ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Метод розпізнавання зображень для подальшої генерації текстів на основі алгоритмів штучного інтелекту»

на здобуття освітнього ступеня магістра
зі спеціальності 121 Інженерія програмного забезпечення
освітньо-професійної програми «Інженерія програмного забезпечення»

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Олександр СОРОКА
(підпис)

Виконав: здобувач вищої освіти групи ПДМ-63
Олександр СОРОКА

Керівник: Ірина ЩЕРБИНА
канд. техн. наук, доц.

Рецензент: _____
науковий ступінь, вчене звання *Ім'я, ПРІЗВИЩЕ*

Київ 2026

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра Інженерії програмного забезпечення

Ступінь вищої освіти Магістр

Спеціальність 121 Інженерія програмного забезпечення

Освітньо-професійна програма «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення
_____ Ірина ЗАМРІЙ
«_____» _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Сорокі Олександру Олександровичу

1. Тема кваліфікаційної роботи: «Метод розпізнавання зображень для подальшої генерації текстів на основі алгоритмів штучного інтелекту»

керівник кваліфікаційної роботи Ірина ЩЕРБИНА, канд. техн. наук, доцент,

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «30» жовтня 2025 р. № 467.

2. Строк подання кваліфікаційної роботи «19» грудня 2025 р.

3. Вихідні дані до кваліфікаційної роботи: алгоритми штучного інтелекту для розпізнавання зображень, методи генерації текстів, великі мовні моделі.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
1. Дослідження алгоритмів штучного інтелекту для розпізнавання зображень
2. Дослідження сучасних методів інтерпретації зображень та генерації текстів
3. Практична реалізація методу розпізнавання зображень для подальшої генерації текстів

5. Перелік ілюстративного матеріалу: *презентація*

1. Порівняльна характеристика існуючих моделей/методів розпізнавання зображень.
2. Математична модель генерації тексту після процесу розпізнавання зображень.
3. Модифікований метод семантичної інтерпретації зображень.
4. Схема процесу методу розпізнавання зображень та подальшої генерації тексту.
5. Загальна архітектура програмної реалізації.
6. Структура та інтерфейс веб-сервісу.
7. Практичний результат.
8. Приклад завантажених документів та екранних форм інтерфейсу застосунку.
9. Згенерований текст опису.

6. Дата видачі завдання «31» жовтня 2025 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Аналіз наявної науково-технічної літератури	31.10-05.11.25	
2	Вивчення основ комп'ютерного зору та розпізнавання зображень	06.11-10.11.25	
3	Дослідження підходів розпізнавання зображень у сучасних системах штучного інтелекту	11.11-16.11.25	
4	Аналіз алгоритмів глибокого навчання для розпізнавання об'єктів та сцен	17.11-23.11.25	
5	Вивчення сучасних моделей трансформерного типу для інтерпретації зображень та генерації текстів	24.11-07.12.25	
6	Застосування моделей та методів для розпізнавання зображень та подальшої генерації текстів.	08.12-12.12.25	
7	Оформлення роботи: вступ, висновки, реферат	13.12-16.12.25	
8	Розробка демонстраційних матеріалів	17.12-19.12.25	

Здобувач вищої освіти _____ Олександр СОРОКА
(підпис)

Керівник
кваліфікаційної роботи _____ Ірина ЩЕРБИНА
(підпис)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 70 стор., 1 табл., 13 рис., 27 джерел.

Кваліфікаційна робота присвячена розробці методу автоматичного створення текстових описів на основі зображень із використанням сучасних підходів комп'ютерного зору та алгоритмів штучного інтелекту.

Мета роботи – оптимізація процесу автоматичного створення текстових описів на основі зображень за рахунок використання сучасних моделей комп'ютерного зору та алгоритмів штучного інтелекту.

Об'єкт дослідження – процес автоматичного опрацювання та інтерпретації візуальної інформації.

Предмет дослідження – засоби, технології та алгоритми штучного інтелекту, що забезпечують розпізнавання зображень і генерацію текстових описів.

У роботі використано комплекс методів і технологій, зокрема алгоритми попередньої обробки зображень, моделі Optical Character Recognition (OCR), трансформерні архітектури, моделі комп'ютерного зору (OpenAI Vision), сучасні мовні моделі GPT-класу, а також методи семантичного аналізу та генерації текстів. Проведено аналіз актуальних підходів до розпізнавання зображень, описано їхні сильні та слабкі сторони, визначено області застосування та обґрунтовано вибір найбільш ефективних моделей.

У ході роботи розроблено та оптимізовано метод автоматичного перетворення фотографій документів на структуровані дані та готові текстові описи**,** що включає етапи попередньої обробки, OCR-розпізнавання, семантичної інтерпретації та генерації тексту. Метод протестовано на зображеннях технічних паспортів транспортних засобів, що дозволило перевірити точність, стабільність та коректність відтворення ключових параметрів.

Проведено експериментальні дослідження для оцінки якості роботи системи: визначено точність вилучення даних, ефективність попередньої обробки та якість сформованих текстових описів. Результати експериментів підтверджують, що запропонований метод значно скорочує час підготовки контенту, мінімізує помилки введення та забезпечує високу інформативність та структурованість згенерованого тексту.

Отримані результати мають практичне значення та можуть бути використані у веб-сервісах подачі оголошень, системах документообігу, сервісах для бізнесу, а також у будь-яких застосунках, де потрібна автоматизація обробки зображень та генерації тексту.

КЛЮЧОВІ СЛОВА: КОМП'ЮТЕРНИЙ ЗІР, МОВНІ МОДЕЛІ, НЕЙРОННІ МЕРЕЖІ, ЦИФРОВА ОБРОБКА ЗОБРАЖЕНЬ, ОПТИЧНЕ РОЗПІЗНАВАННЯ ТЕКСТУ.

ABSTRACT

Text part of the master's qualification work: 66pages, 1 table, 13 figures, 27 sources.

The thesis is devoted to the development of a method for automatically generating text descriptions based on images using modern computer vision approaches and artificial intelligence algorithms.The purpose of the work is to simplify and optimize the process of automatically generating text descriptions based on images using modern computer vision models and artificial intelligence algorithms.The object of research is the process of automatic processing and interpretation of visual information.The subject of research is the means, technologies, and algorithms of artificial intelligence that provide image recognition and text description generation.

The work uses a set of methods and technologies, including image preprocessing algorithms, Optical Character Recognition (OCR) models, transformer architectures, computer vision models (OpenAI Vision), modern GPT-class language models, as well as methods of semantic analysis and text generation. An analysis of current approaches to image recognition is conducted, their strengths and weaknesses are described, areas of application are identified, and the selection of the most effective models is justified.

In the course of the work, a method for automatically converting photographs of documents into structured data and ready-made text descriptions was developed and optimized, which includes the stages of preprocessing, OCR recognition, semantic interpretation, and text generation. The method was tested on images of vehicle technical passports, which made it possible to verify the accuracy, stability, and correctness of the reproduction of key parameters.

Experimental studies were conducted to evaluate the quality of the system: the accuracy of data extraction, the efficiency of pre-processing, and the quality of the generated text descriptions were determined. The results of the experiments confirm that

the proposed method significantly reduces content preparation time, minimizes input errors, and ensures high informativeness and structure of the generated text.

The results obtained are of practical importance and can be used in web services for submitting advertisements, document management systems, business services, as well as in any applications where image processing and text generation automation is required.

KEYWORDS: COMPUTER VISION, LANGUAGE MODELS, NEURAL NETWORKS, DIGITAL IMAGE PROCESSING, OPTICAL CHARACTER RECOGNITION.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

NLP - Обробка природньої мови (Natural Language Processing)

LLM - Великі мовні моделі

CNN - Згорткова нейронна мережа (Convolutional Neural Network)

YOLO - Дивишся лише раз (You Only Look Once).

GPT - Мовна модель штучного інтелекту (Generative Pre-trained Transformer)

ResNet - Залишкова нейронна мережа(Residual Neural Network)

ViT - Модель глибокого навчання (Vision Transformer)

OpenAI Vision – Мультимодальна модель

API – Прикладний програмний інтерфейс (Application Programming Interface)

Yii2 - фреймворк для веб-розробки

DETR – Трансформер розпізнавання об’єктів(DEtection TRansformers)

CLIP - мережа обробки природної мови (Contrastive Language-Image Pre-training)

JSON - Формат даних (JavaScript Object Notation)

OCR - Оптичне розпізнавання символів (Optical Character Recognition)

SEO - Пошукова оптимізація (Search Engine Optimization)

IDE - інтегроване середовище розробки (Integrated Development Environment)

ВСТУП

У сучасному світі децифривізація охоплює всю сферу життя, реклама стала основним інструментом комунікації між бізнесом та споживачами. Завдяки розвитку технологій створення рекламного контенту перемістилося з традиційних друкованих носіїв у цифровий простір, а вимоги до якості та адаптивності цього контенту постійно зростають.

Разом із цим, зростає попит на автоматизацію створення рекламних оголошень, адже процес вимагає значних фінансових, часових та професійних ресурсів. Особливо складно створювати якісний контент представникам малого бізнесу, стартапам або особам, які не мають відповідних навичок у дизайні чи копірайтингу.

Крім того, проблема ускладнюється соціальним аспектом: багато людей з обмеженими можливостями — літні люди, або користувачі з вадами моторики — стикаються з труднощами при створенні навіть найпростішого оголошення. Наявність доступних інструментів для цієї категорії користувачів обмежує їхню участь у сучасній цифровій економіці.

У цьому контексті використання штучного інтелекту є революційним рішенням, яке може змінити підхід до створення реклами. Технології штучного інтелекту, зокрема комп’ютерний зір та обробка природної мови (NLP), дозволяють автоматизувати процес створення оголошень, базуючись на аналізі зображення. Завантажуючи зображення, користувач може отримати автоматично згенероване текстове оголошення, адаптоване до конкретної платформи або цільової аудиторії.

Попри те, що на ринку вже є інструменти автоматизації створення оголошень, такі як Canva або Adobe Spark, вони не вирішують низку важливих проблем:

- Відсутність повної інтеграції аналізу зображення та генерації тексту;
- Низький рівень персоналізації для спеціальних аудиторій;

– Складність у використанні для людей без досвіду роботи з такими платформами;
– Відсутність адаптивних функцій для людей з обмеженими можливостями.

Ціль даної роботи — розробити метод, який дозволяє на основі зображення автоматично створювати текстові оголошення з урахуванням контексту.

Запропонований підхід базується на глибоких нейронних мережах для аналізу зображення, а також на сучасних мовних моделях для генерації тексту. Це дозволяє створювати якісний контент швидко, точно та з урахуванням індивідуальних потреб користувача.

1 ДОСЛІДЖЕННЯ АЛГОРИТМІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ

1.1. Теоретичні основи комп'ютерного зору та розпізнавання зображень

Комп'ютерний зір є однією з най динамічніших галузей сучасного штучного інтелекту. Його головною метою є навчання комп'ютерних систем автономно сприймати, аналізувати та інтерпретувати зображення подібно до людини. На відміну від класичних алгоритмів обробки зображень, які працювали з жорстко визначеними правилами, сучасні системи базуються на моделях глибинного навчання[1]. Такі моделі здатні самостійно навчатися виділяти релевантні ознаки, розрізняти складні об'єкти, розуміти структуру сцени та передбачати взаємозв'язки між її елементами.

В умовах стрімкого розвитку цифрових платформ та великих обсягів візуальної інформації саме автоматизація аналізу зображень стає важливим етапом у створенні інтелектуальних сервісів. Однією з таких задач є автоматичне створення описів об'єктів на основі фотографій. Для систем оголошень, маркетингових платформ або документальних порталів це дозволяє суттєво спростити процес взаємодії користувача з сервісом, знизити кількість ручної роботи та підвищити точність і структурованість даних.

Будь-яке цифрове зображення складається з елементарних одиниць — пікселів[2]. Кожен піксель містить інформацію про колір або його інтенсивність. Тому зображення є тензором трирозмірної структури (1.1):

$$X \in R^{H \times W \times C}, \tag{1.1}$$

а значення кожного пікселя можна подати як (1.2):

$$X_{i,j} = \{I_{i,j}^1, I_{i,j}^2, ..., I_{i,j}^C\}. \tag{1.2}$$

У традиційній системі RGB:

- C=3;
- значення кожного каналу знаходиться в межах [0,255].

З математичної точки зору зображення представляє собою дискретну функцію двох змінних (1.3):

$$f : (x, y) \to I, \tag{1.3}$$

де I — яскравість.

Однак для обробки нейронними мережами зображення піддаються нормалізації (1.4):

$$X_{norm} = \frac{X - \mu}{\sigma}, \tag{1.4}$$

що спрощує процес навчання моделей. Це важливо, оскільки значення пікселів мають різні діапазони, а нейронні мережі працюють ефективніше, коли вхідні дані вирівняні статистично.

Ключові задачі комп'ютерного зору й їх роль у сучасних застосуваннях поділяються на:

- класифікацію зображень;
- локалізацію об'єкта;
- детекцію об'єктів;
- сегментацію;
- оптичне розпізнавання тексту (OCR).

Класифікація передбачає визначення класу об'єкта (1.5):

$$\hat{y} = \arg\max_k P(y = k|X). \tag{1.5}$$

У задачах оголошень це може бути визначення категорії об'єкта: автомобіль, побутова техніка, документ тощо.

Локалізація передбачає визначення положення об’єкта (1.6):

$$b = (x_{min}, y_{min}, x_{max}, y_{max}). \tag{1.6}$$

Це дозволяє виділити саме той фрагмент, який містить корисну інформацію (номер документа, модель авто, серію паспорта тощо).

Детекція об’єктів поєднує класифікацію та локалізацію. Детектори, такі як YOLO, SSD, Faster R-CNN, прогнозують (1.7):

$$(\hat{y}, \hat{b}) = f_\theta(X). \tag{1.7}$$

Детекція є важливою для визначення областей документа, логотипів, номерів, маркувань[3].

Сегментація поділяє зображення на піксельні маски (1.8):

$$M_{i,j} = c, \tag{1.8}$$

де c — клас.

Це дозволяє відділяти документ від фону, що важливо у розпізнаванні паспортів, техпаспортів, документів авто.

