Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи визначення
емоційного стану спікера

(тема)

Виконав:

студент II курсу, групи СПм-22-4

Уваров Г.О.
(прізвище, ініціали)

Спеціальність

123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

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

Освітня програма

Системне програмування
(повна назва освітньої програми)

Керівник: доц. Барковська О.Ю.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

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

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Уварову Георгію Олексійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи визначення емоційного стану спікера

затверджена наказом по університету від " 01 " квітня 2024 р. № 257Ст

2. Термін подання студентом роботи до екзаменаційної комісії 15 червня 2024 р.

3. Вхідні дані до роботи

Вхідний датасет IEMOCAP для розпізнавання емоцій у розмові, розмір якого становить 24 GB, кількість файлів - 302

Бібліотека машинного навчання TensorFlow

Обчислювач на базі центрального процесора

4. Перелік питань, що потрібно опрацювати у роботі

1. Аналіз існуючих систем класифікації емоцій спікера
2. Аналіз нейромережевих моделей для визначення емоцій спікера
3. Вибір модальностей
4. Створення моделі системи класифікації емоцій
5. Тестування запропонованої системи
6. Аналіз отриманих результатів

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 13

6. Консультанти розділів роботи (заповнюється 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	Термін виконання етапів роботи	Примітка
1	Огляд методів екстракції та обробки ключових	02.04.24-08.04.24	
	ознак в мультимедійних даних		
2	Аналіз нейромережевих класифікаторів	09.04.24-16.04.24	
3	Створення функціональної моделі системи	17.04.24-22.04.24	
4	Розробка методології проведення досліджень	23.04.24-06.05.24	
5	Проведення експериментів	07.05.24-23.05.24	
6	Оформлення матеріалів кваліфікаційної роботи	24.05.24-03.06.24	
7	Подання кваліфікаційної роботи керівникові та	04.06.24-07.06.24	
	її попередній захист		
8	Подання кваліфікаційної роботи на рецензу-	08.06.24-12.06.24	
	вання		

Дата видачі завдання 01 квітня 2024 р.

Студент _____
(підпис)

Керівник роботи _____ доц.Барковська О.Ю.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 75 с., 32 рис., 7 табл., 1 дод., 23 джерел.

НЕЙРОННА МЕРЕЖА, РОЗПІЗНАВАННЯ ЕМОЦІЙ, ШІ, SER, ERC.

Метою кваліфікаційної роботи є дослідження методів визначення емоційного стану спікера, що може бути застосоване під час онлайн-зустрічей із психологом, при здачі екзаменів онлайн, при проходженні співбесід тощо.

У ході виконання кваліфікаційної роботи запропонована система визначення емоційного стану спікера, яка є інноваційним програмно-апаратним комплексом, аналізує аудіо- та відеодані для ідентифікації емоційних реакцій учасників онлайн-зустрічей, іспитів або співбесід. Система використовує алгоритми машинного навчання, розпізнавання мови та голосу для визначення емоційного стану в реальному часі. Розглянуті під час проведення порівняльного аналізу роботи дають точність визначення емоційного стану (злість, щастя, нейтраль, сум) до 82,7% на основі трьох різних модальностей – відео, такст та аудіо. Запропонована в роботі система забезпечує точність детектування емоції до 82,9%, спираючись на аудіодані, відеодані та текст, завдяки використанню методів машинного навчання, а саме нейромережевого аналізатора Bi-LSTM.

ABSTRACT

Master’s thesis: 75 pages, 32 figures, 7 tables, 1 appendices, 23 sources.

NEURAL NETWORK, EMOTION RECOGNITION, AI, SER, ERC.

The major goal of this thesis is to study methods for determining the emotional state of a speaker, which can be applied during online meetings with a psychologist, online exams, interviews, and so on.

In order to do so, a system for determining the emotional state of the speaker was proposed. An innovative software and hardware complex that analyzes audio and video data to identify the emotional reactions of participants in online meetings, exams or interviews. The system uses machine learning algorithms, speech, and voice recognition to determine the emotional state in real-time. The comparative analysis conducted during the work shows that the system can determine the emotional state (anger, happiness, neutral, sadness) with an accuracy of up to 82.7% based on three different modalities – video, text, and audio. The proposed system ensures emotion detection accuracy of up to 82.9%, relying on audio data, video data, and text, through the use of machine learning methods, specifically the Bi-LSTM neural network analyzer.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ANN – штучна нейронна мережа (англ, Artificial Neural Network)

CNN – згорткова нейронна мережа (англ, Convolutional neural network)

DANN – Domain adversarial neural networks

DNN – глибока нейронна мережа (англ, Deep neural network)

ERC – розпізнавання емоцій в розмові (англ, Emotion recognition in conversation)

FFT – швидке перетворення Фур’є (англ, Fast Fourier Transform)

GRU – вентильний рекурентний вузол (англ, Gated Recurrent Unit)

LLM – велика мовна модель (англ, Large language models)

LSTM – довга короткочасна пам'ять (англ, Long-Short term Memory)

MAE – маскований автокодувальник (англ, Masked autoencoder)

MFCC – мел-частотні кепстральні коефіцієнти (англ, Mel-frequency cepstral coefficients)

RNN – рекурентна нейронна мережа (англ, Recurrent neural network)

SER – розпізнавання емоцій у мовленні (англ, Speech emotion recognition)

ВСТУП

Розпізнавання проявів емоцій за допомогою штучного інтелекту є актуальним для різноманітних застосувань, включаючи взаємодію людини з комп’ютером, діагностику психічного здоров’я та емоційну аналітику в маркетингу та рекламі. Ця технологія допомагає покращити користувацький досвід у технологічних інтерфейсах, роблячи їх більш інтуїтивними та адаптованими до потреб користувачів.

У сфері психічного здоров’я розпізнавання емоцій можна використовувати для оцінки та моніторингу емоційних станів, що може допомогти в діагностиці та лікуванні розладів психічного здоров’я. Це також може бути корисним для виявлення ранніх ознак психічних розладів, таких як депресія або тривожність, що дозволяє вчасно втрутитися та надати необхідну допомогу.

У маркетингу та рекламі розпізнавання емоцій можна використовувати для розуміння реакцій та настроїв споживачів, що дозволяє підвищити ефективність рекламних кампаній. Аналіз емоційних реакцій на рекламу може допомогти маркетологам створювати більш привабливі та цільові повідомлення, що резонують з аудиторією. Це може бути використано для поліпшення взаємодії з клієнтами, надаючи їм більш персоналізовані послуги та продукти, які відповідають їхнім емоційним потребам.

Таким чином, розпізнавання проявів емоцій за голосовим сигналом має важливе значення для емоційного аналізу, досліджень психологічних станів та настроїв, маркетингу, розвитку технологій та покращення охорони здоров’я. Емоції можуть виражатися не лише за допомогою голосу, але й через обличчя, мовлення, жести та інші канали. Люди виражають свої емоції по-різному, існує велика різноманітність виразів обличчя, тону голосу, мовленнєвих відтінків тощо. Ця варіація може ускладнювати автоматичне розпізнавання емоцій, оскільки моделі повинні бути здатні адаптуватися до

широкого спектру емоційних виразів.

Автоматичне розпізнавання емоцій може знайти застосування у багатьох інших сферах, таких як освіта, де воно може допомогти викладачам краще розуміти емоційний стан учнів і адаптувати свої методи викладання, або в автомобільній промисловості, де воно може підвищити безпеку водіїв, виявляючи ознаки втоми або стресу.

Крім того, розпізнавання емоцій може відігравати важливу роль у соціальній роботі та допомозі людям з особливими потребами, допомагаючи їм краще спілкуватися з оточуючими. Технології, що здатні автоматично ідентифікувати та реагувати на емоції, можуть значно покращити якість життя таких людей, надаючи їм інструменти для більш ефективної соціальної взаємодії.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Обґрунтування актуальності обраної теми

Емоції відіграють важливу роль у прийнятті нами життєвих рішень. Розуміння їх викликає зацікавленість через їх потенційні застосування, оскільки знання про те, як почуваються інші, дозволяє нам взаємодіяти та передавати інформацію ефективніше. За допомогою розпізнавача емоцій інші системи можуть виявляти втрату довіри або зміни в емоціях, спостерігаючи за поведінкою людей. Ця здатність допоможе конкретним системам, таким як чат-боти, реагувати на ці події та адаптувати свої рішення для поліпшення розмов за допомогою налаштування свого тону або виразу обличчя для створення кращого соціо-афективного користувацького досвіду.

Ще одним важливим застосуванням розпізнавання виразів обличчя є безпека в автомобільному транспорті. Виявлення стресу, гніву або втоми може бути вирішальним у запобіганні дорожнім аваріям на розумних автомобілях, дозволяючи автомобілям приймати рішення на основі психологічного стану водія.

В області охорони здоров’я застосуванням цих систем є взаємодія людини з машинами у досвіді підтримуваного проживання для літніх людей. Розпізнавач емоцій може відстежувати емоційний стан людини для виявлення аномалій у її поведінці. Коли виникає аномалія, це може означати, що людині потрібна увага. Крім того, розпізнавач емоцій може бути практичним у діагностиці певних захворювань (депресивні розлади, паркінсонізм і т.д.) за допомогою виявлення дефіцитів у вираженні певних емоцій, що прискорює діагностику, а також лікування пацієнта.

Розпізнавачі емоцій також будуть необхідні для майбутнього створення соціальних роботів. Ці роботи повинні знати, як розпізнавати емоції людей та передавати та виражати свій власний емоційний стан, щоб відображати

ближчі особисті відносини з людьми.

У розпізнавання емоцій є багато практичних застосувань, деякі з них дозволяють значно поліпшити досвід людини при взаємодії з машиною або службами. Наприклад, система може визначати якість роботи агентів кол-центру за допомогою розпізнавання емоційного відклику клієнту, наприклад радість чи злість. Ця інформація може допомогти кол-центру підвищити якість обслуговування [1].

Із збільшенням доступності високопродуктивних обчислювальних ресурсів використання інформаційних технологій у системах візуального розпізнавання зображень і надалі залишатиметься важливою сферою досліджень і розробок. Доступність великих обсягів даних дозволяє навчати складніші моделі, а високопродуктивні обчислювальні ресурси дозволяють запускати ці моделі за розумний проміжок часу [2].

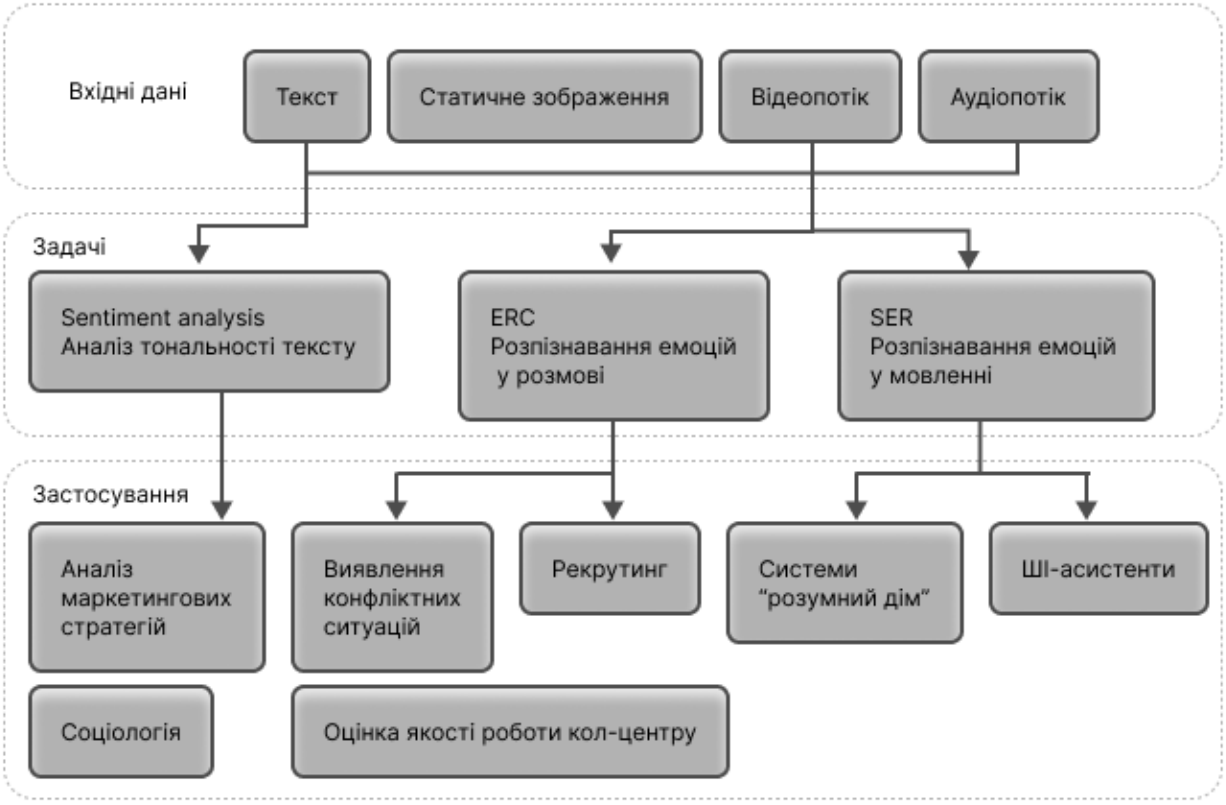

Рисунок 1.1 – Сучасні застосування розпізнавання емоцій

У психології існує безліч систем систематизації основних емоцій та їх відтінків. Від класичних 7 емоцій відомих ще давньогрецьким філософам (рисунок 1.2), до багатовимірних моделей з урахуванням хімічних процесів мозоку (Куб емоцій Левгейма, рисунок 1.3).

Рисунок 1.2 – Класифікації емоцій

Підходи до класифікації емоцій можна розділити на дискретний та безперервний підходи.

У безперервному підході (як куб Левгейма) емоції розглядаються як континуум, де кожна емоція може мати безліч відтінків та ступенів. Замість того, щоб обмежитися кількома основними емоціями, цей підхід дозволяє враховувати більш широкий спектр емоційних станів.

Така модель, як колесо Плутчика (рисунок 1.4) виділяє 8 основних дискретних емоцій та розподіляє відтінки за силою впливу (безперервний параметр).

