Зміст

[Обґрунтування вибору теми/напрямку 1](#_Toc192873578)

[Постановка задачі/мета розробки і кінцевий результат 3](#_Toc192873579)

[Огляд та якісний аналіз існуючих рішень/подібних реалізацій 5](#_Toc192873580)

[Традиційні методи OCR 5](#_Toc192873581)

[Сучасні методи OCR на основі нейронних мереж 5](#_Toc192873582)

[Приклади сучасних рішень: 6](#_Toc192873583)

[Розпізнавання рукописного тексту 6](#_Toc192873584)

[Комерційні та Open-Source рішення 7](#_Toc192873585)

# Обґрунтування вибору теми/напрямку

Тема/напрямок: «Математичне та програмне забезпечення для розпізнавання та отримання тексту із зображень»

Дана тема була обрана, як тема бакалаврської роботи для напрацювання практичних навичок пов’язаних з:

* Мета розробки – напрацювання практичних навичок пов’язаних з:
  + 1. Проектуванням та втіленням архітектури ПЗ
    2. Побудовою архітектури, відлагодженням, підбором гіперпараметрів нейронної мережі.
    3. Аналізом якості роботи мережі за допомогою обраних метрик.
    4. Порівняльним аналізом роботи власноруч побудованої та відлагодженої нейронної мережі з розробленими раніше та більш досконалими моделями.

Оскільки в сучасному світі є потреба в автоматизації:

1. Ведення документообігу та, зокрема, сканування документів, з мінімальною похибкою.
2. Вилучення тексту написаного іноземною мовою в повсякденному житті (текст на вивісках, текст надрукований чи написаний на товарних пакунках)
3. Вилучення тексту в рамках наукової діяльності (транскрипція як древніх історичних текстів, написаних латиною, шумеро-аккадським клинописом, так і відносно сучасних, написаних за останні 100 років) для подальшого перекладу або оцифровування.
4. Вилучення тексту в рамках професійної діяльності (розпізнавання написаних власноруч викладачем оцінок/поміток на паперовій роздруківці перевірених тестів, сканування чеків, документів - заяв/довідок написаних від руки)

Задача розпізнавання та вилучення як друкованого, так і рукописного тексту залишається актуальною. Особливого значення в рамках вирішення цієї задачі набуває досягнення високого рівня точності роботи моделей, оскільки ціна помилки може мати вирішальне значення. Помилки в розпізнаванні можуть призводити до втрати важливих даних, юридичних неточностей або навіть фінансових збитків, особливо в сферах документообігу, архівування та автоматизації бізнес-процесів. Саме тому сучасні системи оптичного розпізнавання символів (OCR) активно розвиваються, використовуючи методи глибинного навчання, що дозволяють підвищувати точність і адаптивність до різних умов (зашумленості, низької якості зображення тощо) за умови наявності великої кількості даних, на яких проводитиметься навчання.

Зокрема, задача вилучення рукописного тексту видається більш складною і цікавою, оскільки друкований текст – за рахунок стандартизації шрифтів – є більш легким для розпізнавання. Рукописний текст може значно відрізнятися залежно від почерку кожної людини, змінюватися в межах одного документа, містити скорочення або нестандартні символи. Це створює додаткові виклики для систем розпізнавання, які мають враховувати контекст та особливості стилю написання, що і представляє цікавість для аналізу.

Сучасні підходи до вилучення рукописного тексту використовують рекурентні та згорткові нейронні мережі, це дозволяє підвищити точність навіть у випадках нестандартного почерку або низької якості сканованого зображення. Вивчення цих підходів, зокрема застосування різних конфігурацій нейронних мереж на практиці для вирішення цієї задачі представляє інтерес для розгляду ще й з точки зору, як саме зміна гіперпараметрів з якими навчалася нейронна мережа (та особливості її архітектури) вплине на її працездатність та якість розпізнавання на різних тестових прикладах (різні почерки, різний рівень зашумленості зображень, різна кольорова гамма).

# Постановка задачі/мета розробки і кінцевий результат

Постановка задачі:

1. Провести аналіз існуючих підходів до розпізнавання рукописного тексту, зокрема методів машинного навчання та нейронних мереж.
2. Дослідити математичні моделі попередньої обробки зображень, що покращують якість текстових даних (фільтрація шуму, бінаризація, сегментація тощо).
3. Розробити модель розпізнавання рукописного тексту на основі нейронної мережі, яка враховує особливості почерку та можливі варіації написання, зашумленість та низьку якість зображення.
4. Реалізувати програмне забезпечення, що забезпечує вилучення та розпізнавання слів із зображень.
5. Провести тестування та оцінити ефективність розробленого підходу на реальних даних, зокрема створених власноруч.
6. Провести порівняльний аналіз впливу різних архітектур/конфігурацій нейронних мереж на якість розпізнавання рукописного тексту.