Оптичне розпізнавання тексту (OCR) трансформує візуальний текст у цифровий (1.9):

$$T = OCR(X). \tag{1.9}$$

OCR-моделі працюють на базі трансформерів (TrOCR)[4], здатних з високою точністю зчитувати текст навіть із деформованих або пошкоджених зображень.

OCR є критично важливим у системах, що розпізнають документи (наприклад, техпаспорт авто, водійське посвідчення).

Розглянемо механізми та принципи нейронних мереж у розпізнаванні зображень Convolutional Neural Networks (CNN) здійснюють згортки (1.10):

$$Y(i,j) = \sum_{m,n} X(i+m, j+n) K(m,n). \tag{1.10}$$

Це дозволяє моделі виявляти патерни різної складності: від простих контурів — до складних структур, таких як номер кузова, логотип автомобіля або серія документа.

Це ключовий інструмент для попередньої обробки та виділення ознак.

У задачах OCR CNN витягує ознаки (1.11):

$$F = CNN(X), \tag{1.11}$$

а LSTM працює як послідовний декодер (1.12):

$$h_t = LSTM(F_t, h_{t-1}). \tag{1.12}$$

Це дозволяє моделі "читати" текст зліва направо.

Трансформери й механізми уваги стали основою моделей OpenAI Vision, Google Gemini Vision, Meta ImageBind та інших. Основна формула самоуваги (1.13):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1.13}$$

До переваг можна віднести глобальну взаємодію між частинами зображення, можливістю аналізувати фото з поганою якістю і робота з нерівномірним освітленням.

Інтерпретація зображень як складова їх розуміння поділяється на виділення ознак (Feature Extraction) та семантичне представлення зображень.

Виділення ознак – це ключовий крок:

- локальні ознаки (градієнти, кути);
- текстурні ознаки;
- семантичні ознаки (марка авто, тип документа, VIN-код).

Семантичне представлення зображень

Сучасні моделі будують семантичний вектор (1.14):

$$z = Encoder(X), \tag{1.14}$$

який містить “розуміння” зображення.

Далі цей вектор надходить у генеративну модель тексту.

Комп'ютерний зір у рамках теми роботи є фундаментом усього методу[5].

Він дозволяє зменшити час заповнення оголошення, уникнути помилок користувача, створювати більш структуровані й професійні описи, розпізнавати параметри документа (рік, номер, модифікація, об'єм двигуна).

Таким чином, комп'ютерний зір виступає першим компонентом у ланцюжку Розпізнавання → Інтерпретація → Генерація тексту, що й дозволяє реалізувати метод автоматичного створення описів на основі фото.

1.2. Методи та підходи до розпізнавання зображень у сучасних системах штучного інтелекту

Розпізнавання зображень є однією з ключових підгалузей комп'ютерного зору, що охоплює сукупність методів і алгоритмів, спрямованих на інтерпретацію візуальної інформації та автоматичне виявлення об'єктів, сцен або певних характеристик. Сучасні підходи до розпізнавання зображень значно еволюціонували — від класичних методів обробки сигналів до глибоких нейронних мереж, здатних виявляти структури у даних на складних багаторівневих представленнях. У цьому підрозділі розглянемо основні групи підходів, принципи їх роботи та математичні основи, що лежать в їх основі.

Класичні методи[6] розпізнавання зображень:

– методи виділення ознак;

– методи класифікації ознак.

До появи глибокого навчання основними інструментами розпізнавання були методи математичної обробки зображень та інженерно встановлених ознак. Вони передбачали ручне створення описів об'єктів, що обмежувало гнучкість алгоритмів, але забезпечувало контрольованість процесів.

Найпоширеніші підходи методів виділення ознак:

– SIFT (Scale-Invariant Feature Transform) Метод створює дескриптори ключових точок, стійкі до змін масштабу, освітлення та поворотів.

– SURF (Speeded-Up Robust Features) Швидша модифікація SIFT, що використовує апроксимації Хаар-вейвлетів.

– HOG (Histogram of Oriented Gradients) Один з фундаментальних методів, що описує об’єкт у вигляді гістограми напрямків градієнтів. Основна ідея — локальна структура зображення визначається напрямками контрастів.

Формально гістограма HOG може бути подана як (1.15):

$$H(\theta_k) = \sum_{(x,y)\in C} m(x,y) \cdot I(\theta(x,y) \in \theta_k), \tag{1.15}$$

де:

m(x,y) — величина градієнта;

θ(x,y) — напрямок;

C — комірка зображення;

I — індикатор належності напрямку до інтервалу θ_k.

Після побудови дескрипторів застосовувалися класифікатори:

- SVM (Support Vector Machine);
- k-NN (k-Nearest Neighbors);
- Decision Trees.

SVM, наприклад, вирішує задачу знаходження гіперплощини (1.16):

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{1.16}$$

,

що максимізує відстань між класами.

З 2012 року, після прориву архітектури AlexNet, глибоке навчання стало домінуючим підходом у розпізнаванні зображень. Головною перевагою є здатність мереж самостійно вивчати ознаки, на відміну від ручного конструювання. Розглянемо ключові компоненти.

Згорткові нейронні мережі (Convolutional Neural Networks, CNN).[7]

CNN імітують принцип роботи зорової кори людини. Основний елемент — згортка, що виділяє патерни на локальних областях.

Операцію згортки можна описати формулою (1.17):

$$S(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n), \quad (1.17)$$

де:

I — зображення;

K — фільтр (ядро);

S — карта ознак.

До переваг CNN можна віднести автоматичне виділення ознак, стійкість до шумів, здатність працювати з великою кількістю даних, ефективність на складних задачах класифікації та детекції.

Мають архітектури:

– AlexNet;

– VGG;

– GoogLeNet;

– ResNet (із залишковими блоками);

– EfficientNet (оптимізація параметрів).

Особливо важливими стали residual connections, що дозволили тренувати сотні шарів. Залишковий блок має вигляд (1.18):

$$y = F(x, W_i) + x, \quad (1.18)$$

де:

F(x) — нелінійне перетворення, а додавання гарантує збереження первинного сигналу.

Трансформери в обробці зображень.

Поява Vision Transformer (ViT) змінила парадигму — модель оперує не пікселями, а "патчами" зображення, застосовуючи механізм уваги.

Базова формула уваги (1.19):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \cdot \quad (1.19)$$

Це дозволяє моделі враховувати віддалені зв'язки між частинами зображення. До переваг ViT віднесемо глобальне бачення сцени, легкість масштабування та краща робота на великих датасетах.

У сучасних рішеннях часто поєднують CNN та Transformer. Приклад — DETR, де CNN виконує екстракцію ознак, а Transformers — обробку та розподіл об'єктів у просторі.

Порівняння класичних та сучасних підходів представлено в таблиці 1.1.

Таблиця 1.1

Порівняння класичних та сучасних підходів

Модель / Метод	Опис	Переваги	Недоліки
CNN (Convolutional Neural Networks)	Глибокі згорткові мережі, що виділяють ознаки зображень через фільтри різних масштабів. Використовуються для класифікації та об'єктного розпізнавання.	Висока точність у задачах обробки зображень; швидке виділення просторових ознак; велика база досліджень.	Погано працюють з абстрактними ознаками; складно адаптувати до мовної генерації без додаткових модулів.
CLIP (Contrastive Language–Image Pre-training)	Модель від OpenAI, що вчить відповідність між зображеннями і текстовими описами, працюючи на мультимодальних	Чудово розуміє семантику зображення; може генерувати релевантні підписи; добре працює для класифікації	Не генерує повноцінний текст самостійно; потребує зв'язки з LLM; залежить від великої кількості даних.

Продовження таблиці 1.1

Порівняння класичних та сучасних підходів

Модель / Метод	Опис	Переваги	Недоліки
Vision Transformer (ViT)	Архітектура, що застосовує механізм Self-Attention для зображень, розбитих на патчі.	Висока точність; ефективна робота зі складною структурою зображень; легко інтегрується з мовними моделями.	Потребує великі набори даних; менш ефективний на низькорівневих ознаках; обчислювально важкий.
LLM-text generation (GPT-подібні моделі)	Генерують структуровані тексти, описи, комерційні оголошення на основі промптів або ознак.	Висока якість тексту; можливість створювати різні стилі; підтримка SEO та оптимізації під ринок.	Потребує чіткої структури вхідних даних; інколи генерує некоректні факти (hallucinations).
Image Captioning (класичні моделі: Show&Tell, Show&Attend&Tell)	Побудовані на поєднанні CNN + RNN/Attention для створення коротких описів фото.	Добре працюють для простих описів; доведена ефективність; легкі в обчисленнях.	Обмежена довжина текстів; неспроможність створювати комерційні або структуровані оголошення.
Meta-CLIP / OpenCLIP	Покращені CLIP-моделі з розширеними даними та кращою узагальнювальною здатністю.	Краще розуміння контексту; точніше виділення семантичних ознак; більш стійка робота зі складними зображеннями.	Не створює текст; використовується лише як енкодер ознак.

Продовження таблиці 1.1

Порівняння класичних та сучасних підходів

Модель / Метод	Опис	Переваги	Недоліки
Proposed Model (Модель з роботи)	Комбінує візуальний енкодер (CLIP/ViT), semantic refinement через базу знань та генерацію повноцінних оголошень через LLM.	Створює не просто опис, а повноцінне комерційне оголошення; висока релевантність; автоматична категоризація; оптимізація під маркетингові метрики.	Потребує складної інтеграції компонентів; залежить від якості бази знань та оптимізатора тексту.

Методи розпізнавання відіграють важливу роль у задачах генерації текстів. У контексті задачі “Метод розпізнавання зображень для генерації текстів” розпізнавання є базовим етапом, що забезпечує виділення структур, об’єктів і контексту сцени. Надалі ці дані передаються до моделей природної мови (LLM), які формують опис або рекламний текст.

Таким чином, алгоритми розпізнавання виконують семантичне структурування зображення**,** виділення суттєвих об’єктів**,** створення основи для подальшої генерації тексту**.**

1.3. Алгоритми глибокого навчання для розпізнавання об’єктів та сцен

Алгоритми глибокого навчання стали фундаментом сучасних систем розпізнавання зображень, забезпечуючи високу точність, універсальність і здатність адаптуватися до широкого спектру задач. На відміну від класичних підходів, де інженер вручну формував набір ознак, сучасні нейронні мережі здатні самостійно вивчати складні шаблони, структури та залежності у даних. Це стало

можливим завдяки розвитку великих обчислювальних потужностей, доступності графічних процесорів, відкритих датасетів і прогресу в архітектурах глибоких мереж.

У цьому підрозділі детально розглянуто ключові підходи глибокого навчання, що використовуються для розпізнавання об’єктів та сцен, зокрема згорткові мережі, сучасні детектори, трансформерні архітектури та комбіновані моделі.

Згорткові нейронні мережі як основа обробки зображень. Згорткові нейронні мережі (CNN) є базовим фундаментом у галузі комп’ютерного зору. Їхня ключова перевага полягає у здатності автоматично виділяти просторові ознаки різного рівня абстракції: від контурів і текстур до складних об’єктів та композицій. Архітектура зображена на рисунку 1.1.

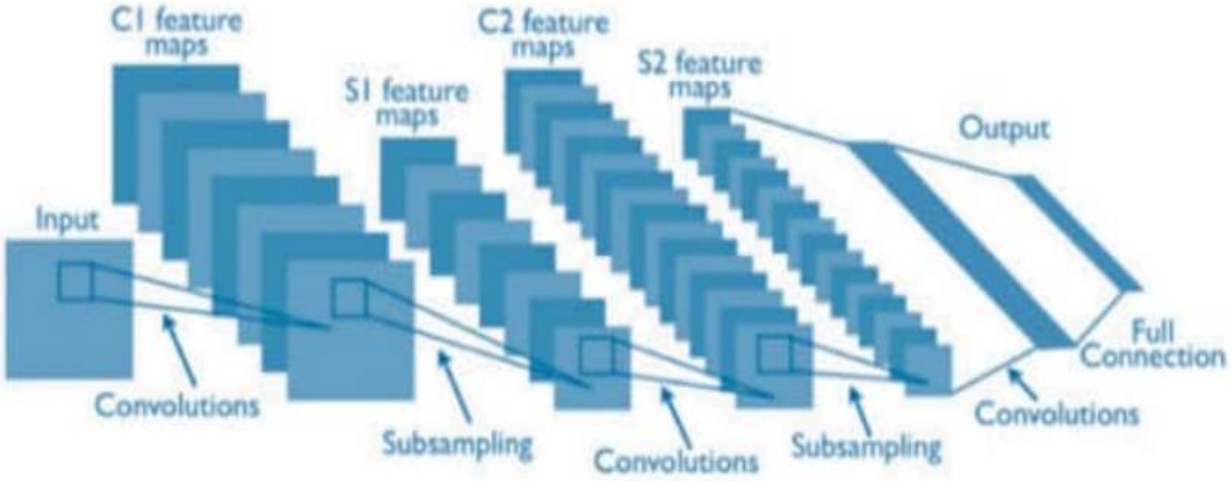


Рис. 1.1 Архітектура згорткової мережі CNN

Базові принципи роботи CNN

Кожен згортковий шар виконує операцію згортки, під час якої ядро проходить по зображенню або карті ознак і виділяє патерни (1.20):

$$Y(i,j) = \sum_{m=1}^{k} \sum_{n=1}^{k} X(i+m, j+n) \cdot K(m,n), \qquad (1.20)$$

де:

X — вхідна матриця (зображення або ознаки);

K — ядро згортки;

k — розмір ядра;

Y — карта ознак.

Чим більше фільтрів у шарі, тим більше різних характеристик може навчитися мережа.

Нелінійність.

Після згортки застосовується активаційна функція, найчастіше ReLU (1.21):

$$f(x) = \max(0, x). \tag{1.21}$$

Вона забезпечує можливість моделі наближати нелінійні залежності.

Пулінг.

Розмір карти ознак зменшується за допомогою максимального або середнього пулінгу (1.22):

$$Y(i,j) = \max_{(m,n)\in R} X(m,n). \tag{1.22}$$

Це збільшує стійкість до зсувів і зменшує обчислення.