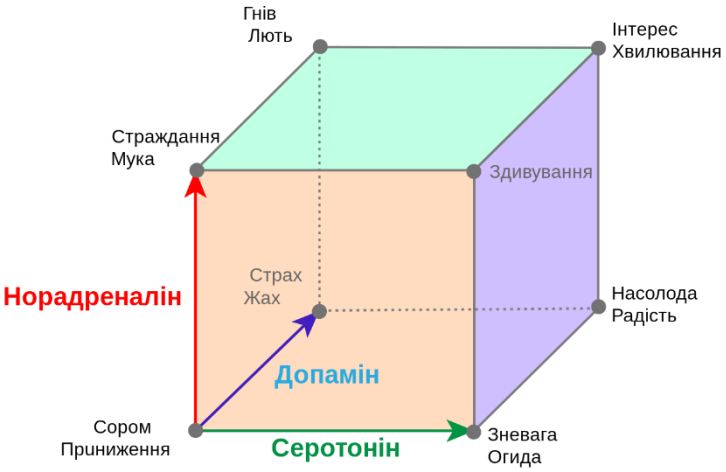

Рисунок 1.3 – Куб Левгейма

У дискретному підході емоції класифікуються як окремі категорії або класи. Зазвичай використовується деяка обмежена кількість основних емоцій, наприклад, радість, сурми, відвідування, смуток тощо. Такий підхід спрощує аналіз емоцій і використовується в багатьох областях, включаючи психологію, медицину, соціологію та інші.

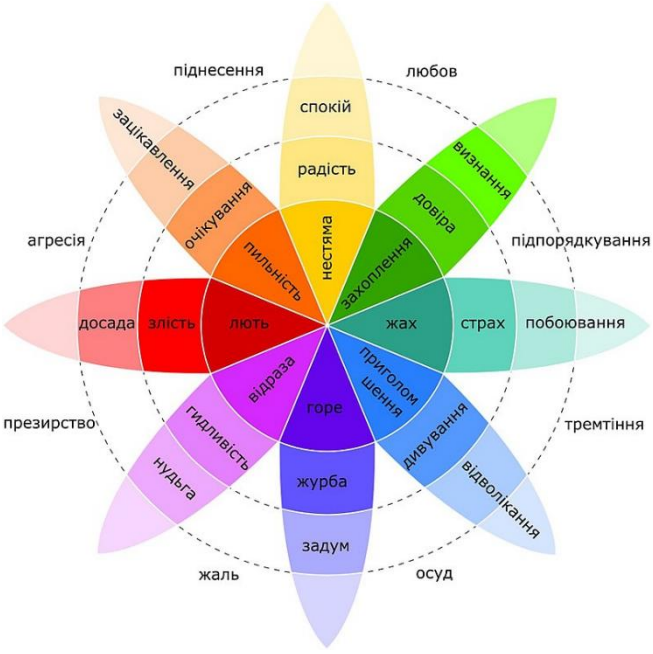

Рисунок 1.4 – Колесо Плутчика

1.2 Огляд проблемної області

Автоматичне розпізнавання емоцій – це задача розпізнавання та категоризації емоційного стану спікера за допомогою комп’ютерних алгоритмів.

SER – використовує аудіопотік як вхідні дані але не семантику висловлівань, тому це задача у рамках обчислювальної паралінгвістики, що постає у розпізнаванні та категоризації емоційного стану спікера, такого як радість, злість, сум чи фрустрація, за такими характеристиками мовлення, як ритм, тон і просодія

Розпізнавання емоцій в тексті полягає в ідентифікації емоційних станів або відчуттів, виражених у тексті. Це може включати емоції, такі як радість, смуток, гнів, страх, подив, тощо. Методи розпізнавання емоцій можуть використовувати правила, статистичні методи або великі лінгвістичні моделі (LLM) для аналізу тексту та визначення його емоційного вмісту.

Пов’язана задача аналізу тональності тексту оцінює загальний настрій або емоційний відтінок текстового документа. Це може бути позитивний, негативний або нейтральний.

Розпізнавання емоцій у тексті реалізується:

- підходами, заснованими на словнику, що знаходять у словнику початкові слова думки чи емоції та шукають їх синоніми й антоніми, щоб розширити початковий список думок чи емоцій;
- підходами, заснованими на корпусі, починаються з початкового списку слів думок або емоцій і розширюють базу даних шляхом пошуку інших слів із контекстно-специфічними характеристиками у великому корпусі;
- підходами машинного навчання, що включають опорні векторні машини (SVM), наївні байєсівські та алгоритм максимальні ентропії;
- підходами глибокого навчання.

Розпізнавання емоцій у розмові (ERC, Emotion Recognition in

Conversation) включає обробку аудіо-відеоряду діалогів. Для цього використовуються системи машинного зору. [3]

В області автоматичного розпізнавання емоцій є складнощі, пов’язані з різницею у вираженні емоцій між культурами та індивідами, шумними даними та суб’єктивною природою емоцій. Багатокультурні великі датасети і різноманіття методів екстракції ознак.

1.3 Аналіз існуючих рішень в області визначення емоційного стану спікера за голосом

Історично, системи розпізнавання емоцій у мовленні (SER) найсучаснішого рівня мали низьку точність та великі витрати обробки. В даний час деякі моделі можуть працювати в реальному часі та демонструють високу продуктивність в таких умовах, як це було продемонстровано у роботі Anvarjon та ін.[4]. Вони запропонували легку модель CNN з простими прямокутними ядрами та модифікованими шарами пулінгу, досягнувши передових результатів на датасетах IEMOCAP та EMO-DB. Для розпізнавання емоцій з аудіо використовують велику кількість підходів. Зокрема, у сфері машинного навчання використовуються згорткові нейронні мережі працюючі із зображенням спектрограм [5], комбінації згорткових та рекурентних мереж.

SER зосереджено на екстракції з потоку мови емоційної забарвленості не враховуючи зміст слів. Типова SER-система є набором методологій, що ізолюють, виділяють та класифікують сигнали мовлення для визначення емоцій, які було вкладено спікером [1]. SER є задачею класифікації часових серій даних. Класами зазвичай виступають 6 основних емоцій: радість, здивування, сум, гнів, страх і огиду. На вхід подається голосовий аудіосигнал або ознаки, що були виділені з аудіосигналу на етапи препроцесингу, такі як рівень енергії та мел-частотні кепстральні коефіцієнти (Mel-frequency cepstral coefficients, MFCCs).

Загальна idef-0 схема алгоритму Speech Emotion Recognition представлена на рисунку 1.5.

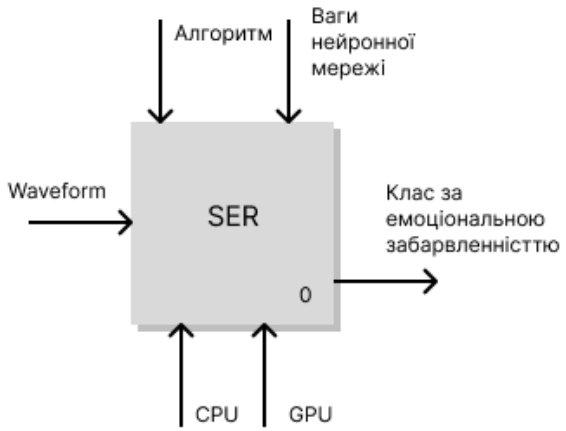

Рисунок 1.5 – SER-система як чорна скринька

Автоматичне визнання емоцій у мові включає кілька етапів для ефективного розпізнавання емоцій в аудіозаписах. Основні етапи можуть включати (рисунок 1.6):

- препроцесинг аудіоданих;
- виділення ознак;
- класифікацію.

Передобробка включає в себе очищення аудіозапису від шуму і нормалізацію гучності, нарізка довгих кліпів на короткі фрейми з якими може працювати класифікатор (рисунок 1.7).

Передобробка аудіосигналу перед використанням його у задачах машинного навчання має кілька важливих цілей та переваг:

Зменшення розміру даних: аудіосигнали можуть бути дуже об'ємними, особливо при високій частоті дискретизації. Передобробка може включати зменшення розміру сигналу шляхом відкидання непотрібної інформації або зменшення частоти дискретизації без втрати суттєвої інформації.

Передобробка може включати фільтрацію для видалення цього шуму та поліпшення якості сигналу.

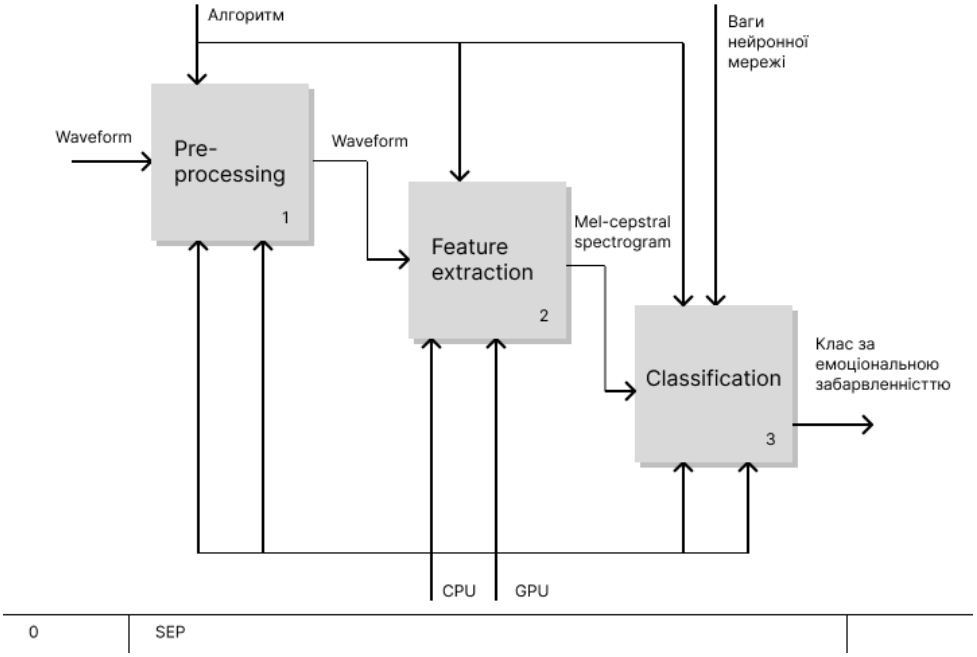

Рисунок 1.6 – Стадії виконання SER

Перед навчанням моделі може бути важливо нормалізувати аудіосигнал для забезпечення сталої динамічного діапазону та однорідності вхідних даних. Нормалізація може допомогти покращити стабільність та швидкість навчання моделі.

Деякі моделі машинного навчання можуть вимагати специфічного формату або типу даних, наприклад, двовимірного масиву з числовими значеннями. Передобробка аудіосигналу може включати перетворення сигналу у такий формат, який підходить для використання в конкретній моделі.

Глибокі нейронні мережі довели свою більш високу точність у визнанні емоцій у мовленні (SER) порівняно з традиційними методами машинного навчання [1]. Однак SER залишається складною задачею класифікації, де потрібно виділити схожі емоційні шаблони; для цього потрібна високодискримінаційна репрезентація ознак.

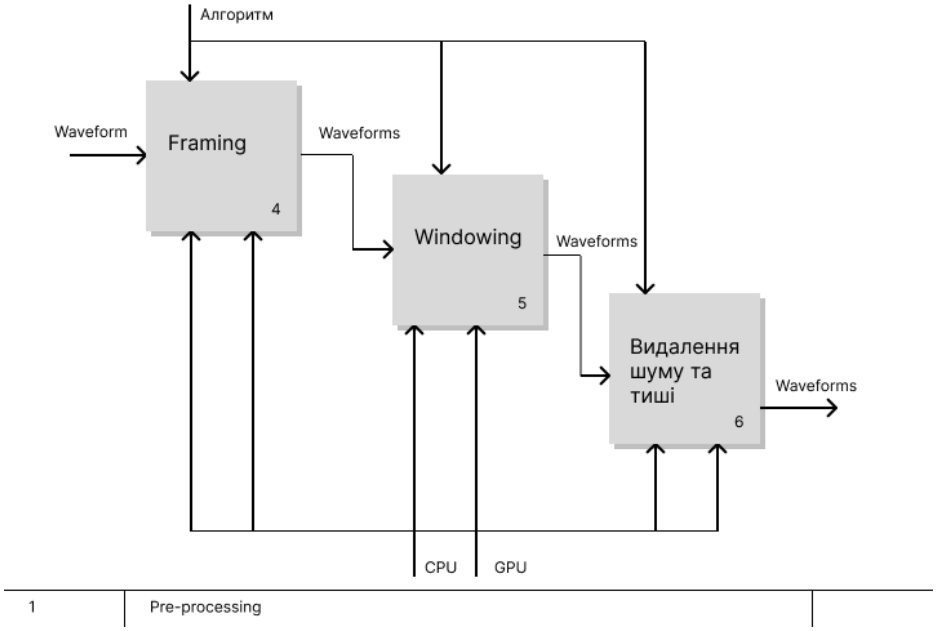

Рисунок 1.7 – Деталізація стадії препроцесингу

У сучасних моделях використовується побудова спектрограм за допомогою перетворення Фур'є та виділення мел-кепстральних коефіцієнтів (MFCC). Така обробка аудіосигналу даватиме двовимірну спектрограму (рисунок 1.8).

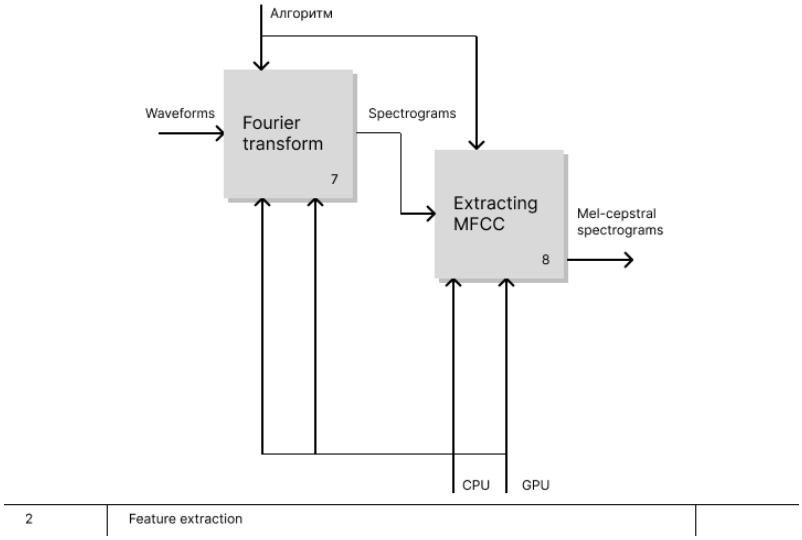

Рисунок 1.8 – Деталізація стадії екстракції ознак

У кінцевому етапі навчається класифікатор машинного навчання на

основі головного вектора ознак для проведення класифікації емоцій. Використовуються різні алгоритми класифікації, такі як k-найближчих сусідів, прихована модель Маркова, опорний векторний класифікатор, штучна нейронна мережа та гаусова змішана модель.

