ҒТАХР 20.19.27

10.51889/2959-5894.2023.84.4.018

Н.А. Тойганбаева[1], А.Н. Алимова[2], М.Ж. Сакыпбекова[1], Ф.Р. Гусманова[1], Ғ.С. Әбдіманап[2]*

[1]Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы қ., Қазақстан
[2]ҚазМұнайГаз Инжиниринг Жауапкершілігі шектеулі серіктестік, Астана қ., Қазақстан
[*]e-mail: bodinaz@mail.ru

НЕЙРОНДЫ ЖЕЛІЛЕР НЕГІЗІНДЕ ҚАЗАҚ-ОРЫС ТІЛДЕРІНДЕГІ ҚОЛЖАЗБА МӘТІНДЕРДІ ТАНУ

Аңдатпа

Мақалада кириллица графикасына негізделген қазақ және орыс тіліндегі қолжазбаны тануға рекурентті нейронды желілерді қолдану қарастырылды. Қолжазба мәтіндерін тану мәтін жазылған қағазды сканерлеуден, алынған ақпаратты мәтіндерге және бейнелерге бөлуден, қолжазбаларды интеллектуалды тану үшін әдістерді қолданудан және нәтижелерді өңдеуден тұратын күрделі құрылымды үдеріс. Бұл ғылыми жұмыста A. Abdallah ұсынған толық жабық конволюциялық нейрондық желілер негізінде құрылған жаңа моделді пайдаланылып, қазақ және орыс қолжазба мәтінін тану мәселесі шешіліп, нәтижелерге талдау жасалды. Жұмыста Gated-CNN-BGRU(CCN, convolutional neural networks, bidirectional gated recurrent unit, конволюциялық нейрондық желілер-екі бағытты басқарылатын блок) архитектурасына негізделген модель сипатталып, қолжазба тану бойынша таңбалар қатесінің жиілігі, сөздер қатесінің жиілігі және сөйлемдер қатесінің жиілігі есептелді. Қолжазбаны тану жүйелерін оқыту және сынау үшін қазақ және орыс тіліндегі HKR (Handwritten Kazakh & Russian) және KOHTD (Kazakh Offline Handwritten Text Dataset) қолжазба деректер жиыны алынды. Ұсынылған модель Python үшін TensorFlow кітапханасын қолдана отырып жүзеге асырылды.

Түйін сөздер: қолжазбаны тану, нейронды желілер, TensorFlow, деректер жинағы, терең оқыту.

Н.А. Тойганбаева[1], А.Н. Алимова[2], М.Ж. Сакыпбекова[1], Ф.Р. Гусманова[1], Ғ.С. Әбдіманап[2]
[1]Казахский Национальный университет имени Аль-Фараби, г. Алматы, Казакстан
[2]Товарищество с ограниченной ответственностью КазМунайГаз Инжиниринг, г. Астана, Казакстан

РАСПОЗНАВАНИЕ РУКОПИСНЫХ ТЕКСТОВ НА КАЗАХСКО-РУССКОМ ЯЗЫКЕ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Аннотация

В статье рассмотрено использование рекурентных нейронных сетей для распознавания рукописного ввода на казахском и русском языках на основе кириллической графики. Распознавание рукописных текстов представляет собой сложный структурированный процесс, состоящий из сканирования бумаги с текстом, разделения полученной информации на тексты и изображения, использования методов интеллектуального распознавания рукописей и обработки результатов. В этой научной работе была решена проблема распознавания казахского и русского рукописного текста с использованием новой модели глубокой нейронной сети на основе полностью закрытого CNN, предложенного Abdallah, и проведен анализ результатов. В работе была описана модель, основанная на архитектуре Gated-CNN-BGRU(CCN, convolutional neural networks, Bidirectional gated recurrent unit, сверточные нейронные сети-двунаправленный управляемый блок), и рассчитана частота ошибок символов, частота ошибок слов и частота ошибок предложений по распознаванию рукописных текстов. Для обучения и тестирования систем распознавания рукописей использованы наборы рукописных данных HKR (Handwritten Kazakh & Russian) и KOHTD (Kazakh Offline Handwritten Text Dataset) на казахском и русском языках. Предложенная модель была реализована с использованием библиотеки TensorFlow для Python.

Ключевые слова: распознавание почерка, нейронные сети, TensorFlow, набор данных, глубокое обучение.