Архітектури CNN, які визначили розвиток галузі

AlexNet

У 2012 році AlexNet стала проривом завдяки глибокій структурі (8 шарів), використанню GPU, застосуванню Dropout, збільшеній кількості фільтрів. Ця архітектура показала, що великі глибокі мережі суттєво перевищують класичні методи.

VGG

Мережі VGG відрізняються послідовним використанням малих згорток 3×3. Їхня основна ідея — глибина важливіша за ширину.

Inception та GoogLeNet

Мережа використовує багатоканальні гілки:

- 1×1 згортки;
- 3×3 згортки;
- 5×5 згортки;
- пулінг.

Це дозволяє моделі аналізувати різні масштаби одночасно.

ResNet

Головний внесок — введення залишкових зв'язків (1.23):

$$y = F(x) + x. \tag{1.23}$$

Це розв'язало проблему «зникання градієнта» і дозволило створювати мережі на 152 шари і більше. Приклад побудови блоків зображено на рисунку 1.2.

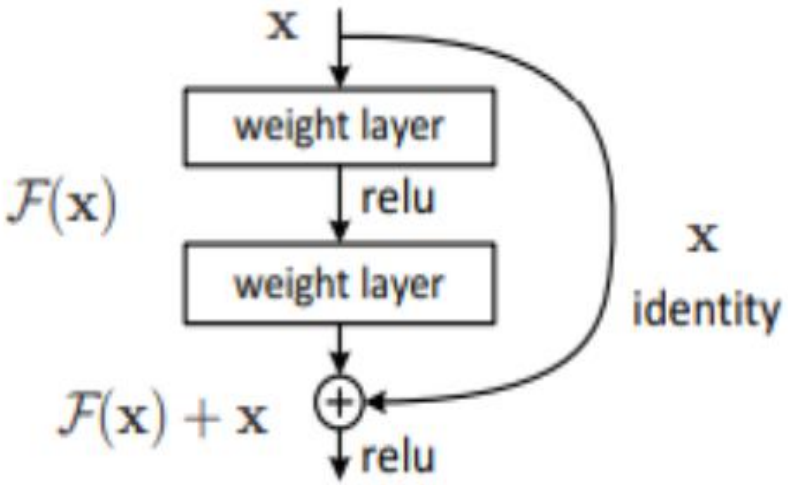


Рис. 1.2 Приклад побудови (residual) блоків

ResNet вважається однією з найефективніших архітектур CNN з точки зору точності обробки великої кількості зображень, у тому числі і для класифікації. Тим не менш, моделі, що побудовані на архітектурі ResNet є досить громіздкими, так як складаються з багатьох прихованих шарів. Через це модель потребуватиме багато ресурсів для навчання. Також, для поліпшення точності роботи CNN на ResNet модель досить великих вибірок зображень. У зворотньому випадку точність класифікації може зменшитись.

До переваг можна віднести те що можна тренувати надглибокі моделі, висока точність у задачах класифікації, стабільний та передбачуваний градієнт, легкість у модифікації (універсальна архітектура).

До недоліків віднесемо великий обсяг обчислень, блоки bottleneck мають значну кількість параметрів, важко інтерпретувати внутрішні рівні, не враховують глобальних залежностей так добре, як Vision Transformers[8].

Методи виявлення об’єктів

Обчислювальні архітектури, спрямовані на знаходження об’єктів у зображенні, поділяють на:

- двоступеневі детектори (R-CNN, Fast R-CNN, Faster R-CNN);
- одноступеневі детектори (YOLO, SSD, RetinaNet).

Двоступеневі детектори

Основна ідея — спочатку запропонувати регіони, потім класифікувати їх.

Faster R-CNN

Обчислення включають (1.24):

$$L = L_{cls} + \lambda L_{reg}, \tag{1.24}$$

де:

L_{cls}— функція втрат класифікації;

L_{reg} — втрати регресії межевих рамок;

Такі моделі точні, але повільніші.

Одноступеневі детектори

Одноступеневі моделі об’єднують виявлення та класифікацію.

YOLO

YOLO ділить зображення на сітку і передбачає:

- координати рамки;
- клас об’єкта;
- впевненість.

Формально, модель мінімізує втрати (1.25):

$$L = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} (x_i - \hat{x}_i)^2 + \sum_{i=0}^{S^2} (p_i - \hat{p}_i)^2 + \dots \quad (1.25)$$

Сучасні версії (YOLOv5–v10) здатні розпізнавати сотні класів у режимі реального часу. Н рисунку 1.3 зображено архітектуру YOLO.

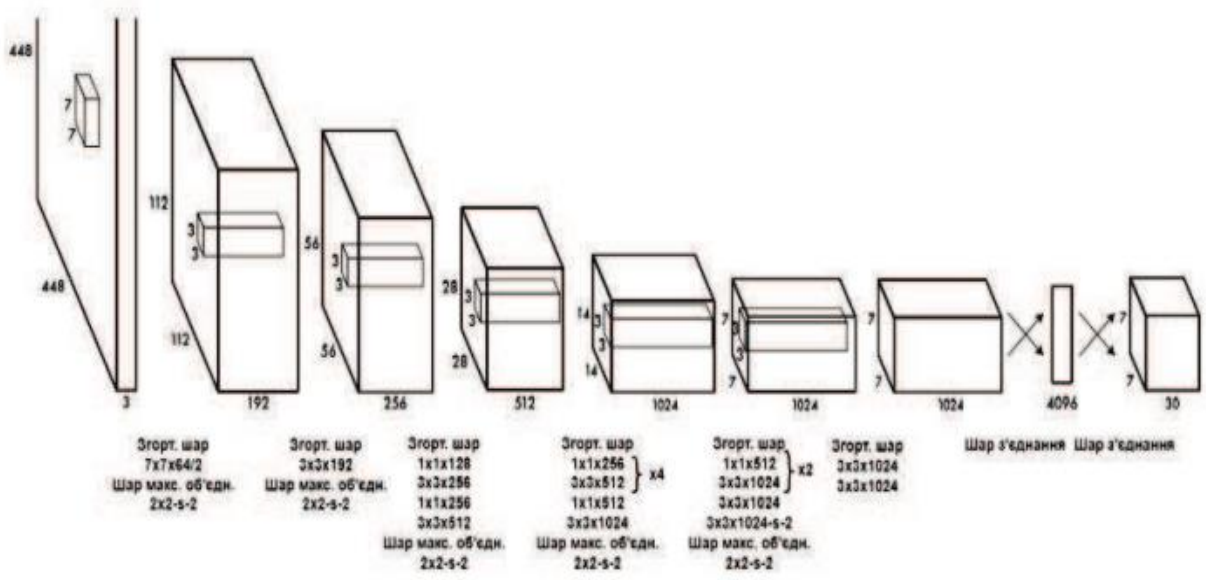


Рис. 1.3 Архітектура YOLO

Трансформери для розпізнавання зображень.

Візуальні трансформери (ViT) працюють за принципом поділу зображення на фрагменти (патчі) та подальшого аналізу через механізм уваги. Архітектура ViT зображена на рисунку 1.4.

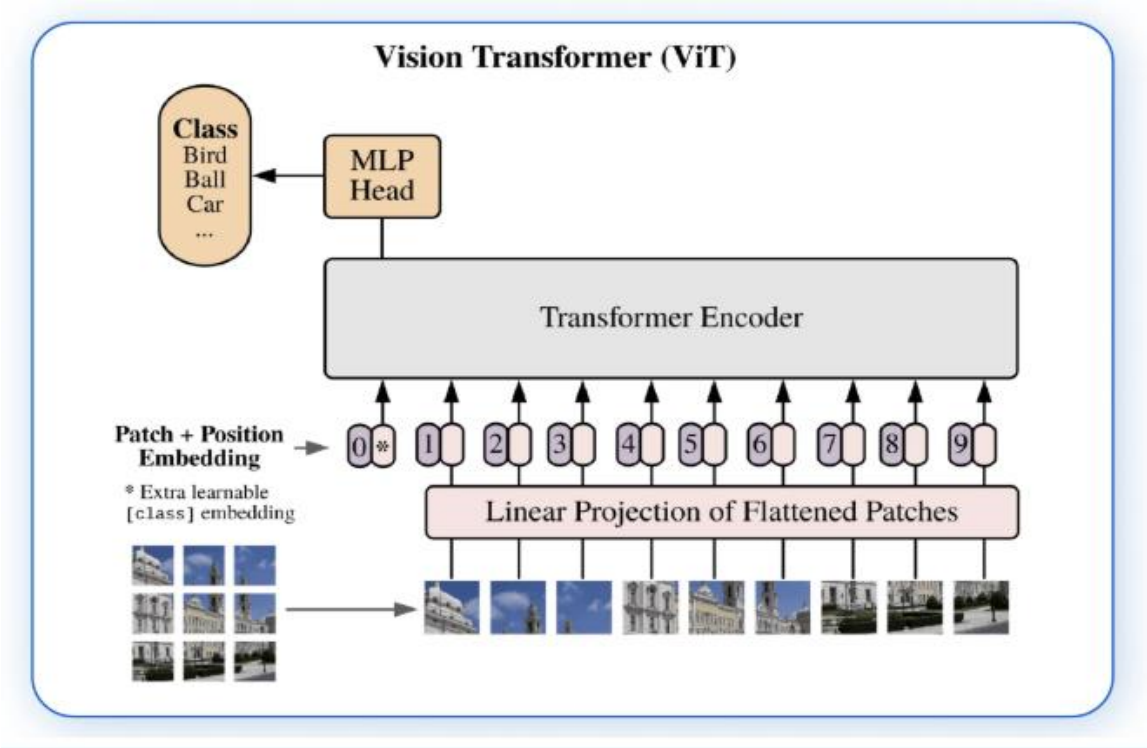


Рис 1.4 Архітектура Vision Transformer (ViT)

Патч розміром 16×16 перетворюється у вектор (1.26):

$$z_0^i = x^i E + E_{pos}^i \,, \tag{1.26}$$

де:

E_{pos} — позиційне кодування.

Механізм уваги (1.27):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \,. \tag{1.27}$$

Комбіновані архітектури DETR зображена на рисунку 1.5.

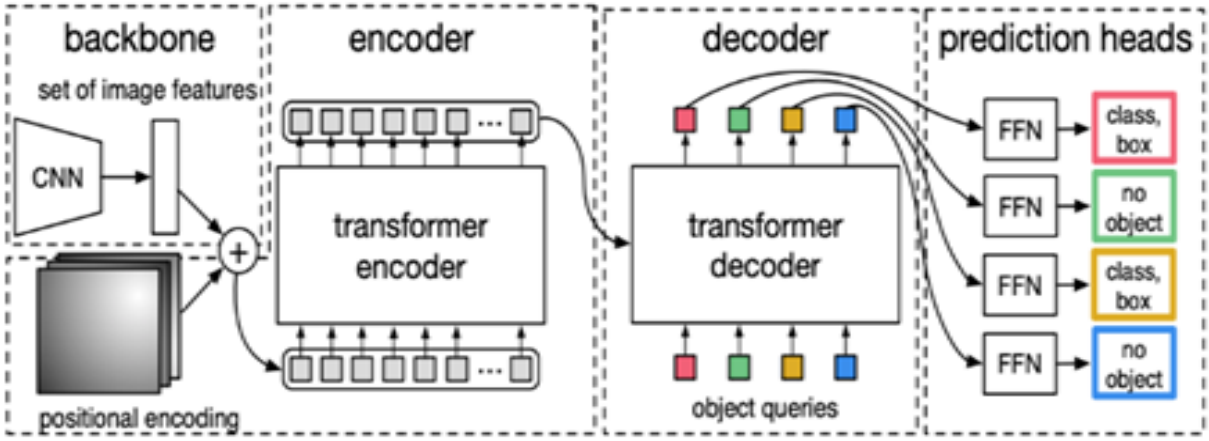


Рис. 1.5 Архітектура DETR

DETR (Detection Transformer) об’єднує CNN та трансформер:

- CNN генерує карту ознак;
- Transformer аналізує глобальні залежності;
- Hungarian loss здійснює оптимальне зіставлення об’єктів.

До переваг алгоритмів глибокого навчання можна віднести автоматичне виділення ознак, здатність працювати з великими даними, висока точність і стійкість, застосування для класифікації, сегментації, детекції, відновлення та генерації.

Алгоритми глибокого навчання дозволяють визначити об’єкти на фото, витягнути контекст сцени, створити структуру для подальшого формування тексту, підвищити точність генерації рекламних, описових та інформаційних повідомлень.

Таким чином, сучасні методи глибокого навчання є ключовим інструментом для виконання поставленої у роботі задачі: створення системи, що отримує зображення та генерує змістовні тексти за допомогою штучного інтелекту.

1.4. Висновки

У першому розділі було проведено всебічне дослідження сучасних алгоритмів штучного інтелекту, що застосовуються для розпізнавання зображень та подальшої генерації текстових описів. Аналіз показав, що за останнє десятиліття технології комп'ютерного зору пройшли шлях від класичних методів обробки зображень та згорткових нейронних мереж (CNN) до складних мультимодальних моделей трансформерного типу, здатних працювати одночасно з текстовою та візуальною інформацією.

Розглянуті класичні CNN-моделі[9], такі як VGG, ResNet і EfficientNet, продемонстрували високу ефективність у завданнях виділення ознак, класифікації та локалізації об'єктів на зображенні. Проте їхні можливості обмежуються переважно локальною обробкою ознак та недостатньою здатністю інтегрувати глобальні взаємозв'язки між частинами зображення, що є критичним фактором у задачах високорівневої інтерпретації.

За результатами проведених досліджень для подальшої практичної реалізації методу було обрано комбінований підхід, який базується на використанні моделей комп'ютерного зору для виділення візуальних і текстових ознак та мовних моделей великого розміру для генерації зв'язного текстового опису. Як основний інструмент розпізнавання було обрано моделі OpenAI Vision, що забезпечують контекстне розуміння зображень і високу точність роботи з документами, таким чином, обрані моделі та алгоритми повністю відповідають меті кваліфікаційної роботи та створюють надійну основу для реалізації запропонованого методу.