Алгоритми машинного навчання мають бути спочатку треновані на розміченому датасеті (аудіокліпи висловлювань з позначеними емоціями), далі протестовані на тестових розмічених даних і тоді можуть бути впроваджені для роботи у прикладні застосунки. Для машинного навчання вкрай важливо, щоб тренувальні дані були якісними: датасет повинен бути репрезентативним для проблеми, яку ви намагаєтеся вирішити. Це означає, що він повинен включати різноманітність даних, які зустрічаються в реальному світі, і відображати різні сценарії, з якими може стикнутися модель. Важливо, щоб кількість прикладів кожного класу була приблизно однаковою. Нерівномірний розподіл класів може призвести до перекосу моделі.

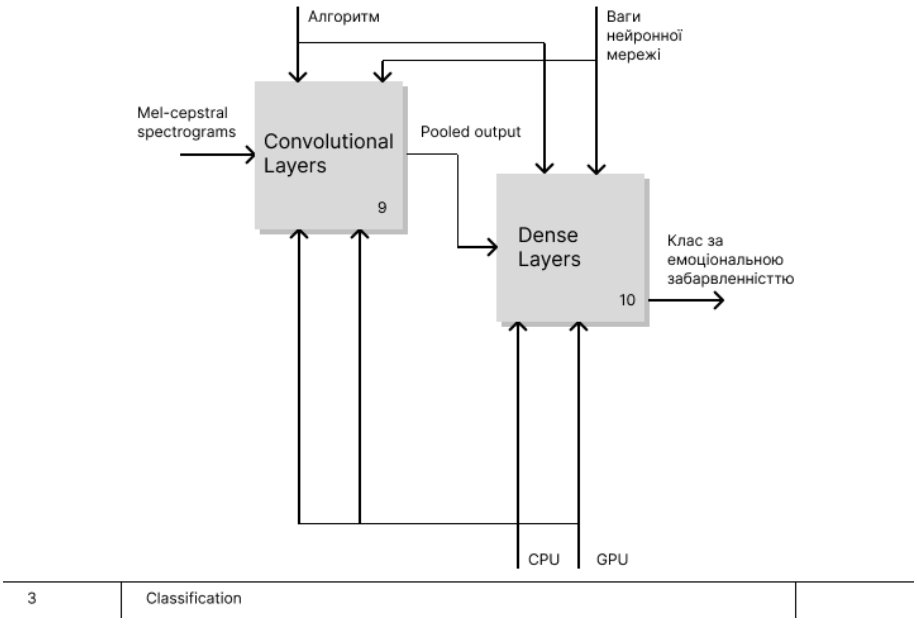

Рисунок 1.9 – Деталізація стадії класифікації

Сучасні класифікатори для розпізнавання емоцій у мовленні використовують методи трансформерів та механізми уваги (attention).

Трансформери – тип архітектури нейронних мереж, який здатний до опрацювання послідовностей даних, таких як текст або зображення.

Трансформери базуються на механізмах уваги, які дозволяють їм ефективно взаємодіяти з різними частинами вхідних послідовностей. Основною ідеєю є те, що кожен елемент вхідної послідовності (токен) може взаємодіяти з усіма іншими елементами, враховуючи їхню важливість. Це дозволяє трансформерам ефективно моделювати довгі залежності та вирішувати складні завдання обробки послідовностей.

Таблиця 1.1 – Порівняння сучасних моделей SER

Модель	Тестовий датасет	Вид	Точність, %	Зважена F1
InstructERC [6]	IEMOCAP	мультимодальна	71.68	71.39
CFN-ESA [7]	IEMOCAP	мультимодальна	70.78	71.04
UniMSE [8]	IEMOCAP	мультимодальна	70.56	70.66
GA2MIF [9]	IEMOCAP	мультимодальна	69.75	70.00
CoordViT [10]	CREMA-D	голос	82.96	-
SepTr+LeRaC [11]	CREMA-D	голос	70.95	-
SepTr [12]	CREMA-D	голос	70.47	-
DANN [13]	IEMOCAP	голос	82.70	-
VQ-MAE-S-12 [14]	RAVDESS	голос	84.10	84.40
CNN-X [5]	RAVDESS	голос	82.99	-
xlsr-Wav2Vec2.0 [15]	RAVDESS	голос	81.82	-
VAVL [16]	CREMA-D	мультимодальна	0.826	-

Це дозволяє моделям ефективно вирішувати завдання розпізнавання емоцій шляхом аналізу контексту та відносин між різними частинами

мовлення. Механізм уваги дозволяє моделі визначати, на які аспекти вхідних даних слід звернути увагу під час прийняття рішення про класифікацію емоцій. Такі підходи дозволяють покращити якість розпізнавання емоцій та зробити моделі більш адаптивними до різноманітних контекстів мовлення.

1.3.1 Порівняння існуючих рішень

В основі CFN-ESA [7] лежить кодер на основі рекурентності для одномодального кодування (recurrence based Uni-Modality Encoding), кодер на основі уваги для крос-модального кодування (attention based Cross-Modality Encoding), класифікатор емоцій (Classifier) та модуля оптимізації зсуву емоцій на основі міток (Emotion-Shift Optimizing) (рисунок 1.10).

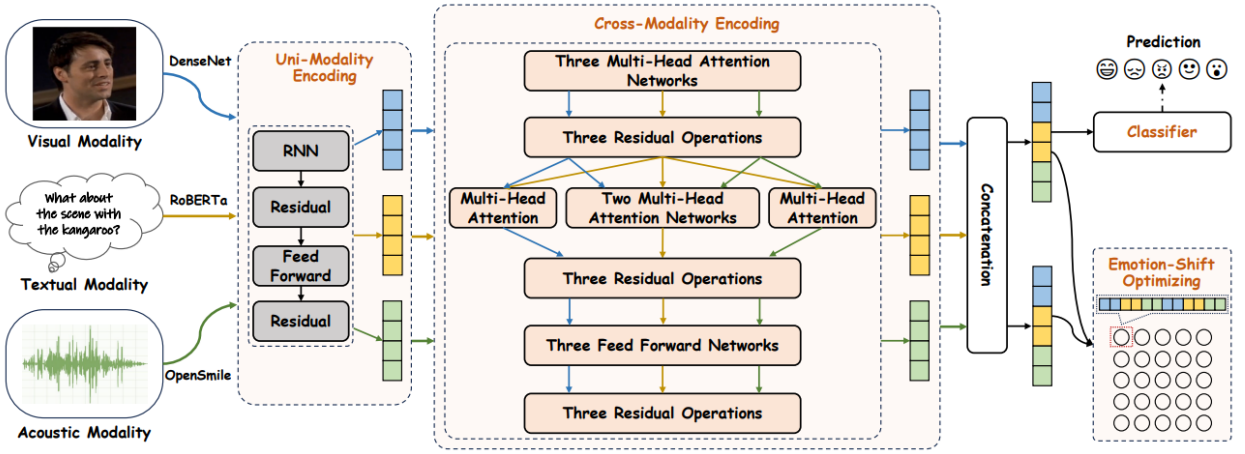

Рисунок 1.10 – Модель CFN-ESA

UniMSE переформулює задачі MSA (Multimodal Sentiment analysis) та ERC як завдання генерації для об'єднання вводу, виводу та завдання. Для цього видобувалася та об'єднувалася аудіо- та відеоознаки та формалізуються мітки MSA та ERC у Загальні Мітки (Universal Labels) для об'єднання тональності та емоцій (рисунок 1.11).

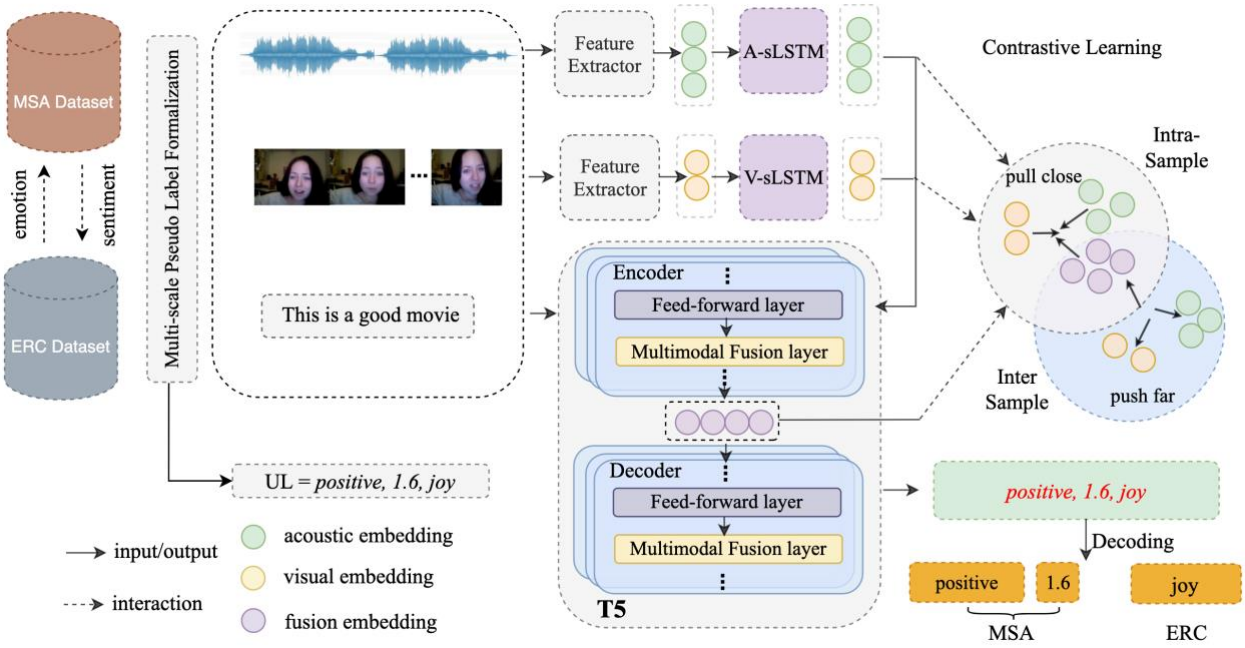

Рисунок 1.11 – Модель UniMSE

У [9] багатомодальний метод під назвою Graph and Attention based Two-stage Multi-source Information Fusion (GA2MIF) використано для виявлення емоцій у розмові. Запропонований підхід обходить проблему прийняття гетерогенного графа на вхід до моделі, водночас усуваючи складні зайві з'єднання при побудові графа. GA2MIF акцентується на контекстному моделюванні та крос-модальному моделюванні за допомогою механізму багатоголової спрямованої уваги на графі (MDGATs) та мереж парної крос-модальної уваги (MPCATs) відповідно (рисунок 1.12). Обширні експерименти на двох загальнодоступних наборах даних (тобто IEMOCAP та MELD) показують, що запропонований GA2MIF має здатність ефективно захоплювати внутрішньо-модальну інформацію про контекст на великій відстані та міжмодальну доповнюючу інформацію, а також перевершує провідні моделі останніх років з вражаючим рівнем.

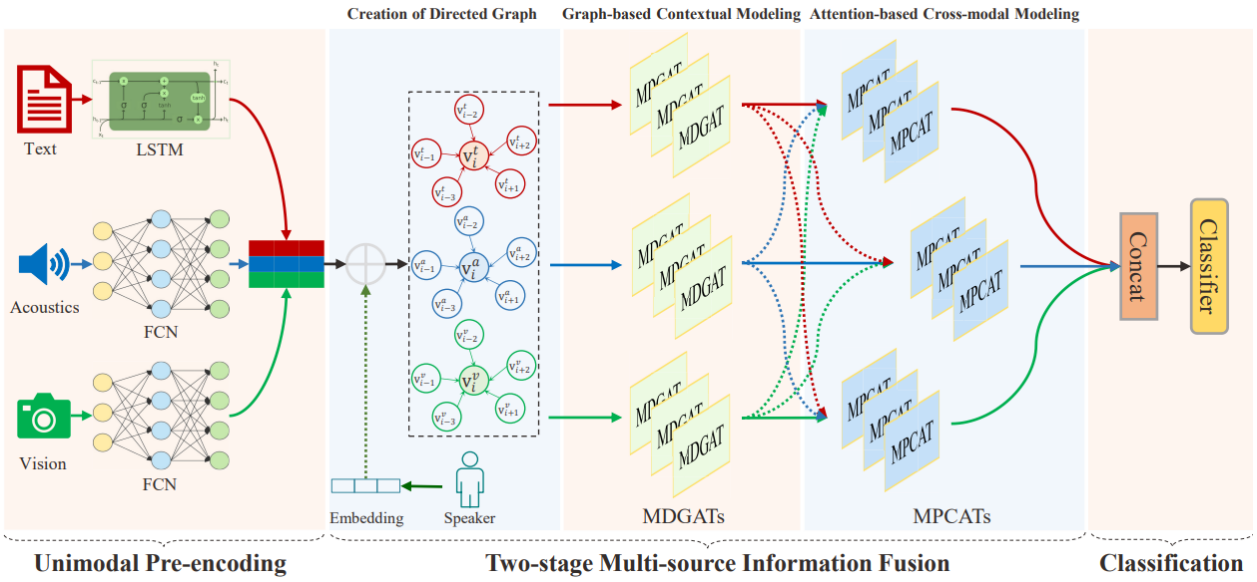

Рисунок 1.12 – Модель GA2MIF

Архітектура Separable Transformer (SepTr) використовує два блоки трансформера послідовно, перший працює на токенах всередині часових інтервалів, а другий – з токенами в одній смузі частот[12].

На відміну від стандартних трансформерів, SepTr лінійно масштабує кількість навчальних параметрів з розміром вводу, тому має менший обсяг пам'яті (рисунок 1.14).

Архітектура передбачає що вхідна спектрограма токенізується (рисунок 1.13) і далі обробляється двома трансформерами (вертикальний і горизонтальний).

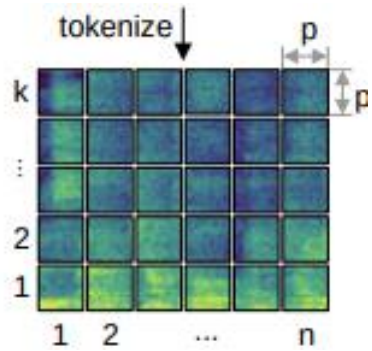

Рисунок 1.13 – Препроцесінг (токенізація спектрограми) для моделі SepTr

Відокремлюваний блок трансформерів (поєднуючий і вертикальний, і горизонтальний) повторюється L разів (глибина моделі). Фінальний шар токенів йде на вхід багатошарового перцептрону, який обирає вихідний клас.

