Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Глибинна рекурентна нейронна мережа та її навчання в завданні розпізнавання мови по губах
(тема)

Виконав:
студент 2 курсу, групи СШІм-20-1

(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки

(код і повна назва спеціальності)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту

(повна назва спеціалізації)

Керівник проф. Бодянський Є.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Онищенко Аллі Георгіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Глибинна рекурентна нейронна мережа та її навчання в завданні розпізнавання мови по губах

затверджена наказом університету від 08 листопада 20 21 р. № 1695Ст

2. Термін подання студентом роботи до екзаменаційної комісії 08 грудня 20 21 р.

3. Вихідні дані до роботи Глибинна рекурентна нейронна мережа та її навчання в завданні розпізнавання мови по губах

4. Перелік питань, що потрібно опрацювати в роботі Вступ; 1 Аналіз предметної галузі та постановка задачі;1.1 Аналіз предметної галузі; 1.2 Постановка задачі; 1.3 Аналіз методів машинного навчання; 1.3.1 Історія машинного навчання; 1.3.2 Методи машинного навчання; 1.3.3 Основні задачі навчання ШНМ; 2 Глибинна рекурентна нейронна мережа та її навчання в завданні розпізнавання мови по губах; 2.1 Автоматичне читання по губах; 2.2 Архітектура нейромережі LipNet; 2.3 Gated Recurrent Unit; 2.4 Корпус GRID; 2.5 Втрати CTC; 3 Оцінювання lipnet на корпусі grid; 3.1 Data Augmentation; 3.2 Baselines; 3.3 Оцінка ефективності; 4 Модифікована активаційна функція для gru що не потерпає від ефекту зникаючого градієнту; 5 Програмна реалізація

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Рисунок 2.1 – Ділянки обличчя розпізнавання салієнтних ознак для слова «please»; Рисунок 2.2 - Ділянки обличчя розпізнавання салієнтних ознак для слова «lay»; Рисунок 2.3 – Архітектура нейромережі LipNet; Рисунок 4.1 – Узагальнений формальний нейрон; Рисунок 4.2 – Активаційні функції узагальненого нейрону; Рисунок 4.4 – Активаційна функція у вигляді степінного ряду; Рисунок 4.5 – Сигмоїдальна функція; Рисунок 4.6 – Сигмоїдальна функція на діапазоні [-2;2]

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання	01.09.2021	виконано
2	Аналіз науково-технічних джерел інформації	02.09.2021 – 10.09.2021	виконано
3	Аналіз проблематики теми кваліфікаційної роботи	11.09.2021 – 25.09.2021	виконано
4	Аналіз предметної області кваліфікаційної роботи	26.09.2021 – 08.10.2021	виконано
5	Постановка задачі дослідження	09.10.2021 – 15.10.2021	виконано
6	Аналіз методів для вирішення поставленої задачі	16.10.2021 – 20.10.2021	виконано
7	Програмна реалізація	21.10.2021 – 19.11.2021	виконано
8	Оформлення пояснювальної записки	20.11.2021 – 24.11.2021	виконано
9	Проходження нормаконтролю	25.11.2021 – 28.11.2021	виконано
10	Перевірка кваліфікаційної роботи на плагіат	29.11.2021 – 30.11.2021	виконано
11	Попередній захист	07.12.2021	виконано
12	Захист	08.12.2021	виконано

Дата видачі завдання 01 вересня 20 21 р.

Студент _____
(підпис)

Керівник роботи _____ _____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 67 с., 15 рис., 2 табл., 2 дод., 85 джерел.

АУГМЕНТАЦІЯ, ЗГОРТКОВА НЕЙРОННА МЕРЕЖА, ЗЧИТУВАННЯ ПО ГУБАХ, МАШИННИЙ ЗІР, МОДЕЛЬ МАРКОВА, СПАТІОТЕМПОРАЛЬНА ХАРАКТЕРИСТИКА, ШТУЧНА НЕЙРОННА МЕРЕЖА, GRID, LIPNET, LONG SHORT-TERM MEMORY

Предмет дослідження – штучні нейронні мережі та методи їх навчання в завданнях розпізнавання відеоданих, що відображають мову глухо-німих людей.

Об’єкт дослідження – процес аналізу масивів відеоданих, що відображають мову людей з обмеженими можливостями.

Методи дослідження базуються на теорії штучних нейронних мереж, теорії оптимізації, теорія розпізнавання образів.

Метою даної роботи є розробка штучної нейронної мережі для вирішення завдання розпізнавання мови глухо-німих на основі відеоспостережень.

РЕФЕРАТ

Записка пояснювальна: 67 с., 15 рис., 2 табл., 2 дод., 85 джерел.

АУГМЕНТАЦИЯ, ИСКУССТВЕННАЯ НЕЙРОННАЯ СЕТЬ, МАШИННОЕ ЗРЕНИЕ, МОДЕЛЬ МАРКОВАЯ, СВЁРТОЧНАЯ НЕЙРОННАЯ СЕТЬ, СПАТИОТЕМПОРАЛЬНАЯ ХАРАКТЕРИСТИКА, СЧИТЫВАНИЕ ПО ГУБАМ, GRID, LIPNET, LONG SHORT-TERM MEMORY

Предмет исследования – искусственные нейронные сети и методы их обучения в задачах распознавания видеоданных, отражающих язык глухонемых людей.

Объект исследования – процесс анализа массивов видеоданных, отражающих язык людей с ограниченными возможностями.

Методы исследования базируются на теории искусственных нейронных сетей, теории оптимизации, теории распознавания образов.

Целью данной работы является разработка искусственной нейронной сети для решения задачи распознавания речи глухонемых на основе видеонаблюдений.

ABSTRACT

Explanatory note: 67 p., 15 fig., 2 tabl., 2 ann., 85 sources.

ARTIFICIAL NEURAL NETWORK, AUGMENTATION, CONVOLUTIONAL NEURAL NETWORK, GRID, LIPNET, LIP-READING, LONG SHORT-TERM MEMORY, MACHINE VISION, MARKOV MODEL, SPATIOTEMPORAL RESPONSE

The subject of research - artificial neural networks and methods of their training in the tasks of video recognition, reflecting the language of deaf and dumb people.

The object of research is the process of analyzing arrays of video data that reflect the language of people with disabilities.

Research methods are based on the theory of artificial neural networks, the theory of optimization, the theory of pattern recognition.

The aim of this work is to develop an artificial neural network to solve the problem of speech recognition of the deaf and dumb on the basis of video surveillance.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ І ТЕРМІНІВ

ШНМ – штучні нейронні мережі;

CER – Character Error Rate – коефіцієнт помилок в символах;

CNN – Convolutional Neural Networks – Згорткові нейронні мережі;

HMM – Hidden Markov Models – Приховані моделі Маркова;

STCNN – Spatiotemporal Convolutional Neural Networks – Спатіотемпоральна згорткова нейронна мережа;

WER – Word Error Rate – коефіцієнт помилок в словах.

ВСТУП

В наші часи досить розвинута середа додатків для спрощування життя людини. В тій чи іншій мірі більшість з них використовують сучасні технології штучного інтелекту. Серед найвідоміших технологій які нам відомі є розпізнавання голосу та управління за допомогою голосу додатком. Найяскравішим прикладом такого управління є Siri у продукції Apple. Siri – це штучний інтелект, який вміє розпізнавати твій голос, виконувати команди за допомогою управління голосом, а також вона вміє самонавчатися. Також досить поширеним направленням для використання штучних нейронних мереж є обробка зображення, тобто розпізнавання певних образів за допомогою навчання нейронних мереж. Досить складно охопити перелік абсолютно всіх існуючих галузей використання штучних нейронних мереж, адже час йде, а їх розвиток не припиняється, а тому перелік можливостей також збільшується.

Все ж таки є інноваційна технологія, яка ще не є досить популярною та відомою, а саме – застосування штучних нейронних мереж для зчитування по губах. Якщо дивитися з точки зору людства, то навичка читання по губах є досить корисною, особливо для людей з обмеженими можливостями, адже саме завдяки поєднанню жестів для читання по губах вони можуть розуміти один одного та оточуючих людей. Але все ж таки більшість людей не знають мову жестів, тому навичка читати по губах стає в нагоді досить часто. Перші експерименти на цю тему почали проводити ще у 1976 році. Проведені експерименти показали, що люди «чують» зовсім інші фонеми, якщо накласти невірний звук на відео, де людина рухає губами, щось кажучи. У воєнні часи навик зчитування по губах використовували для «прослуховування» ворогів через бінокль та інші пристрої. В свою чергу, зараз такий навик поширено використовується у розвідці або сфері безпеки.

Щодо автоматичних систем читання по губах, вони мають досить великий потенціал для їхнього майбутнього розвитку. Таку систему можна застосувати будь-де. Це може бути і слухові апарати, які будуть розпізнавати мовлення; системи для беззвучних лекцій у публічних містах, якщо для цього є потреба; біометрична ідентифікація; системи скритої передачі інформації для розвідки; для поліції було б дуже корисно мати можливість зчитати, що кажуть на відеозаписі з камер відеоспостереження та багато іншого.

Метою даної магістерської кваліфікаційної роботи є розробка глибинної рекурентної нейронної мережі та її навчання в завданні розпізнавання мови по губах.

Методи дослідження базуються на рекурентній нейронній мережі LipNet типа LSTM (long short-term memory).

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз предметної галузі

Предмет дослідження – штучні нейронні мережі та методи їх навчання в завданнях розпізнавання відеоданих, що відображають мову глухо-німих людей.

Об’єкт дослідження – процес аналізу масивів відеоданих, що відображають мову людей з обмеженими можливостями.

Методи дослідження базуються на теорії штучних нейронних мереж, теорії оптимізації, теорія розпізнавання образів.

Метою даної роботи є розробка штучної нейронної мережі для вирішення завдання розпізнавання мови глухо-німих на основі відеоспостережень.

1.2 Постановка задачі

Вихідною інформацією є масиви відеоданих, що відображають процес «мовлення» людей з фізичними вадами (глухо-німих), на основі якої вирішується задача розпізнавання мови (по губах) за допомогою модифікованої рекурентної штучної нейронної мережі типу LSTM.

1.3 Аналіз методів машинного навчання

В наші часи досить швидко розвивається таке поняття, як Машинне Навчання (МН). Машинне навчання – це метод аналізу даних, який автоматизує побудову аналітичної моделі. Це галузь штучного інтелекту, заснована на ідеї, що системи можуть вчитися на основі даних, визначати закономірності та приймати рішення з мінімальним втручанням людини.

Машинне навчання базується на такому понятті, як штучні нейронні мережі (ШНМ). Принцип роботи ШНМ намагається максимально точно та повністю повторити функціонал біологічних нейронних мереж людини. Основний принцип роботи МН в тому, що комп’ютери можуть самонавчатися без заздалегідь запрограмованих конкретних завдань.

1.3.1 Історія машинного навчання

Першу програму на основі алгоритмів, здатних самонавчатися, розробив Артур Самуель (Arthur Samuel) в 1952 році, призначена вона була для гри в шашки. Самуель дав і перше визначення терміну «машинне навчання»: це «область досліджень розробки машин, які не є заздалегідь запрограмованими». Більш точне визначення терміну «навчання» дав набагато пізніше Т. М. Мітчелл: кажуть, що комп'ютерна програма навчається на основі досвіду E по відношенню до деякого класу задач T і заходу якості P, якщо якість вирішення завдань з T, виміряний на основі P, поліпшується з набуттям досвіду E.

Вже в 1957 році була запропонована перша модель нейронної мережі, що реалізує алгоритми машинного навчання, схожі на сучасні. В даний час ведеться розробка самих різних систем машинного навчання, призначених для використання в таких технологіях майбутнього, як Інтернет Речей, Промисловий Інтернет Речей, в концепції «розумне» місто, при створенні безпілотного транспорту і в багатьох інших.

Про те, що на машинне навчання зараз покладають великі надії, свідчать такі факти:

– у компанії Google вважають, що скоро її продукти «перестануть бути результатом традиційного програмування – в їх основу буде покладено машинне навчання»;

– компанії Google, Facebook, Apple, Amazon, Microsoft і китайська фірма Baidu вступили в боротьбу за талановитих фахівців у сфері штучного інтелекту;

– Марк Цукерберг, генеральний директор Facebook, особисто – по телефону і по відеочату – бере участь в спробах його компанії переманити найкращих випускників;

– відвідуваність на найважливіших академічних конференціях в цій сфері збільшилася майже в чотири рази;

– такі нові продукти, як Siri від Apple, M від Facebook, Echo від Amazon були створені за допомогою машинного навчання.