2 ДОСЛІДЖЕННЯ СУЧАСНИХ МЕТОДІВ ІНТЕРПРЕТАЦІЇ ЗОБРАЖЕНЬ ТА ГЕНЕРАЦІЇ ТЕКСТІВ

2.1. Теоретичні аспекти генерації текстових описів

Після завершення етапів розпізнавання зображення та семантичної інтерпретації отриманої інформації система переходить до одного з ключових етапів розробленого методу — автоматичної генерації текстового опису. Метою цього етапу є перетворення формалізованих, структурованих даних у зв’язний текст, який є зрозумілим для користувача та придатним для практичного використання.

На відміну від традиційних підходів, де текстові описи формуються вручну або за допомогою жорстко заданих шаблонів, у запропонованому методі використовується підхід інтелектуальної генерації, що базується на сучасних мовних моделях та алгоритмах обробки природної мови.

Результатом семантичної інтерпретації є внутрішнє подання змісту у вигляді абстрактної моделі, що описує об’єкт на зображенні. Така модель включає ідентифіковані сутності, їхні властивості, числові параметри та взаємозв’язки між ними. З теоретичної точки зору, цей етап можна розглядати як перехід від простору візуальних ознак до простору смислових концептів.

Отримана семантична модель не є текстом у прямому розумінні, а радше є інформаційною основою, яка задає, що саме має бути описано. Вона слугує вхідним контекстом для мовної моделі, яка визначає, як саме ця інформація буде представлена у текстовій формі.

Принципи керованої генерації природної мови у сучасних інтелектуальних системах[10] базується на статистичних і нейромережевих моделях мови, які реалізують імовірнісний підхід до формування послідовностей слів. Основним завданням мовної моделі є побудова такої послідовності, яка має максимальну ймовірність за умови заданого контексту (2.1):

$$P(T \mid S) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, ..., w_{i-1}, S) \quad (2.1)$$

,

де:

T — згенерований текст;

S — семантична модель об’єкта;

wi — окремі слова тексту.

Використання семантичної моделі як контексту дозволяє обмежити генерацію лише релевантною інформацією, запобігаючи появі зайвих або некоректних тверджень. Таким чином, текст формується не довільно, а у відповідності до змісту, отриманого з зображення.

Однією з важливих теоретичних задач є забезпечення логічної структури текстового опису. Людина сприймає інформацію не як набір окремих фактів, а як послідовний наратив. Тому генерація тексту відбувається з урахуванням композиційної побудови:

- вступна частина, що загально описує об’єкт;
- основний опис, який містить ключові характеристики;
- додаткова інформація, що уточнює або доповнює попередні дані.

Такий підхід дозволяє підвищити інформативність та читабельність згенерованого тексту, а також наблизити його за стилем до текстів, створених людиною.

Важливим аспектом генерації є забезпечення семантичної узгодженості між окремими фрагментами тексту. Мовна модель повинна не лише правильно описати кожну характеристику окремо, але й зберегти логічні зв’язки між ними. Це включає коректне узгодження числових значень, уникнення суперечностей та повторів.

З теоретичної точки зору, семантична цілісність досягається за рахунок глобального контексту генерації, який зберігається протягом усього процесу формування тексту. Це дозволяє моделі враховувати вже згенеровані фрагменти при створенні наступних речень.

Ще одним важливим теоретичним аспектом є можливість адаптації тексту під конкретне призначення. Один і той самий набір семантичних даних може бути представлений у різних стилях залежно від контексту використання. Наприклад, текст може бути нейтрально-інформаційним або мати більш описовий характер.

Зміна стилю досягається шляхом керування параметрами генерації та контекстними інструкціями, які задають бажаний рівень деталізації, тональність та структуру тексту. Таким чином, генерація стає гнучким інструментом, що підлаштовується під потреби користувача.

Отже, генерація текстових описів після семантичної інтерпретації є багатокомпонентним процесом, що поєднує методи комп'ютерного зору, семантичного аналізу та обробки природної мови. Теоретичні засади цього етапу забезпечують перехід від формалізованих даних до зрозумілого людині тексту, зберігаючи при цьому смислову точність, логічну послідовність та адаптивність.

Саме цей етап дозволяє реалізувати основну ідею роботи — автоматичне створення якісних текстових описів на основі візуальної інформації із застосуванням алгоритмів штучного інтелекту.

У контексті розробленого методу генерація тексту є логічним завершальним етапом обробки візуальної інформації. Якщо на початкових етапах система працює з пікселями, ознаками та візуальними патернами, то на фінальному етапі відбувається перехід до високорівневої когнітивної обробки, що імітує процес людського опису побаченого об'єкта[11].

Людина, аналізуючи зображення, спочатку ідентифікує об'єкт, далі виокремлює його ключові характеристики, після чого формує цілісний вербальний опис. Запропонований метод відтворює цю послідовність у цифровій формі, використовуючи алгоритми штучного інтелекту, які поєднують комп'ютерний зір та обробку природної мови.

Ключовою передумовою якісної генерації тексту є наявність структурованої семантичної моделі. Вона виконує роль проміжного шару між візуальними даними та мовною репрезентацією. Така модель містить узагальнену інформацію

про об’єкт, яка може включати ідентифіковані атрибути, числові та текстові параметри, логічні зв’язки між характеристиками, контекстні обмеження.

З теоретичної точки зору, семантична модель є формою абстракції, що дозволяє відокремити зміст від способу його подання. Це дає змогу повторно використовувати одну і ту ж модель для генерації різних типів текстів, змінюючи лише параметри генерації.

Генерація тексту реалізується за допомогою мовної моделі, яка працює на основі імовірнісного прогнозування послідовностей слів. Формально процес генерації можна описати як задачу максимізації умовної ймовірності тексту за наявного контексту (2.2):

$$T = \arg\max_{T} P(T \mid S), \tag{2.2}$$

де:

T — текстовий опис;

S — семантична модель, отримана після інтерпретації зображення.

Мовна модель оцінює ймовірність появи кожного наступного слова, враховуючи як попередній текст, так і семантичний контекст. Завдяки цьому досягається плавність, граматична коректність та стилістична узгодженість згенерованого опису.

Однією з принципових відмінностей сучасних мовних моделей від класичних шаблонних підходів є здатність враховувати глобальний контекст. Це означає, що модель не формує речення ізольовано, а постійно аналізує вже згенеровану частину тексту.

Контекст можна подати у вигляді латентного вектора стану, який оновлюється після кожного згенерованого слова. Таким чином, кожен новий фрагмент тексту формується з урахуванням загальної логіки та структури опису.

Важливим теоретичним завданням є запобігання появі семантичних помилок. До таких помилок належать суперечливі твердження, дублювання інформації, логічно несумісні характеристики або неправильне узгодження числових значень.

Для мінімізації таких ризиків семантична модель виконує роль «фільтра», який обмежує простір допустимих тверджень. Мовна модель, у свою чергу, використовує цю інформацію як опорну структуру, що підвищує достовірність тексту.

Теоретично процес генерації можна розглядати як побудову наративу — зв'язного текстового повідомлення, яке має початок, розвиток і завершення. Навіть у коротких описах ця структура зберігається на інтуїтивному рівні.

У запропонованому методі текст зазвичай формується у такій послідовності:

– загальне представлення об'єкта;
– опис основних характеристик;
– деталізація важливих параметрів;
– підсумкове узагальнення.

Такий підхід забезпечує природність тексту та підвищує його сприйняття користувачем.

Однією з переваг сучасних методів генерації є можливість адаптації результату під конкретні вимоги. Теоретично це досягається шляхом модифікації вхідного контексту та параметрів генерації.

Адаптація може відбуватися за такими напрямами як рівень деталізації тексту, стиль викладу, довжина тексту та акцент на певних характеристиках.

Це дозволяє використовувати один і той самий метод у різних сценаріях без необхідності змінювати алгоритмічну основу.

Генерація тексту є не просто завершальним етапом, а ключовим компонентом усього методу. Саме на цьому етапі результат роботи алгоритмів комп'ютерного зору стає корисним для кінцевого користувача.

З теоретичної точки зору, генерація виконує роль інтерфейсу між машинною обробкою та людським сприйняттям. Вона забезпечує трансформацію складних внутрішніх представлень у зрозумілу, логічну та змістовну форму.

Отже, процес генерації текстових описів після семантичної інтерпретації зображень є складною багаторівневою задачею, що поєднує імовірнісні мовні

моделі, семантичне представлення знань та принципи людського мовлення. Запропонований підхід дозволяє забезпечити високу якість, адаптивність і практичну цінність згенерованих текстів, що підтверджує доцільність його використання у прикладних веб-системах.

2.2. Сучасні моделі трансформерного типу для інтерпретації зображень та генерації текстів

Останні роки стали переломним етапом у розвитку комп'ютерного зору завдяки появі трансформерних архітектур, які продемонстрували значно кращу ефективність у завданнях високорівневої інтерпретації зображень, порівняно як із класичними CNN, так і з гібридними моделями. Архітектури типу Vision Transformer (ViT), CLIP, BLIP, Flamingo, Kosmos-2, GPT-4o та OpenAI Vision змінили підхід до роботи з візуальними даними: замість локальної обробки ознак вони навчилися працювати з глобальними зв'язками в межах усього зображення, що наблизило моделі до людського сприйняття[12].

У межах цього підрозділу розглядаються ключові принципи, що лежать в основі таких систем, наведено математичні формулювання, а також пояснюється роль трансформерів у задачі генерації тексту на основі зображення, що безпосередньо відповідає темі кваліфікаційної роботи.

Механізм уваги — ключовий елемент трансформерів полягають в тому що вони працюють за принципом self-attention, який дозволяє моделі аналізувати взаємозв'язки між будь-якими елементами вхідної послідовності. У випадку Vision Transformer такими елементами є візуальні патчі.

Механізм уваги формально визначається так (2.3):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.3}$$

де:

Q— матриця запитів (queries);

K — матриця ключів (keys);

V — матриця значень (values);

d_k — розмірність ключа.

Ця формула визначає, на які частини зображення модель повинна звернути більше уваги, коли генерує текст або виконує аналіз вмісту.

Фундамент сучасних моделей це Vision Transformer (ViT) який запропонував революційну ідею — розбити зображення на патчі, а потім передавати їх у трансформер так само, як слова в реченні (2.4):

$$x_i \in \mathbb{R}^{(P \cdot P \cdot C)}, \tag{2.4}$$

де:

P×P — розмір патча;

C — кількість каналів (RGB = 3).

Кожен патч лінійно проектується у вектор ознак (2.5):

$$z_0^i = x_i E + E_{pos}^i, \tag{2.5}$$

де:

E — матриця ембеддингу;

E_{pos} — позиційне кодування, що зберігає інформацію про розташування патча.

Після проходження через L блоків трансформера (2.6):

$$z_L = \text{Transformer}(z_0). \tag{2.6}$$

а CLS-токен використовується для класифікації або глобального розуміння зображення (2.7):

$$y_{cls} = MLP(z_L^{cls}). \tag{2.7}$$

Модель може інтерпретувати складні сцени, працювати з великими зображеннями та зберігати контекст на рівні усього кадру, а не лише локальних ознак, як це робили CNN.

Справжнім проривом стало поєднання зображення і тексту в єдиному латентному просторі Мультимодальні моделі (CLIP, BLIP, Flamingo, GPT-4o, OpenAI Vision) це дозволило їм розуміти зв’язок між об’єктами і словами,

виконувати описи, відповідати на запитання, порівнювати та класифікувати візуальні дані за текстовими інструкціями.

Однією з таких моделей є CLIP вона навчається за допомогою контрастивної функції втрат (2.8):

$$L = -\log \frac{\exp(\text{sim}(I, T)/\tau)}{\sum_j \exp(\text{sim}(I, T_j)/\tau)}, \quad (2.8)$$

де:

I — вектор зображення;

T — вектор тексту;

sim — косинусна подібність,

Модель вчиться відповідати: *“який текст відповідає цьому зображенню?”*.

Це дозволяє трансформеру формувати семантичне розуміння, а не просто виділяти об’єкти.

Також є моделі BLIP спеціально оптимізовані для генерації описів, візуального питання-відповіді (VQA), текстового уточнення, розпізнавання деталей на зображенні. Їхня робота базується на системі *інструкційного навчання* (instruction tuning) — модель розуміє запит “опиши”, “поясни”, “переліч”, “знайди помилки” тощо.

Моделі GPT-4o та OpenAI Vision є універсальними мультимодальними трансформерами[13], які здатні читати документи, розуміти дрібний текст, аналізувати складні сцени, працювати із зображеннями нерівної якості, генерувати структурований текст, опис, JSON, марковані списки. Саме ця група моделей відповідає потребам роботи, оскільки забезпечує найкращу якість розпізнавання документів (наприклад, техпаспорта авто) та генерації оголошень.

Для генерації текстових описів широко використовується архітектура Encoder–Decoder:

Encoder аналізує зображення та формує набір ознак (2.9):

$$h = Encoder(I). \quad (2.9)$$

Decoder генерує текст, використовуючи механізм уваги до елементів зображення (2.10):

$$y_t = Decoder(y_{t-1}, h) \quad (2.10)$$

Процес продовжується до отримання завершеного тексту. Це дозволяє моделі послідовно описувати наявні об’єкти, структурувати інформацію, дотримуватися логіки побудови речень.

Особливості та переваги трансформерів у задачі генерації описів

Переваги:

- Глобальний контекст: трансформер бачить усе зображення цілісно;
- Стійкість до шумів та різних умов: нерівні документи, плями, тіні, розмиття;
- Можливість інтеграції з LLM: текст генерується природною, зрозумілою мовою;
- Гнучкість: працює з техпаспортами, фото товарів, об’єктів, сцен;

Особливо важливо для даної роботи:

- можливість структурувати інформацію у вигляді JSON;
- витягування тексту (OCR) без спеціальних додаткових моделей;
- коректне тлумачення полів документа;
- генерування повноцінного оголошення (рік, пробіг, стан, опис).

У рамках кваліфікаційної роботи трансформерні моделі застосовуються як центральний компонент системи. Вони забезпечують точне витягування ключових даних із зображень техпаспортів, проводять семантичну інтерпретацію вмісту, автоматично будують структурований текст опису транспортного засобу.

На відміну від CNN, які лише розпізнають об’єкти, трансформери дозволяють розуміти зображення, інтерпретувати його, створювати текст, структуровано подавати інформацію.

