УДК 004.93

П. С. Супрун
М. А. Очкуров

ОПТИЧНЕ РОЗПІЗНАВАННЯ СИМВОЛІВ ТА СТРУКТУРИЗАЦІЯ ТЕКСТОВИХ ДОКУМЕНТІВ

Вінницький національний технічний університет

Анотація
Запропоновано підхід по розпізнаванню символів та структуризації текстових документів із використанням згорткової нейронної мережі.
Ключові слова: розпізнавання символів, структуризація текстових документів, згорткова нейронна мережа.

Abstract
An approach to character recognition and structuring of text documents using a convolutional neural network is proposed.
Keywords: recognition of printed characters, text documents structuring, convolutional neutral network.

Вступ

Одним із напрямів розпізнавання образів є їх реалізація у системах оптичного розпізнавання текстів (Optical Character Recognition, OCR-системах). Система OCR реалізує автоматичне розпізнавання при допомозі спеціально розроблених програм зображень символів друкованого або ж рукописного тексту й переведення його в формат, який можна використати для подальшого оброблення редакторами текстів або текстовими процесорами [1]. Розгляду одного із підходів по виділенню та розпізнавання текстових документів присвячений даний матеріал.

Розпізнавання текстових символів

Для розпізнавання текстових символів на теперішній час сформувалася певна послідовність обробки сканованого текстового документу. Ця послідовність включає етапи виділення фрагменту тексту, попередню обробку цієї частини документу і потім приступають до самої важливої роботи по обробці документів — власне розпізнавання із використанням класифікаторів. У системах розпізнавання текстів використовуються такі класифікатори, як растровий, структурний і ознаковий [2]. Растровий класифікатор порівнює символ з набором еталонів, по черзі накладаючи зображення одне на одне. Ознаковий класифікатор висуває гіпотези, виходячи із ступеня схожості параметрів символу з еталонними значеннями. Структурний класифікатор проводить структурний аналіз символу, розкладаючи останній на елементарні складові (точки, лінії, дуги) і формуючи точну схему аналізованого знаку. Потім отримана схема у вигляді структурного опису символів порівнюється з еталоном. Цей класифікатор працює повільніше растрового і ознакового, зате відрізняється високою точністю.

На теперішній час вказані функції класифікаторів все частіше покладають на нейронні мережі, які залежно від їх структури і налаштування поєднують у собі характеристики цих класифікаторів і дозволяють здійснювати процес розпізнавання символів текстових документів [3]. Тому для створення програмного засобу по розпізнаванню текстових документів було вирішено використати згорткову нейронну мережу. Було вибрано згорткову нейронну мережу типу Fast R-CNN, яка дозволяє формувати бажану конфігурацію із трьох типів шарів: згорткового шару, шару підвибірки та вихідного повнозв’язного шару нейронної мережі, а також з механізму отримання та трансформації регіонів на зображенні.

Структура програмного засобу складається із ряду модулів, що послідовно із отриманого зображення текстового документа виділяють сторінки, потім у них виділяють текстовий фрагмент і виконують розпізнавання символів. Процес розпізнавання символів покладається на нейронну мережу. Попередньо нейронну мережу слід налаштувати на розпізнавання символів тексту [4]. Також у складі програмного засобу є ще модуль для навчання згорткової нейронної мережі на виконання процедури розпізнавання друкованих символів і модуль виведення результатів розпізнавання символів.

Розпізнавання текстових символів в даному підході відбувається за допомогою згорткової нейронної мережі, що згортає вихідне зображення розмірності $m_0 \times n_0 \times 3$ до необхідної $m \times n \times c$, де c — кількість каналів, за допомогою двох основних операцій — конволюції (2-dimensional Convolution) та пулінга (2-dimensional Max Pooling). Додаткові канали утворюються за допомогою накладання фільтрів двовимірною конволюцією на вхідну матрицю. Розмірність зображення зменшується через операцію пулінга, що обирає найбільше значення в певному заданому регіоні, зазвичай розміром 2×2, 3×3 або 4×4 та записує у нову матрицю. Пулінг вирішує дві основні проблеми: зменшує кількість ознак, відмовляючись від тих, що несуть мало інформації, і тим самим запобігає збільшенню кількості математичних операцій та збільшує щільність робочої матриці ознак, відмовляючись від більшості нульових значень та запобігаючи проблемі зникаючого градієнту.

Кожний вихідний канал несе інформацію про певний регіон, що дозволяє перейти від матриці зображення, яка описує кольорові значення, до матриці ознак. Дані ознаки є прихованими і несуть суттєву інформацію для математичної моделі нейронної мережі.

При розпізнаванні вхідного зображення воно розбивається на регіони різного розміру $h_x \times w_x \times c$. Дані регіони проходять певний ланцюжок трансформацій для отримання матриці ознак однакової розмірності $h \times w \times c$. Усі трансформовані регіони об'єднуються у матрицю $n \times h \times w \times c$, де n — кількість отриманих регіонів. Отримана матриця обробляється повноз'єднаним шаром (Dense Layer), вихідне значення якого — матриця $n \times d$, де d — кількість нейронів. Матриця $n \times d$ формує два вихідних значення — матрицю $n \times q$ з нормовано експоненційною активацією (softmax), де q — кількість можливих класів, та $n \times 4$, що є матрицею координат чотирикутників що описують об'єкт.

Структуризація розпізнаних символів

Структуризація тексту відбувається на основі середніх значень відстаней між розпізнаними символами. Якщо відстані певної кількості символів менші за деяке середнє значення, то можна припустити, що це є одним словом. Схожим чином визначаються пробіли та переноси у тексті, що дасть у вихідному результаті структурований текст розпізнаного документу.

Для вирішення задачі розпізнавання та структуризації символів текстових документів створена програмна реалізація запропонованого підходу з використанням мови програмування Python, фреймворку TensorFlow та бібліотеки OpenCV. TensorFlow надає повний набір інструментів для створення, тренування та використання моделі нейронної мережі. OpenCV дозволяє маніпулювати зображеннями різних форматів, змінюючи вже наявні пікселі або створюючи нові. Дані програмні засоби використані для реалізації розпізнавання та структуризації символів текстових документів.

Висновки

Запропонований підхід може бути використаний у комп'ютерних системах для виділення, розпізнавання і структуризації текстових документів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Оптичне розпізнавання символів [Електронний ресурс] — Режим доступу: https://ua.wikipedia.org/wiki/ Оптичне_розпізнавання_символів.
2. Жихаревич В. В. Аналіз методів розпізнавання символів тексту / В. В. Жихаревич, С. Е. Остапов, І. В. Миронів // Радіоелектронні і комп'ютерні системи. 2016, № 5. — С. 137 — 142.
3. Субботін С. О. Нейронні мережі : теорія та практика: навч. посіб. / С. О. Субботін. — Житомир : Вид. О. О. Євенок, 2020. — 184 с.
4. Тимченко О. В. Нейромережеві методи розпізнавання зображень текстів /О. В. Тимченко, Б. М. Гавриш, Б. В. Дурняк // Поліграфія і видавнича справа, 2021, № 1 (81). — С.72—88.

Супрун Павло Сергійович — студент групи 1КІ-22мс факультету інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: sirgenus47@gmail.com

Очкуров Микола Андрійович — старший викладач кафедри обчислювальної техніки, Вінницький національний технічний університет, м. Вінниця.

Suprun Pavlo S., student of group 1KI-22ms, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: sirgenus47@gmail.com

Ochkurov Mykola A. — Senior lecturer of the Computer Techniques Chair, Vinnytsia National Technical University, Vinnytsia.