N.A. Toiganbaeva[1], A.N. Alimova[2], M.Zh. Sakypbekova[2], F.R. Gusmanova[1], G.S Abdimanap[2]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]KazMunayGas Engineering Limited Liability Company, Astana, Kazakhstan

RECOGNITION OF HANDWRITTEN TEXTS IN KAZAKH-RUSSIAN BASED ON NEURAL NETWORKS

Abstract

The article considers the use of recurrent neural networks for handwriting recognition in Kazakh and Russian languages based on Cyrillic graphics. Handwritten text recognition is a complex structured process consisting of scanning paper with text, dividing the received information into texts and images, using methods of intelligent recognition of manuscripts and processing the results. In this scientific work, the problem of recognition of Kazakh and Russian handwritten text was solved using a new deep neural network model based on a fully closed CNN proposed by Abdallah, and the results were analyzed. The paper describes a model based on the Gated-CNN-BGRU architecture (CCN, convolutional neural networks, Bidirectional gated recurrent unit), and calculates the error rate of characters, the error rate of words and the error rate of sentences for recognizing handwritten texts. Russian and Kazakh handwritten data sets HKR (Handwritten Kazakh & Russian) and KOHTD (Kazakh Offline Handwritten Text Dataset) in Kazakh and Russian were used for training and testing of manuscript recognition systems. The proposed model was implemented using the TensorFlow library for Python.

Keywords: handwriting recognition, neural networks, TensorFlow, dataset, deep learning.

Кіріспе

Цифрландыру заман таман талабы болғандықтан бизнес үрдістердегі құжатайнылм сандық форматқа көшіп жатыр. Алайда, шот-фактуралар, салықтар, жадынамалар мен сауалнамалар, тарихи деректер және емтихан сұрақтарына жауаптар сияқты көптеген құжаттарда қолжазба енгізу қажеттілігі әлі де бар. Осыған байланысты қолжазба мәтінді тану қажеттілігі туындайды. Қолжазба мәтінді тану (ағылш. Handwritten Text Recognition, HTR) - компьютер арқылы қолжазба мәтіндерді сандық форматқа автоматты аудару әдісі. Кириллица графикасына негізделген қазақ және орыс тіліндегі қолжазбаны тану бойынша зерттеу жұмыстары аз, сондықтан да зерттеу жұмыстарын жүргізу қажеттілігі туындайды.

Қолжазбаны тану тәсілдерін жасырын Марков модельдерге (HMMs, Hidden Markov models) және рекурентті нейронды желілерге (RNNs – Recurrent neural network) негізделген әдістер деп екі санатқа бөлуге болады. Жасырын Марков модельдері – бұл машиналық оқыту мен сигналдарды өңдеуде қолданылатын қуатты ықтималдық моделі. Мәтінді тану үшін жасырын Марков модельдеріне негізделген тәсілдердің бірқатар артықшылықтары бар. Бұл модельдер шуға төзімді және емледегі өзгерістерді тасымалдай алады және осы модельдерін оқытудың автоматтандырылған алгоритмдері бар, ал HMM құралдары еркін қол жетімді. Қолжазба мәтіндерін сегменттеу қателіктерге бейім және уақытты қажет етеді, жасырын Марков модельдеріне бұл қажет емес. Жасырын Марков модельдерін пайдаланып қолжазбаны тану мәселелері [1], [2], [3] ғылыми жұмыстарында қарастырылып, талқыланған. AlKhateeb [4] жасырын Марков модельдерін (HMM) қолдана отырып, сөзге негізделген оффлайн тану жүйесі ұсынылады. Қолжазбаны тану әдісі алдын ала өңдеуді, белгілерді алуды және жіктеуді қамтитын үш кезеңнен тұрады. Бірінші қадамда кіріс сценарийлерінен сөздерді сегменттеу және қалыпқа келтіру іске асырылады. Содан кейін, сөздің әр айна бейнесінде қозғалатын жылжымалы терезені пайдаланып, әр бөлінген сөзден қарқындылық белгілерінің жиынтығы жиналады. Сонымен қатар, ішкі сөздер мен диакритиктердің саны сияқты құрылымдық ақпарат алынады. Ақырында, бұл сипаттамалар жіктеу схемасына біріктіріледі. Intensity функциялары HMM классификаторын оқыту үшін пайдаланылып, нәтижелер жоғары тану жылдамдығы үшін құрылымдық функцияларды қолдана отырып қайта бағаланады. Зерттеу жұмысында 32492 қолмен жазылған араб сөздерін қамтитын IFN / ENIT дерекқорын қолдана отырып, ауқымды сынақтар жүргізілген.

Рекурентті нейрондық желілер – элементтер арасындағы байланыстар бағытталған тізбекті құрайтын нейрондық желілердің бір түрі. Осы бағытталған тізбектер арқасында уақыт бойынша оқиғалар сериясын немесе дәйекті кеңістіктік тізбектерді өңдеу мүмкіндігі пайда болады. Рекурентті желілер өздерінің ішкі жадын еркін ұзындықтағы тізбектерді өңдеу үшін пайдалана алады. Сондықтан рекурентті нейронды желілері қолжазба мәтінін тану немесе сөйлеуді тану есептерінде сәтті қолданысқа ие. Басқарылатын рекурентті блок (Gated recurrent unit, GRU) [5] және ұзақ қысқа мерзімді жады (Long short-term memory, LSTM) [6] сияқты рекурентті нейронды желілер сөйлеуді тану, машинадық аударма жасау, бейнелерді тану есептерін шешуде керемет нәтижелер көрсетті. Қолжазба мәтін жазылған кескіндегі мәтіндерді тану үшін екі өлшемді кескінді векторға түрлендіріп, кодер мен декодерге жіберу керек. Осы мәселені шешу үшін GRU және LSTM рекурентті желілері көптеген көздерден алынған ақпарат пен мүмкіндіктерді біріктіріп, қолжазба тізбегін алады. Кіріс функциясы байланысты уақытша қолдайтын классификациясы (Connectionist Temporal Classification, CTC) модельдерін қолдануға байланысты кескіндегі мәтіндерді сегменттеу қажет емес [7]. Ақырында деректерді шығыс деректерімен сәйкестендіру іске асырылады.