Це робить їх незамінним інструментом для розробки методу автоматичної генерації текстів із зображень.

2.3. Висновки

Особливу увагу було приділено трансформерним моделям, зокрема Vision Transformer (ViT), CLIP, BLIP, GPT-4o та OpenAI Vision. Завдяки механізму самоуваги вони здатні враховувати глобальний контекст, встановлювати семантичні відповідності між зображенням та текстом, а також генерувати структуровані й логічні описи. Математичний аналіз механізму уваги та принципів роботи трансформерів показав, що їхня архітектура краще пристосована до задач, де необхідна інтелектуальна інтерпретація зображення, а не лише розпізнавання об’єктів.

Проведене дослідження дозволило визначити, що мультимодальні моделі трансформерного типу найповніше відповідають вимогам кваліфікаційної роботи, оскільки забезпечують одразу три ключові властивості:
— точне OCR-розпізнавання тексту на складних документах**,**
— семантичну інтерпретацію отриманої інформації**,**
— генерацію логічно зв’язаного текстового опису**.**

Таким чином, результати аналізу є підґрунтям для формування власного методу автоматичного створення текстових описів на основі зображень, що розробляється у наступних розділах роботи.

3 ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДУ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ ДЛЯ ПОДАЛЬШОЇ ГЕНЕРАЦІЇ ТЕКСТІВ

3.1. Загальна архітектура та логіка роботи методу

У межах даної кваліфікаційної роботи розроблено метод розпізнавання зображень для подальшої автоматичної генерації текстових описів із використанням алгоритмів штучного інтелекту та сучасних моделей комп'ютерного зору[16]. Практична реалізація методу орієнтована на прикладну задачу — спрощення процесу створення оголошень на основі фотографій документів або об'єктів, зокрема транспортних засобів.

Розроблений метод поєднує в собі кілька взаємопов'язаних етапів, кожен із яких виконує чітко визначену функцію та забезпечує поступовий перехід від необробленого зображення до готового текстового опису. Загальна логіка роботи методу ґрунтується на модульному підході, що дозволяє розділити складну задачу на окремі підзадачі та підвищити гнучкість і масштабованість системи.

Архітектура методу представлена на рисунку 3.1, передбачає використання клієнт–серверної моделі. Користувач взаємодіє із системою через веб-інтерфейс, за допомогою якого здійснюється завантаження зображень та отримання згенерованого текстового результату. Серверна частина відповідає за обробку зображень, виконання алгоритмів розпізнавання, семантичний аналіз отриманих даних та генерацію тексту.

Загальна схема роботи методу включає такі основні етапи:

- Завантаження зображення користувачем;
- Попередня обробка зображення;
- Розпізнавання тексту та візуальних елементів;
- Семантична інтерпретація отриманих даних;
- Генерація текстового опису;
- Відображення результату в інтерфейсі користувача.

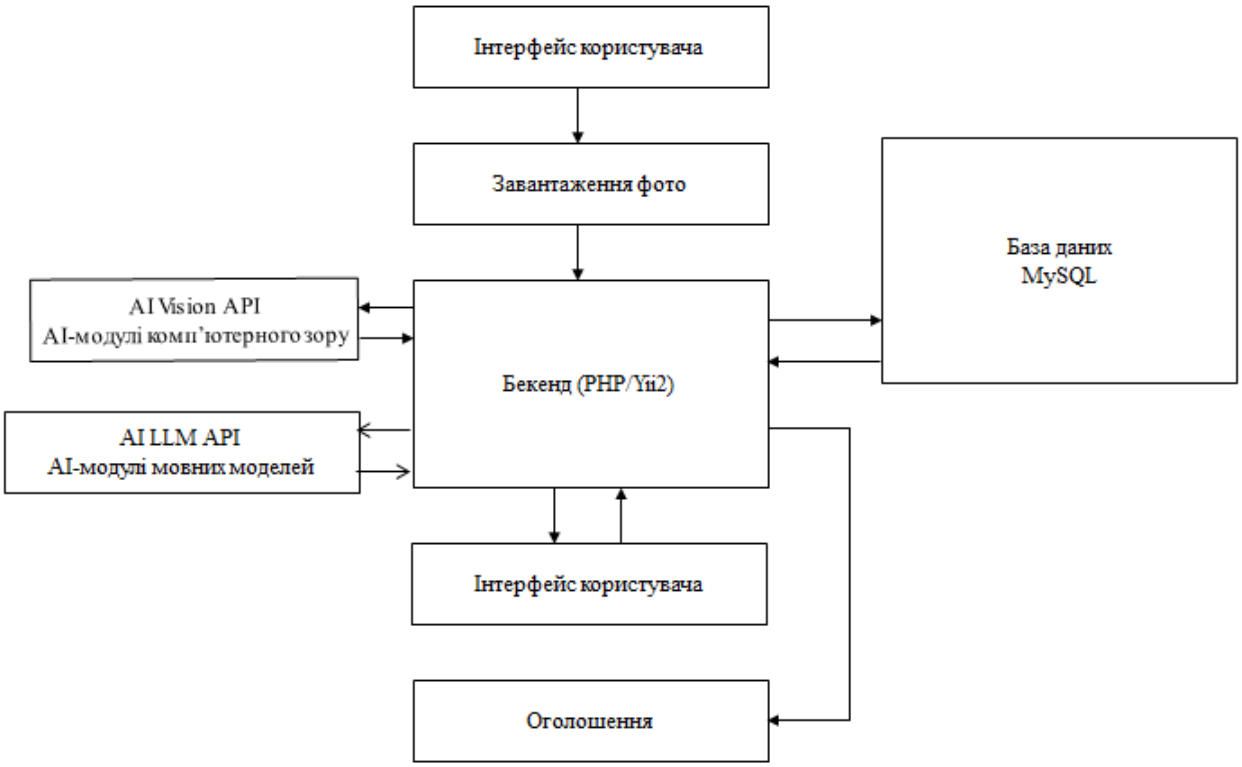


Рис. 3.1 Загальна архітектура практичної реалізації методу

Кожен із зазначених етапів реалізований у вигляді окремого логічного модуля, що дозволяє ізолювати обчислювальні процеси та спростити налагодження і подальший розвиток системи.

Особливістю розробленого методу є використання моделей комп'ютерного зору, здатних працювати з реальними фотографіями, зробленими в неконтрольованих умовах. Це означає, що вхідні зображення можуть мати різну якість, освітлення, нахил, наявність шумів та сторонніх об'єктів. Саме тому в архітектурі методу значна увага приділяється етапу попередньої обробки, який суттєво впливає на кінцеву точність розпізнавання.

На рівні логічної взаємодії між компонентами система функціонує наступним чином. Після завантаження зображення через веб-інтерфейс воно передається на сервер, де ініціюється модуль обробки зображень. Цей модуль виконує нормалізацію, вирівнювання перспективи та підготовку зображення до подальшого аналізу. Оброблене зображення передається до модуля розпізнавання,

який використовує модель комп’ютерного зору для вилучення текстових та візуальних характеристик.

Результати розпізнавання зберігаються у структурованому вигляді, у форматі JSON[17], що забезпечує зручну передачу даних між модулями. Далі ці дані надходять до модуля семантичної інтерпретації, який аналізує значення розпізнаних полів, перевіряє їх на коректність та логічну узгодженість.

Завершальним етапом є генерація текстового опису оголошення. На цьому етапі використовується мовна модель, яка формує зв’язний, зрозумілий та граматично коректний текст на основі отриманих структурованих даних. Згенерований опис повертається користувачеві через інтерфейс, де його можна переглянути, відредагувати або одразу використати для публікації оголошення.

Таким чином, розроблений метод реалізує повний цикл автоматизованої обробки візуальної інформації — від завантаження зображення до формування готового текстового результату. Такий підхід дозволяє суттєво зменшити часові витрати користувача, мінімізувати кількість ручних операцій та підвищити якість текстових описів.

Важливою особливістю практичної реалізації запропонованого методу є його орієнтація на реальні сценарії використання[18], у яких користувач не володіє спеціальними технічними знаннями у сфері комп’ютерного зору чи обробки зображень. Саме тому архітектура методу була спроєктована з урахуванням принципів простоти, автоматизації та мінімальної участі людини у процесі аналізу вхідних даних.

Метод не вимагає від користувача ручного виділення областей інтересу, введення параметрів розпізнавання або попереднього налаштування алгоритмів. Усі ключові операції — від підготовки зображення до генерації текстового опису — виконуються автоматично. Це дозволяє використовувати систему у широкому спектрі прикладних задач, зокрема під час створення оголошень для онлайн-платформ, маркетплейсів або інформаційних сервісів.

Архітектурно метод побудований за принципом послідовної обробки даних[19], де кожен наступний етап використовує результати попереднього.

Такий підхід забезпечує логічну цілісність процесу та дозволяє легко ідентифікувати джерело можливих помилок на будь-якому етапі роботи системи. У разі потреби кожен модуль може бути доопрацьований або замінений без суттєвих змін у загальній структурі.

На концептуальному рівні розроблений метод можна подати у вигляді функціональної залежності (3.1):

$$T = F(I), \tag{3.1}$$

де:

I — вхідне зображення;

F — сукупність алгоритмів обробки, розпізнавання та інтерпретації;

T — згенерований текстовий опис.

Функція F є композицією кількох підфункцій, кожна з яких відповідає окремому етапу обробки (3.2):

$$F = F_{gen} \circ F_{sem} \circ F_{rec} \circ F_{prep}, \tag{3.2}$$

де:

F_{prep} — попередня обробка зображення;

F_{rec} — розпізнавання тексту та візуальних ознак;

F_{sem} — семантична інтерпретація даних;

F_{gen} — генерація текстового опису.

Такий формальний опис дозволяє чітко визначити роль кожного етапу в загальному процесі та підкреслює модульність розробленого підходу.

Окрему увагу під час проєктування архітектури було приділено питанням стабільності та надійності роботи методу. У реальних умовах зображення можуть містити частково закриті дані, відблиски, тіні або нечіткі символи[19]. Тому система не лише виконує розпізнавання, але й аналізує достовірність отриманих результатів. У разі виявлення сумнівних або неповних даних вони можуть бути або позначені як необов’язкові, або оброблені з використанням евристичних правил.

З погляду інформаційних потоків метод реалізує чітко визначений ланцюг передачі даних між компонентами системи. Після завантаження зображення формується запит до серверної частини, який містить сам файл та супровідні метадані. Далі всі проміжні результати зберігаються у структурованому форматі, що забезпечує прозорість процесу та можливість логування або аналізу результатів на кожному етапі.

Важливо зазначити, що розроблений метод не є жорстко прив’язаним до конкретного типу зображень. Хоча в межах даної роботи основна увага приділяється прикладу з технічними документами транспортних засобів, архітектура методу дозволяє адаптувати його і для інших типів візуальних даних. Для цього достатньо змінити правила семантичної інтерпретації та шаблони[20] генерації тексту, не втручаючись у базову логіку розпізнавання.

Завдяки такій універсальності метод може бути використаний не лише для створення оголошень, але й для автоматичного заповнення форм, попереднього аналізу документів або підготовки описів товарів. Це підвищує практичну цінність запропонованого рішення та робить його придатним для подальшого розвитку.

Узагальнюючи, можна стверджувати, що загальна архітектура та логіка роботи методу забезпечують:

- повну автоматизацію процесу обробки зображень;
- мінімальну участь користувача;
- модульність і масштабованість системи;
- можливість адаптації до різних прикладних задач.

Таким чином, цей підрозділ закладає теоретичну та практичну основу для детального розгляду реалізації окремих етапів методу, які будуть проаналізовані у наступних підрозділах.

3.2. Реалізація етапу завантаження та попередньої обробки зображень

Структура та інтерфейс вебзастосунку для практичної реалізації методу представлена на рисунку 3.2.

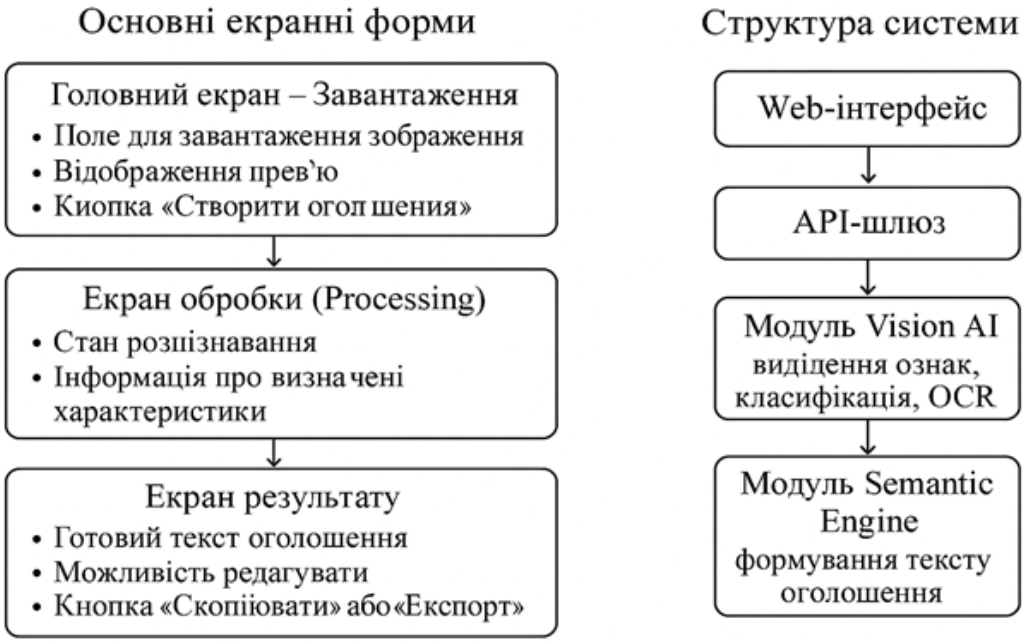


Рис. 3.2 Структура та інтерфейс вебзастосунку

Етап завантаження та попередньої обробки зображень є одним із ключових у практичній реалізації розробленого методу, оскільки саме від якості підготовки вхідних даних значною мірою залежить точність подальшого розпізнавання та коректність згенерованого текстового опису. У реальних умовах зображення, які надаються користувачами, часто мають різну роздільну здатність, нерівномірне освітлення, шуми або геометричні викривлення. Тому застосування етапу попередньої обробки є обґрунтованою необхідністю.

