КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет комп'ютерних наук та кібернетики
Кафедра математичної інформатики
Центр науково-дослідних і дослідно-конструкторських
розробок ТОВ «Самсунг Електронікс Україна Компані»

Дипломна робота

на здобуття ступеня магістра

за спеціальністю 122 Комп'ютерні науки

на тему:

АНАЛІЗ СТРУКТУРИ ЗОБРАЖЕНЬ ДОКУМЕНТІВ ЗАСОБАМИ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Виконав студент 2 курсу магістратури
Захарчук Дмитро Юрійович _____
(підпис)

Науковий керівник:
професор, д.ф.-м.н.
Терещенко Василь Миколайович _____
(підпис)

Консультант (ментор ЦНДДКР
ТОВ «Самсунг Електронікс Україна Компані»):
доцент, к.т.н.
Дегтяренко Ілля В'ячеславович _____
(підпис)

Засвідчую, що в цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент _____
(підпис)

Роботу розглянуто й допущено до захисту
на засіданні кафедри
математичної інформатики
«___»_____ 2021 р.,
протокол №____
Завідувач кафедри математичної інформатики
професор Терещенко Василь Миколайович _____
(підпис)

Київ – 2021

АННОТАЦІЯ

Робота складається із 76 сторінок, 16 ілюстрацій, 9 таблиць, 60 джерел посилань, 7 додатків.

Ключові слова: АНАЛІЗ СТРУКТУРИ ДОКУМЕНТА, ВИЯВЛЕННЯ ОБ'ЄКТІВ, СЕМАНТИЧНА СЕГМЕНТАЦІЯ, ГЛИБОКІ НЕЙРОННІ МЕРЕЖІ, MEAN AVERAGE PRECISION, PUBLAYNET.

Процес аналізу структури та розмітки документа визначається як декомпозиція зображення на складові компоненти для розуміння їх функціональних зв'язків та залежностей.

Об'єктом дослідження є зображення, на яких містяться документи.

Предметом дослідження є моделі архітектур глибоких нейронних мереж, за допомогою яких здійснюється виявлення структури документа, а також інші алгоритми структурного аналізу документів.

Метою роботи є: розробка модифікованої архітектури глибокої нейронної мережі для семантичної сегментації об'єктів структури документів на зображеннях.

Поставлена мета потребує вирішення наступних задач:

- Дослідження існуючих алгоритмів аналізу структури документів;
- Аналіз існуючих наборів даних (датасетів) для здійснення сегментації зображень, що містять друковані документи;
- Порівняння існуючих архітектур моделей глибоких нейронних мереж, призначенням яких є візуальне розпізнавання об'єктів на зображеннях.

Матеріалом дослідження слугували набори даних (датасети) зображень, на яких розміщено документ.

У ході виконання роботи проаналізовано архітектури моделей глибоких нейронних мереж для візуального розпізнавання об'єктів – однофазні та двофазні детектори об'єктів. Здійснено навчання моделей даних типів детекторів на частині набору даних PubLayNet – датасету, присвяченому семантичній сегментації структурних елементів документів на зображеннях. В умовах проведених експериментів, результати засвідчили, що найкращу коректність

розпізнавання має модель мережі YOLOv5. До структури шарів даної мережі були внесені певні модифікації з метою покращення якості розпізнавання. Незважаючи на незначне погіршення усередненого значення за класами (0.911 – для оптимізованого підходу проти 0.914 для оригінального підходу), все ж таки вдалося досягти кращої якості у виявленні об’єктів класу «Текст». (0.840 – для оригінального підходу; 0.853 – оптимізований підхід, SGD; 0.855 – оптимізований підхід, Adam).

ЗМІСТ

СПИСОК СКОРОЧЕНЬ

Adam – Adaptive Moment Estimation

AP – Average Precision

CSPNet – Cross Stage Partial Network

DLA – Document Layout Analysis

EfficientDet – Efficient Detection

FCN – Fully Convolutional Network

FN – False Negative

FP – False Positive

FPN – Feature Pyramid Network

GELU – Gaussian Linear Error Unit

GPU – Graphical Processing Unit

IOU – Intersection over Union

LR – Learning Rate

mAP – mean Average Precision

PDF – Portable Document Format

R-CNN – Regions with Convolutional Neural Networks

ReLU – Rectified Linear Unit

RoI – Region of Interest

RPN – Region Proposal Network

SGD – Stochastic Gradient Descent

SPP – Spatial Pyramid Pooling

SVM – Support Vector Machine

TN – True Negative

TP – True Positive

XML – eXtensible Markup Language

YOLO – You Only Look Once

ВСТУП

Розпізнавання зображень сканованих документів з метою виявлення структурних частин документа у наш час стає все більш затребуваним завданням. Дана задача полягає у виявленні (локалізації) певних наперед визначених елементів структури документа та їх подальшій сегментації. Зазвичай, найпопулярнішим класом об’єктів, що їх необхідно розпізнати в документі, є таблиця. Це пов’язано з тим, що даний клас представляє собою велику цінність з точки зору інформаційної наповненості та необхідності швидкого редагування даних, представлених у таблицях. Поряд з цим класом об’єктів також виділяють такі як: зображення (фігури), абзаци тексту. Останній структурний елемент дуже часто поділяють на декілька окремих класів, зокрема: заголовки, списки (нумеровані, марковані), колонтитули. Залежно від тематики документів, представлених на зображеннях виділяють також деякі спеціальні класи структури документів: формули, матриці.

Актуальність роботи полягає у зростанні потреби видобування інформації із зображень з метою її збору, кластеризації та швидкого редагування.

Об’єктом дослідження є зображення, на яких містяться документи.

Предметом дослідження є моделі архітектур глибоких нейронних мереж, за допомогою яких здійснюється виявлення структури документа, а також інші алгоритми структурного аналізу документів.

Метою роботи є: розробка модифікованої архітектури глибокої нейронної мережі для семантичної сегментації об’єктів структури документів на зображеннях.

Поставлена мета потребує вирішення наступних задач:

- Дослідження існуючих алгоритмів аналізу структури документів;
- Аналіз існуючих наборів даних (датасетів) для здійснення сегментації зображень, що містять друковані документи;

- Порівняння існуючих архітектур моделей глибоких нейронних мереж, призначенням яких є візуальне розпізнавання об’єктів на зображеннях.

Матеріалом дослідження слугували набори даних (датасети) зображень, на яких розміщено документ.

Можливі сфери застосування. Отримана модель глибокої нейронної мережі може бути використана у системах розпізнавання структури документів та системах розпізнавання тексту.

РОЗДІЛ 1 АНАЛІЗ КЛЮЧОВИХ ЕТАПІВ РОЗВ'ЯЗАННЯ ЗАДАЧІ АНАЛІЗУ СТРУКТУРИ ДОКУМЕНТІВ

Процес аналізу структури та розмітки документа визначається як декомпозиція зображення на складові компоненти для розуміння їх функціональних зв'язків та залежностей. [1]

Задача аналізу структури (розмітки) документів включає в себе такі етапи та показана на рис. 1.1.: попередня обробка, аналіз структури та розмітки документа, класифікація сегментованих регіонів та оцінка результату.

Рис. 1.1 Етапи аналізу структури документів на зображенні

Попередня обробка включає в себе такі етапи: усунення шумів на зображенні, корекція кута нахилу зображення, бінаризація зображення (за потреби).

Фізична розмітка документа означає фізичне розташування та межі певних регіонів на зображенні документа [2]. Аналіз розмітки – це декомпозиція зображення документа в ієрархію гомогенних регіонів, таких як: текстові блоки, текстові лінії, фігури. Алгоритми аналізу розмітки діляться на два загальні підходи: висхідний (Bottom-up approach) та низхідний (Top-down approach). Відповідно, bottom-up підхід з маленьких зв'язних регіонів вибудовує більші гомогенні регіони, top-down підхід – єдине зображення поділяється на менші, логічно пов'язані між собою частини, що у підсумку формують даний регіон. Кожен з цих підходів має свої переваги у різних ситуаціях, однак інколи допускають певний гібридний підхід, що поєднує в собі дані два підходи.

Також, документи містять деяку додаткову інформацію, таку як заголовки, текст в колонтитулах, підписи таблиць, фігур та ін. Мітки таких елементів мають логічний, або функціональний характер. Сюди також відносять поняття порядку читання документа – побудова логічної послідовності текстових блоків документа. Сукупність логічних, функціональних елементів разом з їх взаємопов'язаними елементами (наприклад, фігура та підпис до неї, або текстовий блок, до якого подано заголовок є взаємопов'язаними елементами) називають логічною структурою документа.

1.1 Постановка задачі та сфери застосування

Аналіз структури документів є одним зі способів конвертації зображень документів до документів електронної форми, що знаходить своє застосування у системах розпізнавання тексту, системах інформаційного пошуку сканованих документів [3].

Формально дану задачу можна описати наступним чином: метою задачі є розпізнавання в документі семантично наповнених регіонів, що збігаються з класами структурної розмітки документа [4].

У документі D, що складається з дискретної множини компонентів $t = \{t_0, t_1 \dots, t_n\}$, кожен компонент $t_i = \big(r, (x_0, y_0, x_1, y_1)\big)$, де r – інформаційно наповнена складова компонента t_i, (x_0, y_0, x_1, y_1) – координати обмежувальної рамки визначеного регіону на зображенні, для яких встановлюється множина семантичних категорій $C = \{c_0, c_1, \dots, c_m\}$; відбувається встановлення відповідності між компонентами t_i та категоріями C.

Результатом виконання задачі є отримання деякої множини S, що описується формулою (1.1):

$$S = \{(\{t_0^0, \dots, t_0^n\}, c_0), \dots, (\{t_k^0, \dots, t_k^n\}, c_k)\} \quad (1.1)$$

Для якої виконується така функція: $F\colon (C, D) \to S$

Обробку документа можна розглядати з двох компонентів: розпізнавання текстової інформації для передачі її до системи розпізнавання тексту,

розпізнавання графічної інформації – видобування ілюстрацій зі сторінки документа, розпізнавання графічного каркасу діаграм та іншої нетекстової інформації. Наприклад, якщо раніше для внесення змін до роздрукованого на папері документа, потрібно було відсканувати документ, розпізнати на сканованому документі текстові регіони, конвертувати отримані документи-зображення до формату, що підтримується текстовими процесорами, та внести відповідні текстові зміни, то системи аналізу розмітки документа дозволять об'єднати всі перелічені етапи в один етап. Таким чином, стане можливим створення електронної копії документа не з його електронної копії, а з паперової копії роздрукованого документа.

1.2 Попередня обробка зображення

Етап попередньої обробки включає в себе такі процеси (рис. 1.2): усунення шумів, які виникли при скануванні документа, розділення зображення на передній та задній плани (foreground & background), корекція кута нахилу зображення документа. [2]

Рис. 1.2 Ключові етапи попередньої обробки зображення

1.2.1 Усунення шумів на зображенні

У роботі [5] описано деякі види шумів та методи їх усунення. Дуже часто рукописні тексти написано на папері, який розграфлено – на такому папері присутні горизонтальні лінії, що також можуть містити одну-дві вертикальні лінії, що позначають поля. Дана попередньо надрукована розмітка на папері становить певні труднощі, а саме: такі лінії перетинаються з текстом, неоднорідність ліній, а саме різна товщина, чи переривання такої лінії в певному

місці ускладнює роботу алгоритмів з виявлення таких ліній. Для даного виду шуму зазвичай застосовують одну із трьох груп методів: математичні методи морфології, що залежать від апріорних знань; друга група – з використанням перетворення Хафа, за допомогою якого визначаються ознаки текстових регіонів та з'ясовується напрямок таких ліній.

Ще один вид шумів – маргінальний, тобто будь-який текстовий або нетекстовий регіон за межами видимої зони сторінки. Найчастіше виникає при скануванні сторінок книжок. Методи боротьби з таким шумом поділяються на дві категорії – знаходження та усунення вищезгаданих регіонів, друга категорія спрямована на пошук оптимальної видимої зони сторінки, що визначається як найменший прямокутник, що включає у себе всі структурні елементи на передньому плані зображення документу.

На рис. 1.3 показано приклад маргінального шуму – чорна лінія, що розділяє сторінки книги при скануванні.

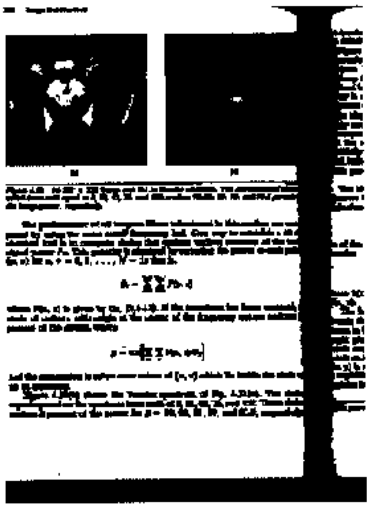


Рис. 1.3 Приклад маргінального шуму [6]

Клаттерний шум, або як його ще називають «шум безладу», – виникає на бінаризованих зображеннях та визначається як певний небажаний регіон, що значно перевищує область тексту. Даний шум має високу ступінь зв'язності з текстовими регіонами через те, що доволі часто перетинає їх. У роботі [6] пропонується такий алгоритм: зменшується роздільна здатність зображення, далі зображення розбивається на блоки, у яких здійснюється пошук клаттерного

шуму на основі таких припущень як розміщення та розміри відповідної області, що містить даний шум. Приклад клаттерного шуму показано на рис. 1.4.

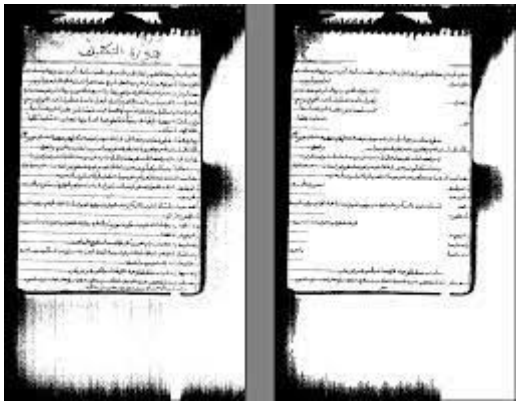


Рис 1.4 Приклад клаттерного шуму [7]

1.2.2 Бінаризація зображень

Бінаризація зображення – процес конвертації мультитонального зображення в бітональне зображення [8]. Іншими словами, бінаризоване зображення містить лише два кольори. У випадку зображень документів – це процес виділення текстових регіонів (переднього плану) від решти зображення.

Частіше за все бінаризація застосовується до зображення у чорно-білих тонах, оскільки при конвертації кольорового зображення у чорно-біле втрачається порівняно мала кількість інформації.

Методи бінаризації для чорно-білих зображень поділяють на дві загальні категорії:

- Методи глобального порогування;
- Методи локального порогування.

1.2.2.1 Бінаризація Отцу

Одним із найпоширеніших алгоритмів глобального порогування є алгоритм Отцу [9]. Нехай розглядається зображення документу у чорно-білих тонах, де $g(x, y) \in [0,255]$ – інтенсивність кольору пікселя з координатами (x, y). Задача глобального порогування полягає у знаходженні деякого порогу T, для якого пікселі, що мають інтенсивність кольору $\leq T$ будуть включені до класу

переднього плану зображення (foreground), а всі інші пікселі – до класу заднього плану (background). Інтенсивності кольору пікселів у вихідному зображенні визначаються формулою (1.2):

$$B(i,j) = \begin{cases} 0, & g(x,y) < T \\ 255, & g(x,y) \geq T \end{cases} \quad (1.2)$$

Поріг T визначається за допомогою гістограми інтенсивності кольору пікселей – h – яка для 8-бітного зображення включає в себе $L = 256$ корзин. Для кожного обраного порогу $0 \leq T \leq L$ гістограма розділяється на два кластери.

Число пікселів кожного класу визначається за формулою (1.3):

$$w_0(T) = \sum_{i=0}^{T-1} h(i) \qquad w_1(T) = \sum_{i=T}^{L-1} h(i) \quad (1.3)$$

Середня інтенсивність кольору пікселів визначається за формулою (1.4):

$$\mu_0(T) = \frac{1}{w_0} \sum_{i=0}^{T-1} ih(i) \qquad \mu_1(T) = \frac{1}{w_1} \sum_{i=T}^{L-1} ih(i) \quad (1.4)$$

Дисперсія кластерів визначається за формулами (1.5) – (1.6) :

$$\sigma_0^2(T) = \frac{1}{w_0} \sum_{i=0}^{T-1} h(i)\big(i - \mu_0(T)\big)^2 \quad (1.5)$$

$$\sigma_1^2(T) = \frac{1}{w_1} \sum_{i=T}^{L-1} h(i)\big(i - \mu_1(T)\big)^2 \quad (1.6)$$

Таким чином, T визначається як поріг, що мінімізує внутрішьокластерну дисперсію:

$$\underset{T}{\operatorname{argmin}}\, w_0(T)\sigma_0^2(T) + w_1(T)\sigma_1^2(T) \quad (1.7)$$

Це еквівалентно максимізації дисперсії між класами:

$$\underset{T}{\operatorname{argmax}}\, w_0(T)w_1(T)\big(\mu_0(T) - \mu_1(T)\big)^2 \quad (1.8)$$

1.2.2.2 Бінаризація Ніблека

Бінаризація Ніблека [10] є прикладом алгоритму адаптивного локального порогування. Поріг визначається для кожного пікселя на основі статистичних даних локального регіону з центром з координатами пікселя, що досліджується.

Середнє значення та дисперсія інтенсивності кольору пікселів визначаються за формулами (1.9) – (1.10):

$$\mu(i,j) = \frac{1}{w^2} \sum_{i'=i-w}^{i+w} \sum_{j'=j-w}^{j+w} I(i',j') \tag{1.9}$$

$$\sigma(i,j) = \sqrt{\frac{\sum_{i'=i-w}^{i+w} \sum_{j'=j-w}^{j+w} \left(I(i',j') - \mu(i,j)\right)^2}{w^2}} \tag{1.10}$$

Де w – розмір регіону, для якого обчислюється вищезгадана статистика.

Поріг Ніблека визначається за формулою (1.11):

$$T_N(i,j) = \mu(i,j) + k\sigma(i,j) \tag{1.11}$$

Де k – параметр, що задається вручну і визначає лінію тренду між precision та recall для класу переднього плану.

Рекомендовано обрати $k = -0.2$ Однак оптимальне значення параметра залежить від зображення та обраного розміру регіону.

Вихідне зображення визначається наступним чином:

$$B(i,j) = \begin{cases} 0, & I(i,j) < T_N(i,j) \\ 255, & I(i,j) \geq T_N(i,j) \end{cases} \tag{1.12}$$

У даному алгоритмі виникають труднощі, коли досліджуваний регіон повністю складається із пікселів, що належать кластеру заднього плану. У цьому випадку пікселі з меншою інтенсивністю кольору потрапляють до кластеру переднього плану.

1.2.2.3 Бінаризація Сауволи

Алгоритм, який запропонував Саувола [11] являє собою модифікацію алгоритму Ніблека.

$$T_S(i,j) = \mu(i,j)\left[1 + k(\frac{\sigma(i,j)}{R} - 1)\right] \quad (1.13)$$

$\mu(i,j)$ та $\sigma(i,j)$ обчислюються так само, як і в алгоритмі Ніблека, параметр k обирається рівним $k = 0.5$; параметр R – максимальне значення дисперсії (Для 256 рівнів інтенсивності кольору $R = 128$)

Значення середньої інтенсивності пікселів та дисперсії залежить від контрасту регіону, який досліджується. Для регіонів з високим контрастом значення дисперсії буде наближено дорівнювати параметру R, а отже, значення порогу T_S наближено дорівнює значенню середньої інтенсивності пікселів $\mu(i,j)$. Звідси можна зробити висновок, що для регіонів, більшість пікселів яких відносяться до класу заднього плану, значення порогу T_S буде меншим, ніж значення порогу за алгоритмом Ніблека, T_N, відповідно у результаті бінаризації менша кількість пікселів заднього плану потрапить до класу переднього плану, і така бінаризація буде ефективнішою.

1.3 Виявлення рамки сторінки

Ще одним етапом попередньої обробки зображення документа є виявлення рамки сторінки. Задача знаходження рамки сторінки полягає в виокремленні області зображення, де зосереджено контент даної сторінки, відкидаючи маргінальний шум по боках сторінки [12].

Маргінальний шум навколо боків сторінки належить до одного з двох типів:

- Нетекстовий маргінальний шум (у вигляді чорних вертикальних смуг та окремих цяточок);
- Текстовий маргінальний шум (частини сусідньої сторінки);

Одним із способів для очищення зображення від нетекстового маргінального шуму є метод зв’язних компонент [13].

Для усунення маргінального текстового шуму застосовується алгоритм визначення видимої зони сторінки. На сьогодні існує безліч OCR-сканерів, що дозволяють користувачу власноруч обрати певний регіон зображення для

сканування. Однак, при великій кількості зображень ручне налаштування регіонів для сканування не є доцільним.

Приклади виявлення рамки сторінки на зображенні показано на рис. 1.5.

