УДК 004.89
DOI https://doi.org/10.32782/2663-5941/2022.5/12

Жученко О.А.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Коротинський А.П.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Цапар В.С.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Федотов В.В.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

НЕЙРОМЕРЕЖЕВИЙ КЛАСИФІКАТОР АВТОМАТИЧНОЇ СИСТЕМИ ОБРОБКИ ДОКУМЕНТІВ

Робота спрямована на вирішення прикладної задачі розробки автоматичної системи обробки електронних документів, а саме однієї з її частин - класифікатора. Для вирішення поставленої задачі запропоновано використовувати підходи машинного навчання та штучного інтелекту. Вирішення поставленої задачі при звичайних умовах не складає труднощів, проте в даній роботі розглядається випадок обмеженості навчальної вибірки, що є поширеним випадком при розробці систем на базі запропонованих підходів. У роботі проведено дослідження початкових даних на основі яких буде проведено навчання моделі та визначено кількість класів, що будуть розпізнаватись, кількість представників у кожному класі та досліджено особливості їх представлення. У роботі представлено підходи, застосування яких дозволяє збільшити точність систем такого типу, в умовах обмеженості початкової начальної вибірки. Серед запропонованих підходів розглядається принцип мінімізації параметрів при формуванні архітектури штучної нейронної мережі, аугментація даних, переднавчання штучної нейронної мережі шляхом застосування автоенкодера. Отримана точність в 94-95 %, після застосування запропонованих підходів на відміну від 70 % початкових, підтверджує можливість оперативної розробки аналогічних класифікаторів такого типу, при обмеженій вибірці та в умовах мінімізації часу, з досягненням високих показників точності.

Ключові слова: *нейронні мережі, аугментація, автоенкодер, переднавчання, класифікатор.*

Постановка проблеми. В даний час, діяльність будь яких закладів і установ неможлива без організації документообігу. Необхідність отримання документів, перевірки достовірності, внесення інформації до відповідних баз даних і т.д., займає багато часу операторів відповідних систем, а також необхідно враховувати людський фактор. Заміна етапів ручної обробки програмною дозволить зменшити час обробки документів та прибере вплив людського фактору на точність виконання робіт. З чого випливає, що вирішення прикладної задачі розробки автоматичної системи обробки документі є досить актуальною в поточний час широкого впровадження електронного документообігу.

Запропоновані системи дозволяють вирішувати ряд задач, таких як: перевірка наявності печатки або підпису, перевірка наявності та валідності дати, парсинг ключових елементів документу, тощо. У роботі вирішується задача розробки автоматичної системи обробки скан/фото атестатів, як прикладу документа що обробляється. Опрацювання даного документа полягає у внесенні оцінок в базу даних, перерахунку середнього балу атестату та перевірки його відповідності до поданого значення з метою уникнення фальсифікації даних. Обробка великої кількості атестатів операторами, тобто ручна обробка, займає велику кількість часу та призводить до неминучих помилок операторів. Відтак, заміна етапу ручної перевірки програмною, дозволить зменшити час обробки документів та підвищить точність їх перевірки.

В даній роботі розглядається вирішення лише однієї складової системи автоматичного розпіз-

навання скан/фото документів, а саме класифікації зображень отриманих після опрацювання документа сегментатором, який виділяє складові документи які потребують класифікації. Класифікатор пропонується виконати на основі штучної нейронної мережі, яка вирішує класичне завдання розпізнавання образів.