Основною метою роботи є не скільки розробити якомога ефективнішу нейронну мережу, скільки проаналізувати вплив різних гіперпараметрів та особливостей архітектури нейронної мережі (кількості шарів згортки, pooling, LSTM). Оскільки, як показує практика останніх років, тут вставити посилання на дослідження де було показано, як точність розпізнавання повищується з імплементацією трансформерів <https://research.samsung.com/news/Samsung-R-D-Institute-Ukraine-s-Handwriting-Recognition-Team-in-Top-at-two-ICDAR2024-Competitions>

Якість розпізнавання (виміряна в метриках CER та WER) зростає, з доданням шарів трансформерів з механізмом уваги. А такі моделі набагато довше тренуються, що робить складнішим підбір потрібних гіперпараметрів, в ході постійних експериментів (оскільки навчання моделі проходить достатньо довго кожного разу).

Кінцевий результат:

* Розроблені моделі нейронної мережі CRNN (згорткової рекурентної нейронної мережі), здатні до розпізнавання та вилучення рукописних слів, написаних англійською мовою.
  + На вхід моделі подається зображення (фотографія) рукописного слова, наприклад *«example»*.
  + На виході модель видає слово в вигляді рядка (тип данних str в Python).
* Точність розробленої моделі для розпізнавання тексту (слів) для обраних метрик:
  + Середній **CER** (Character Error Rate) має бути **не більшим за 0,05**.
  + Середній **WER** (Word Error Rate) має бути **не більшим за 0,1**.
  + Модель повинна оцінюватись на **тестовому наборі даних**.
* Розроблене програмне забезпечення, обгортка для тренування та аналізу якості роботи нейронної мережі. Структура частини проєкту орієнтованої на це має бути організованою та легкою для розуміння й використання, оскільки модель не може бути досконалою, її архітектура буде постійно дороблюватися, а гіперпараметри змінюватися. Тому важливо добре організувати структуру проєкту, зробити її модульною та мінімізувати повтори коду, зробити код «чистішим». Крім того процес навчання моделі має бути оптимізованим, для мінімізації часових та матеріальних витрат.
* Розроблене програмне забезпечення – REST API та веб-застосунок з користувацьким інтерфейсом для легкої роботи користувача з моделлю. ПЗ має передбачати функцію завантаження зображення з рукописним текстом (словом) через форму, та отриманням транскрипції зображеного слова в відповідному текстовому полі.

# Огляд та якісний аналіз існуючих рішень/подібних реалізацій

Системи оптичного розпізнавання символів (OCR) – важливі компоненти автоматизації процесів обробки текстових даних. Вони використовуються для перетворення друкованого та рукописного тексту на цифровий формат, що дозволяє зберігати, редагувати та аналізувати інформацію. Розвиток OCR-технологій відбувався в кілька етапів, від простих алгоритмів шаблонного порівняння до складних моделей глибинного навчання.

### Традиційні методи OCR

Перші системи OCR базувалися на технологіях символьного розпізнавання, використовуючи шаблонне порівняння символів із заздалегідь визначеними еталонними зразками. Основні особливості таких систем:

* Висока залежність від типу шрифту.
* Потреба у великій кількості зразків для навчання.
* Відносно низька адаптивність до змін у тексті та зображенні.

Популярні програмні продукти, такі як **Tesseract OCR** (розроблений HP, зараз підтримується Google), використовували алгоритми сегментації зображення, виділення контурів символів та їх порівняння з вбудованими шаблонами.

### Сучасні методи OCR на основі нейронних мереж

Сучасні підходи OCR широко застосовують машинне навчання, особливо згорткові нейронні мережі (CNN) та рекурентні нейронні мережі (RNN), а також комбінації CNN + LSTM. Завдяки цим технологіям OCR-системи можуть розпізнавати текст із високою точністю навіть у складних умовах, таких як:

* Погана якість зображення.
* Варіативність шрифтів та рукописного тексту.
* Присутність шуму або артефактів на зображенні.

### Приклади сучасних рішень проблеми вилучення тексту:

**Tesseract OCR** (Google) – рушій, який можна використовувати для тренування власної моделі розпізнавання тексту за допомогою спеціальної мови *Variable Graph Specification Language (VGSL),* що дозволяє гнучке створення і налаштування нейронної мережі у вигляді рядка:

Наприклад, [1,0,0,1 Ct5,5,16 Mp3,3 Lfys64 Lfx128 Lrx128 Lfx256 O1c105]

Де кожний компонент визначає конкретний шар нейронної мережі (Згортковий шар, max-pooling, LSTM - шари). Перевага цього підходу полягає в тому, що не потрібно вивчати фреймворки для машинного навчання (PyTorch, Tensorflow) для створення конфігурації нейронної мережі.

Починаючи з версії 4.00, Tesseract OCR використовує складну архітектуру нейронних мереж, розроблену для підвищення точності розпізнавання тексту.