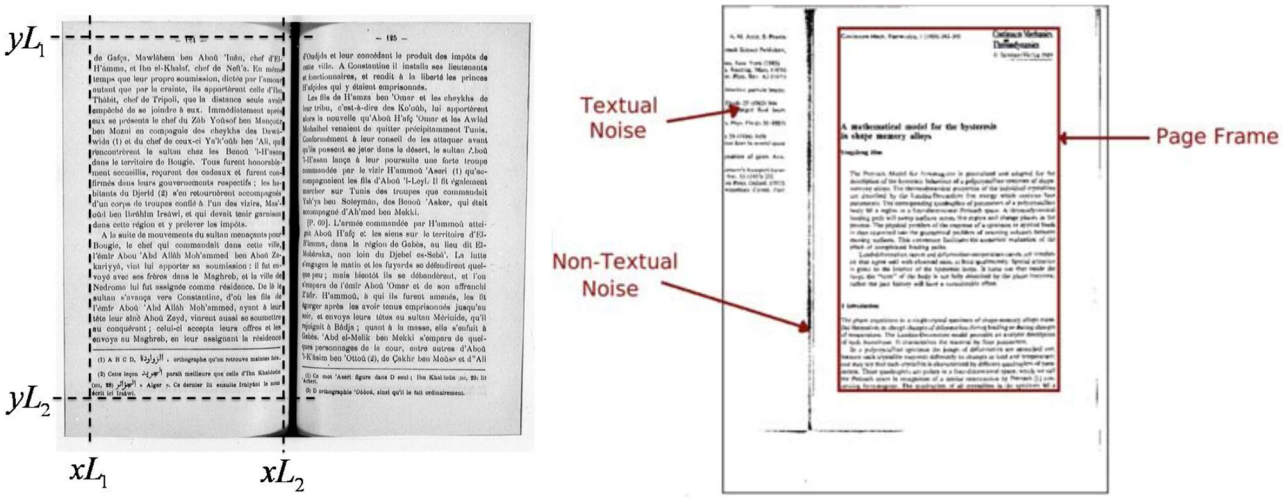


Рис. 1.5 Виявлення рамки сторінки [12, 14]

Більшість алгоритмів спрямовані на одночасне усунення двох вищезгаданих типів маргінального шуму. Так, у роботі [15] запропоновано алгоритм на основі правил, що використовує деякі евристики для виявлення меж сторінки. Здійснюється класифікація рядків та стовпчиків документа на три класи: «текст», «не текст» та «інше», після чого застосовується аналіз проекцій профіля. Даний алгоритм базується на припущенні, що межі документа не перетинаються з краями рамки сторінки. Однак, дане припущення у багатьох випадках не виконується.

У роботі [16] описано алгоритм для усунення маргінального шуму на чорно-білих зображеннях на основі статистичних даних зображення, таких як вертикальні та горизонтальні різницеві вектори. Але даний алгоритм не розрахований на роботу з уже бінаризованими зображеннями. Також, даний алгоритм є чутливим до кількості шуму, оскільки похибка алгоритму зростає монотонно зі збільшенням площі зашумленої зони.

У роботі [17] пропонується алгоритм розпізнавання меж зашумлених регіонів на основі аналізу застосовування підходу проекції профілів для країв зображення.

Даний алгоритм передбачає, що документи були відскановані в однакових умовах, і неможливо забезпечити коректну роботу алгоритму в «реальних» умовах використання.

У роботі [12] представлено алгоритм виявлення зони сторінки за допомогою геометричного пошуку.

Будується геометрична модель для рамки сторінки документа, після чого застосовується метод геометричного пошуку з використанням деякої функції якості.

Структура структурованих документів може бути представлена у вигляді ієрархії, де кожен рівень ієрархії позначає деякий рівень інформативності, прикладами яких є зони, текстові лінії, текстові блоки. Модель документа, представлена у роботі [18] подана як ієрархія зон, текстових блоків та текстових ліній з визначеним порядком читання між даними елементами ієрархії. Однак, така модель враховує лише контент, наявний на сторінці документу, та не враховує розміщення можливих зашумлених регіонів.

У роботі [12] представлено ієрархічну модель документа, у якій введено додатковий рівень ієрархії – рамка сторінки документа. Дана модель містить такі рівні ієрархії:

- Бінарне зображення документа D визначається як об'єднання множини пікселів переднього плану p_f та пікселів заднього плану p_b;
- Множина пікселів переднього плану подається у вигляді розбиття на зв'язні компоненти $C = \{C_1, \dots C_M\}$ таке, що $C_i \cap C_j = \emptyset \; \forall i \neq j$ та $\bigcup_{i=1}^{M} C_i = p_f$;
- Множина зон $Z = \{Z_1, \dots Z_R\}$, де зона визначається як $Z_i \subseteq C \;\; Z_i \cap Z_j = \emptyset \; \forall i \neq j$; Кожна зона містить лише один елемент структури документа;
- Множина текстових ліній визначається як розбиття множини зв'язних компонент C: $L = \{L_1, \dots L_N\}$ таке, що $L_i \subseteq C \;\; L_i \cap L_j = \emptyset \; \forall i \neq j$; деякі зв'язні компоненти можуть не належати жодній з текстових ліній;

- Рамка сторінки F визначається як прямокутник з найменшою площею, що включає в себе усі зв'язні компоненти, що належать документу, поданому на зображенні.

Поділ на текстові лінії здійснено за допомогою алгоритму, описаного у роботі [13], поділ на зони здійснено за допомогою алгоритму на основі діаграми Вороного, описаного в роботі [19].

Рамка сторінки визначається як прямокутник, що описується п'ятьма параметрами: $\vartheta = \{l, t, r, b, a\}$. Ці параметри позначають координати лівої нижньої та правої верхньої точок; a – кут нахилу зображення.

Оскільки на вхід алгоритму, що виявляє рамку сторінки, подаються вже повернуті зображення, рамка сторінка визначатиметься за чотирма параметрами: $\vartheta = \{l, t, r, b\}$. Для задачі виявлення рамки потрібно підібрати такі параметри, щоб функція якості Q, що визначається за формулою (1.14), набувала максимального значення:

$$\hat{\vartheta}(C, L, Z) = \underset{\vartheta \in T}{\operatorname{argmax}} Q(\vartheta, C, L, Z) \qquad (1.14)$$

Проектування вигляду функції якості виконано за допомогою дослідження властивості вирівнювання тексту. У структурованих документах текст зазвичай вирівняно або за шириною, або за лівим краєм. Це дозволяє вважати, що переважна більшість зв'язних компонент, виявлених по боках сторінки, може бути одним із критеріїв функції якості, а саме обмежувальні рамки (bounding box) текстових символів, які, у свою чергу, перетинають межу рамки сторінки.

Однак, даний підхід не є оптимальним з таких причин: текстові лінії, що розміщуються з самого верху та низу, зазвичай, містять меншу кількість символів, у порівнянні з іншими текстовими лініями на сторінці, особливо це стосується текстової лінії, що містить в собі номер сторінки. Крім того, зверху та внизу сторінки може міститися нетекстова інформація, наприклад, фігури. Отже, точність параметрів t та b не слід брати уваги на символьному рівні, тоді як параметри l та r мають сенс лише для тексту, що вирівняний за шириною сторінки. Таким чином, замість зв'язних компонент символьного рівня можна

використати текстові лінії. На рівні текстових ліній функція якості розглядається як кількість текстових ліній, що дотикаються до межі рамки сторінки. Відповідно, здійснюється декомпозиція множини параметрів $\vartheta = \{l, t, r, b\}$ на дві множини: $\vartheta_v = \{t, b\}$ та $\vartheta_h = \{l, r\}$ Спершу параметри ϑ_v встановлюють рівними своїм максимальним значенням $t = 0, b = H$, де $H -$ висота сторінки, після цього здійснюється пошук оптимальних параметрів ϑ_h. Дана декомпозиція дозволяє зменшити розмірність простору для пошуку оптимальних параметрів. Тоді формулу (1.14) можна переписати у вигляді:

$$\hat{\vartheta}_h(L) = \underset{\vartheta_h \in T}{\operatorname{argmax}} Q(\vartheta_h, L) \tag{1.15}$$

Верхня межа функції якості Q записується у вигляді суми локальних функцій якості:

$$Q(\vartheta_h, L) = \sum_{j=1}^{N} q(\vartheta_h, L_j) \tag{1.16}$$

Нехай $L = \{x_0, y_0, x_1, y_1\}$ – обмежувальна рамка текстової лінії, $d(l, x_i)$ та $d(r, x_i)$ – інтервали імовірних відстаней від точки з координатою x_i до параметризованої точки з координатами l та r відповідно. Тоді, локальна функція якості q для заданої текстової лінії в просторі параметрів визначається за формулою (1.17):

$$q_1(\vartheta_h, (x_0, x_1)) = max\left(0, 1 - \frac{d^2(l, x_0)}{\varepsilon^2}\right) + max\left(0, 1 - \frac{d^2(r, x_1)}{\varepsilon^2}\right) \tag{1.17}$$

Де ε означає граничну відстань, за якої текстова лінія робить внесок у рамку сторінки. У межах однієї колонки тексту початкові та кінцеві позиції текстових ліній можуть відрізнятись, залежно від вирівнювання тексту. Для документів з кількома текстовими колонками, алгоритм з функцією якості у такому вигляді оптимальним розв'язком вважатиме колонку тексту з найбільшою кількістю текстових ліній. Щоб вирішити дану проблему, вводяться подібні до тих, що наявні у формулі, доданки зі знаком мінус: для тих текстових ліній, які вказують на «інший» бік рамки сторінки (тобто ті, де x_1 корегує

параметр l замість r, та для ліній, де x_0 корегує параметр r замість l). Дані доданки описуються наступною формулою (1.18):

$$q_2(\vartheta_h, (x_0, x_1)) = -max\left(0, 1 - \frac{d^2(l, x_1)}{(2\varepsilon)^2}\right) - max\left(0, 1 - \frac{d^2(r, x_0)}{(2\varepsilon)^2}\right) \quad (1.18)$$

Керуючись формулами (1.17) – (1.18), формула функції якості набуває вигляду:

$$q(\vartheta_h, (x_0, x_1)) = q_1(\vartheta_h, (x_0, x_1)) + q_2(\vartheta_h, (x_0, x_1)) \quad (1.19)$$

1.4 Сегментація сторінки

Сегментація сторінки – один із ключових етапів попередньої обробки зображення документа. Сегментація – процес поділу зображення на гомогенні регіони, тобто такі, що містять в собі лише один тип інформації, наприклад, фігура, текстовий блок, таблиця.

Алгоритми сегментації сторінки поділяються на три типи [20]:

- Висхідні методи (Bottom-up approaches);
- Низхідні методи (Top-down approaches);
- Гібридні методи (Hybrid approaches).

У Top-down методах єдине цілісне зображення ітеративно розбивається на менші регіони. У bottom-up методах обирається деякий рівень ієрархії (найнижчий рівень ієрархії – пікселі зображення), далі структурні елементи обраного рівня ієрархії об’єднуються у більші за розміром гомогенні регіони – пікселі утворюють множину окремих зв’язних компонент, дані зв’язні компоненти об’єднуються в слова, текстові блоки, аж до моменту об’єднання в наперед визначеними певним критерієм зони. Гібридні методи містять в собі ознаки, що поєднують Top-down та Bottom-up підходи.

До висхідних методів належать:

- Алгоритм Docstrum [21];
- Алгоритм на основі діаграми Вороного [19];
- Алгоритм Уола [22];

- Алгоритм Флетчера та Кастурі [23].

До низхідних методів належать:

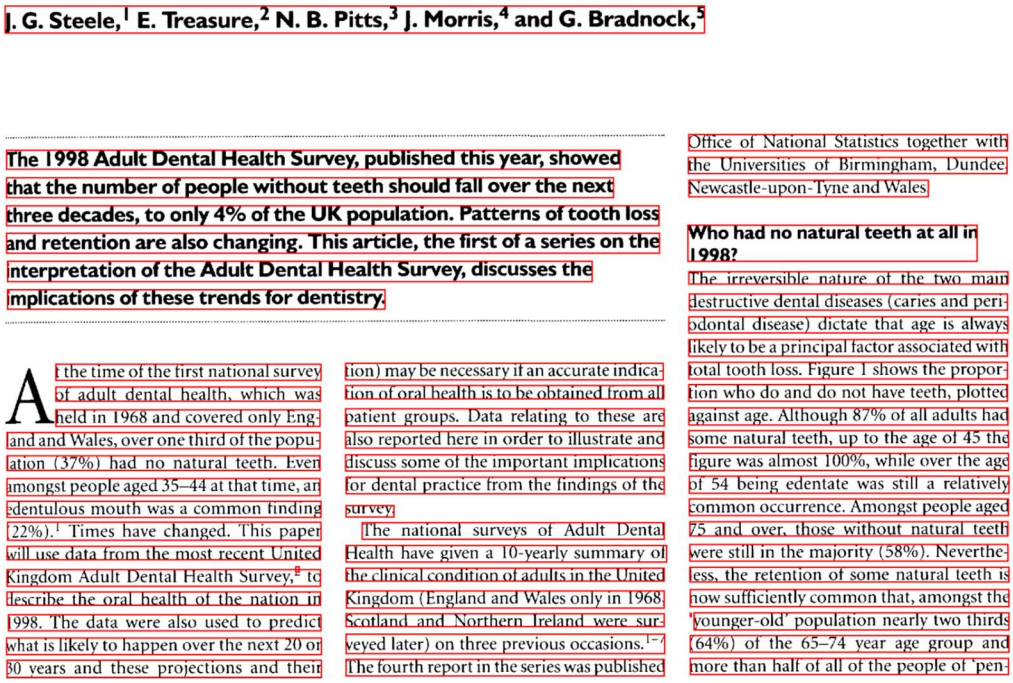
- X-Y Cut [24];
- Алгоритм Бейрда [25].

До гібридних методів належить метод, запропонований Павлідісом та Жоу [26].

1.4.1 Алгоритм Docstrum

Алгоритм Docstrum працює із зображеннями сторінок документів із неманхетенською розміткою. (Манхетенська розмітка сторінки документу виникає тоді, коли між структурними елементами сторінки наявні перпендикулярні лінії роздільники, що не несуть із собою інформаційної наповненості; відповідно неманхетенька розмітка сторінки – структурні елементи сторінки можуть перетинатися один з одним – наприклад, деяка фігура обтикає блок тексту або перетинає його).

Приклад сегментації сторінки алгоритмом Docstrum на рівні текстових ліній показано на рис. 1.6.


J. G. Steele,[1] E. Treasure,[2] N. B. Pitts,[3] J. Morris,[4] and G. Bradnock,[5]

The 1998 Adult Dental Health Survey, published this year, showed that the number of people without teeth should fall over the next three decades, to only 4% of the UK population. Patterns of tooth loss and retention are also changing. This article, the first of a series on the interpretation of the Adult Dental Health Survey, discusses the implications of these trends for dentistry.

At the time of the first national survey of adult dental health, which was held in 1968 and covered only England and Wales, over one third of the population (37%) had no natural teeth. Even amongst people aged 35–44 at that time, an edentulous mouth was a common finding (22%).[1] Times have changed. This paper will use data from the most recent United Kingdom Adult Dental Health Survey,[8] to describe the oral health of the nation in 1998. The data were also used to predict what is likely to happen over the next 20 or 30 years and these projections and their implications for dentistry are discussed.

The national surveys of Adult Dental Health have given a 10-yearly summary of the clinical condition of adults in the United Kingdom (England and Wales only in 1968, Scotland and Northern Ireland were surveyed later) on three previous occasions.[1–7] The fourth report in the series was published by the Office of National Statistics together with the Universities of Birmingham, Dundee, Newcastle-upon-Tyne and Wales.

Who had no natural teeth at all in 1998?

The irreversible nature of the two main destructive dental diseases (caries and periodontal disease) dictate that age is always likely to be a principal factor associated with total tooth loss. Figure 1 shows the proportion who do and do not have teeth, plotted against age. Although 87% of all adults had some natural teeth, up to the age of 45 the figure was almost 100%, while over the age of 54 being edentate was still a relatively common occurrence. Amongst people aged 75 and over, those without natural teeth were still in the majority (58%). Nevertheless, the retention of some natural teeth is now sufficiently common that, amongst the 'younger-old' population nearly two thirds (64%) of the 65–74 year age group and more than half of all of the people of 'pen-

Рис. 1.6. Сегментація сторінки алгоритмом Docstrum [2]

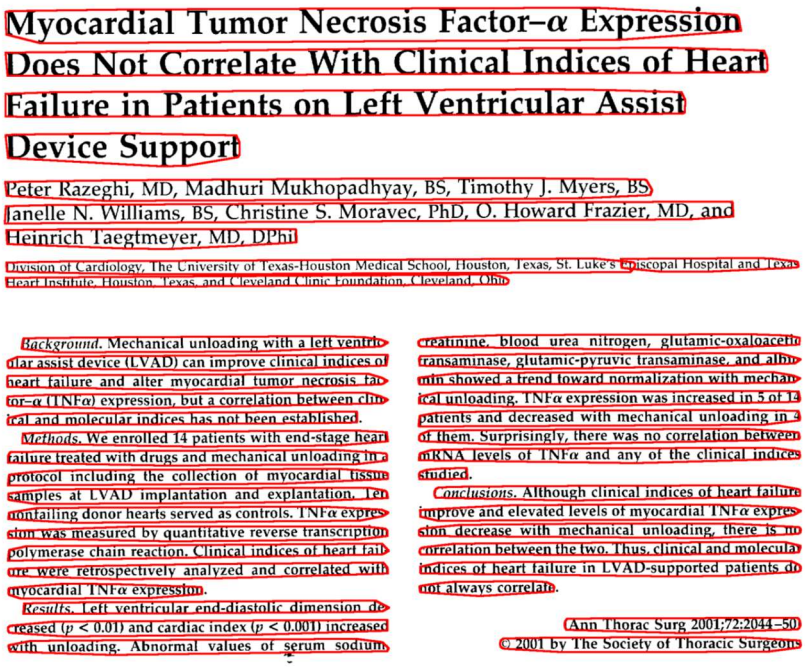
Алгоритм Docstrum:

1. Знайти зв’язні компоненти за допомогою two-pass алгоритму [27];
2. Прибрати замалі або завеликі зашумлені регіони, або зв’язні компоненти нетекстових регіонів за допомогою порогів l та h;
3. Розділити зв’язні компоненти текстових регіонів C_i на два кластери – перший – ті текстові блоки, які належать до заголовків (titles, headings), другий кластер – усі інші текстові регіони;
4. Для кожного регіону із C_i Знайти K найближчих сусідів $NN_K(i)$ за допомогою алгоритму сортування;
5. Обчислити відстань та кут між кожним регіоном із C_i та його K-тим найближчим сусідом: $\left(p_j^i, \theta_j^i\right), j \in NN_K(i)$;
6. Побувати гістограму відстаней між внутрішьорядковими інтервалами для регіонів, визначених у кроці 4 на основі множини W_p, що визначається наступним чином: $W_p = \left\{p_j^i \middle| \, j \in NN_K(i);\ -\theta_h \le \theta_j^i \le +\theta_h\right.$, де θ_h – поріг допустимого інтервалу для горизонтального кута. Присвоїти значення cs інтервалу між текстовими знаками в межах рядка рівним піковому значенню побудованої гістограми;
7. Побувати гістограму відстаней міжрядкових інтервалів для регіонів, визначених у попередньому кроці на основі множини W_p, що визначається наступним чином: $B_p = \left\{p_j^i \middle| \, j \in NN_K(i);\ 90 - \theta_v \le \theta_j^i \le 90 + \theta_v\right.$, де θ_v – поріг допустимого інтервалу для вертикального кута. Присвоїти значення ls міжрядкового інтервалу рівним піковому значенню побудованої гістограми.
8. Виконати транзитивне замикання пар вищезгаданих регіонів. За допомогою порогу внутрішньорядкових відстаней T_{cs}, що визначається: $T_{cs} = f_t \cdot cs$; отримати текстові лінії L_i;
9. Виконати транзитивне замикання текстових ліній з метою отримання структурних блоків Z_i, застосовуючи поріг паралельних відстаней $T_{pa} = f_{pa} \cdot cs$ та поріг перпендикулярних відстаней $T_{pe} = f_{pe} \cdot ls$;

1.4.2 Алгоритм сегментації Кісе на основі діаграми Вороного

Алгоритм Кісе на основі діаграми Вороного також належить до сімейства Bottom-up.

Приклад сегментації сторінки даним алгоритмом показано на рис 1.7.



Myocardial Tumor Necrosis Factor–α Expression Does Not Correlate With Clinical Indices of Heart Failure in Patients on Left Ventricular Assist Device Support

Peter Razeghi, MD, Madhuri Mukhopadhyay, BS, Timothy J. Myers, BS, Janelle N. Williams, BS, Christine S. Moravec, PhD, O. Howard Frazier, MD, and Heinrich Taegtmeyer, MD, DPhil

Division of Cardiology, The University of Texas-Houston Medical School, Houston, Texas, St. Luke's Episcopal Hospital and Texas Heart Institute, Houston, Texas, and Cleveland Clinic Foundation, Cleveland, Ohio

Background. Mechanical unloading with a left ventricular assist device (LVAD) can improve clinical indices of heart failure and alter myocardial tumor necrosis factor–α (TNFα) expression, but a correlation between clinical and molecular indices has not been established.

