Әл-Фараби атындағы Қазақ ұлттық университеті

ӘОЖ 004.4:811.512.122+004.4:811.161.1 Қолжазба құқығында

ТОЙГАНБАЕВА НАЗГУЛЬ АБЕНОВНА

Қазақ-орыс тілінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу және деректер базасын құру

6D060200 - Информатика

Философия докторы (PhD)
дәрежесін алу үшін дайындалған диссертация

Отандық ғылыми кеңесші:
физика-математика ғылымдарының кандидаты,
қауымдастырылған профессор Нурсеитов Д.Б.,
шетелдік ғылыми кеңесші:
доктор PhD, Астеро П.
«Demokritos» ұлттық ғылыми зерттеулер орталығы, Грекия

Қазақстан Республикасы
Алматы, 2025

МАЗМҰНЫ

НОРМАТИВТІК СІЛТЕМЕЛЕР..... 4
АНЫҚТАМАЛАР..... 5
БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР 6
КІРІСПЕ..... 7
1 ҚОЛЖАЗБА МӘТІНДЕРДІ ОПТИКАЛЫҚ ТАНУ ЕСЕПТЕРІНЕ ТЕРЕҢ ОҚЫТУДЫ ҚОЛДАНҒАН ШЕШІМДЕРГЕ ШОЛУ ЖӘНЕ ТАЛДАУ ЖАСАУ 12
1.1 Қолжазба мәтіндерінің деректер жинағына шолу 12
1.2 Кириллица графикасына негізделген қазақ және орыс тіліндегі қолжазбаны тану 17
1.3 Қолжазбаны тану классификаторларына шолу 18
2 ҚАЗАҚ ТІЛІНДЕГІ ОФФЛАЙН ҚОЛЖАЗБА МӘТІНДЕРІНІҢ ДЕРЕКТЕР ЖИНАҒЫН «KAZAKH OFFLINE HANDWRITTEN TEXT DATASET (KOHTD)» ҚҰРУ 23
2.1 Қолжазба мәтіндерді сегменттеуге генетикалық алгоритмді қолдану..... 23
2.2. KOHTD деректер жиынтығын құру, толтыру..... 30
2.3 KOHTD деректер жиынтығына статистикалық сипаттама 35
3 **ҚОЛЖАЗБАНЫ ТАНУҒА АРНАЛҒАН НЕЙРОНДЫҚ ЖЕЛІЛЕР**..... 40
3.1 Нейрондық желілер..... 40
3.2 Рекурентті нейрондық желілер 44
3.3 Ұзақ қысқа мерзімді жады (LSTM, Long Short-Term Memory) блоктары..... 45
3.4 Конволюциялық нейрондық желілер 47
3.5 Нейрондық желіні оқыту..... 48
3.6 Байланысты орнатуды уақытша қолдайтын классификациясы 50
4 ҚОЛЖАЗБА МӘТІНДЕРДІ ТАНУҒА ҰСЫНЫЛҒАН МОДЕЛЬДЕР ЖӘНЕ НӘТИЖЕЛЕР..... 54
4.1 CNN-LSTM-CTC әдісі негізіндегі Scheidl моделі..... 54
4.2 Gated-CRNN әдісі негізінде Bluche моделі..... 59
4.3 2-LSTM әдісі негізінде Puigcerver моделі..... 62
4.4 Gated-CNN әдісі негізінде Флор моделі..... 64
4.5 Attention-Gated-CNN-BGRU әдісі негізінде Abdallah моделі..... 67
4.6 Бағалау әдістері 71
4.6 Қазақ-орыс тіліне икемделіп оқытылған модельдердің нәтижелері 73
ҚОРЫТЫНДЫ 77
ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ.....**Ошибка! Закладка не определена.**

ҚОСЫМША А - KOHTD қолжазба мәтіндеріеің жиынындағы таңбалардың үлестірімі 866
ҚОСЫМША Ә - Авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік 91
ҚОСЫМША Б - Scopus базасындағы журналдық мақала 92
ҚОСЫМША В - KOHTD: Kazakh Offline Handwritten Text Dataset 93
ҚОСЫМША Г - Программа кодының бөлігі 94

НОРМАТИВТІК СІЛТЕМЕЛЕР

Берілген диссертацияда келесі стандарттарға сәйкес сілтемелер қолданылды:

Қазақстан Республикасы МЖМБС 5.04.034-2011. «Қазақстан Республикасы Мемлекеттік жалпыға білім беру стандарты. Жоғары оқу орнынан кейінгі білім. Докторантура». Негізгі ережелер ҚР БҒМ бекітілген. 17.06.2011 ж. №261. Астана, 2011 ж.

«Диссертацияны безендіру нұсқаулығы», Қазақстан Республикасы БҒМ ЖАК 28 қыркүйек, 2004 жыл. №377-3 ж.

ГОСТ 7.32-2001. Ғылыми зерттеу жұмысының есебі. Безендіру ережелері мен құрылымы.

ГОСТ 7.1-2003. Библиографиялық таспа. Библиографиялық сипаттау.

АНЫҚТАМАЛАР

Қолмен жазылған мәтінді тану (ағылш. Handwriting recognition, HWR немесе Handwritten Text Recognition, HTR) – бұл қағаз құжаттар, суреттер және сенсорлық экрандар және басқа құралдар көздерден түсінікті қолжазба ақпаратын алу және тану қабілеті.

Мәліметтер қоры – бұл белгілі бір құрылымға ие, өзара байланысты мәліметтерді (ақпараттарды) сақтау, өңдеу және басқару үшін арналған жүйе.

Датасет – бұл мәліметтер жиынтығы, яғни қандай да бір зерттеу, есептеу, талдау немесе машиналық оқыту (machine learning) үшін жинақталған құрылымды ақпарат.

Сегменттеу – бастапқы кескіндегі мәтінді құрылымдық және мағыналық бірліктерге: жолдарға, сөздерге және жеке символдарға бөлу арқылы кейінгі тану процесін оңтайландыру.

Генетикалық алгоритмдер – бұл табиғаттағы эволюция принциптеріне негізделген іздеу алгоритмдерінің тобы

Популяция – бұл хромосомалар ұрпақтарының жиынтығы, яғни екі сызық арасындағы бөлу сызықтарының жиынтығы.

OCR (Optical Character Recognition) — бұл оптикалық таңбаларды тану технологиясы, яғни сурет немесе сканерленген құжаттағы мәтінді автоматты түрде оқып, сандық мәтінге айналдыру процесі.

Нейрондық желілер – бұл адам миының жұмыс принципіне ұқсас етіп жасалған жасанды интеллект моделінің бір түрі. Олар үлгілерді танып-білу, болжам жасау, мәтінді, дауысты, суретті өңдеу сияқты күрделі есептерді шешу үшін қолданылады.

Байланысты орнатуды уақытша қолдайтын классификациясы (Connectionist Temporal Classification, CTC) – нейрондық желілер арқылы сегменттелмеген деректерді таңбалау міндетін іске асырады. Байланысты орнатуды уақытша қолдайтын классификациясының идеясы кіріс тізбегіне қолданылатын нейрондық желінің шығысы тікелей қажет таңбаларының тізбегі болып табылады, біздің жағдайда таңбалар тізбегі.

Рекурентті нейрондық желілер – элементтер арасындағы байланыстар бағытталған тізбекті құрайтын нейрондық желілердің бір түрі.

Конволюциялық нейрондық желілер (ағылш. Convolutional Neural Networks, қысқаша CNN) — бұл суреттерді және визуалды деректерді өңдеуге өте тиімді нейрондық желілердің бір түрі. Олар компьютерлік көру (computer vision) міндеттерінде кеңінен қолданылады.

БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР

HWR – Handwriting recognition, қолжазба мәтінді тану

HTR – Handwritten Text Recognition, қолжазба мәтінді тану

KOHTD – Kazakh Offline Handwritten Text Dataset, Қазақ тіліндегі оффлайн қолжазба мәтіндерінің деректер жиыны

CER – Character Error rate, таңбалардағы қателер жиілігі

WER – Word Error Rate, сөзіндегі қателер жиілігі

SER – Sentence Error Rate, сөйлемдеріндегі қателер жиілігі

HMMs – Hidden Markov models, жасырын Марков модельдері

ANNs – Artificial neural network, Жасанды нейронды желілер

RNNs – Recurrent neural networks, рекурентті нейронды желілер

GRU – Gated recurrent unit, басқарылатын рекурентті блок

LSTM – Long short-term memory, ұзақ қысқа мерзімді жады

CTC – Connectionist Temporal Classification, байланысты орнатуды уақытша қолдайтын классификациясы

CNNs – convolutional neural networks, конволюциялық нейрондық желілер

BLSTM – Bidirectional LSTM, екібағытты LSTM

DL– deep learning, терең оқыту

KOHTD – Kazakh Offline Handwritten Text Dataset (қазақ тіліндегі оффлайн қолжазба мәтіндерінің деректер жинағы

OCR – optical character recognition, оптикалық таңбаларды тану

GA – Genetic Algorithm, генетикалық алгоритм

CRNNs – Convolutional Recurent Neural Networks,конволюциялық қайталанатын нейрондық желілерге

Gated-CNN – Gated Convolutional Neural Networks, конволюциялық қайталанатын нейрондық желілер

MDLSTM – Multidimensional Long Short-Term Memory, көпөлшемді ұзақ қысқа мерзімді жады

CRNN – Convolutional Recurrent Neural Networks, конволюциялық қайталанатын нейрондық желілер

MLPs – Multi-Layer Perceptrons, көп қабатты перцептрон

КІРІСПЕ

Зерттеу тақырыбының өзектілігі. Әртүрлі салалар сандық ақпарат алмасуға көшірілгеніне қарамастан, сауалнамалар, жадынамалар, емтихан сұрақтарына жауап беру және мұрағаттық деректер сияқты көптеген құжаттарда әлі де қолмен енгізуді қажет етеді. Осыған байланысты қолжазбаларды компьютер арқылы транскрипциялаудың автоматты әдісі болып табылатын қолжазба мәтінін тануды жүзеге асыру қажет. Қолмен жазылған мәтінді тану (ағылш. Handwriting recognition, HWR немесе Handwritten Text Recognition, HTR) – бұл қағаз құжаттар, суреттер және сенсорлық экрандар және басқа құралдар көздерден түсінікті қолжазба ақпаратын алу және тану қабілеті. Қолмен жазылған мәтінді оффлайн тану – кірісі екі өлшемді кескін, шығысы таңбалар тізбегі болып табылатын кескіндегі әріптерді немесе сөздерді сандық мәтінге түрлендіру әдісі. Қазіргі уақытта қол жетімді қолжазбаны тану бағдарламалық шешімдерін онлайн және оффлайн тану деп екі негізгі бөлікке бөлуге болады: Онлайн тану әртүрлі құрылғылардың экранында арнайы құрылғылармен (стилустармен) жазылған мәтінді тануды қамтиды. Оффлайн қолжазбаны тану суреттегі мәтінді компьютерлік және мәтінді өңдеу қосымшаларында қолдануға болатын әріптік кодтарға автоматты түрде түрлендіруді қамтиды. Өз кезегінде, қолжазба мәтінін оффлайн тану мәселесі әлі толық шешілген жоқ. Соңғы 10 жылда орын алған баспа және қолжазба мәтіндерін тану саласындағы ілгерілеу байқалады. Есептеулердің күрделілігіне қарамастан, кейбір өте күрделі есептеулерді машиналық оқыту және терең нейрондық желілер үшін жасалған құрылымдар арқылы шешуге болады. Кириллица негіздегі қазақ-орыс тіліндегі қолжазбаларды тану мәселесі әлі де толық зерттелмеген. Сондықтан, қазақ-орыс тілдеріндегі оффлайн қолжазба мәтіндерін тану алгоритмдерін жасау өте маңызды және осы салада зерттеулер жүргізу қажет, яғни диссертация тақырыбы өзекті болып отыр. Сондықтан оптикалық таңбаларды тану (OCR) жүйелеріне негізделген тиімді автоматты қолжазба мәтіндерін тану алгоритмдерін жасау өте маңызды және осы салада зерттеулер жүргізу қажет. Бұл, әсіресе, қазақ тіліндегі қолжазба мәтіні үшін деректер жиынтығының жоқтығын ескерсек, зерттеу жұмыстарын жүргізу қажеттігі айқын. Осыған орай, бұл тақырып бүгінгі таңда өзекті және практикалық маңызы зор. Қолжазба мәтіндерін тану мәтін жазылған қағазды сканерлеуден, алынған ақпаратты мәтіндерге және бейнелерге бөлуден, қолжазбаларды интеллектуалды тану үшін әдістерді таңдаудан және нәтижелерді өңдеуден тұратын күрделі құрылымды үдеріс. Қолжазба мәтінін тану есептерін шешу үшін нейронды желілер қолданылады. Адам миының биологиялық құрылымына негізделген нейронды желілер өзінің жоғарғы деңгейде есептеу мүмкіндігімен басқа оқыту алгоритмдерінен бірнеше есе тиімді болып есептеледі

Қазіргі уақытта терең оқыту қатерлі ісікті диагностикалау, анықтау және жіктеу [1], бет фотоботтарын криминалистикалық тану [2], сөйлеуді тану [3] бейнеге негізделген адамның іс-әрекетін тану [4], қолмен жазылған тарихи жазбаларды тану [5], кестелерді анықтау және тану [6,7], сондай-ақ, терең оқыту

бағдарламалық жасақтама жобаларының уақыты мен кестесін оңтайландыру [8, 9] үшін қолданылады. Терең оқыту мен қарастыратын қолжазба мәтіндерді тану саласында сәтті қолданыс тапқан. Соңғы жылдары тану әдістері мен жүйелерінің жетілгеніне қарамастан, кескіндердегі қолжазба мәтіндерін автоматты түрде тану әлі де күрделі мәселе болып табылады. Қолмен жазылған жан-жақты және шексіз мәліметтер жиынтығы бұрынғыдан да маңызды бола түсуде. Қолжазба мәтіндері жадынамаларда, медициналық жазбаларда, тарихи құжаттарда, стилус енгізген мәтіндерде, емтихан жауаптарында және т.б. құжаттарда болады. Сондықтан кескіндердегі қолжазба мәтінін тану мәселесін толық шешілуі қамтамасыз етілуі керек. Бірнеше тілдер үшін қолжазбаны танудың кең ауқымды жүйелерін әзірлеу саласындағы зерттеулердің қажеттілігін көрсетеді.

Диссертациялық жұмыстың мақсаты: тереңдетілген оқыту модельдерін қолдана отырып, қазақ-орыс тілдерінің қолжазба мәтінін оптикалық тану үшін құралдар әзірлеу және деректер базасын құру.

Зерттеудің міндеттері: қойылған мақсатқа қол жеткізу үшін келесі міндеттерді іске асыру қажет:

1. Мәтінді оптикалық тану есептеріне терең оқытуды қолданған шешімдерге шолу және талдау жасау.
2. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер жинағын құру.
3. Сканерленген беттердегі мәтінді жолдарға, жолдарды сөздерге бөлетін сегментация үдерісін іске асыру үшін Генетикалық алгоритмді қолдану.
4. Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер қоры негізінде терең оқытудың әртүрлі әдістерімен эксперименттер жүргізу.

Зерттеу нысаны: қазақ-орыс тіліндегі қолжазба мәтіндерін тану.

Зерттеу пәні: қазақ-орыс тіліндегі қолжазба мәтіндерінің деректер жиынын құру және тану үшін терең оқыту модельдерін қолдану.

Зерттеу әдістері: зерттеу әдісі ретінде қолжазба мәтіндерін бөлу үшін генетикалық алгоритм, қолжазба мәтіндерін тану мәселесін шешуге рекурентті және конволюциялық нейрондық желілерді қолданылады.

Алынған нәтижелердің ғылыми жаңалығы:

- Нейрондық модельдерді оқыту мақсатында қолжазба мәтіндер корпусын краудсорсинг әдісі арқылы жинауға арналған автоматтандырылған технологиялық жүйе әзірленді.
- Қазақ тіліндегі қолжазба мәтіндерден сөздерді дәл және тиімді түрде сегменттеуге бейімделген генетикалық алгоритм жасалды.
- Алғаш рет қазақ тіліндегі сөз деңгейіндегі оффлайн қолжазба мәтіндерден тұратын құрылымды корпус — Kazakh Offline Handwritten Text Dataset (KOHTD) құрылды.
- Қазақ және орыс тілдеріндегі қолжазба мәтіндерді оптикалық тануға арналған Flor, Puigcerver, Bluche және Scheidl нейрондық архитектуралары қазақ тіліне бейімделіп, икемделген оқыту (fine-tuning) әдісімен үйретілді.

Жұмыстың теориялық және практикалық маңызы. Қолжазба жазбаларының цифрланған мәтіні көптеген компаниялардың бизнес-процестерін автоматтандыруға мүмкіндік береді және адамның жұмысын жеңілдетеді.

Сонымен қатар, қазақ-орыс тілдеріндегі қолжазба мәтіндерін тану пошта орталықтарының жұмысын автоматтандыруға мүмкіндік береді. Қазақ тілінде алғаш құрылған қолжазба мәтіндерінің деректер жинағы KOHTD көмегімен алгоритмдердің өнімділігін бағалауға және салыстыруға болады.

Қорғауға ұсынылатын негізгі тұжырымдар:

- Қолжазба мәтіндердің деректерін жинау, алдын ала өңдеу, анотациялау және корпусқа енгізу үдерістерін автоматты түрде орындайтын толығымен автоматтандырылған жүйе жасалды және тәжірибелік негізде іске қосылды.
- Қазақ тіліндегі қолжазба мәтіндерден сөздерді дәл сегменттеуге арналған генетикалық алгоритм бейімделіп, оның эвристикалық мүмкіндіктері арқылы әріпаралық шекараларды анықтау дәлдігі арттырылды.
- Қазақ тіліндегі оффлайн қолжазба мәтіндерден тұратын Kazakh Offline Handwritten Text Dataset (KOHTD) атты ашық датасет жасалды.
- Қазақ-орыс тілдеріндегі қолжазба мәтіндерді оптикалық тануға арналған Flor, Puigcerver, Bluche және Scheidl нейрондық модельдері қазақ тілінің графикалық ерекшеліктеріне байланысты икемделіп оқытылды.

Сенімділік дәрежесі және апробациялау нәтижелері: Диссертация тақырыбы бойынша зерттеу нәтижелері 5 мақала жарияланды және 1 авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік алынды. Зерттеу нәтижелері әл-Фараби атындағы ҚазҰУ компьютерлік ғылымдар кафедрасында ғылыми семинарларда, сонымен қатар, халықаралық конференцияларда баяндалды.

Scopus базасындағы журналдық мақала:

1) 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, 2021, vol.108, pp. 116827. doi 10.1007/s11042-021-11399-6 (Scopus: Q2; CiteScore-2.5; Percentile- 73%))

Ғылыми еңбектің негізгі нәтижелерін жариялау үшін Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым және жоғары білім саласында сапаны қамтамасыз ету комитеті ұсынатын ғылыми басылымдар тізбесіндегі журналдардағы мақалалар:

2) Тойганбаева Н.А., Әбдіманап Ғ.С., Алимова А.Н., Жунусова Ж.Х. Қазақ тіліндегі қолжазба мәтіндерінің мәліметтер қорын құру мен тану. «Алматы энергетика және байланыс университетінің Хабаршысы», 3(62), 2023, 105-116б.

3) Тойганбаева Н.А., Алимова А.Н., Әбдіманап Ғ.С., СакыпбековаМ.Ж., Гусманова Ф.Р. Нейронды желілер негізінде қазақ-орыс тілдеріндегі қолжазба мәтіндерді тану. Абай атындағы ҚазҰПУ. Хабаршы. Физика-математика ғылымдары сериясы. Том 84 № 4 (2023). 183-191б.

4) Тойганбаева Н.А. Терең оқыту модельдері негізінде қазақ-орыс тіліндегі оффлайн қолжазба мәтіндерін тану. Қазақстан Республикасы Ұлттық инженерлік академиясының хабаршысы. № 4 (90) – 2023.

Халықаралық конференция:

5) Тойганбаева Н.А., Тойганбаев Б.А. Нейронды желілерді қолжазбаларды интеллектуалды тану есептерінде қолдану мәселелеріне шолу // САТПАЕВ ОҚУЛАРЫ – 2021, Том II, 86-89б.

Авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік:

6) Нурсеитов Д.Б, Бостанбеков К.А. Әбдіманап Ғ.С., Тойганбаева Н.А., Алимова А.Н. "Kazakh Offline Handwritten Text Dataset (KOHTD)" қазақ тіліндегі қолжазба мәтінінің деректер қорына авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік. № 34615. 2023.

Зерттеушінің жеке үлесі. Диссертациялық жұмыстың қойылған мақсаттарына сәкес міндеттер шешілді. Қолжазба мәтіндерін тану үшін алдын-ала өңдеу жұмыстары және мәтін жолдарын генетикалық алгоритм көмегімен сегментациялау, KOHTD деректор қорын құру, толтыру жұмыстары іске асты. Flor, Puigcerver, Bluche және Scheidl модельдерімен тәжірибелер жүргізіліп, CER, WER, SER стандартты тиімділік көрсеткіштері бойынша бағаланды. Барлық модельдер Python және Tensorflow, Keras, OpenMP Matplotlib кітапханаларының көмегімен жүзеге асырылды.

Жұмыстың көлемі мен құрылымы. Диссертация қазақ тілінде жазылды. Диссертация кіріспеден, 4 тараудан және қорытындыдан тұрады. Диссертацияның жалпы көлемі 97 беттен, 42 суреттен, 10 кестеден, 95 қолданылған әдебиеттен тұрады.

Кіріспеде жұмыстың өзектілігі анықталып, қолжазбаны танудың кең ауқымды жүйелерін әзірлеу саласындағы зерттеулердің қажеттілігі көрсетілді. Сонымен қатар, диссертациялық жұмыстың мақсаты, міндеттері, әдістері, алынған нәтижелердің ғылыми жаңалығы, жұмыстың теориялық және практикалық маңызы жазылып, жарияланымдар тізімі берілген.

Бірінші тарауда терең оқыту мүмкіндіктерін қолжазба мәтіндерді оптикалық тану қолданған есептердің шешімдерге шолу жасалды. Қолжазбаны тану саласындағы деректер жиынтықтарына сандық талдау жасалды. Сонымен қатар, қолжазбаны тану саласында кеңінен қолданылатын деректер жиынтықтарына сандық талдау жүргізіліп, олардың көлемдік, құрылымдық және тілдік сипаттамалары сарапталды. Бұл талдау HTR жүйелерінің тиімділігін бағалау және салыстырмалы зерттеу жүргізу үшін негіз бола алады.

Екінші тарауда алғаш құрылған қазақ тіліндегі оффлайн қолжазба мәтіндерінің деректер жинағын Kazakh Offline Handwritten Text Dataset (KOHTD) құру, мәтінді сегменттеу, аннотация жазу қарастырылды.Осы бөлімде KOHTD деректер жиынтығына статистикалық талдау жасалды.

Үшінші тарауда қолжазбаларды тануда кеңінен қолданылатын нейрондық желілердің негізгі түрлері мен олардың жұмыс принциптері қарастырылды. Жасанды, рекуррентті (RNN), ұзақ-қысқа мерзімді жады (LSTM) және конволюциялық нейрондық желілердің (CNN) архитектуралары мен мүмкіндіктері сипатталды. Сонымен қатар нейрондық желілерді оқыту әдістері

мен Connectionist Temporal Classification (CTC) алгоритмінің қолжазбаны танудағы рөлі түсіндірілді.

Төртінші тарауда Bluche ұсынған Gated-CRNN әдісі, Puigcerver ұсынған 2-LSTM әдісі, Flor ұсынған Gated-CNN әдісі, Abdallah ұсынған Attention-Gated-CNN-BGRU әдісі қарастырылды. Сонымен қатар, модельдер KOHTD деректер жиынтығында оқытылып, стандартты өнімділік көрсеткіштері CER, WER және SER бойынша бағалау жасалды.

Қорытындыда диссертацияның негізгі нәтижелері мен қорытындылары берілген.

1 ҚОЛЖАЗБА МӘТІНДЕРДІ ОПТИКАЛЫҚ ТАНУ ЕСЕПТЕРІНЕ ТЕРЕҢ ОҚЫТУДЫ ҚОЛДАНҒАН ШЕШІМДЕРГЕ ШОЛУ ЖӘНЕ ТАЛДАУ ЖАСАУ

Қолжазба мәтіндерді оптикалық тану (Handwritten Text Recognition — HTR) мәселесі – соңғы жылдары жасанды интеллект, компьютерлік көру және табиғи тілдерді өңдеу салаларында кеңінен зерттеліп жатқан өзекті бағыттардың бірі. Аталған есеп түрі, әсіресе, көптілді және құрылымы күрделі тілдер үшін өзекті болып табылады, өйткені қолжазбаның вариативтілігі, әріптердің жалғануы, әртүрлі жазу стилдері мен графикалық ерекшеліктер тану дәлдігіне тікелей әсер етеді. Танудың сапасын арттыру мақсатында дәстүрлі әдістерден гөрі терең оқытуға (deep learning) негізделген тәсілдер кеңінен қолданыла бастады. Әсіресе, конволюциялық нейрондық желілер (CNN), қайталанатын нейрондық желілер (RNN), LSTM және Transformer архитектуралары қолданылып жүрген шешімдер ішінде жетекші орынға ие. Сонымен қатар, кіріс пен шығыстың ұзындықтары әртүрлі болған жағдайларда тиімді жұмыс істейтін Connectionist Temporal Classification (CTC) әдісі де HTR жүйелерінің ажырамас бөлігіне айналды.

Қолжазба деректер жинағын әртүрлі мақсаттарда пайдалануға болады. Атап айтсақ, қолжазбаны тануға арналған машиналық оқыту модельдерін үйретуге, әр түрлі алгоритмдердің дәлдігін бағалауға болады. Осы мүмкіндіктерді іске асыру үшін қолжазба деректер жиынтығы құнды деректер көзі болып табылады.

1.1 Қолжазба мәтіндерінің деректер жинағына шолу

Соңғы отыз жылда зерттеушілер ағылшын [10-12], орыс [13, 14], араб [15, 16], малаялам [17], жапон [18] және т. б. сияқты көптеген тілдер үшін қолжазба мәтінін танудың әртүрлі жүйелерін жасады. Терең және машиналық оқыту арқылы қолжазба мәтінін танумен айналысатын зерттеушілердің эксперименттеріне негіз ретінде қолжазба деректер жиыны (датасет) қажет.

IAM деректер жинағында [19, 20] ағылшын тілінде қолмен жазылған сөйлемдер бар. Авторлар ұсынған деректер жинағын қолжазбаны тану мәселелерін шешу үшін пайдалануға болады. Деректер жиынтығында шамамен 400 түрлі авторлар дайындаған 1066 бланк бар. Жинақта барлығы 82 227 сөз бар. Мәліметтер базасы толық ағылшын сөйлемдерінен тұрады. Дерекқор қолжазбаны танудың әртүрлі тапсырмалары үшін негіз бола алады. Бұл IAM деректер жинағындағы қолжазба деректер толықтырылып, жетілдіріліп отырды. IAM 3.0 қолжазба дерекқорын 657 түрлі авторлардың қолжазба мәтіндері құрайды, 1539 сканерленген қолжазба беттері, 5685 қолжазба сөйлемдері, 13 353 қолжазба жолдары бар және бұл дерекқор қолжазба мәтінін офлайн тану үшін пайдаланылды [21]. Мәтінді шығару құжаттарды тану жүйелеріндегі маңызды қадам болып табылады. Беттік құжаттан мәтінді сегменттеу үшін қолжазба мәтінінің барлық мүмкін аймақтарын анықтау қажет. Осы мәселелер [22] мақалада қарастырылып, қолмен жазылған мәтін жолдарын сегменттеудің тиімді алгоритмін ұсынған. Мәтіндік жол алгоритмі кескін сипаттамаларын алу үшін морфологиялық операторды қолданады. Әрі қарай, сызықпен сегменттелген

мәтін аймағын алу үшін гистограмманы проекциялау және қалпына келтіру реті ұсынылады. IAM дерекқорын қолданатын эксперименттердің нәтижелері бұл жаңа тәсілдің сенімді, жылдам және өте жақсы бағалау көрсеткіштерін беретінін көрсетті.

RAMIS [23] - өнеркәсіптік қолдану дерекқоры. Мәліметтер базасын құрудың негізгі себебі әр түрлі компанияларға жеке адамдар пошта және факс арқылы жіберген мәтіндерге ұқсас қолжазба мәтіндерінің үлгілерін жинау болды. Деректер базасы 1300 еріктілермен толтырылды, олар 12 723 беттен тұратын 5605 электрондық пошта арқылы мәліметтер жинауға үлес қосты. Әр электрондық поштада екі-үш парақ бар, оның ішінде ерікті хат, хат туралы ақпарат формасы және оңтайлы факс парағы бар. Содан кейін беттер сканерленіп, пошта классификациясы [24], қолжазба мәтінін тану [25] және авторды тану [26] сияқты тапсырмаларды тексеруді қолдау үшін мәліметтер базасы жарияланды.

IFN/ENIT [27] – араб қалаларының/ауылдарының қолжазба атауларын қамтитын деректер жинағы. Бұл жинақта 411 автордың қолмен жазылған формулярлық үлгілерінің 2200-ден астам екілік суреттері бар, формулярлардан 26000-ға жуық екілік сөз суреттері оқшауланған және оңай қол жеткізу үшін жеке сақталған. Мәліметтер жиынында әр сөз үшін негізгі ақиқат файлы жасалды. Бұл файлда негізгі сөздердің орналасуы, әлем туралы ақпарат, сондай-ақ сөзде қолданылатын жеке таңбалар туралы ақпарат бар. Дерекқор қолжазба араб сөздерін тану жүйелерін оқытуға және анықтауға арналған.

KHATT - Араб қолжазба мәтіндерінің дерекқоры. Ұсынылған араб тіліндегі қолжазба мәтіндерінің дерекқор жинағы мәтінді оффлайн тану үшін пайдалануға болады. Бұл дерекқорда 1000 түрлі авторлар жазған 1000 қолжазба формасынан тұрады. Мәліметтер базасының формалары кездейсоқ түрде сәйкесінше 70%, 15% және 15% оқыту, тестілеу және тексеру жиынтықтарына бөлінген. Дерекқор бүкіл әлем бойынша зерттеушілерге мәтінді тану, авторды анықтау және тексеру, пішіндерді талдау, алдын ала өңдеу, сегменттеу және т. б. сияқты қолжазба мәтініне қатысты әртүрлі мәселелерді зерттеу үшін қол жетімді.