Основою цієї системи є комбінація згорткових нейронних мереж (CNN) та довготривалої короткочасної пам’яті (LSTM). Згорткові мережі відповідають за вилучення просторових особливостей із вхідних зображень, ефективно розпізнаючи форми та структури символів. Отримані характеристики передаються до LSTM – типу рекурентної нейронної мережі (RNN), здатного працювати з послідовними даними. Це дозволяє аналізувати порядок символів, розуміти контекст і покращувати точність розпізнавання.

Можливість CNN витягувати детальні просторові особливості, у поєднанні зі здатністю LSTM моделювати послідовності символів, робить Tesseract стійким та високоточним інструментом для розпізнавання тексту.

**EasyOCR** — це система оптичного розпізнавання символів з відкритим кодом, яка використовує складну архітектуру нейронної мережі для точного розпізнавання тексту в зображеннях. Основні компоненти EasyOCR включають: алгоритм CRAFT для виявлення текстових регіонів шляхом прогнозування символьних регіонів та оцінки спорідненості між символами; згорткову рекурентну нейронну мережу (CRNN), яка поєднує згорткові шари для вилучення просторових особливостей та рекурентні шари (LSTM) для моделювання послідовностей; і шар транскрипції з використанням Connectionist Temporal Classification (CTC) для перетворення виходу з рекурентних шарів на читабельний текст.

Хоча EasyOCR ефективно працює з друкованим текстом, він наразі офіційно не підтримує розпізнавання рукописного тексту.

**Google Cloud Vision OCR** та **Microsoft Azure OCR** – хмарні сервіси, які пропонують потужні алгоритми розпізнавання тексту, проте є комерційними рішеннями, тому архітектурні особливості OCR систем є невідомими.

**PaddleOCR** — це потужна система розпізнавання тексту з відкритим кодом, розроблена компанією Baidu на основі фреймворку глибокого навчання PaddlePaddle. Система створена для забезпечення багатомовних, надійних та практичних інструментів OCR, які допомагають користувачам тренувати кращі моделі та застосовувати їх на практиці. Архітектура PaddleOCR включає різноманітні модулі, три рівні:перший – це детектор тексту, який використовує модель DB (Differentiable Binarization) для точного виявлення текстових регіонів на зображенні; другий рівень — розпізнавач тексту, що використовує CRNN (Convolutional Recurrent Neural Network) з механізмом уваги для перетворення виявлених текстових областей у символи; третій - включає модуль корекції напрямку тексту та модуль покращення якості зображення, що дозволяє ефективно обробляти зображення з різною орієнтацією та якістю. Система постійно оновлюється — останні версії (2.9 і 2.10) включають десятки нових моделей, включаючи PP-DocLayout для макетів, PP-FormulaNet для розпізнавання формул та SLANeXt для обробки таблиць.

Особливістю PaddleOCR є широкий спектр можливостей: підтримка понад 80 мов, наявність різних конфігурацій моделей (від легких для мобільних пристроїв до високоточних для серверів), повний цикл розробки від навчання до розгортання, а також інтеграція з PaddleX — універсальним інструментом для розробки AI-рішень. Система представляє цілісні рішення, такі як PP-OCRv4, PP-Structure та PP-ChatOCRv4, які дозволяють обробляти складні документи, включаючи таблиці, формули та різні макети. Завдяки активній спільноті розробників і постійним оновленням, PaddleOCR залишається однією з найпопулярніших систем OCR з відкритим кодом, що знаходить застосування в різноманітних галузях — від обробки документів до аналізу зображень у реальному часі.

|  |  |  |  |
| --- | --- | --- | --- |
| **Назва OCR-системи** | **Тип (Open-Source/Комерційний)** | **Підтримка рукописного тексту** | **Використані технології** |
| Tesseract OCR | Open-Source | Так | На початку – шаблонне порівняння.  Зараз CNN + LSTM + CTC |
| EasyOCR | Open-Source | Ні | CRAFT – Character Region Awareness для знаходження участків зображення, де є текст  Архітектура моделі  CNN + LSTM + CTC |
| PaddleOCR | Open-Source | Так | DB (Differentiable Binarization) для виявлення участків зображення де є текст.  Модуль обробки зображень  Модель:  CNN + RNN + Attention |
| Google Cloud Vision OCR | Комерційний | Так | Невідомо |
| Microsoft Azure OCR | Комерційний | Так | Глибинне навчання |

Оскільки поєднання CNN + LSTM + CTC є найбільш часто використовуваним серед вищеописаних реалізацій OCR систем, показує високу ефективність, й не займає велику кількість часу для тренування моделі (на відміну від моделей трансформерів з увагою) було прийнято рішення побудувати аналогічну модель власноруч, експериментуючи з різними значеннями гіперпараметрів та різними архітектурними особливостями (кількість шарів згортки, розмір ядра згортки, кількість шарів LSTM, введення двонаправленої LSTM, тощо).