точності на гетерогенних наборах даних.

Зведеним результатом випробувань став показник загальної точності (Accuracy), який для гібридного підходу виявився вищим на 5,2–17,8% порівняно з автономним використанням моделей, залежно від складності візуальної сцени. Найбільш виражений приріст ефективності зафіксовано на наборі даних Flowers102, де точність зросла з 74,5% (чиста CNN) до 92,3% (гібрид).

Порівняння Gemma 3:4b та Gemma 3:12b виявило наступні закономірності:

- Модель 4b: Демонструє оптимальний баланс між швидкістю та використанням пам'яті в середовищі Google Colab, проте схильна до помилок у розпізнаванні дрібних морфологічних ознак рослин.

- Модель 12b: Показує значно вищу стійкість до візуальних шумів та кращу здатність до контекстуального висновку. Зокрема, у завданнях розпізнавання цифр (MNIST) модель 12b рідше припускалася логічних помилок при інтерпретації розмитих контурів.

В результаті експериментів зі створеними моделями встановлено здатність гібридного підходу до нівелювання явища «галюцинацій», притаманного великим мовним моделям.

1. Механізм стримування: Застосування результатів MobileNetV2 як попереднього фільтра дозволило обмежити простір пошуку для MLLM. Це змусило мовну модель обирати відповіді виключно з валідного переліку категорій датасету, що повністю усунуло помилки типу «вигадування несучих об'єктів».

2. Проблема низької роздільної здатності: Обговорення підтвердило, що MLLM (навіть версії 12b) все ще стикаються з труднощами при обробці зображень 28x28 пікселів (MNIST). У таких випадках гібридна система автоматично надавала пріоритет висновку CNN, що забезпечило структурну достовірність результату.

Точність розпізнавання вдалося значно покращити за рахунок спільного використання трансферного навчання і динамічного промпту, який містив промпт з переліком можливих назв класів.

Аналіз часових витрат показав, що запропонована стратегія узгодженого голосування є обчислювально виправданою. Оскільки важка мультимодальна модель залучалася лише у випадках низької впевненості MobileNetV2 (поріг $\tau < 0{,}85$), середній час обробки одного зображення збільшився лише на 12–15% порівняно з базовою CNN, що є прийнятним для більшості систем інтелектуального моніторингу.

Основні часові витрати гібридного конвеєра зосереджені в модулі LLM-інтерпретації: MobileNetV2 виконує класифікацію зображення менш ніж за 5 мс на GPU, тоді як виклик LLM займає від 8 до 55 с залежно від моделі та апаратного забезпечення. Тому гібридна система не підходить для застосувань реального часу (зокрема для обробки відеопотоків зі швидкістю понад 1 кадр/с), але є доцільною для пакетної обробки даних та інтерактивних застосунків, де виправлення помилок має вищий пріоритет, ніж швидкість роботи.

Результати дослідження доводять, що майбутнє систем комп'ютерного зору лежить не в нарощуванні потужності окремих архітектур, а в їхній інтеграції. Гібридизація дозволяє подолати проблему «закритого світу» традиційних мереж, надаючи їм здатність до семантичного розуміння об'єктів через природну мову, зберігаючи при цьому математичну точність аналізу піксельних структур.

Висновки

1. Систематизація отриманих експериментальних даних підтверджує гіпотезу, що конвергенція структурно-орієнтованих (CNN) та семантично-орієнтованих (MLLM) методів забезпечує вищу достовірність розпізнавання (на 5,2–17,8% вище), ніж автономне використання цих технологій.

2. Гібридна стратегія забезпечує ефективне розв'язання проблеми обмеженості категорій класичних мереж та дозволяє мінімізувати ймовірність виникнення неіснуючих об'єктів у відповідях великих мовних моделей. Зокрема, застосування згорткових нейронних мереж як попереднього фільтра звужує межі пошуку для мультимодальних моделей до переліку підтверджених категорій, що усуває хибні ідентифікації.

3. Запропонована стратегія узгодженого голосування є обчислювально виправданою: залучення ресурсомісткої мультимодальної моделі лише у випадках низької впевненості CNN (поріг $\tau < 0{,}85$) збільшує середній час обробки зображення лише на 12–15%.