1.3.2 Методи машинного навчання

Взагалі, розрізняють два види машинного навчання:

– навчання по прецедентах (індуктивне навчання);

– дедуктивний навчання.

Оскільки дедуктивний тип навчання прийнято відносити до області експертних систем, то терміни «машинне навчання» і «навчання по прецедентах» можна вважати синонімами. На сьогоднішній день даний метод навчання зараз, як то кажуть, в тренді. Що не скажеш про експертні системи, які нині переживають кризу. Бази знань, які в свою чергу є основою експертних систем, важко узгоджувати з реляційною моделлю даних, тому промислові СУБД неможливо ефективно використовувати для наповнення баз знань експертних систем.

Навчання за прецедентами, в свою чергу, можна поділити на три основних типи:

– контрольоване навчання (навчання з учителем (supervised learning));

– неконтрольоване навчання (unsupervised learning), або навчання без учителя;

– навчання з підкріпленням (reinforcement learning) .

Також ще існують інші методи навчання, такі як:

– активне;

– многозадачне;

– різноманітне;

– трансферне;

– та ін.

Особливо успішно розвивається в останні роки «глибинне навчання», при використанні якого можуть успішно поєднуватися алгоритми навчання з вчителем і без вчителя.

1.3.2.1 Навчання з учителем

Даний вид навчання схематично представлено на рисунку 1.1:

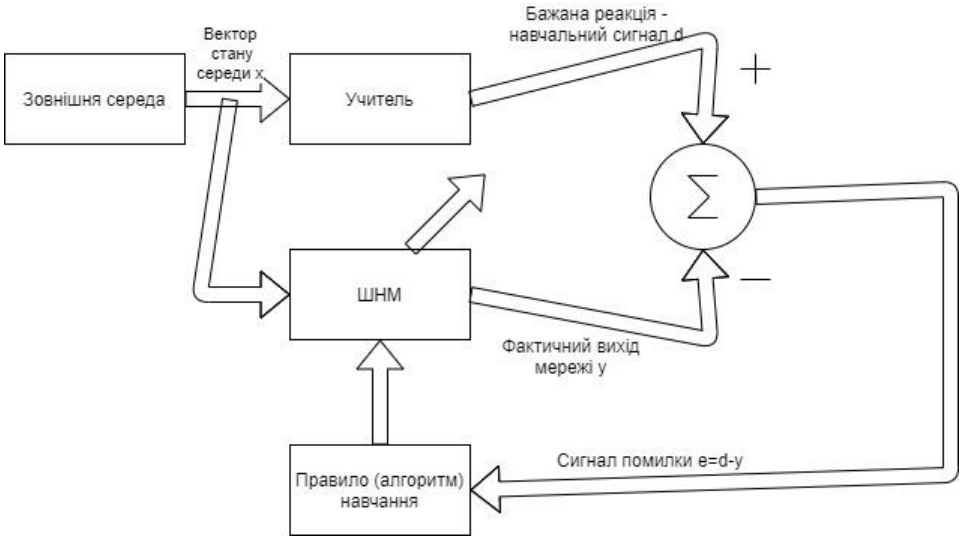

Рисунок 1.1 – Схема навчання з учителем

В даній схемі «вчителю» відома інформація про зовнішнє середовище, задана у вигляді послідовності або пакета вхідних векторів, а також «правильна реакція» на ці сигнали, представлена у вигляді навчального сигналу. Природно, що реакція ненавченої мережі відрізняється від

«правильної» реакції вчителя, в результаті чого виникає помилка. У процесі навчання необхідно так налаштувати параметри ШНМ, щоб деяка скалярна функція від помилки (критерій якості) досягла свого мінімального значення. Навченою вважається мережа, яка в деякому, як правило, статистичному сенсі повторює реакцію вчителя. Оскільки інформація про зовнішнє середовище зазвичай має нестаціонарний характер, процес навчання йде безперервно, для чого використовуються ті чи інші рекурентні процедури.

Даний метод доцільно використовувати при досить великих обсягах даних. Розглянемо приклад з багатьма фотографіями домашніх тварин з маркерами: це кішка, а це собака. Головною задачею є створення алгоритму, за допомогою якого машина могла б за допомогою фотографії, яку «не бачила» раніше, визначити, хто на ній зображений: кішка або собака. У ролі «вчителя» в даному випадку виступає людина, яка заздалегідь проставила маркери. Машина самостійно обирає ознаки, за якими вона відрізняє кішок від собак.

Завдяки такому методу знаходження алгоритму навчання машини, в подальшому знайдений нею алгоритм може бути також застосовано та переналаштовано на рішення іншої задачі, наприклад, на розпізнавання курей і качок. Машина знову сама виконає складну і копітку роботу по виділенню ознак, за якими буде розрізняти цих птахів. А нейромережу, яку навчили розпізнавати кішок, можна швидко навчити обробляти результати комп'ютерної томографії.

1.3.2.2 Навчання без учителя (самонавчання)

Альтернативою для навчання з учителем є навчання без учителя або самонавчання. При навчанні без учителя, реакція на сигнали оточуючого середовища невідома. Процес самонавчання представлено на рисунку 1.2:

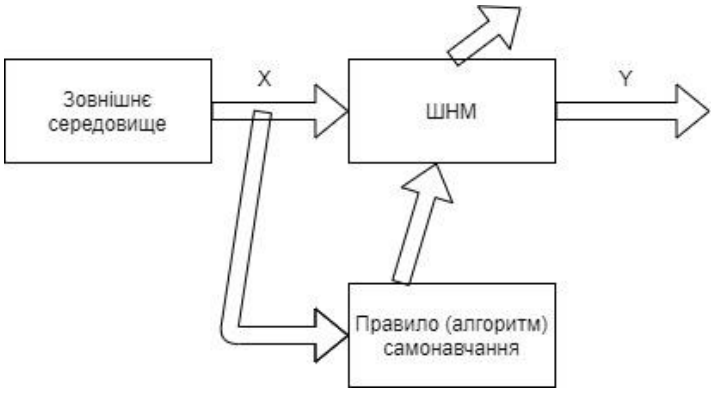

Рисунок 1.2 – Схема навчання без учителя (самонавчання)

Мережі, які реалізують парадигму самонавчання, призначені, як правило, для аналізу внутрішньої латентної структури вхідної інформації і вирішують завдання автоматичної класифікації, кластеризації, факторного аналізу, компресії даних.

Окрім маркерованих даних, для яких використовується навчання з учителем, також існує багато, але даних без міток все ж таки більше. Це зображення без підписів, аудіозаписи без коментарів, тексти без анотацій. Завдання машини при неконтрольованому навчанні – знайти зв'язку між окремими даними, виявити закономірності, підібрати шаблони, упорядкувати дані або описати їх структуру, виконати класифікацію даних. Для таких цілей використовується самонавчання. Навчання без учителя використовується, наприклад, в рекомендаційних системах, коли в інтернет-магазині на основі аналізу попередніх покупок покупцеві пропонуються товари, які можуть зацікавити його з більшою ймовірністю, ніж інші. Або коли після перегляду якогось відеокліпу на порталі YouTube відвідувачеві пропонують десятки посилань на ролики, чимось схожі на переглянутий. Або коли Google у відповідь на один і той же запит ранжує посилання в результатах пошуку для одного користувача інакше, ніж для іншого, оскільки враховує історію пошуків.

1.3.2.3 Навчання з підкріпленням

Навчання з підкріпленням є своєрідною золотою серединою (компромісом) між двома попередніми парадигмами. Частіше за все навчання з підкріпленням плутають з навчанням з заохоченням. Але потрібно ці поняття розрізняти, адже при навчанні з заохоченням доступна лише непряма інформація про правильну реакцію на вхідний сигнал.

На рисунку 1.3 приведена схема процесу навчання з підкріпленням.

Нейронна мережа створює відображення вхідної інформації x у вихідний вектор y у вигляді $y = F(x)$, проте, оскільки навчальний сигнал d в явному вигляді не заданий, неможливо отримати помилку $e = d - y$, на підставі якої відбувається навчання. Передбачається, що є деякі апріорні знання, що дозволяють зв'язати евристичний сигнал підкріплення $\tilde{d}$ з неспостережуваних бажаним виходом d з допомогою деякої функції $\tilde{F}$, що відображає d в $\tilde{d}$. Зазвичай ця функція враховує зв'язок вихідних сигналів мережі y із подіями, які спостерігаються у зовнішньому середовищі, для чого в схему навчання вводиться додатковий блок – «критик» [1], що відображає поведінку мережі в сигнал $\tilde{y} = \tilde{F}(F(x))$. Далі обчислюється евристична помилка $\tilde{e} = \tilde{d} - \tilde{y}$, на основі якої і реалізується процес навчання.

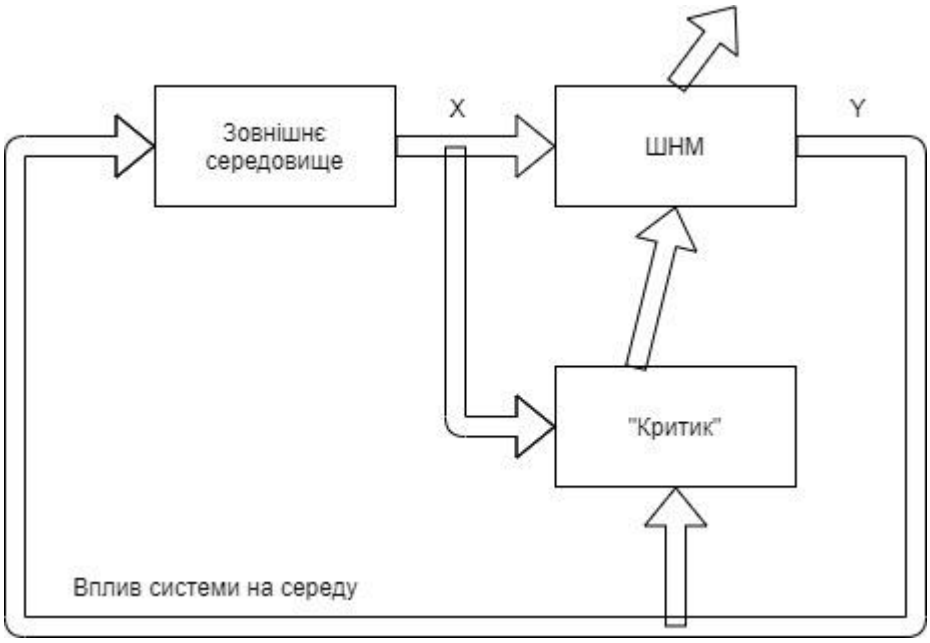

Рисунок 1.3 – Схема навчання з підкріпленням

Процес навчання з підкріпленням розбивається на два відносно незалежних етапи:

– навчання того, як вихідний сигнал мережі впливає на спостережувані змінні середовища, тобто відновлення відображення ;

– навчання мережі на основі мінімізації прийнятого критерію .

Ця парадигма тісно пов'язана з ідеями динамічного програмування [2] і в теорії штучних нейронних мереж відома також як нейродинамічне програмування [1].

Досить широке поширення набула також парадигма змішаного навчання, коли частина параметрів мережі налаштовується за допомогою навчання з учителем, а інша частина або архітектура в цілому – за допомогою самонавчання. Цей підхід набув найбільшого поширення при навчанні радіально-базисних ШНМ.

Таке навчання є окремим випадком контрольованого навчання, але вчителем в даному випадку є «середовище». Машина (її в цій ситуації часто називають «агент») не має попередньою інформацією про середовище, але має можливість виконувати в ній будь-які дії. Середа реагує на ці дії і тим

самим надає агенту дані, які дозволяють йому реагувати на них і вчитися. Фактично агент і середовище утворюють систему зі зворотним зв'язком.

Навчання з підкріпленням використовується для вирішення більш складних завдань, ніж навчання з учителем і без вчителя. Воно використовується, наприклад, в системах навігації для роботів, які навчаються уникати зіткнень з перешкодами досвідченим шляхом, отримуючи зворотний зв'язок при кожному зіткненні. Навчання з підкріпленням використовується також в логістиці, при складанні графіків і плануванні завдань, при навчанні машини логічним іграм (покер, нарди, го та інші).

1.3.3 Основні задачі навчання ШНМ

Звичайно, існує велика різноманітність задач, які розв’язуються за допомогою навчання ШНМ, але також є базові завдання, які найчастіше цікавлять спеціалістів комп’ютерних наук, інженерії та управління.