Methods. We enrolled 14 patients with end-stage heart failure treated with drugs and mechanical unloading in a protocol including the collection of myocardial tissue samples at LVAD implantation and explantation. Ten nonfailing donor hearts served as controls. TNFα expression was measured by quantitative reverse transcription polymerase chain reaction. Clinical indices of heart failure were retrospectively analyzed and correlated with myocardial TNFα expression.

Results. Left ventricular end-diastolic dimension decreased ($p < 0.01$) and cardiac index ($p < 0.001$) increased with unloading. Abnormal values of serum sodium, creatinine, blood urea nitrogen, glutamic-oxaloacetic transaminase, glutamic-pyruvic transaminase, and albumin showed a trend toward normalization with mechanical unloading. TNFα expression was increased in 5 of 14 patients and decreased with mechanical unloading in 4 of them. Surprisingly, there was no correlation between mRNA levels of TNFα and any of the clinical indices studied.

Conclusions. Although clinical indices of heart failure improve and elevated levels of myocardial TNFα expression decrease with mechanical unloading, there is no correlation between the two. Thus, clinical and molecular indices of heart failure in LVAD-supported patients do not always correlate.

(Ann Thorac Surg 2001;72:2044–50)

© 2001 by The Society of Thoracic Surgeons

Рис. 1.7 Сегментація сторінки алгоритмом Кісе на основі діаграми Вороного [2]

Кроки алгоритму наступні:

1. Здійснити маркування зв'язних компонент. Точки-генератори для діаграми Вороного взяти з точок, що утворюють контур промаркованих зв'язних компонент. Параметр sr визначає кількість точок-генераторів;
2. Здійснити усунення шумів на регіонах зв'язних компонент, керуючись такими параметрами: максимального розміру зони шуму nm; максимальної ширини C_w; максимальної висоти C_h; максимального відношення висоти до ширини C_r;
3. Згенерувати діаграму Вороного для визначених зв'язних компонент та точок-генераторів;
4. Видалити ребра побудованої діаграми Вороного, які перетинають області зв'язних компонент;

5. Видалити зашумлені текстові зони, застосувавши значення порогу мінімальної площі зони A_z для всіх зон, а також значення порогів мінімальної площі зони A_l та порогу максимального відношення сторін B_r для тих зон, де відношення висоти до ширини зони перевищує значення порогу B_r.

1.4.3 Алгоритм Recursive X-Y Cut

Алгоритм X-Y-Cut належить до групи низхідних методів (Top-down approaches) та є деревовидним алгоритмом. Корінним вузлом дерева є сторінка документа, а сукупність листкових вузлів становить сегментацію сторінки. Документ розділяється на менші прямокутні зони, кожна з яких є вузлом дерева. У процесі рекурсії, для кожної нової прямокутної зони обчислюються горизонтальні та вертикальні проекції профілів.

Сегментація сторінки алгоритмом Recursive X-Y-Cut показана на рис. 1.8.

Hierarchical Image Analysis
Using Irregular Tessellations

Рис. 1.8 Сегментація сторінки алгоритмом Recursive X-Y-Cut [20]

Кроки алгоритму Recursive X-Y-Cut наступні:

1. Побудувати таблиці префіксних сум горизонтальних та вертикальних пікселів, H_X та H_Y, що визначаються наступним чином:

   $H_X[i][j] = \#\{\, p \epsilon D(I) | X(p) = j, Y(p) \leq i, I(p) = 1\}$

$H_Y[i][j] = \#\{\, p \epsilon D(I) | X(p) \leq j, Y(p) = i, I(p) = 1\}$

Де $D(I) \subseteq Z^2$ – область визначення зображення I, $X(p)$, $Y(p)$ – X, Y-координати пікселя p, $I(p)$ – бінарне значення інтенсивності кольору пікселя p (0 – чорний піксель, 1 – білий піксель);

2. Кореневим вузлом дерева призначити зображення всього документа. Для кожного вузла виконати наступне:

a) Обчислити гістограми проекцій профілів для чорних пікселів поточного вузла:

$HIS_X[i] \leftarrow H_X[Y_2(Z)][i] - H_X[Y_1(Z)][i]$

$HIS_Y[j] \leftarrow H_Y[j][X_2(Z)] - H_Y[j][X_1(Z)]$

Де Z – зона поточного вузла, $((X_1(Z), Y_1(Z))$ – координати верхньої лівої точки обмежувальної рамки зони, $((X_2(Z), Y_2(Z))$ – координати нижньої правої точки обмежувальної рамки зони;

b) Виконати усунення шумів для пікселів заднього плану за допомогою параметрів T_X^n та T_Y^n. Оскільки, за припущенням, зашумлені пікселі розподілені рівномірно, для кожного окремого вузла дані параметри також додатково лінійно масштабуються відповідно до ширини та висоти зони;

c) Повторити крок 2.a);

d) Для гістограм проекцій профілів HIS_X та HIS_Y знайти найдовшу частину гістограми між двома піками: V_X та V_Y відповідно;

e) Поділ вузла поточної зони на два нові вузли здійснюється за таким критерієм: $V_X > T_X^C$ або $V_Y > T_Y^C$, де T_X^C та T_Y^C – покоординатні пороги ширини.

1.5 Алгоритми класифікації контенту сегментованих регіонів сторінки

Алгоритми класифікації вмісту документа призначені для встановлення відповідності між множиною сегментованих регіонів сторінки документа та

множиною переліку категорій структурної розмітки документа. Найпростішим варіантом класифікації регіонів сторінки є поділ на текстові та нетекстові регіони [28]. Автори вищезазначеної роботи класифікують регіони на основі геометричних та ентропійних ознак текстури. У роботі [29] розглядається класифікація трьох регіонів: регіони друкованого тексту, регіони рукописного тексту, зашумлені регіони. Дані регіони отримані за допомогою методу зв'язних компонент, враховуючи геометричну близькість та розміри окремих підрегіонів зв'язних компонент. До побудованої множини зв'язних компонент застосовується видобування ознак, серед яких: розмір регіону зв'язної області, довжини штрихів та ін. Всього таких ознак – 31, навчання яких здійснюється за допомогою класифікатора Фішера.

Перелічені методи прості у застосуванні та демонструють велику швидкість обробки, але можуть бути застосовані до зображень сторінок, на яких текст розміщується лише в одному напрямку.

У роботі [30] запропоновано видобування ознак на рівні пікселів та навчання на цих ознаках за допомогою класифікатора методу k-найближчих сусідів, а також апроксимації даного методу за допомогою k-d дерев.

Автори роботи [31] запропонували класифікатор дерев прийняття рішень на основі навчання з учителем. У якості ознак використано такі концепти, як: відсоткове співвідношення пікселів регіонів відповідних класів, відносний розмір кегля шрифту для точок певного класу, а також статистичні ознаки зв'язних компонент, що не потребують апріорної інформації про визначений клас.

Висновки до першого розділу

Задача розпізнавання зображень зі сканованими документами полягає у локалізації деяких наперед визначених елементів структури документа та подальшій сегментації цих зображень. Аналіз структури документів складається з таких етапів: попередня обробка, аналіз структури та аналіз розмітки документа, класифікація сегментованих регіонів та оцінка результату.

Попередня обробка складається з даних етапів: усунення шумів на зображенні, корекція кута нахилу, бінаризація зображення. Основними видами шумів, що містяться на зображеннях документів є маргінальний та клаттерний шуми. Методи бінаризації чорно-білих зображень поділяються на дві категорії: методи глобального порогування (Отцу) та методи локального порогування (Ніблек, Саувола).

Виявлення рамки сторінки також є важливим етапом при аналізі структури документа, одним із способів виявлення рамки – метод зв’язних компонент, однак також існують інші, більші функціонально навантажені алгоритми.

Сегментація сторінки документа являє собою один із основних етапів попередньої обробки зображення документа. У результаті зображення поділяється на зони, що містять тільки єдиний тип інформації.

Алгоритми сегментації сторінки належать до таких типів: висхідні методи та низхідні методи. Прикладами висхідних методів є алгоритми Docstrum, Kice на основі діаграми Вороного. Прикладом низхідного алгоритму є алгоритм Recursive X-Y-Cut.

Регіони сегментованої сторінки документа потрібно правильно класифікувати до одного з класів структурної розмітки документа. Зазвичай, з сегментованих регіонів видобувають певні незалежні від приналежності до класу ознаки, які подають класифікатору.

РОЗДІЛ 2 ДАТАСЕТИ ДЛЯ ЗАДАЧІ АНАЛІЗУ СТРУКТУРИ ДОКУМЕНТІВ НА ЗОБРАЖЕННЯХ

Залежно від типу письма, набори даних (датасети) для аналізу структури документів поділяються на три типи:

- Друковані документи;
- Рукописні документи;
- Змішані (поєднання друкованих та рукописних частин в одному документі).

2.1 Датасет TableBank

Велика різноманітність варіацій розміток табличних даних у документах зумовлює значний внесок у важливість виявлення та розпізнавання таблиць у процесі аналізу структури документів. Загальноприйняті методи розпізнавання таблиць спираються на певні ознаки, притаманні структурі документа, що робить ці ознаки залежними від розмітки сторінки. Розвиток методів глибинного навчання та комп'ютерного зору розширив можливості візуального розпізнавання таблиць на зображеннях. Візуальний аналіз таблиць полягає у його робастності до типів документів.

TableBank [32] являє собою колекцію зображень друкованих документів для візуального розпізнавання табличних даних на сторінках документів.

Такі формати документів як Microsoft Word (.docx) та Latex (.tex) містять в собі проанотовану структуру для табличних даних. Наприклад, XML-код docx-документу можна модифікувати таким чином, щоб виявляти межі таблиць. Подібним чином, Latex-код містить теги, всередині яких розміщують табличні дані.

Датасет TableBank складається з 417 234 проанотованих таблиць та їх документів. Даний датасет містить документи двох форматів: Word (.docx) та Latex (.tex). Документи формату Latex були отримані шляхом скачування Latex-коду та їх PDF-відповідників статей із каталогу наукових статей Arxiv.

Кожен файл формату .docx являє собою архів, що містить файл document.xml

У цьому xml-файлі міститься розмітка Word-документа, таблиця міститься в коді, обмеженим тегами <w:tbl> та </w:tbl>. Дану частину коду було модифіковано таким чином, щоб включити до розмітки візуальну обмежувальну рамку для таблиці. Після чого Word-документи, що містять таблицю хоча б на одній сторінці, були конвертовані до формату PDF.

Для аналогічної модифікації частини коду, що містить таблицю, у документах формату Latex застосовано спеціальну команду fcolorbox.

У таких документах таблиці зберігаються у вигляді [32]:

```
\begin{table}[]
\centering
\begin{tabular}{}
...
\end{tabular}
\end{table}
```

Здійснено наступну модифікацію [32]:

```
begin{table}[]
\centering
\setlength{\fboxsep}{1pt}
\fcolorbox{bordercolor}{white}{
\begin{tabular}{}
...
\end{tabular}}
\end{table}
```

Розпізнавання табличних структур здійснюється за допомогою побудови датасету структур, який отримано з XML-коду документа Word та Latex-коду.

Таблична структура записується у вигляді HTML-коду з наступними тегами: <tabular>, </tabular>, <tr>, </tr>, <cell_y>, <cell_n>, де теги <tr>, </tr> позначають межі рядка таблиці, <cell_y> – непорожню комірку таблиці, <cell_n> – порожню комірку таблиці. Для розпізнавання табличних структур далі використано image-to-text модель нейронної мережі на основі encoder-decoder

(тобто на вхід подається зображення документа, на виході отримуємо список розпізнаних табличних структур у вигляді HTML-коду).

Автори роботи [32] для навчання датасету TableBank використали image-to-markup модель, описану у роботі [33].

Для виявлення таблиць на зображенні використано архітектуру моделі глибокої нейронної мережі Faster R-CNN. Дана архітектура є третьою у сімействі двофазних детекторів об’єктів. Спочатку, у 2013 році було запропоновано модель R-CNN [34]. Потім, у 2015 році з’явилась модель Fast R-CNN [35]. Дані дві моделі використовують техніку вибіркового пошуку для генерації множини регіонів-кандидатів для виявлення. У Моделі Faster R-CNN для генерації регіонів-кандидатів застосовується мережа пропонування регіонів-кандидатів – Region Proposal Network (RPN), що поєднується з мережею Fast R-CNN.

Приклад виявлення таблиці на зображенні із датасету TableBank показано на рис 2.1.

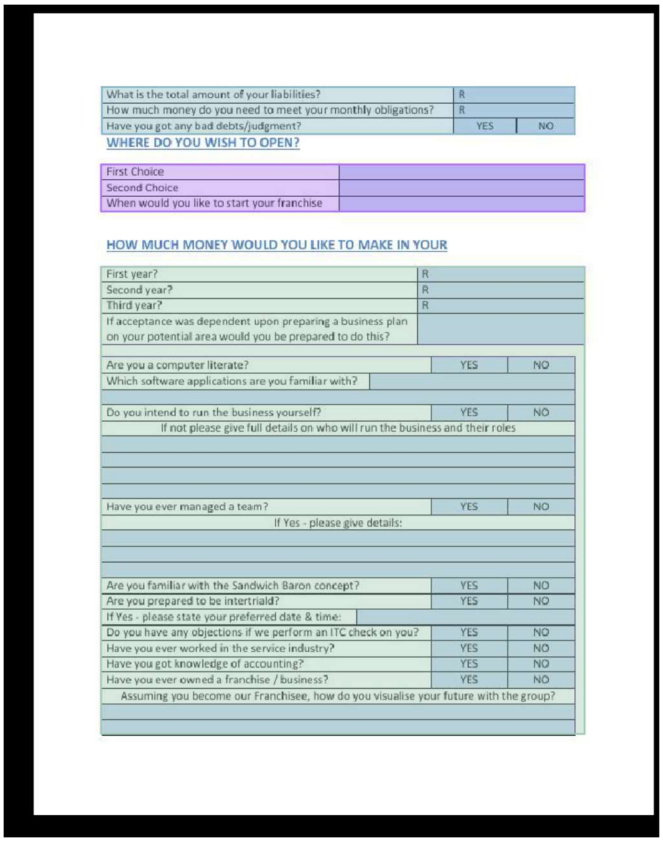


Рис. 2.1 Виявлення таблиць на зображенні із датасету TableBank [32]

Статистика датасету TableBank [32] подана в таблиці 2.1.

Таблиця 2.1.

Задача	Word	Latex	Всього
Виявлення таблиці	163 417	253 817	417 234
Розпізнавання табличної структури	56 866	88 597	145 463

2.2 Датасет Marmot

Датасет Marmot, що описаний у роботі [36] містить близько 2000 PDF сторінок та складається з двох частин: сторінки англійською та китайською мовами у пропорції близько 1:1. Також, сторінки поділено за принципом наявності на ній таблиці – близько 1000 сторінок містять хоча б одну таблицю, інші сторінки – не містять таблиць. Кожна одиниця датасету (тобто сторінка) характеризується такими об’єктами: зображення сторінки з роздільною здатністю 600 dpi, XML файл з істинною (ground truth) розміткою структурних елементів сторінки, де містяться координати їх обмежувальних рамок. XML-файл також містить деревовидну структуру логічних зв’язків елементів на сторінці. До них належать дві групи – листкові елементи та композиційні елементи. До листкових елементів належить найменші структурні одиниці сторінки – текстовий символ, ілюстрація. До композиційних типів належать внутрішні вузли дерева розмітки сторінки – матриця, формула, фігура, текстова лінія, список, заголовок таблиці, підпис таблиці, таблиця, параграф тексту, колонтитули сторінки.

2.3 Датасет PubLayNet

Геометричні підходи аналізу розмітки документів [37] у поєднанні з оптичним розпізнаванням тексту були домінантними методами аналізу вмісту документів до появи аналітичних методів візуального розпізнавання зображень на основі машинного навчання [38].

Ручне анотування структурних елементів на зображенні документа займає немало часу та потребує значних людських ресурсів, особливо при великих

наборах даних. У роботі [39] запропоновано метод автоматичного розмічання структурних елементів на зображеннях сторінок документа датасету PubLayNet, що містить основні категорії об’єктів для аналізу розмітки документів, зокрема: блоки тексту, текстові заголовки сторінки, списки, таблиці, ілюстрації. На початковому етапі створення датасету використано 1 162 856 PDF-документів, кожен з яких додатково представлений XML-файлом, у якому наявна структура даного документа; у процесі створення датасету, описаного у секції C “Data partition” розділу III “AUTOMATIC ANNOTATION OF DOCUMENT LAYOUT” роботи [39], для остаточної версії датасету було відібрано близько 364 тисяч PDF-документів-сторінок, про що свідчить таблиця 2.2. Процес генерації анотації розмітки для PDF-документу складається з таких етапів:

1. Попередня обробка XML-файлів. Поділ XML-вузлів на п’ять груп: 1) структуровані дані, 2) неструктуровані дані, 3) ілюстрації включно з підписами до них; 4) таблиці включно з підписами до них, 5) списки. До структурованих даних віднесено заголовок документу, заголовки абзаців, реферат документу, список ключових слів та текст в межах абзаців. До неструктурованих даних віднесено перелік авторів, список скорочень.
2. Обробка PDF-сторінки за допомогою програми PDFMiner – отримано розмітку сторінки з трьома категоріями 1) текстовий блок, 2) зображення 3) геометрична фігура (ламані, криві, многокутники).
3. Нормалізація Unicode-рядків до KD-форми.
4. Зв’язування частин PDF-документа з XML-документом за допомогою нечіткого пошуку. У ситуації, коли один текстовий блок PDF-документа покриває декілька XML-вузлів, здійснюється послідовний пошук текстових ліній усередині текстового блоку. Якщо досягнуто кінця XML-вузла, але дана текстова лінія не є кінцем текстового блоку, даний текстовий блок розділяється на два текстові блоки.

Статистика датасету PubLayNet [39] подана у таблицях 2.2–2.3.

Таблиця 2.2.

	Навчальна вибірка	Валідаційна вибірка	Тестова вибірка
Сторінки лише з текстовими блоками	87608	1138	1121
Сторінки з заголовками	46480	2059	2021
Сторінки зі списками	53793	2984	3207
Сторінки з таблицями	86950	3772	3950
Сторінки з фігурами	96656	3734	3807
Всього	340391	11858	11983

Таблиця 2.3.

	Навчальна вибірка	Валідаційна вибірка	Тестова вибірка
Тестові блоки	2376702	93528	95780
Заголовки	633359	19908	20340
Списки	81850	4561	5156
Таблиці	103057	4905	5166
Фігури	116692	4919	5333
Всього	3311660	127815	131775

Навчання моделі глибокої нейронної мережі на датасеті PubLayNet здійснено за допомогою моделей Faster R-CNN та Mask R-CNN, імплементації яких наявні в бібліотеці виявлення об’єктів Detectron [40].

Приклади виявлення структурних елементів сторінок документів із датасету PubLayNet показано на рис. 2.2.

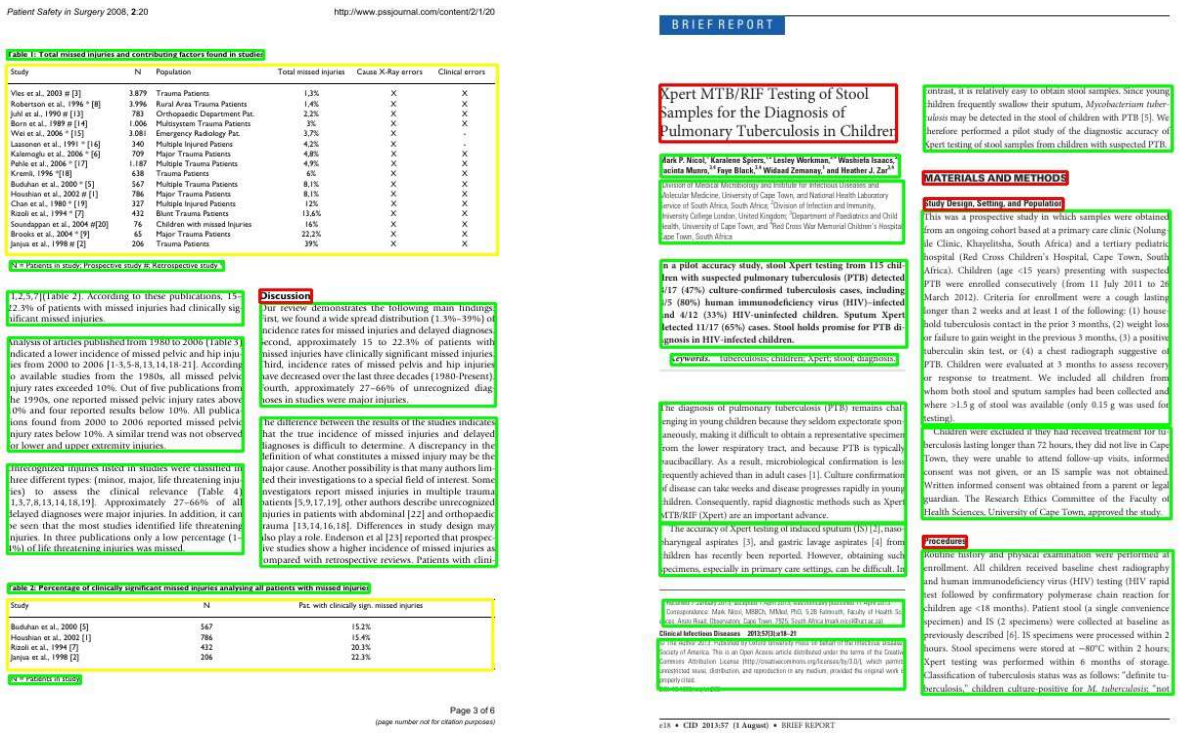


Рис. 2.2 виявлення структурних елементів

сторінок документів із датасету PubLayNet [39]

Результати навчання авторами роботи подані у таблиці 2.4. [39]

Таблиця 2.4.