У [11] SepTr навчали за підходом "Learning Rate Curriculum" (LeRaC), що використовує різні швидкості навчання для кожного шару нейронної мережі, щоб створити куррикулум без використання даних під час початкових епох навчання. LeRaC призначає більші швидкості навчання нейронним шарам, які знаходяться ближче до входу, поступово зменшуючи швидкості навчання, коли шари віддаляються від входу.

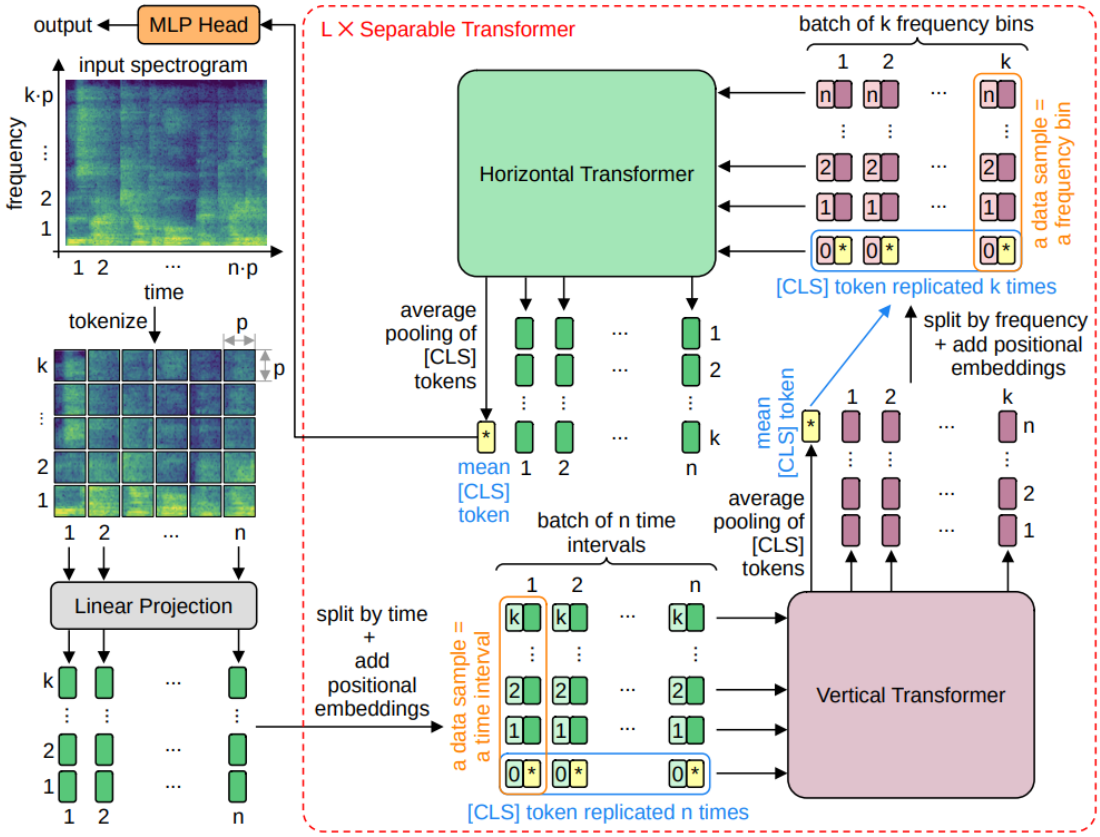

Рисунок 1.14 – Модель SepTr

Швидкості навчання збільшуються з різною швидкістю під час перших ітерацій навчання, поки вони не досягнуть однакового значення. З цього моменту нейронна модель навчається як зазвичай. Це створює стратегію навчання за навчальним планом на рівні моделі, яка не потребує сортування

прикладів за складністю і сумісна з будь-якою нейронною мережею, що генерує вищі рівні продуктивності незалежно від архітектури.

У [13] застосовано контекстозалежну доменну адверсаріальну нейронну мережу (DANN) для багатомодального розпізнавання емоцій для розпізнавання емоцій. Основне завдання полягає в передбаченні емоційних міток. Вторинне завдання – навчитися загальному представленню, де ідентифікатори мовця не можуть бути розрізнені. Цей підхід наближає представлення різних мовців. Тим часом, за допомогою непозначених даних у процесі навчання, зменшено вплив обмежених навчальних зразків. Попередні роботи виявили, що контекстуальна інформація та багатомодальні ознаки є важливими для розпізнавання емоцій. Однак попередні підходи, основані на DANN, ігнорують цю інформацію, обмежуючи їхню ефективність.

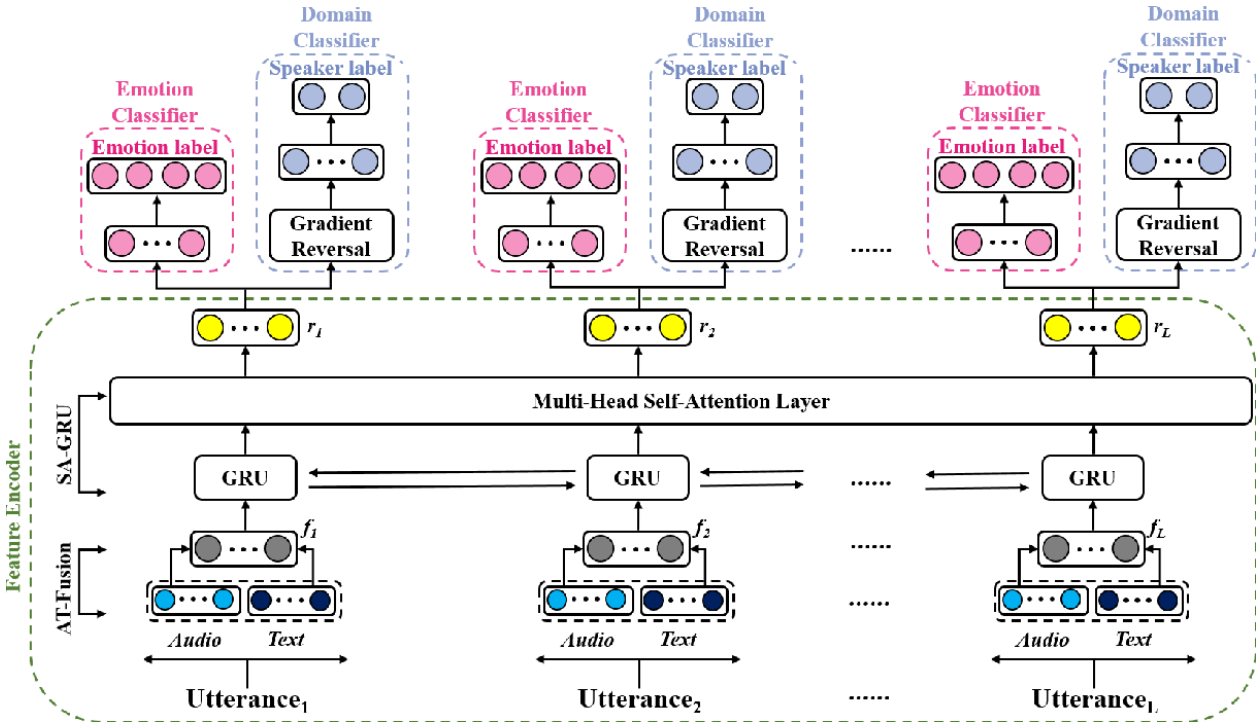

Рисунок 1.15 – Модель DANN для розпізнавання емоцій у аудіо

VQ-MAE-S-12 – модель самонавчання, яка налаштована для розпізнавання емоцій з мовних сигналів. Модель VQ-MAE-S базується на

маскованому автокодувальнику (MAE), який працює у дискретному латентному просторі векторизованого квантувального варіаційного автокодувальника[15].

Кодер VQ-MAE-S, аналогічно до архітектури ViT, складається з одного трансформера (рисунок 1.16). Цей кодер є стеком L резидуальних блоків, який включає шар самоуваги, шар нормалізації та блок багатошарового персептрона. VQ-MAE-S-12 також токенізує віхдну спектрограму перед блоком трансформеру.

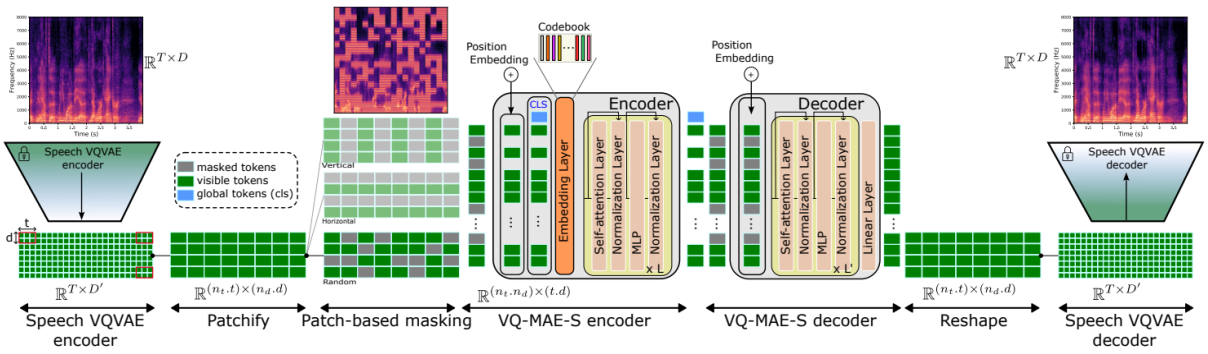

Рисунок 1.16 – Модель VQ-MAE-S-12

Експериментальні результати показують, що запропонована модель VQ-MAE-S, передзавантажена на набір даних VoxCeleb2 та налаштована на емоційні мовні дані, перевершує MAE, що працює на сирому представленні спектрограм та інші передові методи у визначенні емоцій звуку.

VAVL (Versatile Audio-Visual Learning) є мультимодальною систему, яка працює навіть коли дані лише однієї модальності доступні і можуть бути реалізовані як взаємозамінні для прогнозування емоційних атрибутів, так і для категоризації емоцій[16].

Як показано на рисунку 1.17 , акустичний і візуальний шари дзеркально відображають один одного. Обидва шари мають однакову базову структуру і механізм навчання. Основними компонентами цих шарів є кодери-конформери, які обробляють всі послідовні відео- або акустичні кадри паралельно, в залежності від модальності, доступної під час навчання або

логічного висновку.

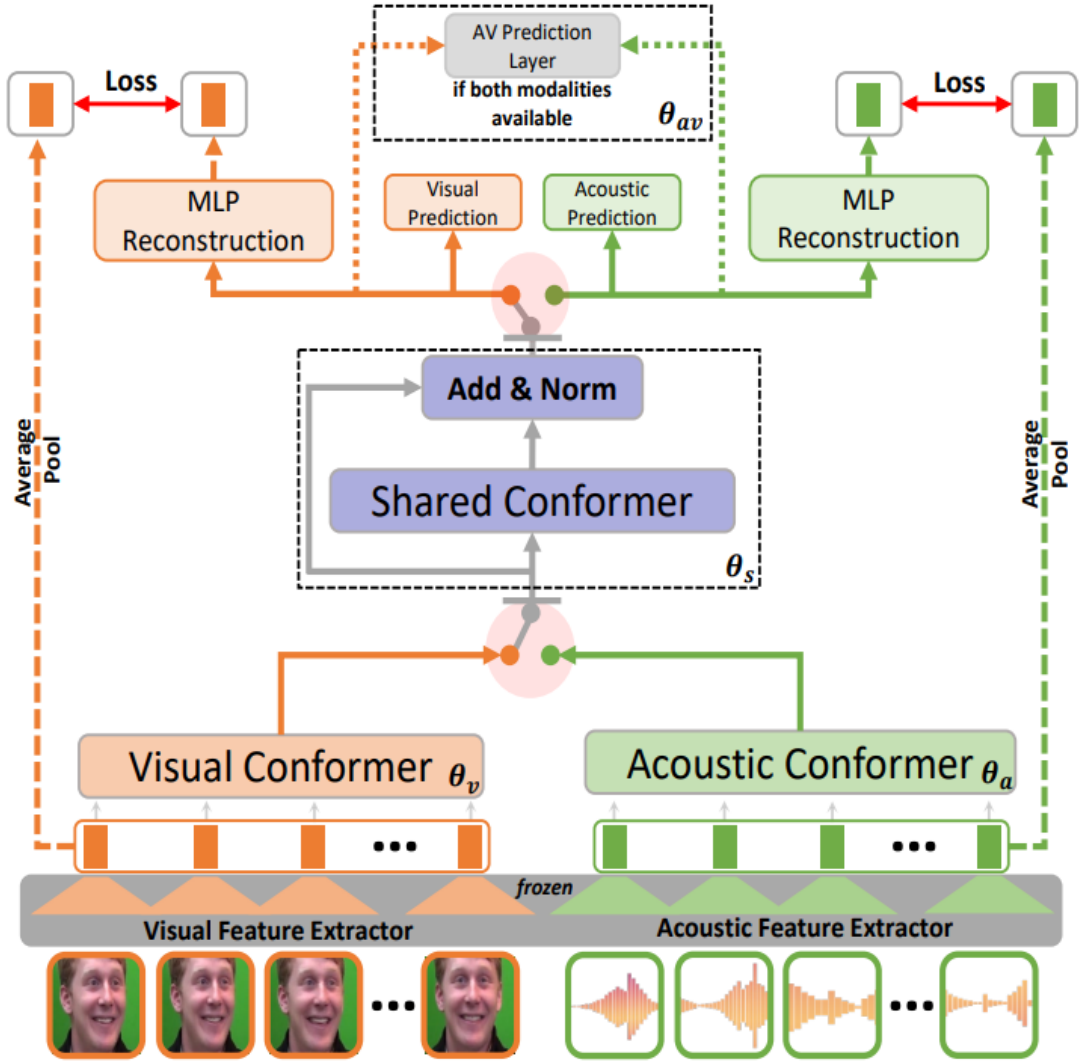

Рисунок 1.17 – Модель VAVL

Для кожної модальності окремо реалізовано два додаткові компоненти (акустичні та візуальні), які реалізуються для того, щоб передбачити емоційні атрибути та реконструювати унімодальні ознаки.

Міттел та ін. [17] запропонували багатомодальне мультимодальне визнання емоцій (M3ER): спочатку з використанням трьох методів отримано вектори ознак. Потім ці ознаки подаються на етап оцінки моделі для отримання ефективних та неефективних ознак, і перші використовуються для побудови проксі-векторів ознак. Нарешті, шляхом злиття обраних ознак вони використовуються для визначення шести емоцій за допомогою злиття остаточного рівня ознак з модулем уваги. Точність визнання M3ER в

IEMOCAP становила 82,7%. Результати показали, що точність визнання представленої багатомодальної моделі перевершувала унімодальну модель та інші багатомодальні моделі в наборі даних.