1.3.3.1 Розпізнавання образів

Штучні нейронні мережі намагаються повністю копіювати функції головного мозку людини. Тому, основні задачі, для яких використовується навчання ШНМ схожі з основними функціями біологічного мозку. Першою такою задачею є розпізнавання образів. Отримуючи дані з навколишнього світу за допомогою біологічних сенсорів, мозок досить просто розпізнає джерело даних і виділяє з нього необхідну інформацію. Таким способом людина, без вжиття особливих зусиль, дізнається знайоме обличчя, хоча бачила його давно і обличчя встигло змінитися, голос, спотворений телефонними перешкодами, місто, в якому не була багато років. Це впізнавання і є результат навчання, причому в ідеальному випадку нейронна

мережа повинна впізнавати представлені їй образи не гірше, ніж це робить живий організм.

Формально розпізнавання образів визначається як процес, в результаті якого отриманий образ (сигнал) відноситься до одного з апріорі призначених класів (категорій) [3, 4-7]. В процесі навчання нейромережі пред'являються різні образи з відомою класифікацією (навчальна вибірка), а в результаті мережа повинна розпізнати об'єкт, який як і раніше не пред'являвся, але який належить тій же сукупності, що і навчальна вибірка. Завдання розпізнавання статистична за своєю природою, при цьому образи представляються випадковими векторами в багатовимірному просторі ознак, а результат навчання полягає в побудові вирішальних гіперповерхонь, які поділяють «в середньому» простір ознак на відповідні класи.

1.3.3.2 Асоціація та кластеризація

Біологічним системам поряд з навчанням і розпізнаванням властива також здатність до асоціацій, тобто відновленню (спогаду) раніше пред'явлених образів за деякими непрямими стимулам. Будь-якій людині знайомі випадки, коли якийсь випадковий звук або запах викликав в уяві складні зорові образи.

У нейромережах асоціації реалізуються в двох формах:

– автоассоціація;

– гетероасоціація.

Під час використання автоассоціації мережу обробляє безліч пред’явлених послідовно їй образів, причому ці образи можуть бути зашумлені або перекручені. Особливість мережі, яка є схожою до властивості природної нейронної мережі – згадування. Штучна нейронна мережа набуває такої властивості відновлення (згадування) під час процесу

виділення та запам’ятовування основних ознак для запропонованих образів, які були показані їй раніше.

В свою чергу, гетероасоціація відрізняється тим, що довільна множина вхідних образів зв'язується (асоціюється) з довільною множиною вихідних прикладів. Основна відмінність між цими формами полягає в тому, що автоассоціація реалізується на основі парадигми самонавчання, а гетероасоціація – навчання з учителем.

Нехай $x(k)$ – вхідний образ-вектор (стимул), в загальному випадку довільно взятий з навчальної вибірки і пред'явлений мережі асоціативної пам'яті, а $y(k)$ – запомненний (вихідний) образ-вектор. Асоціація образів, виконувана мережею, описується відношенням $x(k) \to y(k), k = 1,2,\ldots,N$, де N – кількість образів, які запам’ятала ШНМ. Вхідний образ $x(k)$ діє як стимул, що викликає відгук $y(k)$, а в наслідку служить ключем до відновлення.

У автоасоціативній пам'яті $y(k) = x(k)$, тобто вхідний і вихідний простір мережі збігаються. У гетероасоціативної пам'яті $y(k) \neq x(k)$, при цьому розмірності просторів, як правило, також не збігаються.

В роботі асоціативних нейромереж виділяють дві фази:

– накопичення, яка відповідає періоду навчання;

– відновлення, яка передбачає спогад запам’ятованого образу після пред'явлення зашумленного або спотвореного стимулу.

Число N образів, накопичених в асоціативної пам'яті, є мірою ємності мережі. При проектуванні таких мереж основною проблемою є вибір і забезпечення максимальної місткості, вираженої як відношення числа запам'ятованих прикладів N до загальної кількості нейронів мережі при мінімальному числі некоректно відновлюваних образів.

До проблеми автоассоціації тісно примикає завдання кластеризації (автоматичної класифікації), коли мережа, аналізуючи навчальну вибірку $x(k)$, розміщує «схожі» образи по групам-кластерам. Пропонований

зашумленний образ, що раніше не показаний мережі, по асоціації з уже після успішної реєстрації повинен бути віднесений до «рідного кластеру». Мережі, які реалізують кластеризацію образів, використовуються зазвичай для стиснення даних і вилучення з них знань.

1.3.3.3 Апроксімація функції

Задача апроксимації функцій, заданих на деякій множині точок досить часто зустрічається з проблемою навчання, адже така задача застосовується на практиці.

Розглянемо нелінійне відображення «вхід-вихід», що описується функціональним співвідношенням:

$$d = f(x), \qquad (1.1)$$

де d і x - $(m\times 1)$ і $(n\times 1)$ - вектори виходів і входів відповідно,

$f(\bullet)$ – невідома вектор-функція, яку необхідно оцінити за допомогою заданої навчальної вибірки $\{x(k), d(k)\}, k = 1,2,\ldots, N$.

Завдання навчання апроксимуючої нейромережі полягає в знаходженні функції $F(x)$ в певному сенсі досить близькою до $f(x)$ такої, що:

$$\|F(x) - f(x)\| \le \varepsilon \quad \textit{для всіх } x(k),\ k = 1,2,\ldots, N, \qquad (1.2)$$

де $F(x)$ – відображення, що реалізовується мережею,

ε – мале позитивне число.

Якщо обсяг вибірки N досить великий, а мережа має достатню кількість параметрів налаштованих синаптичних ваг, помилка апроксимації

ε може бути зроблена як завгодно малою, хоча тут є небезпека перетворення мережі з апроксимуючої в інтерполюючу.

Нескладно бачити, що проблема апроксимації в даному контексті повністю збігається із завданням навчання з учителем, де послідовність $x(k)$ грає роль вхідного сигналу ШНМ, а $d(k)$ - навчає сигналу.

Здатність нейромереж апроксимувати невідомі відображення «вхід-вихід» знаходить два найважливіших додатки в задачах інтелектуального управління [8, 9, 10]. Перше з них - ідентифікація об'єктів управління [11, 12, 2], або емуляція - в термінах нейроуправління [13]. Схема системи ідентифікації (емуляції) приведена на рисунку 1.4, при цьому передбачається, що багатовимірний статичний об'єкт описується співвідношенням $d = f(x)$, а нейронна мережа, включена паралельно об'єкту, навчається в реальному часі, «підганяючи» свої вихідні сигнали до виходів реального об'єкта.

Другий додаток – це зворотне моделювання, яке використовується в деяких адаптивних системах управління [14], [15], [16], [17] і полягає в тому, що для об'єкта управління $d = f(x)$, потрібно побудувати «зворотню систему», яка генерує вектор $x(k)$ як відгук на вхідний сигнал $d(k)$. У загальному вигляді зворотна система має форму:

$$x = f^{-1}(d), \qquad (2.3)$$

Однак, оскільки функція f або не відома, або занадто складна, кращим рашенням є використання ШНМ в якості зворотної моделі так, як це показано на рисунку 1.5.

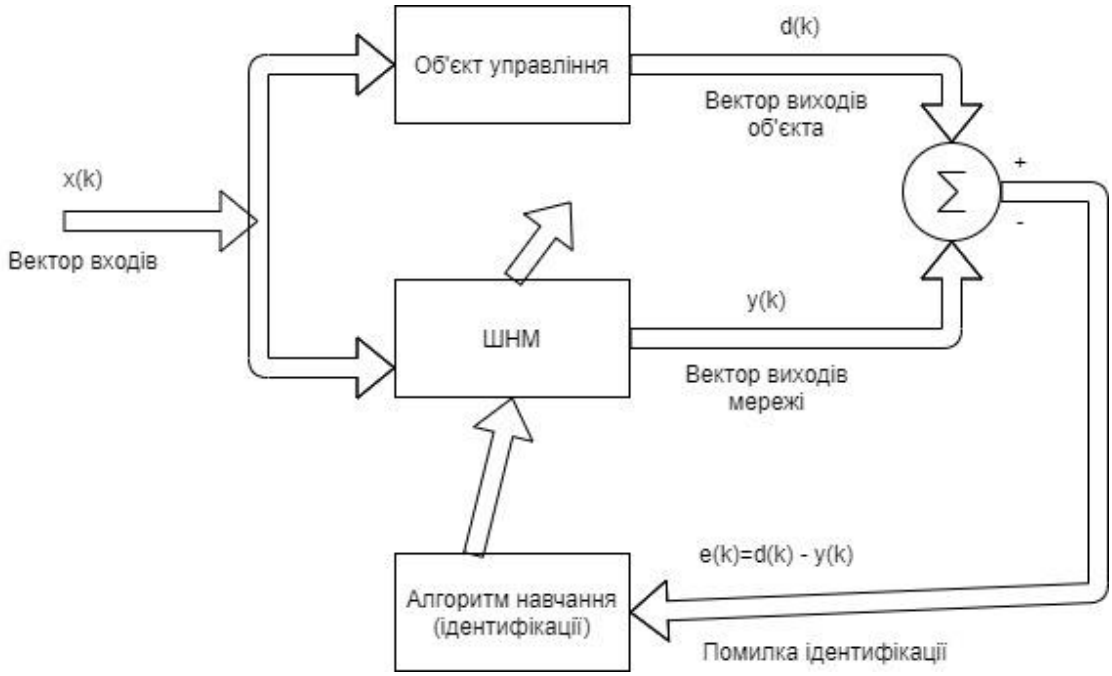

Рисунок 2.4 – Схема системи ідентифікації

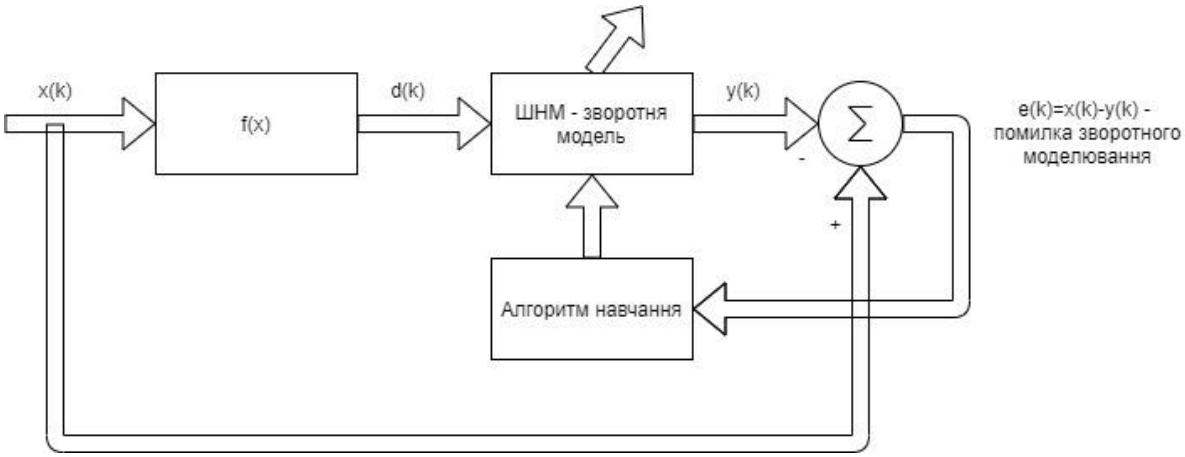

Рисунок 2.5 – Схема зворотного моделювання

У цій схемі ролі сигналів $x(k)$ і $d(k)$ помінялися: вектор $d(k)$ використовується як вхід мережі, а $x(k)$ – як бажаний відгук (навчальний сигнал). Подібно системі ідентифікації сигнал помилки $e(k) = x(k) - y(k)$ використовується для навчання ШНМ.

1.3.3.4 Управління та оптимізація

Управління об'єктами в умовах структурної і параметричної невизначеності – ще одна задача, пов'язана з навчанням нейромереж. На рисунку 1.6 приведена схема управління зі зворотним зв'язком, при цьому передбачається, що в розпорядженні проектувальника системи управління немає інформації ні про структуру нелінійного об'єкта, ні тим більше про його параметри.