Клас	Валідаційна вибірка		Тестова вибірка	
	Faster R-CNN	Mask R-CNN	Faster R-CNN	Mask R-CNN
Текстові блоки	0.910	0.916	0.913	0.917
Заголовки	0.826	0.840	0.812	0.828
Списки	0.883	0.886	0.885	0.887
Таблиці	0.954	0.960	0.943	0.947
Фігури	0.937	0.949	0.945	0.955
Усереднене значення	0.902	0.910	0.900	0.907

Висновки до другого розділу

У даному розділі проаналізовано три датасети (набори даних) зображень для аналізу структури документів – TableBank, Marmot та PubLayNet. Описано способи побудови даних наборів даних для кожного датасету.

Датасет TableBank призначений для розпізнавання табличних даних на зображеннях сторінок документів. Всього у датасеті – понад 417 тисяч таблиць на зображеннях та анотації таблиць. Для виявлення таблиць автори застосували архітектуру моделі глибокої нейронної мережі Faster R-CNN.

Датасет PubLayNet містить близько 364 тисяч сторінок-документів. Автори даного датасету використали дві архітектури моделей глибоких нейронних мереж – Faster R-CNN та Mask R-CNN.

РОЗДІЛ 3 ГЛИБОКІ НЕЙРОННІ МЕРЕЖІ ЯК ЗАСІБ АНАЛІЗУ СТРУКТУРИ ЗОБРАЖЕНЬ ДОКУМЕНТІВ

Детектори об’єктів (object detectors) поділяються на два типи [41]: однофазні детектори (one-stage detectors) та двофазні детектори (two-stage detectors). У двофазних детекторах генерація регіонів-кандидатів відбувається за допомогою Region Proposal Network (RPN). Згенеровані регіони-кандидати передаються до Region Convolutional Network (R-CNN) і на виході отримується розподіл ймовірності приналежності мітки об’єкта для кожного з класів та просторові зміщення об’єктів (тобто координати їх обмежувальних рамок на зображенні).

До однофазних детекторів належать:

- YOLO [42];
- RetinaNet [43];
- EfficientDet [44];

До двофазних детекторів належать:

- Fast R-CNN [35];
- Faster R-CNN [45];
- Mask R-CNN [46];

3.1 Метрики оцінювання візуального розпізнавання об'єктів

Виведення інформації на зображення про виявлені на ньому об’єкти зазвичай складається з таких компонентів: координати обмежувальної рамки об’єкта, мітка класу об’єкта, коефіцієнт достовірності детектора щодо приналежності об’єкта до визначеного класу. У більшості випадків обмежувальна рамка представлена чотирма координатами верхньої лівої та правої нижньої точок обмежувального прямокутника: $(x_{left}, y_{left}, x_{rig}\ \ , y_{right})$. Виняток становлять моделі YOLO, у яких координати обмежувальної рамки задаються відносно центру обмежувальної рамки, нормованої за шириною та висотою зображення: $(\frac{x_c}{w_i}, \frac{y_c}{h_i}, \frac{w_b}{w_i}, \frac{h_b}{h_i})$, де (x_c, y_c) – координати центру

обмежувальної рамки, w_b, h_b – ширина та висота обмежувальної рамки, w_i, h_i – ширина та висота оригінального зображення.

Найпоширенішою метрикою для візуального розпізнавання об'єктів є метрика середньої точності (average precision) та її варіації, зокрема усереднена середня точність (mean average precision) [47].

Для пояснення метрики average precision потрібно пригадати базові концепти визначення точності у машинному навчанні, зокрема: істинно-позитивний результат (True Positive), хибно-позитивний результат (False Positive), хибно-негативний результат (False Negative). Далі ці терміни розглядаються з точки зору розпізнавання об'єктів на зображенні:

- True Positive – клас та місцезнаходження визначені правильно відповідно до істинного результату.
- False Positive – місцезнаходження об'єкта визначено неправильно, або правильно визначено місцезнаходження об'єкта, віднесеного до іншого класу, ніж істинний результат, або виявлено об'єкт, що не збігається із жодним із істинних результатів.
- False Negative – об'єкт не виявлено відповідно до істинного результату.

Правильність виявлення об'єктів визначається за коефіцієнтом Intersection over union (IOU). Даний коефіцієнт визначається як площа перетину обмежувальної рамки, виявленої детектором, ($B_{predicted}$) та істинної обмежувальної рамки (B_{ground_truth}), поділеної на площу об'єднання даних двох обмежувальних рамок. Коефіцієнт IOU визначає за формулою (3.1) [48]

$$IOU(B_{predicted}, B_{ground\_truth}) = \frac{area(B_{predicted} \cap B_{ground\_truth})}{area(B_{predicted} \cup B_{ground\_truth})} \qquad (3.1)$$

Оскільки в метриках для візуального виявлення об'єктів до класу True Negative належить нескінченна кількість об'єктів, застосування метрик, у яких фігурує TN, не є можливим. Метрики для візуального розпізнавання об'єктів базуються на таких поняттях, як точність та повнота (Precision, Recall).

Precision та Recall визначаються за формулами (3.2-3.3):

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Для забезпечення великої точності потрібно, щоб якомога менше об’єктів було віднесено до класу FP. Для забезпечення великої повноти потрібно, щоб якомога менше об’єктів було віднесено до класу FN.

Графік кривої точності-повноти зазвичай являє собою зигзагоподібну криву. Згладжування кривої точності-повноти відбувається за допомогою деякої апроксимації, а саме двох видів інтерполяції.

Перший спосіб апроксимації – інтерполяція одинадцятьма точками. Береться до уваги максимальне значення повноти на одинадцятьох рівномірних відрізках при проекції на вісь, що позначає повноту. Усереднення полягає в тому, що береться не значення точності в кожній фіксованій точці осі повноти, а береться максимальне значення точності, де значення повноти більше за відповідне значення повноти у попередній фіксованій точці на осі повноти.

Даний процес описується наступними формулами:

$$AP_{11} = \frac{1}{11} \sum_{R \in \{0, 0.1, 0.2 \ldots 0.9, 1\}} P_{interp}(R) \tag{3.4}$$

Де

$$P_{interp}(R) = \max_{\tilde{R}:\tilde{R} \geq R} P(\tilde{R}) \tag{3.5}$$

Другий спосіб апроксимації – інтерполяція n точок. У такому випадку формули (3.4-3.5) набувають вигляду:

$$AP_{all} = \sum_{n} (R_{n+1} - R_n) P_{interp}(R_{n+1}) \tag{3.6}$$

Де

$$P_{interp}(R_{n+1}) = \max_{\tilde{R}:\tilde{R} \geq R_{n+1}} P(\tilde{R}) \tag{3.7}$$

Метрика усереднена середня точність (mean average precision) – це середнє значення average precision за класами, що фігурують, та описується формулою:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{3.8}$$

де AP_i – значення за метрикою average precision для $i-$того класу, де N – загальна кількість класів, що розглядаються.

3.2 Архітектури глибоких нейронних мереж для візуального розпізнавання

Сучасні детектори об’єктів на основі глибинного навчання поділяються на дві категорії: двофазні детектори об’єктів та однофазні детектори. Виявлення об’єктів двофазним детектором складається з двох складових: спершу генеруються певні регіони-кандидати. На другому кроці, вектори ознак, отримані з даних регіонів, подаються на вхід нейронної мережі, де відбувається класифікація згенерованих регіонів до визначених класів.

Однофазні детектори не мають окремої складової для генерації регіонів-кандидатів та розглядають кожен регіон інтересу як потенційний об’єкт.

3.2.1 Двофазні детектори об’єктів

У даному розділі розглядаються двофазні детектори об’єктів R-CNN: Fast R-CNN, Faster R-CNN та Mask R-CNN.

3.2.1.1 Fast R-CNN

Модель Fast R-CNN [35] є наступником моделі R-CNN (Region-based Convolutional Network) [34], що є різновидом згорткової глибокої нейронної мережі.

R-CNN має певні недоліки, зокрема: навчання мережі є повільним з точки зору часу та ресурсів. Модель R-CNN складається з трьох окремих модулів: генерація регіонів-кандидатів методом вибіркового пошуку, другий модуль являє собою згорткову нейронну мережу, яка з кожного регіону-кандидата видобуває вектор ознак фіксованої довжини. Третій модуль являє собою

навчання множини лінійних класифікаторів на основі методу опорних векторів (далі – SVM-класифікатори). З метою покращення результатів локалізації об’єктів застосовано етап постобробки – навчання регрессорів обмежувальних рамок (bounding box regressors).

Видобування ознак для SVM-класифікаторів та регрессорів відбувається для кожного регіону-кандидата на кожному зображенні. Наприклад, для VGG16 для 5 тисяч зображень датасету VOC07 даний процес займає 2.5 GPU-днів [35], а виявлення об’єктів на одному зображенні в середньому займає 47 секунд (на GPU).

Модель Fast R-CNN запропонована у 2015 році та покликана виправити недоліки моделі R-CNN. На вхід подається ціле зображення, що пропускається через декілька згорткових шарів та шарів субдисретизації (pooling layers), у даному випадку – max pooling. Таким чином, отримано згорткову карту ознак та множину регіонів-кандидатів. Кожен регіон-кандидат передається шару, що називається регіон інтересу (Region of Interest pooling layer, RoI), суть якого полягає у видобуванні вектора ознак фіксованої довжини зі згорткової карти ознак. Отримані вектори ознак передаються сукупності повнозв’язних шарів (fully-connected layers). Дана модель нейронної мережі містить два вихідні шари (output layers). Перший з них містить softmax-ймовірності приналежності об’єктів до кожного із K класів, інший – чотири дійсні числа, що позначають координати обмежувальної рамки об’єкта на зображенні.

Субдискретизований шар регіону-інтересу ділить кожен регіон, що визначається четвіркою (r, c, h, w), де r, c – координати вершини лівого верхнього кута прямокутної області регіону, h, w – висота та ширина регіону відповідно; на менші підрегіони, розміром $H \times W$, які, у свою чергу, є гіперпараметрами ROI-шару. До отриманих підрегіонів застосовується операція max pooling.

3.2.1.2 Faster R-CNN

На відміну від R-CNN та Fast R-CNN, де для генерації множини регіонів-кандидатів використовувався метод вибіркового пошуку, у моделі Faster R-CNN дану функцію покладено на окремий модель – мережу пропонування регіонів-кандидатів – Region Proposal Network (RPN).

У мережі пропонування регіонів-кандидатів (RPN) на вхід подається зображення будь-якого розміру, на виході отримується вищезгадані регіони, кожен з яких має коефіцієнт достовірності – ймовірність, що позначає неналежність об'єкта до класу заднього плану, відповідно, це відповідає ймовірності належати до одного із попередньо визначених класів для здійснення класифікації об'єкта.

Отже, мережа пропонування регіонів-кандидатів приймає на вхід зображення, що обробляється повністю згортковою мережею (fully convolutional network) [49] – проходить через послідовність згорткових шарів, отримуючи на виході деяку карту ознак зображення розміром $c \times h \times w$. У якості FCN використовується згорткова мережа VGG16. Далі, для карти ознак за допомогою методу сковзаючого вікна формуються області розміром $c \times n \times n$

Відповідно, кожна із n^2 ознак даної області кодується вектором $1 \times c$. Всього для карти ознак формується n^2 даних областей, так, щоб кожна ознака була центром області. Навколо центру формується k фіксованих обмежувальних рамок (anchors) – прямокутники з різними відношеннями сторін (1:1, 1:2, 2:1) та розмірами сторони (128, 256, 512). Далі вектор ознаки $1 \times c$ буде паралельно передано двом шарам – класифікаційному (для здійснення класифікації належності до одного з класів або класу заднього фону) та регресійному (для передбачення координат обмежувальної рамки – 4 дійсних числа). Відповідно, до вектора ознак застосовується згортка 1×1 з кількістю вихідних каналів згортки $c_1 = 2$ для класифікаційного шару та $c_2 = 4$ для регресійного.

Таким чином, за допомогою RPN отримуємо координати регіонів-кандидатів на зображенні та їх класифікацію приналежності до класу заднього

фону (регіонами-кандидатами вважаються ті об’єкти, які були класифіковані як такі, що не належать класу заднього фону).

3.2.1.3 Mask R-CNN

Модель Mask-RCNN [46] є доповненням моделі Faster R-CNN. У моделі Faster R-CNN після проходження вектору ознак, отриманого на виході ROI-шару, через сукупність повнозв’язних шарів на виході міститься два вихідні шари. Для моделі Mask R-CNN додано третій вихідний шар для передбачення сегментаційної маски об’єкта на піксельному рівні, що представлений прямокутною бінарною матрицею розмірності, що відповідає вхідному зображенню. Відповідно, елементами матриці є 1 на тих місцях, що відповідають координатам пікселів зображення, де модель передбачила наявність об’єкта, усі інші елементи матриці – нулі.

3.2.2 Однофазні детектори об’єктів

У даному розділі розглядаються однофазні детектори об’єктів, як RetinaNet, YOLO, EfficientDet.

3.2.2.1 RetinaNet

Модель RetinaNet [43] є одноетапним детектором та складається з трьох частин: основної мережі (backbone network) та двох допоміжних частин – класифікаційної та регресійної мереж.

На вхід backbone мережі подається ціле зображення, на виході отримуємо конволюційну карту ознак, на основі якої класифікаційна згорткова мережа визначає приналежність виявленого об’єкта до певного класу, а за допомогою регресійної мережі визначаються координати обмежувальної рамки виявленого об’єкта.

У якості backbone мережі автори пропонують Пірамідальну Мережу Ознак (Feature Pyramid Network – FPN). [50]. Особливістю Пірамідальної мережі ознак є те, що після проходження через згорткові шари методом прямого поширення

(feedforward) (формування висхідного шляху) для кожного згорткового шару формується окрема карта ознак (формування низхідного шляху). Відповідно, вихід кожного згорткового шару у висхідному шляху формує окремий рівень піраміди. Оскільки stride кожного наступного згорткового шару вдвічі більший за stride попереднього шару, просторова розмірність ознак зменшується, водночас збільшується їх семантична наповнюваність (значимість).

Висхідний шар представлений архітектурою Resnet та містить п'ять згорткових шарів: $Conv_1 \dots Conv_5$. До виходу шару $Conv_5$ застосовується згортка 1×1 і таким чином отримується карта ознак M_5. Для формування карт ознак подальших рівнів піраміди застосовується процес збільшення розмірності (upsampling) методом найближчого сусіда. Розмірність карти M_i після процесу збільшення розмірності збігається з розмірністю виходу згорткового шару $Conv_{i-1}$, тож до виходу цього шару також застосовується згортка 1×1. Однакова розмірність дозволяє здійснити поелементне додавання двох частин для формування карти ознак M_{i-1}. Таким чином отримуємо карти ознак M_4, M_3, M_2.

До кожної із карт ознак M_2, M_3, M_4, M_5 застосовується згортка 3×3 з метою отримання остаточних карт ознак P_2, P_3, P_4, P_5.

За таким алгоритмом працює класична FPN. Для детектора RetinaNet були внесені деякі модифікації до Feature Pyramid Network. Карта ознак P_2 не використовується з міркувань великої кількості обчислень. Карта ознак P_6 отримується за допомогою застосування згортки 3×3 з кроком stride=2 до згорткового шару $Conv_5$. Карта ознак P_7 отримується за допомогою застосування до P_6 функції активації ReLu та згортки 3×3 з кроком stride=2.

Подібно до Faster R-CNN діє система фіксованих обмежувальних рамок. У кожного рівня піраміди свій розмір фіксованої обмежувальної рамки. Для P_3 розмір становить 32, для $P_4 - 64$, і так далі з поступовим збільшенням розміру вдвічі на кожному рівні піраміди; розмір для $P_7 - 512$.

Архітектура мережі RetinaNet показана на рис 3.1.

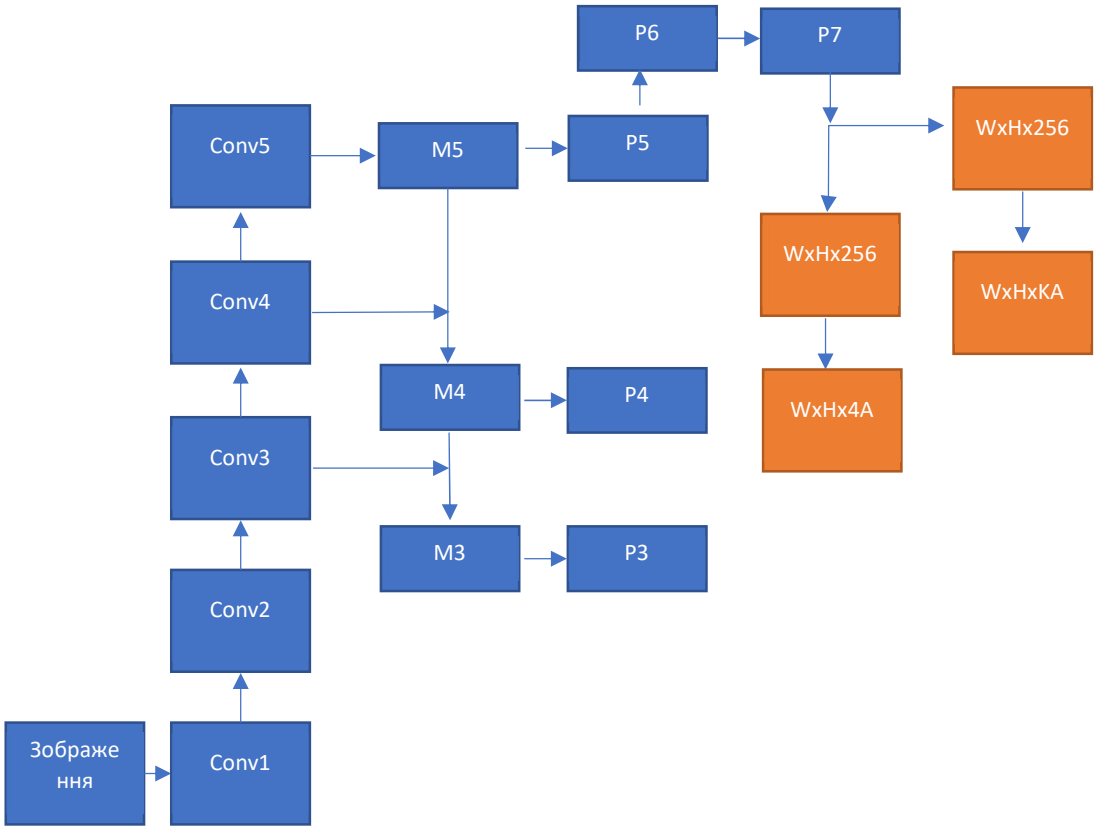


Рис 3.1 Архітектура мережі RetinaNet

Зміни також стосуються класифікаційної частини фіксованих обмежувальних рамок. Якщо в Faster R-CNN до карти ознак, представленої у вигляді вектора розмірності $1 \times c$ застосовувалась згортка з кількістю вихідних каналів $c_1 = 2$, то в RetinaNet $c_1 = k$, де k – кількість класів для виявлення об'єктів без врахування класу заднього фону. Визначення приналежності до заднього фону тепер визначається за допомогою призначення кожній фіксованій обмежувальній рамці деякої істинної (ground-truth) обмежувальної рамки із застосуванням порогу Intersection-over-Union (IoU) між істинною та фіксованою обмежувальними рамки. Якщо IoU перебуває в межах [0; 0.4) – об'єкт належить до класу заднього фону, якщо ж IoU в межах [0,5;1] – об'єкт належить до одного із k класів для виявлення. У випадку, коли IoU перебуває в межах [0.4;0.5) – об'єкт відкидається та не враховується під час навчання мережі.

3.2.2.2 YOLO

У даному підрозділі розглянуто порівняння двох версій одноетапного детектора YOLO (You Only Look Once), а саме YOLOv3 [42] та YOLOv5 [51].