Загальна архітектура моделі [23] розпізнавання емоцій з використанням мультимодальної фузії для мовних виразів на основі глибинного навчання показана на рисунку 1.18. Вона головним чином розділена на три частини: модуль вилучення ознак, модуль вибору ознак і модуль класифікації емоцій. Ця архітектура лягла в основу поточної роботи, однак вона не бере до уваги текстові дані – транскрипцію розмови.

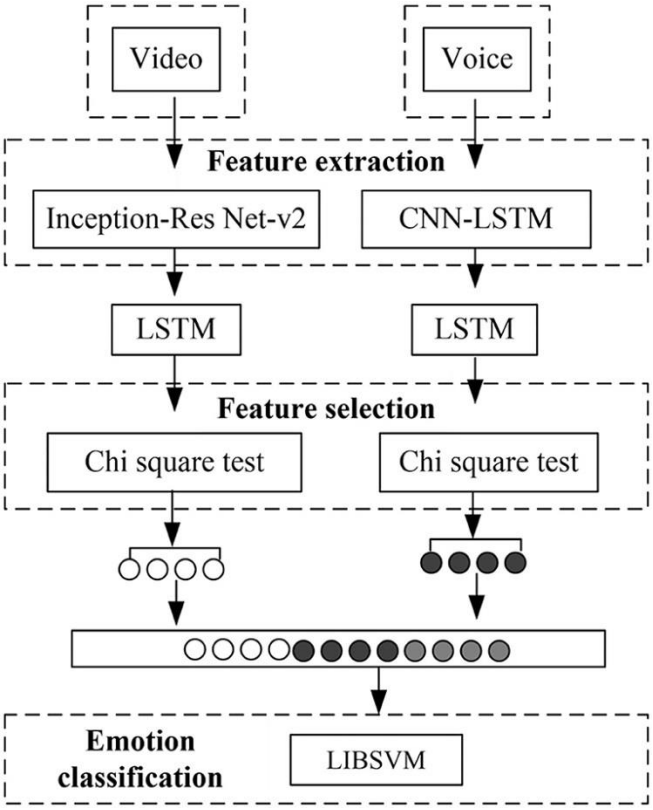

Рисунок 1.18 – Модель злиття аудіо- та відео-ознак

У [22] побудована модель мультимодального розпізнавання на основі об'єднання аудіо- та текстової інформації. Модель злиття складається з двох паралельних гілок, які обробляють аудіо та текст окремо до рівня, на якому інформація з обох гілок об'єднується (рисунок 1.19). Моделі відрізняються

місцем розташування шару об'єднання, мережею, що додається після об'єднання, та підходом до навчання.

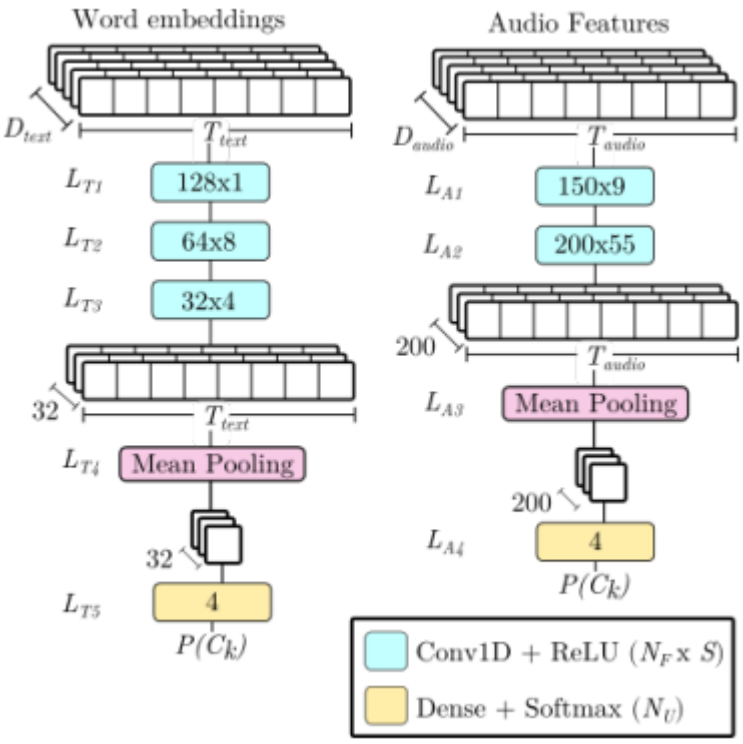

Рисунок 1.19 – Моделі для розпізнавання емоцій з тексту та аудіо розмови

1.4 Обґрунтування доцільності вдосконалення існуючих рішень

Моделі глибокого навчання досягли значних успіхів у задачах SER за останні роки, та їх точність на розглянутих датасетах не сягає 85%. Це пов’язано зі складністю самої задачі розпізнавання емоцій, з цією складністю так само стикається людська інтуїція.

Розпізнавання емоцій залишається складним завданням через варіації у говорців та обмежені навчальні вибірки.

На відміну від людської інтуіції, системи SER стикаються з проблемами узагальнення, при переході від тренування на спеціально підготовлених записах у виконанні акторів до практичного застосування.

1.5 Мета та задачі дослідження

Метою роботи є дослідження методів визначення емоційного стану спікера. Запропоноване рішення має давати перевагу перед існуючими за наступними критеріями: підвищена точність, отримане прискорення, вивільнення обчислювального ресурсу центрального процесору.

Для досягнення поставленої мети мають бути вирішені наступні задачі:

- провести аналіз методів, способів та підходів до ідентифікації емоційного стану людини за її мовленням та обрати найкращий;
- застосувати обраний спосіб ідентифікації змін емоційного стану людини за її мовленням до розв’язання задачі автоматизованого розпізнавання проявів емоцій;
- реалізувати обраний спосіб у вигляді інформаційної системи автоматизованого розпізнавання проявів емоцій;
- провести експериментальне тестування інформаційної системи за еталонними наборами даних.

Розвиток більш точних та ефективних моделей машинного навчання для розпізнавання емоцій в мовленні є однією з головних цілей. Додавання контекстуальної інформації до процесу розпізнавання емоцій може покращити точність [13]. Це може включати аналіз мовного контексту, використання додаткових сенсорів (наприклад, відео або сенсорів фізіологічних показників) та розуміння ситуацій. Об'єднання даних з різних джерел, таких як мовлення, обличчя, жести, може допомогти у створенні більш повних та точних моделей розпізнавання емоцій.

Збільшенням застосування систем розпізнавання емоцій важливо вивчати етичні та приватні аспекти їхнього використання. Це включає в себе забезпечення конфіденційності даних та врахування можливих негативних наслідків застосування таких технологій.

2 АНАЛІЗ ТЕХНОЛОГІЧНОГО ТА МЕТОДОЛОГІЧНОГО ПІДҐРУНТЯ ДЛЯ ВИРІШЕННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ

2.1 Аналіз технологій для вирішення поставленої задачі

Для вирішення задачі було обрано мову програмування Python, оскільки для неї існує величезна кількість бібліотек для машинного навчання та наукових обчислень, таких як NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch та інші. Ці бібліотеки забезпечують потужні інструменти для роботи з даними, побудови моделей та їхнього навчання.

Python 3, це остання стабільна версія мови програмування Python на момент мого останнього оновлення. Python має простий та легкий для вивчення синтаксис, що дозволяє швидко розробляти та тестувати алгоритми машинного навчання. Також Python має ряд потужних бібліотек для візуалізації даних, таких як Matplotlib та Seaborn, які допомагають аналізувати дані та результати моделей.

TensorFlow та PyTorch – це два з найпопулярніших фреймворків глибокого навчання (deep learning) у сучасному світі. Обидва фреймворки мають свої переваги та недоліки.

Таблиця 2.1 – Порівняння фреймворків машинного навчання

TensorFlow	PyTorch
Граф обчислень статичний, розробник описує його заздалегідь. Вбудовані засоби для візуалізації графів, профілювання виконання та інші інструменти для полегшення розробки та налагодження моделей	Граф обчислень динамічний; розробник прописує саме кроки алгоритму як звичайну послідовність команд-операцій над тензорами; при виконанні цих команд рушій Torch будує граф обчислень для градієнтного спуску.

TensorFlow використовує статичні обчислювальні графи: граф обчислень спочатку визначається, а потім виконується. PyTorch використовує динамічні обчислювальні графи: граф створюється "на льоту" під час виконання коду, що дозволяє більшу гнучкість у визначенні моделей та експериментах.

Numpy – це одна з найпопулярніших бібліотек Python для роботи з колекціями даних і TensorFlow має інтеграцію з її структурами даних. Тип масиву NDarray автоматично конвертується у тензор TensorFlow під час виконання.

TensorFlow та PyTorch спроектовані для глибокого навчання за алгоритмами заснованими на обчислені градієнтів, вони не призначені для еволюційних алгоритмів.

Neuralfit – це фреймворк для побудови і навчання нейронної мережі за алгоритмом NEAT.

2.2 Аналіз методологічного підґрунтя для рішення поставленої задачі

Штучні нейронні мережі – це технологія апроксимації заздалегідь невідомої функції, натхненна біологічними нейронними мережами. Вона складається з безлічі штучних нейронів поєднаних одне з одним, зазвичай організовані у шари.

Штучний нейрон – математичний об’єкт, що має функцію активації f та матрицю вагових коефіцієнтів для вхідних сигналів (вхідні дані представлені у вигляді тензору, або вихідні сигнали з нейронів попереднього шару).

$$o(x) = f\left(\sum x_i \times w_i\right), \tag{2.1}$$

де o – сигнал на виході нейрону;

f – функція активації (часто нелінійна, як relu або tanh);

x_i – вхідні сигнали;

w_i – вагові коефіцієнти нейрона.

Якщо сигнали поширюються в одному напрямку, починаючи від вхідного шару нейронів, через приховані шари до вихідного шару і на вихідних нейронах отримується результат опрацювання сигналу – це називається нейронна мережа прямого поширення (feed-forward network). Протилежним типом нейронних мереж із зворотніми зв’язками є рекурентні нейронні мережі (recurrent neural network, RNN).

Підбір таких вагових коефіцієнтів кожного нейрона, які дозволяють краще апроксимувати цільову функцію (мінімізувати функцію помилки), називається тренуванням мережі. Оскільки навчання є задачею оптимізації, існує безліч алгоритмів його проведення. Найпопулярніші для нейронних мереж є алгоритми на основі градієнтного спуску (gradient descent), що використовує обчислення похідних функції помилки від ваг нейронів.

Глибокими нейронними мережами називаються такі, де існує один чи більше прихованих шарів – нейронів, які беруть на вхід сигнали, обчислені попереднім шаром та видають сигнали не на вихід мережі, а на входи наступного шару нейронів. Обчислення градієнтів для таких мереж є складнішим процесом, що потребує обчислення похідних за правилом ланцюга. Глибокі нейромережі є найпотужнішим видом машинного навчання на сьогодні, їх популярність зумовлена здатністю апроксимувати функції будь-якої складності.

Автоматичне диференціювання є важливою складовою кожної бібліотеки глибокого навчання. До складу PyTorch Рушій автоматичного диференціювання Autograd.

Тренування нейронної мережі складається з двох кроків:

- прямий прохід, коли обчислюються вихідні значення і функція помилки;
- зворотній, коли рахується градієнт помилки на нейронах;
- зворотній прохід потребує обчислювати градієнти за правилом ланцюга.

Алгоритми, засновані на диференціюванні можуть порівняно швидко знаходити локальні мінімуми функції складної помилки. До недоліків можна

віднести:

- обмеження на функції активації – вони всі мають бути диференційовані;
- складність обчислень (може бути прискорена за допомогою GPU, тому розвиток глибокого навчання послідував за розвитком обчислень загального призначення на GPU);
- ризик застрягнення у локальних мінімумах;
- проблема затухаючих градієнтів у рекурентних шарах.

Конкурентною альтернативою є еволюційні алгоритми, такі як NEAT, та їх поєднання з алгоритмами на основі градієнтів.

На рисунку 2.1 показано втрати в навчанні та тестуванні нашого набору даних. Як видно з графіка, помилки як «навчання, так і тестування» зменшуються зі збільшенням числа епох навчальної моделі.

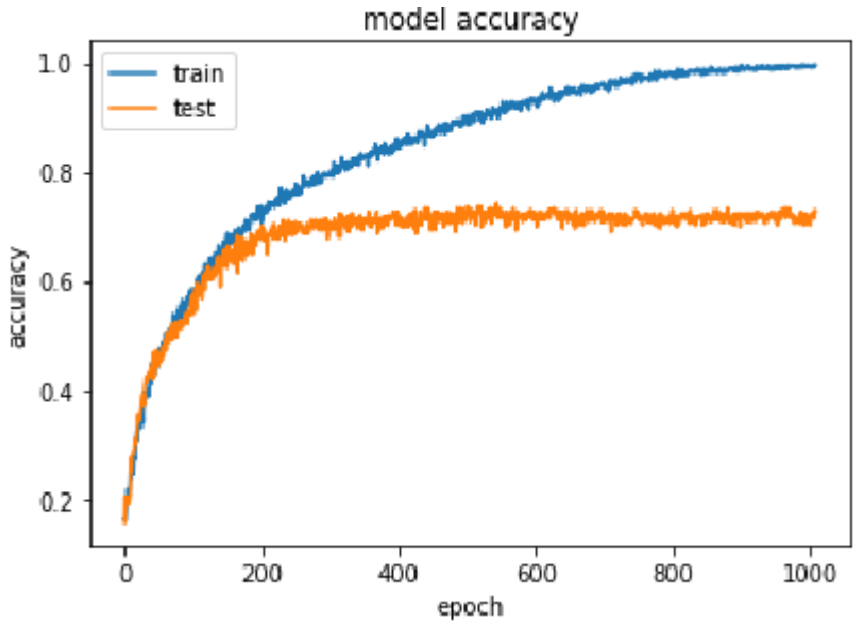

Рисунок 2.1 – Графік тренування, залежність точності на тренувальному і тестовому наборі даних від кількості епох тренування

Зазвичай структура нейронної мережі визначається людиною заздалегідь, і тому може знадобиться деякий процес спроб і помилок, щоб

дійти до архітектури, яка буде здатна навчитися апроксимувати цільову функцію, та буде робити це ефективно. Алгоритми, такі як NEAT, автоматизують цей процес, починаючи з простої одношарової топології і поступово нарощуючи складність.

2.2.1 Використані архітектури нейронних мереж

Для розпізнавання емоцій у мовленні (SER) зазвичай використовуються різні архітектури нейронних мереж, зокрема: згорткові, рекурентні, автокодувальники, трансформери.