4. Встановлено, що використання Gemma 3:12b порівняно з версією 4b підвищує точність на складних природничих об'єктах на 8,5%, хоча й потребує у 2,4 раза більше часу на обробку запиту.

5. Хмарна інфраструктура Google Colab із застосуванням інструментів квантизації (Ollama) є ефективним та доступним середовищем для розгортання

гетерогенних систем із параметрами до 12b, що не потребує використання локальних суперкомп'ютерів.

Перелік використаних джерел

[1] Visual instruction tuning / Liu H., Li C., Wu Q., Lee Y. J. *Advances in Neural Information Processing Systems*. 2023. Pp. 34892–34916. DOI: https://doi.org/10.48550/arXiv.2304.08485.
[2] GPT-4 technical report / J. Achiam et al. arXiv preprint. arXiv:2303.08774. 2023. DOI: https://doi.org/10.48550/arXiv.2303.08774.
[3] Gemma 3 technical report / G. T. Kamath et al. arXiv preprint. arXiv:2503.19786. 2025. DOI: https://doi.org/10.48550/arXiv.2503.19786.
[4] MiniGPT-4: Enhancing vision-language understanding with advanced large language models / D. Zhu et al. arXiv preprint. arXiv:2304.10592. 2023. DOI: https://doi.org/10.48550/arXiv.2304.10592.
[5] MobileNetV2: Inverted residuals and linear bottlenecks / M. Sandler et al. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18-23 June 2018. Pp. 4510–4520. DOI: https://doi.org/10.1109/CVPR.2018.00474.
[6] Deep residual learning for image recognition / He K., Zhang X., Ren S., Sun J. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27-30 June 2016. Pp. 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90.
[7] Learning transferable visual models from natural language supervision / A. Radford et al. *Proceedings of the International Conference on Machine Learning*, 18-24 July 2021. Vol. 139. Pp. 8748–8763. DOI: https://doi.org/10.48550/arXiv.2103.00020.
[8] A survey on multimodal large language models / S. Yin et al. arXiv preprint. arXiv:2306.13549. 2024. DOI: https://doi.org/10.48550/arXiv.2306.13549.
[9] Hallucination of multimodal large language models: A survey / Z. Bai et al. arXiv preprint. arXiv:2404.18930. 2024. DOI: https://doi.org/10.48550/arXiv.2404.18930.
[10] BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models / Li J., Li D., Savarese S., Hoi S. *Proceedings of the International Conference on Machine Learning*, Honolulu, Hawaii, USA, 23–29 July 2023. DOI: https://doi.org/10.48550/arXiv.2301.12597.
[11] Rethinking pre-training and self-training / B. Zoph et al. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 6–12 December 2020. Vol. 33. Pp. 3833–3845. DOI: https://doi.org/10.48550/arXiv.2006.06882.
[12] Flamingo: A visual language model for few-shot learning / J. Alayrac et al. arXiv preprint. arXiv:2204.14198. 2022. DOI: https://doi.org/10.48550/arXiv.2204.14198.
[13] Ollama Team. Ollama: Get up and running with large language models. URL: https://ollama.com (дата звернення: 15.09.2025).
[14] Vake D., Vičič J., Tošić A. Hive: A secure, scalable framework for distributed Ollama inference. *SoftwareX*. 2025. Vol. 30. Article 102183. DOI: https://doi.org/10.1016/j.softx.2025.102183.
[15] Google. Google Colaboratory – Free GPU/TPU for machine learning. URL: https://colab.research.google.com (дата звернення: 15.09.2025).

IMAGE RECOGNITION USING A HYBRID APPROACH BASED ON MULTIMODAL LARGE LANGUAGE MODELS AND CONVOLUTIONAL NEURAL NETWORKS

Chychkarov Y.A. *D.Sc. (Engineering), professor, State University of Information and Communication Technologies, Kyiv, ORCID: https://orcid.org/0000-0002-4362-5129, e-mail: e.chychkarov@duikt.edu.ua;*

Semenov O.V. *postgraduate student, State University of Information and Communication Technologies, Kyiv, ORCID: https://orcid.org/0009-0005-9749-3343, e-mail: sasha.sem.fti@gmail.com*