Відомо, що основна частина часу, при розробці нейронних мереж, складає саме опрацювання вхідних даних для їх зручного представлення. В даному випадку необхідно з кожного документу вирізати визначений, унікальний по своєму наповненні сегмент, який буде в подальшому оброблятися, зберегти його у новий файл та згенерувати йому відповідну мітку. Необхідно зазначити, що описана вище робота, з підготовки достатньої кількості даних для навчання нейронної мережі, потребує наявності великої кількості оригіналів документів, що обробляються, та займає досить багато часу. Для деяких завдань, забезпечення зазначених умов може бути неможливим (відсутність достатньої кількості документів для обробки), або не доцільним (час підготовки даних та навчання нейромережі в декілька разів перевищує час виконання обробки в ручному режимі). Враховуючи зазначене а саме, складність виконання умов щодо кількості вхідної інформації та часу, пропонується розробити класифікатор в умовах обмеженості вибірки, тобто зменшивши час ручної підготовки вибірки для навчання мережі. Практична реалізація штучної нейронної мережі, дослідження роботи та аналіз результатів пропонується виконати на прикладі класифікатора атестатів оцінок шкільної дванадцятибальної системи оцінювання.

Необхідність навчання моделі класифікації зображень на невеликому обсязі даних – звичайна ситуація, з якою часто стикаються в практиці розпізнаванням образів за допомогою технологій комп'ютерного зору на професійному рівні. Під «невеликим» об'ємом розуміється від декількох сотень до декількох десятків тисяч зображень [1].

У цьому матеріалі розглядається проста стратегія вирішення даного завдання:

В частині 1, розроблено невелику нейронну мережу за принципом мінімізації параметрів, щоб задати базовий рівень точності класифікації. Після чого представлено варіант ефективного способу розширення початкової навчальної вибірки для задачі розпізнавання образів – data augmentation. Розглянуто два основних прийоми глибокого навчання на невеликих наборах даних: виділення ознак з використанням попередньо навченої мережі і донавчання попередньо навченої мережі.

В частині 2, буде розглянуто можливість застосування генеративно-змагальної мережі (GAN – Generative Adversarial Networks) як способу розширення початкової вибірки з метою покращення рівня точності розпізнавання. Фінальне значення точності роботи розробленої моделі буде отримано після проведення регуляризації.

Аналіз останніх досліджень і публікацій. На сьогоднішній день існує досить велика кількість матеріалів, що описують способи використання нейронних мереж для схожих задач.

Наприклад, робота [2] присвячена розробці системи розпізнавання писемних символів за допомогою штучної нейронної мережі. Проте у самому матеріалі не описана структура нейронної мережі, а відтак не зрозумілі вимоги та потреби на її навчання.

Авторами роботи [3] досліджено основні методи використання згорткових нейронних мереж для вирішення задачі класифікації текстів. Експерименти на текстових даних великого обсягу показали, що згорткові нейронні мережі для задачі класифікації текстів дозволяють досягти якості, аналогічної або кращої в порівнянні з традиційними методами. Проте аналогічно до вищесказаного у самому матеріалі не описана структура запропонованої нейронної мережі.

У роботі [4] описано синтез згорткової нейронної мережі для розпізнавання рукописних цифр на базі класичного датасету MNIST. Матеріали описують вирішення класичної задачі з урахуванням досить великої вибірки. При вирішення задачі використовувався принцип регуляризації.

У роботі [5] проведено аналіз роботи чотирьох різноманітних моделей розпізнавання рукописних цифр на базі класичного датасету MNIST при навчанні даних моделей на різному по величині об’ємові вибірки, а саме 25, 50, 75 та 100 %. Навіть при 25 % об’ємові вибірки кількість екземплярів складає 15000 значень, що складно назвати обмеженим дата сетом. Проте аналогічно до вищесказаного у самому матеріалі не описана структура нейронної мережі.

У роботі [6] роботі розглядається вдосконалення традиційного методу розпізнавання зображень на базі загорткових нейронних мереж. Було зроблено висновки, щодо доцільності вибору однієї з трьох функцій активації. З результатів дослідження зроблено висновки, що ручне коригування параметрів все ще потрібно в процесі

експерименту для визначення найкращих параметрів розробки та використання нейронних мереж.

У роботі [7] представлена реалізація штучної нейронної мережі для реалізації класифікації рукописної бази даних MNIST. Крім того, у роботі, намагались досягти стиснення зображення за допомогою Autoencoder та визначити його ефективність. Це дозволяє зменшити розмір мережі та таким чином збільшують покращити продуктивність її роботи. Матеріали описують вирішення класичної задачі з урахуванням досить великої вибірки.

