УДК 004.89

КЛАССИФИКАЦИЯ СКАНИРОВАННЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ СВЕРТОЧНОЙ НЕЙРОСЕТИ

[1,2]Котюжанский Л.А., [1]Четверкин Н.В., [1,2]Протасевич А.А., [1,2]Кочеров Р.В., [2]Рыжкова Н.Г.

[1]ООО «Нексус», Артемовский, e-mail: nexus077@gmail.com;
[2]ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», Екатеринбург, e-mail: nexus077@gmail.com

В настоящее время одной из актуальных задач автоматизации документооборота организации в условиях поступления разнообразной документации от большого количества контрагентов является проверка и классификация сканированных материалов. В статье представлен анализ и основные характеристики существующих способов решения данной задачи. Целью исследования является разработка программного модуля, позволяющего классифицировать документы с точностью не менее 97% в режиме реального времени, что актуально для электронного документооборота в крупных и средних компаниях. Приведено описание решения поставленной задачи на основе сверточной нейросети (CNN – Convolutional Neural Network). Входными данными для программного модуля является pdf-файл сканированного документа, выходными данными является xml-файл с классом документа. Для повышения точности и скорости работы программы были решены задачи по кодированию сигнала для нейронной сети и определению ее структуры. Приведено описание этапов обработки сканированных документов и архитектуры разработанной нейросети. Предложенный метод классификации позволяет классифицировать страницы с высокой точностью на небольшом датасете. Проведено тестирование программы на датасете из 9628 страниц и 22 возможных классов. Точность составила 99,1%. Время классификации одной страницы без учета чтения файла и копирования в GPU составляет 2 мс на GeForce 780TI. Полное время классификации страницы составляет примерно 22,3 мс.

Ключевые слова: автоматизация документооборота, интеллектуальный документооборот, классификация документов, сверточная нейросеть, распознавание изображений

CLASSIFICATION OF SCANNED DOCUMENTS USING A CONVOLUTIONAL NEURAL NETWORK

[1,2]Kotyuzhanskiy L.A., [1]Chetverkin N.V., [1,2]Protasevich A.A., [1,2]Kocherov R.V., [2]Ryzhkova N.G.

[1]LLC «Nexus», Artemovsky, e-mail: nexus077@gmail.com;
[2]Ural Federal University, Ekaterinburg, e-mail: nexus077@gmail.com

Currently, one of the urgent tasks of automating the organization's document flow in the context of receiving a variety of documentation from a large number of counterparties is the verification and classification of scanned materials. The article presents the analysis and main characteristics of the existing methods of solving this problem. The purpose of the study is to develop a software module that allows you to classify documents with an accuracy of not less than 97% in real time, which is relevant to the electronic document management in large and medium-sized companies. The description of the solution of the problem based on the convolutional neural Network (CNN – Convolutional Neural Network) is given. The input data for the program module is a pdf file of the scanned document, the output data is an xml file with the document class. To improve the accuracy and speed of the program, the tasks of encoding the signal for the neural network and determining its structure were solved. The stages of processing scanned documents and the architecture of the developed neural network are described. The proposed classification method allows you to classify pages with high accuracy on a small dataset. The program was tested on a dataset of 9628 pages and 22 possible classes. The accuracy was 99.1%. The classification time of a single page without considering file reading and copying to the GPU is 2 ms on the GeForce 780TI. The total page classification time is approximately 22.3 ms.

Keywords: document management automation, intelligent document management, document classification, convolutional neural network, image recognition

Автоматизация документооборота является важным этапом автоматизации производственных процессов предприятия. Одной из задач автоматизации документооборота является классификация документов, при которой программой производится выбор для документа одного из нескольких возможных классов. Наибольшая востребованность для делопроизводства организаций и сложность в обработке возникает для сканированных изображений, поэтому далее под документом будем понимать отдельную страницу.

Для автоматической классификации документов используются два основных метода: классификация страниц по шаблону и классификация с помощью сверточных нейронных сетей. Первый метод заключается в том, что необходимо для каждого вида документа описать шаблон страницы, то есть четко определить расположение текстовых полей и ключевых слов в этих

полях. Это требует кропотливого описания шаблонов, а также их модификации при изменении формата документа.