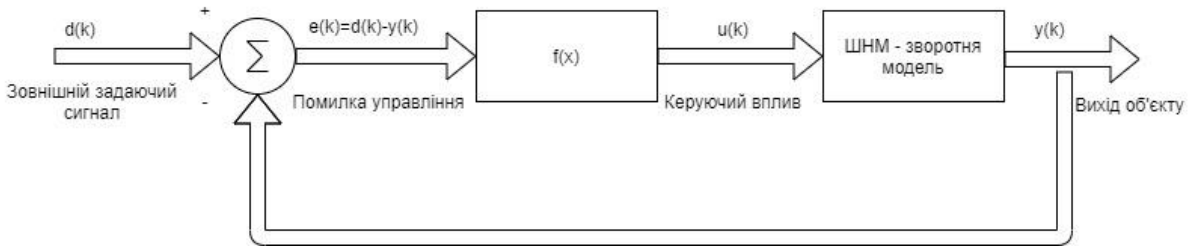

Рисунок 1.6 – Система управління зі зворотним зв’язком

Метою управління є вироблення керуючих сигналів $u(k)$, які забезпечують стале стеження виходу об'єкта $y(k)$ за зовнішнім сигналом, який задає бажаної траєкторією руху $d(k)$.

Оскільки про об'єкт управління нічого не відомо, як регулятор можна використовувати нейромережу, входом якої є вектор помилок управління $e(k) = d(k) - y(k)$, а виходом – сигнал управління $u(k)$, що подається на об'єкт.

Синтез оптимального управління пов'язаний з оцінкою якобіана об'єкта $J = \{\partial y_j / \partial u_i\}$ [18], для визначення якого знову-таки можуть бути використані апроксимуючі властивості навченої ШНМ.

В теорії адаптивного управління сформувалося два основних напрямки:

– перший – непрямий або ідентифікаційний підхід, при якому в схему вводиться налаштована модель, навчається в темпі з процесом управління і параметри якої в лінійному випадку є оцінками елементів матриці-якобіана;

– альтернативою ідентифікаційному є прямий підхід до синтезу регулятора, при якому передбачається, що проектувальнику доступна інформація про знаки приватних похідних $\partial y_j / \partial u_i$. У прямій системі управління присутня одна нейромережа-регулятор, навчається за допомогою алгоритмів, що використовують тільки знаки оброблюваних сигналів.

Поруч до задачі управління примикає завдання оптимізації, коли потрібно визначити екстремум багатовимірної неявно заданої функції при наявності обмежень. І хоча для вирішення проблеми оптимізації спроектовані спеціальні архітектури нейромереж [19], досить великий клас задач може бути вирішено в рамках систем нейроуправління [13], коли в якості цільової функції використовується лагранжіан, що враховує різноманітні обмеження, що накладаються на змінні об'єкта.

1.3.3.5 Фільтрація. Згладжування. Прогнозування

Системи обробки «зашумлених» сигналів в умовах невизначеності в даний час знаходять широке застосування в найрізноманітніших додатках [20], [21], [22], [23]. Власне поняття «обробка сигналів» традиційно включає в себе три завдання: фільтрація, згладжування і прогнозування. Якщо в розпорядженні дослідника є вибірка «забруднених» спостережень $x(1), x(2), \ldots x(k), \ldots, x(N)$, то задача фільтрації зводиться до знаходження найкращої оцінки процесу $\hat{x}(N \mid N)$ в момент часу N за інформацією N про спостереження, згладжування – оцінки $\hat{x}(k \mid N)$ при $k < N$ і прогнозування – $\hat{x}(N + l \mid N)$ при $N + l > N$, де l – горизонт попередження.

Останні роки увагу дослідників притягнуто ще до однієї нетрадиційної задачі обробки – «сліпої» сепарації та ідентифікації сигналів [24]. Передбачається, що є безліч невідомих джерел сигналів $\{u_i(k)\}_{i=1}^{n}$, які не залежать одне від одного. Сенсори сприймають ці сигнали покомпонентно, а в суміші, що представляє собою невідому лінійну комбінацію $x(k) = Au(k)$ так, як це показано на рисунку 1.7.

Завдання зводиться до відновлення вектора $y(k) \approx u(k)$ за даними спостережень вектора $x(k)$ при невідомої $(n \times n)$ – матриці A.

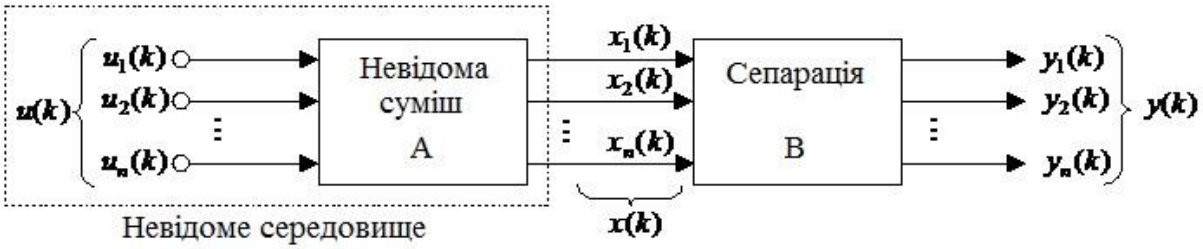

Рисунок 1.7 – Схема сліпої сепарації

Наглядно показано, що перші три завдання дуже близькі до проблеми ідентифікації, а завдання сліпої сепарації практично збігається із завданням зворотного моделювання і зводиться до знаходження оператора сепарації $B = A^{-1}$. Отже, застосування ШНМ для вирішення цих завдань принципових труднощів не викликає.

2 ГЛИБИННА РЕКУРЕНТНА НЕЙРОННА МЕРЕЖА ТА ЇЇ НАВЧАННЯ В ЗАВДАННІ РОЗПІЗНАВАННЯ МОВИ ПО ГУБАХ

Всі ми спілкуємося за допомогою мови, яка допомагає нам зрозуміти один одного, але мови бувають різні, не лише ті, які складаються зі слів. Найпершім способом, який з’явився між людьми для комунікації є саме мова жестів та звуків. Такий спосіб спілкування досить примітивний, але й має свої складності. Людині потрібен час, щоб навчитися розуміти мову жестів, а штучній нейронній мережі ще більше часу, адже штучна нейронна мережа поки що не може в повній мірі навчатися повністю подібно до людського навчання. Коли люди спілкуються один з одним, то ми на рівні підсвідомості скануємо всі зовнішні сигнали: що ми бачимо, що ми чуємо, що ми відчуваємо та інше, адже головною складністю є зовнішня схожість багатьох голосних та приголосних звуків. Не усвідомлюючи цього, при спілкуванні ми не лише слухаємо людини, але й через зір наш мозок отримує сигнали як рухаються губи людини, коли вона говорить, яка міміка та жести людини.

Якщо звернутися до тварин, то ми зможемо побачити, що вони «спілкуються» один з одним через рухи, жести, міміку та звуки, тобто використовуючи примітивні способи комунікації. Хоча люди і винайшли мову, багато різних мов, але ми все одно на підсвідомості звертаємося до наших «первісних» відчуттів, щоб проаналізувати навколишнє оточення. Ми не помічаючи робимо це кожної хвилини, кожної секунди. Також є люди, в яких обмежені певні можливості, вони або не можуть говорити, або не чують. В таких людей максимально загострюються інші сенсори для аналізу оточення.

Ввжають, що Goldschen (1997) були одними з перших, хто зміг відтворити візуальне читання по губах людей. Вони також займають місця перших, хто зміг навчити нейронну мережу розпізнавати мовлення людини на рівні речень. Для цього було використано приховані моделі Маркова

(Hidden Markov Models), які базувалися на обмеженому наборі даних, які були вручну сегментовані.

Надалі, після Голдшен Neti (2000) також спробували розвивати цю ідею, але вони пішли ще далі та стали першими, хто виконав аудіовізуальне розпізнавання мовлення людини, також на рівні речень. Для цього вони використали комбінацію HMM та функцій, яку було розроблено вручну базуючись на наборі даних IBM ViaVoice (Neti 2000). Також вони надалі намагалися покращити результати розпізнавання мовлення у більш шумних середовищах. Набір даних для такого виду навчання у шумному середовищу складався із 17111 висловлювань, які були промовлені 261 доповідачами. Це близько 34,9 годин аудіо. Цей набір даних не є у вільному доступі.

Отже, продуктивність читання з губ людини погана. Люди з вадами слуху досягають точності лише 17±12% навіть для обмеженої підмножини з 30 однокореневих слів і 21±11% для 30 складених слів (Істон і Базала, 1982). Взаємозв’язок між мовленням та рухами вважається відкритою проблемою. Тому важливою метою є автоматизація читання з губ. LipNet вирішує вказані проблеми.

2.1 Автоматичне читання по губах

Зараз вчені розроблюють нейронну мережу, яка б навчилася зчитувати мову людини по її губах. Такий вид навчання нейронної мережі є однією із усіх задач машинного зору. Така задача називається «Автоматичне читання по губах». Цей процес полягає в тому, що нейронна мережа обробляє кожен кадр відео. Найбільшою проблемою може стати саме погана якість відео, тоді нейронна мережа не зможе правильно розпізнати спатіотемпоральні характеристики обличчя людини. Вказані характеристики відображають просторово-часові характеристики обличчя у відео під час якоїсь розмови. На таких відео людина повинна щось говорити та повертати обличчя в різні сторони, щоб нейронна мережа навчалася розпізнаванню рис обличчя з

різних ракурсів. Такий метод можна порівняти із розпізнаванням обличчя в айфонах. Щоб внести свої дані про риси обличчя людина повинна повертати голову з якомога більшою амплітудою, щоб айфон зафіксував всі можливі ракурси. Це зроблено для більшої зручності користувачів, щоб в не залежності від пози людини та розташування телефона (в розумних межах), айфон зміг розпізнати ваше обличчя та розблокуватись.

Одними з перших розробок в сфері зчитування по губах досягли лише спроможності з усієї промови розпізнавати лише окремі слова, які не пов’язані між собою та які не містять сенс всього речення. Досить не так давно, розробники з Оксфордського університету змогли досягнути значного прориву у розробках в даній області. Вони змогли створити нейрону мережу, яка була навчена з метою розпізнавання мови за допомогою методу зчитування по губах. Вони назвали цю мережу LipNet. Саме нейромережа LipNet стала першим відкриттям, адже сама вона вперше змогла розпізнати цілі речення за допомогою методу зчитування по губах за допомогою передбачення послідовності слів та речень. Мережа робила це обробляючи ціле відео, також вона змогла не загубити ланцюг сенсу речення. На рисунку 2.1 та рисунку 2.2 ми зможемо побачити принцип роботи цієї нейронної мережі. На цих рисунках відображені кадри з відео, де певна людина говорить англійські слова «please» та «lay». LipNet синім кольором позначає ділянки обличчя, які найбільш змінюються і тим самим привертають увагу. Саме ці ділянки дозволяють відрізняти промовлені звуки людиною.

Рисунок 2.1 – Ділянки обличчя розпізнавання салієнтних ознак для слова «please»

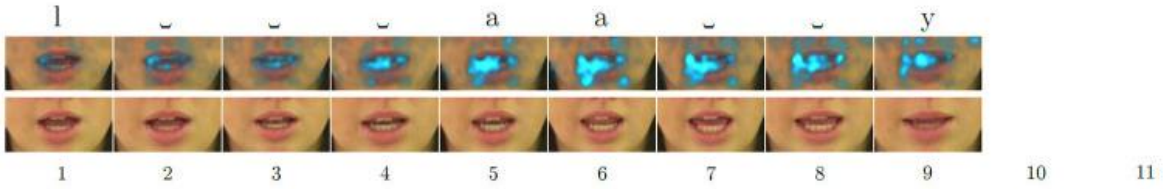

Рисунок 2.2 - Ділянки обличчя розпізнавання салієнтних ознак для слова «lay»

Звичайно, існує досить велика кількість схожих звуків, тому це є головною проблему нейронної мережі для зчитування по губах, адже міміка людини повинна бути досить активною, що не часто зустрічається в нашому повсякденному житті.

Продукція слова «please» будь ласка, на початку вимагає значної артикуляційного руху: губи міцно притиснуті один до одного для білабіального вибуху «p». При цьому кінчик язика стикається з альвеолярним відростком в очікуванні наступного латерального «l». Потім губи розходяться, дозволяючи стисненому повітрю вийти між губами. Потім щелепа та губи розкриваються далі, а губи розсуваються (збільшуючи відстань між куточками рота), для близького голосного «еа». Оскільки це відносно стійкий голосний, положення губ залишається незмінним до кінця його тривалості, коли рівень уваги значно падає. Потім щелепа і губи злегка змикаються, оскільки лезо язика потрібно наблизити до альвеолярного відростка для «se», де відновлюється увага.