HIT - MW [28] - қытайлық қолжазба мәтінінің дерекқоры, оны қытайлық қолжазба мәтінін офлайн тану үшін пайдалануға болады. HIT-MW дерекқорының қазіргі нұсқасында 853 пішін және 186 444 таңба бар. Деректер базасы пошта арқылы жиналады және делдал бетпе-бет емес, ол нақты қолжазбаны тануға қатысты көптеген қосымшаларға қызмет ете алады.

HIT_OR3C – қолмен жазылған Қытай таңбаларынан алынған қолжазба деректер жиыны. Ақпарат онлайн және офлайн режимінде қол жетімді. HIT_OR3C деректер жиыны 5 жиынтықтан тұрады: 4 жиынтықта сандар мен таңбалар [Цифр (1-10), (11-62), GB1 (63-3817), GB2 (3818-6825)] бар және құжаттардың 1 жиынтығы. Сандар мен таңбалар жиынын 122 еріктілер толтырған, 6 825 бет және барлығы 832 650 үлгі бар. Құжат жиынтығын 20 ерікті жазған және 2442 сыныптан 77 168 үлгіні қамтитын 10 жаңалық мақаласына сәйкес келеді. Құжаттардағы мәліметтер жиналғаннан кейін, жеке таңбаларға бөлініп, таңбалардың өлшемі 128 x 128 пиксельге дейін өзгертіліп қайта өңделді

және алғашқы төрт ішкі жиынға ұқсас бір суретте және бір векторлық файлда дәйекті түрде сақталды. Деректер жинағы 909 818 кескінді қамтиды. Деректер жинағының жалпы өлшемі 15,5 ГБ (сығымдалған 1125 Мб) [29].

Parzival [30] дерекқорында XIII ғасырда үш автор жазған 4477 мәтіндік жол, 23478 сөз нұсқалары, 4934 сөз жолдары, 93 әріптен тұратын ортағасырлық қолжазбаның 47 беті бар. Бұл дерекқорда 96 түрлі таңба бар. Қолжазбалар кескіндері неміс тілі тілінде пергаментті сиямен жазылған орта ғасырлардағы ең маңызды эпикалық шығармалардың бірі болып табылады. Washington [30] дерекқорында ағылшын тілінде 20 бет, 656 мәтін жолы, 4894 сөз нұсқалары, 1471 сөз топтары және 18 ғасырда екі автордың қолымен жазылған 82 хат бар. Мәтін қолмен жазылса да, кескіндер бинаризацияланған. Сандық кескінді өңдеу саласында кескінді бинаризациялау – бұл түрлі-түсті кескінді немесе сұр түсті кескінді екілік кескінге айналдыру үдерісі, яғни тек екі түсті кескін (әдетте қара және ақ). Бұл әдіс нысандарды, пішіндерді және әсіресе таңбаларды тану үшін қолданылады.

SHIBR [31] Швецияның көптеген махаллаларының мұрағаттарында сақталған туу туралы қолжазбалар жиынтығының кескіндерінен тұратын тарихи жиын. Бұл тарихи өолжазба деректер жиыны құжаттарды талдау алгоритмдерінің тиімділігін бағалауды жеңілдететін метадеректерді қамтиды. Бұл швед тіліндегі алғашқы және ең үлкен ашық қол жетімді деректер жиынтығы (1800-1840 жылдар аралығындағы 15000 жоғары сапалы түрлі-түсті кескіндер). Шведтер шежірешілер құнды деп санайтын статистика мен ақпаратты алу үшін деректер жинағына кейбір смарт-талдаулар жасайды. Сондай-ақ, осы тақырып бойынша жалпыға қол жетімді заманауи деректер жиынтығына жан-жақты шолу жасалған. Әр суретте мәтіндік транскрипция файлында 17 деректер бағаны бар. SHIBR деректер жиынтығын сөздерді тану мен құжаттарды талдау жарыстарында қолдануға болады.

Saint Gall [32] – 9 ғасырдағы каролинг шрифтімен жазылған латын қолжазбасының суреттері мен транскрипциясы бар қолжазба деректер жинағы. Аталған транскрипция қолмен жазылған кескінмен салыстырғанда, қол жетімді және оқырман үшін тамаша сапаға ие, бірақ сөздердің жазылуы дәл емес. Нәтижесінде транскрипцияда жол үзілімдері, аббревиатуралар және бас әріптер сияқты әртүрлі туралау мәселелері туындайды. Сондай-ақ, осы мәселелерді шеше алатын және құжаттың бүкіл беттерін тиімді туралай алатын жасырын Марков таңбаларының модельдеріне негізделген туралау жүйесі ұсынылған. Saint Gall дерекқоры тану жүйелерін оқыту үшін пайдаланылуы мүмкін.

ESPOSALLES [33] – бұл зерттеу жұмысында некеге рұқсат туралы буклеттер жинағынан алынған жаңа мәліметтер базасын ұсынылған. Бұл жинақта XX ғасырдың басына дейінгі жарты мыңжылдықты қамтитын қолжазба кітаптар бар. Сондай-ақ, берілген дерекқорға қатысты қолжазба мәтінін танудың заманауи жүйелерінің мүмкіндіктерін зерттеу ұсынылады. Аталған деректер жиынтығы әлеуметтік жазбаларды қамтитын тарихи құжаттарға арналған қолжазба мәтінін автоматты түрде тану бағдарламалық жасақтамасын жасау және сынау үшін пайдалы.

Germany [34] – мәтін жолдарын бөлу мен қолжазба мәтіндерін оффлайн танудың әртүрлі тәсілдерін эмпирикалық салыстыруды жеңілдету үшін ұсынылған деректер жиынтығы. Жиынтық 1891 жылғы 764 беттік испан қолжазбасын цифрландыру мен аннотациялаудың нәтижесі болып табылады. Бұл негізінен испан тілінде жазылған және стандартты мәліметтер базасымен салыстыруға болатын қолжазбаны зерттеуге арналған жалпыға қол жетімді мәліметтер базасы. Дерекқордың құрылымы кітап құрылымының арқасында ол интерактивті қолжазбаны тану жүйелерін шынайы бағалау үшін де жақсы жұмыс істейді. Бұл дерекқор зерттеулерде бастапқы нәтижелерді қамтамасыз ету және стандартты алдын ала өңдеу әдістері мен құралдарын, ерекшеліктерді айқындауды, HMM негізіндегі кескінді модельдеуді және тілдік модельдеуді қолданатын эмпирикалық нәтижелер алу үшін қолданылады.

RODRIGO [35] – көне тарихи құжатқа негізделген қолжазба мәтіндерінің дерекқоры. RODRIGO дерекқоры 1545 жылғы «Дон Родригоның Испания тарихы» деп аталатын қолжазбаға сәйкес келеді және оны бір автор толығымен ескі Кастилия (испан) тілінде жазған. Бұл испан тарихының шежіресін сипаттайтын 307 тарауға бөлінген 853 беттен тұратын том. Деректер жиынында 853 бет, 20357 жолдар бар. Ұсынылған қолжазба мәтіндерінің деректер жинағы тарихи құжаттардың типтік кемшіліктерін көрсетеді.

CArDIS [36] шведтік тарихи қолжазба таңбалары мен суреттерге негізделген жаңа жалпыға қолжетімді деректер жинағы. Деректер жиынтығында RGB түс кеңістігінде Швед алфавитінің 116 000 суреті бар, швед атауларының 30 000 сурет үлгісі, сондай-ақ Швеция аймақтарының 1 000 атауы бар. CArDIS деректер жинағы негізінде әртүрлі Машиналық оқыту классификаторларының өнімділігін зерттеу үшін эксперименттер жүргізіліп, нәтижесі жарияланды.

EPARCHOS - шамамен 1500-1530 жылдарға жататын грек қолжазба кодексінен алынған – қолжазба деректер жиынтығы. Бұл қолжазбалар Британ мұражайының Addit кодексінің бөлігі. Бұл мәтіндерді Antonius Eparchos және Camillos Zanettus жазған. Мәтінде ең маңызды аббревиатуралар, логограммалар мен жалғаулықтар келтірілген, олар қолжазбалардың транслитерациясы кезеңінен бастап (9 ғасыр) Византиядан кейінгі жылдарға дейінгі барлық грек қолжазба кодексінде келтірілген. Бұл мәліметтер жиынтығы 9285 мәтін жолын, 18809 сөзді (6787 ерекше сөз) қамтитын 120 сканерленген қолжазба мәтінінен тұрады [37].

EMNIST – арнайы NIST 19 дерекқорынан алынған, 28x28 пиксель кескін пішіміне және MNIST деректер жинағына тікелей сәйкес келетін деректер жиынының құрылымы түрлендірілген қолжазба таңбалар сандарының жинағы [38]. NIST 19 дерекқорында 3600 автордың қолмен жазылған бланк үлгілері, 810 000 таңба суреттері және осы суреттерге арналған негізгі сенімділік жіктемелері, деректерді одан әрі жинауға арналған анықтамалық пішіндер және суреттерді басқаруға арналған бағдарламалық жасақтама утилиталары бар. MNIST дерекқоры (Modified National Institute of Standards and Technology) сандарды қолмен жазу үлгілерінің көлемді дерекқоры. Кеңейтілген EMNIST қолжазба

деректер жиынтығы оқыту, жіктеу және компьютерлік көру жүйелері үшін стандартты сілтеме болып табылады.

Қолжазба мәтінінің үлгілерін бір немесе бірнеше адам әртүрлі уақытта жинауы мүмкін. Ғылыми жұмыста қолжетімді қолжазба деректер жиыны бойынша зерттеу жұмыстары сараланды. Қолжазбаны тануға арналған танымал деректер жиынтығы арасындағы салыстыру келтірілді (Кесте 1). Бұл кестеде әр деректер жиынындағы беттер, жолдар, сөздер, авторлар және тіл көрсетілген.

Кесте 1 – Кесте қолжазбаны тану саласындағы деректер жиынтығын сандық салыстыру

Ғалымдар	Деректер жиыны	Беттер саны	Жолдар саны	Сөздер саны	Қолжазба иелерінің саны	Қолжазба тілі
1	2	3	4	5	6	7
Marti U. және басқалар	IAM	657	13353	115320	657	ағылшын
Augustin E. және басқалар	RAMIS	12723	-	250000	13000	ағылшын
Pechwitz M. және басқалар	IFN/ENIT	-	-	26400	411	араб
Mahmoud S. A. және басқалар	KHATT	1000	9327	-	1000	араб
Su T. және басқалар	HIT-MW	853	86664	-	780	қытай
Fischer A. және басқалар	Parzival	47	4477	23478	-	неміс
Fischer A. және басқалар	Washington	20	656	4894	2	ағылшын
Cheddad A. және басқалар	SHIBR	1500	191301	-	1	швед
Fischer A. Ж. Және басқалар	Saint Gall	60	1410	11597	-	ағылшын
Romero V. және басқалар	ESPOSALLES	29	1563	8893	-	ағылшын
P´erez D. және басқалар	Germana	764	21000	217000	-	испан
Serrano N. және басқалар	RODRIGO	1545	20000	231000	-	испан

1 –кестенің жалғасы

1	2	3	4	5	6	7
Yavariabdi A. және басқалар	CArDIS	116000	-	30000	-	швед
Papazoglou A. және басқалар	EPARCHOS	120	9285	18809 (6787 ерекше сөздер)	-	грек
G. Cohen G. және басқалар	EMNIST		70000	814255	-	Ағылшын
Nurseitov D. және басқалар	*HKR*	*1500*	*15245*	*106718*	*200*	*Қазақ-орыс*
Біздің қолжазба деректер жиыны	*KOHTD*	*3000*	*21287*	*149015*	*1000*	*Қазақ-орыс*

Қолжазбаны тану алдын-ала өңдеу, сегментациялау, деректер жинағын құру, толтыру немесе дайын деректер жинағын пайдалану және тану нәтижелерін объективті сандық, сапалық бағалау үдерістерінен тұрады. Қолжазба мәтіндерінің деректер жинағындағы таңбаларды, сөздерді, сөйлемдерді тану нәтижелерін объективті сандық бағалау – өзекті мәселелердің бірі. Сондықтан да, қолжазба мәтіндерінің деректер жинағын жан-жақты зерттеу, талдау мәселесі өзекті және практикалық маңызды.

1.2 Кириллица графикасына негізделген қазақ және орыс тіліндегі қолжазбаны тану

Кириллица графикасына негізделген қазақ және орыс тіліндегі қолжазбаны тану бойынша зерттеу жұмыстары аз, сондықтан да зерттеу жұмыстарын жүргізу қажеттілігі туындайды. Осы өзекті мәселе «Электроника, компьютерлер және есептеу техникасы» бойынша халықаралық конференцияда талқыланып, нәтижелер жарияланды [39]. Қазақ және орыс тілдері бойынша жалпыға қолжетімді қолжазба деректер жиыны құрылып, модельдер оқытылды және салыстырмалы бағалау бойынша бірқатар эксперименттер жүргізілді [40]. Кириллица графикасына негізделген қолжазбаны тануды іске асыратын әртүрлі тәсілдер мен жетістіктер [41] жұмыста қарастырылған. Бұл мақалада сканерленген құжаттардағы орыс және қазақ тілдеріндегі қолжазба мәтінін анықтау мен танудың күрделі мәселелерін шешу үшін (Handwritten Kazakh and Russian) HKR деп аталатын ауқымды деректер қорын ұсынады (тиісінше 95% орыс және 5% қазақ сөздері/сөйлемдерінен тұрады). Зерттеу жұмысында орыс тілінде жазылған қолжазбаларды тануға басымдылық жасалған. Құрылған деректер қорында 200 аса автор жазған шамамен 63000 сөйлем бар. [42] жұмыста

кириллица негіздегі сөздерден тұратын деректер қорын пайдаланып, қолжазбаларды тану үшін әртүрлі архитектуралар қарастырылған. Қазақ-орыс тілдерінде жазылған қолжазбаларды тан бойынша эксперименттер әр түрлі терең оқыту әдістерін қолдана отырып жүргізілген және деректер жиынтығы үшін тану дәлдігі бойынша мәліметтер кестеде беріледі (Кесте 2):

Кесте 2 – Модельдердің таңбалардағы және сөздердегі қателер жиілігі

Модельдер	Таңбалардағы қателер жиілігі	Сөздердегі қателер жиілігі	Таңбалардағы қателер жиілігі	Сөздердегі қателер жиілігі
	1 тест		2 тест	
Bluche	16,15%	82,91%	10,15%	37,49%
Puigcerver HTR	73,43%	96,89%	54,75%	82,91%
The SimpleHTR	20,13%	58,97%	1,55%	11,09%

Кез-келген тілде көптеген сөздер бар. Мысалы, Оксфорд ағылшын сөздігінде 300 000-нан астам сөз бар. Қазақ тілінің сөздігінде 380 000-нан астам сөз бар [43], сондықтан барлық сөздерді қамтитын қолжазба сөздердің дерекқорын жинау мүмкін емес сияқты. Біздің білуімізше, қазақ тілі үшін жалпыға қолжетімді деректер жиынтығы жоқ. Зерттеу жұмысында сканерленген құжаттарда қазақ тіліндегі қолжазба мәтінді анықтау мен танудың күрделі мәселелерін шешу үшін Kazakh Offline Handwritten Text Dataset (KOHTD) деп аталатын үлкен деректер жинағын сипатталады. Қолжазбаны оффлайн тану үшін жаңа қазақ дерекқоры ұсынылды [44]. Деректер жиынтығы кириллицада жазылған және қазақ тілінде бірдей 42 таңбадан тұрады. KOHTD деректер жинағы студенттердің емтихан жазба жұмыстарының жинағы болып табылады. KOHTD деректер жиынтығында шамамен 922010 таңба және 140335 сегменттелген кескін бар. KOHTD көптеген себептерге байланысты ұсынылады. Біріншіден, бұл деректер қоры арқылы зерттеушілер терең және машиналық оқыту арқылы қолжазбаны тану мәселелерін шеше алады. Екіншіден, бұл әртүрлі алгоритмдердің өнімділігін бағалауға және салыстыруға арналған стандартты және таза деректер қоры. Үшіншіден, бұл қазақ тіліндегі қолжазбалардың алғашқы деректер қоры. KOHTD дерекқоры Сәтбаев университеті мен Әл-Фараби атындағы Қазақ ұлттық университетінің студенттері толтырылған емтихан жұмыстарының үлкен жинағынан тұрады. Деректер жинағындағы қолжазбаладардың 99% қазақ тілінде және 1% орыс тіліндегі сөздерден тұрады.

1.3 Қолжазбаны тану классификаторларына шолу

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. 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.

Бунке [45] кең сөздік қоры бар шексіз қолжазба мәтіндерін оффлайн тану жүйесін ұсынады. Қолжазба мәтінді тану жасырын Марков модельдеріне және статистикалық тілдік модельдерге негізделген (униграммалар, биграммалар және триграммалар-трограммалар). Деректерге қатысты мәтін ағылшын тілінде жазылғаны туралы бір ғана болжам жасалған. Бұл болжам мәтінді тану үшін статистикалық тілдік модельдерді қолдануға мүмкіндік береді. Зерттеу барысында бір және бірнеше авторлардың қолжазбасы негізінде эксперименттер жүргізілген. Әр түрлі көлемдегі лексика қолданылды (10 000-нан 50 000 сөзге дейін). Тілдік модельдерді қолдану жүйенің дәлдігін арттыратыны анықталды. Қолжазба мәтінді тану жүйесінің өнімділігі дәлдік пен тану тұрғысынан бағаланды.

Safabakhsh [46] парсы тілінде қолжазба сөздерді танудың толық әдісін ұсыну үшін үздіксіз тығыздығы бар жасырын Марковтың айнымалы ұзындық моделін, CDVDHMM (Continuous density variable duration hidden Markov model) [47] қолданады. CDVDHMM оңнан солға қарай ретті анықтау арқылы алынған ішкі таңбалар тізбегін модельдейді. Белгілер кеңістігінде таңбаларды бейнелеу үшін сегіз белгі қолданылады, оның ішінде үш Фурье сипаттамасы және бес құрылымдық және дискретті сипаттамалар. Бұл нысан векторы өлшем мен сдысуға сезімтал емес. Модель таза таңбаларды (қайталама соққыларсыз) және nastaaligh қолжазба стиліндегі таңбалардың кейбір құрама пішіндерін қарастырады. Алдын ала өңдеу кезеңінде, бинаризациядан, шуды басудан және қосылған компоненттердің сипаттамасынан кейін, бастапқы кескіннен жоғары және төмен сызықтарды, нүктелерді және басқа да кішігірім соққыларды табу және жою үшін жаңа алгоритмдер қолданылды. Содан кейін жоғарғы тізбекті талдауға және басқа екі процеске негізделген жаңа сегменттеу алгоритмі қолданылады. Бұл алгоритмнің негізгі мақсаты-жеткіліксіз сегменттеу мәселесін болдырмау. Жүйеде өзгермелі ұзақтық күйлерін есепке алу шамадан тыс сегменттеу мәселесін шешуге мүмкіндік береді. Қолмен жазылған үлгілермен

және 50 сөзден тұратын сөздікпен тәжірибе жасау жүйенің өте жақсы өнімділігін көрсеткен.

AlKhateeb [48] 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.

Басқарылатын рекурентті блок (Gated recurrent unit, GRU) [49] және ұзақ қысқа мерзімді жады (Long short-term memory, LSTM) [50] 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 [51]. Ақырында деректерді шығыс деректерімен сәйкестендіру іске асырылады.

Ingle R.R. [52] 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 [53] 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 [54] 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 [55] 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 [56] 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 [57] 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.

A. Yeleussinov жұмысында GAN (generative adversarial network) үлгілерін пайдалана отырып, қазақ қолжазба мәтінін тану кезінде мәтінді тану дәлдігін арттыру мәселесі қарастырылған. GAN негізіндегі ұсынылған модель дәстүрлі тану әдістерімен салыстырғанда қазақ қолжазбасын тану дәлдігін 11% - ға арттыра алғанын көрсеткен. Ұсынылған тәсіл таңбалардағы қателік жиілігін (CER) 11,15% деңгейінде және сөздегі қателік жиілігін (WER) 25,65% деңгейінде қабылдайды [58].

2 ҚАЗАҚ ТІЛІНДЕГІ ОФФЛАЙН ҚОЛЖАЗБА МӘТІНДЕРІНІҢ ДЕРЕКТЕР ЖИНАҒЫН «KAZAKH OFFLINE HANDWRITTEN TEXT DATASET (KOHTD)» ҚҰРУ

Қолжазба мәтіндерін тану мәтін жазылған қағазды сканерлеуден, алынған ақпаратты мәтіндерге және бейнелерге бөлуден, қолжазбаларды интеллектуалды тану үшін әдістерді таңдаудан және нәтижелерді өңдеуден тұратын күрделі құрылымды үдеріс. Осы бөлімде Kazakh Offline Handwritten Text Dataset (KOHTD) қазақ тіліндегі оффлайн қолжазба мәтіндерінің деректер жинағын құру, мәтінді сегменттеу қарастырылады.

Қазақ-орыс тілінің қолжазба мәтіндерін тану толық зерттелмеген күйінде қалып отыр. Осыған байланысты қазақ-орыс тілінің қолжазба мәтінін танудың жаңа тиімді алгоритмдерін әзірлеу және зерттеу өзекті мәселе. Нейрондық желілерді пайдалануға негізделген қазақ-орыс тілінің қолжазбасын тану мәселесіне көзқарас ұсынылып, зерттеу жұмыстары жүргізілді. Қазақ-орыс тілінің қолжазбасы туралы деректерді жинау мынадай негізгі кезеңдерден тұрады:

- Қолжазбаны тану үшін деректерді алдын-ала өңдеу: бұл кезеңде кескіннің сапасын жақсартылады және сегменттеуге ыңғайлы пішінге келтіріледі. Алдын ала өңдеу кезеңінде қолжазба мәтіні UFRII сериялы Canon MF 4400 сканерімен сканерленеді. Сканерленген емтихан парақтарының сапасы дюйміне 300 нүктені, ал түс тереңдігі 24 битті құрайды. Осы кезеңде қағаз құжаттарды цифрлық графикалық көрініске аударылады.
- Сканерленген қолжазба мәтіндерді сөздерге сегменттеу: осы кезеңде сканерленген қолжазба мәтіні талдауға ыңғайлы бөліктерге бөлінеді немесе сегменттеледі. Бұл кезеңдегі ең негізгі әрекеттер – мәтінді жеке жолдарға бөлу (жолдарды сегментациялау) және жолдарды сөздерге бөлу (сөздерді сегментациялау), мұнда бос орын олардың бөлгіші болып табылады. Ол үшін мәтінге шуды жою және сөз шекараларын анықтау үшін сүзгілер дәйекті түрде қолданылады.
- Әр суретті JSON форматындағы файлдағы мәтінімен салыстыратын сегменттелген сөздердің аннотациясын құру.

2.1 Қолжазба мәтіндерді сегменттеуге генетикалық алгоритмді қолдану

Қолжазба мәтіндерін сегменттеу оптикалық таңбаларды тану (optical character recognition, OCR) үдерісінің маңызды кезеңдерінің бірі болып табылады. Жол сегментациясы – бұл таңбалар тізбегін қамтитын кескіннің жеке таңбалары бар фрагменттерге ыдырауы. Толық бетті және жолдарды сегменттеу бойынша зерттеу жұмыстары бар [59, 60].

Жалпы генетикалық алгоритмнің жұмыс барысы екі кезеңнен тұрады: мәліметтерді жинау және құжаттағы мәтінді сөздеоге сегменттеу (Сурет 1).



Сурет 1 – Генетикалық алгоритм кезеңдері

Қолжазба мәтіндерінің деректер жинағына сегменттелген сөздермен жұмыс жасалуы қажет. Сондықтан, біздің жағдайда сканерленген беттердегі мәтінді сөздерге бөлу үшін генетикалық алгоритм қолданылады. Қазақ тілінде қолжазбаны тану үшін әр түрлі қолжазба үлгілерінен А4 форматындағы 3 мыңнан астам бет сканерленді (студенттердің сұрақтарына қолжазба емтихан жауаптары). Сканерленген емтихан парақтарының жолдарын сөздерге бөлу үшін генетикалық алгоритм қолданылды. Генетикалық алгоритмдер – бұл табиғаттағы эволюция принциптеріне негізделген іздеу алгоритмдерінің тобы [61]. Генетикалық алгоритм (Genetic Algorithm, GA) - кездейсоқ параметрлерді асыра пайдалануға негізделген классикалық эволюциялық алгоритм. Мұнда кездейсоқ ұғымы Генетикалық алгоритм көмегімен шешім табу үшін қазіргі шешімдерге жаңаларын құру үшін кездейсоқ өзгерістер қолданылған параметр. Бұл кішігірім және баяу өзгерістер енгізу арқылы жұмыс істейтін біртіндеп жүретін үрдіс. Сонымен қатар, Генетикалық алгоритм ең жақсы шешімді алғанға дейін өз шешімдеріне баяу өзгерістер енгізеді. Генетикалық алгоритмнің мақсаты қолжазба мәтін жазылған құжаттарда жолдарды кесу кезінде әріптердің төменгі немесе жоғарғы бөлігін кесіп алмау керек (мысалы: қолжазба әріптер р, в, б, д, з, қ, ң және т.б.). Хромосома гендерден тұрады. Біздің алгоритмде әрбір ген қалыпты жағдайда 50 пикселден тұратын белгілі бір ұзындықтағы жолдан және

екі жол арасындағы кездейсоқ биіктіктен тұрады. Популяция – бұл хромосомалар ұрпақтарының жиынтығы, яғни екі сызық арасындағы бөлу сызықтарының жиынтығы. Ұрпақ ата-анасының біреуінен 50% екіншісінен 50% гендерін алады. Нәтижеде алынған хромосомалар тізбегі ұрпақ болып табылады.

Генетикалық алгоритмнің бастапқы параметрлері (Кесте 3) алдын-ала анықталады.

Кесте 3 – Генетикалық алгоритмнің бастапқы параметрлері

Белгілеу	Мәні	Атауы
Pop_num	80	Популяция саны
L	28	Бір индивидтің ұзындығы (ген саны)
△x	50	Бір геннің ені(пикселмен)
epoch	70	Эпохалар саны
a_i	[0,H]	Индивид і нүктеснің у координатасы
I(y,x)	[0,255]	Суреттегі пиксель мәні

- image – 2d массив
- delta_x – әрбір қадамда өңделетін баған саны, ген ұзындығы (50 пиксель)
- length – қадам саны
- individ – әрбір баған бойынша позициялар
- Pop_num – популяция көлемі (80)
- epoch – эпохалар саны, яғни циклдер

Генетикалық алгоритм кезеңдері:

Алдын ала өңдеу. Бұл кезеңде суреттерді оқу және алдын ала өңдеу үрдісі жүргізіледі. Python тілінде жазылған және OpenCV кітапханасын қолданылды:

- glob.glob – көрсетілген жол бойынша барлық .jpg форматындағы суреттердің тізімін береді;
- f'{IMAGES_PATH}/*.jpg' – суреттер сақталған буманың жолын көрсетеді
- for imagename in images: – цикл арқылы суреттер өңделеді;
- fstart_time = time.time() – өңдеу уақыты қанша екенін есептеу үшін суретті өңдеудің басталу уақытын белгілейді;
- cv2.imread(...) – суретті жүктеу үшін қолданылады;

imagename – цикл арқылы алынған сурет файлының жолы;

image –суреттің пиксельдік мәліметтерін сақтайтын массив (матрица);

image.shape[:2] – суреттің биіктігі мен ені, height — суреттің биіктігі (жол саны), width –ені (бағана саны);

cv2.cvtColor(...) – бір түсті форматтан екіншісіне түрлендіреді;

cv2.COLOR_BGR2GRAY – бұл параметр BGR (түрлі-түсті сурет) форматынан GRAY (сұр реңкті сурет) форматына ауыстырады;

gray – өңделген, сұр реңкті сурет

Алдын ала өңдеу кезеңінде белгілі бір бумадан барлық .jpg суреттерді табады, әр суретті жүктейді, оның өлшемін (ені мен биіктігін) анықтайды және суретті сұр түске түрлендіреді.

$I \in R^{H \times W}$- кескін(матрица), H – биіктігі, W – ені, $\bar{I}_i$ – i-ші жолдағы пиксель мәндерінің орташа шамасы, y – кері өңделген және тегістелген бір өлшемді сигнал, σ – Гаус сүзгісінің дисперсиясы.

Жол бойынша орташа мән:

$$\bar{I}_i = \frac{1}{W}\sum_{j=1}^{w} I_{ij}, \text{ мұнда } i=1,\ldots,H \quad (2.1)$$

Кері өңдеу:

$$y_i = 255 - \bar{I}_i \quad (2.2)$$

Сигналды Гаус фильтрімен тегістеу:

$$\tilde{y}_i = \sum_{k=-K}^{K} y_{i-k} \cdot G(k;\sigma) \quad (2.3)$$

Мұнда k – бұл сүзгінің көрші индекстерін білдіреді (қадамдар саны), сигналдағы ағымдағы нүктеден ±көрші элементтерді қамтитын ағымдағы орын ауыстыру. K – бұл сүзгінің жарты терезе ені.

$$G(k;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{k^2}{2\sigma^2}} \quad (2.4)$$

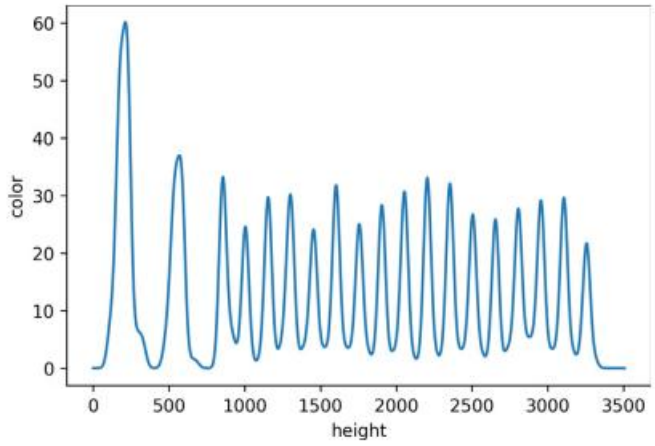
$G(k;\sigma)$ – Гаус функциясы.