Ingle R.R. [8] зерттеу жұмысында қолжазба мәтіндерін тану жүйесін құруда кездесетін деректер, тиімділік және интеграция мәселелерін қарастырған. Зерттеуші ауқымды онлайн қолжазба деректерін пайдалануды және қолжазбаларды тануда қайталанатын байланыссыз нейрондық желілерге негізделген сызықты тану моделін ұсынады. Модель LSTM негізіндегі модельдермен салыстырылатын дәлдікке қол жеткізеді, сонымен бірге оқыту мен логикалық қорытындыда жақсы параллелизмді қамтамасыз етеді. Бұл компоненттер қолжазбаны танудың көлемді мәтіндерін тану жүйесіне біріктіру шешімін құрайды.

Espana-Boquera [9] шексіз, дербес қолжазба мәтіндерін тану үшін жасырын Марков моделінің (HMM) және жасанды нейрондық желінің (ANN) гибридті модельдерін ұсынады. Марков тізбектері оптикалық модельдердің құрылымдық элементтерін сипаттау үшін пайдаланылды, ал сәулелену ықтималдығын бағалау үшін көп қабатты перцептрон қолданылды. Бақыланатын оқыту тәсілдерінің арқасында бұл жұмыс қолжазба мәтінінен көлбеуді алып тастаудың және мәтіндік кескіндердің өлшемін қалыпқа келтірудің жаңа стратегияларын ұсынады. Көлбеуді түзету және өлшемді қалыпқа келтіру мәтіндік контурлардың жергілікті экстремумдарын жіктеу үшін көп қабатты перцептрондарды қолдану арқылы жүзеге асырылады. Жасанды нейрондық желілер көлбеуді біркелкі емес жолмен азайту үшін де қолданылады. Эксперименттер IAM дерекқорынан қолмен жазылған мәтіннің автономды жолдарын қолдану арқылы жүргізілді және қол жеткізілген тану көрсеткіштері жарияланған нәтижелермен салыстырғанда бірдей жұмыс үшін ең жақсы көрсеткіштердің бірі болды.

Ф. Абдурахман [10] конвульсиялық қайталанатын нейрондық желілерге негізделген амхар тілінде (Эфиопия федералды үкіметінің тілі) автономды қолжазба сөздерді тану жүйесін ұсынады. Зерттеу жұмысында қолмен жазылған амхар сөздерін тану үшін нейрондық желілер қолданылған. Конволюциялық нейрондық желілер (CNNs, convolutional neural networks) сөздердің кіріс кескіндерінен белгілерді алу үшін, қайталанатын нейрондық желілер (RNNs) тізбекті кодтау үшін және байланысты уақытша қолдайтын классификациясы (CTC) жоғалту функциясы ретінде ұсынылған. Зерттеушілер HARD-I қолмен жазылған амхар сөздерінің деректер жинағын жасады, ең тиімді тану моделі WER 5,24% және CER 1,15% көрсетті. Ұсынылған модельдер амхар тіліндегі ресми қолжазба сөздерді тану үшін қолданыстағы модельдермен салыстырғанда бәсекеге қабілетті өнімділікті қамтамасыз етеді.

J.C. Aradillas бастаған зертеушілер тобы [11] тарихи құжаттардағы қолжазба мәтіндерді оффлайн тану мәселесін қарастырып, негізгі үш мәселені шешті. Біріншіден, тану моделінің қай деңгейлерін дәл баптауды қажет ететінін талдай отырып, массивті дерекқордан кішірек тарихи дерекқорға тасымалдауды оқытуды (transfer learning, TL) жүзеге асырды. Екіншіден, біз тасымалдауды оқытуды тиімді біріктіру және деректер көлемін (data augmentation, DA) көбейту әдістерін талдады. Соңында, оқу жинағындағы қате таңбалаудың салдарын азайту

алгоритмін ұсынды. Қолжазбаны тану ICFHR 2018 competition database, Washington және Parzival деректер жиыны негізінде талданды және біз сынақтар нәтижесі таңбалардағы қателер жиілігі айтарлықтай төмендеген (кейбір жағдайларда 6 пайыздық тармаққа дейін).