У якості backbone мережі у YOLOv3 використовується повністю згорткова мережа Darknet-53. Із детальним описом шарів backbone мережі можна ознайомитись у роботі [42]. Дана мережа містить 106 згорткових шарів. Із тих 53 шарів, що призначені для задачі виявлення об'єктів, п'ять шарів застосовують крок stride рівним двом. Тож, розмірність карт ознак в результаті проходження через всі згорткові шари backbone мережі будуть рівними 1/32 від роздільної здатності зображення, що подається на вхід цій мережі. Саме тому конфігурація архітектури YOLO-мереж передбачає використання зображень з роздільною здатністю кратною 32 пікселям. Найпоширеніші варіанти - 416×416 та 608×608, менш поширені - 320×320 та 640×640.

Особливістю даної архітектури є побудова передбачень обмежувальних рамок з трьох різних карт ознак у процесі проходження через шари мережі. Отже, передбачення обмежувальних рамок здійснюється у трьох місцях, а саме після третього, четвертого та п'ятого згорткового шару з кроком stride рівним двом. Позначимо дані шари як $Conv_{s1}$ ... $Conv_{s5}$ Відповідно, застосування шару $Conv_{s3}$ зменшує розмірність вхідного зображення у 8 разів, застосування $Conv_{s4}$ – у 16 разів, $Conv_{s5}$ – у 32 рази. Чим більше зменшення кожної з карт ознак, тим більші об'єкти доступні для виявлення. Отже, $Conv_{s3}$ виявляє найменші об'єкти, $Conv_{s4}$ – об'єкти середніх розмірів, $Conv_{s5}$ – найбільші об'єкти.

Ще однією відмінністю YOLOv3 з попередніми версіями є відмова від softmax-апроксимації при визначенні ймовірностей класової приналежності. Softmax-ймовірності передбачають, що об'єкт може належати тільки одному класу, тобто не передбачається, що певний клас може бути підкласом деякого більшого класу. У YOLOv3 ймовірність класів визначається за допомогою логістичної регресії та деякого порогу, що регулює приналежність об'єкта одразу до декількох класів.

Основною відмінністю YOLOv4 від YOLOv3 є використання іншої backbone-мережі – YOLOv3 використовує Darknet-53, тоді як у YOLOv4 та YOLOv5 використовується CSP-Darknet-53 [52].

Для побудови карти ознак застосовується Spatial Pyramid Pooling (SPP) [53].

Основні покращення версії YOLOv5 [54] у порівнянні з попередніми версіями YOLOv4 та YOLOv3 спрямовані на дослідження використання різних функцій активацій, експерименти з аугментацією даних та постобробкою отриманих результатів для удосконалення результатів виявлення об'єктів. Також моделі версії YOLOv5 займають набагато менше місця на диску, тому з'являється більша ймовірність використання YOLOv5 на мобільних пристроях у подальшому.

3.2.2.3 EfficientDet

У роботі [44] запропоновано нову архітектуру згорткових нейронних мереж під назвою EfficientDet (скорочено від Efficient Detection).

У якості backbone мережі використовується модель нейронної мережі EfficientNet [55].

Автори роботи запропонували внести зміни до Пірамідальної Мережі Ознак (FPN). Якщо в класичній FPN [50] поєднання ознак відбувається лише під час формування низхідного шляху, то особливістю реалізації Пірамідальної Мережі Ознак в EfficientDet є двонапрямленість у способі поєднання ознак – після проходження низхідним шляхом з'являється додатковий висхідний шлях подібно до мережі PaNet [56]. Відмінністю від PaNet є те, що низхідний та висхідний шляхи поєднання ознак розглядаються як один шар мережі EfficientDet з можливістю долучення декількох таких шарів з метою більш ефективного видобування ознак.

Висновки до третього розділу

У даному розділі проаналізовано основні архітектури глибоких нейронних мереж для візуального розпізнавання об’єктів, що також називають детекторами об’єктів. Існує дві основні категорії детекторів об’єктів: однофазні детектори та двофазні детектори об’єктів. До двофазних детекторів належать Fast R-CNN, Faster R-CNN, Mask R-CNN та ін., до однофазних детекторів належать детектори RetinaNet, YOLO, EfficientDet та ін.

Перевагою двофазних детекторів є краща точність виявлення за рахунок того, що більше часу відводяться на навчання даної моделі глибокої нейронної мережі, тоді як однофазні детектори менш точні у виявленні, але потребують менше часу для навчання моделі, у порівнянні з двофазними детекторами.

РОЗДІЛ 4 ЗАСТОСУВАННЯ МОДЕЛЕЙ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ АНАЛІЗУ СТРУКТУРИ ЗОБРАЖЕНЬ ДОКУМЕНТА

У роботі здійснено експерименти навчання різних моделей глибоких нейронних мереж на частині набору даних (датасету) PubLayNet [39], що містить основні категорії об'єктів для аналізу розмітки документів, зокрема: блоки тексту, текстові заголовки сторінки, списки, таблиці, ілюстрації. Оригінальний датасет містить 340391 та 11858 зображень сторінок документів у навчальній та валідаційній вибірках. Для навчання було обрано 1200 зображень з навчальної вибірки та 300 з валідаційної.

Проведено експерименти наступних згорткових нейронних мереж (див. розділ 4.2):

- RetinaNet [57];
- EfficientDet [58];
- Mask R-CNN (на основі фреймворку MMDetection [59]);
- YOLOv5 [51].

Спочатку експерименти з навчанням нейронної мережі на датасеті PubLayNet були проведені з використанням архітектури мережі RetinaNet та EfficientNet. Результати навчання даних моделей виявилися незадовільними (0.534 та 0.563 відповідно – див. таблицю 4.2). У зв'язку з цим, мною було прийнято рішення про перехід на використання двофазних детекторів об'єктів – здійснено навчання архітектури Mask R-CNN. Вдалося досягти кращих результатів розпізнавання (див. таблицю 4.5), однак результати виявлення об'єктів класу «Текст» залишали бажати кращого.

Було прийнято рішення про навчання ще однієї архітектури однофазного детектора об'єктів – YOLOv5 від Ultralytics [51]. У результаті проведення експериментів з навчання даної моделі було встановлено, що якість виявлення об'єктів класу «Текст» покращилась. Для досягнення ще більшої якості розпізнавання об'єктів класу «Текст» експерименти з YOLOv5 продовжились. У підрозділі 4.1 запропоновано оптимізацію YOLOv5, що дозволила досягти кращого результату у виявленні текстових блоків на зображеннях документів.

4.1 Оптимізація нейронних мереж

Після проведення навчання моделі YOLOv5, у конволюційних шарах нейронної мережі було замінено функцію активації – функцію SiLU (Sigmoid Linear Unit) на функцію GELU (Gaussian Linear Error Unit).

Функція активації GELU має такий вигляд (формула (4.1)) [60]:

$$GELU(x) = x\Phi(x) = 0.5x(1 + tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3))) \quad (4.1)$$

Функція активації SiLU має такий вигляд (формула (4.2)) [60]:

$$SILU(x) = x \cdot sigmoid(x) = \frac{x}{1 + e^{-x}} \quad (4.2)$$

Графіки функцій активації ReLU, SiLU та GELU показано на рис.4.1.

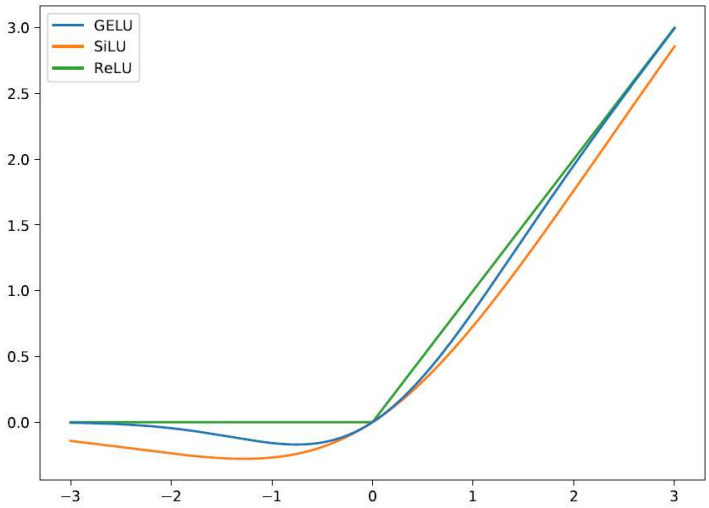


Рис. 4.1. Графіки функцій активації GELU, SiLU, ReLU [60]

Автори роботи [60] показали, що запропонована ними функція активації GELU є ефективнішою за SiLU для задач візуального розпізнавання об'єктів – навчання на датасетах MNIST, CIFAR10, CIFAR100.

Також, внесено зміни до CSP-групи шарів мережі: замість суперпозиції трьох згорток $conv_3(conv_2(conv_1(x)))$ застосовується така суперпозиція: $conv_3(conv_1(conv_1(x)))$. Така оптимізація дозволила скоротити кількість параметрів на 3% з 87,2 млн параметрів до 84,6 млн параметрів.

Порівняльна характеристика кількості параметрів для кожної з груп шарів показана в таблиці 4.1.

Таблиця 4.1.

№	Назва шару	Кількість фільтрів	Розмір згортки	Кількість параметрів (оригін.)	Кількість параметрів (оптиміз.)
1	Focus	12	3x3	8800	8800
2	Conv	160	3x3	115520	115520
3	CSP	160	1x1, 1x1, 3x3	309120	296160
4	Conv	320	3x3	461440	461440
5	CSP	320	1x1, 1x1, 3x3	3285760	3234240
6	Conv	640	3x3	1844480	1844480
7	CSP	640	1x1, 1x1, 3x3	13125120	12919680
8	Conv	1280	3x3	7375360	7375360
9	SPP	–	–	4099840	4099840
10	CSP	1280	1x1, 1x1, 3x3	19676160	18855680
11	Conv	640	1x1	820480	820480
12	CSP	640	1x1, 1x1, 3x3	5332480	4922240
13	Conv	320	1x1	205440	205440
14	CSP	320	1x1, 1x1, 3x3	1335040	1232320
15	Conv	320	3x3	922240	922240
16	CSP	640	1x1, 1x1, 3x3	4922880	4717440
17	Conv	640	3x3	3687680	3687680
18	CSP	1280	1x1, 1x1, 3x3	19676160	18855680
19	Detection	–	–	67290	67290

4.2 Проведення експериментів

Результати всіх вищезгаданих методів показано у таблиці 4.2 (показано результат з найкращою точністю для кожного методу).

Таблиця 4.2.

#	Метод	@Map_IOU
1	RetinaNet [57]	0.534
2	EfficientDet [58]	0.563
3	Mask R-CNN [59]	0.886
4	YOLOv5 [51]	0.914

Варто зазначити, що дані результати отримано внаслідок усереднення значення декількох запусків кожного із зазначених методів.

Зведені результати наведено у **додатку Є**.

Вкрай низькі результати перших двох моделей обумовлені двома основними причинами: обмеження кількості даних для навчання та недостатня кількість епох в процесі навчання. Рекомендується збільшити вибірку даних та провести навчання моделей з більшою кількістю епох. Результати останніх двох моделей можна пояснити тим, то ці моделі менше адаптовані до дрібних об'єктів на зображенні, таких як заголовки абзаців. Наприклад, об'єкти площею менше ніж 23x23 пікселі автоматично потраплять до класу False Negative, що також негативно впливає на результат.

Результати роботи EfficientDet подано у **додатку Д**.

Результати роботи RetinaNet подано у **додатку Е**.

4.2.1 YOLOv5

На час початку проведення експериментів з YOLOv5 (січень 2021 року) актуальною версією даної архітектури була четверта релізна версія [51]. Дана версія пропонує чотири моделі з різними конфігураціями та кількістю параметрів: yolov5s, yolov5m, yolov5l, yolov5x. Першими експериментами були експерименти з моделями yolov5s та yolov5x оригінального підходу. У зв’язку з

тим, що результати yolov5x виявились кращими за yolov5s, було прийнято рішення відкинути yolov5s та зосередити свою увагу на моделі yolov5x.

Усі подальші описані результати YOLOv5 – це результати навчання yolov5x.

Порівняльні результати метрики mAP за кожним класом для оригінального та оптимізованого підходів нейронної мережі YOLOv5 подано в таблиці 4.3.

Під час навчання мережі поданих у таблиці 4.3 результатів було використано оптимізатор SGD з початковим коефіцієнтом швидкості навчання (learning rate), що дорівнює 0.01.

Таблиця 4.3.

Клас	Оригінальний підхід	Оптимізований підхід
(Усі класи)	**0.914**	0.911
Текст	0.840	**0.853**
Заголовок	0.865	**0.870**
Список	**0.927**	0.920
Таблиця	**0.974**	0.969
Фігура	**0.966**	0.943

Нижче наведено графіки метрики mAP та метрик Precision та Recall для кожної епохи (рис 4.2.):

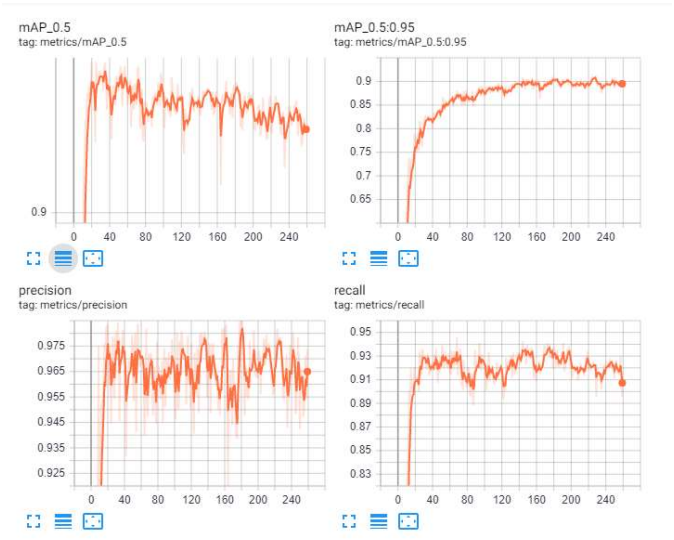


Рис. 4.2. Графіки метрик mAP, Precision, Recall для оптимізованого підходу

Нижче наведено графіки для різноманітних функцій втрат: за обмежувальними рамками, за виявленням об’єктів, за класифікацією об’єктів (рис 4.3):

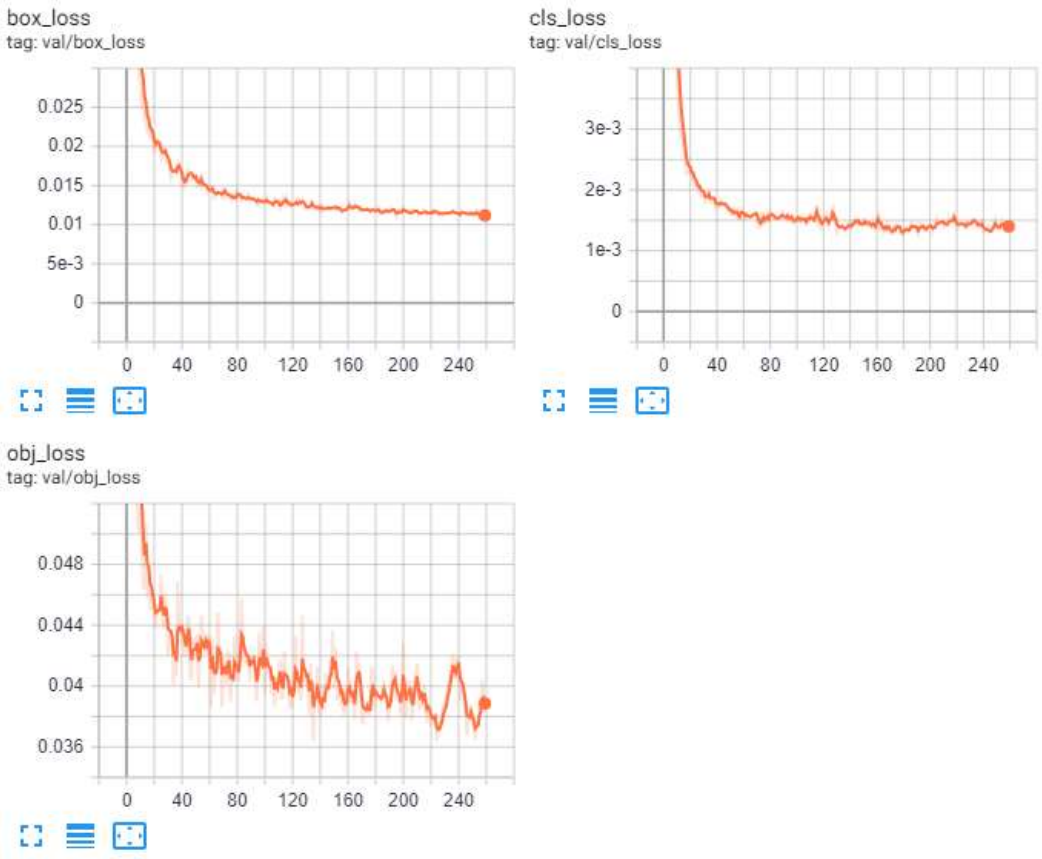


Рис. 4.3. Графіки функцій втрат для оптимізованого підходу YOLOv5

Результати навчання моделі YOLOv5 оптимізованого підходу з використанням оптимізатора Adam та початковим коефіцієнтом швидкості навчання рівним 0.001 наведено в таблиці 4.4

Таблиця 4.4.

Клас	Оптимізований підхід YOLOv5, LR=0.001
(Усі класи)	0.911
Текст	0.855
Заголовок	0.874
Список	0.922
Таблиця	0.969
Фігура	0.936

Додаткові графічні матеріали, пов’язані з оцінюванням точності розпізнавання подані у **додатках А-Б**. Результати роботи подані у **додатку В**.

У порівнянні з навчанням з оптимізатором SGD покращились класи «Текст», «Заголовок», «Список», але, водночас, погіршились результати для класу «Фігура».

4.2.2 Mask R-CNN

Результати моделі Mask R-CNN з розподілом по класах подано в таблиці 4.5. N – означає кількість епох.

Таблиця 4.5.

#	Клас	Mask R-CNN; N=12	Mask R-CNN; N=24
1	(Усі класи)	0.880	**0.886**
2	Текст	**0.794**	0.772
3	Заголовок	0.885	**0.891**
4	Список	0.864	**0.889**
5	Таблиця	0.944	**0.960**
6	Фігура	0.914	**0.915**

Графік зміни коефіцієнта навчання (Learning Rate) – навчання протягом 12 епох – показано на рис.4.4.

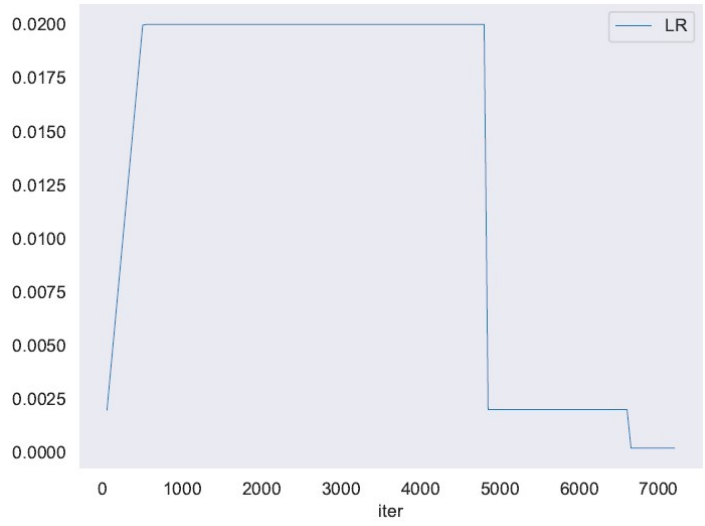


Рис. 4.4. Графік зміни коефіцієнта навчання (Learning Rate) для Mask R-CNN для 12 епох

Під час однієї епохи здійснюється 600 ітерацій (batch size дорівнює 2). Шаблон розкладу зміни коефіцієнта навчання (Learning Rate) для експерименту з кількістю епох N=24 має такий же вигляд, як і для N=12. (Тобто зміна коефіцієнта з 0.02 до 0.002 відбувається після завершення восьмої (4800-та ітерація) та шістнадцятої (9600-та ітерація) епохи відповідно та зміна від 0.002 до 0.0002 відбувається після завершення одинадцятої (6600-та ітерація) та двадцять другої (13200-та ітерація) епох відповідно).

Графіки функцій втрат – класифікаційної та регресійної обмежувальних рамок показано на рис.4.5: а) навчання 12 епох, б) навчання 24 епохи.