Максимумды табу:

$$peaks = \{i \in [1, H-1] | \tilde{y}_{i-1} < \tilde{y}_i > \tilde{y}_{i+1}\} \quad (2.5)$$

- Генетикалық алгоритмді жүзеге асыру үшін, ең алдымен, өзгерістер енгізуді бастайтын координаттарды анықтау керек. Дұрыс сегменттеу үшін координаттар екі жолдың арасында орналасуы шарт. Егер жол толығымен ақ болса, онда пикселдердің қосындысы 255 болады, ыңғайлылық үшін инверсия жасалады, содан кейін ақ түс нөлге тең болады.

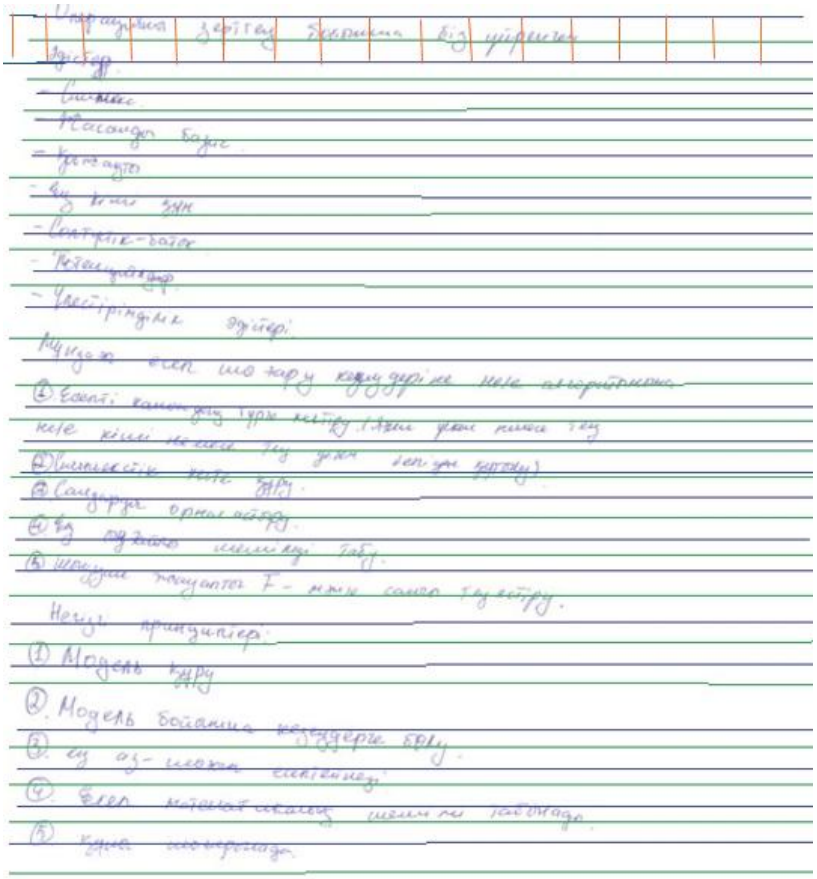
Шеткі нүктелерді анықтау үшін біз суреттің гистограммасын биіктігі бойынша аламыз (Сурет 2).



Сурет 2 – Әр жол бойынша пикселдер қосындысы

Бұл гистограмма бізге суреттің әр жолының пикселдерінің қосындысын береді, айқын шыңдарды алу үшін деректерді Гаусс сүзгісімен тегістелді. Мұнда height-парақтың биіктігі, color – әр жолдың орташа жарықтық мәні.

Генетикалық алгоритм орындалатын координатттар көк сызықтардың арасында жатыр (Сурет 3). Бастапқы жуықтау осы екі көк сызықтың арасын 50 пикселмен бөлу арқылы орындалады. Жасыл сызық хромосома. Хромосома өлшемі 50 пиксель болатын гендерден тұрады. Генетикалық алгоритм әр жол бойынша жеке-жеке тәуелсіз орындалады.



Сурет 3 – Максимальды және минимальды мәндер бойынша сызықтар сызу

Алғашқы популяция кездейсоқ түрде жасалады. Популяция мүмкін болатын шешімдер жиынтығы. Популяция құрылған соң фитнес функция есептеледі. Fitness функциясының мақсаты - әр индивидтің (жеке шешім) қаншалықты жарамды екенін бағалау. Фитнес функциясын есептеу үшін белгілеулер келесі түрде анықталды:

I(y,x)- суреттегі инверсияланған пиксель мәні

$\triangle x$-бір геннің ені

L – индивид ұзындығы

A=$[a_0, a_1, \dots, a_{L-1}]$ – индивидтегі y координаталар тізімі

Фитнес функциясы келесі түрде есептеледі:

$$Fitness(A) = \sum_{i=0}^{L-1} H_i + \sum_{i=1}^{L-1} V_i \rightarrow min \quad (2.6)$$

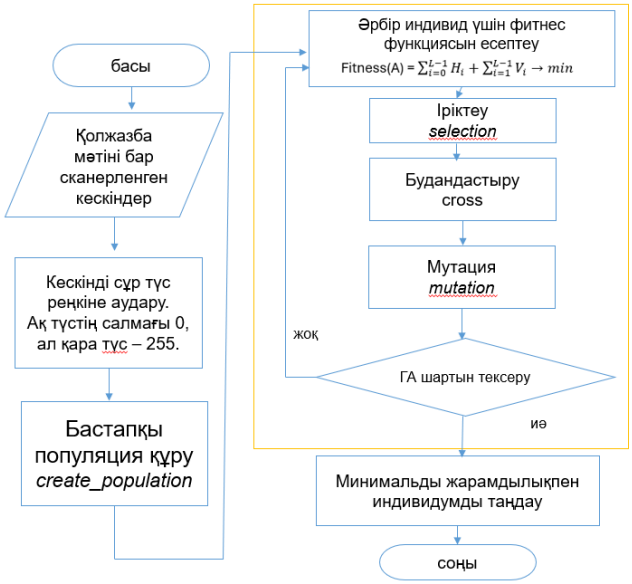
Горизонталь сегмент:

$$\mathrm{H}_i = \sum_{j=0}^{\triangle x-1} I(a_i, i \triangle x + j), i = 0, \ldots, L-1 \quad (2.7)$$

Вертикаль сегмент:

$$V_i = \sum_{k=\min(a_i,a_{i-1})}^{\max(a_i,a_{i-1})-1} I(k, i\Delta x), i = 1, \ldots, L-1 \quad (2.8)$$

Қазақ тіліндегі қолжазба мәтіндерден сөздерді дәл және тиімді түрде сегменттеуге бейімделген генетикалық алгоритмнің блок-схемасы келесі түрде жасалды (Сурет 4).



Сурет 4 – Генетикалық алгоритмнің блок-схемасы

Фитнес мәнінің өзгерісі бойынша тиімді EPOCH таңдалды (Сурет 5).



Сурет 5 – Фитнес мәнінің өзгерісі бойынша тиімді EPOCH таңдау

Генетикалық алгоритмнің үш негізгі операциясы бар:

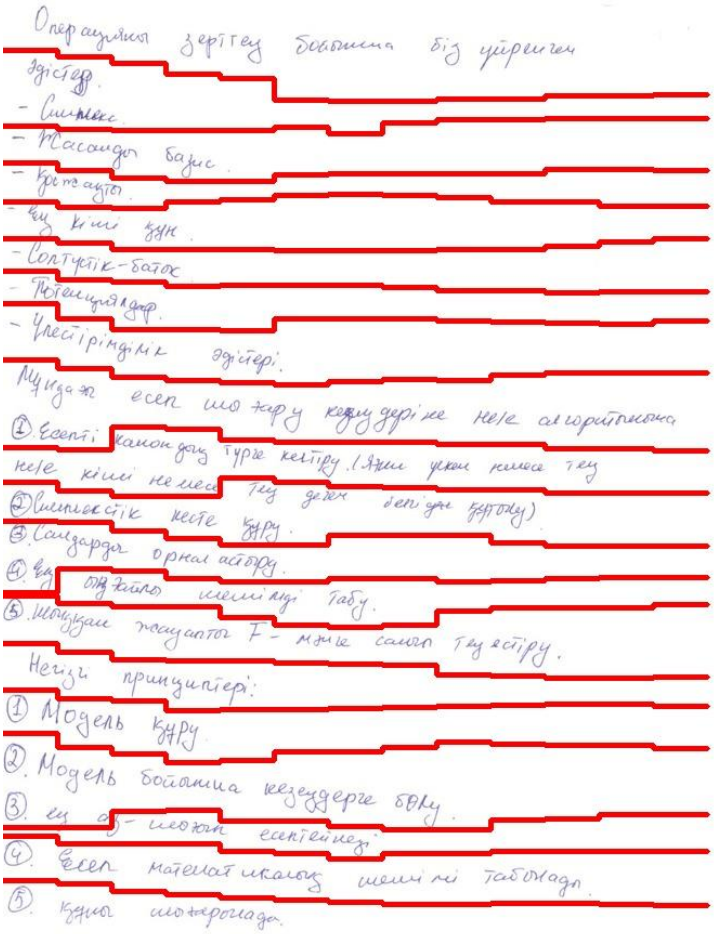
- Таңдау (Selection)
- Кроссовер (Crossover)
- Мутация (Mutation)

Полиморфизм қасиетін пайдалану үшін алгоритм интерфейсін құру. Интерфейс бастапқы популяция құру (create_population), іріктеу алгоритмі (selection), мутация алгоритмі (mutation), будандастыру алгоритмі (cross) және алдынғы үш функцияны шақыратын call функцияларынан тұрады.

Интерфейс келесі функциялардан тұрады:

- create_population - бастапқы популяция құру ;
- selection – іріктеу алгоритмі;
- mutation - мутациялар алгоритмі;
- cross – будандастыру алгоритмі;
- call – алдынғы 3 функцияны шақырады.

Генетикалық алгоритмді пайдаланып, нәтижесінде жолдарға сегментация жасалды (Сурет 6).



Сурет 6 – Генетикалық алгоритм нәтижесінде жолдарға бөлу

Жолдар бойынша сегментеция жасалғаннан кейін, жолдағы сөйлемді сөздерге бөлу үшін осы қолданылған әдіс қайталанады (Сурет 7).

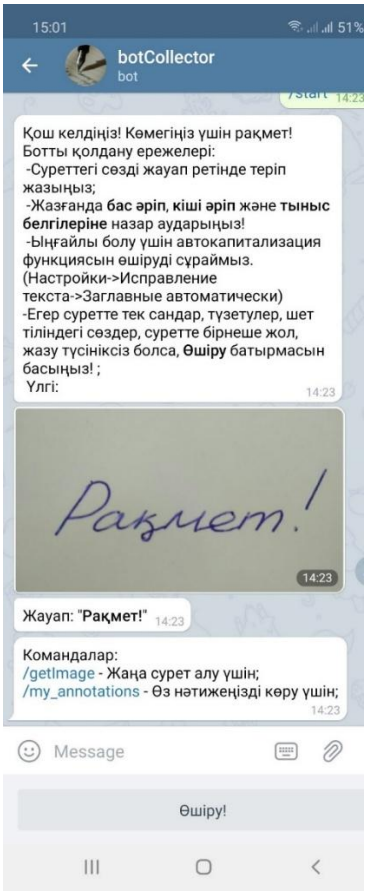


Сурет 7 – Генетикалық алгоритм нәтижесінде жолдарды сөздерге бөлу

Генетикалық алгоритм қайталанатын есеп болғандықтан, көп уақытты алады, уақыттың көп бөлігі фитнес функцияларды есептеуге кетеді. Өңдеу уақытын қысқарту үшін numba кітапханасын фитнес функциялары үшін қолданылды, нәтижеде өңдеу жылдамдығы 27 есе азайтылды (бір бейне үшін 190 секундқа 7 секунд). Сондай-ақ операцияларды параллельдеу үшін pump кітапханасын қолданылды, жалпы 1 бетті сегменттеуге 3,5 секунд уақыт кететін болды.

2.2 KOHTD деректер жиынтығын құру, толтыру

KOHTD деректор қорын толтыру үшін сегменттелген сөздердің суреттері ерікті пайдаланушыларға @collectorOfdataset_bot (botCollector) деп аталатын Telegram боты арқылы жіберілді. botCollector – бұл Python-да бағдарламалар жасауға және оларды бот ретінде тіркеуге мүмкіндік беретін ұялы телефондарға және компьютерге арналған мессенджер қосымшасы. Ерікті пайдаланушылар botCollector ұсынған сөзді енгізеді (Сурет 8).

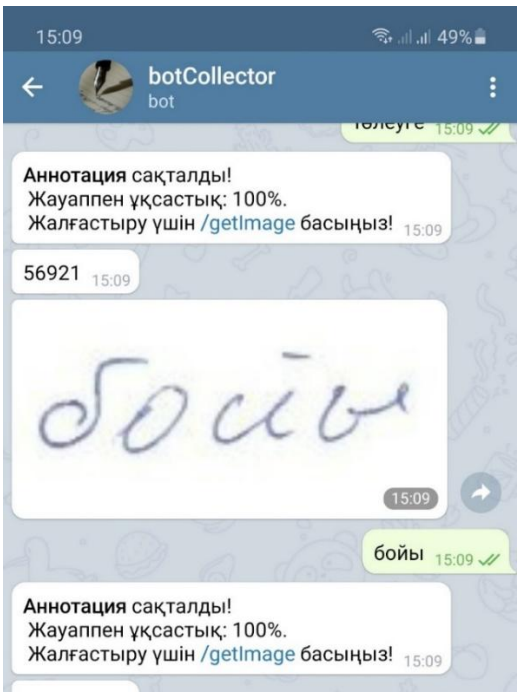


Сурет 8 – bot Collector

Арнайы құрылған botCollector-да бірнеше командалар бар:

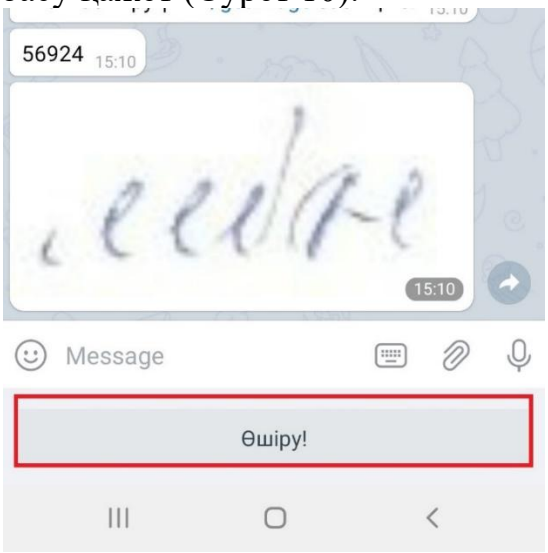
- */getImage* – жаңа сурет алу;
- */start* – бастапқы бетке келу;
- */my_annotations* – өз нәтижеңізді көру;
- */ all_ annotations* – барлық еріктілер тізімін көру;
- */get_graph* – аннотация таңбаларының санына байланысты гистограмма алу.

Егер ерікті пайдаланушы сөзді дұрыс жазса, сөзді жазсаңыз, 100% болады. Жалғастыру үшін */getImage* батырмасын басу қажет (Сурет 9).



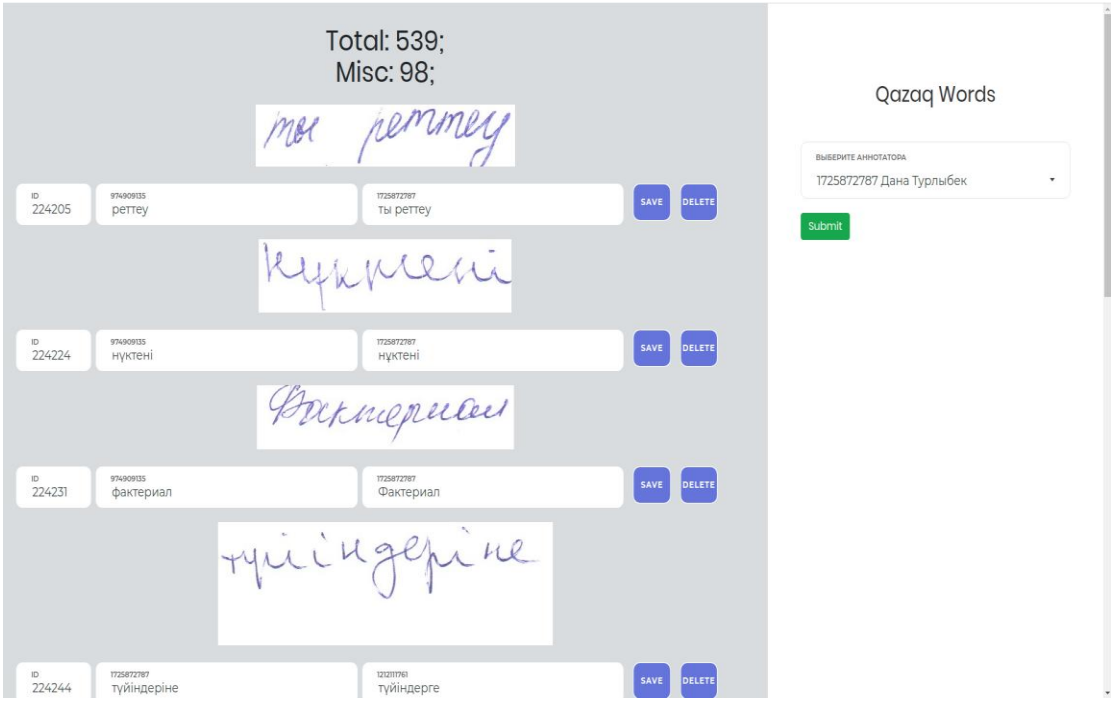
Сурет 9 – Ерікті пайдаланушының толтыруы

Егер сөз түсініксіз болса, яғни сандар, ағылшын сөздер, кесте, формулалар болса *Өшіру* батырмасын басу қажет (Сурет 10).



Сурет 10 – Өшіру батырмасы

Аннотацияларды жинау процесін жеңілдету үшін бір суретті екі кездейсоқ пайдаланушыға жіберу туралы шешім қабылданды. Егер екі пайдаланушыдан жіберілген аннотациялар абсалютті бірдей болса, біз бұл нәтижені сенімді деп таптық. Деректер жиынтығын толық тексерілді. Жалпы суреттердің жартысын жіберіп, нәтижелерді алғаннан кейін, пайдаланушылардың көпшілігі аннотацияларды өте ұқыпсыз толтырғаны, кейде тіпті дұрыс емес аннотациялар жазғаны белгілі болды. Мұндай жағдайларды болдырмау үшін нейрондық желі мүмкіндіктерін пайдаланып, бұрыннан бар толық емес деректер жиынтығында қолжазбаны тану үйретілді. Әрі қарай, әр аннотацияны енгізіп, жібергеннен кейін біз пайдаланушының нәтижесі мен тану нәтижесі арасындағы Левенштейн қашықтығын есептедік. Бізді таң қалдырғаны, тіпті ішінара деректер жиынтығында да нейрондық желі субъективті түрде тамаша нәтижелер берді (Сурет 11).



Сурет 11 – Нейрондық желілер негізінде жауаптарды түзету құралы

Левенштейн қашықтығы – екі таңбалар тізбегі арасындағы айырмашылықты модуль бойынша өлшейтін метрика. Ол таңбалардың бір тізбегін екіншісіне айналдыру үшін қажет бір таңбалы операциялардың (атап айтқанда, кірістіру, жою, ауыстыру) ең аз саны ретінде анықталады. Жалпы жағдайда, осы конверсияда қолданылатын операцияларға әртүрлі бағалар берілуі мүмкін. Левенштейн қашықтығы ақпарат теориясында және компьютерлік лингвистикада кеңінен қолданылады. Бұл жерде жол элементтері 1-ден бастап нөмірленеді. S_1 және S_2 сәйкес M және N ұзындықтағы екі жол болсын, онда Левенштейн қашықтығы келесі рекурентті формуламен есептеуге болады:

$$d(S_1, S_2) = D(M, N) \quad (2.9)$$

Мұнда

$$D(i,j) = \begin{cases} 0, i = 0, j = 0 \\ i, j = 0, i > 0 \\ j, i = 0, j > 0 \\ \min\{ \\ D(i, j-1) + 1 \\ D(i-1, j) + 1, \quad j > 0, i > 0 \\ D(i-1, j-1) + m(S_1[i], S_2[j]) \\ \} \end{cases} \tag{2.10}$$

Егер *a=b* болса, онда *m(a,b)=0*, кері жағдайда *m(a,b)=1*. Min{a,b,c} аргументтердің ең кішісін қайтарады. Мұнда i қадам бірінші жолдан (D) жоюды, j - бірінші жолға (I) кірістіру, ал екі индекс бойынша қадам (R) ауыстыруды немесе өзгеріссіз (M) білдіреді және келесі тұжырым ақиқат:

$$d(S_1, S_2) \geq \big||S_1| - |S_2|\big| \tag{2.11}$$

$$d(S_1, S_2) \leq \max(|S_1|, |S_2|) \tag{2.12}$$

$$d(S_1, S_2) = 0 \Leftrightarrow S_1 = S_2 \tag{2.13}$$

Левенштейн арақашықтығы – мәтіндер арасындағы ұқсастықты өлшеудің қарапайым, бірақ тиімді тәсілі. Ол әсіресе OCR, транскрипция және мәтіндік түзету жүйелерінде кеңінен қолданылады [62].

Сондай-ақ, */my_annotations* батырмасын таңдағаннан кейін, пайдаланушы ID, барлық жіберген сөздері, тексерілгені, қате сөздер, жауаппен ұқсастық, яғни сенімділіктің орташа мәні туралы мәліметтер шығады (Сурет 12):

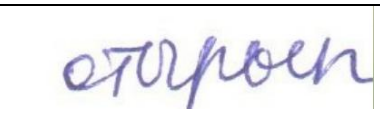
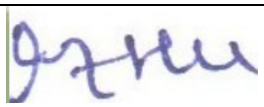
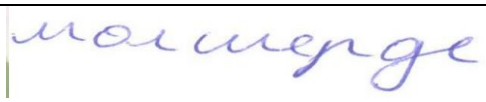
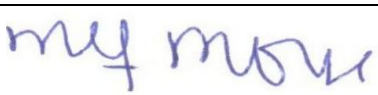
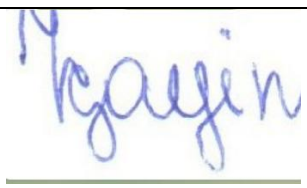


Сурет 12 – /my_annotations батырмасының нәтижесі

KOHTD деректер жиыны мәліметтер базасынан экспортталған JSON форматындағы файлдар түріндегі сөздердің суреттері мен аннотацияларынан тұрады. JSON (ағылш. JavaScript Object Notation) - JavaScript негізінде деректер алмасудың мәтіндік форматы. Бұл формат кез-келген бағдарламалау тілінде қолданыла алады.

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 3).

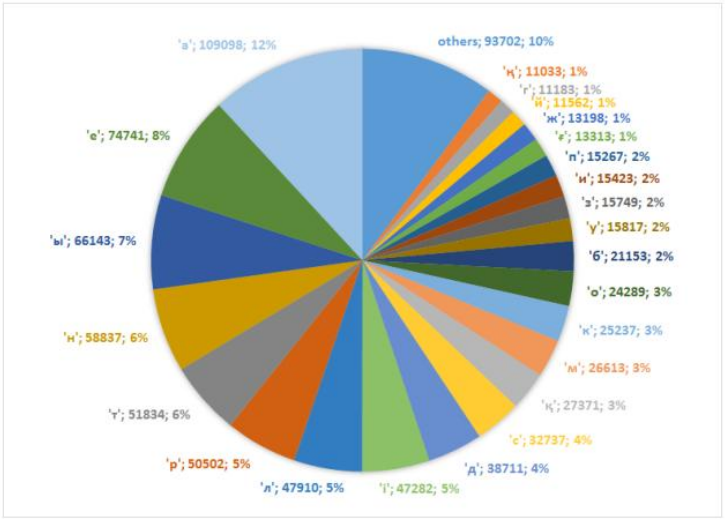
Кесте 4 – Әртүрлі авторлардың қолжазба мәтіндерін тану пайызы

№	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 4, 7 жол), *м* және *ш* (Кесте 4, 3 жол), *н* және *и* (Кесте 4, 6 жол), *л* және *е* (Кесте 4, 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. Қазақ тілінде қолжазба мәтіндерін тану үшін әр түрлі қолжазба үлгілерінен А4 форматындағы 3 мыңнан астам бет сканерленді (студенттердің сұрақтарына қолжазба емтихан жауаптары). Қазақ тілінде қолжазба мәтіндерін жинауда 643 адам қатысты (негізінен университет студенттері мен оқытушылары).

2.3 KOHTD деректер жиынтығына статистикалық сипаттама

Деректер базасы 3000-нан астам сканерленген емтихан жұмыстарының сегменттелген сөздерінен тұрады. Орта есеппен бір адам 2 емтихан парағын толтырған, сонда шамамен 1000 түрлі қағаз стилі бар. Суреттерге түсініктеме берген адамдардың саны бойынша біз бұл жазу кезінде 643 болды деп сенімді түрде айтуға болады, өйткені мәліметтер қорында барлық жазбалар бар. Сканерленген құжаттардағы мәтін жолдары генетикалық алгоритм көмегімен сегментацияланды және шамамен 922010 таңба және 140335 сегменттелген кескін жинақталды (Сурет 13).



Сурет 13 – KOHTD деректер жиынындағы таңбалар гистограммасы

Бұл бөлімде мәліметтер базасының статистикасы берілген. Деректер базасы үш эксклюзивті бөлікке бөлінеді (оқыту, валидация және тест жиынтығы). Осы жиындардың және бүкіл Дерекқордың uni-, bi - және tri-граммдары, сондай-ақ

бүкіл деректер жиынтығы мен оның бөліктері үшін қанша сөз, бірегей сөздер және таңбалар саны көрсетілген (Кесте 5).

Кесте 5 – KOHTD деректер жиынындағы N-грамм статистикасы

Жинақ	Сөздер саны	Бірегей сөздер	Таңбалар саны	Uni грамм	Bi-грамм	Tri-грамм
Оқытылғаны	104278	24943	98258	24943	100142	104242
Бекітілгені	22386	8870	21054	8870	22076	22376
Тестіленгені	22351	8984	21054	21054	22040	22346
Барлығы	149015	31483	140366	31483	130912	147238

Uni-грамм – құжаттағы әрбір жеке сөз өлшем болып табылады. Bi-грамм – дәйекті сөздердің әрбір жұбы өлшем болып табылады. Tri-грамм – дәйекті сөздердің әрбір үштігі өлшем болып табылады.

Келесі кестеде оқытуға арналған мәліметтер жиынтығымен салыстырғанда тексеру және тестілеу үшін мәліметтер жиынтығында сөздіктің болмауы (OOV, out of vocabulary) статистикасы көрсетілген (Кесте 6). Сөздік қорға кірмейтін сөздер (OOV, out of vocabulary) - бұл тексерілген сөйлеуде кездесетін, бірақ тану сөздігінде жоқ белгісіз сөздер. Бұл әдетте сөйлеуді танудың көптеген тапсырмаларының сәтті болуы үшін маңызды ақпаратты қамтитын атаулар мен орындар сияқты маңызды мағыналы сөздер. Дегенмен, сөйлеуді тану жүйелерінің көпшілігі бекітілген соңғы сөздік қорындағы сөздерді ғана танитын жабық сөздік тану жүйелері болып табылады.

Кесте 6 – KOHTD деректер жиынындағы сөздік қорға кірмейтін сөздер

Сөздік қорға кірмейтін сөздер	Сөздік қорға кірмейтін сөздер (токендер)	Сөздік қорға кірмейтін сөздер (пайызбен)
Бекітілгені	3408	10.36%
Тестіленгені	3442	10.46%

KOHTD деректер жиыны қолжазба мәтіндердің JSON форматындағы файлдар түріндегі сөздердің суреттері мен аннотацияларынан тұрады.

KOHTD құрамында:

- 1926 қолжазба емтихан жұмыстары;
- 140335 сегменттелген кескін (Сурет 14);
- 922010 таңбадан бар.



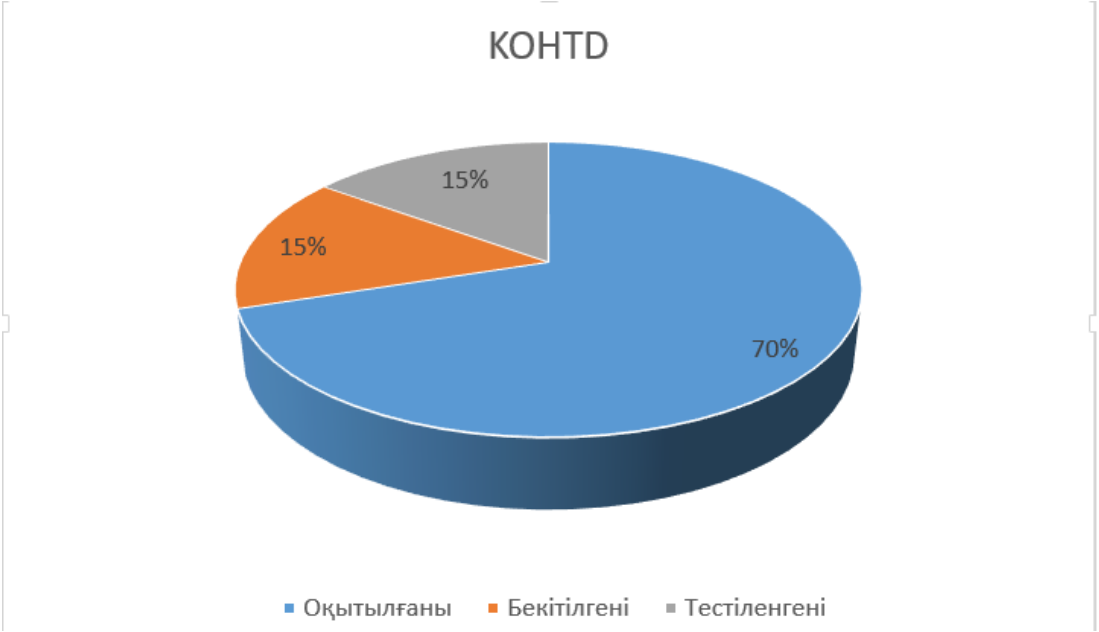
Сурет 14 – KOHTD деректер жиыны қолжазба мәтін кескіні

KOHTD қолжазба мәтіндеріеің жиынындағы таңбалардың үлестірімі кестеде көрсетілген (Қосымша А). KOHTD қолжазба мәтіндеріеің жиынында:

- Қазақ әліпбиінің 42 кіші әрпі;
- Қазақ әліпбиінің 38 бас әрпі;
- 10 араб цифры;
- 30 түрлі таңбалар, оның ішінде бос орын, тыныс белгілері және басқа белгілері;
- Латын әліпбиінің 26 кіші әрпі;
- Латын әліпбиінің 26 бас әрпі бар.