T.Ngo [12] жапон және қытай тілдерінде оффлайн режимде қолмен жазылған мәтін жолдарының кескіндерін анықтау үшін рекурентті нейронды желінің түрлендіргіш моделін ұсынады. Бұл ұсынылған модель үш негізгі компоненттен тұрады:

- CNN көмегімен кіріс кескінінен визуалды белгілерді шығаратын, содан кейін BLSTM көмегімен визуалды белгілерді кодтайтын визуалды белгілер кодтаушысы;
- кірістірілген қабаттар мен LSTM көмегімен кіріс кескінінен лингвистикалық белгілерді шығаратын және кодтайтын лингвистикалық контекстті кодтаушы;
- толық қосылған softmax қабаттары мен қабаттары арқылы визуалды және лингвистикалық ерекшеліктерді соңғы тегтер тізбегіне біріктіретін, содан кейін декодтайтын бірлескен декодер.

Ұсынылған модель кіріс кескінінің визуалды және лингвистикалық ақпаратын пайдаланады. Модельдің өнімділігі Kuzushiji және SCUT-EPT деректер жиынтығында бағаланды.

Екі бағытты LSTM (Blstm) - бұл негізінен табиғи тілді өңдеу үшін қолданылатын қайталанатын нейрондық желі. Стандартты LSTM-ден айырмашылығы, кіріс екі бағытта да келеді және ол ақпаратты екі жағынан да қолдана алады. Бұл сонымен қатар тізбектің екі бағытындағы сөздер мен сөз тіркестері арасындағы дәйекті тәуелділіктерді модельдеудің қуатты құралы.

H. M. Balaha [13] оқыту, тестілеу және валидация үшін араб қолжазба таңбаларының HMBD (handwritten characters' dataset, HMBD) үлкен және күрделі деректер жиынтығын жинау, дайындау, тазарту және алдын-ала өңдеуді талқылайды. Ғылыми жұмыста Араб қолжазба таңбаларының тану үшін HMB1 және HMB2 деп аталатын конволюциялық нейрондық желі архитектурасынан (CNN) бар терең оқыту жүйесінен (deep learning DL) тұратын жүйені ұсынады. Сипатталған жүйенің архитектурасының дәлдік көрсеткіштері жоғары және бұрын жарияланған мәліметтер жиынтығы негізінде жалпылауға болады.

Зерттеудің мақсаты мен міндеттері

Зерттеудің мақсаты - рекурентті нейронды желілер негізінде қазақ-орыс тілдеріндегі қолжазбаны тануға бағдарламалық кешенді құру. Қойылған мақсатқа қол жеткізу үшін келесі міндеттерді жүзеге асыру керек:

1. Мәтінді оптикалық тану есептеріне рекурентті нейронды желіні қолданған шешімдерге шолу және талдау жасау,
2. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер жинағы негізінде қазақ-орыс тілінде қолжазба мәтінді тануды жүзеге асыру .
4. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер қоры негізінде рекурентті нейронды желі әдісімен эксперименттер жүргізу.

Нейрондық желілерді пайдалануға негізделген қазақ-орыс тілінің қолжазбасын тану мәселесі [14, 15] ғылыми жұмыстарда қарастырылған. Кириллица графикасына қатысты қазақ және орыс тілдеріндегі қолжазбаларды тану есептері шешіліп, нәтижелері жарияланды [16, 17]. Бұл жұмыстарда орыс тіліндегі қолжазба мәтіндерді тануға басымдылық берілген. Қазақ тіліндегі қолжазба мәтінін тану әлі де толық зерттелмеген күйінде қалып отыр. Осыған байланысты қазақ тілінің қолжазба мәтінін танудың жаңа тиімді алгоритмдерін әзірлеу және зерттеу өзекті болып отыр. Осыған орай, қазақ тіліндегі қолжазбаларды тануға назар аударып, қазақ-орыс тілінің қолжазба мәтінін тану мәселесін шешуге, нейрондық желілерді қолдануға негізделген тәсіл осы мақалада қарастырылады. Қазақ-орыс тілінің қолжазба мәтінін танудың негізгі кезеңі мынадай кезеңдерден тұрады:

1. Кескінді алдын-ала өңдеу (алдын-ала өңдеу) қолжазба мәтінін тану: бұл кезеңде кескіннің сапасын жақсарту және оны сегменттеуге ыңғайлы түрге келтіру мақсатында өңдеу жүреді.

Алдын ала өңдеу кезеңінде қолжазба мәтіні сканерленеді. Қолжазба мәтін жазылған қағаз құжаттар цифрлық графикалық көскінге айналады.

2. Сканерленген қолжазба мәтінін сөздерге сегменттеу. Бұл кезеңде сканерленген қолжазба мәтіні талдауға ыңғайлы бөліктерге бөлінеді немесе сегменттеледі. Бұл кезеңдегі негізгі әрекеттер – мәтінді жеке жолдарға бөлу (жолдарды сегментациялау) және жолдарды сөздерге бөлу (сөздерді сегментациялау), мұнда бос орын олардың бөлгіші болып табылады. Ол үшін шуды жою және сөздердің шекараларын анықтау үшін мәтінге сүзгілер дәйекті түрде қойылады.