Формулювання цілей статті. Метою даної роботи є розроблення та дослідження роботи однієї із складових автоматичної системи оброки документів, а саме класифікатора, на прикладі класифікатора оцінок в атестаті, в умовах обмеженості навчальної вибірки. Робота спрямована на визначення доцільності застосування підходів підвищення рівня точності таких систем в умовах обмеженості вибірки та визначення точності роботи розробленого класифікатора після застосування запропонованих підходів.

Виклад основного матеріалу. Аналіз навчальної вибірки. Аналіз та опрацювання вибірки початкових даних, на основі яких буде проведено навчання моделі, є надзвичайно важливим етапом при розробці систем на базі машинного навчання. Саме на цьому етапі розробки можна визначити кількість класів, що будуть розпізнаватись, кількість їх представників, особливості представлення, тощо.

В нашому випадку об'єктом розпізнавання виступають зображення, що будуть отримані в результаті опрацювання документів сегментатором, наприклад: табелів успішності, додатків до атестату про середню освіту та інше. Зазначені документи містять у собі перелік предметів та відповідних оцінок, причому оцінки можуть бути як рукописними цифрами або словами так і друкованими (див рис. 1).

Рис. 1. Різні типи представлення оцінок

Прийнято рішення, що друковані та рукописні екземпляри можна розглядати як один клас, оскільки дані елементи повинні мати схожі патерни. Відтак, кількість унікальних класів для запропонованої моделі складатиме 24, а саме 12 класів рукописних та друкованих словесних оцінок та 12 класів рукописних та друкованих цифр. Можлива ситуація, коли в сформованому сегментатором області з тих чи інших причин оцінки не буде, а тому для пустого поля необхідно визначити окремий клас. Відповідно до вище сказаного, попередньо для вирішення задачі класифікації було визначено 25 класів.

Попередній аналіз 30 додатків до атестатів показав, що загальна кількість унікальних за представленням (різний шрифт, почерк і .т.д.) оцінок в атестатах не рівномірна для кожного з класів, а для деяких класів взагалі складає нуль. Класи що відповідають оцінками «один», «два», «три», «1», «2», «3», «4», «5» взагалі не мають представників у вибірці 30 атестатів. Прослідковується наявність двох груп класів оцінок, а саме цифр та рукописних, при чому загальна кількість представників в групах відрізняється майже в 8 разів.

З попереднього аналізу даних випливає дві проблеми: відсутність представників деяких класів та нерівномірність представників поміж всіма класами.

Зрозуміло, що навчання нейронної мережі без представників деяких класів не має сенсу. Вирішення даної проблеми можливе різними способами, наприклад: розширення вибірки за рахунок існуючих датасетів, наприклад таких як «MNIST», що не є універсальним вирішенням для схожих задач; розширення вибірки шляхом обробки більшої кількості атестатів або штучне створення представників класів. Оскільки задача вирішується саме при умові обмеженості матеріалів для навчання мережі, будемо вважати що вже існуючих вибірок для розширення початкової вибірки, як і додаткових документів для опрацювання не існує, відтак обираємо останній варіант.

Значна нерівномірність представників поміж всіма класами передбачає, що нейромережа навчиться краще розпізнавати ті класи в яких представників більше, що також можна вирішити описаним вище варіантом. Проте, цікавим є факт, що деякі групи класів мають схожу розподіленість. Відтак можливо розробити окремо класифікатор для кожної з груп класів, що вирішить проблему нерівномірності представників поміж групами класів.

Після ручного вирівнювання розподілення представників поміж всіма класами отримано наступне представлення даних: для класів 13-24 з середнім значенням кількості представників 12, для класів 1-12 з середнім значенням кількості представників 26.

Вибір структури нейронної мережі.

Навчання нейронної мережі це процес оптимізації параметрів обраної моделі на навчальній

вибірці з метою досягти високого рівня узагальнення, тобто точності роботи даної моделі на даних, які раніше не застосовувались.