KOHTD қолжазба мәтіндерінің жиынындағы таңбалардың үлестірімінің кестесі Қосымша-1-де беріледі.

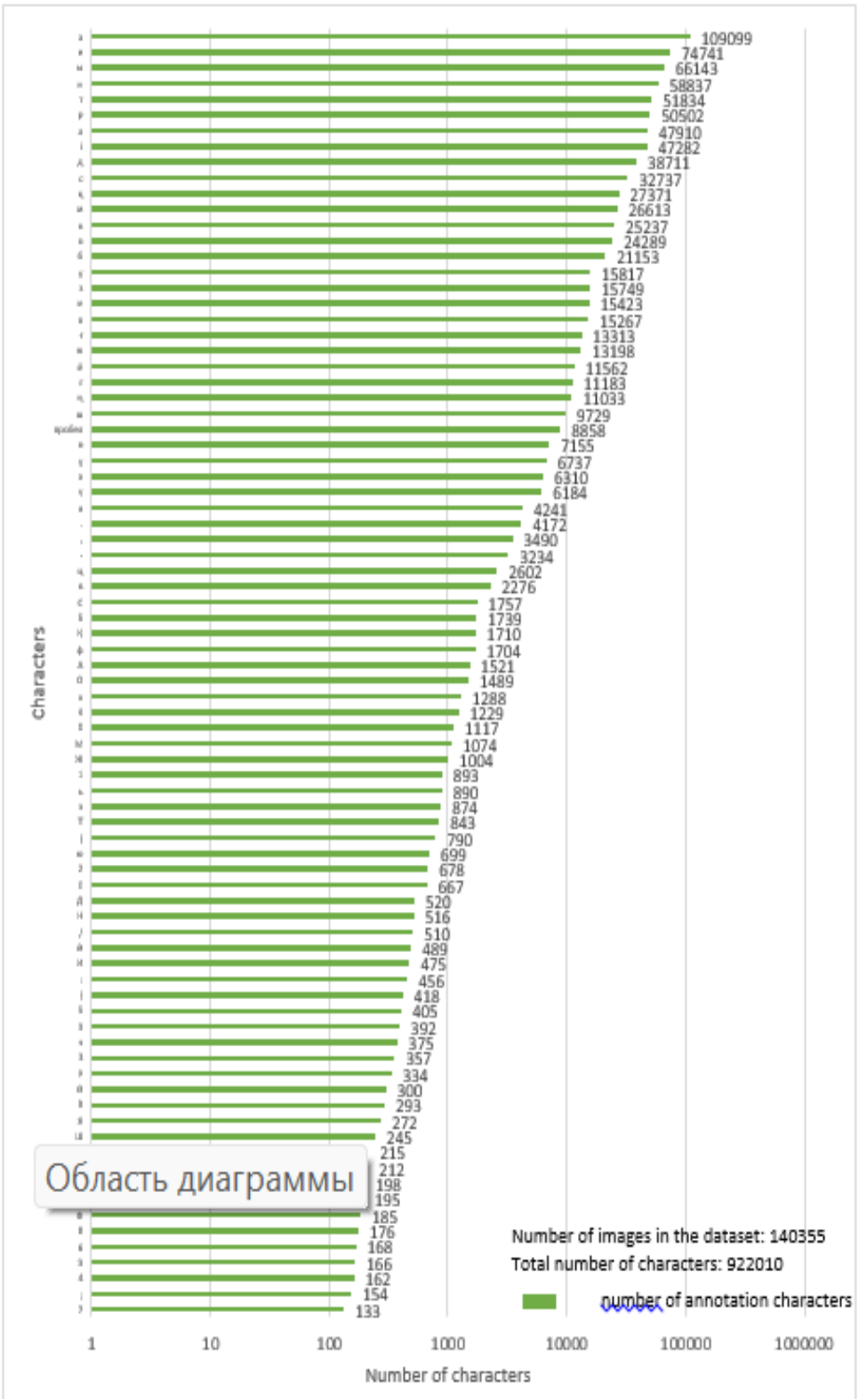
Kazakh Offline Handwritten Text Dataset (KOHTD) деректер жиынтығы құрылымдық түрде үш негізгі бөлікке бөлінген: оқыту (training) жиынтығы, валидация (validation) жиынтығы, және тестілеу (testing) жиынтығы (Сурет 15).



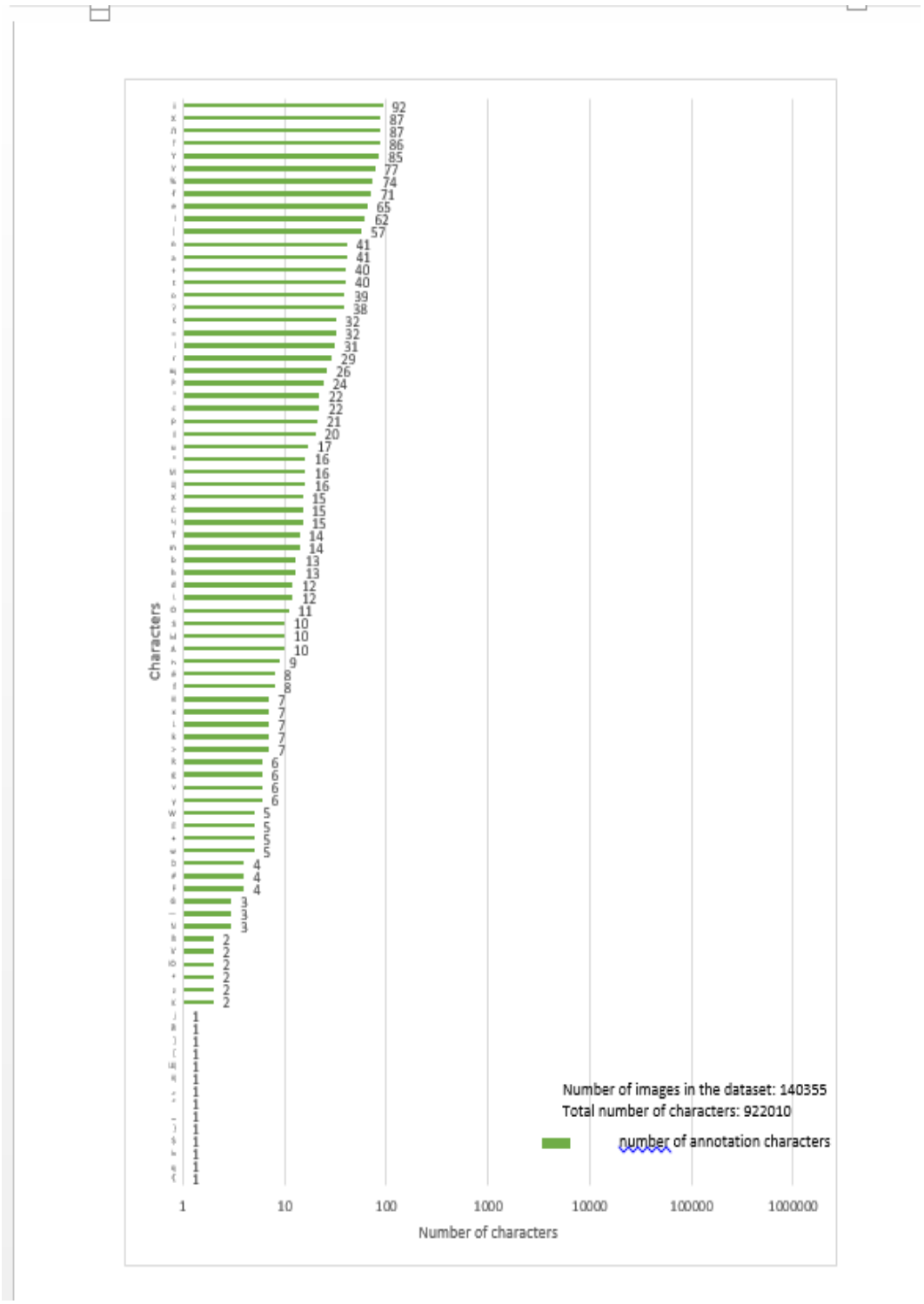
Сурет 15 – KOHTD жиынтығының оқыту, валидация және тестілеу жиынтықтары бойынша бөлінуі

Бұл деректер құрылымын оқыту, валидация және тестілеу жиынтықтарына бөлу әдістемесі модельдің жалпылау қабілетін арттырып, оны нақты қолдану жағдайларындағы тиімділігін объективті бағалауға мүмкіндік береді. Сонымен қатар, мұндай сегментация деректердің ішкі құрылымдық біртұтастығын сақтауға, эксперимент нәтижелерінің қайталанбалылығын қамтамасыз етуге және модельді оқыту үдерісінде жүйелік қателердің алдын алуға ықпал етеді.

KOHTD деректер жиынымен жұмыс жасау үшін Python 3 және numpy, matplotlib кітапханаларын орнату керек. KOHTD деректер жиынындағы әріптердің үлестірілу гистограммасы 15-суретте көрсетілген (Сурет 16).



Сурет 16 – KOHTD деректер жиынындағы әріптердің үлестірілу гистограммасы



Сурет 17 – KOHTD деректер жиынындағы әріптердің үлестірілу гистограммасы (жалғасы)

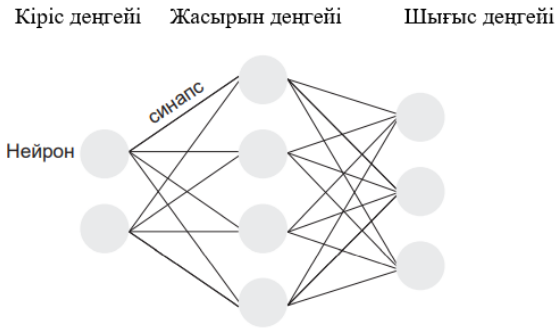
Мәтінді офлайн режимде тану модельдеріне арналған «Kazakh Offline Handwritten Text Dataset (KOHTD)» қазақ тіліндегі қолжазба мәтінінің деректер қорына авторлық құқықпен қорғалатын объектілерге құқықтардың мемлекеттік тізімге мәліметтерді енгізу туралы куәлік алынды [63].

3 ҚОЛЖАЗБАНЫ ТАНУҒА АРНАЛҒАН НЕЙРОНДЫҚ ЖЕЛІЛЕР

3.1 Нейрондық желілер

Нейрондық желілер (NNS Neural Networks, NNs) – жалпы үлгіні танудың танымал жүйелері. Нейрондық желілер бір-бірімен өлшенген және бағытталған қосылыстармен байланысқан негізгі процессорлық блоктардан тұрады, осылайша кейбір блоктардың шығысы басқаларға кіріс болып табылады. «Жасанды нейрондық желі» атауы осы модельдердің элементтері мен биологиялық нейрондардың ұқсастығынан шыққан. Жасанды нейрондық желілердің алғашқы ресми сипаттамасын McCulloch & Pitts (1943) ұсынды [64]. Қосылыстардың салмағын реттеуге арналған алгоритмдер перцептронға [65] және көп қабатты перцептрондарға (Multi-Layer Perceptrons, MLPs) негізделеді [66]. Нейрондық желілердің әртүрлі түрлерінің ішінде қайталанатын нейрондық желілер (Recurrent Neural Networks, RNNs) тізбектерді өңдеуге қабілетті, конволюциялық нейрондық желілер (Convolutional Neural Networks, CNNs) кескіндермен оңтайлы жұмыс жасайды.

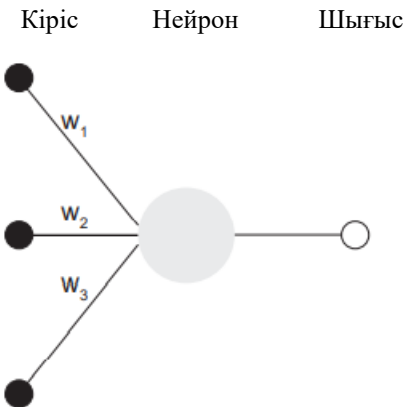
Жасанды нейрондық желі (немесе жай нейрондық желі) – бұл біздің миымыз жұмыс істейтін принциптер бойынша жұмыс істейтін бағдарламалық жасақтама. Біздің биологиялық жүйке жүйесін басқару синапстар деп аталатын арналар арқылы бір-бірімен өзара әрекеттесетін нейрондар арқылы жүзеге асырылады. Нейрондық желіде өзара байланысты жасанды нейрондар желіні оқыту үшін адам миының нейрондарының жұмысын модельдейді. Оқу процесінде қажетті нәтижеге жету үшін белгілі бір нейрондар арасындағы байланыс күшейтіледі. Келесі диаграммада үш деңгейлі нейрондық желі көрсетілген (Сурет 18). Әр шеңбер нейронды білдіреді, ал олардың арасындағы сызықтар синапстарды модельдейді.



Сурет 18 – Үш деңгейлі нейрондық желі

Осы суретте толық байланысқан желі көрсетілген – берілген деңгейдегі әрбір нейрон келесі деңгей нейрондарымен байланысады. Нейрондық желілерді оқыту итеративті принцип бойынша жүреді. Әрбір итерация дәуір (epoch) деп аталады. Оқу жиынтығындағы үлгі бір рет өңделеді және «дұрыс» дәуірлер саны жоқ. Бұл гиперкөрсеткішті оқу деректеріне және моделіне байланысты баптау

қажет болды. Желінің кіріс мәліметтері ретінде оқу үлгілерінің белгілері алынады. Бір деңгейлер алдынғы деңгейлердің шығысы бойынша жаңа белгілерді зерттесе, басқалары осы белгілерді болжам жасау үшін қолданады. Келесі суретте барлық нейрондарға (немесе олардың бөліктеріне) берілетін үш кірістен, бір шығыстан тұратын нейрон бейнеленген (Сурет 19):



Сурет 19 – Үш кірістен, бір шығыстан тұратын нейрон.

Мұнда w1, w2 және w3 мәндері салмақ деп аталады. Жаңадан оқылатын модельдерде бұл мәндер модель арқылы кездейсоқ қосылады. Оқыту барысында желі болжамды модель белгілері мен нақты үлгі белгілері арасындағы қателіктердің үлесін азайтуға тырысады. Қателіктердің үлесі шығындар деп аталады, ал шығындар мөлшерін анықтау формуласы шығындар функциясы деп аталады. Оқу процесінде желі әр нейронның жалпы шығындарға әкелетін мөлшерін анықтайды, содан кейін ол деңгейлерден кері бағытта өтеді және шығындарды азайтуға тырысып, салмақты реттейді. Бұл әдіс кері таралу деп аталады. Салмақтарды оңтайландыру градиентті түсу деп аталатын үдерісте біртіндеп іске асады. Нейрондық желілер қолжазба мен сөйлеуді тануда заманауи жүйелерінің негізгі құрамдас бөлігі болып отыр. Дербес жағдайда, қолжазба мен сөйлеуді тануда ең тиімді болып танылған рекурентті нейронды желілер.

Перцептрон – «иә/жоқ» шешімін қабылдауға бағытталған екілік классификатор. Ү шығысы теріс және оң шешімге сәйкес келетін екі мәнді қабылдай алады және оны келесідей тұжырымдауға болады:

$$y = f(x) = sign(b + \omega_1 x_1 + \cdots + \omega_n x_n) \quad (3.1)$$

Мұнда x=$x_1, \ldots, x_n$ – кіріс векторы, $b, \omega_1, \ldots, \omega_n$ – еркін параметрлер (салмақ), егер $x > 0$ болса, онда $sign(x) = 1$ басқа жағдайда -1-ге тең. Белгі функциясының орнына тек екі мәнді қабылдай алатын басқа функция опцияларын қолдануға болады, мысалы, Heaviside функциясы. Алайда,

практикалық себептерге байланысты дифференциалданатын (және үздіксіз) функцияға артықшылық беріледі. Сигма тәрізді функция таңдалады:

$$\sigma(\mathrm{t}) = \frac{1}{1+\mathrm{e}^{-\mathrm{t}}} \tag{3.2}$$

бұл Хевисайд (Heaviside) функциясына ұқсайды, бірақ үздіксіз және дифференцияланады. Сигма тәрізді нәтижелер бойынша шешім қабылдау арқылы екілік жіктеуді орындай аламыз:

$$y = \begin{cases} 0, \text{егер } \sigma(w^T x + b) < 0.5 \\ 1, \text{басқа жағдайдарда} \end{cases} \tag{3.3}$$

Мұнда $w = [\omega_1 \dots \omega_n]$ салмақ векторы, b – болжам. Сигма тәрізді нәтиженің шекті мәні кіріс кеңістігінде шешім қабылдау шекарасы бинарлы классификаторын гипержоспар ретінде келесі теңдеумен анықтайды:

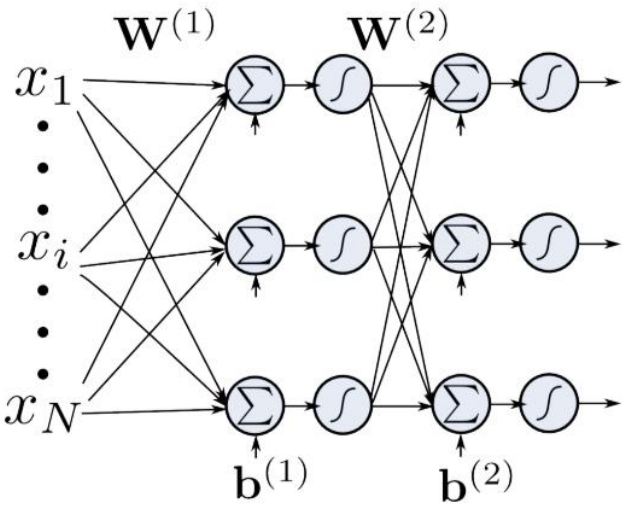
$$\mathrm{b} + \omega_1 \mathrm{x}_1 + \cdots + \omega_\mathrm{n} \mathrm{x}_\mathrm{n} = 0 \tag{3.4}$$

Демек, перцептрон жіктеу тапсырмасын кеңістікте сызықтық түрде бөлуге болатын кезде ғана оңтайлы болады. Бұл модельде кіріс мәліметтері x_i басқа нейрондардан синапстар арқылы келеді, ал ω_i салмақтары синаптикалық байланыстың күші мен қоздырғыш/тежегіш сипатын білдіреді. Жасуша денесінде барлық кіріс қосылыстарының үлесі жинақталады және егер ол алдын ала белгіленген шектен асып кетсе, нейрон іске қосылады. Бұл нейронның жеңілдетілген және дәл емес моделі болғанымен, бұл модель (жасанды) нейрон деп аталады, ал мұндай нейрондардың ассоциациясы олардың біреуінің шығысын басқаларының кірісіне қосу (жасанды) нейрондық желілерді анықтайды.

Ары қарай $a = \sum_{i=1}^{N} \omega_i x_i + b$ арқылы нейронды белсендіруді (сызықтық емес активтендіру функциясына дейін) және оның z шығысы деп белгілейміз.

Көп қабатты перцептрон (Multi-Layer Perceptrons, MLPs) бір-бірімен байланысқан жасанды нейрондық желілер, яғни қабаттарда ұйымдастырылған нейрондарды қамтиды (Сурет 19). Кірісте бір перцептронның орнына әр түрлі салмақпен бірнеше $x_1, \dots, x_n$ нейрон қосылады. Бұл нейрондардың шығысы нейрондардың жаңа қабаты үшін кіріс болып табылады.

Егер біз бәрін бірге қарастыратын болсақ, онда әр нейронның салмағы $k\left(w_k^{(i)}\right)$, L_{i-1} қабатынан L_i қабатына дейінгі $W^{(i)}$салмақ матрицасымен анықталады. L_i қабатының шығыс сигналы (векторы) $y^{(i-1)}$ кіріс векторын $W^{(i)}$ салмақ матрицасына көбейтіп, болжамды қосу арқылы анықталады және f_i сызықтық емес функциясы элементтік қолданыс ретінде есептеледі:



Сурет 20 – Көп қабатты перцептрон

$$y^{(1)} = f_1(W^{(1)}x + b^{(1)})$$

...

$$y^{(i)} = f_i(W^{(i)}x + b^{(i)}) \tag{3.5}$$

...

$$y^{(N)} = f_N(W^{N)i}x + b^{(N)})$$

Мұнда $x_1, \dots, x_n$ кірістер, $W^{(i)}$ – салмақ матрицасы, $b^{(i)}$ – i қабаты бойынша болжамдар. Нейрондық желілердің тізбектелген деңгейлерінің артықшылықтары да бар. Осы тұрған деңгей мәліметтерді алдынғы деңгейден алатын болғандықтан және кіріс мәліметтерін келесі деңгейге ұсынады, сондықтан желінің шығыс мәліметтері тізбектелген түрде бір тура жолмен есептеледі.

Соңғы деңгейдегі MLP нейрондары бірдей кіріс сипаттамаларын қолданатын сызықтық екілік жіктеуіштер болып табылады. Сондықтан, бірнеше шығысы бар MLP көп класты жіктеуіш болып табылады. Нейронды желінің шығыстары апостериорлы ықтималдықтар ретінде қарастрыуға болады [67]. Softmax функциясы (Bridle, 1990b) көбінесе шығыс деңгейіндегі сигма тәрізді функцияның орнына қолданылады [68]. $a_1, \dots, a_n$ активациямен n нейрондар үшін softmax функциясы келесідей анықталады:

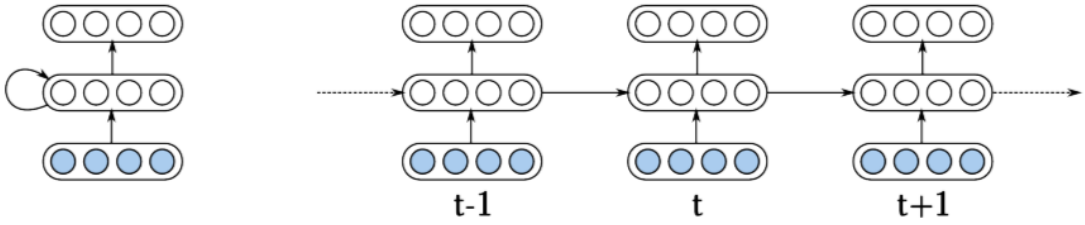
$$z_i = softmax(a_i) = \frac{e^{a_i}}{\sum_{k=1}^{n} e^{a_k}} \tag{3.6}$$

Мұнда $z_i \in [0,1]$, бұл функция арқылы z_i шығысы бірлікке жақындайды, яғни $\sum_{i=1}^{n} z_i = 1$ және желінің кірісіне байланысты әртүрлі класс бойынша

ықтималдықтар таралуын анықтайды. Бұл қасиеттің артықшылығы, біріншіден, дәйекті интерпретациямен сенімділікті бағалаумен байланысты және екіншіден, бұл желі үлкен жүйенің құрамдас бөлігі бола алады, мұнда апостериорлы ықтималдықтар маңызды.

3.2 Рекурентті нейрондық желілер

Қайталанатын нейрондық желілер (ағылш. Recurrent neural network, RNN) – элементтер арасындағы байланыстар бағытталған тізбекті құрайтын нейрондық желілердің бір түрі. Рекурентті нейрондық желілер (RNNS) - уақыт өте келе дамитын және қайталанатын байланыстар арқылы қол жеткізілетін ішкі күй ұғымы бар желілер (Сурет 21).

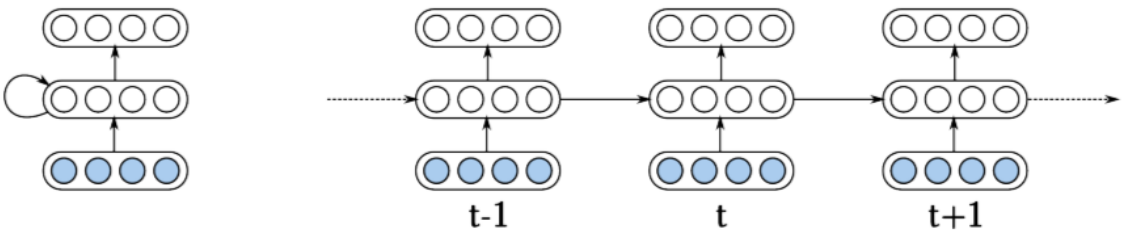


Сурет 21 – Қайталанатын нейрондық желілер, қарапайым пішін

Рекурентті нейрондық желілер (RNNs) – уақыт өте келе дамитын және қайталанатын байланыстар арқылы қол жеткізілетін ішкі күй ұғымы бар желілер. Хопфилд желілері (Хопфилд, 1982) қайталанатын нейрондық желінің ерте түрі болып табылады, мұнда қайталану уақыт тізбегін өңдеу үшін емес, тұрақты күйді табу үшін пайдаланылады. Hopfield желілері [69] қайталанатын нейрондық желінің ерте түрі болып табылады, мұнда желі қайталану уақыт тізбегін өңдеу үшін емес, тұрақты күйді табу үшін пайдаланылады.

Saul & Jordan [70], Elman[71] кіріс векторларының тізбегі үшін нейрондық желі архитектураларын ұсынды. Бұл желілерде бір жасырын деңгей, сондай-ақ желідегі жадты жүзеге асыратын бөлек контексттік деңгей бар.

Қарапайым түрінде рекурентті нейрондық желілер – бұл көп қабатты перцептрон (MLP), яғни нейрондар қайталанатын қабаттармен. әртүрлі қабаттарда ұйымдастырылады. Ағымдағы қабат тек алдыңғы қабаттан ғана емес, сонымен қатар, өзінен де кіріс алады (Сурет 22).



Сурет 22 – Қайталанатын нейрондық желілер, қарапайым пішін

Қабаттың активациясы a_k^t уақыт бойынша келесі формуламен анықталады:

$$a_k^t = \sum_{i=1}^{I} \omega_{ki}^{in} x_i^t + \sum_{h=1}^{H} \omega_{kh}^{rec} z_h^{t-1} \quad (3.7)$$

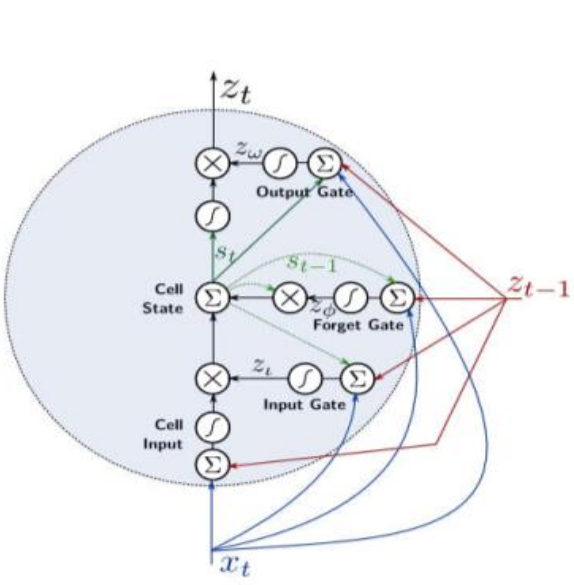
Мұнда x_i^t – кірістер, ω_{ki}^{in} – тиісті салмақтар, z_h^{t-1} – алдыңғы уақыт қадамындағы қабат шығысы және ω_{kh}^{rec} – тиісті салмақтар.

Екі бағытты рекурентті нейронды желі (Bidirectional RNNs, BRNNs) [72] екі бағытта да реттілікті өңдейді. Бұл желілерде екі қайталанатын деңгей бар: алдыңғы уақыт қадамынан кірістерді қабылдайтын тікелей деңгей және келесі уақыт қадамына қосылған кері деңгей. Екі қабат бірдей кіріс және шығыс қабаттарына қосылған. Graves және басқалар қайталанатын қабаттарда төрт бағытты кіріс кескінін өңдейтін көп өлшемді рекурентті нейронды желілерді (Multi-Dimensional RNNs , MDRNNs) құру үшін конволюциялық аспектілер қосты.

Robinson сөйлеуді тану үшін қарапайым рекурентті нейрондық желілерді ұсынды [73], Lee мен Kim [74] және Senior мен Robinson қолжазбаны тану үшін қарапайым қайталанатын нейрондық желілерді қолданды [75].

3.3 Ұзақ қысқа мерзімді жады (LSTM, Long Short-Term Memory) блоктары

Рекурентті нейронды желілерде ұзақ мерзімді тәуелділіктерді зерттеу мүмкін емес. Сондықтан, Hochreiter мен Schmidhuber ұзақ қысқа мерзімді жады (LSTM, Long Short-Term Memory) деп аталатын жетілдірілген қайталанатын нейрондарды ұсынды [76]. Ұзақ қысқа мерзімді жадыда ақпараттар ағынын ағымдағы жағдайда кіріс ақпаратын масштабтайтын, кіріс және қайталанатын ақпарат және ұяшықтағы ішкі жағдай негізінде алдынға уақыт қадамындағы шығыс активациясын іске асыратын шлюздік жүйе басқарады. Ұзақ қысқа мерзімді жады ұяшығы көрсетілген (Сурет 23).