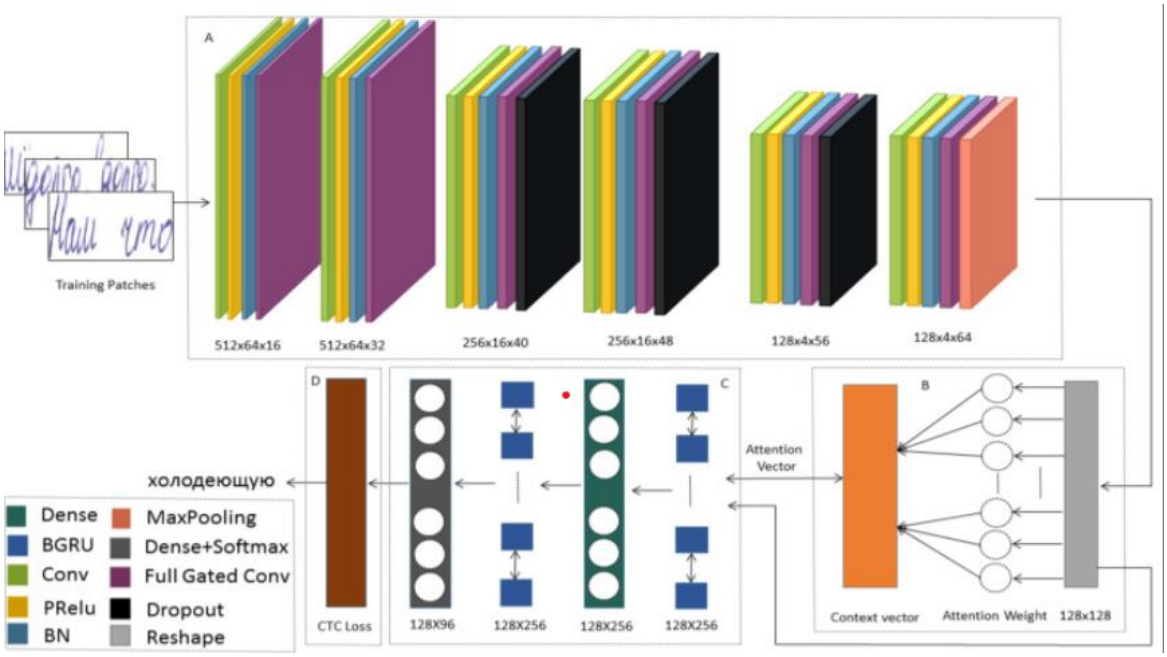
3. Тану үшін қолмен жазылған сканерленген және бөлінген сөздердің дерекқорларын жасау, толтыру.

4. Нейрондық желілердің көмегімен қазақ-орыс тілінің қолжазба мәтінін тану үдерісін зерттеу.

5. Қазақ-орыс тілінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу

8. Ұсынылған алгоритмдердің тиімділігін бағалау үшін қазақ-орыс тілінің қолжазба мәтінін танудың эксперименттік жүйесін әзірлеу.

Бұл мақалада A. Abdallah ұсынған [18] толық жабық CNN негізінде терең нейрондық желінің жаңа моделін пайдаланып, қазақ және орыс қолжазба мәтінін тану есебі қарастырылады. Жабық бағытталған CNN-BGRU (convolutional neural networks- bidirectional gated recurrent unit, конволюциялық нейрондық желілер-екі бағытты басқарылатын блок) архитектурасы сипаты келесі суретте беріледі (Сурет 1).



Сурет 1. Қолжазбаны тануға арналған жабық бағытталған CNN-BGRU жүйесі

Жүйе төрт негізгі бөліктен тұрады:

- (A) кодтаушы;
- (B) назар аудару блогы;
- (C) декодер;
- (D) Байланысты орнатуды уақытша қолдайтын классификация (CTC).

Кодтаушы көмегімен қолжазба жазылған кескіндер тұрақ белгілер векторлық қатарға түрлендіріледі. Кодтаушы желісі суреттерден тиісті белгілерді алуға үйретуге сәйкес келетін 6 конволюциялық блоктан тұрады. Әрбір блок бірінші, екінші, төртінші және алтыншы блоктарда (3, 3) және үшінші және бесінші блоктарда (2, 4) өлшемді сүзгі ядросын қолданатын