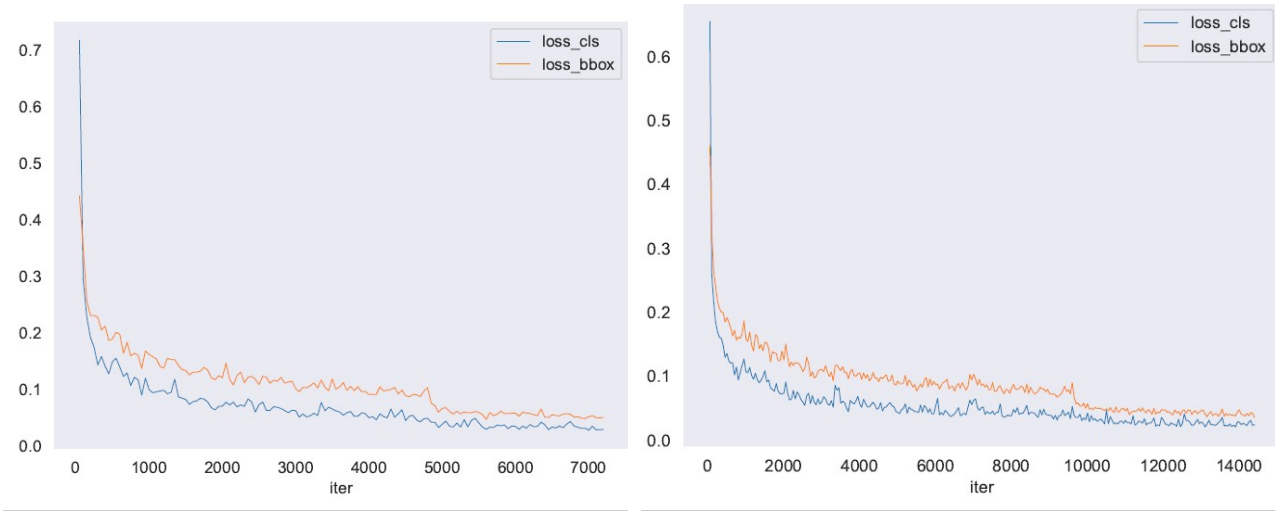


Рис. 4.5. Графіки функцій втрат для Mask R-CNN

а) – навчання протягом 12 епох б) навчання протягом 24 епох

Результати роботи Mask R-CNN подано у **додатку Г**.

Висновки до четвертого розділу

У роботі здійснено експерименти навчання різних моделей глибоких нейронних мереж на частині набору даних (датасету) PubLayNet, як двофазних детекторів об’єктів (Mask R-CNN), так і однофазних (RetinaNet, EfficientDet, YOLOv5). Проведені експерименти свідчать, що найкращу точність показав однофазний детектор YOLOv5. Однак, отримані результати щодо детекторів RetinaNet, EfficientDet свідчать про замалий час навчання цих моделей.

Рекомендовано збільшити час навчання даних моделей та розширити розміри вибірок даних датасету для досягнення кращої точності.

Для архітектури двофазного детектора Mask R-CNN здійснено експерименти з налаштування розкладу коефіцієнта навчання (learning rate) для різної кількості епох – дванадцяти та двадцяти чотирьох епох. Для 24 епох вдалося досягти кращого усередненого за класами результату (0.886 проти 0.880), однак для класу «Текст» результати на дванадцяти епохах виявились кращими (0.794 проти 0.772). Але, навіть такий результат для класу «Текст» не є прийнятним.

Експерименти з навчання однофазного детектора YOLOv5 засвідчили покращення у якості виявлення об’єктів класу «Текст». Зокрема, було проведено експерименти з різним початковим коефіцієнтом навчання та застосуванням різних оптимізаторів навчання – Stochastic Gradient Descent (SGD) та Adaptive Moment Estimation (Adam) для двох підходів – оригінального та оптимізованого (оптимізований підхід описано в підрозділі 4.1). Результати показали, що, незважаючи на незначне погіршення усередненого значення за класами (0.911 – для оптимізованого підходу проти 0.914 для оригінального підходу), все ж таки вдалося досягти кращої якості у виявленні об’єктів класу «Текст». (0.840 – для оригінального підходу; 0.853 – оптимізований підхід, SGD; 0.855 – оптимізований підхід, Adam).

ВИСНОВКИ

Завдяки стрімкому розвитку цифрових технологій задача аналізу структури документа стає все більш затребуваною, зокрема у випадку необхідності конвертації паперової сторінки до електронного вигляду з метою її редагування.

Алгоритми аналізу розмітки діляться на два типи: висхідний (Bottom-up approach) та низхідний (Top-down approach).

У даній роботі розглянуто декілька моделей глибоких нейронних мереж для розв'язання задачі локалізації та сегментації структурних елементів документа на зображенні сторінки.

Задовільні результати продемонстрували двофазний детектор Mask R-CNN та однофазний детектор YOLOv5. Завдяки оптимізації нейронної мережі YOLOv5 вдалося досягти найкращої якості виявлення об’єктів класу «Текст» з-поміж усіх проведених експериментів.

У подальшому, може бути здійснена інтеграція отриманої моделі нейронної мережі до систем розпізнавання структури документів та систем розпізнавання тексту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Haralick, R.M.: Document Image Understanding: Geometric and Logical Layout. in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Seattle, WA (1994) 385–390.
2. Namboodiri, Anoop & Jain, Anil. (2007). Document Structure and Layout Analysis. 10.1007/978-1-84628-726-8_2.
3. T. M. Breuel, “High performance document layout analysis,” in Symposium on Document Image Understanding Technology, Greenbelt, MD, 2003.
4. M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, et al., "DocBank: A benchmark dataset for document layout analysis", Proc. 28th Int. Conf. Comput. Linguistics, pp. 949-960, 2020.
5. Atena Farahmand, Abdolhossein Sarrafzadeh, & Jamshid Shanbezadeh (2013). Document Image Noises and Removal Methods. Lecture Notes in Engineering and Computer Science. 2202. 436-440.
6. K.-C. Fan, Y.-K. Wang, and T.-R. Lay, “Marginal noise removal of document images,” Proceedings of Sixth International Conference on Document Analysis and Recognition (ICDAR’01), pp. 317–321, 2001.
7. M. Agarwal and D. Doermann, “Clutter noise removal in binary document images,” in Proceedings of International Conference on Document Analysis and Recognition, pp. 556–560, 2009.
8. Tensmeyer, C., Martinez, T. Historical Document Image Binarization: A Review. SN COMPUT. SCI. 1, 173 (2020).
9. Otsu N. A threshold selection method from gray-level histograms. Trans Syst Man Cybern. 1979;9(1):62–6.
10. Niblack W. An introduction to digital image processing. Birkerod: Strandberg Publishing Company; 1985.
11. Sauvola J, Pietikäinen M. Adaptive document image binarization. Pattern Recognit. 2000;33(2):225–36.
12. Shafait, F., van Beusekom, J., Keysers, D. et al. Document cleanup using page frame detection. IJDAR 11, 81–96 (2008).

13. Breuel, T.M.: Two geometric algorithms for layout analysis. In: Proceedings of Document Analysis Systems. Lecture Notes in Computer Science, vol. 2423, Princeton, NY, USA, pp. 188–199 (2002).
14. N. Stamatopoulos, B. Gatos and T. Georgiou, Page Frame Detection for Double Page Document Images, in Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, 2010, pp. 401 – 408.
15. Le, D.X., Thoma, G.R., Wechsler, H.: Automated borders detection and adaptive segmentation for binary document images. In: 13th International Conference on Pattern Recognition, Vienna, Austria, pp. 737–741 (1996).
16. Cinque, L., Levialdi, S., Lombardi, L., Tanimoto, S.: Segmentation of page images having artifacts of photocopying and scanning. Pattern Recognit. 35(5), 1167–1177 (2002).
17. Peerawit,W., Kawtrakul, A.: Marginal noise removal from document images using edge density. In: 4th Information and Computer Engineering Postgraduate Workshop, Phuket, Thailand (2004).
18. Liang, J., Phillips, I.T., Haralick, R.M.: Performance evaluation of document structure extraction algorithms. Comput. Vis. Image Underst. 84(1), 144–159 (2001).
19. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area Voronoi diagram. Computer Vision and Image Understanding 70(3), 370–382 (1998).
20. S. Mao and T. Kanungo. Empirical performance evaluation methodology and its application to page segmentation algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence, 23(3):242–256, 2001.
21. O'Gorman, L.: The Document Spectrum for Page Layout Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1993), 1162–1173.
22. Wahl, F. and Wong, K. and Casey, R.: Block Segmentation and Text Extraction in Mixed Text/Image Documents. Graphical Models and Image Processing 20 (1982), 375–390.

23. L.A. Fletcher and R. Kasturi. A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 10, pp. 910-918, 1988.
24. G. Nagy, S. Seth, and M. Viswanathan. A Prototype Document Image Analysis System for Technical Journals, Computer, vol. 25, pp. 10-22, 1992.
25. H.S. Baird, S.E. Jones, and S.J. Fortune. Image Segmentation by Shape-Directed Covers, Proc. Int'l Conf. Pattern Recognition, pp. 820-825, June 1990.
26. Pavlidis, T. and Zhou, J. Page Segmentation and Classification. Graphical Models and Image Processing 54 (1992), 484–496.
27. R.M. Haralick and L.G. Shapiro, Computer and Robot Vision. Reading, Mass.: Addison-Wesley, 1992.
28. J. Duong, M. Ct, H. Emptoz and C. Suen. "Extraction of Text Areas in Printed Document Images". In ACM Symposium on Document Engineering: DocEng'01, November 9-10, pp. 157–165. Atlanta, USA, November 2001.
29. Y. Zheng, H. Li and D. Doermann. "Machine printed text and handwriting identification in noisy document images". Tech. rep., LAMP Lab, University of Maryland, College Park, 2002.
30. H. Baird, M. Moll, C. An, and M. Casey, "Document image content inventories," in Proceedings of SPIE/IS&T Document Recognition & Retrieval XIV Conference, vol. 6500, 2007.
31. C. Shin, D. Doermann, and A. Rosenfeld, "Classification of document pages using structure-based features," International Journal on Document Analysis and Recognition, vol. 3, no. 4, pp. 232–247, 2001.
32. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: TableBank: Table benchmark for image-based table detection and recognition. In: ICDAR (2019).
33. Deng, Y., Kanervisto, A., and Rush, A. M. (2016). What you get is what you see: A visual markup decompiler. CoRR, abs/1609.04938.
34. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
35. R. Girshick, "Fast R-CNN," in Proceedings of ICCV, 2015, pp. 1440 1448.

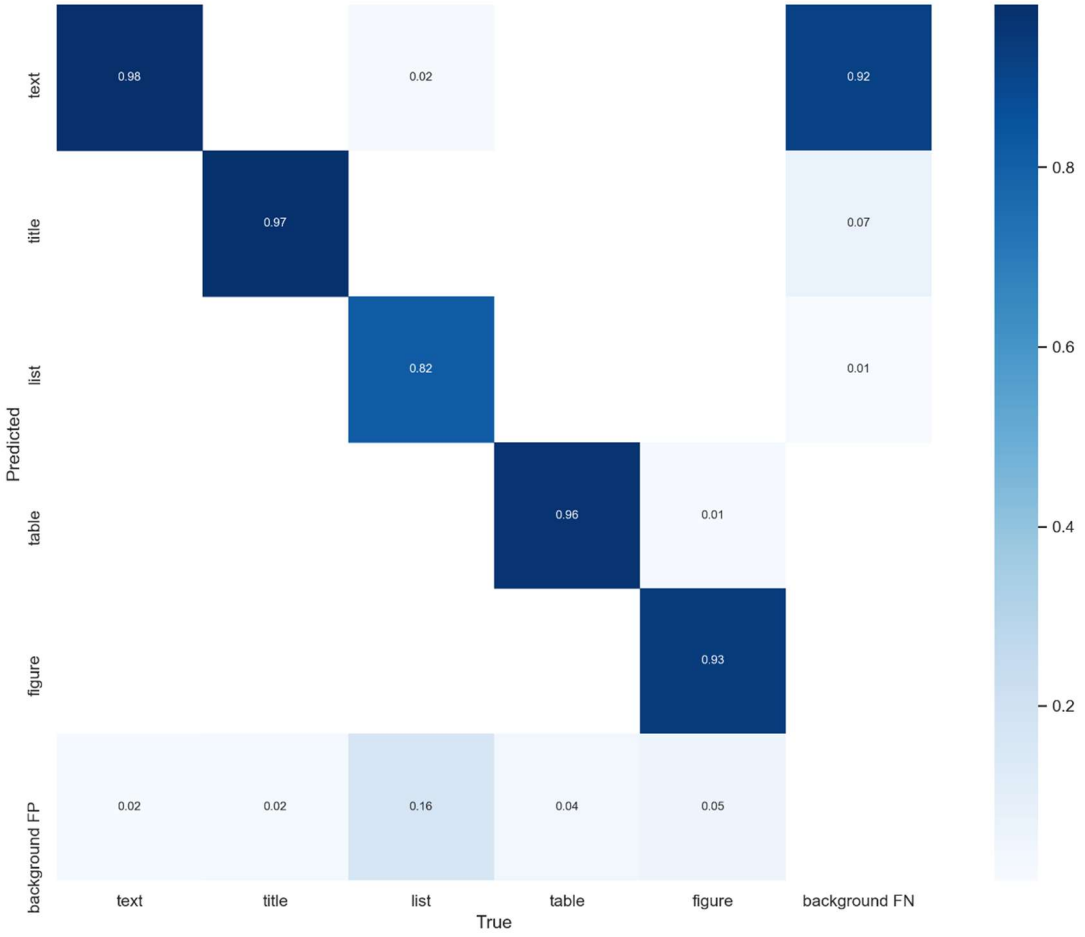
36. J. Fang, X. Tao, Z. Tang, R. Qiu, and Y. Liu. Dataset, groundtruth and performance metrics for table detection evaluation. In 2012 10th IAPR International Workshop on Document Analysis Systems, pages 445–449, 2012.
37. R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena, “Geometric layout analysis techniques for document image understanding: a review,” ITC-irst Technical Report, vol. 9703, no. 09, 1998.
38. P. W. J. Staar, M. Dolfi, C. Auer, and C. Bekas, “Corpus conversion service: A machine learning platform to ingest documents at scale,” in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD ’18. New York, NY, USA: ACM, 2018, pp. 774–782.
39. Zhong, Xu & Tang, Jianbin & Jimeno-Yepes, Antonio. (2019). PubLayNet: Largest Dataset Ever for Document Layout Analysis. 10.1109/ICDAR.2019.00166.
40. R. Girshick, I. Radosavovic, G. Gkioxari, P. Doll´ar, and K. He, “Detectron,” 2018. / [Електронний ресурс]
Режим доступу: https://github.com/facebookresearch/detectron.
41. Lu, Xin & Li, Quanquan & Li, Buyu & Yan, Junjie. (2020). MimicDet: Bridging the Gap Between One-Stage and Two-Stage Object Detection.
42. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016).
43. T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999-3007.
44. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781–10790 (2020).
45. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards RealTime Object Detection with Region Proposal Networks,” in Proceedings of NIPS, 2015, pp. 91–99.

46. K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in ´ Proceedings of ICCV, 2017, pp. 2961–2969.
47. Padilla, R.; Netto, S.L.; da Silva, E.A.B. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the 27th International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020; pp. 237–242.
48. D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, “IOU loss for 2d/3d object detection,” in 2019 International Conference on3D Vision (3DV). IEEE, 2019, pp. 85–94.
49. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
50. Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He,Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. CVPR, 2017.
51. Ultralytics. YOLOv5 in PyTorch. [Електронний ресурс] / Режим доступу: https://github.com/ultralytics/yolov5/tree/v4.0
52. Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of cnn. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPR Workshop), 2020.
53. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 37(9):1904–1916, 2015.
54. Yap MH., Hachiuma R, Alavi A, Brungel R, Goyal M, Zhu H, Cassidy B, Ruckert J, Olshansky M, Huang X, et al. Deep learning in diabetic foot ulcers detection: a comprehensive evaluation; 2020.
55. Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. ICML, 2019.

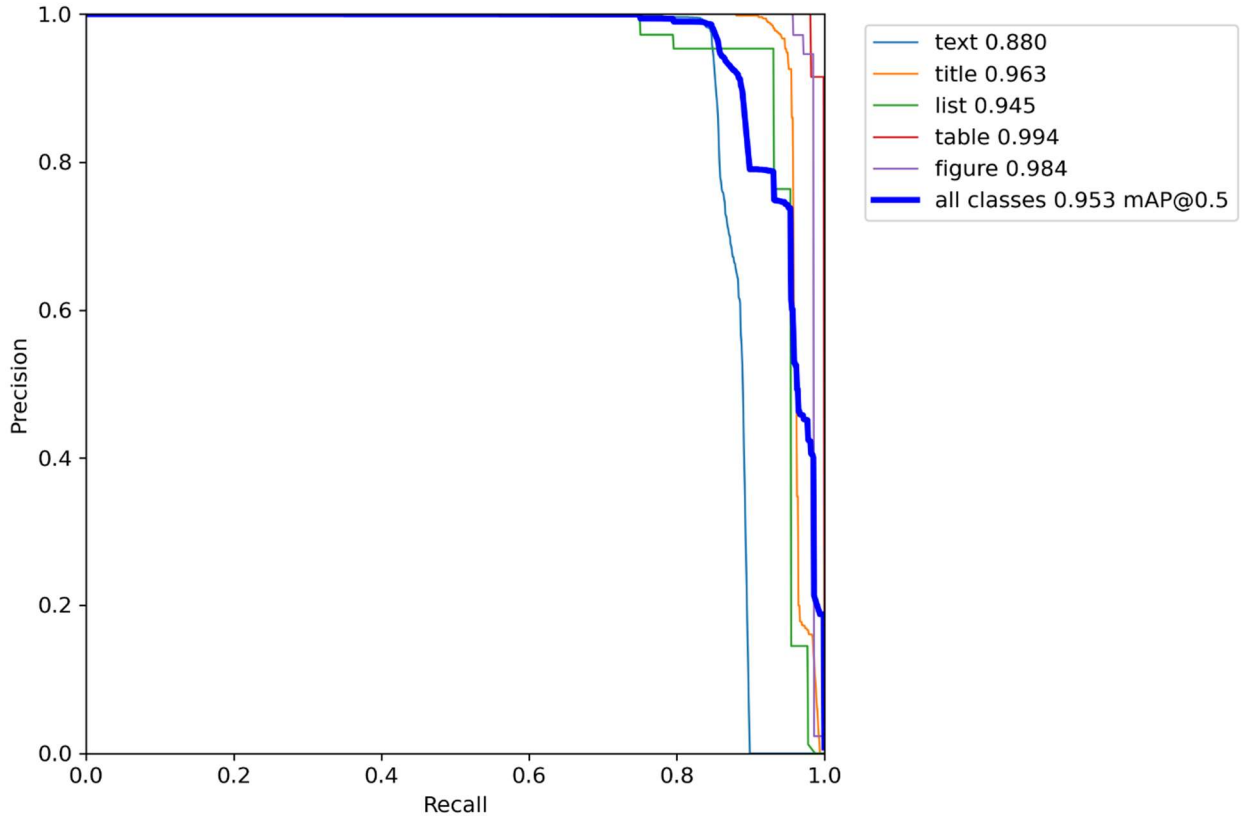
56. Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. CVPR, 2018.
57. Henon, Y. Pytorch implementation of RetinaNet object detection. [Електронний ресурс]. Режим доступу: https://github.com/yhenon/pytorch-retinanet
58. Signatrix GmbH. A Pytorch Implementation of EfficientDet Object Detection. 2020 [Електронний ресурс].
Режим доступу: https://github.com/signatrix/efficientdet
59. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019.
60. Hendrycks, Dan & Gimpel, Kevin. (2016). Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units.