Сурет 23 – Ұзақ қысқа мерзімді жады (LSTM, Long Short-Term Memory)

Ұяшық кірісі мен барлық шлюздар активацияны төменгі деңгейден және уақыт бойынша қадамның алдынғы қабатынан алады. Келесі теңдеу (3.8) LSTM модулінің жұмысын анықтайды. Кіріс элементі ұяшық кіріс сигналының ұяшық күйіне интегралданғанын басқарады:

$$a_l^t=\sum_{i=1}^{I}\omega_{il}x_i^t+\sum_{h=1}^{H}\omega_{hl}z_h^{t-1}+\sum_{c=1}^{C}\omega_{cl}s_c^{t-1} \quad (3.8)$$

$$z_l^t=f(a_l^t)$$

Ұмыту элементі алдыңғы күйдің ұяшық күйіне біріктірілгенін немесе ұмытылғанын басқарады.

$$a_\Phi^t=\sum_{i=1}^{I}\omega_{i\Phi}x_i^t+\sum_{h=1}^{H}\omega_{h\Phi}z_h^{t-1}+\sum_{c=1}^{C}\omega_{c\Phi}s_c^{t-1} \quad (3.9)$$

$$z_\Phi^t=f(a_\Phi^t)$$

Ұяшық күйі – ұмыту элементімен масштабталатын алдынға жағдай қосындысы және кіріс элементімен масштабталатын ұяшықтардың кіріс элементтерінің қосындысы.

$$a_c^t=\sum_{i=1}^{I}\omega_{ic}x_i^t+\sum_{h=1}^{H}\omega_{hc}z_h^{t-1} \quad (3.10)$$

$$s_c^t=z_\Phi^t s_c^{t-1}+z_l^t g(a_c^t)$$

Шығыс элементі LSTM блогының $h(s_c^t)$ белсендіру сигналын беретіндігін басқарады

$$a_\omega^t=\sum_{i=1}^{I}\omega_{i\omega}x_i^t+\sum_{h=1}^{H}\omega_{h\omega}z_h^{t-1}+\sum_{c=1}^{C}\omega_{c\omega}s_c^t \quad (3.11)$$

$$z_\omega^t=f(a_\omega^t)$$

Ұяшықтың шығыс мәліметтері h белсендіру функциясын шығыс элементімен масштабталатын ұяшық күйіне қолдану арқылы есептеледі:

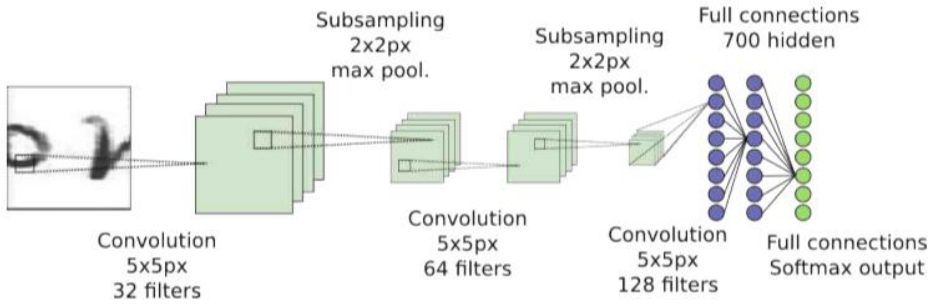
$$z_c^t=z_\omega^t h(s_c^t) \quad (3.12)$$

Шлюз үшін f белсендіру функциясы әдетте сигма тәрізді функция болып табылады. Doetsch және басқалар [77] әр түрдегі барлық шлюздерге ортақ активтендіру масштабтау коэффициентін ұсынады: $z_l^t = f(\lambda_l a_l^t), z_\omega^t = f(\lambda_\omega a_\omega^t), z_\Phi^t = f(\lambda_\phi a_\phi^t)$. Бұл факторлар зерттеліп, элементті таңдаулы етеді. Белсендіру функциялары g және h әдетте сигма тәрізді немесе гиперболалық тангенс функциясы болып табылады. Шлюз активацияларының a_l^t, a_ϕ^t және a_ω^t ұяшық күйіне тәуелділігі «көру байланысы» деп аталатын ұяшықтарға сәйкес келеді, бұл нақты уақытпен жұмыс жасағанда маңызы зор [78].

Қайталанатын қабаттардағы ұзақ мерзімді қысқа мерзімді есте сақтау нейрондарының арқасында екі бағытты (Bidirectional) және көп өлшемді (Multi-Dimensional) рекуретті нейронды желілер қолжазба мәтінін тануда өте жақсы нәтижелерге қол жеткізеді және осы саладағы ең заманауи шешімді ұсынады.

3.4 Конволюциялық нейрондық желілер

Конволюциялық нейрондық желілер көп қабатты перцептронға ұқсас құрылымға ие, бірақ әр қабат алдыңғы қабатпен толық байланыспайды. Конволюциялық нейрондық желілер локальді байланыстар мен салмақты бөлу арқылы жергілікті рецепторлар тұжырымдамасын жүзеге асырады. Кіріс екі өлшемді кескін болып табылады және жасырын қабаттардағы нейрондар әрқайсысы бір нысанды іздейтін екі өлшемді карталардан тұрады. Әрбір карта нейроны алдынғы карталармен бірдей байланыс салмағымен шағын аймақта байланысады (Сурет 24).



Сурет 24 – Конволюциялық нейрондық желіге мысал [79]

Мұны оқытылатын сүзгілері бар кескін конволюциялары ретінде түсіндіруге болады. Конволюциялық қабаттар көбінесе желінің көлемін шектеу және бұрмалануға кейбір инварианттылықты жүзеге асыру үшін кіші үлгі (subsampling) қабаттарымен жалғасады. Кіші үлгі (subsampling) операциялары – шағын аймақ бойынша максималды біріктіру (max-pooling) немесе орташалау. Әдетте төменгі қабаттарда бірнеше нысандар (немесе карталар немесе оқытылатын сүзгілер) алынады, ал жоғарғы қабаттарда көбірек нысандар алынады, ал карта өлшемдері кішірейеді.

Жіктеуге арналған конволюциялық нейрондық желілерде кескін айнымалыларын оңай кеңейтуге болады, бұл жергілікті байланыстар мен жалпы салмақтардың арқасында болжамдар тізбегін алуға мүмкіндік береді.

Осындай құрылым кеңістіктік қозғалысты нейрондық желісі (Space-Displacement Neural Network, SDNN) деп аталады. Конволюциялық нейрондық желілер күрделі архитектураларда қолданылуы мүмкін.

Қолжазбаны тану мәселесін шешу үшін нейрондық желілердің әртүрлі түрлері қолданылды:

- көп қабатты перцептрондарға негізделген әдістер;
- конволюциялық нейрондық желілерге негізделген әдістер;
- рекурентті нейронды желілерге негізделген әдістер.

Қолжазбаны тануға арналған терең нейрондық желілердің қосымшалары конволюциялық архитектуралармен шектеледі, мұнда параметрлер саны жергілікті рецепторлармен және салмақтың таралу аспектілерімен шектеледі, сонымен қатар төменгі деңгейлерде бірнеше белгілерді шығарады. Бір немесе екі жасырын қабаты бар тығыз байланысқан нейрондық желілер қолжазбаны тану кезінде қолданылады, бірақ оқшауланған таңбаларды танумен шектеледі.

3.5 Нейрондық желіні оқыту

Нейрондық желіні оқыту оның параметрлерін, қосылыс салмағын модель тапсырманы орындай алатындай етіп конфигурациялаудан тұрады. Нейрондық желіні оқыту – кіріс мәлімететтері бар оңтайландыру есебі, яғни шығыс мәліметтері класификацияға сәйкес келуі үшін нейронды желі бірнеше қарапайым есептерулерді жүзеге асырады. Нейрондық желі сапасын бағалау үшін $y(x)$ нақты нәтижені z күтілетін нәтижемен салыстырып, қателік өлшемді есептеуге болады. Бұл $(y(x) - z)^2$ екілік жіктеу үшін квадраттық қате болуы мүмкін. Көп класты жіктеу үшін шығыс мәліметтері $y(x)$ векторын құрайды. Күтілетін шығыс сигналы z векторы ретінде кодталады, бұл жағдайда тек ақиқат класқа сәйкес келетін компоненттен басқа барлық компоненттер 0-ге тең, ал қате квадраты былай өрнектеледі:

$$E_{MSE} = \|y(x) - z\|^2 \tag{3.13}$$

Мұнда y компоненттері апостериорлы ықтималдақтар. Сонымен қатар, softmax қолданған кезде шығыс мәліметтері оң болады және бір-біріне қосылады, сондықтан кросс-энтропия (Cross-Entropy, CE) критерийін қолдануға болады:

$$E_{CE}(x, z) = -\sum_{i=1}^{C} z_i log y_i(x) \tag{3.14}$$

$S = \left(x^{(i)}, z^{(i)}\right), i = 1,2, \dots, N$ күтілетін шығыс мәліметтері екенін ескере отырып, глобальды шығын функциясын есептеуге болады:

$$E(S) = \frac{1}{N}\sum_{i=1}^{N} E(x^{(i)}, z^{(i)}) \quad (3.15)$$

Қателік функциясын азайту арқылы $\theta = \theta_1, \dots \theta_N$ оңтайлы параметрлерді аламыз. Осы мақсатқа жету үшін градиентті түсіру сандық оңтайландыру әдістері қолданылады. Градиенттің түсуі келесі қадамдардан тұрады:

1. Әрбір оқу $\left(x^{(i)}, z^{(i)}\right)$мысалы үшін $y(x^{(i)})$ және $E\left(x^{(i)}, z^{(i)}\right)$ есептеледі.
2. *E(S)* қатесі және оның туындысы $\frac{\partial E}{\partial \theta_i}$ параметрлер бойынша есептеледі
3. Параметрлерді градиент бағыты бойынша қайта есептеу керек $\theta_i \leftarrow \theta_i - \eta \frac{\partial E}{\partial \theta_i}$, мұнда η оқыту жылдамдығы.

Қатені есептеу үшін параметрлерді конфигурациялауда бір қадам жасау керек, бұл барлық мәліметтер жиынтығын қарастыруға әкеледі және баяу жинақталуы мүмкін. Стохастикалық градиенттің түсуінде (Stochastic gradient descent, SGD) параметрлер әр оқу мысалынан кейін немесе бірнеше мысалдар арқылы жаңартылып, параметрлер кеңістігін жақсырақ зерттеуге және пайдалануға мүмкіндік береді.

Импульс әдісі [80] алдыңғы жаңартудың бір бөлігін ағымдағы жаңартуға қосу болып табылады:

$$\Delta\theta(t) = -\eta \frac{\partial E}{\partial \theta} + \rho\Delta\theta(t-1) \quad (3.16)$$

Оқу жылдамдығын таңдау өте маңызды. Оқу жылдамдығын тұрақты ұстаудың орнына оқыту кезінде оны өзгертуге болады. Оқу жылдамдығы қарқынының графигін алдын-ала анықтауға немесе азайтуға болады. Оқытылатын жиынтықта қатені азайтумен қатар, қайта оқытуды болдырмау қажет. Жалпылау қабілетін басқарудың бір жолы – тексеру жинағын оқу жинағынан бөлек ұстау, тексеру жинағы негізінде қатені де есептеуге болады. Ерте тоқтату әдісі – бұл тексеру деректеріндегі қате көбейген кезде оқу процедурасын тоқтату.

Көптеген жасырын қабаттары бар нейрондық желілер мидың иерархиялық ұйымына ұқсастығына байланысты тартымды. Қарапайым белгілер терең нейрондық желілердің төменгі деңгейлерінен алынады және күрделі құрылымдарды тану үшін жоғары деңгейлерде біріктіріледі. Терең нейрондық желілер қазіргі уақытта көптеген машиналық оқыту қосымшаларында, соның ішінде компьютерлік көру, сөйлеуді тану және табиғи тілді өңдеуде өте танымал. Мұндай модельдермен қателік деңгейі күрт төмендейді.

3.6 Байланысты орнатуды уақытша қолдайтын классификациясы

Байланысты орнатуды уақытша қолдайтын классификациясы (Connectionist Temporal Classification, CTC) ұғымын Graves және басқалар ұсынды [81] және нейрондық желілер арқылы сегменттелмеген деректерді таңбалау міндетін іске асырады. Байланысты орнатуды уақытша қолдайтын классификациясының идеясы кіріс тізбегіне қолданылатын нейрондық желінің шығысы тікелей қажет таңбаларының тізбегі болып табылады, біздің жағдайда таңбалар тізбегі. Ұсынылған әдістің негізгі артықшылықтары біріншіден, оқыту деректерін алдын-ала сегменттеудің қажеті жоқ, екіншіден, шығыс мәліметтері ешқандай өңдеуден кейінгі өңдеуді қажет етпейді, бірден таңбалар тізбегі алынады, ал әдетте нейрондық желілер болжам жасайды декодтауды қажет етеді. Бұл табалар тізімін алу үшін бірнеше артефактілер қажет. Кіріс тізбегінің белгілі бір ұзындығы T = |x| болу керек. Болжау тізбегінің ұзындығы (softmax-тан кейін) да T-ға тең болады. Ал шығыс тізбегінің ұзындығы |z| ≤ T. Осылай, болжау тізбегінің ұзындығы (softmax-тан кейін) өлшемге де байланысты болады, ал күтілетін шығыс тізбегінің ұзындығы әдетте |z| = T-ден аз болады (Сурет 24).

Егер күрделі кейінгі өңдеуді қажет етпесе, желіні таңбаларды тікелей болжауға, содан кейін болжамдарды қарауға мәжбүрлеу оңай жол болып табылады. Шығыс мәліметтері көшірмелерді жою арқылы алынады, мысалы $AAAABBB \rightarrow AB$. Екі тізбек элементтері бірдей болғанды қиындық туындайды, мысалы AAB екенін болжау қажет. Осы мәселеге байланысты Graves және басқалар бос таңба ∅ ұғымын енгізді. Байланысты орнатуды уақытша қолдайтын классификациясы желісінде L алфавитінде әрбәр таңбаға бір орын және бір бос орын бар, яғни шығыс алфавиті келесі түрде болады:

$$L' = L \cup \{\varnothing\} \tag{3.17}$$

Картаға түсіру B: $L'^{T} \rightarrow L^{\leq T}$ анықталады, ол қайталануларды, содан кейін желіні болжаудағы бос орындарды жояды. Мысалы For example: B(∅AA∅∅B) = B(AAA∅BB) = AB.

Кіріс деректерін ескере отырып, әр түрлі уақыт аралықтары үшін желінің шығысы тәуелсіз болған жағдайда (өйткені шығыс деңгейі мен аралық деңгейлер арасында байланыс жоқ), берілген x үшін RNN шығыс деректері бойынша $\pi \in L'^{T}$ белгілерінің реттілік ықтималдығы мына өрнекпен анықталады:

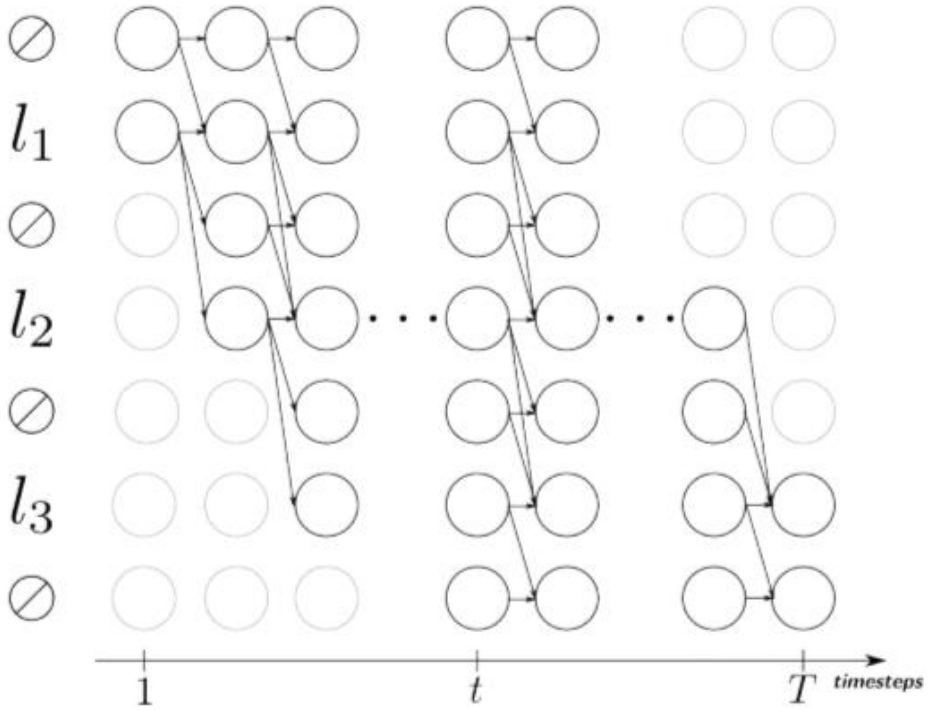
$$\rho(\pi|x) = \prod_t y_{\pi_t}^t(x) \tag{3.18}$$

Картаға түсіру B барлық мүмкін сегменттерді қосындылау арқылы белгілердің (таңбалардың) $l \in L^{\leq T}$ реттілігінің апостериорлық ықтималдығын есептеуге мүмкіндік береді:

$$\rho(l|x) = \sum_{\pi \in B^{-1}(l)} \rho(\pi|x) \tag{3.19}$$

Осы формула (3.19) негізінде желіні оқытып, шығындарды азайту және градиентті түсіру әдістерін қолдану арқылы сегменттелмеген оқыту деректерін $S = \{(x, z), z \in L^{\leq |x|}$ дұрыс таңбалау ықтималдығын барынша арттыруға болады:

$$E_{CTC} = -\sum_{(x,z)\in S} \log \quad p(z|x) \tag{3.20}$$



Сурет 24 – Байланысты орнатуды уақытша қолдайтын классификациясы (CTC) графы

$p(z|x)$ есептеуі $B^{-1}(z)$-гі барлық жолдардың қосындысын білдіреді, олардың әрқайсысының ұзындығы T = |x|. Graves және басқалар таңбалаудың барлық мүмкін нұсқаларын көрсететін «алға-артқа» алгоритмін қолдануды ұсынады (Сурет 3.8). Тура (сәйкес кері) айнымалылар префикстердің (сәйкес суффикстердің) ықтималдығын білдіреді. Олар келесі формуламен анықталады:

$$\alpha_t(s) = \rho(\pi_{l:t}: B(\pi_{l:t}) = l_{1:s/2}, \pi_t = l'_s | x = \sum_{\pi:B(\pi_{1:t})=l_{1:s/2}} \prod_{t'=1}^{t} y_{\pi_{t'}}^{t'} \tag{3.21}$$

$$\beta_t(s) = \rho(\pi_{t+1:T}: B(\pi_{l:T}) = l_{\frac{s}{2}:|l|}, \pi_t = l'_s | x = \sum_{\pi:B(\pi_{1:t})=l_{s/2:|l|}} \prod_{t'=t+1}^{T} y_{\pi_{t'}}^{t'} \tag{3.22}$$

Мұнда $l \in L^{\leq T}$ қарастырылып отырған белгілер тізбегі және $l' \in L'^{T}$ l-дан алынған,бос орындармен толықтырылған, ұзындық $|l'| = 2|l|+1$. Бос орын міндетті емес болғандықтан және көшулер B картасын жасауға рұқсат етілгендіктен, біз қайталау аламыз:

$$\alpha_t(s) = \begin{cases} \text{егер } l'_s = \varnothing \text{ немесе } l'_s = l'_{s-2} \text{ болса, онда } y_{l'_s}^{t} \sum_{n=0}^{1} \alpha_{t-1}(s-n) \\ \text{болмаса, } y_{l'_s}^{t} \sum_{n=0}^{2} \alpha_{t-1}(s-n) \end{cases} \tag{3.23}$$

Тура айнымалылар үшін

$$\beta_t(s) = \begin{cases} \text{егер } l'_s = \varnothing \text{ немесе } l'_s = l'_{s-2} \text{ болса } y^{t+1}_{l'_{s+1}}\beta_{t+1}(s+1) + y^{t+1}_{l'_s}\beta_{t+1}(s) \\ \text{басқа жағдайда } y^{t+1}_{l'_{s+1}}(s+1) + y^{t+1}_{l'_{s+1}}\beta_{t+1}(s) + y^{t+1}_{l'_{s+2}}\beta_{t+1}(s) \end{cases} \quad (3.24)$$

Таңбалау ықтималдығы l тура айнымалылармен беріледі:

$$\rho(l|\text{x}) = \alpha_T(|l'|) + \alpha_T(|l'|-1) \quad (3.25)$$

және кез келген t уақытта αβ көбейтіндісі:

$$\rho(l|x) = \sum_{s=1}^{|l'|} \alpha_t(s)\beta_t(\text{s}) \quad (3.26)$$

бұл классикалық "алға-артқа" нәтижелері. Желінің шығым мәліметтеріне қатысты құнның туындылары:

$$\frac{\partial E}{\partial y_k^t} = -\frac{1}{p(Z|X)}\frac{\partial p(Z|X)}{\partial y_k^t} = -\frac{1}{\sum_{s'=1}^{|z'|}\alpha_t(s')\beta_t(s')}\frac{\partial\left(\sum_{s=1}^{|z'|}\alpha_t(s)\beta_t(s)\right)}{\partial y_k^t} \quad (3.27)$$

Қарапайым математикалық талдау арқылы біз мынаны көрсете аламыз:

$$\frac{\partial p(Z|X)}{\partial y_k^t} = \frac{1}{y_k^t}\sum_{s\in lab(z,k)}\alpha_t(s)\beta_t(s) \quad (3.28)$$

Мұнда $lab(z,k) = \{s: z'_s = k\}$.

Осылайша, бізде softmax-қа дейінгі активацияларға қатысты келесі туындылар бар:

$$\frac{\partial E}{\partial a_k^t} = y_k^t - \sum_{s\in lab(z,k)}\frac{\alpha_t(s)\beta_t(s)}{\sum_{s'}\alpha_t(s')\beta_t(s')} \quad (3.29)$$

Қолжазбаны тану туралы соңғы басылымдарда нейрондық желілер бастапқы жүктеу процедурасы арқылы кадр бойынша оқытылады немесе олар

RNNS болып табылады және CTC арқылы оқытылады. Бір жағынан, CTC теңдеулері тікелей және кері оқыту теңдеулеріне ұқсас және CTC оқытқан RNN өте жақсы нәтижелерге қол жеткізеді.

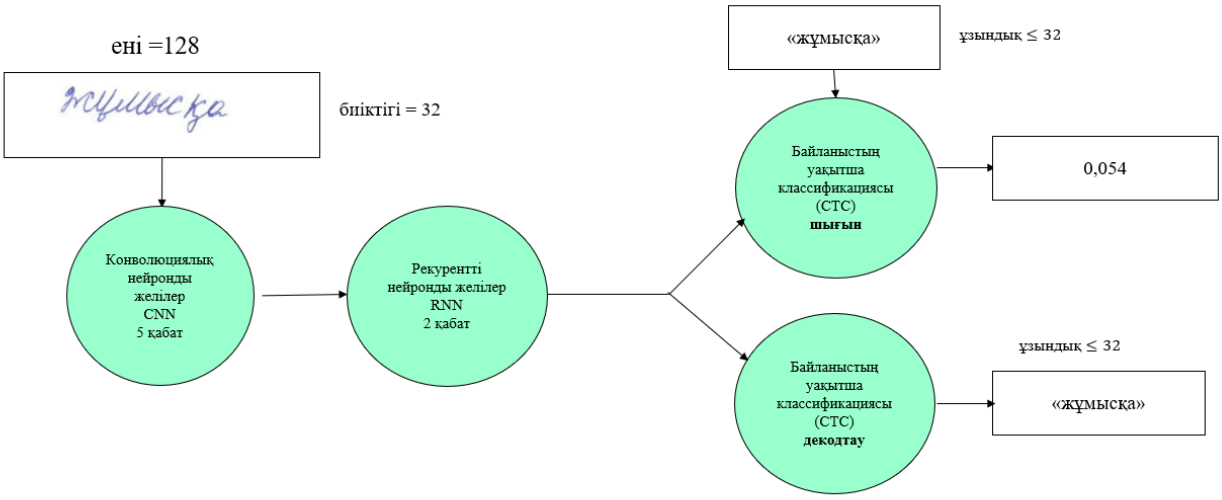
Кириллица графикасына негізделген қазақ және орыс тіліндегі қолжазбаны тануға рекурентті нейронды желілерді қолдану қарастырылды [82]. Қолжазбаны тану жүйелерін оқыту және сынау үшін қазақ және орыс тіліндегі HKR(Handwritten Kazakh & Russian) және KOHTD (Kazakh Offline Handwritten Text Dataset) қолжазба деректер жиыны алынды. Ұсынылған модель Python үшін TensorFlow кітапханасын қолдана отырып жүзеге асырылды.

4 ҚОЛЖАЗБА МӘТІНДЕРДІ ТАНУҒА ҰСЫНЫЛҒАН МОДЕЛЬДЕР ЖӘНЕ НӘТИЖЕЛЕР

Соңғы жылдары машиналық оқыту мен жасанды интеллекттің дамуы, әсіресе терең оқыту (deep learning) әдістерінің, оның ішінде конволюциялық нейрондық желілер (Convolutional Neural Networks, CNN) және қайталанбалы нейрондық желілердің (Recurrent Neural Networks, RNN) қолданылуы, HTR жүйелерінің тану дәлдігін едәуір арттырды. Бұл әдістер күрделі графикалық құрылымы бар қолжазбаларды тиімді талдауға мүмкіндік береді және мәтіннің мазмұнын автоматты түрде тану мен түсінуді жетілдіреді. Қазақ тілінде жазылған қолжазбаларды тану — ерекше назар аударуды қажет ететін мәселе, себебі қазақ графикасының фонетикалық және морфологиялық ерекшеліктері HTR жүйелерінің бейімделуін талап етеді. Сондықтан бұл зерттеу HTR жүйелерінің көптілді ортада, әсіресе қазақ тіліне бейімделу мүмкіндіктерін зерттеу арқылы тілдік әртүрлілікке бейімделген әмбебап шешімдерге жол ашады.

4.1 CNN-LSTM-CTC әдісі негізіндегі Scheidl моделі

Қазақ-орыс тіліндегі оффлайн қолжазбаны тану жүйелері сканерленген суреттердегі мәтінді сандық мәтінге түрлендіреді. Қазақ-орыс тіліндегі оффлайн қолжазбаны тану үшін Harald [83] ұсынған SimpleHTR моделі пайдаланылды. Ұсынылған модель кіріс фотосуретінен нысандарды алу үшін көптеген конволюциялық нейрондық желілердің (CNNs) қабаттары пайдаланылатын жасанды нейронды желілерді (ANNs) пайдаланады. Содан кейін бұл қабаттардың шығысы рекурентті нейронды желілерге (RNN-ге) жіберіледі. Рекурентті нейронды желілер ақпаратты реттілік арқылы таратады. Рекурентті нейронды желілер шығысы кезектегі әрбір таңба үшін ықтималдықтарды есептейді. Соңғы мәтінді болжау үшін декодтау алгоритмдері рекурентті нейронды желілер шығысында жүзеге асырылады. Байланысты орнатуды уақытша қолдайтын классификация (CTC) функциялары ықтималдықтарды соңғы мәтінге декодтауға жауап береді (Сурет 25).



Сурет 25– SimpleHTR моделі

Нейронды желіні (NN) формальды түрде ұзындығы 0-ден L-ға дейінгі (c1, c2, ...) таңбалар тізбегініен тұратын W×H өлшемді суретті M матрицасына сәйкестендіретін функция ретінде қарастыруға болады (4.1).

$$\underset{Wxx}{NN}: M \rightarrow \underset{0 \leq n \leq 1}{(c_1, c_2, \ldots, c_n)} \qquad (4.1)$$

.

Мәтін таңба негізінде танылады. Сондықтан Simple моделі оқытылған мәліметтерде кездеспейтін сөздердер мен мәтіндерді тануды іске асырады (жеке таңбалар дұрыс жіктелген жағдайда).

CNN: кіріс кескіні CNN қабаттарына жіберіледі. Бұл қабаттар кескіннен сәйкес нысандарды алуға үйретілген. Әр қабат үш операциядан тұрады. Біріншіден, алғашқы екі қабатта 5×5 және соңғы үш қабатта 3×3 өлшемді сүзгі ядросын енгізуге қолданылатын конволюция операциясы. Содан кейін сызықтық емес RELU функциясы қолданылады. Соңында, біріктіру қабаты кескін аймақтарын қорытындылайды және кіріс деректерінің кішірек нұсқасын шығарады. Кескіннің биіктігі әр қабатта 2-ге азайғанымен, объектілердің Шығыс картасы (немесе реттілігі) 32×256 болатындай етіп объектілер карталары (арналар) қосылады.

RNN: Нысандар тізбегі уақыт қадамында 256 нысанды қамтиды, RNN осы реттілік арқылы тиісті ақпаратты таратады. Ұзақ мерзімді қысқа мерзімді жадымен (LSTM) танымал RNNs енгізу қолданылады, өйткені ол ақпаратты ұзақ қашықтыққа жібере алады және әдеттегі RNN-ге қарағанда сенімді оқу сипаттамаларын ұсынады. RNN Шығыс тізбегі 32×80 матрицаға айналады. IAM деректер жиынтығы 79 түрлі таңбадан тұрады, сонымен қатар CTC операциясы үшін бір қосымша таңба қажет (Бос CTC белгісі), сондықтан 32 уақыт қадамының әрқайсысы үшін 80 жазба бар. Ал KOHTD деректер жиынында қазақ-орыс әріптері болғандықтан матрица өлшемі өзгереді.

CTC: Нейронды желілер оқыту кезінде CTC рекурентті нейронды желіден шығыс матрицасын және негізгі ақиқат мәтінін алады және шығынның мәнін есептейді. Шығару кезінде CTC тек матрицаны алады және ол оны соңғы мәтінге декодтайды. Негізгі мәтіннің де, танылған мәтіннің де ұзындығы 32 таңбадан аспауы қажет.

Мәліметтер

Кіріс суреттерін KOHTD мәліметтер жиынтығынан алуға болады (Сурет 26).



Сурет 26 – KOHTD мәліметтер жиынтығынан алынған қолжазба мәтін

Кіріс: бұл 128×32 сұр түсті сурет. Әдетте, деректер жиынтығындағы кескіндер дәл осындай өлшемге ие емес, сондықтан біз олардың өлшемін (бұрмаланбай) ені 128 немесе биіктігі 32 болғанша өзгертіледі. Содан кейін біз суретті 128×32 өлшемді (ақ) мақсатты кескінге көшіреміз. Бұл процесс көрсетілген (Сурет 27).Соңында, біз сұр кескіннің мәндерін қалыпқа келтіреміз, бұл нейронды желі үшін жағдайды жеңілдетеді. Деректерді үлкейту кескінді сол жаққа туралаудың орнына ерікті позицияларға көшіру немесе кескіннің өлшемін кездейсоқ өзгерту арқылы оңай біріктірілуі мүмкін.



Сурет 27 – 128×32 өлшемді мақсатты кескінге сәйкес масштабталуы

Еркін өлшемдегі KOHTD деректер жиынтығынан алынған сурет (Сурет 26) әр түрлі өлшемде болады. Ол 128×32 өлшемді мақсатты кескінге сәйкес масштабталады, мақсатты кескіннің бос бөлігі ақ түспен толтырылады.