Відомо, що найкращий спосіб отримання моделі без перенавчання, тобто коли модель завчила всі шаблони тренувальних вибірки, що не обов'язково характерні тестовій – збільшення об'єму тренувальних даних. Модель, навчена на великому обсязі даних, матиме велику узагальненість, до того ж велика вибірка дозволяє збільшувати глибину самої мережі. З іншої сторони обмежений об'єм тренувальної вибірки обумовлює мінімізацію параметрів для навчання, а відтак глибини мережі, тобто мережі доведеться вивчати зжаті представлення. У той же час, модель повинна мати достатню кількість параметрів, щоб не виник ефект недонавчання. В зв'язку з обмеженістю даних, для розроблення нейромережі класифікації пропонується використати принцип мінімізації структури нейронної мережі.

Для вирішення класичної задачі розпізнавання малюнків зазвичай використовують добре зарекомендувавший себе підхід чередування згорткових та агрегувальних шарів. Відтак, враховуючи все вище сказане запропонована наступна структура/глибина класифікатора для розпізнавання сегментів додатків до атестатів, що складає в себе двічі чередовані згорткові (convolutional layer) та агрегувальні шари (max pooling layer), вирівнювальний шар (flatten layer) та два повноз'єднаних шара (dense layer).

Розглядається два варіанта реалізації, як один загальний класифікатор на 25 класів та два окремих класифікатора для цифр та слів.

Відповідно до сказаного вище, глибина нейронних мереж визначена, а параметри шарів такі як кількість фільтрів, їх розміри, кількість нейронів, функції активації і т.д. підбиралися індивідуально до кожного з класифікаторів. В результаті підбору було отримано дві структури: перша структура відповідає класифікатору цифр; друга структура відповідає класифікатору слів та загальному класифікатору. Більш детальний опис структури з числовими характеристиками можна побачити на рис. 2 [8].

Загальна кількість параметрів описаних вище структур нейронних мереж досить велика, наприклад для класифікатора цифр складає 16614, для загального класифікатора 37124. Оптимізації такої кількості параметрів передбачає велику кількість екземплярів для навчання, а тому без програмного розширення навчальної вибірки не обійтись.

Аугментація, попередня обробка даних та навчання нейронної мережі.

Для розширення навчальної вибірки використовується підхід аугментації даних (data augmentation) – це методика створення додаткових навчальних даних з наявних даних. У цьому випадку до екземпляру з навчальної вибірки застосовуються виконання повороту картинки на випадковий з заданого діапазону градус, випадкове масштабування, здвиг по всім осям. Відтак, з початкових 380 унікальних екземплярів, після аугментації данних було отримано вибірку з розміром 4000.

Перед передачею в мережу дані повинні бути перетворені в тензори. В даний час дані зберігаються у вигляді файлів JPEG, тому їх потрібно підготувати для передачі в мережу, виконавши наступні кроки: декодувати вміст з формату JPEG в таблиці пікселів; провести зріз порогових значень з метою прибрати шум на зображеннях, змінити масштаб значення пікселів з діапазону [0, 255] в діапазон [0, 1].