Згорткові нейронні мережі (CNN), часто використовуються для SER завдяки їхній здатності екстрагувати ознаки зі спектрограм або інших часово-частотних представлень мовленнєвих сигналів.

CNNs використовуєть згорткові операції лінійної алгебри для екстракції характерних ознак і знаходження форми у зображеннях.

В CNN фільтр (або ядро) – це невелика матриця ваг нейронів, що ковзає по вхідному тензору, робить поелементне перемноження з підматрицею вхідної матриці над якою фільтр знаходиться, і потім сумує результат множення, записуючи значення одного пікселя вихідного тензору (рисунок 2.2). В циклі або паралельно, фільтр ковзає по вхідних даних, щоразу переміщаючись на певну кількість пікселів, визначену «кроком» (stride).

Під час навчання CNN вивчає оптимальні значення вагових коефіцієнтів фільтрів для вирішення поставленого. Це робиться за допомогою зворотного поширення та алгоритмів оптимізації, як і будь-які інші ваги в нейронній мережі.

Розмір фільтра є важливим гіперпараметром. Найчастіше вживані розміри включають 3x3, 5x5 і 7x7.

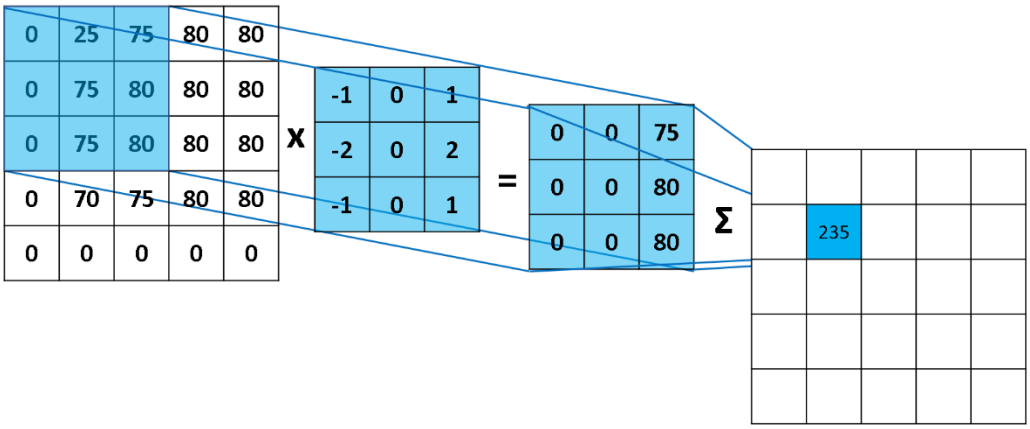

Рисунок 2.2 – Принцип роботи фільтра CNN

Рекурентні нейронні мережі (RNN), такі як LSTM (Long Short-Term Memory) або GRU (Gated Recurrent Unit), ефективні для захоплення часових залежностей у мовленнєвих сигналах і часто використовуються для моделювання послідовних даних у SER [7][8][9]. Оскільки RNN приймає як вхідні дані свій вихід з попереднього кроку, її застосування до послідовності даних (такої як звукова хвиля, фрейми спектрограми тощо) не може бути обчислено паралельно. Паралельно можна обчислювати прохід RNN по кільком незв’язаним послідовностям.

LSTM – це RNN спроектована для кращого вивчення довгострокових залежностей. Моделі LSTM складаються з трьох різних компонентів (шлюзів, gates). Є вхідний шлюз, вихідний шлюз та шлюз забування.

Вхідний шлюз приймає рішення про те, які значення є важливими та мають бути пропущені через LSTM. У вхідному шлюзі використовується сигмоїдна функція, яка визначає, які значення передавати далі на вихід. Шлюз забування займається видаленням інформації, яку модель вважає непотрібною для подальшого аналізу. LSTM зберігають «прихований» (hidden state) і «стан клітини» (cell state).

GRU – це мережі LSTM зі спрощеним механізмом шлюзів. Вони мають два шлюзи: reset та update, і зберігають лише один прихований стан.

Гібридні моделі: Деякі системи SER поєднують як CNN, так і RNN,

щоб використовувати їхні переваги у вилученні ознак та моделюванні часу.

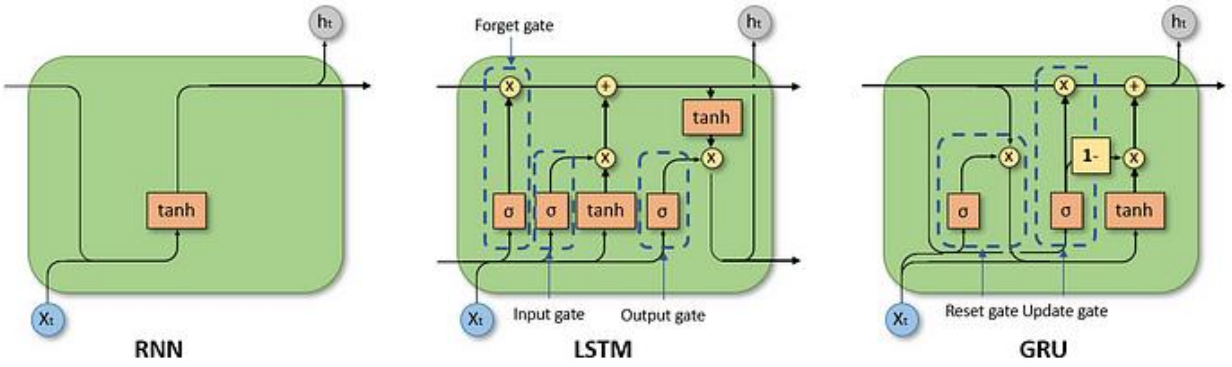

Рисунок 2.3 – Моделі рекурентних нейронних мереж

Архітектури трансформерів, схожі на ті, які використовуються у завданнях обробки природної мови, також застосовуються до SER останнім часом і досягають передових результатів [11], захоплюючи далекі залежності у мовленнєвих сигналах ефективно.

Архітектура трансформеру складається з кодера і декодера. Кодер отримує на вході векторизовану послідовність з позиційною інформацією. Декодер отримує на вході частину цієї послідовності і вихід кодера. Кодер і декодер складаються з шарів. Шари кодера послідовно передають результат наступного шару в якості його входу. Шари декодера послідовно передають результат наступного шару разом з результатом кодировщика в якості його входу. Кожен кодер складається з механізму самоуваги (вихід з попереднього шару) і нейронної мережі з прямим зв'язком (вихід з механізму самоуваги). Кожен декодер складається з механізму самоуваги (вихід з попереднього шару), механізм уваги до результатів кодування (вихід з механізму самоуваги і кодування) і нейронної мережі з прямим зв'язком (вихід з механізму уваги).

Механізми уваги часто включаються у архітектури нейронних мереж [13][15] для SER, щоб зосередитися на важливих частинах вхідного мовленнєвого сигналу і покращити здатність моделі розпізнавати емоції.

2.2.2 Допоміжні функції

Для побудови спектрограм потрібен алгоритм швидкого перетворення Фур'є (Fast fourier Transform, FFT) – це алгоритм для ефективного обчислення дискретного перетворення Фур'є і його інверсії. Він здатний прискорити обчислення перетворення Фур'є на ряд порядків за рахунок використання спеціальної структури симетрії вхідних даних. Це особливо корисно при обробці сигналів та інших ситуаціяї, де потрібно швидко обробляти великі обсяги даних, таких як тренування нейронних мереж FFT дозволяє обчислити результат дискретного перетворення Фур'є з N точок даних використовуючи O(N log N) операцій, радше за O(N 2) потрібних для прямого алгоритму.

Дискретне перетворення Фур'є перетворює ряд чисел $a_{0,}\ldots,a_{n-1}$ в ряд $b_{0,}\ldots,b_{n-1}$ такий що

$$b_i = \sum_{j=0}^{n-1} a_j \varepsilon^{ij}, \tag{2.2}$$

де $\varepsilon^n = 1$, $\varepsilon^k \neq 1$, при $0 < k < n$.

Коефіцієнти мел-частотного кепстрального аналізу (MFCC) – це характеристика, яка широко використовується у обробці мовлення та аудіо для завдань, таких як розпізнавання мови та ідентифікація диктора. Для виділення MFCC:

- аудіосигнал розділяється на короткі фрейми зазвичай 20-40 мілісекунд, з невеликим перекриттям між фреймами. Це дозволяє нам аналізувати сигнал протягом коротких проміжків часу, припускаючи, що характеристики сигналу залишаються відносно сталі в межах кожного фрейму;

- до кожного фрейму застосовується функція вікна, наприклад, вікно Хеммінга, для зменшення спектрального витоку;

- швидке перетворення Фур'є (FFT): Потужність спектра кожного фрейму обчислюється шляхом взяття модуля перетворення Фур'є віконованого сигналу. Це дає нам представлення сигналу в частотному

домені;

- потужність спектра потім проходить через набір трикутних фільтрів, які розташовані рівномірно на мел-шкалі. Ці фільтри розроблені для імітації відповіді людського вуха на різні частоти. Вихід кожного фільтра представляє енергію в різних частотних діапазонах;

- береться логарифм енергії в кожному фільтрі. Це служить для імітації людського сприйняття гучності, яке приблизно логарифмічне.

- дискретне косинусне перетворення (DCT): до логарифму енергії кожного фільтра застосовується дискретне косинусне перетворення, а отримані коефіцієнти є MFCC. Зазвичай лише коефіцієнти нижчих порядків залишаються, оскільки вони містять більшість важливої інформації про спектральну оболонку сигналу.

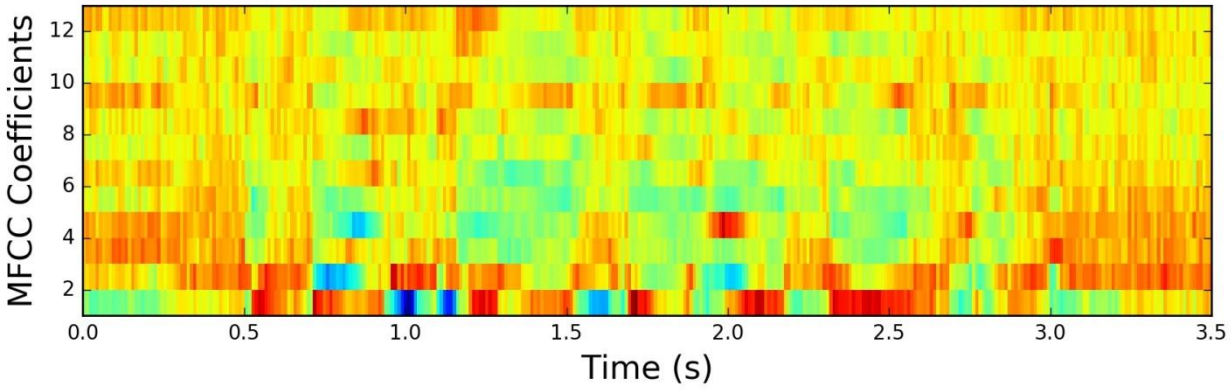

Рисунок 2.4 – Візуалізація отриманих мел-частотних кепстральних коефіцієнтів

Класифікатор Softmax — це функція активації, яка використовується в нейронних мережах, зокрема для задач класифікації. Вона перетворює логіти (незбалансовані виходи нейронної мережі) у вірогідності, які сукупно складають 1. Кожен елемент вектора на виході Softmax відповідає ймовірності належності вхідного зразка до певного класу.

Softmax для вектора z = [z1,z2,…,zK] визначається так:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \quad (2.3)$$

де zi — і-ий елемент вхідного вектора логітів, K — кількість класів.

Softmax має властивість нормаліації. Всі виходи Softmax знаходяться в інтервалі (0, 1) і їхня сума дорівнює 1. Це дозволяє інтерпретувати виходи як ймовірності.

В нейронних мережах Softmax зазвичай використовується разом з крос-ентропійною функцією втрат (cross-entropy loss) для навчання моделей класифікації.

Cross-entropy loss вимірює розбіжність між розподілом прогнозованих ймовірностей та фактичним розподілом (цільовими мітками).

Для багатокласової класифікації, де K — кількість класів, формула крос-ентропії виглядає так:

$$L = \sum_{j=1}^{K} y_i \log(\hat{y}_i), \quad (2.4)$$

де yi — індикатор (0 або 1) того, чи є клас i правильним класом для даного зразка, y^i — передбачена ймовірність належності до класу i.

Cross-entropy loss високо оцінює моделі, які мають високу впевненість у правильному прогнозі. Чим ближче передбачена ймовірність до істинної мітки, тим менша втрата.

Завдяки логарифмічній природі, крос-ентропія значно збільшується для неправильних передбачень, що допомагає моделі швидше навчатися на помилках. Функція крос-ентропії є диференційованою, що дозволяє використовувати її для градієнтного спуску та інших методів оптимізації.

Критерій хі-квадрат (χ^2) — це статистичний тест, який використовується для визначення, чи існує статистично значуща різниця між очікуваними та спостережуваними частотами у категоричних даних. Він часто застосовується для аналізу таблиць спряженості, де досліджуються

зв'язки між двома або більше категоріальними змінними. Він може застосовуватися для відбору статистично значущих ознак у нейромережі.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad (2.5)$$

де O і — спостережувані частоти,

Eі — очікувані частоти.

Ознаки з високими значеннями χ^2 є більш релевантними для цільової змінної.

2.2.3 Розглянуті датасети

Якісний датасет для навчання містить точні та надійні дані, що сприяє покращенню точності моделі. Якщо дані мають багато шуму або помилок, модель може навчитися неправильно і давати некоректні прогнози. Якщо датасет є репрезентативним і охоплює широкий спектр можливих сценаріїв, модель буде здатна краще узагальнювати та застосовувати знання до нових, невідомих даних, що і є кінцевою ціллю навчання класифікатора.

RAVDESS включає записи 24 різних акторів, що вимовляють дві фрази (dogs are sitting by the door, kids are talking by the door) з різним емоційним забарвленням (спокій, радість, сум, злість, страх, здивованість та огида) у двох інтенсивностях.

У CREMA-D 91 актор (різний вік, стать, національність) повторює 12 речень у 6 різних емоціях (злість, огида, радість, спокій, сум) у 3 рівнях інтенсивності.