CNN шығысы: сурет. Ұзындығы 32 тізбегі болып табылатын CNN қабаттарының шығысы. Әр жазбада 256 нысан бар. Әрине, бұл нысандар одан әрі RNN қабаттарымен өңделеді, дегенмен кейбір нысандар кіріс кескінінің белгілі бір жоғары деңгейлі қасиеттерімен жоғары корреляциялы көрсеткіштерге ие болады.

RNN шығысы: мәтіні бар кескін үшін RNN Шығыс матрицасының мәндері сәйкестендіріледі. Әрбір таңбаларға мәндер сәйкестендірілген матрица құрылады:
()*+,./0123456789:;=>?ABCDEFGHIKLMNOPRSTVWX[]_abcdefghijklmnopqrstu

vwxyz|}ӘәІАБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЫЬЭЮЯабвгдежзийклмнопрстуфхцчшщъыьэюяёіҒғҚқҢңҮүҰұһӘәӨө——,”„•… 《

Әр уақыт қадамынан ең ықтимал таңба алынады, бұл ең жақсы жол деп аталады, содан кейін қайталанатын таңбалар алып тасталынады.

SimpleHTR моделі арқылы қолжазбаны тануды TensorFlow көмегімен іске асыру 4 модульден тұрады:

1. SamplePreprocessor.py: Нейронды желі үшін KOHTD деректер жиынтығынан кескіндер дайындайды.

2. DataLoader.py : үлгілерді оқиды, оларды пакеттерге бөледі және деректерді қарау үшін интерфейс жасау

3. Model.py : жоғарыда сипатталғандай модель жасайды, модельдерді жүктейді және сақтайды, TensorFlow сеанстарын басқарады және оқыту мен логикалық шығыс интерфейсін ұсынады

4. main.py : бұрын аталған барлық модульдерді біріктіреді

Басқа бастапқы файлдар негізгі енгізу-шығару файлдарымен (DataLoader.py) және кескінді өңдеумен (SamplePreprocessor.py) байланысты болғандықтан, Model.py файлы қарастырылады.

CNN: Әрбір CNN қабаты үшін конволюция операциясында қолданылатын k×k өлшемді ядро жасалады.

```
kernel = tf.Variable(tf.truncated_normal([k, k, chIn, chOut], stddev=0.1))

conv = tf.nn.conv2d(inputTensor, kernel, padding='SAME', strides=(1, 1, 1, 1))
```

Содан кейін конволюция нәтижесін RELU операциясына жіберу керек, артынан қайтадан px×py өлшемімен және sx×sy қадам өлшемімен біріктіру қабатына жіберіледі.

```
elu = tf.nn.relu(conv)

pool = tf.nn.max_pool(relu, (1, px, py, 1), (1, sx, sy, 1), 'VALID')
```

Бұл қадамдар for цикліндегі барлық қабаттар үшін қайталанады.

RNN: Әрқайсысында 256 бірліктен тұратын екі RNN қабатын жасау керек және қабаттастыру қажет. Содан кейін одан екі бағытты RNN жасалып, кіріс тізбегі алдыңғыдан артқа және керісінше беріледі. Нәтижесінде біз 32×256 өлшемді fw және bw екі шығыс тізбегін аламыз, оларды кейінірек объектілер осі бойымен біріктіріп, 32×512 өлшемді тізбекті қалыптастырамыз. Соңында, ол CTC қабатына берілетін 32×80 шығыс тізбегіне (немесе матрицаға) сәйкес келетіріледі.

```
((fw, bw),_) = tf.nn.bidirectional_dynamic_rnn(cell_fw=stacked, cell_bw=stacked, inputs=inputTensor, dtype=inputTensor.dtype)
```

CTC: Шығындарды есептеу үшін біз операцияға негізгі ақиқат мәтінін де, матрица да енгізіледі. Ақиқаттың негізгі мәтіні сирек тензор түрінде кодталады. Кіріс тізбегінің ұзындығы CTC екі операциясына да берілуі керек.

```
    gtTexts = tf.SparseTensor(tf.placeholder(tf.int64, shape=[None, 2]),
tf.placeholder(tf.int32, [None]), tf.placeholder(tf.int64, [2]))
    seqLen = tf.placeholder(tf.int32, [None])
```

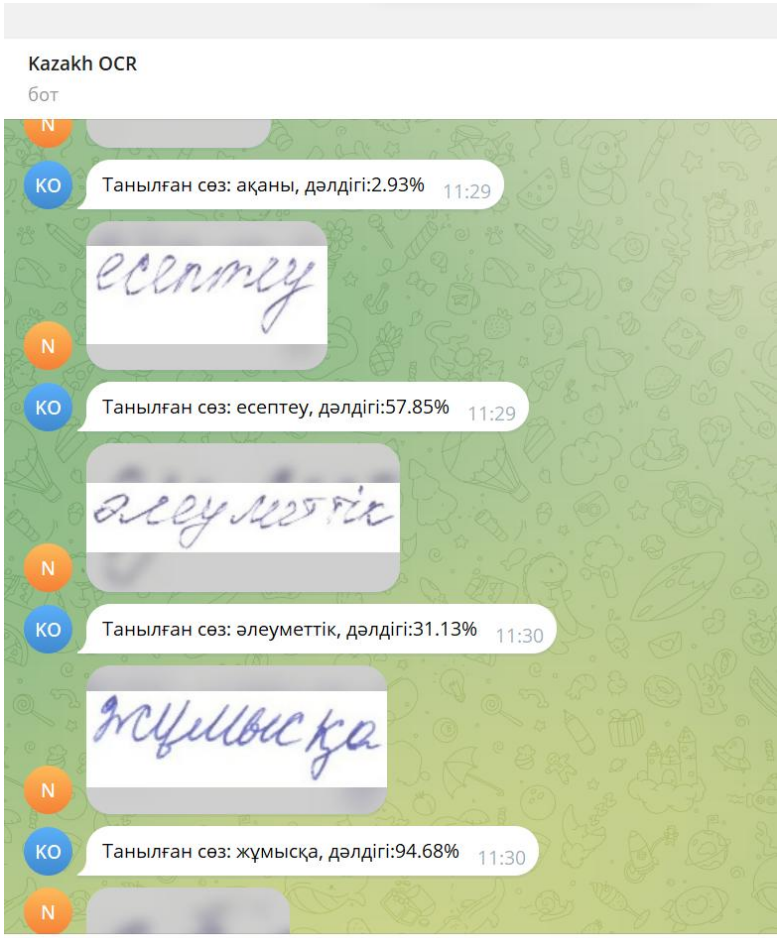
Енді жоғалту операциясын және декодтау операциясын жасау үшін барлық кірістер бар.

```
    loss = tf.nn.ctc_loss(labels=gtTexts, inputs=inputTensor,
sequence_length=seqLen, ctc_merge_repeated=True)
    decoder = tf.nn.ctc_greedy_decoder(inputs=inputTensor,
sequence_length=seqLen)
```

Оқыту: Пакет элементтерінің жоғалу мәндерінің орташа мәні NN оқыту үшін пайдаланылады: ол RMSProp оңтайландырғышқа беріледі

```
optimizer = tf.train.RMSPropOptimizer(0.001).minimize(loss)
```

Қазақ-орыс тіліндегі оффлайн қолжазбаны тану жүйелері сканерленген суреттердегі мәтінді сандық мәтінге түрлендіру үшін интерфейс телеграм ботпен байланыстырылды (Сурет 28).



Сурет 28 – SimpleHTR моделі негізінде қолжазба мәтіндерді тану

Қолжазба мәтіндердің тану дәлдігін арттыру үшін келесі әрекеттерді іске асыру қажет:

– Деректер жиынтығын үлкейту: кіріс кескіндеріне қосымша (кездейсоқ) түрлендірулерді қолдану арқылы деректер жиынының көлемін ұлғайтуға болады

- Кіріс суреттеріндегі курсивті жазу мәнерін жою керек
- Кіріс көлемін ұлғайту
- Қосымша CNN қабаттарын қосу
- LSTM-ді 2D-LSTM - ге ауыстыру
- Мәтінді түзету: егер танылған сөз сөздікте болмаса, ең ұқсасын іздеу керек.

Суреттердегі мәтінді тануға қабілетті нейронды желі қарастырылды. талқыладық. Нейронды желі 5 CNN қабатынан және 2 RNN қабатынан тұрады және таңбалардың ықтималдық матрицасын шығарады. Бұл матрица CTC шығынын есептеу үшін немесе CTC декодтау үшін қолданылады. TF көмегімен іске асыру енгізілді, сонымен қатар кодтың кейбір маңызды бөліктері ұсынылды. Соңында, тану дәлдігін жақсарту бойынша ұсыныстар берілді.

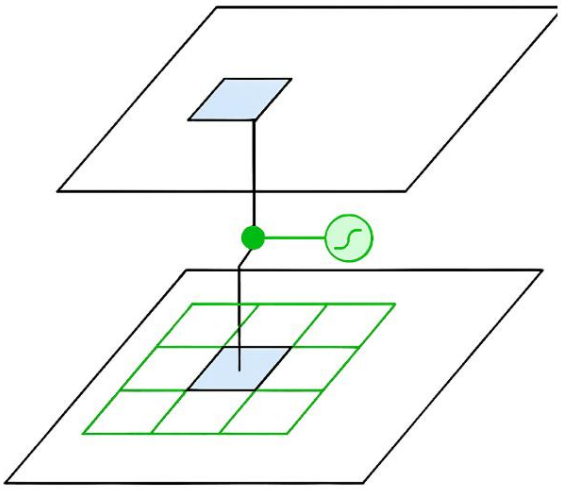
4.2 Gated-CRNN әдісі негізінде Bluche моделі

Bluche T. мен Messina R. ғылыми жұмыстарында негізгі үш бөліктен тұратын терең нейронды желіні құрған [84]. Бірінші бөлімде конволюциялық қабаттар кіріс кескіндерінің кодерінен тұрады. Бұл кодерлер екі өлшемді көріністермен жұмыс істейді және объектілердің 2D карталарын ұсынады. Бұл бөлікте модельдің шамамен 20% бос параметрлері бар, бірақ олардың архитектурасының ең баяу құрамдас бөлігі болып табылады. Мақсат – өңдеу уақытын ұзарту үшін оны қайта пайдаланып, әмбебап ету. Екінші бөлім – 2D кескінін 1D көрінісіне түрлендіретін интерфейс. Үшінші бөлім – таңбалар тізбегін болжау үшін белгілер тізбегін өңдейтін ұзақ қысқа мерзімді жадымен екі бағытты рекурентті нейроды желі (LSTM RNN) декодері. Бұл бөлім желінің ең үлкен өткізу қабілеттілігіне ие, шамамен 80%, бірақ өңдеу жылдамдығы жоғары. Модель құру барысында басқа салалардағы зерттеу тәжірибесін пайдаланылды, функциялар қайта қолданысқа енгізілді, оқыту графикалық процессорда іске асты. Графикалық процессормен үйлесімді жақсы модельдерді табу модельді дамытудың алғышарты болды.

Құрылған қолжазбаны тану белгілі бір тілге тән мәліметтер негізінде оқытылады. Мысалы, француз нейрондық желісі үшін қол жетімді деректерді тек француз тілінде қолданылады. Ресурстары төмен тілдер үшін біз басқа тілде оқытылған нейрондық желіні баптауға болады.. Дегенмен, бұл тәсілдің күрделі мәселесі бар: бізде бір тілге арналған деректер көп болмаса, онда тек бір ғана құжаттар жинағы бар дегенді білдіреді. Осылайша, модельде жаңа тілге бейімдеп қана қоймай, оны белгілі бір коллекцияға бейімдеуге болады. Модельде алдымен желіні барлық қол жетімді деректер негізінде оқытылған. Сөйтіп, кодер әртүрлі жинақтарда жақсы функцияларды шығаруы керек. Келесі қадамда тек әр тіл үшін декодер бапталады. Бұл тәсілдің бірқатар артықшылықтары бар. Біріншіден, кодер есептеу уақытының көп бөлігін алады. Тек декодерді бейімдеу өте жылдам оқуға әкеледі. Модельде ортақ кодер болғандықтан, ол белгілі бір тілге немесе жинаққа маманданбайды, сондықтан салыстырмалы түрде ортақ мүмкіндіктерге үміттенуге болады. Сонымен қатар, модель қазір кеңейіп келеді, өйткені оның бірінші бөлігі әр түрлі тілдерге ортақ, бұл сақтау көлемін үнемдейді. Танылатын

тіл алдын-ала белгісіз болуы мүмкін. Модельде әмбебап желіні оқытылғандықтан, оны осындай белгісіз жағдайларда қолдануға болады.

Кодерде конволюциялық элемент қолданылады (Сурет 29). Басқару элементі объектінің келесі қабатқа таралуын басқарады.



Сурет 29 – Конволюциялық элемент

Негізінде, элемент берілген позициядағы объектінің мәнін және іргелес мәндерді қарастырады, кейіннен сол нысанды сол күйде сақтау немесе одан бас тарту туралы шешім қабылдайды. Бұл бүкіл суреттегі жалпы белгілерді есептеуге және контекстке сәйкес бұл белгілер маңызды болған кезде сүзуге мүмкіндік береді.

Нүктелік механизмнің шығу сигналы кіріс сигналын шығыс сигналына нүктелік көбейтіндісіне тең:

$$y = g(x) \cdot (x) \tag{4.2}$$

мұнда

$$g(x_{ij}) = \sigma(\omega_{00}x_{i-1,j-1} + \omega_{01}x_{i-1,j} + \omega_{02}x_{i-1,j+1} + \omega_{10}x_{i,j-1} + \omega_{11}x_{i,j} + \omega_{12}x_{i,j+1} + \omega_{20}x_{i+1,j-1} + \omega_{21}x_{i+1,j} + \omega_{22}x_{i+1,j+1}) \tag{4.3}$$

Бұл LSTMs ішіндегі кіріс немесе шығыс элементтеріне ұқсайды, тек контекст іргелес позициялардағы қабаттың шығуынан емес, кіріс кеңістігінде болады. Басқару элементі көршілес позициялардағы объектілердің бүтін векторларына сүйене отырып, әр объект үшін [0, 1] мәнін есептейді (яғни әр позициядағы объектілердің векторларын өлшеу үшін). Осылайша, берілген мәннің өзектілігі туралы шешім қабылдау үшін ескерілген контекст салыстырмалы түрде өте көп.

Bluche T. мен Messina R. Ұсынған модельді оқыту үшін әртүрлі көздерден мүмкіндігінше көп қолжазба деректер жиынын қолданды. Жалпы, жеті тілде

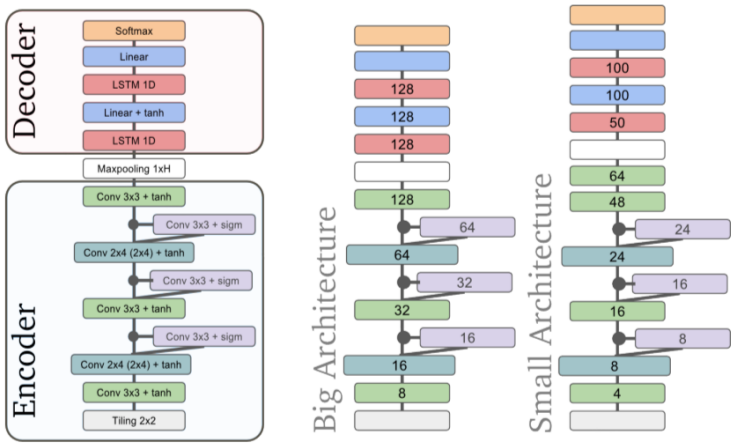
мәтін жолдарының шамамен 133 мың суреті бар, оның ішінде жеке деректердің 30% (негізінен орыс тілінде). Қолжазба деректер жиыны келесі кестеде көрсетіледі (Кесте 7).

Кесте 7 – Көптілді жүйеге арналған оқыту қолжазба деректер жиыны

Тіл	Қолжазба деректер жиыны	Жолдар саны
Ағылшын	Maurdora[a]	32475
	IAM	6482
	IBM†	4910
	Private.EN*	9404
Француз	Rimes	10532
	Maurdor	26870
Испан	Germana	11 505
	Cristo Salvador	971
	Private.SP*	452
Неміс	Private.DE†*	6,939
Итальян	Private.IT*	2,261
Португал	Private.PT*	870
Орыс	Private.RU*	19,312
Барлығы		132983

Мұнда * белгісімен белгіленген деректер жиынтығы жеке деректерге сәйкес келеді және жолдардың жалпы санының 30% құрайды. Ал † белгісімен белгіленген деректер жиынтығы кескіндерді транскрипциялармен туралау арқылы автоматты түрде белгіленді және қателерді қамтиды және *a* белгісімен белгіленген деректер жиынтығы сызықтық деформациялармен толықтырылды.

Құрылған модель нейронды желілерге негізделген. Әр түрлі тапсырмаларды (дәлдік, жылдамдық және өлшем) оңтайландыру үшін біз екі архитектураны жасалды. Олардың екеуі де конволюциялық кодерден, максималды тік өлшеу бірлестігінен және қайталанатын декодерден тұрады (Сурет 30).



Сурет 30 – Ұсынылған нейрондық желілердің архитектурасы

Бұл архитектура Bluche T. мен Messina R. ғылыми жұмыстарында келтірілген.

Кодер 3x3 конволюциялық қабаттан тұрады, 8 (сәйкес. 4) Нысандар, 16-дан 2x4 өлшемді конволюциялық қабат (сәйкес. 8) элементтер, 3x3 конволюциялық элемент, 3x3 конволюциялық қабаты 32 (сәйкес. 16) элементтер, 3x3 конволюциялық элемент, 64 бар 2x4 конволюциялық қабат (сәйкес. 24) нысандар, көлемі 3x3-тен 128-ге дейінгі конволюциялық қабат (сәйкес. 32) үлкендер үшін Ерекшеліктер (сәйкес. шағын) желі. Шағын желі кодерінде 64 функциясы бар қосымша 3x3 конволюциялық деңгейі бар.

Декодерлер әрқайсысы 128 болатын 2 екі бағытты LSTM қабаттарынан тұрады (сәйкесінше. 50 және 100) бірлік, сызықтық қабаты 128 (сәйкес. 100) аралық нейрондар. Әрбір архитектура үшін екі желі жасалады: бастапқы кескіндермен жұмыс істейтін баяу желі және бастапқы өлшемінің шамамен 70% - на дейін кішірейтілген кескіндермен жұмыс істейтін жылдам желі. Біз бұл желілерді үлкен архитектураның екі данасы үшін дәл (A) және жылдам (F), ал шағын архитектураның екі данасы үшін FastSmall (s) және FasterSmall (X) деп атаймыз. Жалпы алғанда, үлкен архитектурада шамамен 750 мың параметр бар, ал кіші архитектурада 12 битке дейінгі салмақты кванттағаннан кейін 300 мың.. бар (бұл өнімділікке әсер етпейді), бұл желілер сәйкесінше 1,1 МБ және 500 КБ дискілік кеңістікті алады.

Theodore Bluche and Ronaldo Messina ұсынған модель қолжазбаны танудың ең заманауи нәтижелерін беретін нейрондық желі архитектурасынан тұрады. Ол қолжазба мәтінінің жалпы белгілерін шығаратын конволюциялық кодерден және таңбалар тізбегін болжайтын әр тілге бейімделген LSTM декодерінен тұрады. Кодтаушының маңызды аспектісі – тиісті функцияларды таңдауға және қалғандарын блоктауға қабілетті конволюциялық қабаттар түрінде жүзеге асырылатын басқару элементтері.

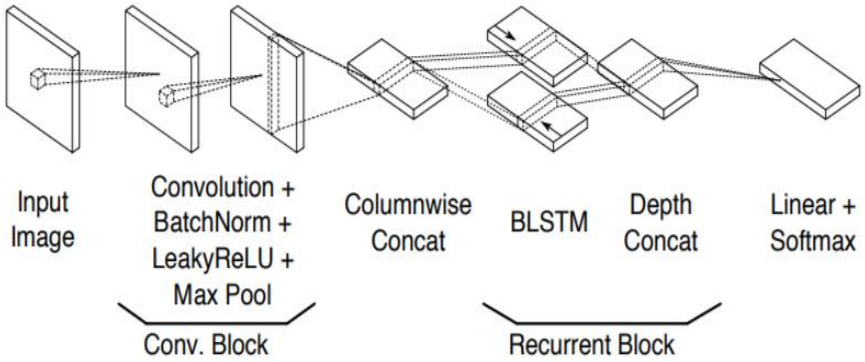
Kazakh Offline Handwritten Text Dataset (KOHTD) негізінде жүргізілген алғашқы эксперимент Bluche нейрондық моделімен жүзеге асырылды. Бұл модель тестілеу жиынтығында орташа алғанда Сөздік қателік көрсеткіші (WER) – 8,36%, Таңбалық қателік көрсеткіші (CER) – 28,95%, және Жолдық қателік көрсеткіші (SER) – 31,45% нәтижесін көрсетті.

4.3 2-LSTM әдісі негізінде Puigcerver моделі

Puigcerver жол деңгейіндегі қолжазба мәтінін (HTR) тану үшін конволюциялық және 1D-LSTM қабаттарына негізделген нейрондық желі архитектурасын ұсынды [85]. Ұсынылған архитектура заманауи 2D-LSTM архитектурасына ұқсас және жақсырақ дәлдікті қамтамасыз ететінін және салыстырмалы түрде жылдамырақ екенін дәлелдейтін статистикалық негізделген эмпирикалық зерттеулер жүргізіледі. Сонымен қатар, оқу суреттерінде сәйкес кездейсоқ бұрмалауларды, қателер жиілігін төмендететіні көрсетілді.

Бұл жұмыста автор MDLSTM желілері сызықтық деңгейдегі үшін ең заманауи өнімділікке қол жеткізу үшін өте қажет пе, әлде оның орнына арзан деңгейлерді (есептеу ресурстары тұрғысынан) пайдалануға болатындығын зерттеген. Келесі зерттеу міндеттері іске асты:

- нейрондық желілерге шолу;
- қолжазба мәтіндерді тану үшін сызықтық деңгейде орындау үшін конволюциялық қабаттар мен 1D-LSTM қабаттарына негізделген нейрондық желі архитектурасын ұсынады (Сурет 31);
- статистикалық негізделген эмпирикалық зерттеулер жүргізіп, ұсынылған архитектур қазіргі 2D-LSTM архитектурасымен салыстырғанда ұқсас немесе жақсырақ дәлдікті қамтамасыз ететінін және ол жылдамдықтары жоғары екенін көрсетті;
- оқыту кескіндеріне адекватты кездейсоқ бұрмалауларды қолдану арқылы қателер жиілігін айтарлықтай төмендетуге болатынын көрсетті
- басқа зерттеушілерге нәтижелерін сенімді түрде қайталауға мүмкіндік беретін бағдарламалық жасақтама мен эксперимент жасады.



Сурет 31 – Екі бағытты 1D-LSTM көмегімен құрылған нейрондық желінің архитектурасы

Архитектурада барлығы 5 конволюциялық блок бар: кіріс кескіндері (Input Image), Конволюциялық блоктар (Conv. Block), Бағандар бойынша біріктіру (Columnwise Concat), Рекурентті блок(Recurrent Block), Linear+SoftMax

Конволюциялық блоктар: әрбір конволюциялық блокта көлденең және тігінен 1 пиксель қадамымен 3×3 пиксель ядросының өлшемі бар екі өлшемді конволюциялық қабаттан тұрады. N-ші конволюциялық қабаттағы сүзгілер саны 16n. Қайта оқытуды азайту үшін кейбір конволюциялық қабаттардың кірісіне 0,2-ге тең ықтималдықпен електен өткізу қолданылады. Конволюциялық деңгейден кейін сызықтық емес активтендіру функцияларының кіріс мәліметтерін қалыпқа келтіру үшін пакеттік қалыпқа келтіру қолданылады. Конвульсиялық блоктарда белсендіру функциясы ретінде бос түзетілген сызықтық бірліктерді (Leaky Rectifier Linear Units, LeakyReLU) пайдаланады. Соңында, активтендіру функциясының шығысы 2 = 2 пиксельді қабаттаспайтын

ядролармен максималды біріктіру деңгейіне (Maxpool layer, Max pool) беріледі. Максималды қабаты әдетте кіріс кескіндерінің өлшемін азайту үшін қолданылады. Әр конволюциялық блокта қолданылатын конфигурация берілген (Кесте 8).

Кесте 8 – Архитектураның конфигурациясы

Конфигурация	Мәндер
Конволюциялық фильтрлер	16 – 32 – 48 – 64 – 80
Макимальды пул (2 × 2)	Yes – Yes – Yes – No – No
Түсуі	0 – 0 – 0.2 – 0.2 – 0.2

Қайталанатын блоктар: қайталанатын блоктар 1D-LSTM екі бағытты қабаттарымен қалыптасады, олар кіріс кескінін бағандар бойынша өңдейді (14-суреттегі «Columnwise Concat» бөлімі) солдан оңға және оңнан солға қарай. Екі бағыттың шығысы тереңдікте біріктіріледі(«Depth Concat». Сурет 3.4). Сонымен, егер D әр бағыттағы жасырын блоктардың саны болса, онда BLSTM блогының шығысы 2•D арналарының тереңдігіне ие болады. Мұнда 1D-LSTM әр қабатының алдында електен өткізу қолданылады (0,5 ықтималдығы бар). Барлық 1DLSTM қабаттарындағы жасырын блоктардың саны D = 256-ға тең. Барлығы 5 қайталанатын блокты қолданылады.

Сызықтық қабат: соңында, қайталанатын 1D-LSTM блоктарынан кейінгі әрбір баған Шығыс белгісіне сәйкес келуі керек. Ол үшін тереңдік 2 * D-ден L-ге аффиндік түрлендіру арқылы түрлендіріледі (мұндағы L бос CTC таңбасы үшін + 1 таңбалар санына тең). Қайталанатын блоктарда жасалғандай, қайта оқытудың алдын алу үшін сызықтық қабаттың алдында електен өткізу қолданылады (сонымен қатар 0,5 ықтималдығы бар).

Puigcerver қолжазбаны тану дәлдігіне қол жеткізу үшін көп өлшемді қайталанатын қабаттар қажет болмауы мүмкін екендігі туралы көптеген дәлелдер келтірген. Дегенмен, пайдаланған дерекқорлар жиынтығының көлемі аз екенін ескеру керек.

Kazakh Offline Handwritten Text Dataset (KOHTD) негізінде Puigcerver моделімен жүргізілген экспериментте 135 дәуір деректерінде оқытылды. Бұл модель сынақ деректер жиынтығында орташа есеппен 8,01% CER, 26,34% WER және 28,94% SER өнімділігін көрсетті.

4.4 Gated-CNN әдісі негізінде Флор моделі

Оффлайн қолжазба мәтінін танудың заманауи тәсілдері танымал конволюциялық қайталанатын нейрондық желілерге (Convolutional Recurent Neural Networks, CRNNs) негізделген. Бұл конволюциялық қайталанатын нейрондық желілер қолданғанда сканерленген құжаттарда жиі кездесетін көлемді мәтіндік кескіндерді өңдеу кезінде жоғалып кету, градиенттердің жоғалуы сияқты мәселелерден қолжазбаны тану ықтималдығы төмендеуі мүмкін. Сонымен қатар, конволюциялық қайталанатын нейрондық желілер

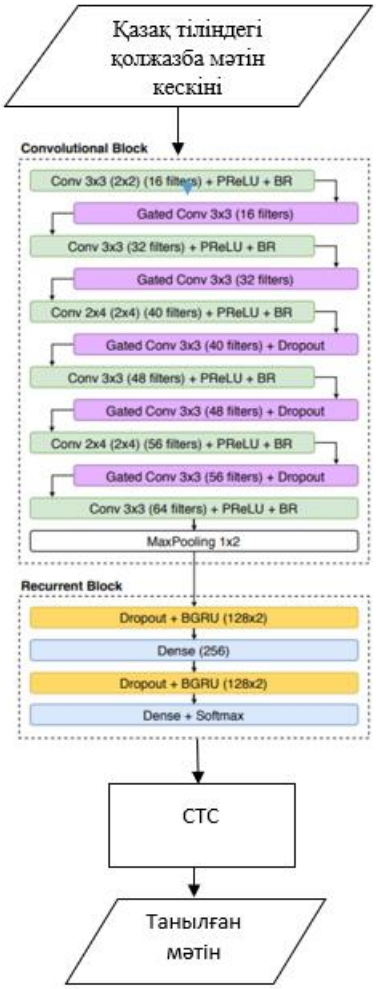
әдетте үлкен көлемдегі деректер мен есептеу ресурстарын қажет ететін миллиондаған параметрлерді қамтиды. Конволюциялық қайталанатын нейрондық желілердің мүмкіндіктерін толықтыратын Жабық Конволюциялық нейрондық желілер (Gated Convolutional Neural Networks, Gated-CNN) деп аталатын нейрондық желілердің жұмыс архитектурасының жаңа класы құрылды. Flor жабық конволюциялық нейрондық желілеріне негізделіп, параметрлері мен қабаттары аз заманауи қолжазбаны тану моделін, архитектурасын ұсынды [86]. Ұсынылған моделмен эксперименттер жасау нәтижесінде қолжазбаны тану бойынша статистикалық мәліметтер өте жақсы нәтижелер берді және басқа модельдермен салыстырғанда тану ықтималдығы жоғары болды. Сонымен қатар, ұсынылған модель оқу деректері аз болса да, қанағаттанарлық тану көрсеткіштеріне қол жеткізе алады. Ақырында, модельдің архитектурасы шағын болғанына қарамастан, есептеулер саны аз және бұл роботтар және смартфондар үшін қолдануға ыңғайлы. Қолжазба мәтіндерді тану үшін машиналық оқытудың соңғы әдістері мен тілдерді өңдеу саласында қолданылатын тәсілдерді біріктіретін Gated Convolutional Recurrent Neural network (Gated-CRNN) жаңа архитектурасының артықшылықтары:

- Оқу деректері шектеулі болса да, әртүрлі стильдермен, вариациялармен және Шумен ұзақ сөйлемдерді өңдеуге қабілетті;
- Жаңа Gated-CNN-BGRU архитектурасы арқылы CNN-BLSTM тәсілімен тану нәтижелерін жақсартуға болады;
- Gated-CNN-BGRU архитектурасы арқылы оқытылатын параметрлердің санын азайтып (мыңдаған) және дәстүрлі CNN-BLSTM (миллиондаған) орнына модельді кішірейтіп, есептеу шығындарын санын азайтуға болады.