Продукція слова «lay» досить цікава, адже основна частина артикуляційного руху, видимого спереду, включає кінчик язика, який контактує з альвеолярним відростком на «l», а потім опускається вниз для голосного «ау». Саме на цьому зосереджена більшість уваги LipNet, оскільки положення губ практично не змінюється.

Застосовуються методи для візуалізації салієнтних ознак для інтерпретації навченої поведінки LipNet. Досить наглядно відображено основний принцип роботи даної нейронної мережі, що вона звертає увагу саме на фонологічно важливі регіони у відео.

2.2 Архітектура нейромережі LipNet

Нейромережа LipNet розроблена на основі рекурентної нейронної мережі, яка відноситься до типу LSTM, тобто Long Short-Term Memory. На рисунку 2.3 можна побачити архітектуру рекурентної нейронної мережі LipNet. Для навчання такого виду нейронної мережі використовується метод нейромережевої класифікації CTC (Connectionist Temporal Classification). Даний метод досить поширений у застосуванні в наші часи. Його використовують у системах, які застосовуються для розпізнавання мовлення людини. Саме метод нейромережевої класифікації CTC є вигідним та зручним, адже він виключає необхідність навчати мережу на основі набору вхідних даних, який має синхронізацію з правильним вихідним результатом.

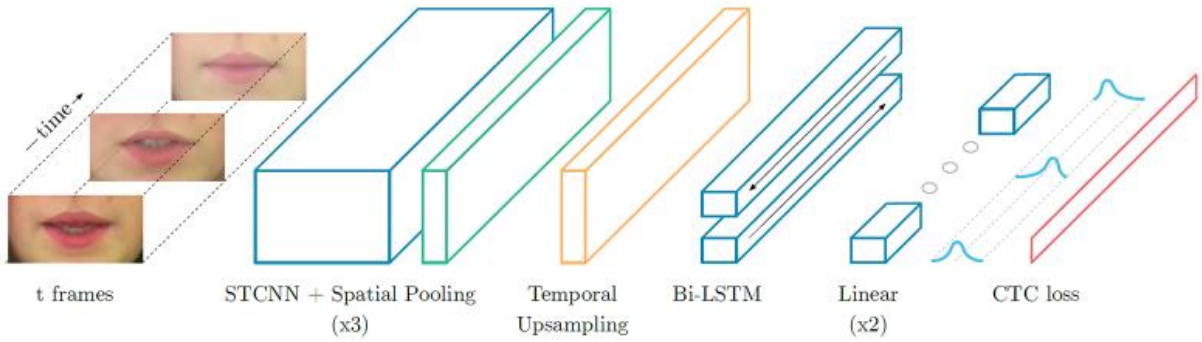

Рисунок 2.3 – Архітектура нейромережі LipNet

Принцип даної архітектури полягає в тому, що в якості вхідних даних надається деяка послідовність кадрів певного відео. Таку послідовність

будемо називати *T*. Обробка кадрів виконується за допомогою тришарової просторово-часової згорткової нейронної мережі STCNN, також її ще називають спатіотемпоральною згортковою нейронною мережею. Кожен шар такої нейромережі має свій шар просторової вибірки. Далі нейромережа отримує стапіотемпоральні ознаки, які допомагають їй розпізнавати промовлені звуки. Для отриманих ознак виконується підвищення частоти дискретизації за часовою шкалою (апсемплінг), надалі вони обробляються подвійною LTSM. Кожен часовий крок проходить через обробку двошаровою мережею прямого розповсюдження. В самому кінці отримані результати проходять обробку останнім шаром – SoftMax. Кожен часовий крок виходу GRU обробляється лінійним шаром та SoftMax. Ця наскрізна модель навчається за допомогою CTC.

Згорткові нейронні мережі CNN (Convolutional Neural Networks) складаються з упорядкованих згорток, які у просторі оброблюють зображення. Вони мають досить важливу роль для підвищення продуктивності та точності виконання завдань «комп'ютерного зору» для розпізнавання багатьох різноманітних об'єктів, зображення яких отримує нейронна мережа як вхідні дані.

Базовий шар 2D згортки від каналів C до C^0 (без зміщення та з одиничним кроком) обчислюється за формулою:

$$[\mathrm{conv}(\mathbf{x}, \mathbf{w})]_{c'ij} = \sum_{c=1}^{C} \sum_{i'=1}^{k_w} \sum_{j'=1}^{k_h} w_{c'ci'j'} x_{c,i+i',j+j'}, \qquad (2.1)$$

де *x* – вхід,

w – ваги $w \epsilon R^{C'xCxkwxkh}$, де ми визначаємо $x_{cij} = 0$ для *i, j*, які поза межами.

Просторово-часові згорткові нейронні мережі STCNNs (Spatiotemporal Convolutional Neural Networks) мають можливість обробляти відеодані шляхом згортання в часі, також як і просторових розмірів. Таким чином:

$$[\text{stconv}(\mathbf{x}, \mathbf{w})]_{c'tij} = \sum_{c=1}^{C} \sum_{t'=1}^{k_t} \sum_{i'=1}^{k_w} \sum_{j'=1}^{k_h} w_{c'ct'i'j'} x_{c,t+t',i+i',j+j'}. \qquad (2.2)$$

Рисунок 2.3 досить наглядно ілюструє принцип роботи та архітектури нейронної мережі LipNet. Ми можемо побачити, що архітектура почитається з 3х(просторово-часові згортки, випадання каналів, просторове максимальне об’єднання). Потім, отримуються ознаки двох Bi-GRUs. Вони мають вирішальне значення щодо ефективності подвльшої агрегації вихідних даних STCNN. Далі, лінійне перетворення застосовується на абсолютно кожному часовому етапі, за яким виконується softmax над словниковим запасом, який доповнений пробілом CTC, а потім втратою CTC. Абсолютно всі шари використовують функції активації випрямленої лінійної одиниці (ReLu).

2.3 Gated Recurrent Unit

Gated Recurrent Unit (GRU) – це тип рекурентної нейронної мережі RNN (Recurrent Neural Network), робота якої складається у покращені попередніх RNN за допомогою додавання осередків і шлюзів для поширення інформації на більшій кількості часових кроків та використовуючи навчання контролю цього інформаційного потоку. За принципом своєї роботи це досить схоже на довготривалу тимчасову пам’ять (LSTM) RNN:

$$\begin{aligned} [\mathbf{u}_t, \mathbf{r}_t]^T &= \text{sigm}(\mathbf{W}_z \mathbf{z}_t + \mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{b}_g) \\ \tilde{\mathbf{h}}_t &= \tanh(\mathbf{U}_z \mathbf{z}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \\ \mathbf{h}_t &= (\mathbf{1} - \mathbf{u}_t) \odot \mathbf{h}_{t-1} + \mathbf{u}_t \odot \tilde{\mathbf{h}}_t \end{aligned} \qquad (2.3)$$

де $z := \{z_1, \ldots, z_T\}$ – вхідна послідовність до RNN,

$\odot$ – позначає поелементне множення та $sigm(r) = 1/(1 + exp(-r))$.

Використовується двонаправлена GRU (Bidirectional GRU (Bi-GRU)), яка була представлена Graves & Schmidhuber (2005) у контексті LSTM: одна RNN карта $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\} \mapsto \{\overrightarrow{\mathbf{h}_1}, \ldots, \overrightarrow{\mathbf{h}_T}\}$, та інша $\{\mathbf{z}_T, \ldots, \mathbf{z}_1\} \mapsto \{\overleftarrow{\mathbf{h}_1}, \ldots, \overleftarrow{\mathbf{h}_T}\}$, тоді $\mathbf{h}_t := [\overrightarrow{\mathbf{h}_t}, \overleftarrow{\mathbf{h}_t}]$. Bi-GRU може гарантувати, що h_t залежить від $z_{t'}$ для всіх t'. Для виконання параметризації розподілу по послідовностях, на часовому кроці t нехай $p(u_t|z) = softmax(mlp(h_t; W_{mlp}))$, где mlp – це мережа прямого зв'язку з вагами W_{mlp} . Надалі ми маємо можливість визначити розподіл по послідовностях довжини-Т як $p(u_1, \ldots, u_T|z) = \Pi_{1 \le t \le T} \rho(u_t|z)$, де T визначається z, вхідним для GRU. У нейронній мережі для зчитування по губам LipNet z є вихідним сигналом STCNN.

2.4 Корпус GRID

Для підвищення ефективності навчання моделі LipNet, було застосовано стандартизований набір даних корпусу GRID – це сучасний загальнодоступний набір даних для зчитування по губах. До цього набору даних входять аудіодані з різною якістю звуку, а також відеодані з різною якістю відео, які розраховані на 34 динаміки. Але модель LipNet спеціалізується на вузькому направленні – лише на візуальному зчитуванні по губах. За допомогою особливого корпусу речень GRID точність розпізнавання мовлення візуально за допомогою зчитування по губах нейромережі досягає 93,4%. Це досить високий показник у порівнянні з іншими програмними розробками у даній сфері, але найбільш значущим фактором є те, що точність розпізнавання мовлення нейромережі LipNet перевищує навіть ефективність читання по губах спеціально навчених цьому людей. У таблиці 2.1 можна побачити результати точності зчитування по губах інших програмних розробок.

Таблиця 2.1 – Порівняння точності результатів зчитування по губах

Метод	**Набір даних**	**Розмір**	**Видача**	**Точність**
Fu et al. (2008)	AVICAR	851	Цифры	37,9%
Zhao et al. (2009)	AVLetter	78	Алфавит	43,5%
Papandreou et al. (2009)	CUAVE	1800	Цифры	83,0%
Chung & Zisserman (2016a)	OuluVS1	200	Фразы	91,4%
Chung & Zisserman (2016b)	OuluVS2	520	Фразы	94,1%
Chung & Zisserman (2016a)	BBC TV	>400000	Слова	65,4%
Wand et al. (2016)	GRID	9000	Слова	79,6%
LipNet	**GRID**	**28853**	**Предложения**	**93,4%**

Проаналізувавши результати, внесені в таблицю, можна побачити, що модель LipNet є найефективнішою серед усіх представлених інших, адже лише вона може розпізнавати цілі речення. Але все ж таки корпус GRID складається за таким шаблоном: *command(4) + color(4) + preposition(4) + letter(25) + digit(10) + adverb(4) (команда (4) + колір (4) + прийменник (4) + буква (25) + цифра (10) + прислівник (4))*, де цифра відповідає кількості варіантів слів для кожної з усіх шести представлених мовленевих категорій.

Потрібно розуміти, що отримані результати проводилися в лабораторних умовах, тому і результати несуть ідеальних характер, що означає, що не бралися в рахунок більш складні умови для розпізнавання мови. Тобто, використовувалися спеціально записані відео, де обличчя людини знято великим планом, люди говорять спеціально чітко кожен звук навмисно досить чітко працюючи мімікою [57]. А це означає, що під час намагань аналізувати довільні речення людини, або відео з поганою якістю, освітленням результат буде свідомо гірше.

2.5 Втрати CTC

Вхідні дані, які поступають до нейронної мережі LipNet мають послідовний характер, тобто відео, яке було розбито на кадри повинно зберігати певну послідовність кадрів, тому необхідно врахувати вибір функції втрат [58, 60, 62]. Враховуючи той факт, що фрази, які були промовлені у відео не мають певного часового вирівнювання, але водночас мають змінну довжину, тому доцільно буде застосовувати функцію втрат CTC (Connectionist Temporal Classification). Саме ця функція допомагає вирішити обидві проблеми, які виникли. Втрата коннекціоністської тимчасової класифікації досить поширена для використання в сцчасних задачах розпізнавання мовлення людини, адже саме вона допомагає усунути потребу в навчальних даних, які вирівнюють вхідні дані та цільові вхідні дані. CTC виконує обчислення ймовірності послідовності шляхом маргіналізації всіх послідовностей, які визначено як еквівалентні, враховуючи модель, яка виводить послідовність дискретних розподілів по класах лексем (словниковим запасом), доповнених спеціальним пустим маркером. Це одночасно усуває необхідність вирівнювання та адресує послідовності змінної довжини. Нехай V позначає набір лексем, які модель класифікує на одному часовому етапі свого виходу (словник), а порожній словник $\tilde{V} = V \cup \{\_\}$.

Отже, нейронна мережа для максимальної ефективності своєї роботи використовує корпус GRID та функцію CTC. Результати перевірки ефективності роботи даної системи розписано в наступному розділі.