Ускорения процесса можно добиться с использованием машинного обучения. Обзор успешного использования нейронных сетей для решения различных прикладных задач приведен в [1]. В статье [2] представлен двухэтапный подход к обучению и тестированию в реальном времени классификации изображений документов, основанный на использовании компьютерного зрения, с конечной точностью 83,24%. Работа [3] посвящена повышению эффективности обучения классификаторов на основе областей и их объединения для классификации изображений документов, метод достигает точности в 92,21%. В [4] предложен подход, основанный на выделении, анализе и объединении текстового и визуального потоков для классификации изображений документов, в визуальном потоке используются глубокие CNN для извлечения структурных особенностей изображений, точность зависит от вида входных данных. В исследовании [5] предлагается двухпоточная нейронная архитектура для выполнения задачи классификации изображений документов, при этом используется подход совместного обучения признаков, объединяющий признаки изображения и текстовые части, подход совместного обучения имеет точность классификации до 97,05%. Преимуществом использования нейросетевого подхода является отказ от шаблонов. Вместо этого для создания обучающего датасета требуется указать лишь класс документа. Это позволит быстро разметить датасет и при необходимости внести оперативные изменения, что сделает систему гибкой. Обзор решений показывает актуальность задачи повышения точности распознавания при работе системы в режиме реального времени.

Основными требованиями к реализации программного модуля классификации документов является точность и скорость его работы. Цель работы состоит в определении подхода, включающего порядок обработки сканированных документов и архитектуру нейросети, позволяющего классифицировать документы с точностью не ниже 97% в режиме реального времени.

Входными данными для модуля является pdf-файл сканированного документа (примером документа является акт освидетельствования скрытых работ [6]), выходными данными является xml-файл с классом документа.

Классификация документов с использованием CNN

Исходя из вышеописанного, для классификации документов выбран подход с использованием нейронной сети. Для повышения точности и скорости работы программы были решены задачи по кодированию сигнала для нейронной сети и определению ее структуры.

Анализ факторов [7], влияющих на производительность CNN для обработки изображений документов, позволяет выделить применение сдвиговых преобразований во время обучения и использование больших входных изображений, которые приводят к наибольшему увеличению производительности, достигая показателей на RVL-CDIP с точностью 90,8%. Для увеличения скорости работы сети и повышения точности классификации были использованы «экспоненциальные линейные модули» – ELU [8], а также использованы слои DropOut[9].

Определен следующий порядок обработки сканированных документов.

1. Выполняется коррекция изображения по гистограмме яркости и по углу поворота. Эти меры должны существенно поднять качество работы Tesseract OCR алгоритма.

2. Выполняется распознавание текстовых полей и их значений. Результат – дескриптор страницы записывается в xml-файл.

3. Для каждого такого дескриптора создается специальная пара изображений (показаны на рис. 1), которые являются входным сигналом для нейросети классификатора и сжимаются до разрешения 128×128 пикселей.