Кескіндерден қолжазба мәтіндерді тану үшін Gated-CNN тәсілін қолданылады. Бұл Gated-CNN механизмі Блюш тәсілімен бірдей мақсатты көздейді, бірақ аздап айырмашылық бар. Ол сигмоидты белсендіруді орындау үшін кіріс функцияларының (h_1) жартысын ғана пайдаланады, ал екінші жартысы (h_2) жасамайды және соңында нәтиже екі жартының арасындағы нүктелік көбейту болып табылады:

$$y = s(h_1) \cdot h_2 \tag{4.4}$$

Бұл тәсіл аз мөлшерде параметрлерді қолдайтын (шамамен 820 мың) жабық механизмді жақсырақ пайдалануға және ұсынылған модельдің өнімділігін арттыруға мүмкіндік береді. Сонымен қатар, дәстүрлі BLSTM орнына BGRU қолданылады. Ұсынылаған архитектура бойынша 11 конволюциялық қабатты (5 жабық қабатты қоса) және 2 BGRU көмегімен жұмыс процесі іске асырылған (Сурет 32).



Сурет 32 – Ұсынылған архитектураның жұмыс процесі

Конволюциялық блок дәстүрлі және жабық конволюциялары бар шағын блоктардан тұрады, осылайша: (1) 3x3 конволюциясы және 3x3 жабық конволюциясы бар (16 объект); (2) 3x3 конволюциясы және 3x3 жабық конволюциясы (32 объект); (3) 2x4 конволюциясы және 3x3 жабық конволюциясы (40 объект); (4) 3x3 конволюция және 3x3 жабық конволюция (48 объект); (5) 2x4 конволюция және 3x3 жабық конволюция (56 объект); және (6) 3x3 конволюция (64 объект). Оның формасы активатор ретінде параметрлік түзетілген сызықтық блок (Parametric Rectified Linear Unit, PReLU) инициализатор ретінде қолданылады. Пакеттік қалыпқа келтіру барлық конволюциялық қабаттарда қолданылады, содан кейін соңғы үш жабық механизмде електен өткізу (0,2 ықтималдығы) қолданылады.

Қайталанатын блокта тығыз қабатпен кезектесетін GRU ұяшықтарында пролапсы бар 2 BGRU (0,5 ықтималдығы) бар. GRUs ішіндегі жасырын бірліктердің саны 128-ге тең. Соңында, модельде таңбалар жиынтығының өлшеміне тең тағы бір тығыз қабат бар (бос CTC таңбасы).

Терең оқытудың мүмкіндіктерін пайдаланып қолжазба мәтіндерді тану сәтті қолданыс табуда. Модельде қолжазба мәтіндерді тану үшін конволюциялық

қайталанатын нейрондық желілер (Convolutional Recurrent Neural Networks, CRNN) әдісі қолданылды.

Жұмыс процесін 3 кезеңге бөлуге болады:

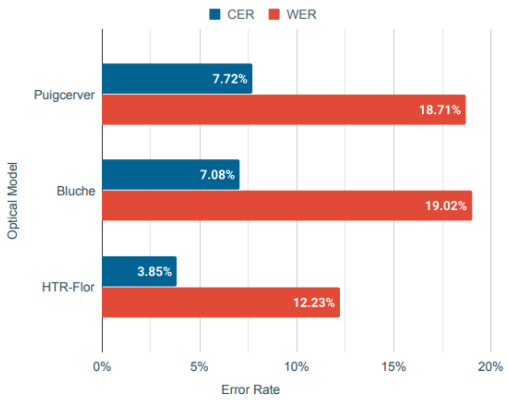
1-қадам: Нысандарды шығару үшін кіріс кескіні конволюциялық нейрондық желілер (CNN) қабаттарына жүктеледі. Шығу кезінде объектілердің картасы алынады.

2-қадам: Ұзақ қысқа мерзімді жады (LSTM) арқылы рекурентті нейронды желілер (RNN) ақпарат сенімді оқытылады.

3-қадам: Рекурентті нейронды желілер (RNN) шығыс матрицасын қолдана отырып, байланысты орнатуды уақытша қолдайтын классификациясы (CTC) шығындар мәндерін есептейді, сондай-ақ соңғы мәтіндегі бөліктерді шығарады.

Ұсынылған модельді қазіргі деңгеймен салыстыру үшін Bentham, IAM, RIMES, Saint Gall және Washington деректер жиынтығын қолдана отырып, эксперименттік бағалау жүргізілді, барлығы сегменттелмеген тәсілді қолданады.

Эксперименттің барлық нәтижелерін қорытындылау үшін барлық мәліметтер жиынтығында алынған орташа қателіктерді талданды. Әрбір оптикалық модель үшін орташа қателік жылдамдығы көрсетілген (Сурет 33)



Сурет 34 – Қате жиілігі туралы қысқаша ақпарат (неғұрлым аз болса, соғұрлым жақсы)

Gated-CNN-BGRU жаңа архитектурасының арқасында тану нәтижелерін жақсартылды. Оқытылатын параметрлер саны азайтылды, модель өлшемі және есептеу шығындарын азайтылды.

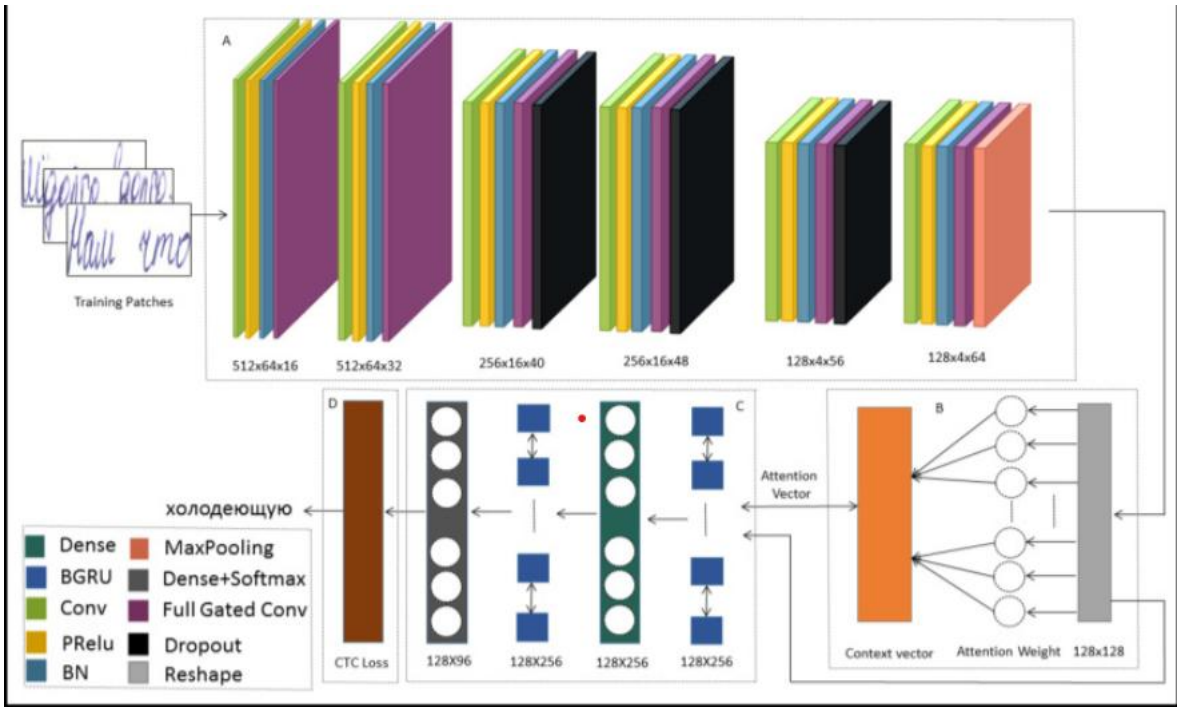
Kazakh Offline Handwritten Text Dataset (KOHTD) негізінде 186 дәуір деректерінде оқытылған Flor моделімен жүргізілді. Бұл модель сынақ деректер жиынтығында орташа есеппен 6,52% CER, 24,52% WER және 26,98% SER өнімділігін көрсетті.

4.5 Attention-Gated-CNN-BGRU әдісі негізінде Abdallah моделі

A. Abdallah толық жабық CNN негізінде терең нейрондық желінің жаңа моделін пайдаланып, қазақ және орыс қолжазба мәтінін тану мәселесі қарастырды [87]. Attention-Gated-CNN-BGRU моделі кириллица графикасына

негізделген қазақ және орыс тіліндегі қолжазба мәтінді тануға бағытталған. Кіріс - 1D символдар тізбегін құрайтын сөздер жазылған кескіндер. Attention-Gated-CNN-BGRU моделі жоғары тану жылдамдығына ие, сонымен қатар, басқа модельдермен салыстырғанда қателік деңгейі төмен.

Attention-Gated-CNN-BGRU (convolutional neural networks - bidirectional gated recurrent unit, екі бағытты басқарылатын блокпен конволюциялық нейрондық желілер) архитектурасы сипаты келесі суретте беріледі (Сурет 35):



Сурет 35 – Қолжазбаны тануға арналған жабық бағытталған CNN-BGRU жүйесі

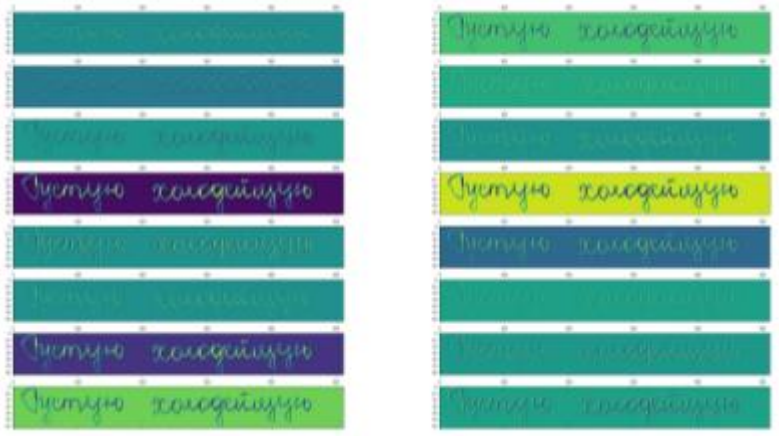
Жүйе төрт негізгі бөліктен тұрады:

- (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. Жабық (Gate) қабатқа дейін (Сурет 4.9) және кейін объектінің нақты мысалының объектілерінің визуализациясынан тану нәтижесінің жақсарғанын көруге болады (Сурет 36).



Сурет 36 – Жабық (Gate) қабаттарға енгізілмегенге дейінгі шығыс нысаналары



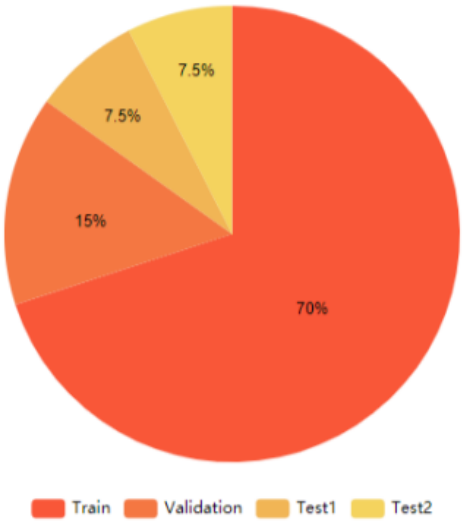
Сурет 37 – Жабық (Gate) қабаттан кейінгі шығыс нысандары

Жабық (Gate) қабаты берілген позициядағы векторлық объектінің мәнін және іргелес мәндерді қарастырады және оны сол күйде ұстау немесе тастау керектігін анықтайды.

Декодер – бұл таңбалар тізбегін анықтау үшін белгілер тізбегін өңдейтін екі бағытты басқарылатын рекурентті блокпен (Gated Recurrent Unit, GRU) рекурентті нейронды желілер. Белгілер векторында 256 белгі бар және қайталанатын нейрондық желі ақпаратты реттілік бойынша таратады. Басқарылатын рекурентті блок (Gru) жаңарту элементі мен қалпына келтіру элементін пайдаланып жоғалып бара жатқан градиенттер мәселесін шеше алады. Жаңарту элементі жадқа келетін ақпаратты басқара алады, ал қалпына келтіру

элементі жадқа келетін ақпаратты басқара алады. Жаңарту және қалпына келтіруді басқару элементтері-шығысқа қандай ақпарат берілетінін анықтайтын екі вектор. Екі вектор өткен оқуды сақтауға немесе болжамға қатысы жоқ ақпаратты жоюға қабілетті болуы мүмкін. Басқарылатын рекурентті блок LSTM-ге ұқсас, бірақ оның параметрлері аз, өйткені оның шығуы үшін басқару элементі жоқ. RNN қабатының Шығыс тізбегі 128 × 96 матрица болып табылады. Назар аудару блогы дереккөздегі ең маңызды ақпаратқа назар аударуға мүмкіндік береді.

Қазақ және орыс тілдеріндегі қолжазба (HKR, handwritten Kazakh and Russian) мәтіндерінің дерекқоры [13, б. 33075] қолжазба мәтінін тану бойынша зерттеулерге негіз бола алады. Онда жүздеген түрлі жазушылар жазған орыс сөздері (аудандар, қалалар, ауылдар, ауылдар, көшелер) бар. Ол сондай-ақ Қазақстан Республикасындағы ең танымал сөздерді қамтиды. Деректер базасымен бірге бірнеше алдын-ала өңдеу және сегменттеу процедуралары жасалған. Олардың құрамында әр түрлі қолжазба формалары бар. Деректер базасы қазақ және орыс сөздерін тану бойынша зерттеулерге арналған оқыту және тест жинағын ұсыну үшін дайындалған. Деректер базасы 1400-ден астам толтырылған нысандардан тұрады. Онда шамамен 63 000 сөйлем, 715 699-дан астам таңба және шамамен 106 718 сөз бар. Қазақ және орыс тілдері бойынша жалпыға қолжетімді деректердің болмауына байланысты бұл жұмыста алғаш рет HKR деректер жинағы пайдаланылады (Сурет 38).



Train images	45,470
Validation images	9359
Test1 images	5057
Test2 images	5057
Total	64,943

Сурет 38 – HKR қолжазба деректер жинағы бойынша оқыту, валидация және тестілеуге арналған мәліметтер жиынтығы.

Деректер жиынтығы үш деректер жиынтығына келесідей бөлінді: оқыту (70%), валидация (15%) және тестілеу (15%). Сынақ деректер жинағы екі ішкі деректер жинағына тең бөлінді (әрқайсысы 7,5%): бірінші деректер жинағы Test1 деп белгіленді және оқу және тексеру деректер жинағында жоқ сөздерден тұрды; екіншісі Test2 деп аталды және оқу деректер жинағында бар, бірақ қолжазба стильдері мүлдем басқа сөздерден тұрды.

Біз модельді назар аударумен, назар аудармай және соңында басқарылатын рекурентті блоксыз (GRU) оқытылды. Назар аудару декодерді қолдану кезінде қателерді азайтуға көмектесетінін көруге болады (Кесте 9).

Кесте 9 – Назар аудару механизмінің HKR деректер жиынтығына әсері

Модель	Test 1			Test 2			Көрсеткіштер
	CER (%)	WER (%)	SER (%)	CER (%)	WER (%)	SER (%)	
Назар аударумен	4.13	18.91	25.27	6.31	23.69	35.16	885K
Назар аударусыз	9.5	37.7	46.1	11.4	41.1	49.4	823K
Басқарылатын рекурентті блоксыз (GRU)	23.5	76.4	83.6	19.5	61.6	80.8	296K

Attention-Gated-CNN-BGRU әдісі қазақ және орыс тілдерінде қолжазбаны танудың ең заманауи нәтижелеріне қол жеткізу үшін кодер–декодер нейрондық желісінің архитектурасын қолдануды қарастырған. Ол қолжазба мәтінінің жалпы белгілерін, назар аудару механизмін, BGRU декодерін және таңбалар тізбегін болжау үшін CTC моделін шығаратын толық жабық конволюциялық кодерден тұрады. Маңызды блок - бұл маңызды функцияларды таңдай алатын және басқаларды бұғаттай алатын кодерде енгізілген Шығыс тізбегінің және жабық деңгейлердің әрбір бөлігі үшін қажетті ақпаратты жинау үшін желінің қабілетін арттыратын назар аудару механизмі.Ұсынылған әдіс аз параметрлермен тану жылдамдығының жоғары деңгейін қамтамасыз етеді. Эксперименттердің нәтижелері ұсынылған әдістің тиімділігін көрсетті.

KOHTD деректер жиынтығы негізінде жүргізілген эксперимент барысында Adallah моделі CER 8,22%, WER 22.60, SER 25.22% көрсетті.

4.6 Бағалау әдістері

Сөйлеуді автоматты түрде тану жүйелерінің жұмысындағы негізгі мәселелердің бірі-тану нәтижелерін объективті сандық бағалау, бұл әзірлеушілер үшін де, жүйенің соңғы пайдаланушылары үшін де маңызды нәтижелерін объективті сандық бағалау, бұл әзірлеушілер үшін де, жүйенің соңғы пайдаланушылары үшін де маңызды. Бұл жұмыста модельдерді бағалау үшін таңбалардағы қателік жылдамдығы (CER) және сөздегі қателік жылдамдығы (WER) өнімділік көрсеткіштерін пайдаланылды [88]. Таңбалардағы қателік жылдамдығы Левенштейн қашықтығы негізінде[89] келесі формуламен анықталады:

$$CER = \frac{S+I+D}{N} \quad (4.5)$$

Мұнда бір жолды екінші жолға түрлендіру үшін S – таңбалар саны, I – кірістіру саны, D –жою саны. N – негізгі ақиқат сөзіндегі таңбалардың жалпы саны.

Сол сияқты, сөздегі қателік жылдамдығы WER мәні былай анықталады:

$$WER = \frac{S_w+I_w+D_w}{N_w} \quad (4.6)$$

Мұнда бір жолды екінші жолға түрлендіру үшін қажетті S_w – сөздер саны, I_w – кірістіру саны, D_w – жоюлар саны, N_w – негізгі ақиқат сөздер саны.

Кейде сөйлемдерді тану қателерінің көрсеткіші де қолданылады (SER, Sentence Error Rate). Сөйлемдерді тану қателерінің көрсеткіші сөйлемдерді тану қателерінің саны сөйлемдердің жалпы санына бөлінеді.

Барлық модельдерді оқыту үшін Python негізіндегі TensorFlow [90] терең оқыту кітапханасы пайдаланылды. Python көмегімен TensorFlow графикалық процессорларда жоғары тиімді математикалық операцияларды ашық пайдалануды қамтамасыз етеді. TensorFlow туралы айтпас бұрын, алдымен тензорлар терминін түсініп алу керек. Тензорлар – бұл терең оқытуда деректерді ұсынудың құралы. Тензорлар – бұл жоғары өлшемді деректерді ұсынуға мүмкіндік беретін көп өлшемді массивтер.

TensorFlow-да тензор термині деректерді көп өлшемді массив ретінде ұсынуды білдіреді, ал ағын термині жоғарыдағы суретте көрсетілгендей адамның тензорлармен орындайтын бірқатар операцияларын білдіреді. Негізінде, TensorFlow бағдарламасын жазудың жалпы процесі екі кезеңнен тұрады:

- Есептеу графигін құру
- Есептеу графигін іске қосу

Есептеу графигі – бұл графиктегі түйіндер түрінде орналасқан тензор ағынының бірқатар операциялары. Әрбір түйін кіріс ретінде 0 немесе одан да көп тензорды қабылдайды және шығыс ретінде тензорды шығарады.

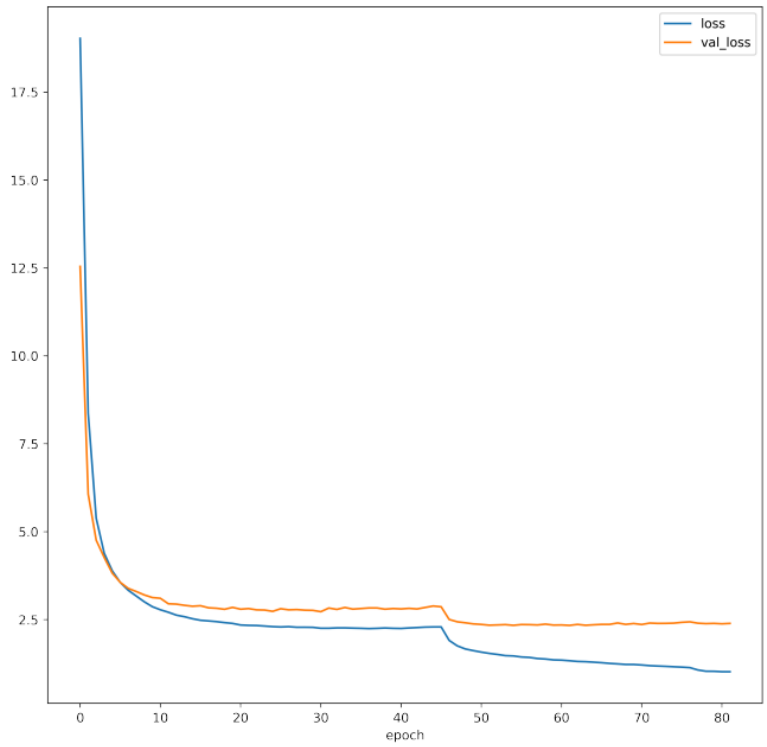
Python сценарийінде есептеу деректері үшін қажетті барлық операцияларды анықтайтын есептеу графигі құрылады. Сынақтар жүргізілген машинада 2x Intel(R) Xeon(R) E-5-2680 процессорлары, 4X NVIDIA Tesla k20x графикалық карталары және 100 ГБ жедел жады бар. GPU пайдалану модельді оқыту уақытын шамамен үш есе қысқартты, бірақ бұл үдеу жоба барысында мұқият бағаланбады, сондықтан ол әртүрлі болуы мүмкін. Есептің графиктері Python-да matplotlib пакеті арқылы салынды. Тексеру кезінде шығын мөлшері барлық модельдерде минимумға дейін азаяды. RMSProp әдісі 0,001 базалық оқу жылдамдығын және 32 шағын пакетті қолдана отырып, стохастикалық градиентті түсіруді оңтайландыру үшін қолданылады. Біз сондай-ақ patience 20 көмегімен ерте аялдаманы қолдандық, өйткені біз әр дәуірде валидацияның

жоғалуын қадағалап, 20 дәуірден кейін валидацияның жоғалуы жақсармаса, оқуды тоқтатқымыз келді.

4.6 Қазақ-орыс тіліне икемделіп оқытылған модельдердің нәтижелері

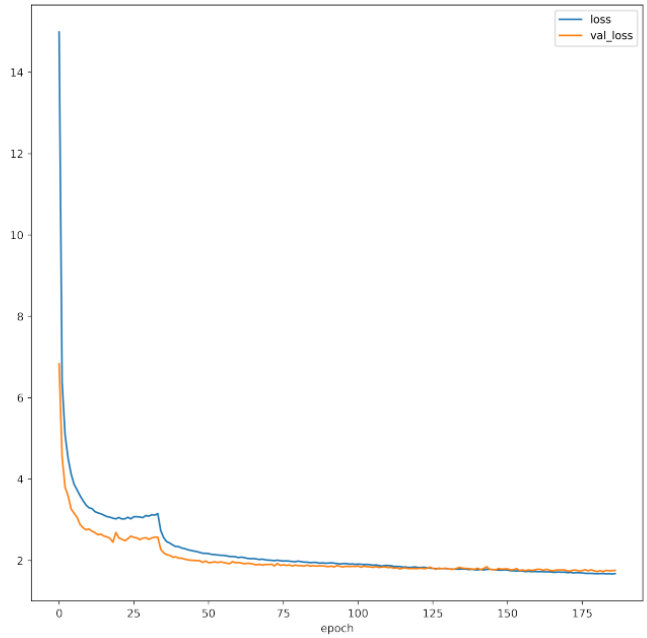
Барлық модельдер KOHTD деректер жиынтығында оқытылды. Біз бұл модельдерді барлық ұсынылған нәтижелер үшін қолданылатын стандартты өнімділік көрсеткіштері бойынша бағаладық: CER, WER және SER. Барлық модельдер үшін 32 өлшемді шағын партия орнатылды және 20 кезеңнен кейін ерте тоқтады, тексеру кезінде шығын мәнін жақсартпады және lr = 0,001. Әр модельді тиімді пайдалану үшін 20 дәуір шамасында 0,2 әлсіреу коэффициентімен график ReduceLROnPlateau графигі [91] пайдаланылды және 10 дәуірден кейін шығын мәні жақсы нәтиже бермеді.

Бірінші эксперимент Bluche моделімен жүргізілді, ол сынақ деректер жиынтығында орташа есеппен 8,36% WER, 28,95% CER және 31,45% SER өнімділігін көрсетті. Bluche моделін оқыту және тексеру кезіндегі шығындар графикте берілген (Сурет 39).



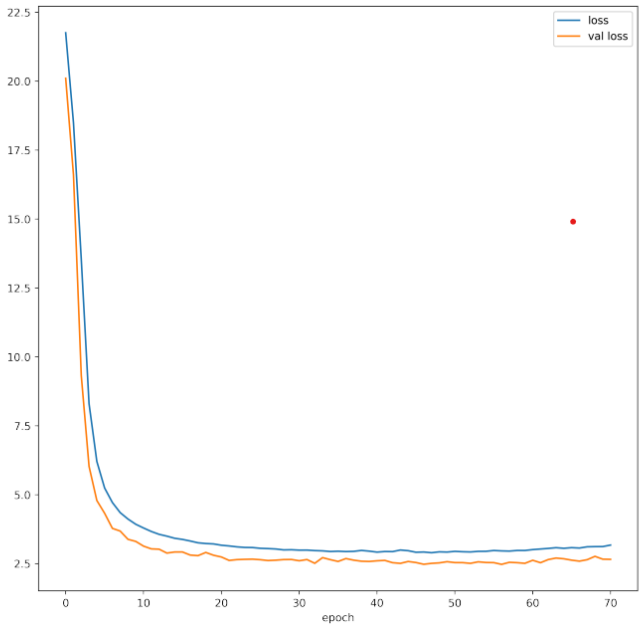
Сурет 39 – Bluche моделі бойынша шығындар

Келесі эксперимент 186 дәуір деректерінде оқытылған Flor моделімен жүргізілді. Бұл модель сынақ деректер жиынтығында орташа есеппен 6,52% CER, 24,52% WER және 26,98% SER өнімділігін көрсетті. Оқыту және тексеру кезіндегі шығындар график түрінде берілген (Сурет 40).



Сурет 40 – Flor моделі бойынша шығындар

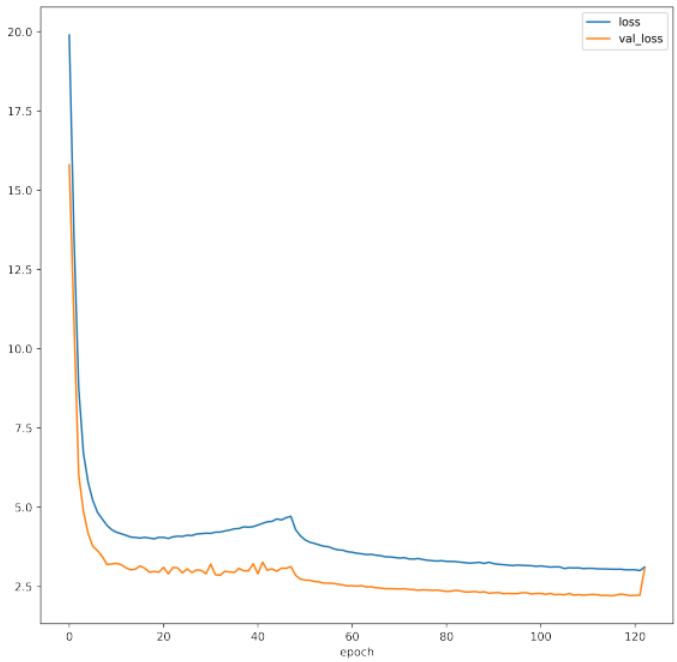
Дәл осындай эксперименттер Абдалланың HTR моделімен жүргізілді. Алдыңғы қарастырылған модельдерден айырмашылығы, бұл модель 8,22% төмен орташа көрсеткішті көрсетті. Сынақ деректер жиынтығында CER, 22,60% WER және 25,22% SER. Келесі суретте Абдаллах моделін оқыту және тексеру кезіндегі шығындар көрсетілген (Сурет – 41).



Сурет 41 – Абдаллах моделін оқыту және тексеру кезіндегі шығындар

Соңғы эксперимент Puigcerver моделімен жүргізілді, ол 135 дәуір деректерінде оқытылды. Бұл модель сынақ деректер жиынтығында орташа есеппен 8,01% CER, 26,34% WER және 28,94% ser өнімділігін көрсетті. Келесі

суретте Puigcerver моделін оқыту және тексеру кезіндегі шығындар көрсетілген (Сурет 42).



Сурет 42 – Puigcerver моделін оқыту және тексеру кезіндегі шығындар

Сонымен қатар, Carion [92], Scheidl [93] модельдерімен тағы екі тәжірибе жасалды. Carion моделі орташа есеппен 12,85% CER, 31,73% WER және 35,16% SER өнімділігін көрсетті. Бұл көрсеткіштер Scheidl моделі үшін 22,99% CER, 62,77% WER және 62,77% SER мәндерге ие болды. Барлық модельдер арасындағы салыстыру нәтижесі көрсетілген (Кесте 10).

Кесте 10 – Модельдер арасындағы салыстыру нәтижесі

Модель	Әдіс	CER	WER	SER
Flor[84]	Gated-CNN	6.52%	24.52%	26.98%
Puigcerver[83]	2-LSTM	8.01%	26.34%	28.94%
Abdallah[85]	Attention-Gated-CNN-BGRU	8.22%	22.60%	25.22%
Bluche[82]	Gated-CRNN	8.36%	28.95%	31.45%
Carion[90]	Transformer	12.85%	31.73%	35.16%
Scheidl[91]	CNN-LSTM-CTC	22.99%	62.77%	-

Қазақ-орыс тіліндегі қолжазба мәтінін оптикалық тану мақсатында Flor, Puigcerver, Bluche және Scheidl ұсынған нейрондық модельдер икемделіп оқытылды.