3 ОЦІНЮВАННЯ LIPNET НА КОРПУСІ GRID

Під час роботи із задачами розпізнавання зображень або як в нашому випадку відео, досить часто виникає проблема через обмеженість даних. Щоб вирішити дану проблему використовується методика для створення додаткових даних із вже існуючих. Така методика називається аугментация даних (Data Augmentation).

3.1 Data Augmentation

Перед аугментацією даних спочатку виконується їх попередня обробка. Як вже було сказано віще, для підвищення ефективності роботи нейронної мережі LipNet було вирішено використати корпус GRID. Для наповнення цього корпусу даними було використано 34 спікера, кожен з яких розповів 1000 речень та це було записано на відео. Надалі виявилось, що записані відео 21-го спікера були відсутні, а декі відео виявилися порожніми або пошкодженими. Тому, залишилось 32746 відео, які можна використовувати. Серед працюючих відео було взято 3971 відео чотирьох спікерів. Ці відео розділили, що отриманні відео містять дані двох мовців чоловіків (1 та 2) і двох мовців жінок (20 та 22) для оцінки ефективності. Залишок відео було використано для навчання, а це 28775 відео.

Також не виключається можливість використання іншого методу, як варіант розділення відео на рівні пропозиції, схоже на метод Wand et al. (2016). В даному методі було використано 255 випадкових речень від кожного спікера. Та потім всі отриманні дані від усіх спікерів було використано для навчання. Кожне відео має тривалість 3 секунди та частоту кадрів 25 кадрів у секунду. Обробка відео виконується за допомогою нейронної мережі, яка виконує роль детектора обличчя DLib, також використано інструмент для прогнозування орієнтирів обличчя iBug. У поєднанні з онлайн-фільтром Калмана було обрано 68 орієнтирів.

Враховуючи обрані орієнтири застосовується афінне перетворення, щоб отримати кадрування обличчя в районі рота за розміром 100x50 пікселів на кадр. Також важливий момент це стандартизація каналів RGB для всього навчального набору для того, щоб мати нульове середнє значення та одиничну дисперсію.

На етапі аугментації даних виконується доповнення набору даних за допомогою простих трансформацій з метою зменшення переобладнання. Тренування нейронної мережі виконується спочатку на звичайній послідовності зображень, а потім на горизонтально-дзеркальній. Також використовуються додаткові навчальні приклади у вигляді відео кліпів окремих слів, адже обраний набір даних враховує час початку та закінчення слів для кожного відео речення. Такі приклади мають швидкість розпаду 0,925. Ще одним з важливих факторів, які потрібно врахувати – це змінна швидкість. Щоб вона не впливала на результат, тобто щоб він був стійким до змін швидкості, необхідно застосувати метод видалення та дублювання кадрів, який буде застосовуватися з ймовірністю на кадр 0,05. Серед усіх запропонованих базових моделей було застосовано одні й ті ж самі методи аугментації.

3.2 Baselines

Для проведення оцінки ефективності нейронної мережі LipNet, її було порівняно із показниками ефективності зчитування по губам трьох людей з вадами слуху, які вміють читати мову по губах, а також було використано три моделі абляції, які використовують технології останніх сучасних роботів (Chung & Zisserman, 2016a; Wand et al., 2016). Для проведення тестування з участю трьох людей, які мають вади слуху було обрано три добровольці серед членів Оксфордської студентської спільноти інвалідів. Добровольцям було надано для ознайомлення граматика корпусу GRID після чого їм було продемонстровано анотовані 10-ти хвилинні відео, які

містили навчальні набори даних. Після цього добровольці по черзі коментували 300 випадкових відео. В моменти, якщо вони були невпевнені, то їх просили обирати найбільш вірогідну відповідь на їхню думку. Для Baseline-LSTM навчальної установки LipNet було використано архітектуру моделі попереднього інноваційного корпусу GRID глибинного навчання. Baseline-2D на основі архітектури LipNet було замінено STCNN просторовими згортками, які були подібні до згорток Chung & Zisserman. Найцікавішим було те, що на відміну від отриманих сучасних результатів під час тестування нейронної мережі LipNet, Chung & Zisserman отримали за результатами досить низьку продуктивність їхніх STCNN на 14% та 31% у порівнянні з двовимірними архітектурами їхніх двох наборів даних. Щодо Baseline-NoLM, то він досить схожий на LipNet, можна навіть сказати, що вони ідентичні, але Baseline-NoLM має вимкнену мовленеву модель, яка використовується в променевому пошуку.

3.3 Оцінка ефективності

Для того, щоб визначити рівень ефективності та продуктивності нейронної моделі LipNet необхідно обчислити коефіцієнт помилок в словах Word Error Rate (WER), також необхідно обчислити коефіцієнт помилок в символах Character Error Rate (CER) і стандартні показники, які враховуються для оцінки продуктивності моделей ASR. Застосовуючи пошук променів CTC ми можемо лише отримати приблизні прогнози щодо максимальної ймовірності ефективності LipNet. Коефіцієнти WER або CER характеризуються мінімальною кількістю вставок, замін та видалень слів або символів. Отримані дані щодо коефіцієнтів потрібні надалі для перетворення прогнозу у реальну істину, яка має підтвердження, а ця істина ділиться на кількість слів або символів в реальній істині. Необхідно зауважити, що WER досить часто дорівнює помилці класифікації, за умови, що передбачене речення містить таку ж саму кількість слів, як і реальна

істина. В нашому випадку це має велике значення, адже майже всі отримані помилки утворені через заміни.

В таблиці 3.2 можна побачити підсумки результатів на перевірку ефективності LipNet у порівнянні з базовими показниками. Згідно з літературними даними, то точність та ефективність «чистого» зчитування по губам дорівнює 20%. Для полегшення розпізнавання контексту речення було використано фіксовану структуру речення, а також обмежену підмножину слів для кожної позиції в корпусі GRID. Як і було очікувано, це спрацювало і дійсно полегшення було. Якщо подивитися на результати при використанні невидимих динаміків, то троє людей із вадами слуху досягають точності 57,3%, 50,4%, 35,5% WER відповідно. Отже, в середньому результати 47,7% WER.

Таблиця 3.1 – Ефективність LipNet у наборі даних GRID порівняно з baselines, виміряна за двома розподілами: (а) оцінка лише на невидимих динаміках та (б) оцінка на підмножині з 255 відео речень кожного промовця.

Метод	**Невидимі динаміки**		**Відео речень промовців**	
	CER	**WER**	**CER**	**WER**
Hearing-Impaired Person (avg)	–	47,7%	–	–
Baseline-LSTM	38,4%	52,8%	15,2%	26,3%
Baseline-2D	16,2%	26,7%	4,3%	11,6%
Baseline-NoLM	6,7%	13,6%	2,0%	5,6%
LipNet	**6,4%**	**11,4%**	**1,9%**	**4,8%**

Архітектури, які було покращено за допомогою згорткових стеків дають свої результати у вигляді найвищої продуктивності під час оцінки ефективності при використанні невидимих динаміків. LipNet має в 2,3 рази вищу продуктивність у порівнянні з іншими тестованими методами. Під час використання тестування з невидимими динаміками, Baseline-2D і LipNet

досягають 1,8x та 4,2x менше WER, відповідно, ніж люди з вадами слуху. Результати отримані досить наглядно демонструють важливість поєднання STCNN з RNN. Така велика різниця в ефективності підтверджує гіпотезу, що вилучення просторово-часових об’єктів за допомогою STCNN краще, ніж агрегування лише просторових об’єктів. Отримані результати абсолютно протилежні від тих, що отримали Chung & Zisserman. Крім того, використання у LipNet STCNN, RNN і CTC дуже добре дає можливість обробляти як вхідні послідовності змінної довжини, так і вихідні послідовності змінної довжини, тоді як архітектури Chung & Zisserman обробляють лише перший вид.

4 МОДИФІКОВАНА АКТИВАЦІЙНА ФУНКЦІЯ ДЛЯ GRU ЩО НЕ ПОТЕРПАЄ ВІД ЕФЕКТУ ЗНИКАЮЧОГО ГРАДІЄНТУ

На рисунку 4.1 приведено архітектуру формального нейрону. Він є деяким узагальненням нейрона Маккаллоха-Піттса, але формальний нейрон відрізняється від нейрона Маккаллоха-Піттса тим, що у нього було введено додаткові блоки посилення γ_j та Γ_j, які можна налаштовувати.

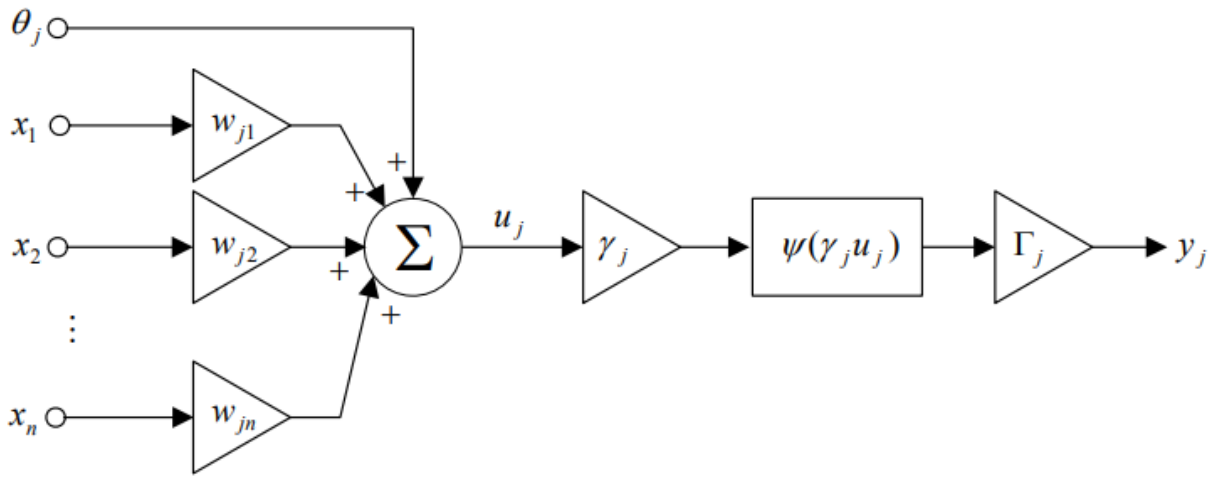

Рисунок 4.1 – Узагальнений формальний нейрон

У приведеній архітектурі формального нейрону крутизною функції активації керує параметр γ_j , а Γ_j – це коефіцієнт посилення, який допомагає визначити максимальні та мінімальні значення вихідного сигналу y_j.

Ми можемо побачити, що можна записати вихідний сигнал узагальненого формального нейрону у вказаному вигляди за формулою 4.1:

$$y_j = \Gamma_j \psi(\gamma_j w_j^T x). \tag{4.1}$$

Однією з перших сигмоїдальних функцій була представлена функція Цибенко, яка мала вигляд формули 4.2:

$$0 < \sigma(\gamma u) = \frac{1}{1+e^{-\gamma u}} < 1, \tag{4.2}$$

така конструкція була визначеною на множині всіх дійсних чисел. Її особливістю було те, що вона приймала лише позитивні значення. Однак, в ході експериментів було визначено, що для використання більш зручна крива гіперболічного тангенсу, формулу якого продемонстровано нижче у формулі 4.3:

$$-1 < \tanh(\gamma u) = \frac{1-e^{-2\gamma u}}{1+e^{-2\gamma u}} < 1, \tag{4.3}$$

виявилося, що така крива гіперболічного тангенсу пов'язана із сигмоїдою певним відношенням, яке відображено у формулі 4.4:

$$\sigma(\gamma u) = \frac{1}{2}\left(\tanh\left(\frac{\gamma u}{2}\right)+1\right). \tag{4.4}$$

Надалі необхідно визначити обмеження функції. Обмеженнями на квадраті визначено $-1 \le u_j \le 1$ та $-1 < y_j < 1$. Потім визначаємо всі можливі функції, які можуть виступати в якості активаційних функцій нейрону:

$$\psi_1(\gamma u) = \tanh(\gamma u) = \frac{1-e^{-2\gamma u}}{1+e^{-2\gamma u}}, \ \Gamma_1 < \frac{1}{\tanh\gamma}, \tag{4.6}$$

$$\psi_2(\gamma u) = \frac{\gamma u}{\sqrt{1+\gamma^2 u^2}}, \ \Gamma_2 < \frac{\sqrt{1+\gamma^2}}{\gamma}, \tag{4.7}$$

$$\psi_3(\gamma u) = \sin(\frac{\pi}{2}\gamma u), \ \Gamma_3 < \frac{1}{\sin\left(\frac{\pi}{2}\gamma\right)}, \tag{4.8}$$

$$\psi_4(\gamma u) = \frac{2}{\pi} arctg(\gamma u),\ \ \Gamma_4 < \frac{\pi}{2arctg\gamma}, \tag{4.9}$$

$$\psi_5(\gamma u) = \gamma u - \frac{\gamma^3}{3} u^3,\ \ \Gamma_5 < \frac{3}{3\gamma - \gamma^3}. \tag{4.10}$$

Розглянемо залежність приведених функцій. Графік цих залежностей продемонстровано на рисунку 4.2.