The article considers the current scientific problem of image recognition under the constraints of modern neural network architectures. The feasibility of using a hybrid approach that combines the structural accuracy of convolutional neural networks (CNN) with the semantic flexibility of multimodal large language models (MLLM) is substantiated. The aim of the study is to develop and test a method that allows eliminating the limitations of categories of classical recognition systems and the tendency of multimodal models to form non-existent objects. The study used methods of hierarchical classification, consensus voting (Hybrid Voting) and text instruction generation (Prompt Engineering). For software implementation, MobileNetV2 models and local MLLMs (Gemma 3, Llama3 series) were used in the Google Colab cloud environment using the Ollama tool. The validation of the approach was carried out on the MNIST, CIFAR-10 and Flowers102 datasets. The key result of the work is the establishment that the hybrid architecture provides higher recognition reliability compared to the autonomous use of models. In particular, on the complex Flowers102 dataset, an increase in accuracy of up to 92.3% was recorded. It was determined that the use of CNN as a base layer allows to level the weakness

of language models in working with low-dimensional data, while MLLM provides deep semantic interpretation in conditions of visual similarity of objects. The scientific novelty lies in the development of a dynamic mechanism for the interaction of neural network architectures, which allows to optimize computing resources and increase the reliability of recognition systems under conditions of uncertainty. The practical significance of the work is confirmed by the possibility of implementing the proposed pipeline in batch data processing systems and interactive applications.
Keywords: *image recognition; convolutional neural networks; multimodal large language models; hybrid approach; MobileNetV2; Gemma 3; recognition without prior training; Google Colab; Ollama.*

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023. **doi: 10.48550/arXiv.2304.08485**.

[2] J. Achiam et al., "GPT-4 technical report," 2023, arXiv:2303.08774. **doi: 10.48550/arXiv.2303.08774**.

[3] G. T. Kamath et al., "Gemma 3 technical report," 2025, *arXiv:2503.19786*. **doi: 10.48550/arXiv.2503.19786**.

[4] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*. **doi: 10.48550/arXiv.2304.10592**.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 18–23, 2018, pp. 4510–4520. **doi: 10.1109/CVPR.2018.00474**.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 27–30, 2016, pp. 770–778. **doi: 10.1109/CVPR.2016.90**.

[7] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. of the Int. Conf. on Machine Learning*, July 18–24, 2021, vol. 139, pp. 8748–8763. **doi: 10.48550/arXiv.2103.00020**.

[8] S. Yin et al., "A survey on multimodal large language models," 2024, *arXiv:2306.13549*. **doi: 10.48550/arXiv.2306.13549**.

[9] Z. Bai et al., "Hallucination of multimodal large language models: A survey," 2024, *arXiv:2404.18930*. **doi: 10.48550/arXiv.2404.18930**.

[10] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. of the Int. Conf. on Machine Learning*, Honolulu, Hawaii, USA, July 23–29, 2023. **doi: 10.48550/arXiv.2301.12597**.

[11] B. Zoph et al., "Rethinking pre-training and self-training," in *Proc. of the 34th Conf. on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, December 6–12, 2020, vol. 33, pp. 3833–3845. **doi: 10.48550/arXiv.2006.06882**.

[12] J. Alayrac et al., "Flamingo: A visual language model for few-shot learning," 2022, *arXiv:2204.14198*. **doi: 10.48550/arXiv.2204.14198**.

[13] Ollama Team. Ollama: Get up and running with large language models. [Online]. Available: https://ollama.com. Accessed on: September 15, 2025.

[14] D. Vake, J. Vičič, and A. Tošić, "Hive: A secure, scalable framework for distributed Ollama inference," *SoftwareX*, vol. 30, article 102183, 2025. **doi: 10.1016/j.softx.2025.102183**.

[15] Google. Google Colaboratory – Free GPU/TPU for machine learning. [Online]. Available: https://colab.research.google.com. Accessed on: September 15, 2025.

Стаття надійшла 11.01.2026
Стаття прийнята 15.02.2026
Стаття опублікована 26.03.2026

Цитуйте цю статтю як: Чичкарьов Є. А., Семенов О. В. Розпізнавання зображень за допомогою гібридного підходу до на основі мультимодальних великих мовних моделей та згорткових нейронних мереж. *Вісник Приазовського державного технічного університету. Серія: Технічні науки*. 2026. Вип. 53, том 1. С. 85–91. DOI: https://doi.org/10.31498/2225-6733.53.1.2026.359780.