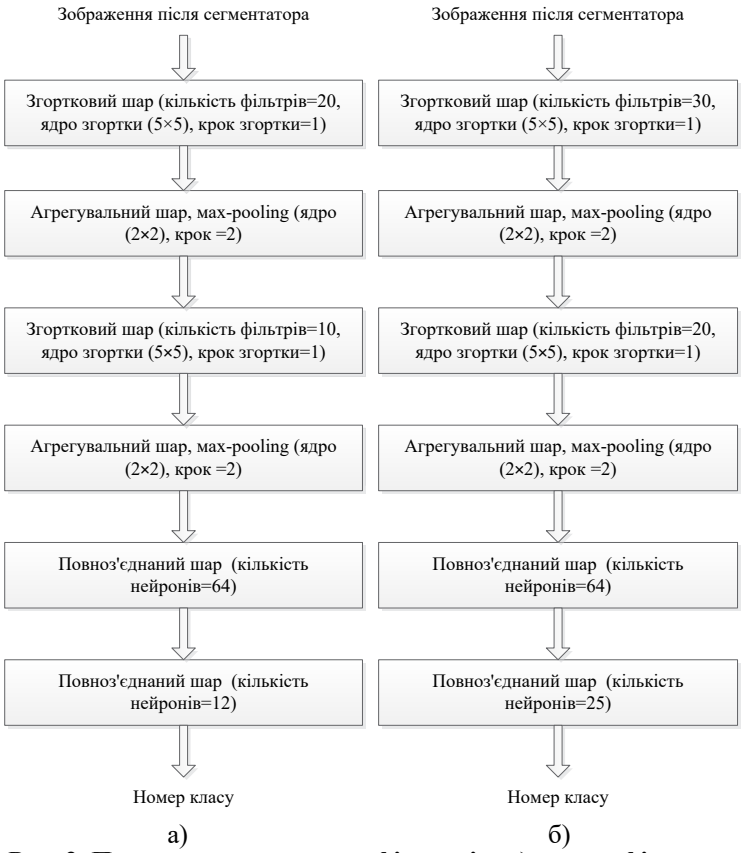

Рис. 2. Повна структура класифікаторів: а) - класифікатора цифр, б) - загального класифікатора та класифікатора слів

У якості метрики для дослідження ефективності запропонованих моделей використовуються стандартні функцію втрат categorical_crossentropy та accuracy – частка правильних відповідей алгоритму.

Після обрання відповідної структури моделі та проведення попередньої обробки даних, проведено навчання загальної нейронної мережі для отримання базового рівня точності розпізнавання, а також проведено навчання нейронних мереж для груп класів для визначеності правильності запропонованого підходу. Останнім було проведено навчання описаних мереж з використанням аугментації даних, результати навчання наведені на рис. 3-5.

З результатів проведеного навчання штучних нейронних мереж видно, що для всіх випадків навчання, кількість епох навчання можливо зменшити без втрати валідаційної точності, а відтак зменшити час необхідний для навчання запропонованих мереж.

Використання автоенкодера для переднавчання класифікаторів.

Автоенкодер (Autoencoder) – це особлива архітектура нейронної мережі, основна задача якої відновлення вхідного сигналу на виході з мережі, шляхом відновлення сигналу з його стиснутої репрезентації. В цьому випадку вхідний сигнал відновлюється з помилками через втрати під час проходження прихованого шару, проте, щоб їх мінімізувати, мережа змушена вчитися відбирати найбільш важливі ознаки в прихованих шарах. Ця особливість лежить в можливості використання автоенкодера для переднавчання. Основною метою роботи автоенкодерів – отримати на вихідному шарі відгук, найбільш близький до вхідного, а відтак структурна характерна особливість автоенкодерів – кількість нейронів у вхідному та у вихідному шарі збігається.

Враховуючи, що структури запропонованих класифікаторів відомі, запропонована наступна структура автоенкодера для переднавчання, що відповідає, двом чередуванням згорткових та агрегувальних шарів, згорткового шару по середині та знову двох чередування загорткових та агрегувальних шарів (див. рис. 6).