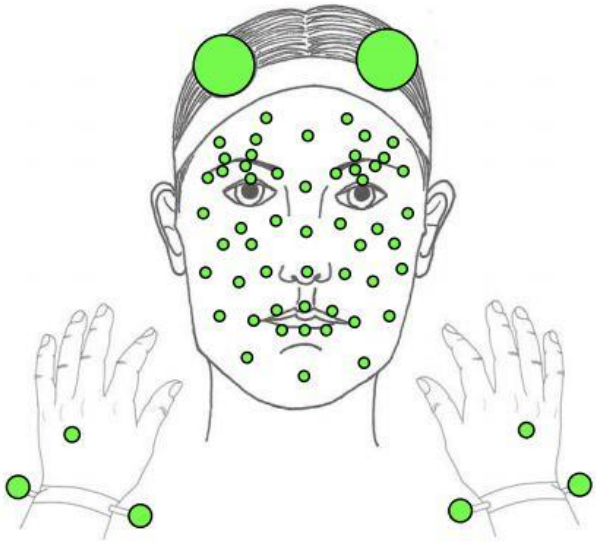

Рисунок 2.5 – Розташування маркерів для захоплення руху на обличчях акторів IEMOCAP

База даних EMODB – це вільно доступна німецька база даних з семи емоцій: 1) гніву; 2) нудьга; 3) тривожність; 4) щастя; 5) смуток; 6) відраза; 7) нейтральний. Десять професійних спікерів (5 чоловіків і 5 жінок) взяли участь у записі даних. База даних містить загалом 535 висловлювань, що складаються в діалоги.

Таблиця 2.2 – Розглянуті датасети

Датасет	Датапоінтів	Модальність	Кількість мовців	Кількість емоцій	мова
IEMOCAP	302	Відео	10	9	Анг.
CREMA-D	7442	Аудіо	91	6	Анг.
RAVDESS	7356	Відео	24	7	Анг.
TESS	2800	Аудіо	2	7	Анг.
EMODB	535	Аудіо	10	7	Нім.

IEMOCAP записаний в Інституті інформаційних наук Університету Південної Каліфорнії (USC) у 2008 і є одним із найпопулярніших наборів даних для вивчення емоцій через його багатогранність і ретельну анотацію, що робить його цінним ресурсом для різних дисциплін. Він складається з 151 запису відео діалогів між акторами, з 2 спікерами на сесію. Кожен сегмент підписаний за наявністю 9 емоцій: злість, збудженість, страх, сум, здивованість, фрустрація, радість, розчарування та спокій. Включає 5 сесій по 5 пар спікерів (Таблиця 2.2).

Відео діалогів записані разом із захопленням рухів, через що у акторів на обличчах нанесені маркери (рисунок 2.5). Це впроваджує додаткову модальність, проте такий тип даних зазвичай не зустрічається у практичних ситуаціях застосування автоматичного розпізнавання емоцій.

TESS записан в Університеті Торонто. Цей датасет містить виключно жіночі голоси і має дуже високу якість аудіо. Більшість інших наборів даних схиляється до чоловічих голосів, що призводить до певної дисбалансованої репрезентації. У наборі даних є 200 цільових слів, які були вимовлені у фразі «Скажи слово _» двома актрисами віком 26 та 64 роки. Записи були зроблені для кожного з семи емоційних станів (злість, відраза, страх, радість, приємний сюрприз, смуток та нейтральний). Загалом є 2800 розмічених аудіофайлів.

У кожному датасеті записи були анотовані людськими експертами, які позначали емоції в кожному фрагменті.

3 ЗАПРОПОНОВАНИЙ МЕТОД

3.1 Запропонована архітектура моделі системи визначення емоційного стану спікера

Ми пропонуємо метод, який відбирає результати, отримані в результаті обробки аудіо, тексту та зображень, вибираючи загальні та ефективні ознаки для розуміння вираженої емоції. Тому нам потрібна модель, здатна витягувати ефективні ознаки іншим способом, ніж існуючі алгоритми. У цій роботі представлено метод (рисунок 3.1) вилучення ефективних ознак в аудіо та тексті, а також метод вилучення ефективних ознак в аудіо та зображеннях. Після вилучення результати цих двох методів об'єднуються за допомогою злиття даних і використовуються як вхідні дані для класифікатора.

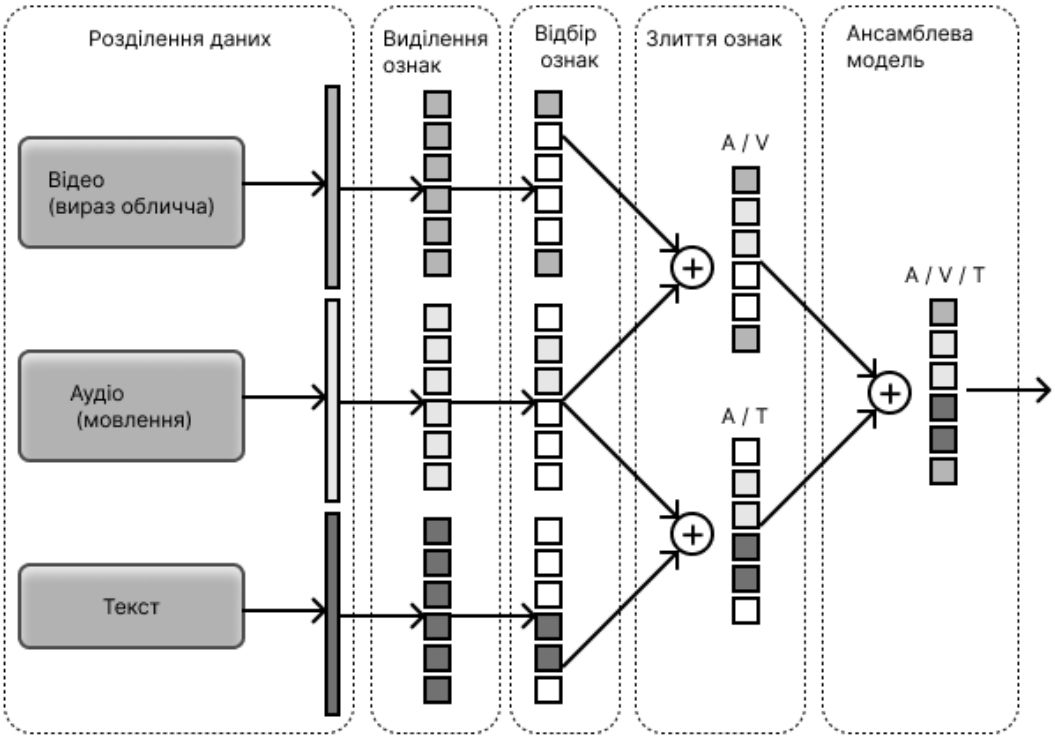

Рисунок 3.1 – Узагальнена схема роботи трьохступеневого методу визначення емоцій людини на основі штучного

3.1.1 Метод екстракції ознак з аудіо- і відеоданих

Загальна архітектура мультимодального композитного методу розпізнавання емоцій для аудіо- та відео-ознак за допомогою глибокого навчання показана на рисунку 3.2.

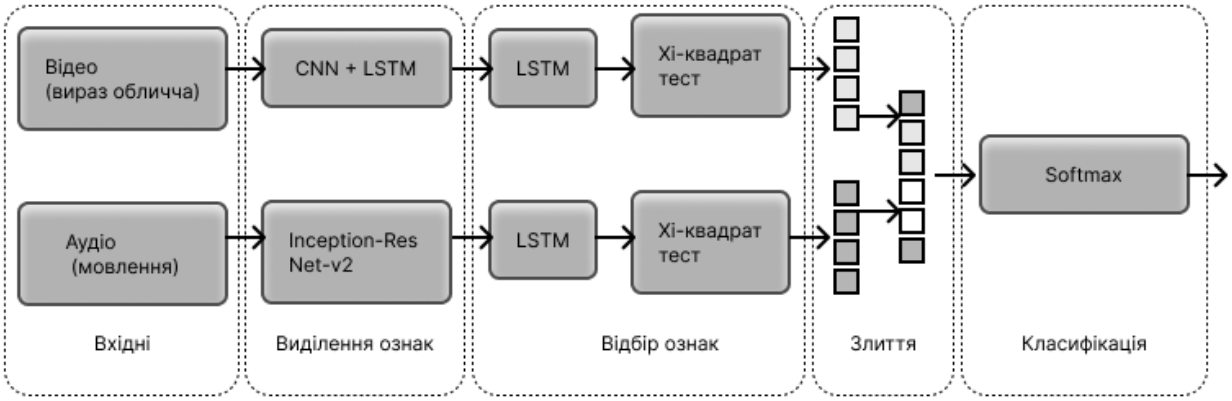

Рисунок 3.2 – Схема роботи методу злиття аудіо-відео даних (A/V)

Загалом він поділяється на модулі вилучення ознак, відбору ознак та класифікації емоцій. Вхідні дані не можна використовувати для категоризації емоцій без підготовки. Тому модальність даних слід змінити. Модуль розділення перетворює відеодані на аудіо- та відеодані, а потім перетворює аудіодані на текст.

Модуль вилучення ознак лежить в основі мультимодального методу розпізнавання емоцій, який спрямований на отримання вхідних даних для обробки інформації та перетворення їх на ознаки, які можуть бути використані в моделі [19].

В модулі розглядалися різні схеми для вилучення ознак. Для аудіоданих використовується мережа (CNN-LSTM), а для відео – мережа Inception-Res Net-v2 для вилучення відповідних ознак.

У багатогранному емоційному стані існує кореляція між структурою LSTM і ознакою, яка може бути використана для розпізнавання часових залежностей від часових рядів. У зв'язку з цим між різними методами існує

інформаційний зв'язок. Ці зв’язки можуть бути використані як загальні та приховані сигнали емоцій, щоб допомогти прийняти рішення щодо класифікації.

Inception-ResNet-v2, глибока нейронна мережа для класифікації зображень, яка поєднує ідеї з двох архітектур: Inception і ResNet. Ця модель, запропонована дослідниками Google, представляє собою удосконалену версію об'єднання цих двох підходів, забезпечуючи покращену продуктивність і ефективність.

Мережа використовує модулі Inception (рисунок 3.3), які об'єднують кілька конволюційних шарів з різними розмірами фільтрів і операціями підсумовування, що дозволяє ефективно витягувати ознаки на різних масштабах. Ця архітектура ефективно використовує параметри, зменшуючи обчислювальні витрати порівняно зі звичайними конволюційними шарами.

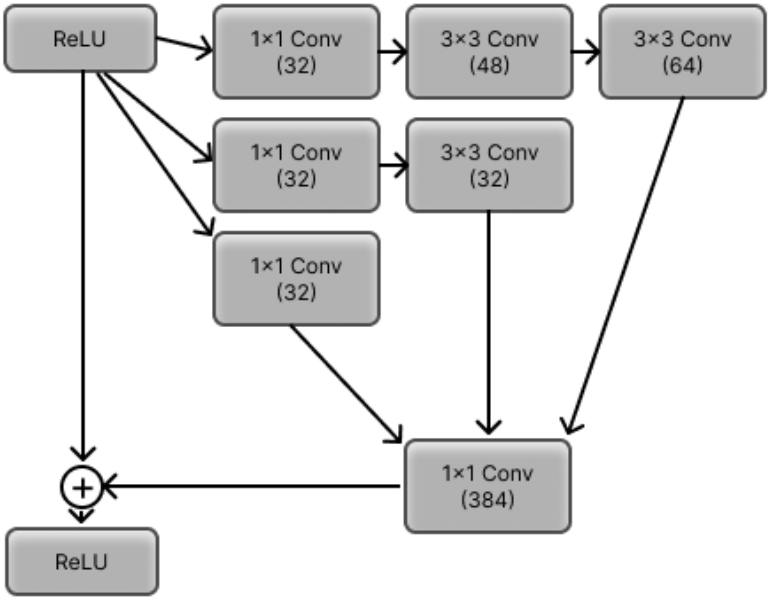

Рисунок 3.3 – Модуль Inception-resnet-A

Мережа поєднує паралельні конволюційні шари з залишковими зв'язками ResNet, що дозволяє моделі вивчати більш складні патерни в даних. Мережа приймає зображення фіксованого розміру 299x299 на вхід.

Inception-ResNet-v2 послідовно з’єднує Stem модуль, 3 “Inception-ResNet” модуля та модулі зменшення розмірності (рисунок 3.4).

Модуль Stem (рисунок 3.5) обробляє вхідне зображення за допомогою серії згорток, пулінгу та шарів нормалізації, щоб зменшити просторові розміри та підготувати карти ознак для подальшої обробки.

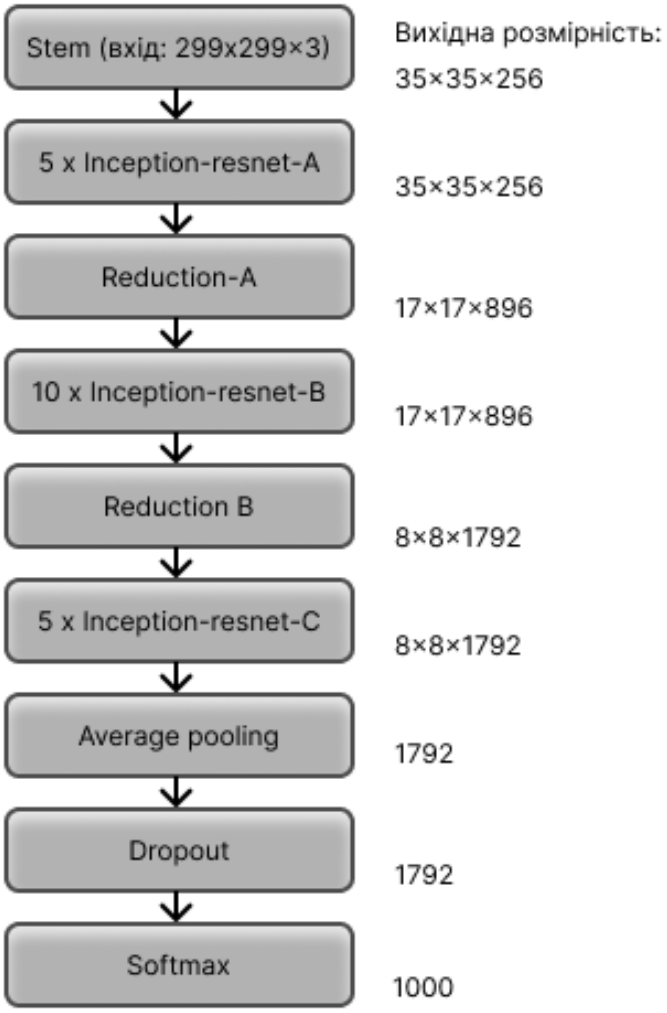

Рисунок 3.4 – Шари Inception-ResNet-v2

Inception-ResNet-v2 сумарно має 1,280 тренованих параметра

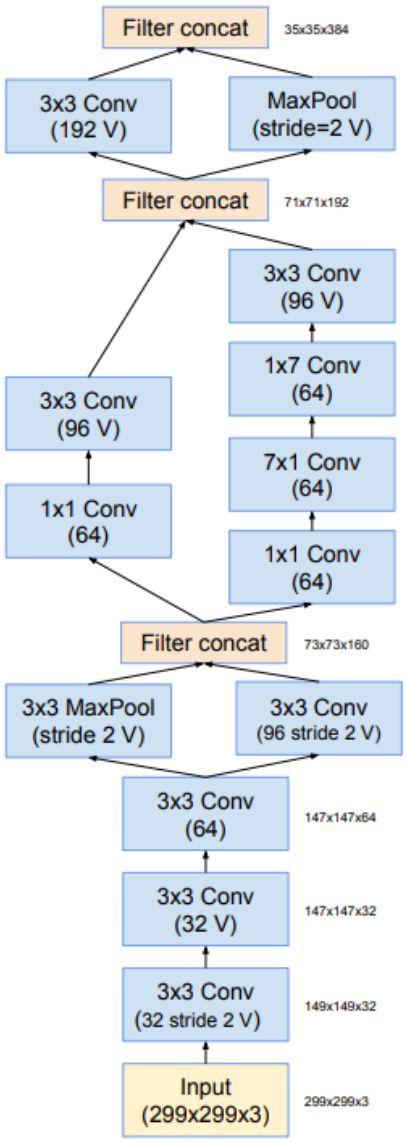

Рисунок 3.5 – Модуль Stem в Inception-Resnet-v2

У певних точках мережі просторові розміри зменшуються за допомогою згорток з кроком або операцій пулінгу, щоб знизити обчислювальну складність і збільшити поле огляду. Це зменшення є важливим для керування глибиною мережі та забезпечення того, щоб кінцеві шари могли працювати з достатньо великими представленнями ознак.