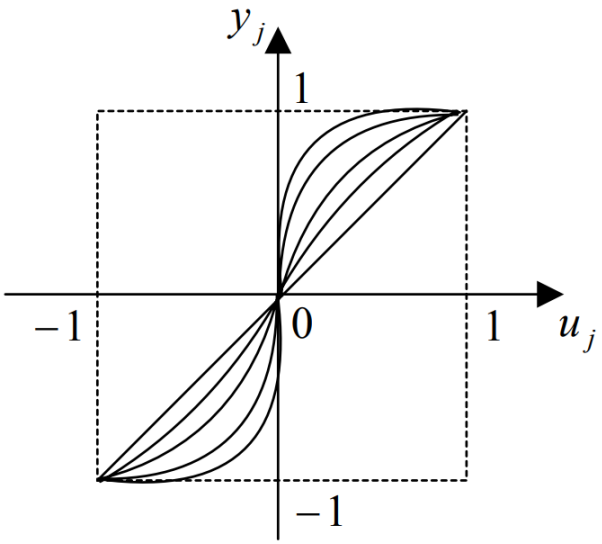

Рисунок 4.2 – Активаційні функції узагальненого нейрону

Проаналізувавши приведений графік ми можемо побачити, що всі задані функції відповідають переліченим вимогам до них. Через те, що під час процесу навчання нейронних мереж ми працюємо безпосередньо як із самими активаційними функціями, а також і з їхніми похідними, то представлення функції буде більш доцільним у вигляді $\psi_j(\bullet)$. Такий вигляд функції буде спрощувати підрахунки похідних, а це в свою чергу буде спрощувати сам процес навчання та налаштування нейронних мереж.

Для отримання такого зручного вигляду функції можна використати один з представлених підходів у вигляді степінного ряду. Отже, ми

розкладаємо наведені функції (4.6-4.10) за допомогою розкладу у ряд Маклорена. Звідси, ці функції мають вигляд:

$$\psi_1(\gamma u) = \gamma u - \frac{1}{3}(\gamma u)^3 + \frac{2}{15}(\gamma u)^5 - \frac{17}{315}(\gamma u)^7 + ..., \tag{4.11}$$

$$\psi_2(\gamma u) = \gamma u - \frac{1}{2}(\gamma u)^3 + \frac{3}{8}(\gamma u)^5 - \frac{39}{336}(\gamma u)^7 + ..., \tag{4.12}$$

$$\psi_3(\gamma u) = \frac{\pi}{2}\left(\gamma u - \frac{\pi^2}{24}(\gamma u)^3 + \frac{\pi^4}{1920}(\gamma u)^5 - \frac{\pi^6}{322560}(\gamma u)^7 + ...\right), \tag{4.13}$$

$$\psi_4(\gamma u) = \frac{2}{\pi}\left(\gamma u - \frac{1}{3}(\gamma u)^3 + \frac{1}{5}(\gamma u)^5 - \frac{1}{7}(\gamma u)^7 + ...\right), \tag{4.14}$$

$$\psi_5(\gamma u) = \gamma u - \frac{1}{3}(\gamma u)^3, \tag{4.15}$$

необхідно звернути увагу, що необхідну точність апроксимації забезпечують перші чотири члени розкладання у розмірі 10^{-2}.

Дослідивши зміни коефіцієнтів рядів, ми зможемо вирази загального виду (4.11-4.14) представити у вигляді ряду:

$$\psi^*(\gamma u) = \varphi_0 \gamma u + \varphi_1(\gamma u)^3 + \varphi_2(\gamma u)^5 + \varphi_3(\gamma u)^7 + ... = \sum_{l=0}^{L} \varphi_l (\gamma u)^{2l+1}, \tag{4.16}$$

Коефіцієнти даного виразу задовольняють рівнянню авторегрессії першого порядку:

$$\varphi_{l+1} = \alpha \varphi_l + \xi_{l+1}, \; -1 < \alpha < 0, \; \varphi_0 = 1, \tag{4.17}$$

У даному рівнянні a та дисперсія випадкової компоненти ξ можна оцінити за допомогою використання відомих співвідношень методу найменших квадратів, який можна побачити у формулі 4.18:

$$\alpha = \frac{\sum_{l=0}^{L-1} \varphi_{l+1}\varphi_l}{\sum_{l=0}^{L-1} \varphi_l^2}, \qquad \sigma_\xi^2 = \frac{\sum_{l=0}^{L-1} (\varphi_{l+1} - \alpha\varphi_l)^2}{L-1}. \tag{4.18}$$

Виявилося, за результатами підрахунків із застосуванням чотирьох членів розкладання, що параметр a , який забезпечує потрібні властивості активаційної функції для всіх розглянутих кривих належить інтервалу $-0.5 < a < -0.3$, а $\sigma_\xi^2 \le 10^{-2}$.

На рисунку 4.3 можна побачити графік представлення активаційної функції, яка представлена за допомогою функції активації (4.16), коли коефіцієнт $a = -0.3$, який позначено цільною лінією, а коефіцієнт $a = -0.5$ позначено пунктирною лінією, а також коефіцієнтом посилення:

$$\Gamma^* = \left(\sum_{l=0}^{L-1} \varphi_l \right)^{-1}. \tag{4.19}$$

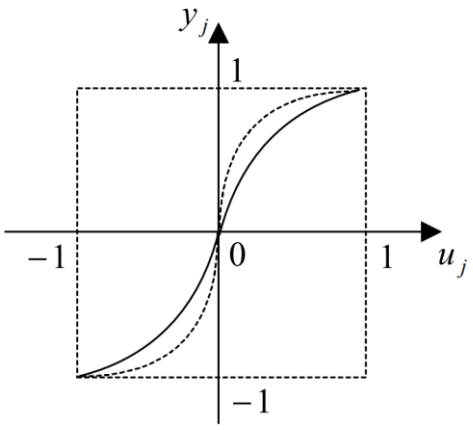

Рисунок 4.4 – Активаційна функція у вигляді степінного ряду

Проаналізувавши графік стає очевидним, що дана функція містить всі вказані властивості сигмоїдальної функції, але при цьому диференціювання поліному у розрахунковому сенсі простіше, ніж диференціювання приведених функцій (4.6-4.10), тому доцільно ввести у розгляд модифіковану сигмоїдальну функцію на основі кубічного поліному, тобто сплайн апроксимація. Нелінійне перетворення елементарного опису персептрону має вигляд:

$$y_j = \psi(\gamma_j \omega_j^T x) = \psi(\gamma_j u_j). \quad (4.20)$$

Задана функція має обмеження у вигляді:

$$0 < \sigma(\gamma_j u_j) = \frac{1}{1+e^{-\gamma_j u_j}} < 1. \quad (4.21)$$

Похідна даної функції має наступний вигляд:

$$\sigma'(\gamma_j u_j) = \gamma_j y_j (1 - y_j) \quad (4.22)$$

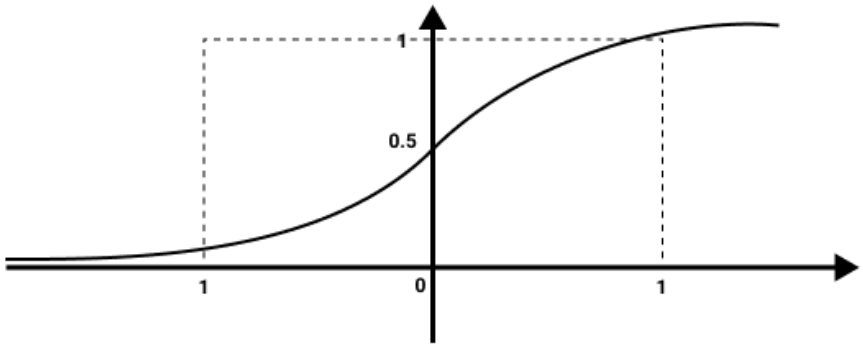

Рисунок 4.5 – Сигмоїдальна функція

За даним графіком ми бачимо, що в точках 1 та 0 похідна прямує до 0, а це означає, що функція потерпає від ефекту зникаючого градієнту, а отже,

навчання нейронної мережі зупиняється. Для того, щоб уникнути потерпання від ефекту зникаючого градієнту, ми вводимо кубічний поліном. Вирішимо наступне рівняння:

$$y_j = a_0 + a_1 u_j + a_2 u_j^2 + a_3 u_j^3 \tag{4.23}$$

Введення даного кубічного поліному дозволяє нам варіювати межі функції не лише від -1 до 1, а й від -2 до 2 і т.д., щоб уникати ефект зникаючого градієнту.

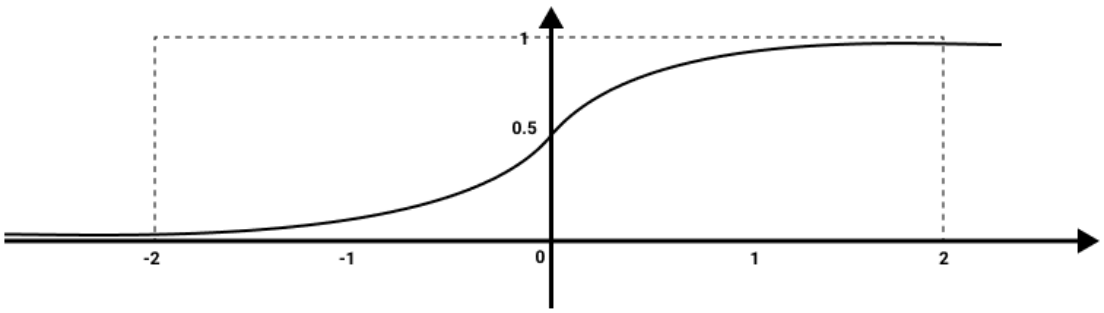

Рисунок 4.6 – Сигмоїдальна функція на діапазоні [-2;2]

За допомогою такого методу ми можемо побудувати функцію на будь-який інтервал та уникнути ефект зникаючого градієнту. Чим простіша похідна, тим легше проходить навчання.

Вирішимо отримане рівняння кубічного поліному:

$$\begin{gathered} y_j = a_0 + a_1 u^* + a_2 u^{*2} + a_3 u^{*3} = 1 \\ y_j = a_0 - a_1 u^* + a_2 u^{*2} - a_3 u^{*3} = 0 \\ y_j' = a_1 + 2a_2 u^* + 3a_3 u^{*2} = 0 \\ y_j' = a_1 - 2a_2 u^* + 3a_3 u^{*2} = 0, \end{gathered} \tag{4.24}$$

нехай $u^* = +- 1$:

$$\begin{gathered} y_j = a_0 + a_1 + a_2 + a_3 = 1 \\ y_j = a_0 - a_1 + a_2 - a_3 = 0 \\ y_j' = a_1 + 2a_2 + 3a_3 = 0 \\ y_j' = a_1 - 2a_2 + 3a_3 = 0, \end{gathered} \tag{4.25}$$

нехай $a_0 = 0{,}5$:

$$\begin{gathered} a_1 + a_2 + a_3 = 0{,}5 \\ -a_1 + a_2 - a_3 = -0{,}5 \\ a_1 + 2a_2 + 3a_3 = 0 \\ a_1 - 2a_2 + 3a_3 = 0, \end{gathered} \tag{4.26}$$

якщо $a_2 = 0$:

$$\begin{gathered} a_1 + a_3 = 0{,}5 \\ -a_1 - a_3 = -0{,}5, \end{gathered} \tag{4.27}$$

$$\begin{gathered} a_1 + a_3 = 0{,}5 \\ a_1 + 3a_3 = 0, \end{gathered} \tag{4.28}$$

$$\begin{gathered} a_1 + 3a_3 = 0 \\ a_1 + 3a_3 = 0, \end{gathered} \tag{4.29}$$

$$\begin{gathered} a_1 = -3a_3 \\ -3a_3 + a_3 = 0{,}5 \\ -2a_3 = 0{,}5 \\ a_3 = -0{,}25 \\ a_1 = 0{,}75. \end{gathered} \tag{4.30}$$

Такі результати ми отримали для відрізку u^* = [-1;+1]. Розширимо обраний відрізок до [-2;+2], звідси:

$$y_j^{'} = 0{,}75 - 0{,}75u_j^2, \tag{4.31}$$

при u^* = 2, то:

$$\begin{gathered} a_0 + 2a_1 + 4a_2 + 8a_3 = 1 \\ a_0 - 2a_1 + 4a_2 - 8a_3 = 0 \\ a_1 + 4a_2 + 12a_3 = 0 \\ a_1 - 4a_2 + 12a_3 = 0, \end{gathered} \tag{4.32}$$

Нехай $a_2 = 0$ та $a_1 = 0{,}5$:

$$\begin{gathered} 2a_1 + 8a_3 = 0{,}5 \\ -2a_1 - 8a_3 = 0{,}5 \\ a_1 + 12a_3 = 0, \end{gathered} \tag{4.33}$$

$$\begin{gathered} 2a_1 + 8a_3 = 0{,}5 \\ a_1 + 12a_3 = 0, \end{gathered} \tag{4.34}$$

$$\begin{gathered} a_1 + 4a_3 = 0{,}25 \\ a_1 + 12a_3 = 0, \end{gathered} \tag{4.35}$$

нехай $a_1 = -12a_3$:

$$\begin{gathered} -12a_3 + 4a_3 = 0{,}25 \\ -8a_3 = 0{,}25 \\ -32a_3 = 1 \end{gathered} \tag{4.36}$$

$$a_3 = -1/32,$$

$$a_1 = 12/32. \quad (4.37)$$

Звідси отримуємо:

$$y_j' = 12/32 - 3/32u_j^2 \quad (4.38)$$

Отже, ми досягли своєї мети і сигмоїда може мати свої параметри, які змінні для діапазону функції. Також, ми змогли уникнути ефект зникаючого градієнту.