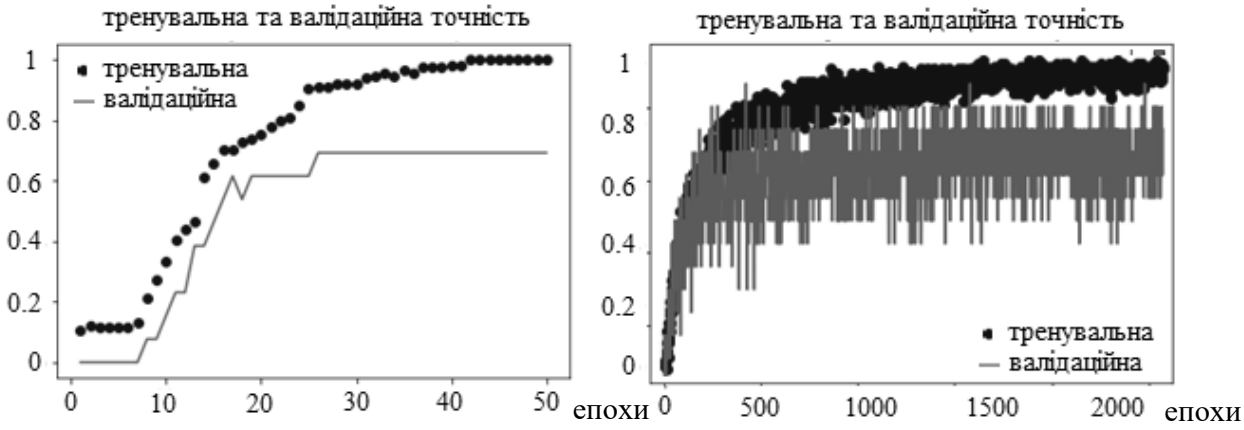

Рис 3. Результати навчання класифікатора цифр без аугментації даних та з аугментацією даних

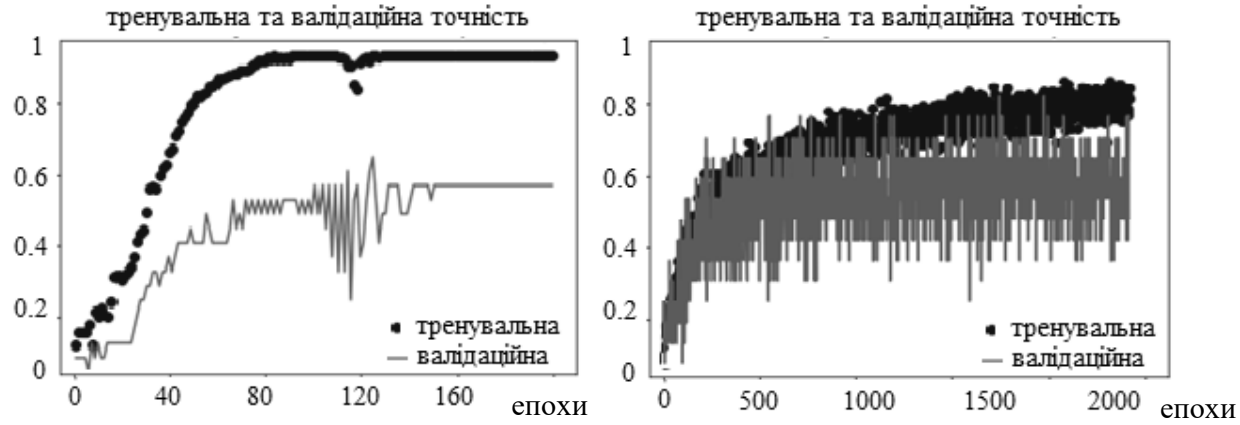

Рис 4. Проміжні дані навчання класифікатора слів без аугментації даних та з аугментацією даних