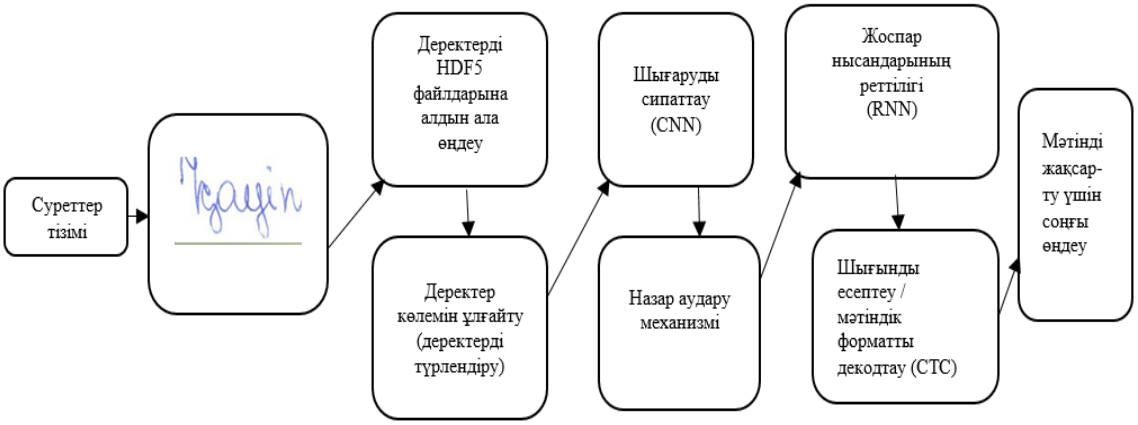
конволюциялық операциядан тұрады. Параметрлік түзетілген сызықтық блок (ReLU) және пакеттік қалыпқа келтіру қолданылады. Қайта оқытуды азайту үшін біз кейбір конволюциялық қабаттарда скринингті қолданамыз (скрининг ықтималдығы 0,2).

Назар аударушы блок – декодер контекстік векторды құруды жеңілдетуді пайдаланып, бастапқы тізбектердің кеңейтілген кодтауын жүзеге асыратын механизм. Декодер таңбалар тізбегін болжау үшін белгілер тізбегін өңдейді. Gate басқару элементтерінің идеясы объектілер векторын келесі қабатқа тарату болып табылады. Gate қабаты берілген позициядағы векторлық объектінің мәнін және іргелес мәндерді қарастырады және оны сол күйде ұстау немесе тастау керектігін анықтайды.

Шығыс деңгейіндегі байланысты орнатуды уақытша қолдайтын классификациясы (CTC – Connectionist Temporal Classification) тізбектерді таңбалау тапсырмаларына рекурентті нейронды желілерді қолданады.

Gated-CNN-BGRU архитектурасына негізделген модельді пайдаланып, кириллица негіздегі қазақ-орыс тілдеріндегі мәтіндер танылды. Алгоритм алты кезеңнен тұрады:

1. Алдын ала өңдеу
2. CNN қабаттары арқылы сипаттамаларды алу
3. Назар аударуға және шығыс функциясымен байланыстыру
4. Жоспар бойынша RNN реттілігі
5. Шығынды есептеу / мәтіндік форматты декодтау (CTC)
6. Соңғы мәтінді жақсарту үшін кейінгі өңдеу.



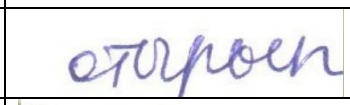
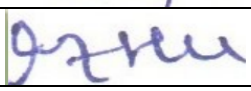
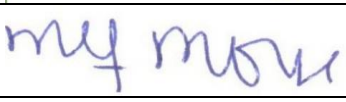
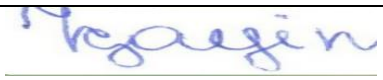
Сурет 2. Gated-CNN-BGRU архитектурасы

Кириллица графикасына негізделген қазақ және орыс тіліндегі HKR [16] және KOHTD [19] қолжазба деректер жиыны құрылып, әртүрлі зерттау жұмыстары жүргізіліп, нәтижелер алына бастады. HKR (Handwritten Kazakh & Russian) – 95% орыс және 5% қазақ сөздері/сөйлемдерінен тұратын қолжазба деректерінің жиыны. Деректер жиыны 1500-ден астам толтырылған нысандардан тұрады. HKR жиынында шамамен 63000 сөйлемді 200 түрлі авторлардың қолжазбасы енгізілген және 715699-нан астам таңба. HKR жиына арқылы зерттеушілер терең және машиналық оқыту арқылы қолжазбаны тану есептерін шеше алады. Kazakh Offline Handwritten Text Dataset (KOHTD) – қазақ тіліндегі алғашқы оффлайн қолжазба мәтіндерінің үлкен деректер жиыны. Бұл деректер жинағын құру үшін студенттердің жазбаша емтихан жұмыстары сканерленіп, генетикалық алгоритм көмегімен сегментацияланды. Деректер жиынында шамамен 922010 таңба және 140335 сегменттелген кескін жинақталды.

Қолжазбаны тану есебінің күрделілігі қолжазбаның, пішіндердің, әріптердің өлшемдерінің және әртүрлі тілдердің алуан түрлілігіне байланысты. Сондай-ақ, қолжазба мәтін бар қағазда "шу" болуы мүмкін, қағаздағы ақаулар, сыртқы дақтар – бұл қолайсыздықтар бүкіл үрдісті қиындатады. Қолжазба мәтіндегі әр сөздің каллиграфиялық түрде шығарылған әріптерінен

бастап, белгілі бір әріпті жазу стандарты болғанына қарамастан, әр адамның өз қолжазбасы бар. Әртүрлі авторлардың қолжазба мәтіндерін тану пайызы келесі кестеде көрсетілген (Кесте 1).