Завантаження зображення користувачем

Процес роботи методу починається із завантаження зображення через веб-інтерфейс системи, який зображений на рисунку 3.3.

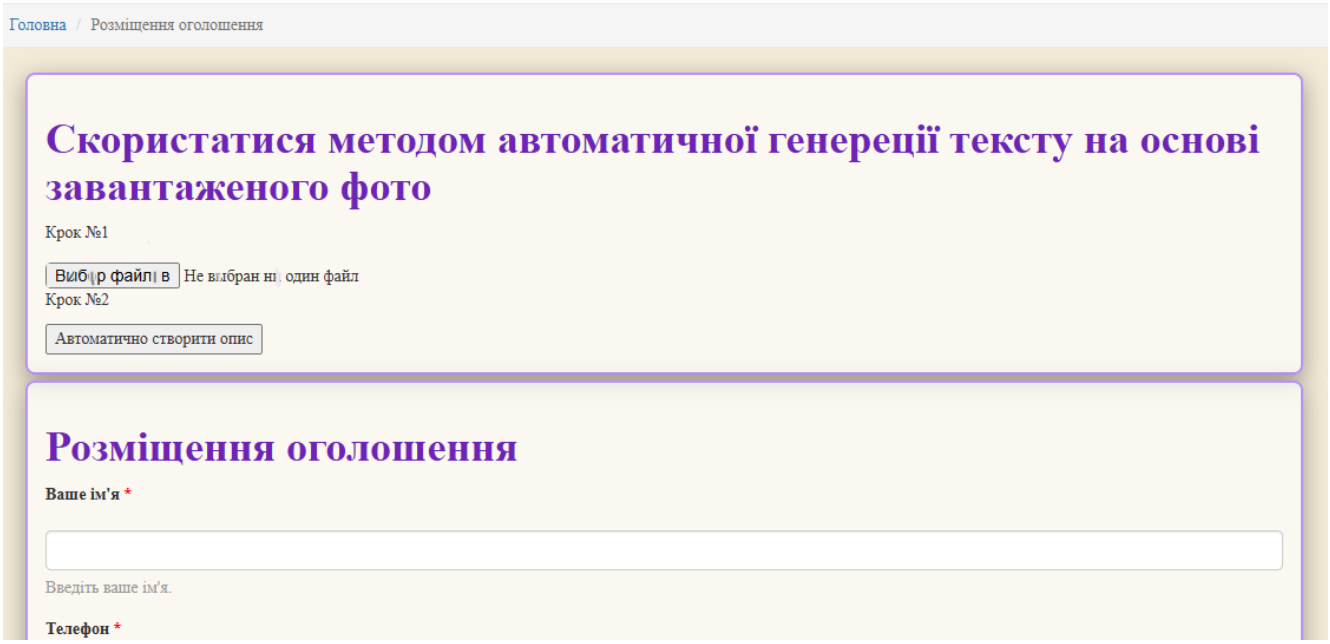


Рис. 3.3 інтерфейс системи

Користувач має можливість додати одне або декілька зображень, наприклад фотографії документа з різних сторін зображені на рисунках 3.4, 3.5 та 3.6.

Рис. 3.4 Фото технічного паспорта (лицева сторона)

Рис. 3.5 Фото технічного паспорта (зворотня сторона)

Рис. 3.6 Фото автомобіля

Під час завантаження виконується базова перевірка коректності вхідних даних, зокрема:

- перевірка формату файлу (JPEG, PNG);
- перевірка розміру зображення;
- перевірка цілісності файлу.

Після успішного завантаження зображення тимчасово зберігається на сервері, а його метадані передаються до модуля попередньої обробки. Такий підхід дозволяє відокремити інтерфейсну частину системи від логіки обробки даних, що підвищує надійність та зручність масштабування.

Геометричне вирівнювання зображення[21]

Однією з поширених проблем є фотографування документа під кутом, що призводить до спотворення перспективи. Для усунення цієї проблеми застосовується процедура геометричного вирівнювання (deskewing), метою якої є приведення зображення до правильної прямокутної форми[22].

На цьому етапі система аналізує контури документа, визначає його межі та виконує корекцію перспективи. Формально цей процес можна описати як перетворення координат пікселів (3.3):

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{3.3}$$

де:

(x,y) — координати пікселя у вихідному зображенні;

(x′,y′) — координати після трансформації;

H — матриця проєктивного перетворення.

Завдяки застосуванню такого перетворення документ набуває правильної орієнтації, що значно спрощує подальше розпізнавання текстових областей.

Фільтрація шумів та підсилення контрасту

Після вирівнювання зображення виконується фільтрація шумів, які можуть виникати внаслідок низької якості камери або недостатнього освітлення. Для цього застосовуються згладжувальні фільтри, зокрема медіанний або гаусовий фільтр (3.4):

$$I'(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x+i, y+j) \cdot G(i,j), \tag{3.4}$$

де:

I(x,y) — значення пікселя вхідного зображення;

G(i,j) — ядро гаусового фільтра;

I′(x,y) — відфільтроване зображення.

Паралельно з фільтрацією виконується підсилення контрасту, що дозволяє чіткіше виділити текстові символи на фоні документа. Це особливо важливо для документів зі складним фоном або водяними знаками.

Автоматичне кадрування області документа

Наступним кроком є автоматичне виділення області документа та усунення зайвого фону[23]. На цьому етапі система аналізує форму та межі зображення, щоб відокремити корисну інформацію від сторонніх об’єктів, таких як стіл, руки або навколишні поверхні.

Процес кадрування дозволяє:

- зменшити обсяг оброблюваних даних;
- зосередити увагу алгоритмів розпізнавання виключно на вмісті документа;
- підвищити швидкість та точність роботи системи.

Нормалізація кольору та освітлення

Останнім етапом попередньої обробки є нормалізація кольору та яскравості. Метою цього кроку є зменшення впливу тіней, відблисків і нерівномірного освітлення. Зображення переводиться у стандартизований колірний простір, після чого виконується вирівнювання яскравості.

Формально нормалізацію можна подати як (3.5):

$$I_{norm} = \frac{I - \mu}{\sigma}, \tag{3.5}$$

де:

μ — середнє значення яскравості;

σ — стандартне відхилення.

Цей підхід забезпечує стабільні умови для подальшого розпізнавання тексту незалежно від зовнішніх факторів зйомки.

У результаті виконання етапу завантаження та попередньої обробки формується стандартизоване зображення високої якості, придатне для подальшого аналізу. Застосування описаних методів дозволяє суттєво знизити кількість помилок на етапі OCR-розпізнавання та підвищити загальну ефективність роботи системи.

Таким чином, етап попередньої обробки є критично важливим компонентом запропонованого методу та створює надійну основу для наступного етапу — розпізнавання тексту та візуальних ознак з використанням моделей комп’ютерного зору, який буде детально розглянуто у наступному підрозділі.

3.3. Реалізація етапу розпізнавання тексту та візуальних ознак з використанням OpenAI Vision

Після виконання етапів завантаження та попередньої обробки зображення система переходить до ключового етапу розробленого методу — розпізнавання текстової та візуальної інформації з використанням моделей комп’ютерного зору на базі OpenAI Vision[25]. Саме на цьому етапі здійснюється перехід від візуального представлення даних до формалізованої структури, придатної для подальшої семантичної інтерпретації та генерації текстового опису.

Загальний принцип роботи модуля розпізнавання

Модуль розпізнавання отримує на вхід підготовлене зображення документа, яке вже вирівняне, очищене від шумів та нормалізоване за кольором і освітленням. Основним завданням цього етапу є автоматичне виділення текстових фрагментів, ключових атрибутів та їх перетворення у структурований вигляд.

На відміну від класичних OCR-систем, які працюють виключно на рівні символів, OpenAI Vision використовує глибокі нейронні мережі, що поєднують аналіз візуального контексту, розташування елементів та семантичний зміст тексту. Це дозволяє не лише зчитувати символи, але й розуміти логіку документа в цілому.

Формально цей етап можна подати у вигляді відображення (3.6):

$$I \rightarrow \{(t_i, p_i, c_i)\}_{i=1}^{N}, \tag{3.6}$$

де:

I — вхідне зображення;

t_i — розпізнаний текстовий фрагмент;

p_i — просторове розташування фрагмента;

c_i — контекстна приналежність (атрибут документа).

Особливістю розробленого методу є підтримка обробки декількох зображень одного документа, наприклад фотографій технічного паспорта транспортного засобу з різних сторін. У такому випадку кожне зображення обробляється окремо, після чого результати інтегруються у єдину логічну структуру.

Для кожного зображення виконується:

- аналіз загальної структури документа;
- виявлення зон із текстовою інформацією;
- розпізнавання текстових полів та числових значень.

Отримані дані не розглядаються ізольовано, а аналізуються у взаємозв’язку між собою. Наприклад, значення року випуску, марки або моделі автомобіля

можуть бути розташовані на різних сторонах документа, але у підсумку формують єдиний опис об’єкта.

Результатом роботи модуля розпізнавання є не просто набір текстових рядків, а структурований набір даних у форматі JSON. Такий формат є зручним для подальшої обробки, зберігання та передачі між модулями системи.

Приклад узагальненого результату розпізнавання можна подати таким чином:

```
# Приклад результату розпізнавання
{
  "document_type": "vehicle_registration_certificate",
  "vehicle": {
    "brand": "Suzuki",
    "model": "Grand Vitara",
    "year": "2006",
    "engine_capacity": "2737",
    "body_type": "універсал",
    "color": "чорний",
    "seats": "5"
  },
  "registration": {
    "country": "Україна",
    "first_registration": "16.05.2013",
    "last_registration": "04.07.2020"
  },
  "status": {
    "owner": "є власником",
    "documents_valid": true
  }
}
```

Формування такої структури дозволяє уникнути дублювання інформації, забезпечити однозначну інтерпретацію даних, спростити подальшу генерацію тексту оголошення.

Контекстне розпізнавання та обробка невизначеностей

Важливою особливістю OpenAI Vision є здатність працювати з неповними або частково пошкодженими даними. У реальних умовах деякі поля документа можуть бути нечіткими, перекритими або відсутніми. У таких випадках система не робить жорстких припущень, а позначає відповідні значення як невизначені або потребуючі уточнення[26].

Це дозволяє уникнути генерації некоректної інформації та забезпечує більш відповідальне використання розпізнаних даних. Формально невизначені значення можуть бути подані як:

"engine_capacity": null

або з додатковою позначкою рівня впевненості:

```
"engine_capacity": {
  "value": "2737",
  "confidence": 0.92
}
```

Застосування OpenAI Vision у межах запропонованого методу забезпечує низку суттєвих переваг:

- високу точність розпізнавання навіть при складних умовах зйомки;
- здатність аналізувати структуру документа, а не лише окремі символи;
- підтримку багатомовних документів;
- можливість інтеграції результатів у єдиний семантичний простір.

У порівнянні з традиційними OCR-підходами, запропоноване рішення демонструє кращу адаптивність та універсальність, що є критично важливим для задач автоматичної генерації текстів на основі зображень.

Використання моделей OpenAI Vision у межах розробленого методу дозволяє перейти від традиційного підходу «зображення в текст» до більш складної та інтелектуальної моделі аналізу, яка враховує не лише вміст зображення, але й його структурну та семантичну організацію. Це особливо важливо у випадку офіційних документів, де значення окремих текстових фрагментів визначається не лише їх змістом, але й розташуванням відносно інших елементів.

На практиці це означає, що система здатна розрізняти, наприклад, серійний номер документа, технічні характеристики об’єкта та службову інформацію, навіть якщо вони представлені однаковим шрифтом або мають схожий формат. Такий рівень аналізу неможливо досягти за допомогою класичних OCR-бібліотек без додаткових евристик і ручних правил.

Окрему увагу у розробленому методі приділено стійкості алгоритму до варіативності вхідних даних. Фотографії документів можуть бути зроблені за різних умов освітлення, різними камерами та з різною якістю. Моделі OpenAI Vision, навчені на великих і різноманітних наборах даних, демонструють високу узагальнювальну здатність і зберігають прийнятний рівень точності навіть у складних умовах.

Крім того, важливою перевагою є здатність моделі працювати з контекстом усієї сторінки. Наприклад, якщо певне текстове поле частково пошкоджене або розмите, система може відновити його значення на основі суміжних полів або загальної логіки документа. Такий підхід значно зменшує кількість помилок та підвищує надійність кінцевого результату.

З точки зору програмної реалізації, використання OpenAI Vision спрощує архітектуру системи, оскільки замість ланцюжка окремих модулів (детектор тексту, сегментатор, OCR, постпроцесор) застосовується єдиний інтелектуальний

компонент. Це зменшує кількість точок відмови та полегшує подальшу підтримку й масштабування рішення.

Важливо також зазначити, що результати розпізнавання зберігаються у формалізованому вигляді, що відкриває можливості для подальшого аналізу, повторного використання даних та інтеграції з іншими інформаційними системами. Наприклад, збережені JSON-структури можуть бути використані не лише для генерації оголошень, але й для статистичного аналізу, фільтрації або автоматичного заповнення форм.

Загалом, етап розпізнавання тексту та візуальних ознак з використанням OpenAI Vision є ключовим для досягнення поставленої мети роботи. Саме на цьому етапі забезпечується перетворення неструктурованої візуальної інформації у структуровані дані, які надалі використовуються мовними моделями для створення зрозумілих, логічно побудованих та релевантних текстових описів.

Таким чином, етап розпізнавання тексту з використанням OpenAI Vision є центральною ланкою розробленого методу. Він забезпечує перехід від візуального образу документа до структурованого представлення даних, яке згодом використовується для семантичної інтерпретації та генерації текстового опису оголошення. Якість та надійність цього етапу безпосередньо впливають на ефективність усієї системи в цілому.