ВИСНОВКИ

В ході написання даної кваліфікаційної роботи магістра, було запропоновано новий метод онлайн навчання штучної нейронної мережі LipNet з можливістю уникнути виникнення ефекту зникаючого градієнту. Такий результат було досягнено за допомогою використання кубічного поліному для сигмоїдальної функції. Завдяки такому підходу, ми можемо обрати будь-який інтервал для нашої функції, що надає нам перевагу у боротьбі з виникненням ефекту зникаючого градієнту. Також було досліджено принцип методу зчитування по губам реальних людей та штучної нейронної мережі. Результати були порівняні, та як висновок можна зробити, що нейронна мережа LipNet лише починає наближатися до рівня ефективності зчитування мовлення по губам реальної людини.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Haykin S. Neural Networks. A Comprehensive Foundation. Upper Saddle River, N.J.: Prentice Hall, Inc., 1999. 842 p.

2. Изерман Р. Цифровые системы управления. М.: Мир, 1984. 541 с.

3. Bishop C. M. Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1995. 482 p.

4. Фу К. Последовательные методы в распознавании образов и обучении машин. М.: Наука, 1971. 256 с.

5. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука,1979. 368 с.

6. Патрик Э. А. Основы теории распознавания образов. М.: Сов. радио, 1980. 408 с.

7. Растригин Л. А., Эренштейн Р. Х. Метод коллективного распознавания. М.: Энергоиздат, 1981. 80 с.

8. Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches / ed. by D. A. White, D. A. Sofge. N.Y.: Van Nostrand Reinhold, 1992. 568 p.

9. Harris C. J., Moore C. G., Brown M. Intelligent Control. Aspects of Fuzzy Logic and Neural Nets. Singapore: World Scientific. 1993. 380 p.

10. Advances in Intelligent Control / ed. by C. J. Harris. London: Taylor and Francis, 1994. 373 p.

11. Цыпкин Я. З. Основы информационной теории идентификации. М.: Наука, 1984. 320 с.

12. Льюнг Л. Идентификация систем. Теория для пользователя. М.:Наука,1991. 432 с.

13. Omatu S., Khalid M., Yusof R., Neuro-Control and its Applications. London: Springer-Verlag, 1995. 255 p.

14. Фомин В. Н., Фрадков А. Л., Якубович В. А. Адаптивное управление динамическими объектами. М.: Наука, 1981. 448 с.

15. Деревицкий Д. П., Фрадков А. Л. Прикладная теория дискретных адаптивных систем управления. М.: Наука, 1981. 216 с.

16. Real-Time Computer Control / ed. by S. Bennet, D. A. Linkens. London: Peter Peregrinus, 1984. 254 p.

17. Industrial Digital Control Systems / ed. by K. Warwick, D. Rees. London: Peter Peregrinus, 1986. 439 p.

18. Первозванский А. А. Курс теории автоматического управления. М.: Наука, 1986. 616 с.

19. Cichocki A., Unbehauen R. Neural Networks for Optimization and Signal Processing. Stuttgart: Teubner, 1993. 526 p.

20. Адаптивные фильтры / ред. К. Ф. Н. Коуэна, П. М. Гранта. М.: Мир, 1988. 329 с.

21. Бабак В. П., Хандецький В. С., Шрюфер Е. Обробка сигналів. К.: Либідь, 1999. 496 с.

22. Haykin S. Adaptive Filter Theory. Upper Saddle River, N.J.: Prentice Hall, 1996. 987 p.

23. Уидроу Б., Стирнз С. Адаптивная обработка сигналов. М.: Радио и связь, 1989. 440 с.

24. Amari S., Cardoso J.-F. Blind source separation semiparametric statistical approach. *IEEE Trans. on Signal Processing*. 1997. № 45. P. 2692–2700.

25. Moody J., Darken C. J. Fast learning in networks of locally-tuned processing units. *Neural Computation*. 1989. № 1. P. 281–294.

26. Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. М.: Наука. 1970. 384 с.

27. Parzen, E. On the estimation of a probability density function and the mode. *Ann. Math. Statist.* 1962. № 38. P. 1065–1076.

28. Надарая Э.А. О непараметрических оценках плотности вероятности и регрессии. *Теория вероятностей и ее применение*. 1965. № 1. С. 199–203.

29. Варядченко Т. В., Катковник В. Я. Непараметрический метод обращения функций регрессии. *Стохастические системы управления*. Новосибирск: Наука. 1979. С. 4–14.

30. Haykin S. Neural Networks. A Comprehensive Foundation. Upper Saddle River, N.J.: Prentice Hall, Inc. 1999. 842 p;

31. Живоглядов В.П., Медведев А.В. Непараметрические алгоритмы адаптаци. Фрунзе: Илим. 1974. 214 с.

32. Раудис Ш. Ю. Оптимизация непараметрического алгоритма классификации. *Адаптивные системы и их приложения*. Новосибирск: Наука. 1978. С. 57–61.

33. Медведев А. В. Непараметрические алгоритмы идентификации нелинейных динамических систем. *Стохастические системы управления*. Новосибирск: Наука. 1979. С. 15–22.

34. Hartman E. J., Keeler J. D., Kowalski J. Layered neural networks with Gaussian hidden units as universal approximations. *Neural Computation*. 1990. № 2. P. 210–215.

35. Park J., Sandberg I. W. Universal approximation using radial-basis-function networks. *Neural Computation*. 1991. № 3. P. 246–257.

36. Leonard J.A., Kramer M. A., Ungar L. H. Using radial basis functions to approximate a function and its error bounds. *IEEE Trans. on Neural Networks*. 1992. № 3. P. 614–627.

37. Sunil E.V.T., Yung C. Sh. Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems. *IEEE Trans. on Neural Networks*. 1994. № 5. P. 594–603.

38. Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности. *Теория вероятностей и ее применение*. 1968. № 14. С. 156–161.

39. Розенблатт Ф. Модель памяти на нейронных сетях. *Автоматика*. 1965. № 5. 1966. № 2, 4.

40. Поляк, Б.Т. Введение в оптимизацию. М: Мир. 1984. 541 с.

41. Nelles O. Nonlinear System Identification. Berlin: Springer. 2001. 785 p.

42. Specht D.E. A general regression neural network. *IEEE Trans. on Neural Networks*. 1991. № 2. P.568–576.

43. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М.: Наука. 1974. 416 с.

44. Вапник, В.Н. Восстановление зависимостей по эмпирическим данням. М.: Наука. 1979. 448 с.

45. Least Squares Support Vector Machines / J.A.K. Suykens et. al. Singapore: World Scientific. 2002. 294 p.

46. Bodyanskiy Y. V., Tyshchenko A. K., Deineko A. A. An evolving radial basis neural network with adaptive learning of its parameters and architecture. *Automatic Control and Computer Sciences*. 2015. № 49 (5). P. 255–260.

47. Kohonen T. Self-Organizing Maps. Berlin: Springer-Verlag. 1995. 362 p.

48. Angelov P., Kasabov N. Evolving computational intelligence systems. Proc. 1st Int. Workshop on Genetic Fuzzy Systems. Granada, Spain, 2005. P. 76–82.

49. Huang G.-B., Siew C.-K. Extreme Learning Machine with Randomly Assigned RBF Kernels. *International Journal of Information Technology*. 2005. Vol. 11. No. 1. P. 16–24.

50. Bifet A. Adaptive Stream Mining: Pattern Leaning and Mining from Evolving Data Streams. IOS Press. 2010. 224 p.

51. Park J., Sandberg I. W. Universal approximation using radial-basisfunction networks. *Neural Computation*. 1991. Vol. 3. P. 246–257.

52. Huang G.-B., Zhu Q.-Y., Siew C.-K. Extreme Learning Machine: Theory and applications. *Neurocomputing*. 2006. Vol. 70. P. 489–501.

53. Гласс Л., Мэкки М. От часов к хаосу. Ритмы жизни. М.: Мир. 1991. 107 с.

54. Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности. Теория вероятностей и ее применение. 1968. № 14. С. 156–161.

55. Айзерман М. А., Браверман Э. М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970. 384 с.

56. Parzen E. On the estimation of a probability density function and the mode. *Ann. Math. Statist*. 1962. 38. P. 1065–1076.

57. https://habr.com/ru/post/398901/

58. Almajai, S. Cox, R. Harvey, and Y. Lan. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2722–2726, 2016.

59. D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595, 2015. P. Ashby. Understanding phonetics. Routledge, 2013.

60. J. S. Chung and A. Zisserman. Lip reading in the wild. In Asian Conference on Computer Vision, 2016a. J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016b.

61. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

62. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.

63. A. Cruttenden. Gimson’s pronunciation of English. Routledge, 2014.

64. G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for largevocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1): 30–42, 2012.

65. F. DeLand. The story of lip-reading, its genesis and development. 1931.

66. R. D. Easton and M. Basala. Perceptual dominance during lipreading. Perception & Psychophysics, 32(6): 562–570, 1982.

67. E. Ferragne and F. Pellegrino. Formant frequencies of vowels in 13 accents of the british isles. Journal of the International Phonetic Association, 40(01):1–34, 2010.

68. C. G. Fisher. Confusions among visually perceived consonants. Journal of Speech, Language, and Hearing Research, 11(4):796–804, 1968.

69. Y. Fu, S. Yan, and T. S. Huang. Classification and feature extraction by simplexization. IEEE Transactions on Information Forensics and Security, 3(1):91–100, 2008.

70. A. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

71. S. Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa. Dynamic stream weighting for turbodecoding-based audiovisual ASR. In Interspeech, pp. 2135–2139, 2016. 9

72. A. J. Goldschen, O. N. Garcia, and E. D. Petajan. Continuous automatic speech recognition by lipreading. In Motion-Based recognition, pp. 321–343. Springer, 1997.

73. A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning, pp. 1764–1772, 2014.

74. A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5):602–610, 2005.

75. A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling ´unsegmented sequence data with recurrent neural networks. In ICML, pp. 369–376, 2006.

76. M. Gurban and J.-P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. IEEE Transactions on Signal Processing, 57(12):4765–4776, 2009.

77. K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In IEEE International Conference on Computer Vision, pp. 1026–1034, 2015.

78. S. Hilder, R. Harvey, and B.-J. Theobald. Comparison of human and machine-based lip-reading. In AVSP, pp. 86–89, 2009.

79. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97, 2012.

80. S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

81. D. Hu, X. Li, et al. Temporal multimodal learning in audiovisual speech recognition. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 3574–3582, 2016.

82. S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1):221–231, 2013.

83. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732, 2014.

84. D. E. King. Dlib-ml: A machine learning toolkit. JMLR, 10(Jul):1755–1758, 2009.

85. D. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.