Рис. 1. Пара изображений, сгенерированных для дескриптора

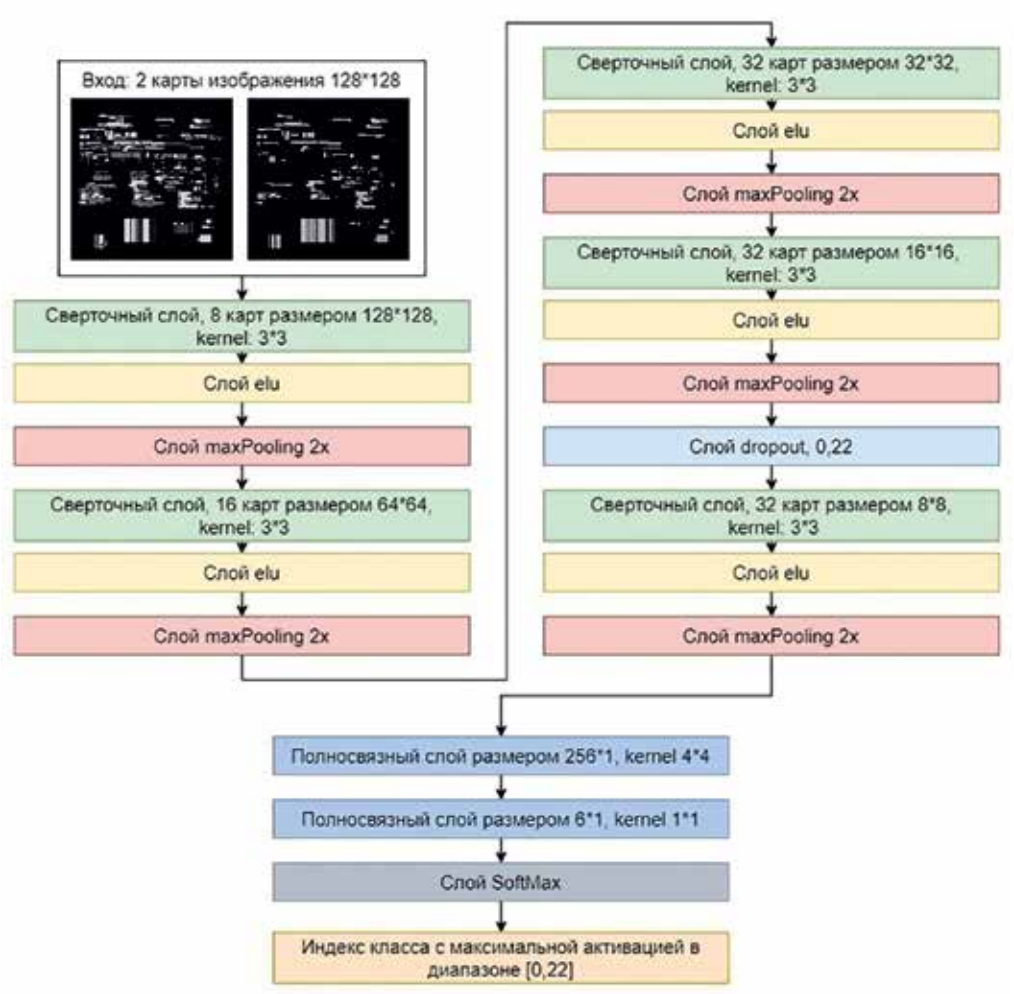

Рис. 2. Архитектура разработанной нейросети

4. Нейросеть классифицирует каждую из страниц-дескрипторов. Причем при обучении этой нейросети вводился специальный класс «неизвестная страница» – страница, которая не относится ни к одному из заданных в датасете типов документов. Нейросеть возвращает целое положительное число, которое является ID класса страницы. Архитектура разработанной нейросети представлена на рис. 2.

В работах по классификации страниц часто используется значительное уменьшение размера исходного изображения. Такой подход приводит к значительной потере информации, это особенно актуально при наличии мелкого текста. Для преодоления этой проблемы ряд авторов использовали подход формирования признаков из исходного изображения, используя для этого нейросеть и, например, скользящее окно. Признаки компактно описывают модель страницы без значительных потерь информации. Однако они используют дополнительные действия, что усложняет процесс разработки и обучения. Для повышения эффективности предлагается использовать подход по созданию изображений из полученного набора текстовых полей и символов, находящихся внутри них: яркость и координаты пикселя в дескрипторе определяются значением и положением символа на странице. Результат – компактное описание скана страницы практически без потери информации.

Результаты исследования и их обсуждение

Проведено тестирование программного модуля на сканированных документах, таких как акт освидетельствования скрытых работ, акт освидетельствования ответственных конструкций, акт о результатах проверки изделий, акты заключений неразрушающего контроля (радиографический метод). Объем датасета, сформированный из этих документов, составил 9628 страниц.

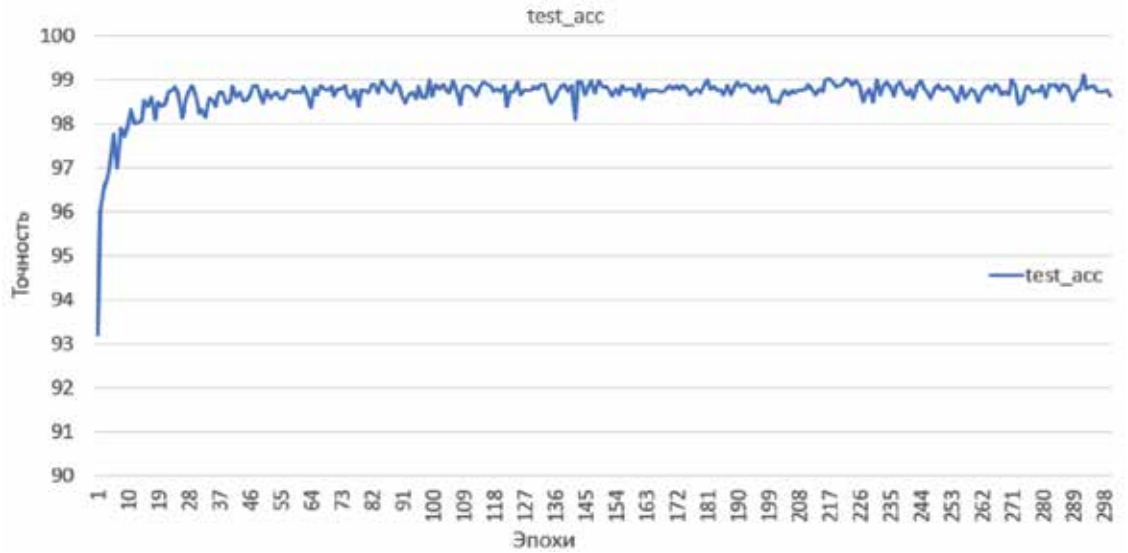

Рис. 3. Доля верных ответов на тестовом наборе в зависимости от эпохи

Доля обучающей выборки составила 7702 страницы, доля тестовой выборки – 1926 страниц. В результате обучения была достигнута точность (доля верных ответов) 99,1% на тестовой выборке (рис. 3). Время обучения составило около 18 с на одну эпоху на видеокарте Geforce GTX 780TI. Время классификации одной страницы без учета чтения файла и копирования в GPU составляет: 2 мс на GeForce 780TI. Время классификации одной страницы с учетом всех этапов обработки документа и записи типа страницы в xml-файл составляет примерно 22,3 мс, что соответствует обработке документов в режиме реального времени.