3.4 Генерація тексту оголошення та адаптація під користувача

Після завершення етапів розпізнавання та семантичної інтерпретації візуальної інформації система переходить до заключного етапу розробленого методу — автоматичної генерації тексту оголошення. Метою цього етапу є створення зрозумілого, інформативного та стилістично коректного текстового опису, який максимально відповідає очікуванням потенційних користувачів та вимогам рекламних платформ.

На відміну від класичних шаблонних підходів, запропонований метод базується на використанні сучасних мовних моделей, здатних формувати зв’язний текст з урахуванням контексту, логіки та специфіки предметної області. Генерація тексту не є прямим копіюванням розпізнаних даних, а являє собою інтелектуальний процес узагальнення та переформулювання інформації.

Вхідними даними для модуля генерації тексту є структурований набір атрибутів, сформований на попередніх етапах та поданий у форматі JSON. Такий підхід дозволяє відокремити процес аналізу зображення від процесу генерації тексту, що підвищує гнучкість та масштабованість методу.

Формально процес генерації можна подати як функцію (3.7):

$$T = f(D, P, U), \tag{3.7}$$

де:

D — структуровані дані, отримані з зображення;

P — параметри генерації (стиль, довжина, тональність);

U — налаштування користувача;

T — згенерований текст оголошення.

Перед безпосередньою генерацією тексту здійснюється внутрішня логічна підготовка даних. Система визначає ключові смислові блоки майбутнього оголошення, зокрема:

- загальний опис об’єкта;
- основні характеристики;
- додаткові переваги;
- інформацію про стан та юридичну коректність документів.

Такий підхід дозволяє уникнути хаотичного переліку характеристик та забезпечує читабельну структуру тексту. У результаті згенероване оголошення виглядає природно та легко сприймається користувачем.

Безпосередня генерація тексту здійснюється за допомогою мовних моделей великого розміру, які попередньо навчені на великих корпусах текстів та здатні формувати граматично правильні та семантично узгоджені речення.

Особливістю запропонованого методу є те, що мовна модель працює не з «сирим» текстом, а з попередньо структурованими даними. Це значно знижує ризик появи неточностей та вигаданих фактів у згенерованому описі.

Крім того, модель може адаптувати стиль тексту залежно від обраного сценарію використання, наприклад:

- нейтральний інформативний опис;
- рекламний стиль з акцентом на перевагах;
- короткий технічний опис.

Важливою складовою розробленого методу є можливість адаптації тексту оголошення під індивідуальні потреби користувача. Користувач може впливати на:

- обсяг тексту (короткий, середній, розширений);
- стиль викладу;
- мову оголошення;
- рівень деталізації характеристик.

Це дозволяє використовувати один і той самий метод для різних платформ та сценаріїв, не змінюючи внутрішню логіку алгоритму.

Обробка невизначених та відсутніх даних

У випадках, коли деякі атрибути не були розпізнані або мають низький рівень достовірності, система враховує це під час генерації тексту. Невизначені поля або не включаються до опису, або формулюються узагальнено, без категоричних тверджень.

Такий підхід забезпечує коректність та етичність згенерованого контенту і запобігає поширенню недостовірної інформації.

На завершальному етапі формується фінальний текст оголошення, який може бути одразу використаний для публікації або за потреби відредагований користувачем. Приклад узагальненого результату зображено на рисунку 3.7 :

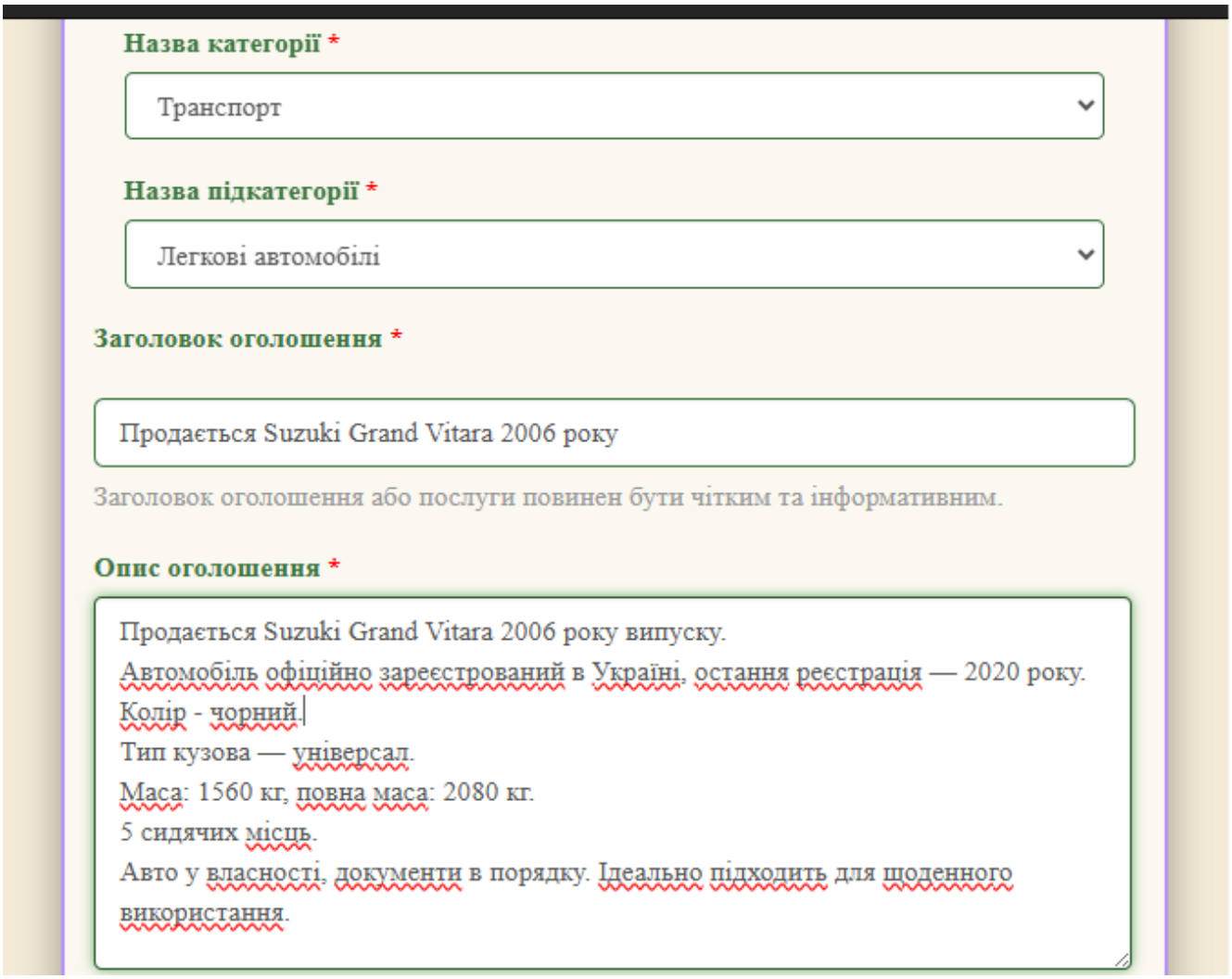


Рис 3.7 Сформований фінальний текст

Отже, етап генерації тексту оголошення та його адаптації під користувача завершує реалізацію розробленого методу. Поєднання структурованих даних, сучасних мовних моделей та механізмів адаптації дозволяє автоматично створювати якісні текстові описи, що відповідають як технічним, так і комунікаційним вимогам сучасних рекламних платформ.

Алгоритм роботи системи представлений на рисунку 3.8.

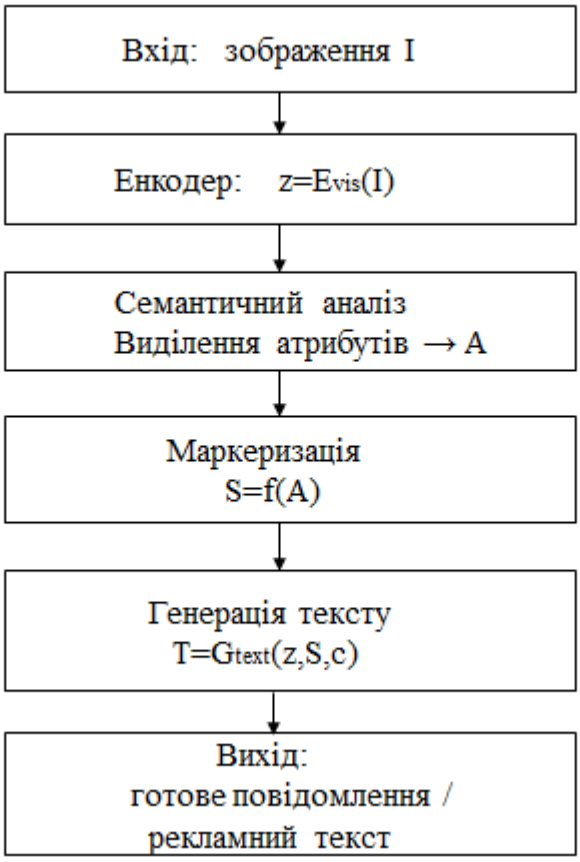


Рис. 3.8 Алгоритм роботи системи

ВИСНОВКИ

У третьому розділі кваліфікаційної роботи здійснено практичну реалізацію запропонованого методу розпізнавання зображень для подальшої генерації текстових описів на основі алгоритмів штучного інтелекту. Розроблений метод охоплює повний цикл обробки візуальної інформації — від моменту завантаження зображення користувачем до формування готового тексту оголошення, придатного для публікації на онлайн-платформах.

У підрозділі 3.1 детально описано етапи розробки методу, зокрема механізми завантаження зображень, попередньої обробки та підготовки вхідних даних. Запропонований підхід до попередньої обробки, що включає вирівнювання перспективи, фільтрацію шумів, нормалізацію кольору та автоматичне кадрування області документа, дозволяє суттєво підвищити якість подальшого розпізнавання. Особливу увагу приділено можливості обробки декількох зображень одного об’єкта, що є характерним для реальних сценаріїв використання, зокрема при роботі з офіційними документами.

У підрозділі 3.2 розглянуто програмні засоби та технології, використані для реалізації методу. Показано, що поєднання веб-фреймворку Yii2, сучасних інструментів обробки зображень та хмарних моделей штучного інтелекту дозволяє створити гнучку та масштабовану систему без необхідності використання складних інфраструктурних рішень. Такий підхід спрощує інтеграцію методу у вже існуючі веб-сервіси та знижує вимоги до апаратних ресурсів.

У підрозділі 3.3 детально описано реалізацію етапу розпізнавання тексту та візуальних ознак із використанням моделей OpenAI Vision. Продемонстровано, що застосування сучасних моделей комп’ютерного зору забезпечує не лише високу точність зчитування тексту, але й можливість контекстного аналізу структури документа. Формування структурованих даних у форматі JSON

створює надійне підґрунтя для подальшої семантичної інтерпретації та автоматизованої генерації текстів.

У підрозділі 3.4 розкрито принципи генерації тексту оголошення та його адаптації під потреби користувача. Показано, що використання мовних моделей у поєднанні зі структурованими даними дозволяє автоматично формувати логічно побудовані, інформативні та стилістично коректні тексти. Реалізовані механізми адаптації забезпечують можливість налаштування стилю, обсягу та змісту оголошення без зміни базової логіки алгоритму.

Таким чином, результати третього розділу підтверджують практичну придатність та ефективність запропонованого методу. Реалізоване рішення дозволяє суттєво спростити та оптимізувати процес створення текстових описів на основі зображень, зменшити трудомісткість ручної роботи та підвищити якість кінцевого результату. Запропонований метод може бути використаний як самостійне рішення або інтегрований у наявні веб-платформи для автоматизації створення оголошень та інших текстових матеріалів на основі візуальних даних.

Робота пройшла апробацію на конференціях та опублікована в наступних тезах доповідей на конференціях:

1. Сорока О.О., Щербина І.С. Соціальне значення методу створення оголошень через фото за допомоги штучного інтелекту. VI Всеукраїнська науково-технічна конференція «Застосування програмного забезпечення в інформаційно-комунікаційних технологіях» 24 квітня 2025 р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.294-297.

2. Сорока О.О., Щербина І.С. Бізнес-аспект розробки методу створення оголошень через фото за допомоги штучного інтелекту. VI Всеукраїнська науково-технічна конференція «Сучасний стан та перспективи розвитку ІОТ». 15 квітня 2025р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.175-179

ПЕРЕЛІК ПОСИЛАНЬ

1. Гнатюк, О. О. Комп’ютерний зір та його застосування. — Київ : Кондор, 2021. — 228 с.

2. Климчук, В. І. Штучний інтелект у прикладних задачах комп’ютерного зору. — Тернопіль : ТНТУ, 2022. — 256 с.

3. Ковальчук, А. Основи штучного інтелекту: сучасні методи та моделі. — Київ : КНУ, 2021. — 248 с.

4. Кравченко, Ю. В. Нейронні мережі та алгоритми штучного інтелекту. — Харків : Ранок, 2020. — 280 с.

5. Мельник, А. Нейронні мережі та глибинне навчання. — Львів : Новий Світ, 2022. — 312 с.

6. Мороз, Р. В. Обробка зображень і відео: сучасні технології. — Київ : КНТ, 2021. — 312 с.

7. Струтинський, В. Б. Основи глибинного навчання: теорія та практичні застосування. — Львів : ЛНУ, 2022. — 340 с.

8. Aggarwal, C. C. Machine Learning for Text. — Cham : Springer, 2021. — 436 c.

9. Bishop, C. Pattern Recognition and Machine Learning. Modernized edition. — New York : Springer, 2020. — 750 c.

10. Chollet, F. Deep Learning with Python. — 2nd ed. — New York : Manning Publications, 2021. — 504 c.

11. Dudek, G. Computational Principles of Mobile Robotics. — MIT Press, 2021. — 400 c.

12. Foster, D. Generative Deep Learning. — 2nd ed. — Sebastopol : O’Reilly Media, 2022. — 380 c.

13. Nakov, P. Natural Language Processing for Social Media. — London : Morgan & Claypool, 2021. — 210 c.

14. Nweke, H. F., Teh, Y. W., Al-Garadi, M. A. Deep Learning Applications: A Practical Approach. — Boca Raton : CRC Press, 2021. — 289 c.