ДОДАТКИ

Додаток А. YOLOv5 – Оптимізований підхід – Confusion Matrix



Додаток Б. YOLOv5 – Оптимізований підхід – крива Precision-Recall



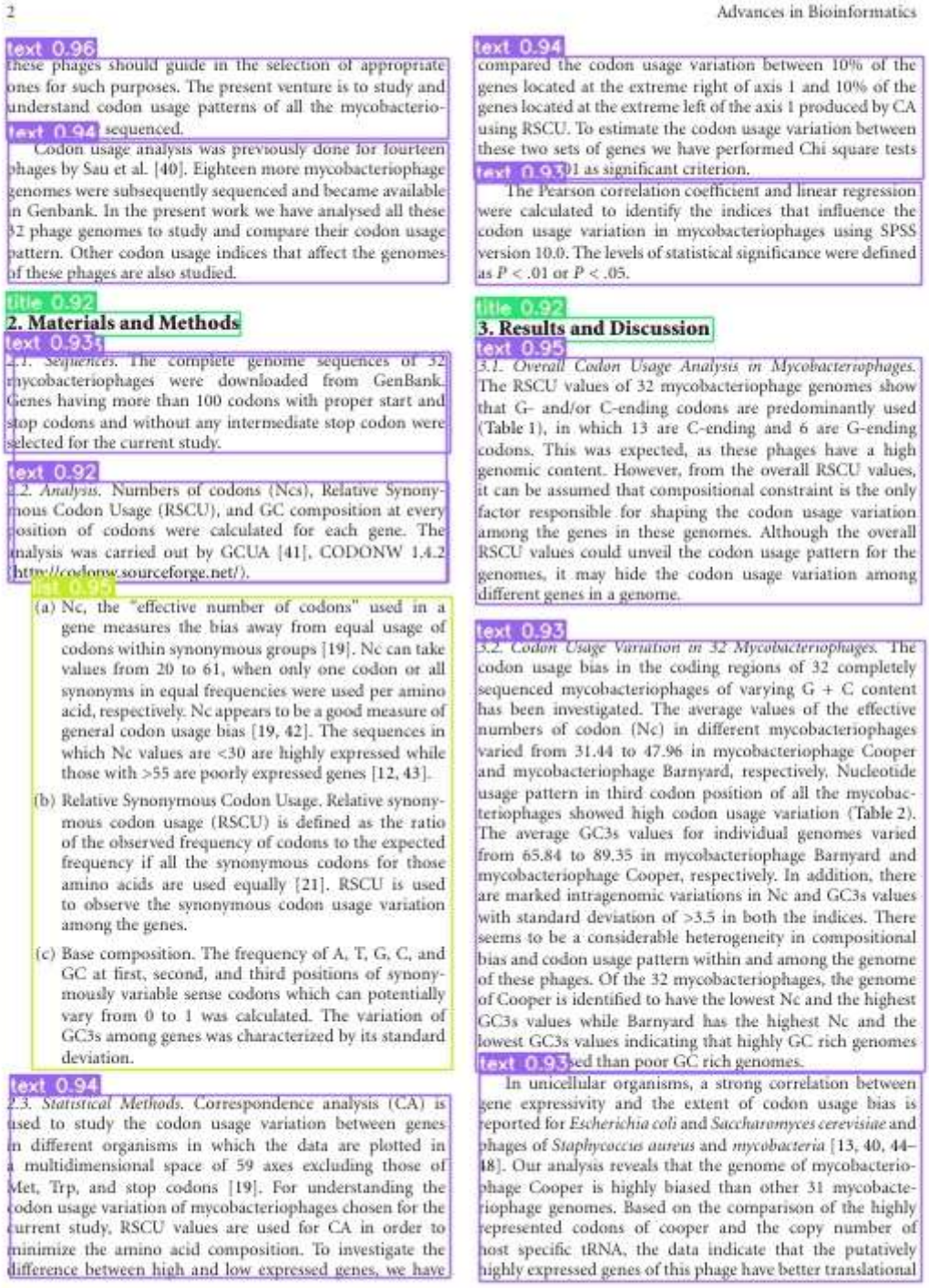

2
Advances in Bioinformatics
text 0.96
these phages should guide in the selection of appropriate ones for such purposes. The present venture is to study and understand codon usage patterns of all the mycobacterio-
text 0.94
sequenced.
Codon usage analysis was previously done for fourteen phages by Sau et al. [40]. Eighteen more mycobacteriophage genomes were subsequently sequenced and became available in Genbank. In the present work we have analysed all these 32 phage genomes to study and compare their codon usage pattern. Other codon usage indices that affect the genomes of these phages are also studied.
title 0.92
2. Materials and Methods
text 0.93
2.1. Sequences. The complete genome sequences of 32 mycobacteriophages were downloaded from GenBank. Genes having more than 100 codons with proper start and stop codons and without any intermediate stop codon were selected for the current study.
text 0.92
2.2. Analysis. Numbers of codons (Ncs), Relative Synonymous Codon Usage (RSCU), and GC composition at every position of codons were calculated for each gene. The analysis was carried out by GCUA [41], CODONW 1.4.2 (http://codonw.sourceforge.net/).
list 0.95
(a) Nc, the "effective number of codons" used in a gene measures the bias away from equal usage of codons within synonymous groups [19]. Nc can take values from 20 to 61, when only one codon or all synonyms in equal frequencies were used per amino acid, respectively. Nc appears to be a good measure of general codon usage bias [19, 42]. The sequences in which Nc values are <30 are highly expressed while those with >55 are poorly expressed genes [12, 43].
(b) Relative Synonymous Codon Usage. Relative synonymous codon usage (RSCU) is defined as the ratio of the observed frequency of codons to the expected frequency if all the synonymous codons for those amino acids are used equally [21]. RSCU is used to observe the synonymous codon usage variation among the genes.
(c) Base composition. The frequency of A, T, G, C, and GC at first, second, and third positions of synonymously variable sense codons which can potentially vary from 0 to 1 was calculated. The variation of GC3s among genes was characterized by its standard deviation.
text 0.94
2.3. Statistical Methods. Correspondence analysis (CA) is used to study the codon usage variation between genes in different organisms in which the data are plotted in a multidimensional space of 59 axes excluding those of Met, Trp, and stop codons [19]. For understanding the codon usage variation of mycobacteriophages chosen for the current study, RSCU values are used for CA in order to minimize the amino acid composition. To investigate the difference between high and low expressed genes, we have
text 0.94
compared the codon usage variation between 10% of the genes located at the extreme right of axis 1 and 10% of the genes located at the extreme left of the axis 1 produced by CA using RSCU. To estimate the codon usage variation between these two sets of genes we have performed Chi square tests
text 0.93
01 as significant criterion.
The Pearson correlation coefficient and linear regression were calculated to identify the indices that influence the codon usage variation in mycobacteriophages using SPSS version 10.0. The levels of statistical significance were defined as $P < .01$ or $P < .05$.
title 0.92
3. Results and Discussion
text 0.95
3.1. Overall Codon Usage Analysis in Mycobacteriophages. The RSCU values of 32 mycobacteriophage genomes show that G- and/or C-ending codons are predominantly used (Table 1), in which 13 are C-ending and 6 are G-ending codons. This was expected, as these phages have a high genomic content. However, from the overall RSCU values, it can be assumed that compositional constraint is the only factor responsible for shaping the codon usage variation among the genes in these genomes. Although the overall RSCU values could unveil the codon usage pattern for the genomes, it may hide the codon usage variation among different genes in a genome.
text 0.93
3.2. Codon Usage Variation in 32 Mycobacteriophages. The codon usage bias in the coding regions of 32 completely sequenced mycobacteriophages of varying G + C content has been investigated. The average values of the effective numbers of codon (Nc) in different mycobacteriophages varied from 31.44 to 47.96 in mycobacteriophage Cooper and mycobacteriophage Barnyard, respectively. Nucleotide usage pattern in third codon position of all the mycobacteriophages showed high codon usage variation (Table 2). The average GC3s values for individual genomes varied from 65.84 to 89.35 in mycobacteriophage Barnyard and mycobacteriophage Cooper, respectively. In addition, there are marked intragenomic variations in Nc and GC3s values with standard deviation of >3.5 in both the indices. There seems to be a considerable heterogeneity in compositional bias and codon usage pattern within and among the genome of these phages. Of the 32 mycobacteriophages, the genome of Cooper is identified to have the lowest Nc and the highest GC3s values while Barnyard has the highest Nc and the lowest GC3s values indicating that highly GC rich genomes
text 0.93
ed than poor GC rich genomes.
In unicellular organisms, a strong correlation between gene expressivity and the extent of codon usage bias is reported for *Escherichia coli* and *Saccharomyces cerevisiae* and phages of *Staphylococcus aureus* and *mycobacteria* [13, 40, 44–48]. Our analysis reveals that the genome of mycobacteriophage Cooper is highly biased than other 31 mycobacteriophage genomes. Based on the comparison of the highly represented codons of cooper and the copy number of host specific tRNA, the data indicate that the putatively highly expressed genes of this phage have better translational

figure 0.93

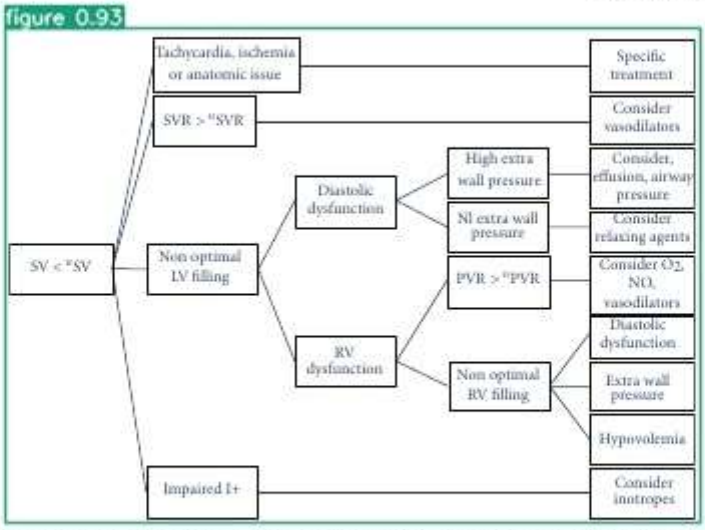


Figure 2: Subsequent algorithm dedicated to SV analysis. The needed SVR is that specific value of SVR allowing the generation of the minimum blood pressure (usually set at a mean value of 65 mmHg) with the needed CO. Similarly for needed PVR.

text 0.95

the bias introduced by the missing input data. There are theoretically three solutions. First, it is sometimes possible to find a surrogate of a given input variable assuming that the information provided is close. This is the case for example replacing SvO_2 by the central venous O_2 saturation ($ScVO_2$) or SaO_2 by the pulse oximetry (SpO_2) or total peripheral resistance instead of SVR when right aerial pressure is missing. Second, it is possible to give a fixed value to a given variable. For example, we may enter in the model directly at the level shown in Figure 2, assuming that the needed SV is the normal SV as a function of age and gender since the adaptative mechanisms of CO to needs are mostly resulting from a regulation of the heart rate. Third, it is possible to estimate a continuous variable from discontinuous clinical or paraclinical analysis. For example, VO_2 is often not monitored and VO_2 needs unknown. Consequently $VO_2/{}^{n}VO_2$ ratio <1 is suspected by the presence of clinical signs of shock or by an increase in blood lactate over time. Alternatively, a change in VO_2 needs (therefore translated to CO and EO_2 needs) can, be estimated by age, gender, body size, body temperature and/or empiric estimation of the change in metabolic needs due to underlying or associated pathologies. In any case, if we restrict our analysis to a part of the global model, we must assume that the blind part is fixed and given an approximate value. The reliability of the model is therefore reduced.

text 0.92

2.3. What Are the Specific Quality Criteria and Tolerance for Each Input Variable? The quality criteria and tolerance for the CO monitoring have been reviewed recently. [8, 9] The same effort must be done for each input variable of the model. Basically these criteria are very similar for all quantitative variables.

list 0.95

1) The accuracy is how close the value is to a gold standard. It is estimated by the mean difference (bias) with the true value given by the gold standard.
2) The linearity is the capability of maintaining constant the ratio between the physiologic signal and the electric output signal. Therefore the bias is constant. It can be verified by comparing the regression curve of the bias with the identity line.
3) The precision is the ability to indicate the same value when the physiologic signal is stable. In other words, it is the variation due to random error in the signal processing. It can be estimated by the standard deviation/mean value when the physiologic signal is stable. The least minimum significant change (smallest change indicating a real change) is a direct consequence of precision.
4) The resolution is the smallest change that the device can detect.
5) The stability is the capability of maintaining the preceding quality criteria unchanged during time (without drift).
6) The measuring range is the boundaries of value where the preceding quality criteria are found acceptable
7) The responsiveness is the delay between a real change in the physiologic signal and a change greater than the least minimum significant change in the observed value. Coupled with the linearity, it determines the accuracy of the amplitude response.

on modern breast imaging [12]. Some histopathological studies have not shown such correlation [13, 14], although these studies did not use large-section histology. Recently, the disease extent was shown to correlate with presence of lymphovascular invasion and lymph node status [9]. Other studies have evaluated the distribution of lesions as markers for lymphovascular invasion and lymph node status, using histology [15, 16].

The aim of this study was to evaluate the disease extent determined with large-section histology, as a prognostic marker for local recurrence in patients treated with breast-conserving surgery, and in addition, to find an appropriate cutoff for defining extensive disease.

2. Materials and Methods

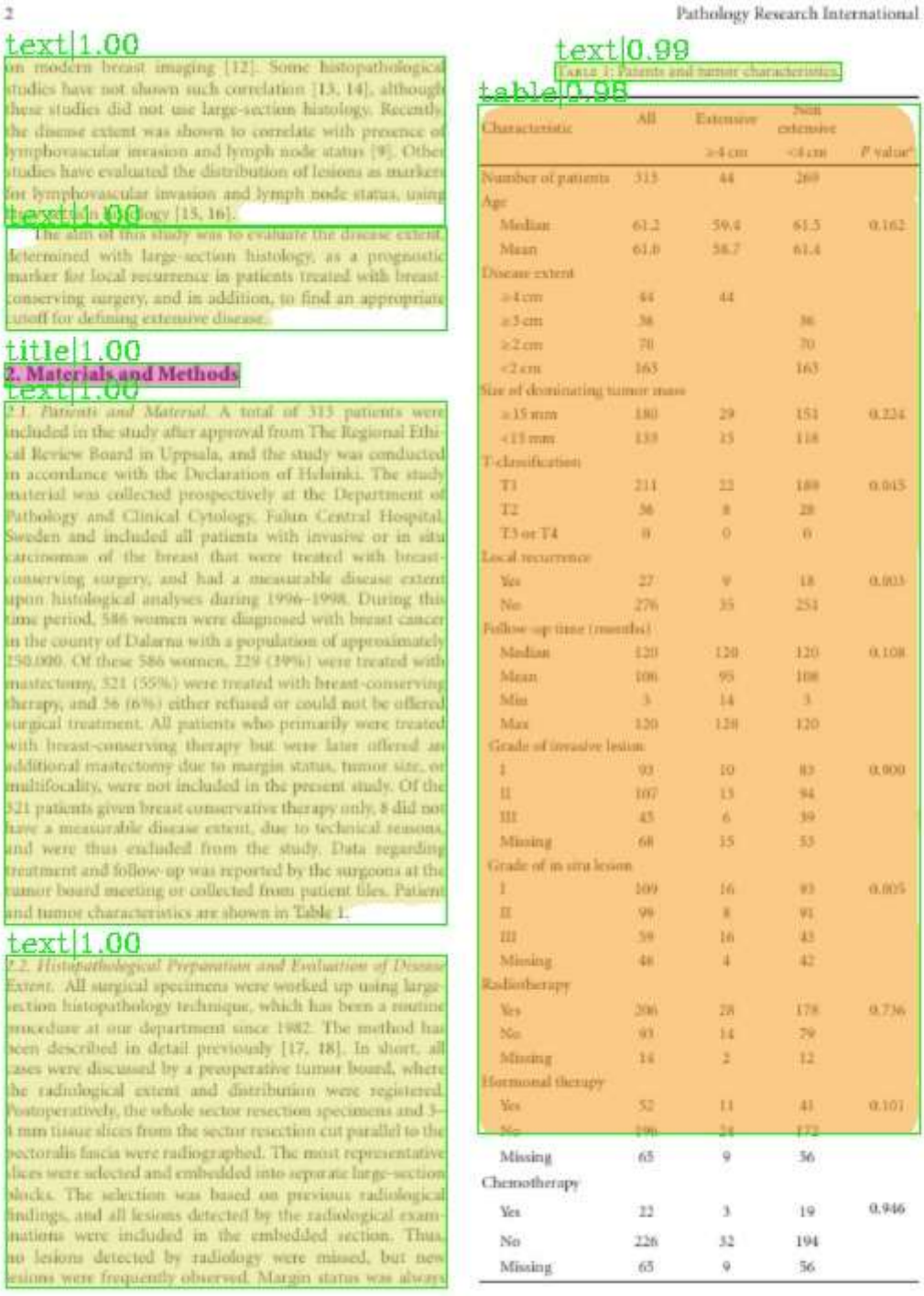
2.1. Patients and Material. A total of 313 patients were included in the study after approval from The Regional Ethical Review Board in Uppsala, and the study was conducted in accordance with the Declaration of Helsinki. The study material was collected prospectively at the Department of Pathology and Clinical Cytology, Falun Central Hospital, Sweden and included all patients with invasive or in situ carcinomas of the breast that were treated with breast-conserving surgery, and had a measurable disease extent upon histological analyses during 1996–1998. During this time period, 586 women were diagnosed with breast cancer in the county of Dalarna with a population of approximately 250.000. Of these 586 women, 229 (39%) were treated with mastectomy, 321 (55%) were treated with breast-conserving therapy, and 36 (6%) either refused or could not be offered surgical treatment. All patients who primarily were treated with breast-conserving therapy but were later offered an additional mastectomy due to margin status, tumor size, or multifocality, were not included in the present study. Of the 321 patients given breast conservative therapy only, 8 did not have a measurable disease extent, due to technical reasons, and were thus excluded from the study. Data regarding treatment and follow-up was reported by the surgeons at the tumor board meeting or collected from patient files. Patient and tumor characteristics are shown in Table 1.

2.2. Histopathological Preparation and Evaluation of Disease Extent. All surgical specimens were worked up using large-section histopathology technique, which has been a routine procedure at our department since 1982. The method has been described in detail previously [17, 18]. In short, all cases were discussed by a preoperative tumor board, where the radiological extent and distribution were registered. Postoperatively, the whole sector resection specimens and 3–4 mm tissue slices from the sector resection cut parallel to the pectoralis fascia were radiographed. The most representative slices were selected and embedded into separate large-section blocks. The selection was based on previous radiological findings, and all lesions detected by the radiological examinations were included in the embedded section. Thus, no lesions detected by radiology were missed, but new lesions were frequently observed. Margin status was always

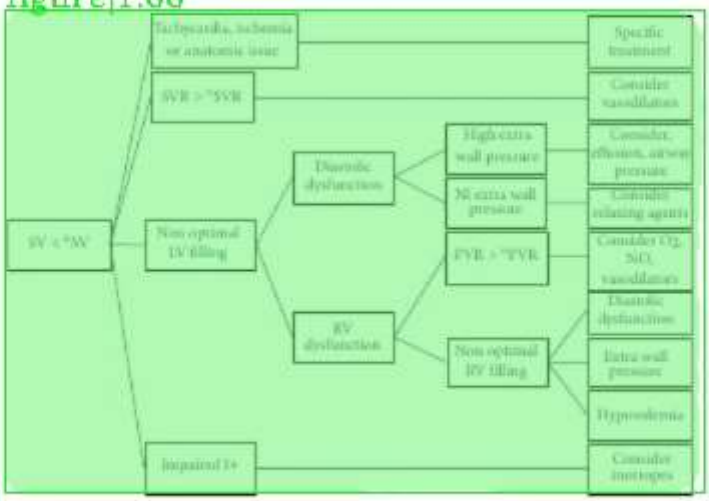
Table 1: Patients and tumor characteristics.

Characteristic	All	Extensive	Non extensive	
		≥4 cm	<4 cm	*P* value[a]
Number of patients	313	44	269	
Age				
Median	61.2	59.4	61.5	0.162
Mean	61.0	58.7	61.4	
Disease extent				
≥4 cm	44	44		
≥3 cm	36		36	
≥2 cm	70		70	
<2 cm	163		163	
Size of dominating tumor mass				
≥15 mm	180	29	151	0.224
<15 mm	133	15	118	
T-classification				
T1	211	22	189	0.045
T2	36	8	28	
T3 or T4	0	0	0	
Local recurrence				
Yes	27	9	18	0.003
No	276	35	251	
Follow-up time (months)				
Median	120	120	120	0.108
Mean	106	95	108	
Min	3	14	3	
Max	120	120	120	
Grade of invasive lesion				
I	93	10	83	0.900
II	107	13	94	
III	45	6	39	
Missing	68	15	53	
Grade of in situ lesion				
I	109	16	93	0.005
II	99	8	91	
III	59	16	43	
Missing	46	4	42	
Radiotherapy				
Yes	206	28	178	0.736
No	93	14	79	
Missing	14	2	12	
Hormonal therapy				
Yes	52	11	41	0.101
No	196	24	172	
Missing	65	9	56	
Chemotherapy				
Yes	22	3	19	0.946
No	226	32	194	
Missing	65	9	56	

Додаток Г. Mask R-CNN – виявлення об’єктів та семантична сегментація



figure|1.00



text|0.81

Figure 2: Subsequent algorithm dedicated to SV analysis. The needed SVR is that specific value of SVR allowing the generation of the minimum blood pressure (usually set at a mean value of 65 mmHg) with the needed CO. Similarly for needed PVR.

text|1.00

The bias introduced by the missing input data. There are theoretically three solutions. First, it is sometimes possible to find a surrogate of a given input variable assuming that the information provided is close. This is the case for example replacing SvO_2 by the central venous O_2 saturation ($ScVO_2$) or SaO_2 by the pulse oximetry (SpO_2) or total peripheral resistance instead of SVR when right atrial pressure is missing. Second, it is possible to give a fixed value to a given variable. For example, we may enter in the model directly at the level shown in Figure 2, assuming that the needed SV is the normal SV as a function of age and gender since the adaptative mechanisms of CO to needs are mostly resulting from a regulation of the heart rate. Third, it is possible to estimate a continuous variable from discontinuous clinical or paraclinical analysis. For example, VO_2 is often not monitored and VO_2 needs unknown. Consequently VO_2/°VO_2 ratio < 1 is suspected by the presence of clinical signs of shock or by an increase in blood lactate over time. Alternatively, a change in VO_2 needs (therefore translated to CO and EO_2 needs) can be estimated by age, gender, body size, body temperature and/or empiric estimation of the change in metabolic needs due to underlying or associated pathologies. In any case, if we restrict our analysis to a part of the global model, we must assume that the blind part is fixed and given an approximate value. The reliability of the model is therefore reduced.

text|1.00

2.3. *What Are the Specific Quality Criteria and Tolerance for Each Input Variable?* The quality criteria and tolerance for the CO monitoring have been reviewed recently. [8, 9] The same effort must be done for each input variable of the model. Basically these criteria are very similar for all quantative variables.

list|0.99

(1) The accuracy is how close the value is to a gold standard. It is estimated by the mean difference (bias) with the true value given by the gold standard.

