УДК 004.93

О. П. Олефір
Т. Б. Мартинюк
М. А. Очкуров

ПРИКЛАДНЕ ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ РОЗПІЗНАВАННЯ СИМВОЛІВ СКАНОВАНИХ ТЕКСТОВИХ ДОКУМЕНТІВ

Вінницький національний технічний університет

Анотація
Запропоновано підхід до розпізнавання символів сканованих текстових документів із використанням попередньої обробки документа, нормалізації, сегментації та нейронної мережі.
Ключові слова: розпізнавання символів, сканований текстовий документ, нейронна мережа.

Abstract
An approach to character recognition in scanned text documents using document preprocessing, normalization, segmentation, and a neural network is proposed.
Keywords: character recognition, scanned text document, neural network.

Вступ

Сучасні системи оптичного розпізнавання текстових документів реалізують автоматичне розпізнавання зображень символів друкованого тексту за допомогою спеціально розроблених програм і переведення його у формат, який можна використати для подальшого оброблення текстовими редакторами або текстовими процесорами [1]. Завдання переведення паперових документів у електронну форму стає більш складним при виконанні операцій автоматизованого розпізнавання текстових документів із деякими дефектами та спотвореннями, яке на даний час вважається одним із найбільш складних і актуальних завдань розпізнавання даних. Виникає нагальна потреба у подальшому вдосконаленні підходів до етапів попередньої обробки зображень, сегментації тексту та формування ознак текстових символів та у впровадженні методів, що здатні адаптуватися до зміни умов подання зображення паперових документів. Розгляду одного з підходів до виділення та розпізнавання символів сканованих текстових документів присвячено цей матеріал.

Основна частина

Процес розпізнавання друкованих символів текстових документів являє собою систему дій, яка поєднує алгоритми цифрової обробки зображень, методи машинного навчання та сучасні підходи штучного інтелекту. У загальному вигляді процес розпізнавання друкованих символів сканованих документів включає декілька основних етапів: попередню обробку, нормалізацію зображення, сегментацію, виділення ознак, власне етап розпізнавання символів та корекцію отриманих результатів. Ці етапи є вже традиційними для систем розпізнавання текстових документів [2].

Програмний засіб розпізнавання символів можна умовно поділити на дві основні функціональні частини. Перша частина відповідає за роботу з вхідними даними: зчитування текстових документів, отримання зображень зі сторінок, застосування фільтрів для покращення якості зображення, а також перетворення графічної інформації у відповідну матрицю пікселів. Друга частина пов’язана безпосередньо з моделлю розпізнавання: ініціалізацією нейронної мережі, її навчанням, а також обробкою результатів, сформованих на виході системи. Для створення програмного засобу розпізнавання символів сканованих текстових документів було вирішено використати гібридну нейронну мережу шляхом поєднання згорткової та рекурентної нейронних мереж [3].

Архітектура програмного комплексу складається з модуля попередньої обробки зображення, модуля оптичного розпізнавання символів, модуля післяобробки отриманого тексту та модуля форматування текстового документа. На етапі попередньої обробки зображення масштабується, очищується від шумів, покращується за контрастом і переводиться у форму, зручну для аналізу. Після цього підготовлене зображення передається до модуля розпізнавання, де за допомогою нейромережевого алгоритму виконується перетворення графічної інформації у текстову. У модулі післяобробки здійснюється остаточне формування результату у вигляді придатного для подальшого використання текстового документа.

Після отримання результатів нейронної мережі виконується завершальний етап — структуризація розпізнаних символів і формування текстового документа. На цьому етапі окремі символи об'єднуються у слова, рядки та фрагменти тексту. Визначаються межі полів документа, розташування абзаців, порожні рядки, переноси слів та інші елементи форматування. Це дозволяє не лише отримати розпізнаний текст, а й наблизити його структуру до вигляду початкового документа.

Попередньо підготовлено вхідні дані, сформовано модель нейронної мережі та виконано її навчання. Для цього було використано набір навчальних даних, який містить основні алфавітні, цифрові та спеціальні символи. Такий набір дозволяє моделі навчитися розпізнавати різні типи символів, що можуть зустрічатися у сканованих текстових документах. Для ефективного використання архітектури LSTM спочатку було виконано налаштування базової нейронної моделі, яка виконує роль класифікатора. Після цього здійснено навчання основної архітектури з урахуванням попередньо підготовленої базової моделі.

Процес навчання нейронної мережі реалізовано за допомогою окремого модуля train. Він приймає як вхідний параметр попередньо створений об'єкт моделі та виконує основні дії, необхідні для організації тренування. На початку роботи функції в консоль виводяться задані гіперпараметри, що дозволяє контролювати поточні налаштування навчального процесу та переконатися у правильності вибраних параметрів. Ініціалізуються два об'єкти датасетів, перший відповідає за тренувальний набір даних, другий — за валідаційний, що є меншим за розміром та містить зразки, жодний з яких не є присутнім у тренувальному наборі.

Оцінювання якості роботи програмного засобу проводилось у два етапи. На першому етапі було перевірено точність відображення та розпізнавання символів у текстовому форматі, а також перевірялися обмежувальні рамки розпізнаних символів. Наступним протестованим модулем був модуль, який здійснює форматування текстового документа.

Висновки

Запропонований підхід до розпізнавання символів сканованих текстових документів поєднує етапи попередньої обробки зображення, нормалізації, сегментації, нейромережевого розпізнавання та післяобробки отриманого тексту. Використання такого підходу дозволяє підвищити якість розпізнавання символів у документах, що можуть містити шуми, спотворення, нерівномірне освітлення або інші дефекти сканування. Застосування нейронної мережі забезпечує адаптивність системи до різних типів символів, шрифтів і структур тексту, а модуль форматування дає змогу не лише отримати розпізнаний текст, а й частково відновити його початкову структуру. Отже, розроблений програмний засіб може бути використаний у комп'ютерних системах для автоматизованого виділення, розпізнавання та структуризації текстової інформації зі сканованих документів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Шелестов А. Ю. Методи та засоби оптичного розпізнавання текстової інформації / А. Ю. Шелестов // Вісник Національного технічного університету України «КПІ». — Серія «Інформатика, управління та обчислювальна техніка». — 2017. — № 65. — С. 88–94.
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.

Олефір Олександр Павлович — студент групи 2КІ-22б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: sashaolefir3008@gmail.com.

Мартинюк Тетяна Борисівна — д.т.н, професор кафедри обчислювальної техніки, Вінницький національний технічний університет, м. Вінниця

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

Olefir Oleksandr P. — student of group 2KI-22b, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: sashaolefir3008@gmail.com.

Martyniuk Tetiana B. — Dr. Sc. (Eng.), Professor at the Department of Computer Engineering, Vinnytsia National Technical University, Vinnytsia.

Ochkurov Mykola A. — Senior Lecturer at the Department of Computer Engineering, Vinnytsia National Technical University, Vinnytsia.