Заключение

Разработан порядок обработки сканированных документов, архитектура нейронной сети, выполнена программная реализация предложенного решения. Обучение и тестирование нейронной сети подтвердило увеличение точности распознавания по сравнению с представленными решениями. Оценка скорости работы программного модуля позволяет использовать систему в организациях в режиме реального времени. Дальнейшие пути развития системы видятся в следующих направлениях: оптимизация хранения данных, предварительная обработка входных данных для уменьшения количества читаемых файлов, использование базы данных.

Список литературы

1. Kotyuzhanskiy L.A., Ryzhkova N.G., Chetverkin N.V. Semantic segmentation in flaw detection. MIP: Engineering-2020. IOP Conf. Series: Materials Science and Engineering 862 (2020) 032056. P. 7. [Electronic resource]. URL: https://iopscience.iop.org/article/10.1088/1757-899X/862/3/032056 (date of access: 22.05.2021). DOI:10.1088/1757-899X/862/3/032056.

2. Kölsch A., M. Afzal M.Z., Ebbecke M., Liwicki M. Real-time document image classification using deep CNN and extreme learning machines. 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR Volume 1, 25 January 2018, P. 1318–1323. [Electronic resource]. URL: https://www.researchgate.net/publication/321124702_Real-Time_Document_Image_Classification_Using_Deep_CNN_and_Extreme_Learning_Machines (date of access: 22.05.2021). DOI: 10.1109/ICDAR.2017.217.

3. Das A., Roy S., Bhattacharya U., Parui S.K. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. 24th International Conference on Pattern Recognition, ICPR 2018. Proceedings – International Conference on Pattern RecognitionVolume 2018 – August, 26 November 2018, p. 8545630, P. 3180–3185. [Electronic resource]. URL: https://ieeexplore.ieee.org/document/8545630 (date of access: 22.05.2021). DOI: 10.1109/ICPR.2018.8545630.

4. Asim M.N., Khan M.U.G., Malik M.I., Razzaque K., Dengel A., Ahmed S. Two stream deep network for document image classification. 5th IAPR International Conference on Document Analysis and Recognition, ICDAR 2019. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR September 2019, p. 8978000, P. 1410–1416. [Electronic resource]. URL: https://ieeexplore.ieee.org/document/8978000 (date of access: 22.05.2021). DOI: 10.1109/ICDAR.2019.00227.

5. Bakkali S., Ming Z., Coustaty M., Rusinol M. Visual and textual deep feature fusion for document image classification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2020. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops Volume 2020-June, June 2020, P. 9150829, P. 2394–2403. [Electronic resource]. URL: https://ieeexplore.ieee.org/document/9150829 (date of access: 22.05.2021). DOI: 10.1109/CVPRW50498.2020.00289.

6. Приказ Федеральной службы по экологическому, технологическому и атомному надзору от 26 декабря 2006 г. № 1128 «Об утверждении и введении в действие Требований к составу и порядку ведения исполнительной документации при строительстве, реконструкции, капитальном ремонте объектов капитального строительства и требований,

предъявляемых к актам освидетельствования работ, конструкций, участков сетей инженерно-технического обеспечения» (с изменениями на 9 ноября 2017 года). Приложение 3. Актуальная форма акта освидетельствования скрытых работ [Электронный ресурс]. URL: http://www.consultant.ru/document/cons_doc_LAW_66762/ (дата обращения: 13.05.2021).

7. Kang L., Kumar J., Peng Y., Li Y. Convolutional neural networks for document image classification // 22nd International Conference on Pattern Recognition (ICPR), 2014. [Electronic resource]. URL: https://www.researchgate.net/publication/286725002_Convolutional_Neural_Networks_for_Document_Image_Classification (date of access: 22.05.2021). DOI:10.1109/ICPR.2014.546.

8. Clevert D., Unterthiner T., Hochreiter S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). ICLR 2016. [Electronic resource]. URL: https://arxiv.org/abs/1511.07289 (date of access: 22.05.2021).

9. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15, 2014, P. 1929–1958. [Electronic resource]. URL: http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf (date of access: 22.05.2021).