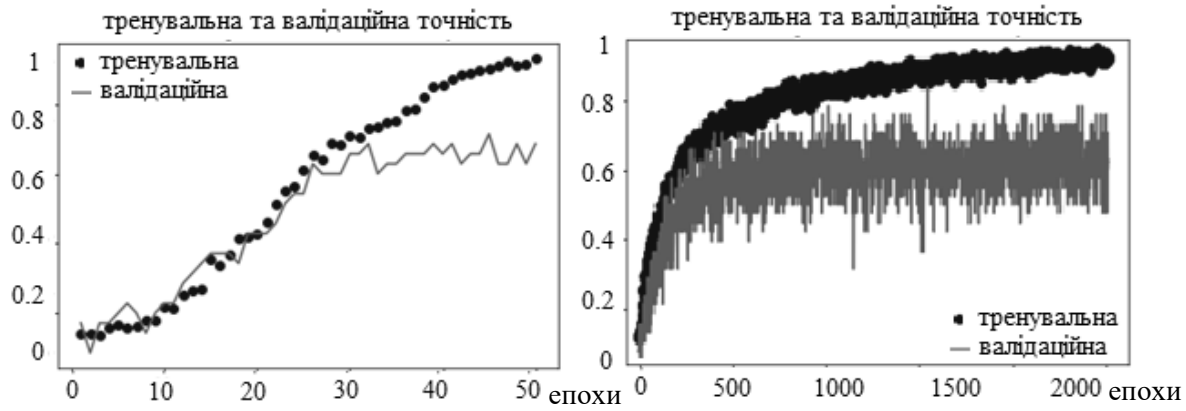

Рис 5. Проміжін дані навчання загального класифікатора без аугментації даних даних та з аугментацією даних

Навчання автоенкодера відбувається на даних отриманих після аугментації, а відтак не обумовлює проблеми їх обмеження, а при доданні до якісно навченого енкодера (саме енкодера – першої половини автоенкодера, що частково відповідає структурі класифікатора) повноз'єднаних шарів дозволяє отримати нейромережу, що відповідає структурі класифікатора з не навченими лише повноз'єднаними шарами [9].

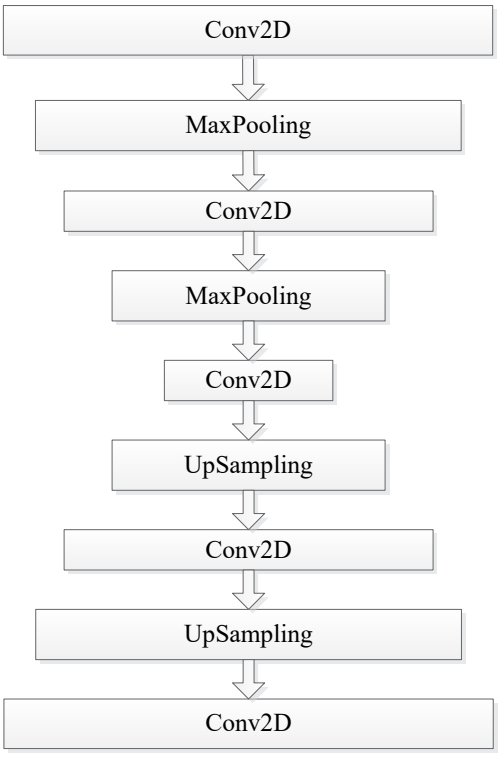

Рис. 6. Запропонована структура автоенкодера

Заморозивши ті шари нейронної мережі, які були попередньо навчені при синтезі автоенкодера та провівши донавчання повноз'єднаних шарів було отримано готові нейронні мережеві класифікаторів з наступними параметрами точності (див. табл. 1).

Таблиця 1

Результати навчання нейронних мереж

Параметри	Класифікатор цифр з передн-авчанням	Класифікатор слів з передн-авчанням	Загальний класифікатор з передн-авчанням
Кількість параметрів мережі	16614	37124	37124
Величина навчальної вибірки	114	266	380
Валідаційна точність моделі	0.9411	0.953	0,938

Висновки. У роботі розглянуто підходи до розробки складової системи автоматичної обробки документів, а саме класифікатора в умовах обмеженості навчальної вибірки. У якості екземплярів для розроблення, дослідження ефективності та точності запропонованих підходів та систем використовувалась вибірка з 30 атестатів оцінок шкільної дванадцятибальної системи оцінювання, а саме 380 представників.

Показано, що оперативне розроблення класифікаторів такого типу можливе при використанні запропонованих підходів аугментації даних та переднавчання на базі автоенкодера. Отримане початкове порогове значення точності класифікатора в межах 65%, отримана точність в 95 % після застосування запропонованих підходів свідчить, що при розробці аналогічних систем, при обмеженій

навчальній вибірці, можливе досягнення високих показників точності. Отримана точність навіть без регуляризації обумовлює отримання класифікаторів з більш високим значенням точності.