(2) The linearity is the capability of maintaining constant the ratio between the physiologic signal and the electric output signal. Therefore the bias is constant. It can be verified by comparing the regression curve of the bias with the identity line.

(3) The precision is the ability to indicate the same value when the physiologic signal is stable. In other words, it is the variation due to random error in the signal processing. It can be estimated by the standard deviation/mean value when the physiologic signal is stable. The least minimum significant change (smallest change indicating a real change) is a direct consequence of precision.

(4) The resolution is the smallest change that the device can detect.

(5) The stability is the capability of maintaining the preceding quality criteria unchanged during time (without drift).

(6) The measuring range is the boundaries of value where the preceding quality criteria are found acceptable.

(7) The responsiveness is the delay between a real change in the physiologic signal and a change greater than the least minimum significant change in the observed value. Coupled with the linearity, it determines the accuracy of the amplitude response.

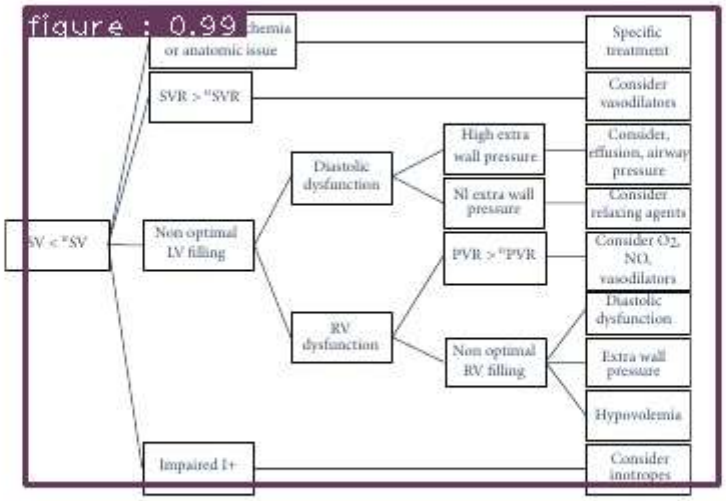


Figure 2: Subsequent algorithm dedicated to SV analysis. The needed SV
minimum blood pressure (usually set at a mean value of 65 mmHg) with th

list : 0.52 of SVR allowing the generation of th
for needed PVR.

text : 1.00 the missing input data. There are theoretically three solutions. First, it is sometimes possible to find a surrogate of a given input variable assuming that the information provided is close. This is the case for example replacing SvO_2 by the central venous O_2 saturation ($ScVO_2$) or SaO_2 by the pulse oximetry (SpO_2) or total peripheral resistance instead of SVR when right aerial pressure is missing. Second, it is possible to give a fixed value to a given variable. For example, we may enter in the model directly at the level shown in Figure 2, assuming that the needed SV is the normal SV as a function of age and gender since the adaptative mechanisms of CO to needs are mostly resulting from a regulation of the heart rate. Third, it is possible to estimate a continuous variable from discontinuous clinical or paraclinical analysis. For example, VO_2 is often not monitored and VO_2 needs unknown. Consequently $VO_2/°VO_2$ ratio <1 is suspected by the presence of clinical signs of shock or by an increase in blood lactate over time. Alternatively, a change in VO_2 needs (therefore translated to CO and EO_2 needs) can, be estimated by age, gender, body size, body temperature and/or empiric estimation of the change in metabolic needs due to underlying or associated pathologies. In any case, if we restrict our analysis to a part of the global model, we must assume that the blind part is fixed and given an approximate value. The reliability of the model is therefore reduced.

text : 0.97 *Quality Criteria and Tolerance for Each Input Variable.* The quality criteria and tolerance for the CO monitoring have been reviewed recently. [8, 9] The same effort must be done for each input variable of the model. Basically these criteria are very similar for all quantitative variables.

list : 0.82 estimated by the mean difference (bias with the true value given by the gold standard.

(2) The linearity is the capability of maintaining constan the ratio between the physiologic signal and th electric output signal. Therefore the bias is constant It can be verified by comparing the regression curv of the bias with the identity line.

(3) The precision is the ability to indicate the sam value when the physiologic signal is stable. In othe words, it is the variation due to random error i the signal processing. It can be estimated by th standard deviation/mean value when the physiologi signal is stable. The least minimum significant chang (smallest change indicating a real change) is a direc consequence of precision.

(4) The resolution is the smallest change that the devic can detect.

(5) The stability is the capability of maintaining th preceding quality criteria unchanged during tim (without drift).

(6) The measuring range is the boundaries of valu where the preceding quality criteria are found accept able

(7) The responsiveness is the delay between a real chang in the physiologic signal and a change greater tha the least minimum significant change in the observe value. Coupled with the linearity, it determines th accuracy of the amplitude response.

text : 1.00 ging [12]. Some histopathological studies have not shown such correlation [13, 14], although these studies did not use large-section histology. Recently, the disease extent was shown to correlate with presence of lymphovascular invasion and lymph node status [9]. Other studies have evaluated the distribution of lesions as markers for lymphovascular invasion and lymph node status, using large-section histology [15, 16].

text : 0.98 ly was to evaluate the disease extent determined with large-section histology, as a prognostic marker for local recurrence in patients treated with breast-conserving surgery, and in addition, to find an appropriate cutoff for defining extensive disease.

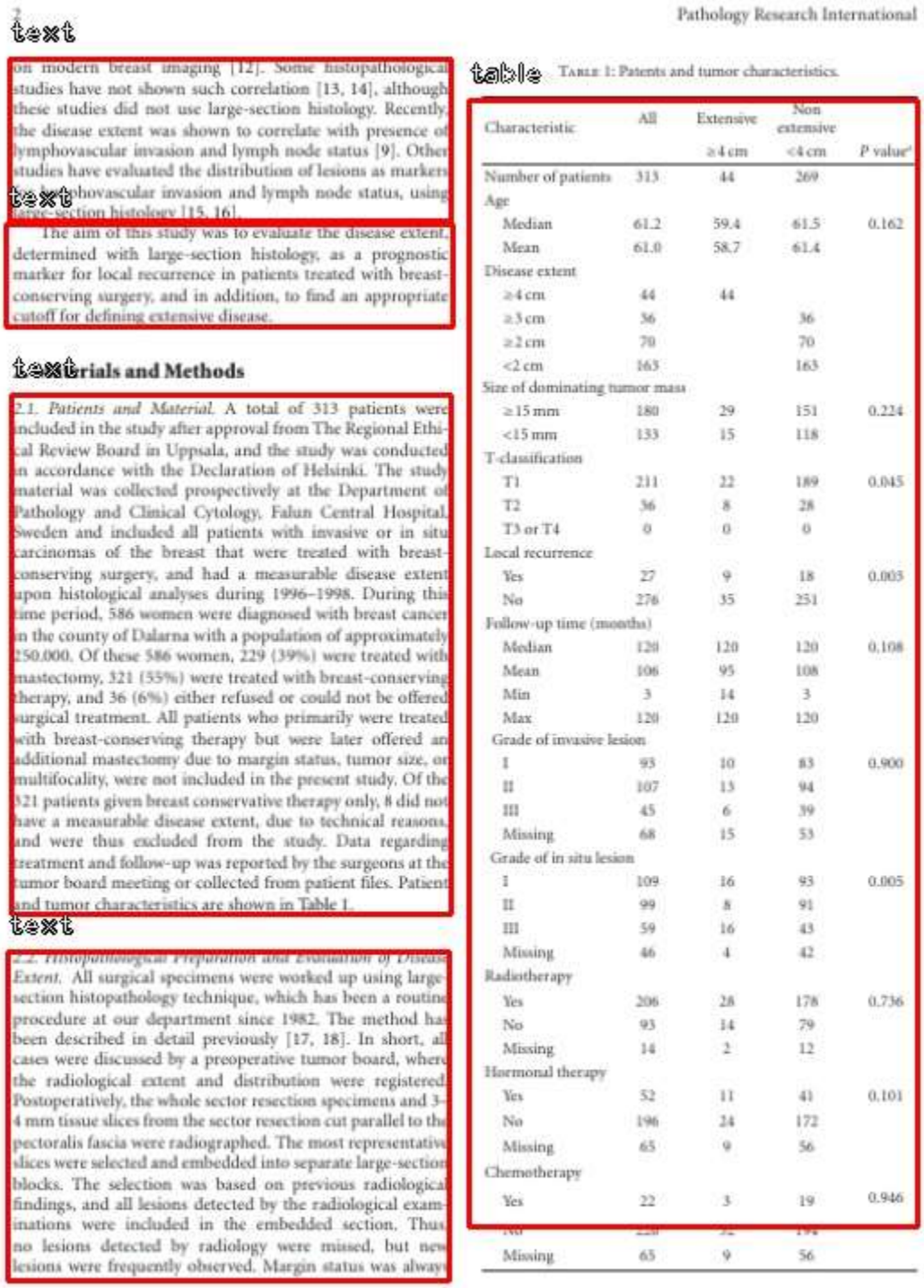
2. Materials and Methods

text : 1.00 *rial.* A total of 313 patients were included in the study after approval from The Regional Ethical Review Board in Uppsala, and the study was conducted in accordance with the Declaration of Helsinki. The study material was collected prospectively at the Department of Pathology and Clinical Cytology, Falun Central Hospital, Sweden and included all patients with invasive or in situ carcinomas of the breast that were treated with breast-conserving surgery, and had a measurable disease extent upon histological analyses during 1996–1998. During this time period, 586 women were diagnosed with breast cancer in the county of Dalarna with a population of approximately 250.000. Of these 586 women, 229 (39%) were treated with mastectomy, 321 (55%) were treated with breast-conserving therapy, and 36 (6%) either refused or could not be offered surgical treatment. All patients who primarily were treated with breast-conserving therapy but were later offered an additional mastectomy due to margin status, tumor size, or multifocality, were not included in the present study. Of the 321 patients given breast conservative therapy only, 8 did not have a measurable disease extent, due to technical reasons and were thus excluded from the study. Data regarding treatment and follow-up was reported by the surgeons at the tumor board meeting or collected from patient files. Patient and tumor characteristics are shown in Table 1.

text : 1.00 *eparation and Evaluation of Disease Extent.* All surgical specimens were worked up using large-section histopathology technique, which has been a routine procedure at our department since 1982. The method has been described in detail previously [17, 18]. In short, all cases were discussed by a preoperative tumor board, where the radiological extent and distribution were registered. Postoperatively, the whole sector resection specimens and 3–4 mm tissue slices from the sector resection cut parallel to the pectoralis fascia were radiographed. The most representative slices were selected and embedded into separate large-section blocks. The selection was based on previous radiological findings, and all lesions detected by the radiological examinations were included in the embedded section. Thus, no lesions detected by radiology were missed, but new lesions were frequently observed. Margin status was always

table : 0.95 and tumor characteristics.

Characteristic	All	Extensive	Non extensive	
		≥4 cm	<4 cm	*P* value[a]
Number of patients	313	44	269	
Age				
Median	61.2	59.4	61.5	0.162
Mean	61.0	58.7	61.4	
Disease extent				
≥4 cm	44	44		
≥3 cm	36		36	
≥2 cm	70		70	
<2 cm	163		163	
Size of dominating tumor mass				
≥15 mm	180	29	151	0.224
<15 mm	133	15	118	
T-classification				
T1	211	22	189	0.045
T2	36	8	28	
T3 or T4	0	0	0	
Local recurrence				
Yes	27	9	18	0.003
No	276	35	251	
Follow-up time (months)				
Median	120	120	120	0.108
Mean	106	95	108	
Min	3	14	3	
Max	120	120	120	
Grade of invasive lesion				
I	93	10	83	0.900
II	107	13	94	
III	45	6	39	
Missing	68	15	53	
Grade of in situ lesion				
I	109	16	93	0.005
II	99	8	91	
III	59	16	43	
Missing	46	4	42	
Radiotherapy				
Yes	206	28	178	0.736
No	93	14	79	
Missing	14	2	12	
Hormonal therapy				
Yes	52	11	41	0.101
No	196	24	172	
Missing	65	9	56	
Chemotherapy				
Yes	22	3	19	0.946
No	226	32	194	
Missing	65	9	56	



text

on modern breast imaging [12]. Some histopathological studies have not shown such correlation [13, 14], although these studies did not use large-section histology. Recently, the disease extent was shown to correlate with presence of lymphovascular invasion and lymph node status [9]. Other studies have evaluated the distribution of lesions as markers for lymphovascular invasion and lymph node status, using large-section histology [15, 16].

text

The aim of this study was to evaluate the disease extent, determined with large-section histology, as a prognostic marker for local recurrence in patients treated with breast-conserving surgery, and in addition, to find an appropriate cutoff for defining extensive disease.

text rials and Methods

2.1. *Patients and Material.* A total of 313 patients were included in the study after approval from The Regional Ethical Review Board in Uppsala, and the study was conducted in accordance with the Declaration of Helsinki. The study material was collected prospectively at the Department of Pathology and Clinical Cytology, Falun Central Hospital, Sweden and included all patients with invasive or in situ carcinomas of the breast that were treated with breast-conserving surgery, and had a measurable disease extent upon histological analyses during 1996–1998. During this time period, 586 women were diagnosed with breast cancer in the county of Dalarna with a population of approximately 250.000. Of these 586 women, 229 (39%) were treated with mastectomy, 321 (55%) were treated with breast-conserving therapy, and 36 (6%) either refused or could not be offered surgical treatment. All patients who primarily were treated with breast-conserving therapy but were later offered an additional mastectomy due to margin status, tumor size, or multifocality, were not included in the present study. Of the 321 patients given breast conservative therapy only, 8 did not have a measurable disease extent, due to technical reasons, and were thus excluded from the study. Data regarding treatment and follow-up was reported by the surgeons at the tumor board meeting or collected from patient files. Patient and tumor characteristics are shown in Table 1.

text

2.2. *Histopathological Preparation and Evaluation of Disease Extent.* All surgical specimens were worked up using large-section histopathology technique, which has been a routine procedure at our department since 1982. The method has been described in detail previously [17, 18]. In short, all cases were discussed by a preoperative tumor board, where the radiological extent and distribution were registered. Postoperatively, the whole sector resection specimens and 3–4 mm tissue slices from the sector resection cut parallel to the pectoralis fascia were radiographed. The most representative slices were selected and embedded into separate large-section blocks. The selection was based on previous radiological findings, and all lesions detected by the radiological examinations were included in the embedded section. Thus, no lesions detected by radiology were missed, but new lesions were frequently observed. Margin status was always

table

Table 1: Patients and tumor characteristics.

Characteristic	All	Extensive ≥4 cm	Non extensive <4 cm	P value[a]
Number of patients	313	44	269	
Age				
Median	61.2	59.4	61.5	0.162
Mean	61.0	58.7	61.4	
Disease extent				
≥4 cm	44	44		
≥3 cm	36		36	
≥2 cm	70		70	
<2 cm	163		163	
Size of dominating tumor mass				
≥15 mm	180	29	151	0.224
<15 mm	133	15	118	
T-classification				
T1	211	22	189	0.045
T2	36	8	28	
T3 or T4	0	0	0	
Local recurrence				
Yes	27	9	18	0.003
No	276	35	251	
Follow-up time (months)				
Median	120	120	120	0.108
Mean	106	95	108	
Min	3	14	3	
Max	120	120	120	
Grade of invasive lesion				
I	93	10	83	0.900
II	107	13	94	
III	45	6	39	
Missing	68	15	53	
Grade of in situ lesion				
I	109	16	93	0.005
II	99	8	91	
III	59	16	43	
Missing	46	4	42	
Radiotherapy				
Yes	206	28	178	0.736
No	93	14	79	
Missing	14	2	12	
Hormonal therapy				
Yes	52	11	41	0.101
No	196	24	172	
Missing	65	9	56	
Chemotherapy				
Yes	22	3	19	0.946
[illegible]	[illegible]	[illegible]	[illegible]	
Missing	65	9	56	

figure

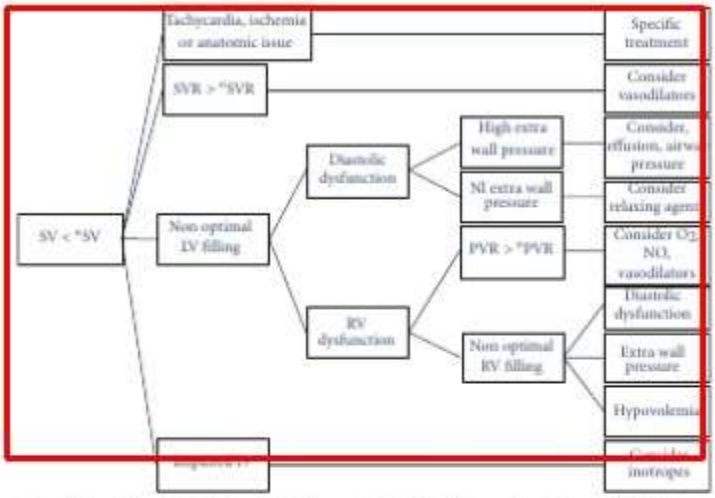


Figure 2: Subsequent algorithm dedicated to SV analysis. The needed SVR is that specific value of SVR allowing the generation of the minimum blood pressure (usually set at a mean value of 65 mmHg) with the needed CO. Similarly for needed PVR.

text

the bias introduced by the missing input data. There are theoretically three solutions. First, it is sometimes possible to find a surrogate of a given input variable assuming that the information provided is close. This is the case for example replacing SvO_2 by the central venous O_2 saturation ($ScVO_2$) or SaO_2 by the pulse oximetry (SpO_2) or total peripheral resistance instead of SVR when right aerial pressure is missing. Second, it is possible to give a fixed value to a given variable. For example, we may enter in the model directly at the level shown in Figure 2, assuming that the needed SV is the normal SV as a function of age and gender since the adaptative mechanisms of CO to needs are mostly resulting from a regulation of the heart rate. Third, it is possible to estimate a continuous variable from discontinuous clinical or paraclinical analysis. For example, VO_2 is often not monitored and VO_2 needs unknown. Consequently $VO_2/{}^*VO_2$ ratio <1 is suspected by the presence of clinical signs of shock or by an increase in blood lactate over time. Alternatively, a change in VO_2 needs (therefore translated to CO and EO_2 needs) can, be estimated by age, gender, body size, body temperature and/or empiric estimation of the change in metabolic needs due to underlying or associated pathologies. In any case, if we restrict our analysis to a part of the global model, we must assume that the blind part is fixed and given an approximate value. The reliability of the model is therefore

text

2.3. What Are the Specific Quality Criteria and Tolerance for Each Input Variable? The quality criteria and tolerance for the CO monitoring have been reviewed recently. [8, 9] The same effort must be done for each input variable of the model. Basically these criteria are very similar for all quantitative variables.

list

(1) The accuracy is how close the value is to a gold standard. It is estimated by the mean difference (bias) with the true value given by the gold standard.

(2) The linearity is the capability of maintaining constant the ratio between the physiologic signal and the electric output signal. Therefore the bias is constant. It can be verified by comparing the regression curve of the bias with the identity line.

(3) The precision is the ability to indicate the same value when the physiologic signal is stable. In other words, it is the variation due to random error in the signal processing. It can be estimated by the standard deviation/mean value when the physiologic signal is stable. The least minimum significant change (smallest change indicating a real change) is a direct consequence of precision.

(4) The resolution is the smallest change that the device can detect.

(5) The stability is the capability of maintaining the preceding quality criteria unchanged during time (without drift).

(6) The measuring range is the boundaries of value where the preceding quality criteria are found acceptable

(7) The responsiveness is the delay between a real change in the physiologic signal and a change greater than the least minimum significant change in the observed value. Coupled with the linearity, it determines the accuracy of the amplitude response.

Додаток Є – зведена таблиця результатів

Клас	YOLOv5, LR=0.01, SGD, Оригінальний підхід	YOLOv5, LR=0.01, SGD, Оптимізований підхід	YOLOv5, LR=0.001, Adam, Оптимізований підхід	Mask R-CNN N=12	Mask R-CNN N=24
(Усі класи)	**0.914**	0.911	0.911	0.880	0.886
Текст	0.840	0.853	**0.855**	0.794	0.772
Заголовок	0.865	0.870	0.874	0.885	**0.891**
Список	**0.927**	0.920	0.922	0.864	0.889
Таблиця	**0.974**	0.969	0.969	0.944	0.960
Фігура	**0.966**	0.943	0.936	0.914	0.915