Кесте 1. Әртүрлі авторлардың қолжазба мәтіндерін тану пайызы

№	Кіріс	Ерікті пайдаланушының енгізген сөзі	Енгізілген сөз бен кіріс кескінде жазылған сөбен сәйкестік
1		отырып	100%
2		яғни	75%
3		адаш	67%
4		әдістеріне	84%
5		мөлшерде	100%
6		тұтын	83%
7		ипотекалық	90%
8		арқылы	83%
9		қауіп	60%

Кестеден енгізілген сөз бен кіріс кескінде жазылған сөздің сәйкестік пайызын анықтауда адамның қолжазбасы маңызды екенін көруге болады. Қазақ тілінде қолжазба мәтіндерде *ң* және *қ* (Кесте 1, 7 жол), *м* және *ш* (Кесте 1, 3 жол), *н* және *и* (Кесте 1, 6 жол), *л* және *е* (Кесте 1, 8 жол) жазылуы ұқсас болуы мүмкін. Бұл тану ықтималдығын азайтады. Қолжазбаны қабылдау мен тануға мәтіндегі бос орындардың болуы да ықпал етеді. Бос орын мәтінді басынан аяғына дейін дәйекті түрде оқуға кері ықпалын тигізеді. Қолжазба деректер жиынын құруға арналған құралда логикалық көрсеткіштерге негізделіп жасалған.

Зерттеу нәтижелері мен талқылау

Бұл бөлімде HKR және KOHTD екі түрлі деректер жиынтығында A. Abdallah ұсынған толық жабық CNN негізінде терең нейрондық желінің жаңа моделін пайдаланып, қазақ және орыс қолжазба мәтінін тану есебі қарастырылып, келесі нәтижелер алынды (Кесте 2).

Зерттеу жұмысында қойылған мақсатқа жету үшін міндеттер толық атқарылды:

1. Мәтінді оптикалық тану есептеріне рекурентті нейронды желіні қолданған шешімдерге шолу және талдау жасалды,

2. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер жинағы негізінде қазақ-орыс тілінде қолжазба мәтінді тану Attention-Gated-CNN-BGRU моделі негізінде жүзеге асырылды.

4. Қазақ-орыс тілдерінде HKR және KOHTD екі түрлі деректер жиынтығында қолжазбаны тану бойынша таңбалар қатесінің жиілігі(CER), сөздер қатесінің жиілігі(WER) және сөйлемдер қатесінің жиілігі(SER) есептеліп, нәтижесі кесте түрінде берілді.

Кесте 2. Қазақ және орыс тіліндегі қолжазбаны тану бойынша таңбалар қатесінің жиілігі(CER), сөздер қатесінің жиілігі(WER) және сөйлемдер қатесінің жиілігі (SER)

Қолжазба деректер жиыны	*Әдіс*	*CER*	*WER*	*SER*
HKR	*Attention-Gated-CNN-BGRU*	*6,40%*	*24%*	*36%*
KOHTD	*Attention-Gated-CNN-BGRU*	*8,22%*	*22,60%*	*25,22%*

Ұсынылған және сыналған модель Python үшін TensorFlow кітапханасын [46] қолдана отырып жүзеге асырылды, бұл Python көмегімен GPU-да жоғары оңтайландырылған математикалық операцияларды пайдалануға мүмкіндік береді.

Кітапханаларда, мұражайларда және мұрағаттарда тарихи құжаттардың үлкен коллекциялары бар, олар қазіргі уақытта сақтау, жариялау мақсатында цифрландырылады және тарихи мәлеметтерді онлайн цифрлық кітапханалар арқылы бүкіл әлемге қол жетімді жариялауға болады. Осындай тарихи құжаттарды нақты ақпараттық мазмұнмен, атап айтқанда мәтіннің транскрипциясымен қамтамасыз ету үшін қолжазбаны тану мәселесі іске асырылу керек. Болашақта тарихи құжаттардағы қолжазбаларды тану бойынша зерттеу жұмыстарын іске асыру жоспарланып отыр.

Қорытынды

Соңғы 10 жылда қолжазба мәтіндерін тану саласындағы ілгерілеу байқалады. Есептеулердің күрделілігіне қарамастан, кейбір өте күрделі есептеулерді машиналық оқыту және рекурентті нейрондық желілер үшін жасалған құрылымдар арқылы шешуге болады. Қолжазба мәтінін тану есептерін шешуде нейронды желілер сәтті қолданылады. Адам миының биологиялық құрылымына негізделген нейронды желілер өзінің жоғарғы деңгейде есептеу мүмкіндігімен басқа оқыту алгоритмдерінен бірнеше есе тиімді болып есептеледі. Осы мақалада ең көп қолданылатын тану модельдері, атап айтқанда Жасырын Марков модельдеріне (HMM), конволюциялық (CNN) және қайталанатын нейрондық желілерге (RNN) негізделген модельдерге шолу жасалынып, талданды. Бұл жұмыста қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер жинағы негізінде қазақ-орыс тілінде қолжазба мәтінді тану Attention-Gated-CNN-BGRU моделі негізінде жүзеге асырылды.