В результаті попереднього налізу початкової вибірки було прийнято рішення по розробці двох окремих класифікаторів для двох окремих груп класів. Згідно з отриманих результатів, запропонований підхід дозволив підняти точність загальної класифікації атестатів на декілька відсотків. Звідки слідує, що для визначення унікальних, або непоширених екземплярів має місце навчання принципово окремої, персональної структури/моделі.

Список літератури:

1. Франсуа Шолле. Глубокое обучение на Python. – СПб.: Питер, 2018. – 400 с.: ил. – (Серия «Библиотека программиста»). ISBN 978-5-4461-0770-4
2. Соколенко Д. Г., Корнага Я. І. «Система розпізнавання писемних символів за допомогою нейронної мережі», *Вчені записки ТНУ імені В.І. Вернадського. Серія Технічні науки* Том 29 (68) Ч. 2 № 5 2018 с. 56-58
3. Карпович Артем Валерійович «Використання згорткових нейронних мереж для задачі класифікації текстів», *International scientific journal «Internauka»* // № 14(54), 2018 // Technical sciences // с. 69-73
4. Orhan G. Yalçın Image Classification in 10 Minutes with MNIST Dataset, URL: https://towardsdatascience.com/image-classification-in-10-minutes-with-mnist-dataset-54c35b77a38d
5. Feiyang Chen, Nan Chen, Hanyang Mao, Hanlin Hu Assessing Four Neural Networks on Handwritten Digit Recognition Dataset (MNIST) / *Chuangxinban journal of computing*, june 2018, URL: https://arxiv.org/pdf/1811.08278. pdf
6. Yifan Wang, Fenghou Li, Hai Sun, Wenbo Li, Cheng Zhong, Xuelian Wu, Hailei Wang, Ping Wang Improvement of MNIST Image Recognition Based on CNN, 7th Annual International Conference on Geo-Spatial Knowledge and Intelligence IOP Conf. Series: Earth and Environmental Science 428 (2020), URL: https://iopscience.iop.org/article/10.1088/1755-1315/428/1/012097/pdf
7. Wan Zhu Classification of MNIST Handwritten Digit Database using Neural Network, URL: http://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_117.pdf
8. Korotynskyi, A., Zhuchenko, O. Development of a classifier for the system of automatic document processing with limited sampling, ATIT 2020 – Proceedings: 2020 2nd IEEE International Conference on Advanced Trends in Information Theory, 2020, стр. 349–352
9. A. Korotynskyi, O. Zhuchenko A system of automated control for the baking process that minimizes the probability of defects, Eastern-European Journal of Enterprise Technologies, 2020, (2-103), стр. 58–67.

Zhuchenko O.A., Korotynskyi A.P., Tsapar V.S. Fedotov V.V. **NERAL NETWORK CLASSIFIER OF AUTOMATION DOCUMENT PROCESSING SYSTEMS**

The work is aimed at solving the applied problem of developing an automatic system for processing electronic documents, namely one of its parts of the classifier. To solve this problem, it is proposed to use approaches to machine learning and artificial intelligence. Solving this problem under normal conditions is not difficult, but this paper considers the case of limited training sample, which is a common case in the development of systems based on the proposed approaches. The study of initial data on the basis of which the model will be taught and the number of classes to be recognized, the number of representatives in each class and the peculiarities of their presentation. The paper presents approaches, the application of which allows to increase the accuracy of systems of this type, in the conditions of limited initial initial sampling. Among the proposed approaches, the principle of minimizing the parameters in the formation of the architecture of the artificial neural network, data augmentation, pre-training of the artificial neural network by using an autoencoder. The obtained accuracy of 94-95%, after the application of the proposed approaches in contrast to 70% of the original, confirms the possibility of rapid development of similar classifiers of this type, with limited sampling and time minimization, achieving high accuracy.

Key words: *neural networks, augmentation, autoencoder, pre – learning, classifier.*