Кириллица графикасына негізделген қазақ тіліндегі қолжазбаны тану мәселесін шешуге бағытталған нейрондық желілерге негізделген әдіс [94,95] жұмыста ұсынылады. Жұмыстың негізгі мақсаты – қазақ тіліндегі мәтін жолдарын дәл әрі тиімді тану мүмкіндігін қамтамасыз ету. Зерттеу нәтижелері болашақтағы ғылыми ізденістердің бағытын айқындауға ықпал етеді. Сонымен қатар, қазақ тіліндегі қолжазба мәтіндерді тану тиімділігін бағалау үшін KOHTD деректер қорындағы сөздер негізінде жаңа әдістерді сынақтан өткізу жоспарланған.

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.

Диссертациялық жұмыста келесідей нәтижелер алынды:

- Қолжазба мәтіндерді оптикалық тану есептеріне терең оқытуды қолданған шешімдерге шолу жасалды. Жасырын Марков модельдеріне (HMM), конволюциялық (CNN) және қайталанатын нейрондық желілерге (RNN) негізделген модельдер сипатталынып, талданды;

- Нейрондық желілерді тиімді оқыту үшін қажетті деректерді жинауға мүмкіндік беретін краудсорсинг әдісіне негізделген автоматтандырылған технологиялық жүйе әзірленді. Бұл жүйе қазақ тіліндегі қолжазба мәтіндер корпусын жинақтаудың тиімділігін арттыруға ықпал етті.

- Қазақ-орыс тіліндегі қолжазба мәтіндерден сөздерді дәл және сенімді түрде сегменттеуге арналған бейімделген генетикалық алгоритм құрылды;

- Алғаш рет сөз деңгейінде құрылымдалған қазақ тіліндегі оффлайн қолжазба мәтіндер корпусын қамтитын Kazakh Offline Handwritten Text Dataset (KOHTD) құрылды. Бұл корпус қазақ тіліне арналған HTR жүйелерін дамытуда маңызды ресурстардың біріне айналды.

- Қазақ-орыс тілдерінде қолжазба мәтіндерінің деректер қоры негізінде терең оқытудың әртүрлі әдістерімен эксперименттер жүргізілді. Bluche ұсынған Gated-CRNN әдісі, Puigcerver ұсынған 2-LSTM әдісі, Flor ұсынған Gated-CNN әдісі, Abdallah ұсынған Attention-Gated-CNN-BGRU әдісі негізінде модельдер KOHTD деректер жиынтығында икемделген оқыту (fine-tuning) әдісі арқылы оқытылып, стандартты өнімділік көрсеткіштері CER, WER және SER бойынша бағалау жасалды. Нәтижесінде қазақ тіліндегі қолжазба мәтіндерді тану дәлдігі арттырылып, HTR жүйелерінің тілдік бейімделу қабілеті көрсетілді.

Жалпы алғанда, бұл зерттеу қазақ тілінде жазылған қолжазбаларды тану мәселесін шешуге бағытталған алғашқы жүйелі жұмыстардың бірі болып табылады және болашақта қазақ тіліндегі цифрлық мұрағаттарды автоматты өңдеу, лингвистикалық талдау және тарихи құжаттарды сақтау салаларында кеңінен қолданылуы мүмкін. Қолжазба жазбаларының цифрланған мәтіні көптеген компаниялардың бизнес-процестерін автоматтандыруға мүмкіндік береді, бұл адамның жұмысын жеңілдетеді. Зерттеу жұмыстарын орындау кезінде алынған нәтижелер пошта жөнелтілімдерін өңдеу орталықтарының жұмысын автоматтандыруға мүмкіндік береді.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1 Fakoor R., Ladhak F., Nazi A., Huber M. Using deep learning to enhance cancer diagnosis and classification // In Proceedings of the international conference on machine learning. ACM, New York, USA. – 2013. – Vol. 28. - P.3937–3949.

2 Galea C., Farrugia R.A. Forensic Face Photo-Sketch Recognition Using a Deep Learning-Based Architecture // IEEE Signal Processing Letters. – 2017. – Vol.24. P.1586-1590.

3 Chorowski J., Bahdanau D., Serdyuk D., Cho K., Bengio Y. Attentionbased models for speech recognition. // arXiv. – 2015. – Vol: 1506. -7503 p.

4 Lakkhanawannakun P., Noyunsan C. Speech Recognition using Deep Learning // 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). – 2019. – P. 1-4.

5 Yavariabdi H., Kusetogullari T., Celik, S., Thummanapally S., Rijwan and J. Hall. CArDIS: A Swedish Historical Handwritten Character and Word Dataset // in *IEEE Access*. – 2022. – Vol. 10. - P.55338-55349. doi: 10.1109/ACCESS. 2022.3175197.

6 Abdallah A., Berendeyev A., Nuradin I, Nurseitov D. TNCR: Table Net Detection and Classification Dataset // Neurocomputing. – *2021. – Vol. 473*. - P.79-97.

7 Prasad D., Gadpal A., Kapadni K., Visave M., Sultanpure, K.A. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2020. – P.2439-2447.

8 Waschneck B., Reichstaller A., Belzner L., Altenmüller T., Bauernhansl T., Knapp A., Kyek A. Optimization of global production scheduling with deep reinforcement learning // Procedia CIRP. – 2018. – Vol.72. - P.1264-1269. doi:10.1016/ j.procir.2018.03.212.

9 Hamada M.A., Abdallah A., Kasem M., Abokhalil M. Neural network estimation model to optimize timing and schedule of software projects // 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), IEEE. – 2021. – P. 1–7. doi:10.1109/SIST50301.2021.9465887.

10 Fischer A., Suen C.Y., Frinken V., Riesen K., Bunke H. A fast matching algorithm for graph-based handwriting recognition // International Workshop on Graph-Based Representations in Pattern Recognition. Springer. – 2013. – P. 194–203.

11 Liu H., Ding, X. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes // Eighth International Conference on Document Analysis and Recognition (ICDAR'05). – 2005. – Vol. 1. - P. 19-23.

12 Zamora-Martínez F., Frinken V., Boquera S.E., Bleda M.J., Fischer A., Bunke H. Neural network language models for off-line handwriting recognition // Pattern Recognit. – 2014. – Vol. 47. - P.1642-1652.

13 Nurseitov D., Bostanbekov K., Kurmankhojayev D., Alimova A., Abdallah A., Tolegenov R. Handwritten kazakh and russian (HKR) database for text recognition // Multimedia Tools and Applications. – 2020. – Vol.80. - P. 33075 - 33097.

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

15 Mahmoud S. A., Ahmad I., Al-Khatib W. G., Alshayeb M., Tanvir Parvez M., Märgner V., Fink, G. A. // KHATT: An open Arabic offline handwritten text database. Pattern Recognition. – 2014. – Vol 47(3). - P. 1096-1112. doi:10.1016/j.patcog.2013.08.009.

16 Parvez M.T., Mahmoud S. A. Arabic handwriting recognition using structural and syntactic pattern attributes // Pattern Recognition. – 2013. – Vol. 46 (1). - P.141–154.

17 Jomy J., Balakrishnan K., Pramod K. A system for offline recognition of handwritten characters in malayalam script // International Journal of Image, Graphics and Signal Processing. – 2013. – Vol.5. - P.53-59.

18 Das S., Banerjee S. An algorithm for Japanese character recognition // International Journal of Image, Graphics and Signal Processing. – 2015. –Vol.1. - P. 9-15.

19 Marti U., Bunke H. A full English sentence database for off-line handwriting recognition // Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99. – 1999. – P.705-708.

20 Marti U., Bunke H. The IAM-database: an English sentence database for offline handwriting recognition // International Journal on Document Analysis and Recognition. – 2002. – Vol.5. - P.39-46.

21 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.

22 Santos R.P., Clemente G.S., Tsang I., Cavalcanti, G.D. Text Line Segmentation Based on Morphology and Histogram Projection // 10th International Conference on Document Analysis and Recognition – 2009. – P. 651-655.

23 Augustin E., Carré M., Grosicki E., Brodin J.M., Geoffrois E., Prêteux, F.J. RIMES evaluation campaign for handwritten mail processing // in: International Workshop on Frontiers in Handwriting Recognition (IWFHR'06). – 2009. – P. 231-235.

24 Kermorvant C., Louradour J. Handwritten mail classification experiments with the rimes database. // in: 2010 12th International Conference on Frontiers in Handwriting Recognition, IEEE. – 2010. – P. 241-246.

25 Guichard L., Toselli A.H., Coüasnon B. Handwritten word verification by SVM-based hypotheses re-scoring and multiple thresholds rejection // in: 2010 12th International Conference on Frontiers in Handwriting Recognition, IEEE. – 2010. – P. 57–62.

26 Siddiqi I., Vincent N. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features // Pattern Recognition. – 2010. – Vol. 43 (11). - P.3853–3865.

27 Pechwitz M., Maddouri S., Märgner V., Ellouze N., Amiri, H. IFN/ENIT: database of handwritten arabic words // in: Proc. of CIFED. – 2002. – Vol. 2. - P. 127–136.

28 Su T., Zhang T., Guan D. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text // IJDAR. – 2007. – Vol. 10. - 27 p. https://doi.org/10.1007/s10032-006-0037-6

29 Zhou Q. Chen X. Wang X. Guo H., Li. An Empirical Evaluation on HIT-OR3C Database // 2011 International Conference on Document Analysis and Recognition. Beijing, China. – 2011 – P. 1150-1154. doi: 10.1109/ICDAR.2011.232.

30 Fischer A., Keller A., Frinken V., Bunke H. Lexicon-free handwritten word spotting using character HMMS // Pattern recognition letters. – 2012. – Vol. 33(7). - P. 934–942.

31 Cheddad A., Kusetogullari H., Hilmkil A., Sundin L., Yavariabdi A., Aouache M., Hall J. SHIBR – the swedish historical birth records: a semiannotated dataset // Neural Computing and Applications. – 2021. – Vol. 33 (22). - P.15863–15875.

32 Fischer A., Frinken V., Forn´es A., Bunke H. Transcription alignment of Latin manuscripts using hidden markov models // in: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. – 2011.- P. 29–36.

33 Romero V., Forn´es A., Serrano N., S´anchez J. A., Toselli A. H., Frinken V., Vidal E., Llad´os J. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition, Pattern Recognition. – 2013. – Vol. 46(6). - P.1658–1669.

34 P´erez D., Taraz´on L., Serrano N., Castro F., Terrades O.R., Juan A. The GERMANA database // 2009 10th International Conference on Document Analysis and Recognition. – 2013. – P. 301–305. doi:10.1109/ ICDAR.2009.10.

35 Serrano N., Castro F., Juan A. The RODRIGO database // Proceedings of the Seventh International Conference on Language Resources and Evaluation.– 2010. – P.2709-2712.

36 Yavariabdi A., Kusetogullari H., Celik T., Thummanapally S., Rijwan S., Hall J. CArDIS: A Swedish Historical Handwritten Character and Word Dataset // IEEE Access. – 2022. – Vol. 10. - P.55338-55349. doi: 10.1109/ACCESS.2022.3175197.

37 Papazoglou A., Pratikakis I., Markou K., Tsochatzidis L. EPARCHOS - Historical Greek Handwritten Document Dataset. (1.0).– 2021. https://zenodo.org/record/4095301#.YaneeroxVPY. 28.10.2023.

38 Cohen G. G., Afshar S.S., Tapson J. J, van Schaik A. EMNIST: Extending MNIST to handwritten letters // 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA. – 2017. – P. 2921-2926. doi: 10.1109/IJCNN.2017.7966217.

39 Daniyar N., Kairat B., Maksat K., Anel A. 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.

40 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.

41 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. - P.33075–33097.

42 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.

43 Sozdikqor: Sh. SHayahmetov atyndagy «til-qazyna» ulttyq gylymipraktikalyq ortalygy, https://sozdikqor.kz. 24.10.2023.

44 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. – 116827 p. doi: https://doi.org/10.1016/j.image.2022.116827.

45 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

46 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.

47 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.

48 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.

49 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 p.

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

51 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.

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

53 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.

54 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.

55 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.

56 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.

57 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.

58 Yeleussinov A., Amirgaliyev Y., Cherikbayeva L. Improving OCR Accuracy for Kazakh Handwriting Recognition Using GAN Models // Appl. Sci. – 2023. – Vol. 13.- 5677 p. https://doi.org/10.3390/ app13095677.

59 Bluche T. Louradour J. Messina R. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention // 2017 14th IAPR international conference on document analysis and recognition (ICDAR). – 2017. – Vol. 1. - P. 1050–1055.

60 Moysset B., Kermorvant C., Wolf C. Full-page text recognition: Learning where to start and when to stop // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). – 2017. – P. 871–876.

61 Wirsansky E. Hands-On Genetic Algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems. Packt Publishing Ltd. – 2020. – 334 p.

62 Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии Наук СССР. - 1965. - Т. 163, № 4. - С. 845–848.

63 Нурсеитов Д.Б, Бостанбеков К.А. Әбдіманап Ғ.С., Тойганбаева Н.А., Алимова А.Н. // База данных рукописного текста на казахском языке «Kazakh Offline Handwritten Text Dataset (KOHTD) » для моделей распознавания текста в офлайн режиме. – 2023. – № 34615.

64 McCulloch W. S., Pitts W. A logical calculus of the ideas immanent in nervous activity // The bulletin f mathematical biophysics. – 1943 – Vol.5 (4). - P.115–133.

65 Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // Psychological review. – 1958. – Vol. 65 (6). - 386 p.

66 Rumelhart D. E., Hinton G. E., Williams. Learning representations by back-propagating errors. Cognitive modeling // Nature. – 1986. – P.533-536.

67 Bourlard H., Wellekens C.J. Links between Markov models and multilayer perceptrons // Pattern Analysis and Machine Intelligence, IEEE Transactions. – 1989 – Vol. 12(12). P. 1167–1178.

68 Bridle J. S. Probabilistic interpretation of feedforward classification network outputs with relationships to statistical pattern recognition // Neurocomputing. Springer. – 1990. – P. 227–236.

69 Hopfield J. J. Neural networks and physical systems with emergent collective computational abilities // Proceedings of the national academy of sciences. – 1982. – Vol. 79. - P. 2554–2558.

70 Saul L. K., Jordan M. I. Artificial Neural Networks. chap. Attractor Dynamics and Parallelism in a Connectionist Sequential Machine, Piscataway, NJ, USA: IEEE Press. – 1990. – Vol. - P. 112–127.

71 Elman J. L. (1990). Finding structure in time. Cognitive science – 1990. – Vol. 14 (2). - P.179–211.

72 Schuster M., Paliwal, K. K. Bidirectional recurrent neural networks // Signal Processing, IEEE Transactions. – 1997. – Vol. 45 (11). - P.2673–2681.

73 Graves A., Fernández S., Schmidhuber, J. Multi-Dimensional Recurrent Neural Networks. CoRR. – 2007. – Vol. abs/0705.2 .

74 Robinson A. J. An application of recurrent nets to phone probability estimation. Neural Networks, IEEE Transactions. – 1994. – Vol. 5 (2). - P.298–305.

75 Lee S.-W., Kim Y.-J. (1995). A new type of recurrent neural network for handwritten character recognition. In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference. – 1995. – Vol. 1. - P. 38–41.

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

77 Doetsch P., Kozielski M., Ney H. Fast and robust training of recurrent neural networks for offline handwriting recognition // 14th International Conference on Frontiers in Handwriting Recognition – 2014. – P. 279-284.

78 Gers F. A., Schmidhuber J. Recurrent nets that time and count // In Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference. – 2000. – Vol. 3. - P. 189–194.

79 LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D. (1989). Backpropagation applied to handwritten zip code recognition // Neural computation. – 1989. – Vol. 1 (4). - P. 541–551.

80 Rumelhart D. E., Hinton G. E., Williams R. J. Learning internal representations by error propagation. // Tech. rep., DTIC Document – 1986.

81 Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // International Conference on Machine learning. – 2006. – P.369–376.

82 Тойганбаева Н.А., Алимова А.Н., Сакыпбекова М.Ж., Гусманова Ф.Р., Әбдіманап Ғ.С. Нейронды желілер негізінде қазақ-орыс тілдеріндегі қолжазба мәтіндерді тану // Хабаршы «физика-математика ғылымдары» сериясы. – 2023. – Т. 84(4). DOI:https://doi.org/10.51889/2959-5894.2023.84.4.018.

83 Scheidl H. Handwritten text recognition in historical documents. Master's thesis, Technische Universitat¨ Wien, Vienna, diplom-Ingenieur in Visual Computing. – 2018. Doi: https://doi.org/10.34726/hss.2018.43931

84 Bluche T., Messina R. Gated convolutional recurrent neural networks for multilingual handwriting recognition // 2017 14th IAPR international conference on document analysis and recognition (ICDAR). – 2017. – Vol. 1. - P. 646–651.

85 Puigcerver J. Are multidimensional recurrent layers really necessary for handwritten text recognition?, // 2017 14th IAPR International 29 Conference on Document Analysis and Recognition (ICDAR) – 2017. – Vol. 1. - P. 67–72.

86 Neto A. F. de Sousa., Bezerra B. L. D., Toselli A. H., Lima E. B. HTR Flor: a deep learning system for offline handwritten text recognition // 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE. – 2020. – P. 54–61. DOI:10.1109/SIBGRAPI51738.2020.00016

87 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.

88 Frinken V., Bunke H. Continuous handwritten script recognition // D. Doermann, K. Tombre (Eds.), Handbook of Document Image Processing and Recognition, Springer London. – 2017. – P. 391–425. doi:10.1007/978-0-85729-859-1_12

89 Levenshtein V. I. Binary codes capable of correcting deletions, insertions and reversals // Soviet physics. Doklady Dokl. – 1965. – Vol. 10. - P. 707—710.

90 Abadi M, Agarwal A., Barham P., Brevdo E., et al., TensorFlow: Large-scale machine learning on heterogeneous distributed systems // CoRR abs/1603. 04467. arXiv:1603.04467. – 2016. http://arxiv.org/abs/1603.04467.

91 Vinciarelli A., Luettin J. A new normalization technique for cursive handwritten words, Pattern Recognit. Lett. – 2001. – Vol. 22 (9). - P.1043–1050.

92 Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. End-to- end object detection with transformers // European Conference on Computer Vision, Springer. – 2020. – P. 213–229.

93 Scheidl H., Fiel S., Sablatnig R. Word beam search: A connectionist temporal classification decoding algorithm // 16th International Conference on Frontiers in Handwriting Recognition, IEEE. – 2018. – P. 253–258.

94 Тойганбаева Н.А., Әбдіманап Ғ.С., Алимова А.Н., Жунусова Ж.Х. Қазақ тіліндегі қолжазба мәтіндерінің мәліметтер қорын құру мен тану // «Алматы энергетика және байланыс университетінің Хабаршысы», 2023. - №3(62). -Б. 105-116.

95 Тойганбаева Н.А. Терең оқыту модельдері негізінде қазақ-орыс тіліндегі оффлайн қолжазба мәтіндерін тану // Қазақстан Республикасы Ұлттық инженерлік академиясының хабаршысы. 2023. - № 4 (90).

ҚОСЫМША А

KOHTD қолжазба мәтіндерінің жиынындағы таңбалардың үлестірімі

Кесте А.1 – KOHTD қолжазба мәтіндерінің жиынындағы таңбалардың үлестірімі

№	Таңба	Таңбалар саны
1	2	3
1.	а	109098
2.	ә	6310
3.	б	21153
4.	в	2276
5.	г	11183
6.	ғ	13313
7.	д	38711
8.	е	74741
9.	ё	8
10.	ж	13198
11.	з	15749
12.	и	15423
13.	й	11562
14.	к	25237
15.	қ	27371
16.	л	47910
17.	м	26613
18.	н	58837
19.	ң	11033
20.	о	24289
21.	ө	7155
22.	п	15267
23.	р	50502
24.	с	32737
25.	т	51834
26.	у	15817
27.	ұ	6737
28.	ү	6184
29.	ф	1704
30.	х	1288
31.	h	13
32.	ц	2602
33.	ч	375
34.	ш	9729
35.	щ	26
36.	ъ	212

А.1-кестенің жалғасы

1	2	3
37.	ы	66143
38.	і	47282
39.	ь	890
40.	э	874
41.	ю	699
42.	я	4241
43.	А	1521
44.	Ә	489
45.	Б	1739
46.	В	293
47.	Г	86
48.	Ғ	71
49.	Д	520
50.	Е	667
51.	Ё	-
52.	Ж	1004
53.	З	357
54.	И	475
55.	Й	1
56.	К	1229
57.	Қ	1710
58.	Л	87
59.	М	1074
60.	Н	516
61.	Ң	1
62.	О	1489
63.	Ө	300
64.	П	215
65.	Р	334
66.	С	1757
67.	Т	843
68.	У	77
69.	Ұ	195
70.	Ү	85
71.	Ф	185
72.	Х	87
73.	Һ	-
74.	Ц	16
75.	Ч	15
76.	Ш	245
77.	Щ	1
78.	Ъ	-
79.	Ы	10
80.	І	62

А.1-кестенің жалғасы

1	2	3
81.	Ь	1
82.	Э	166
83.	Ю	2
84.	Я	272
85.	1	893
86.	2	678
87.	3	392
88.	4	162
89.	5	405
90.	6	168
91.	7	133
92.	8	176
93.	9	198
94.	0	1117
95.	пробел	8858
96.	.	4172
97.	,	3490
98.	:	456
99.	;	154
100.	!	12
101.	?	38
102.	-	3234
103.	—	3
104.	„	1
105.	”	1
106.	"	16
107.	'	22
108.	(	418
109.	)	790
110.	[	1
111.	]	1
112.	}	1
113.	_	1
114.	+	40
115.	•	5
116.	*	2
117.	/	510
118.	=	32
119.	#	4
120.	|	57
121.	>	7
122.	$	1
123.	《	1

А.1-кестенің жалғасы

1	2	3
124.	%	74
125.	a	41
126.	b	13
127.	c	22
128.	d	12
129.	e	65
130.	f	8
131.	g	6
132.	h	9
133.	i	92
134.	j	1
135.	k	7
136.	l	20
137.	m	14
138.	n	41
139.	o	39
140.	p	21
141.	q	1
142.	r	29
143.	s	32
144.	t	40
145.	u	17
146.	v	6
147.	w	5
148.	x	7
149.	y	6
150.	z	2
151.	A	10
152.	B	2
153.	C	15
154.	D	4
155.	E	5
156.	F	4
157.	G	3
158.	H	7
159.	I	31
160.	J	-
161.	K	2
162.	L	7
163.	M	16
164.	N	3
165.	O	11
166.	P	24

А.1-кестенің жалғасы

1	2	3
167.	Q	-
168.	R	6
169.	S	10
170.	T	14
171.	U	-
172.	V	2
173.	W	5
174.	X	15
175.	Y	-
176.	Z	-

ҚОСЫМША Ә

Авторлық құқықпен қорғалатын объектілерге құқықтардық мемлекеттік тізімге мәліметтерді енгізу туралы куәлік



ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2023 жылғы «12» сәуір № 34615

Автордың (авторлардың) жеке, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
НУРСЕИТОВ ДАНИЯР БОРИСОВИЧ, БОСТАНБЕКОВ КАЙРАТ АРАТОВИЧ, ӘБДІМАНАП ҒАЛЫМЖАН СЕЙТАХМЕТҒАЛИҰЛЫ, ТОЙҒАНБАЕВА НАЗГУЛЬ АБЕНОВНА, АЛИМОВА АНЕЛЬ НУРДАНБЕКОВНА

Авторлық құқық объектісі: деректер базасы

Объектінің атауы: База данных рукописного текста на казахском языке «Kazakh Offline Handwritten Text Dataset (KOHTD)» для моделей распознавания текста в офлайн режиме

Объектіні жасаған күні: 19.05.2022

ЭЦҚ қол қойылды

Е. Оспанов

ҚОСЫМША Б

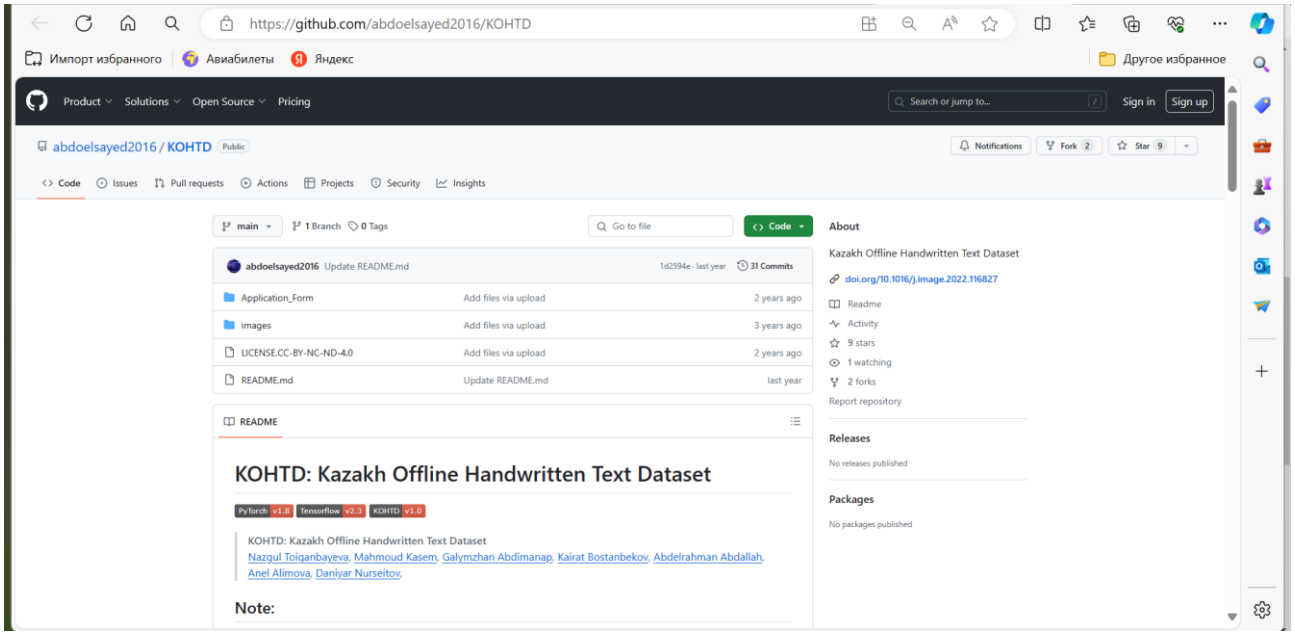
Scopus базасындағы журналдық мақала



https://www.scopus.com/authid/detail.uri?authorId=56069463400

ҚОСЫМША В

KOHTD: Kazakh Offline Handwritten Text Dataset



https://github.com/abdoelsayed2016/KOHTD

ҚОСЫМША Г
Программа кодының бөлігі

```
from __future__ import division

from __future__ import print_function


import os, json

import random

import numpy as np

import cv2

from SamplePreprocessor import preprocess

import codecs


class Sample:

  "sample from the dataset"

  def __init__(self, gtText, filePath):

    self.gtText = gtText

    self.filePath = filePath


class Batch:

  "batch containing images and ground truth texts"

  def __init__(self, gtTexts, imgs):

    self.imgs = np.stack(imgs, axis=0)

    self.gtTexts = gtTexts


class DataLoader:

  "loads data which corresponds to IAM format, see: http://www.fki.inf.unibe.ch/databases/iam-handwriting-database"

  def __init__(self, filePath, batchSize, imgSize, maxTextLen):

    """ Loader for dataset at given location, preprocess images and text according to parameters """


    assert filePath[-1] == '/'
```

```
        self.currIdx = 0
        self.batchSize = batchSize
        self.imgSize = imgSize
        self.samples = []
        chars = set()

        # Read json lables file
        # Dataset folder should contain a labels.json file inside, with key is the file name of images and value is the label
        with open(filePath + 'labels.json') as json_data:
            label_file = json.load(json_data)

        # Log
        print("Loaded", len(label_file), "images")

        # Put sample into list
        for fileName, gtText in label_file.items():
            self.samples.append(Sample(gtText, filePath + fileName))
            chars = chars.union(set(list(gtText)))

        #self.charList = list(open(FilePaths.fnCharList).read())

        # Random shuffle
        random.shuffle(self.samples)

        # Split into training and validation set: 90% - 10%
        splitIdx = int(0.9 * len(self.samples))
        self.trainSamples = self.samples[:splitIdx]
        self.validationSamples = self.samples[splitIdx:]
        # self.test1Val = self.samples # Experience

        print("Train on", len(self.trainSamples), "images. Validate on",
```

```
            len(self.validationSamples), "images.")

        # Number of randomly chosen samples per epoch for training
        self.numTrainSamplesPerEpoch = 10000

        # Start with train set
        self.trainSet()

        # List of all chars in dataset
        self.charList = sorted(list(chars))

        # save characters of model for inference mode
        #open(FilePaths.fnCharList, 'w').write(str().join(self.charList))



    def truncateLabel(self, text, maxTextLen):
        # ctc_loss can't compute loss if it cannot find a mapping between text label and input
        # labels. Repeat letters cost double because of the blank symbol needing to be inserted.
        # If a too-long label is provided, ctc_loss returns an infinite gradient
        cost = 0
        for i in range(len(text)):
            if i != 0 and text[i] == text[i-1]:
                cost += 2
            else:
                cost += 1
            if cost > maxTextLen:
                return text[:i]
        return text

    def trainSet(self):
        "switch to randomly chosen subset of training set"
        self.dataAugmentation = True
```

```
        self.currIdx = 0
        random.shuffle(self.trainSamples)
        self.samples = self.trainSamples[:self.numTrainSamplesPerEpoch]


    def validationSet(self):
        "switch to validation set"
        self.dataAugmentation = False
        self.currIdx = 0
        self.samples = self.validationSamples
        # self.samples = self.test1Val # experience

    def getIteratorInfo(self):
        "current batch index and overall number of batches"
        return (self.currIdx // self.batchSize + 1, len(self.samples) // self.batchSize)

    def hasNext(self):
        "iterator"
        return self.currIdx + self.batchSize <= len(self.samples)


    def getNext(self):
        "iterator"
        batchRange = range(self.currIdx, self.currIdx + self.batchSize)
        gtTexts = [self.samples[i].gtText for i in batchRange]
        imgs = [preprocess(cv2.imread(self.samples[i].filePath, cv2.IMREAD_GRAYSCALE), self.imgSize,
self.dataAugmentation) for i in batchRange]
        file_path_t = [self.samples[i].filePath for i in batchRange]
        self.currIdx += self.batchSize
        return Batch(gtTexts, imgs), file_path_t
```