ABSTRACT

on the dissertation work on the topic:
"Development of an information model and recognition and machine learning algorithms for handwritten texts in the Kazakh language", PhD student in the specialty "6D075100 - Computer Science, Engineering and Management",
Yeleussinov Arman

Relevance of the work: Handwritten text recognition (HTR) is the key to many applications that depend on document digitization, including archival document processing, librarianship, healthcare, banking, etc. The biggest challenge arises from the wide variety of styles. Machine learning methods offer a potential solution to this problem. These methods are easily scalable to millions of documents, are faster than human transcription, and provide reproducible results. Many software applications already implement HTR, but they are not suitable for the Kazakh language due to the peculiarity of the language, and research in this area is still relevant.

The Archives of the President of the Republic of Kazakhstan is the custodian of unique documents on the modern political history of Kazakhstan. As of January 1, 2020, the Archives contain more than 830 thousand storage units for 1918-2008, of which 762 thousand storage units are paper-based, including handwritten ones. Digitization of archival documents is a technically complex process that depends on many factors: the availability of a special hardware and software complex, determination of a rational work strategy, qualifications of scanning operators and availability of the necessary regulatory and methodological documentation. Currently, instead of the term "digitization", a similar concept is sometimes used - scanning.

Digitization allows for safe and more economical storage, convenient search for the necessary information, as well as prompt access to archival documents. Digitization of archival documents allows you to solve several problems at once:

- bringing order to the fund accounting;
- ensuring the safety of originals by gradually withdrawing them from circulation;
- expanding and simplifying access to documents of the archival fund;
- providing access to the electronic fund;
- accelerating the process of providing public services;
- working with the scientific reference apparatus of the archive

To solve these problems, a comprehensive study and implementation of the latest technological and research results are required, including recognition of handwritten documents, where large sets of specially collected data will be required.

Objective of the study: Development of machine learning algorithms and neural network architecture for recognizing archival manuscripts in the Kazakh language

Research objectives: To achieve this objective, the following tasks are considered:

- Collection and processing of a database for training and testing machine learning methods for recognizing handwritten text in the Kazakh language.
- Research and analysis of systems, methods and algorithms for recognizing handwritten text;
- Development of algorithms and segmentation models for pre-processing archival documents;
- Development of a machine learning model based on generative adversarial neural networks (GAN) for artificially supplementing handwritten text data in the Kazakh language;
- Development of a machine learning model based on a transformer for text recognition;
- Development of a system for searching information in documents

Object of the study: Images of handwritten texts in the Kazakh language.

Subject of the research: Machine learning algorithms, neural networks and information models for handwriting recognition in Kazakh

Research methods: Computer vision algorithms, deep neural networks, convolutional networks, generative adversarial networks (GAN), transformer architecture

Scientific novelty of the work:

- A hybrid CNN+Transformer architecture adapted for handwriting recognition in Kazakh has been developed.
- A method based on the GAN architecture is proposed to supplement handwriting data in Kazakh by creating realistic handwriting variations. This method has shown that recognition accuracy can be improved even with a small amount of data.

Theoretical significance of the study: The study contributes to the development of machine learning methods and neural networks used for handwriting recognition. The development and adaptation of architectures (for example, CNN, Transformer) to the specifics of the Kazakh writing system expands the theoretical knowledge base in the field of intelligent processing of text information.

Unlike widely studied languages, the Kazakh language has an agglutinative structure, complex morphology and alphabetic features. The study allows us to theoretically substantiate and propose new methods for recognizing and processing texts in the Kazakh language.

Practical significance of the results: The development of an autonomous system for intelligent document processing is a promising direction in the development of information and telecommunication technologies in the Republic of Kazakhstan. The implementation of the results of this project will primarily increase the efficiency of the process of digitizing handwritten texts. Accordingly, the main socio-economic effect of the implementation of the results of the work is associated with saving time and manual labor for processing documents of various types and volumes.

Thesis provisions submitted for defense:

- The hybrid architecture of a convolutional neural network and a transformer for handwriting recognition allows to reduce the wer (word error rate) and cer (symbol error rate) indicators by 10.44% and 2.85%, respectively.

- The method of artificially augmenting handwritten text data based on GAN allows to increase the accuracy of word recognition from 75.8% to 86.3%.

The volume and structure of the work: The dissertation consists of an introduction, three sections, a conclusion, a list of references and appendices. The full volume of the dissertation is: 84 pages of typewritten text, including 54 figures, 6 tables, 83 references and 2 appendices.

The Introduction describes the problem, provides a brief overview of the subject area. The relevance of the dissertation is substantiated, and a review of literary sources is provided.

The First chapter considers the problem of handwriting recognition and general approaches to its solution. The methods for creating and processing datasets for handwriting recognition are described in detail. The article shows ways to improve image quality and increase recognition accuracy using pre-processing methods (filtering, histogram equalization, noise removal). This stage is considered an important preparatory step to improve the efficiency of text recognition.

The Second chapter discusses modern machine learning methods used for handwriting recognition. First, line and word segmentation methods, neural networks that implement them, and evaluation results are presented. Next, models based on CNN and Transformer architectures for letter classification and word recognition are described. In addition, the use of the GAN (Generative Adversarial Networks) method for increasing the data volume and its results are described.

The Third chapter describes the architecture, components, and technical implementation of the developed handwriting recognition system. The system components are optimized using Nvidia Triton Inference Server. In addition, a system for automatic search and extraction of necessary information from documents is considered, as well as the possibilities of integration and deployment of the finished system.

Approbation of dissertation work results: The validity and reliability of the study meet the requirements. The main provisions of the dissertation have been tested and are presented in 6 printed works, including:

- 2 articles in international scientific journals indexed in Scopus (Q2, Q3):

1. **Yeleussinov, A.**, Amirgaliyev, Y., Cherikbayeva, L. Improving OCR Accuracy for Kazakh Handwriting Recognition Using GAN Models. *Appl. Sci.* **2023**, *13*, 5677 (SJR 0.492, процентиль 75) https://doi.org/10.3390/app13095677

2. Amirgaliyev B., **Yeleussinov A.**, Taizo M. Kazakh handwritten recognition. Journal of Theoretical and Applied Information Technologies, 2020, 98(14), pp. 2744–2754.

- 1 article in a journal from the list recommended by the Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan:

1. Амиргалиев Е. Н., Миачи Т., **Елеусинов А.И.** Методы и алгоритмы формирования базы и обработки рукописных текстов на казахском языке, Вестник КазНИТУ, 2020г., №138, сс. 653-660.

- 3 publications in the materials of international and national conferences and reports at scientific seminars:

1. **Yeleussinov A.**, Islamgozhayev T., Kozhirbayev Z. Creation and Annotation of a Handwritten Text Database for the Kazakh Language: Methodologies and Preliminary Results. In2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE) 2024 Oct 30 (pp. 1-6). IEEE.

2. **Елеусинов А.И.**, Бурибаев Ж.А. Сегментация рукописных символов с исправлением наклона текста. V-International Conference "Informatics and Applied Mathematics", Almaty, IIVT, 09.29.2020-10.01.2020.

3. **Елеусинов А.И.,** Амиргалиев Е. Н., Мажитов Ш. С. О разработке базы изображений рукописных символов казахского языка. IV International Scientific and Practical Conference "Informatics and Applied Mathematics" dedicated to the 70th anniversary of professors T.N. Biyarov, Valdemar Vujcik

and the 60th anniversary of professor E.N. Amirgaliev, Institute of Information and Computing Technologies, September 25-29, 2019 year.

The dissertation was also discussed at scientific seminars of the department.