15. Pal, S. K., Samanta, S., Dutta, P. Image Understanding Using Deep Learning. — Singapore : Springer, 2021. — 321 с.

16. Prince, S. J. D. Understanding Deep Learning. — Cambridge : MIT Press, 2023. — 624 с.

17. Raj, B., Singh, V. Machine Learning in Action: Advanced Concepts. — London : CRC Press, 2021. — 298 с.

18. Sarker, I. H. Machine Learning for Intelligent Data Analysis. — London : Springer, 2021. — 350 с.

19. Szeliski, R. Computer Vision: Algorithms and Applications. — 2nd ed. — Cham : Springer, 2022. — 1100 с.

20. OpenAI. OpenAI API Documentation. Документація OpenAI 2024. URL: https://platform.openai.com/docs (дата звернення 20.11.2025).

21. Keras. Keras Applications Overview. Огляд API моделей 2024. URL: https://keras.io/api/applications (дата звернення 25.11.2025).

22. Google AI Blog. Computer Vision Research Articles. Найновіші дослідження 2024. URL: https://ai.googleblog.com (дата звернення 25.11.2025).

23. Meta AI Research. Vision Transformer Research Materials. 2024. URL: https://ai.meta.com/research (дата звернення 27.11.2025).

24. Jason Brownlee. Convolutional Neural Networks Tutorials. 2023. URL: https://machinelearningmastery.com (дата звернення 27.11.2025).

25. Papers With Code. State-of-the-art Computer Vision Models. 2025. URL: https://paperswithcode.com (дата звернення 01.12.2025).

26. Towards Data Science. Image Recognition and AI Articles. 2024. URL: https://towardsdatascience.com (дата звернення 01.12.2025).

27. NVIDIA Developer Blog. Deep Learning Optimization Materials. 2024. URL: https://developer.nvidia.com/blog (дата звернення 01.12.2025).

ДОДАТОК А. ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Магістерська робота

«Метод розпізнавання зображень для подальшої генерації текстів на основі алгоритмів штучного інтелекту»

Виконав: студент групи ПДМ-63 Олександр СОРОКА

Керівник: канд. техн. наук, доцент кафедри ІПЗ Ірина ЩЕРБИНА

Київ - 2025

МЕТА, ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: оптимізація процесу автоматичного створення текстових описів на основі зображень за рахунок використання сучасних моделей комп'ютерного зору та алгоритмів штучного інтелекту.

Об'єкт дослідження: процес автоматичного опрацювання та інтерпретації візуальної інформації.

Предмет дослідження: засоби, технології та алгоритми штучного інтелекту, що забезпечують розпізнавання зображень і генерацію текстових описів.

ПОРІВНЯЛЬНА ХАРАКТЕРИСТИКА ІСНУЮЧИХ МОДЕЛЕЙ/МЕТОДІВ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ

Модель / Метод	Опис	Переваги	Недоліки
CNN (Convolutional Neural Networks)	Глибокі згорткові мережі, що виділяють ознаки зображень через фільтри різних масштабів. Використовуються для класифікації та об'єктного розпізнавання.	Висока точність у задачах обробки зображень; добре виділяють просторові ознаки; великий бази досліджень.	Погано працюють з абстрактними ознаками; складно адаптувати до нової генерації без додаткових модулів.
CLIP (Contrastive Language–Image Pre-training)	Модель від OpenAI, що вчить відповідність між зображеннями і текстовими описами, працюючи на мультимодальних даних.	Чудово розуміє семантику зображення; може генерувати релевантні підписи; добре працює для класифікації та фільтрації контенту.	Не генерує повноцінний текст самостійно; потребує зв'язки з LLM; залежить від великої кількості даних.
Vision Transformer (ViT)	Архітектура, що застосовує механізм Self-Attention для зображень, розбитих на патчі.	Висока точність; ефективна робота зі складною структурою зображень; легко інтегрується з мовними моделями.	Потребує великі набори даних; менш ефективний на низькороздільних ознаках; обчислювально важкий.
LLM-text generation (GPT-подібні моделі)	Генерують структуровані тексти, описи, комерційні оголошення на основі промптів або ознак.	Висока якість тексту; можливість створювати різні стилі; підтримка SEO та оптимізації під ринок.	Потребує чітких структурованих даних; іноді галюцинує некоректні факти (hallucinations).
Image Captioning (класичні моделі: Show&Tell, Show&Attend&Tell)	Побудовані на поєднанні CNN + RNN/Attention для створення коротких описів фото.	Добре працюють для простих описів; доведена ефективність; легкі в обчисленнях.	Обмежена довжина тексту; недостатність створювати комерційні або структуровані оголошення.
Meta-CLIP / OpenCLIP	Покращені CLIP-моделі з розширеними даними та краще узагальнювальною здатністю.	Краще розуміння контексту; точніше виділення семантичних ознак; більш стійка робота зі складними зображеннями.	Не створює текст; використовується лише як енкодер ознак.
Proposed Model (Модель з роботи)	Комбінує візуальний енкодер (CLIP/ViT), semantic refinement через базу знань та генерацію повноцінних оголошень через LLM.	Створює не просто опис, а повноцінне комерційне оголошення; висока релевантність; автоматична категоризація; оптимізація під маркетингові метрики.	Потребує складної інтеграції компонентів; залежить від якості бази знань та оптимізатора тексту.

3

МАТЕМАТИЧНА МОДЕЛЬ ГЕНЕРАЦІЇ ТЕКСТУ ПІСЛЯ ПРОЦЕСУ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ

Формальна постановка задачі

На вхід подається зображення

$$I \in \mathbb{R}^{H \times W \times 3}$$

Необхідно отримати текстовий опис – T, який максимально точно відображає зміст зображення.

Модель візуального подання

$$z = E_{vis}(I)$$

де:

E_{vis} - CNN/Vision Transformer

z — семантичний вектор ознак

Модель генерації тексту

$$T = G_{text}(z, c)$$

де:

G_{text} — трансформерний генератор

c — параметри стилю (формат оголошення)

Ключова ідея:

Розділення процесу розуміння зображення та генерації тексту, що підвищує якість результату.

4

МОДИФІКОВАНИЙ МЕТОД СЕМАНТИЧНОЇ ІНТЕРПРЕТАЦІЇ ЗОБРАЖЕНЬ

Семантична сегментація об'єкта

Виділення основних елементів сцени
Усунення фону
Виявлення ключових характеристик об'єкта
(форма, колір, тип, стан)

Результат: структурований набір атрибутів

$$A = \{a_1, a_2, ..., a_n\}$$

Перетворення атрибутів у текстові семантичні маркери

$$S = f(A)$$

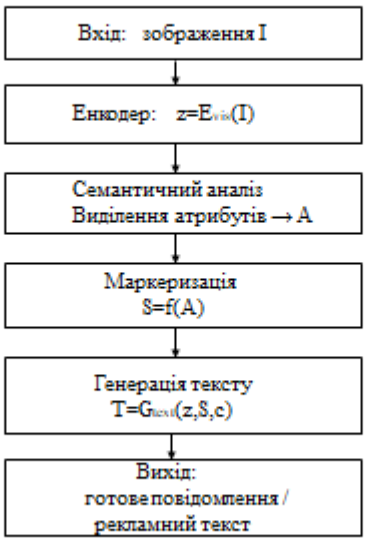
Маркеризація створює структуру, зрозумілу мовній моделі.

Керована генерація опису

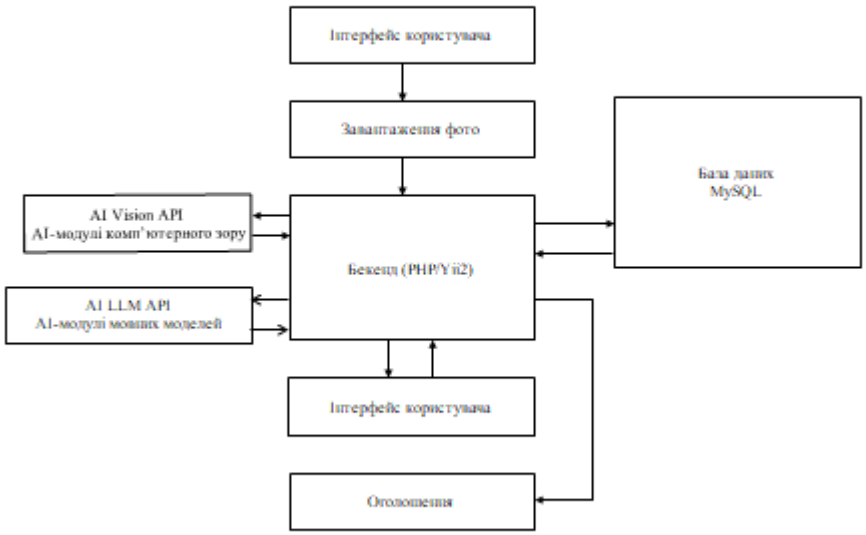
$$T = G_{text}(z, S, c)$$

Модель формує не загальний опис, а структурований рекламний текст.

СХЕМА ПРОЦЕСУ МЕТОДУ РОЗПІЗНАВАННЯ ЗОБРАЖЕНЬ ТА ПОДАЛЬШОЇ ГЕНЕРАЦІЇ ТЕКСТУ



ЗАГАЛЬНА АРХІТЕКТУРА ПРОГРАМНОЇ РЕАЛІЗАЦІЇ



Структура та інтерфейс веб-сервісу

Основні екранні форми

Головний екран – Завантаження
- Поле для завантаження зображення
- Відображення прев'ю
- Кнопка «Створити оголошення»

Екран обробки (Processing)
- Стан розпізнавання
- Інформація про визначені характеристики

Екран результату
- Готовий текст оголошення
- Можливість редагувати
- Кнопка «Скопіювати» або «Експорт»

Структура системи

Web-інтерфейс

API-шлюз

Модуль Vision AI
виділення ознак, класифікація, OCR

Модуль Semantic Engine
формування тексту оголошення

ПРАКТИЧНИЙ РЕЗУЛЬТАТ

Розроблено веб-сервіс, який автоматично створює рекламні оголошення на основі завантаженого фото.

Система поєднує моделі комп'ютерного зору та мовні моделі (LLM) для отримання структурованої інформації з зображення та генерації маркетингового тексту.

Основний функціонал сервісу:

Завантаження фотографії товару, автоматичне розпізнавання об'єктів, виділення характеристик та ключових параметрів, генерація маркетингового опису, формування готового оголошення (текст + оброблене фото).

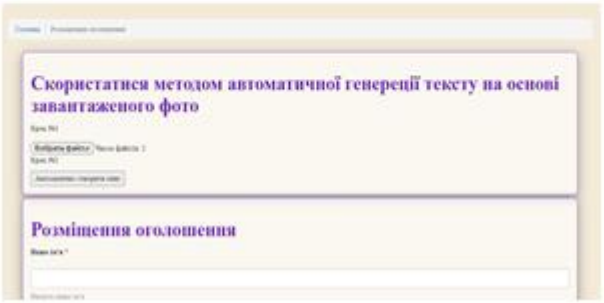
Крок 1 – завантажуємо фотографію

Крок 2 – розпізнаємо фотографію

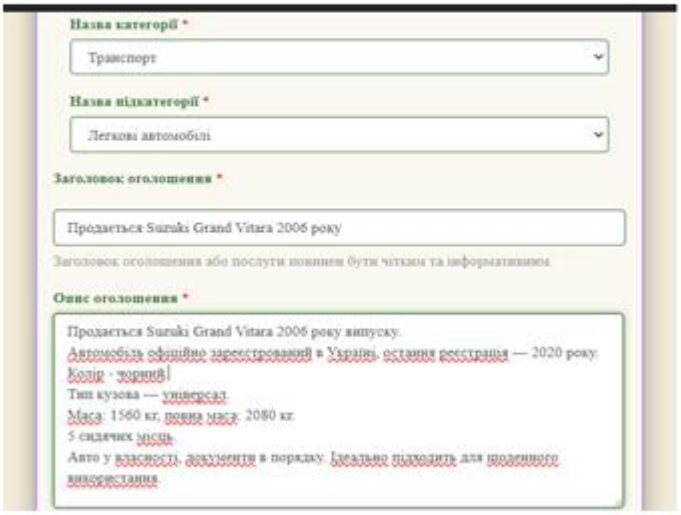
Крок 3 – генеруємо текст

9

ПРИКЛАД ЗАВАНТАЖЕНИХ ДОКУМЕНТІВ ТА ЕКРАННИХ ФОРМ ІНТЕРФЕЙСУ ЗАСТОСУНКУ



ЗГЕНЕРОВАНИЙ ТЕКСТ ОПИСУ



ВИСНОВКИ

1. Проаналізовано існуючі моделі та визначено соціальну значущість задачі: автоматичне створення оголошення на основі фото спрощує процес публікації для користувачів, зокрема для людей з обмеженими можливостями моторики, які не можуть швидко або якісно вводити текст вручну.

2. Розроблено метод перетворення зображення на рекламний текст, що поєднує аналіз візуальних ознак, семантичне структурування та генерацію маркетингово орієнтованого опису за допомогою моделей штучного інтелекту.

3. Створено алгоритм та архітектуру системи, які забезпечують автоматичний цикл «фото → структурований маркетинговий опис», а також реалізовано веб-модуль на PHP/Yii2 для практичної перевірки роботи методу.

4. Проведено моделювання та порівняльний аналіз, які показали, що розроблений метод збільшує інформативність описів (≈92%), покращує їх маркетингову якість (до 4.7/5) та зменшує час створення оголошення з 5-10 хвилин до 30 секунд, що підтверджує досягнення мети роботи.

ПУБЛІКАЦІЇ ТА АПРОБАЦІЯ РОБОТИ

Тези доповідей:

1. Сорока О.О., Щербина І.С. Соціальне значення методу створення оголошень через фото за допомоги штучного інтелекту. VI Всеукраїнська науково-технічна конференція «Застосування програмного забезпечення в інформаційно-комунікаційних технологіях» 24 квітня 2025 р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.294-297.

2. Сорока О.О., Щербина І.С. Бізнес-аспект розробки методу створення оголошень через фото за допомоги штучного інтелекту. VI Всеукраїнська науково-технічна конференція «Сучасний стан та перспективи розвитку ІОТ». 15 квітня 2025р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.175-179