Пайдаланған әдебиеттер тізімі:

1 Bunke H., Bengio S., Vinciarelli A. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models, IEEE transactions on Pattern analysis and Machine intelligence. – 2004. – Vol. 26 (6), P. 709–720

2 Safabakhsh R., Adibi P. Nastaaligh handwritten word recognition using a continuous-density variable-duration HMM // Arabian Journal for Science and Engineering. – 2005. – Vol. 30 (1). P. 95–120

3 Chen M.-Y., Kundu A., Srihari S. N. Variable duration hidden Markov model and morphological segmentation for handwritten word recognition // IEEE transactions on image processing – 2005. – Vol 4 (12). P. 1675–1688

4 AlKhateeb J. H., Ren J., Jiang J., Al-Muhtaseb H. Offline handwritten arabic cursive text recognition using hidden Markov models and re-ranking // Pattern Recognition Letters – 2005. – Vol 32 (8). P.1081–1088.

5 Chung J., Gulcehre C., Cho K., Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling // arXiv preprint arXiv. – 2014. –Vol. 1412.3555.

6 Hochreiter S., Schmidhuber J. Long short-term memory // Neural computation – 1997. – Vol. 9 (8). P. 1735–1780.

7 Graves A., Fern´andez S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // Proceedings of the 23rd international conference on Machine learning. – 2006. – P. 369–376.

8 Ingle R.R., Fujii Y., Deselaers T., Baccash J., Popat, A. A Scalable Handwritten Text Recognition System // 2019 International Conference on Document Analysis and Recognition (ICDAR). – 1997. – P. 17-24.

9 Espana-Boquera S., Castro-Bleda M. J., Gorbe-Moya J., ZamoravMartinez F. Improving offline handwritten text recognition with hybrid HMM/ANN models // IEEE transactions on pattern analysis and machine intelligence. – 2010. – Vol. 33. P. 767–779.

10 Abdurahman F., Sisay E., Fante K. A. AHWR-net: offline handwritten amharic word recognition using convolutional recurrent neural network // SN Applied Sciences. – 2021. – Vol. 3 (8). P. 1–11.

11 Aradillas J. C., Murillo-Fuentes J. J., Olmos P. M. Boosting offline handwritten text recognition in historical documents with few labeled lines // IEEE Access 9. – 2 021. – P. 76674–76688. doi:10.1109/ACCESS.2021. 3082689.

12 Ngo T. T., Nguyen H. T., Ly N. T., Nakagawa M. Recurrent neural network transducer for Japanese and Chinese offline handwritten text recognition // in: International Conference on Document Analysis and Recognition, Springer, 2021, P. 364–376.

13 Balaha H. M., Ali H. A., Saraya M., Badawy M. A. New arabic handwritten character recognition deep learning system (AHCR-dls) // Neural Computing and Applications. – 2021. – Vol 33 (11). P. 6325–6367

14 N. Daniyar, B. Kairat, K. Maksat, A. Anel. Classification of handwritten names of cities using various deep learning models // 15th International Conference on Electronics, Computer and Computation. Abuja, Nigeria. – 2019. – P. 1–4. doi:10.1109/ICECCO48375.2019.9043266

15 Abdallah A., Hamada M., Nurseitov, D.B. Attention-Based Fully Gated CNN-BGRU for Russian Handwritten Text // Journal of Imaging. – 2020. – Vol.6. doi: https://doi.org/10.3390/jimaging6120141

16 Nurseitov D., Bostanbekov K., Kurmankhojayev D., Alimova A., Abdallah A., Tolegenov R. Handwritten Kazakh and Russian (HKR) database for text recognition. Multimedia Tools Applications. – 2021. – Vol. 80, pp. 33075–33097.

17 Daniyar N., Kairat B., Maksat K., Anel A. Classification of handwritten names of cities using various deep learning models. Advances in Science, Technology and Engineering Systems – 2021. – Vol 5, p.934-943, doi 10.25046/aj0505114

18 Abdallah A., Hamada M., Nurseitov D. Attention-based fully gated CNN-BGRU for russian handwritten text // Journal of Imaging. – 2020. – Vol. 6. P. 141. doi:10.3390/jimaging6120141.

19 Toiganbayeva N., Kasem M, Abdimanap G., Bostanbekov K., Abdallah A., Alimova A., Nurseitov D. KOHTD: Kazakh Offline Handwritten Text Dataset // Signal Processing: Image Communication. Elsevier.- 2022. – Vol.108. – P. 116827. doi: https://doi.org/10.1016/j.image.2022.116827

20 Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin, M. et al // Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv. – 2021. – Vol. 1603.04467.