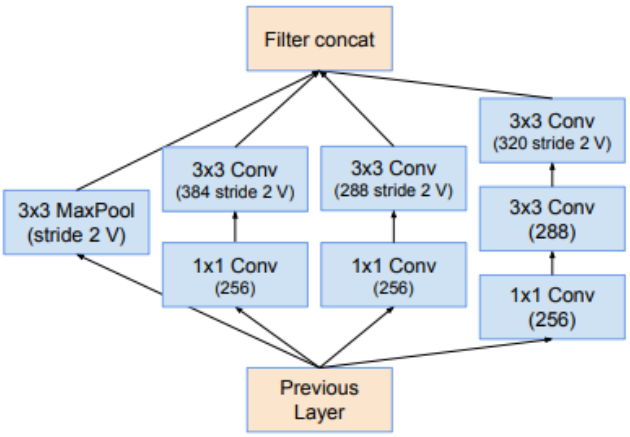

Рисунок 3.6 – Модуль зменшення розмірності

3.1.2 Вилучення ознак із аудіосигналів

Аудіо — це сигнал із властивістю нелінійної послідовності часу, яка повністю залежить від часу. Таким чином, рекурентна мережа LSTM підходить для вилучення аудіо-ознак. Під час моделювання інформації з урахуванням контексту дуже корисно знати часову інформацію про емоційні ознаки звуку. Однак у мережі LSTM немає нелінійного прихованого шару, що збільшує коефіцієнт прихованого режиму (hidden mode coefficient). Тому тут використовується комбінована мережа LSTM і CNN для вивчення емоційних особливостей аудіо (рисунок 3.7).

Структурні особливості CNN мережі об’єднуються в мережу CNN-LSTM і додається рівень LSTM. Щоб перетворити вхідні дані з рівня згортки, а потім для повторного навчання, ми вводимо дані до рівня LSTM. Кореляційна частина складається з максимального шару агрегації та двох згорткових шарів. Рівень згортки має параметри, подібні до параметрів мережі CNN, а у функції активації використовується лінійна одиниця (ReLU).

Вихід CNN надходить безпосередньо до LSTM у спільній мережі CNN-LSTM. Таким чином можна отримати розширену інформацію, включаючи довгострокові контекстні та локальні інформаційні залежності. Однак CNN залишають багато даних і зосереджуються на локальній інформації. Було

представлено метод із використанням каналів CNN і Bi-LSTM для запобігання втраті важливих даних. Набір із чотирьох одновимірних згорткових шарів використовувався в каналі CNN з різною кількістю фільтрів для одновимірного тимчасового введення аудіо-ознак. Крім того, шари максимізаційного агрегування та загального усереднювального агрегування були застосовані для виконання операцій об'єднання середнього та максимального об'єднання даних.

Набір комірок Bi-LSTM із 256 аргументами було розміщено в каналі Bi-LSTM для вилучення інформації про довготривалі текстові залежності, а механізм уваги використовувався для пошуку більш ефективних ознак. Нарешті дані з двох каналів об'єднували, а вихідні дані поміщали в щільний шар. Після індукування нелінійної зміни кореляція цих характеристик була визначена та відображена у вихідному просторі.

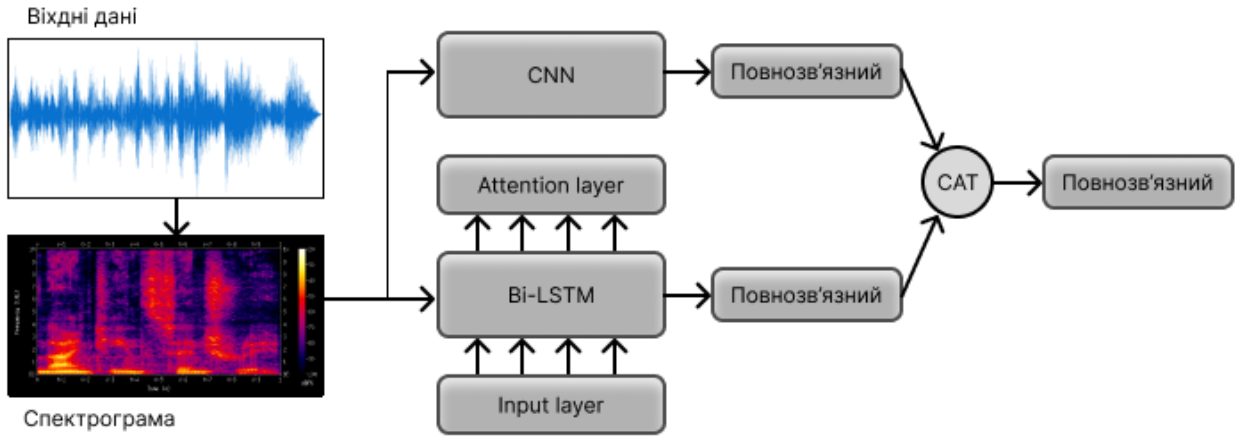

Рисунок 3.7 – Схема аудіо-модуля

3.1.3 Метод виділення ознак з відео обличчя

Обсяг відеоданих зазвичай дуже великий. Тому нам потрібно визначити контур обличчя в кожному кадрі відео та зменшити розмір зображення кадру за допомогою кадрування контуру обличчя, що необхідно

для обробки даних і отримання ознак.

Тут інформація про емоцію з відео визначається шляхом короткочасного виділення ознак за допомогою мережі Inception-Res Net-v2 і перетворення Фур'є сигналу (STFT) для глибокого виділення ознак.

Inception-Res Net-v2 означає CNN, розроблену Google у 2016 році, представляючи мережу ResNet на основі моделі Inception. Мережа вважається прототипом Inception V3, який використовує ідею підключення в моделі Microsoft ResNet, де нейронна мережа отримує більш глибоке навчання. Можна суттєво спростити модуль Inception, використовуючи кілька згорток 3 × 3 замість згорток 5 × 5 і 7 × 7. Як наслідок, обчислювальна складність і розміри параметрів значно зменшуються, а швидкість навчання мережі збільшується.

3.1.4 Вибір аудіо- та відео-ознак для злиття (A/V)

Структура LSTM-RNN використовується для розуміння внутрішньої залежності ознак у кожній модальності. Залежність існує як між ознаками різних модальностей, так і між внутрішніми ознаками однієї модальності. З обмеженою кількістю ознак в одній модальності наявна інформація з іншої модальності може бути використана для прийняття рішень класифікатором.

Оскільки унімодальна структура є h-вимірною ми можемо представити ознаки в мультимодальній за допомогою вектора ознак $X_{i,t} \in R^h$, де R^h та t представляють t-ту множину t у відео i. Усі вектори у відео можна зібрати для створення векторної матриці $X_i = [x_1, x_2, \ldots, x_{iL}] \in R^{h.Li}$, Li представляє всю кількість речень у відео та може використовуватися як вхідні дані для LSTM.

Цей підхід ґрунтується на підході фільтрації для вибору ознак, щоб покращити можливість узагальнення методу розпізнавання емоцій і зменшити складність обчислень. Підмножини ознак вибираються з даних загальних ознак, а статистичні методи використовуються для присвоєння

балів ознакам. Потім ці ознаки ранжируються в порядку спадання на основі їхніх балів, і ознаки з вищим рейтингом зберігаються в підмножині ознак, а непов’язані ознаки відфільтровуються. Крім того, тест хі-квадрат використовується для відображення ознак і відповідних кореляцій класів, де чим вищий бал, тим більша залежність від відповідного класу, а ознаки з нижчими балами містять менше інформації та повинні бути виключені. Таким чином, ознаки з низькими оцінками містять менше інформації та їх потрібно вилучити. У цьому контексті усувається помилкова інформація, надлишкова інформація та шум.

3.1.5 Класифікація аудіо- та відео-ознак (A/V)

Аудіофайли та вирази обличчя зливаються як остання ознака мультимодального сигналу. У цьому методі Softmax використовується як класифікатор для класифікації емоцій. Кожен канал розглядається як група окремих мультимодальних сигналів для класифікації емоцій. Злиття (fusion) виконується на рівні прийняття рішень для класифікації результатів, створених кожним класифікатором Softmax.

3.2 Виділення ознак для злиття аудіо- та текстових ознак (A/T)

Аудіосигнал і текстова інформація спочатку обробляються, щоб виділити ознаки емоцій низького рівня (рисунок 3.8).

Потім аудіо-ознаки вводяться в модель для представлення локальної та загальної інформації. Щоб отримати ознаки високого рівня, текстові ознаки додаються до нейронних мереж Bi-LSTM. Потім використовується об’єднання ознак для об’єднання емоційних особливостей голосу та тексту.

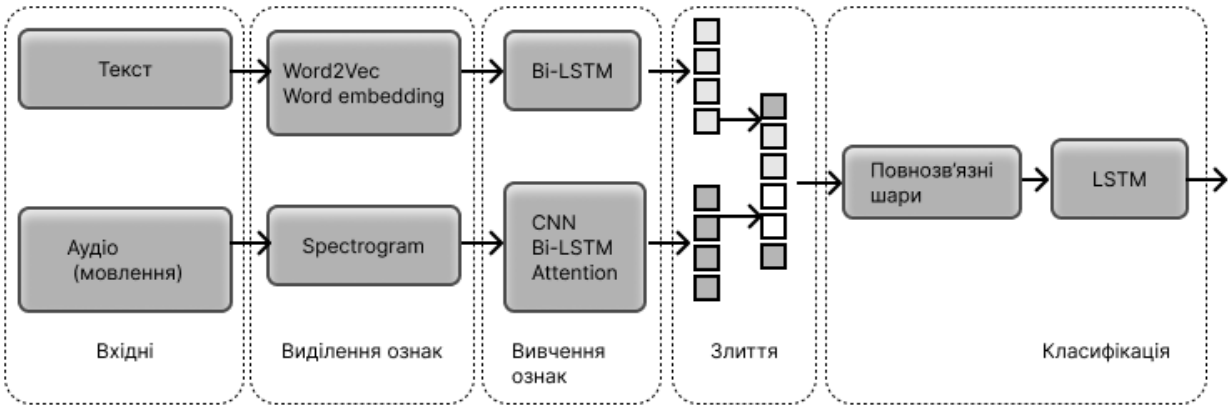

Рисунок 3.8 – Схема роботи методу здиття аудіо- та текстових даних (A/T)

Нарешті, глибока нейронна мережа використовується для вивчення та класифікації ознак після злиття У мультимодальному методі мережа DNN використовується для навчання та оптимізації злиття ознак.

3.2.1 Виділення аудіо-ознак

Процес починається з обробки сирого аудіосигналу за допомогою даних з бази даних. Потім витягуються 34-вимірні низькорівневі аудіоознаки зі спектральної, часової та кепстральної областей, такі як рівень енергії, швидкість перетину нуля, ентропія енергії, спектральний центр, спектральне розширення, спектральна ентропія, спектральний потік і Мел-частотні кепстральні коефіцієнти (MFCC). Максимальна кількість вхідних даних становить 100 кадрів, і вектор отримується для кожного виразу (100, 34). Багато ознак, таких як MFCC, енергія, швидкість перетину нуля і спектральний потік, враховуються для представлення сенсорних характеристик портативної акустики.

3.2.2 Виділення текстових ознак

Векторне представлення слів (word embedding) є популярним методом у вилученні ознак тексту, використовуючи щільні вектори для представлення

документів та слів. Word2vec вважається одним з найбільш поширених методів векторного рпедставлення слів. Це спосіб створення матриці слів, яка поєднує інформацію про загальні та локальні статистичні дані слів для отримання лінгвістичних моделей та векторів слів [20].

Метод векторного представлення (embedding) Word2vec використовується для текстових даних і отримує вектори емоційних ознак тексту. 300-вимірні вектори Word2vec з максимальною довжиною послідовності 500 використовуються для отримання векторів для кожного виразу (300, 500). У Word2vec нелінійні активаційні функції (сигмоїдна, tanh, ReLU тощо) не використовуються в обчисленнях Softmax на останньому шарі, і виходи надсилаються як зважені комбінації входів.

$$j` = u_j \left(\sum x_i \times w_{i,j} \right), \tag{3.1}$$

де u_j – сигнал j-го прихованого нейрону, x_i – вхідні сигнали, (для векторів слів лише один x_i дорівнює 1), $w_{i,j}$ – вагові коефіцієнти між і-м віхдним та j-м прихованим нейроном.

3.2.2 Злиття аудіо- та текстових ознак

Після введення аудіоознак у модель і текстових ознак у мережу Bi-LSTM можна отримати високорівневі текстові ознаки $H=\{h_1, h_2, ..., h_t\}$ та акустичні ознаки $S=\{s_1, s_2, ..., s_t\}$, які складаються із загальної та локальної інформації. Тут застосовано підхід злиття на рівні ознак. Основною перевагою є те, що емоційні ознаки, отримані з різних станів, безпосередньо пов'язані з кінцевим рішенням, а злиття результатів може значно зберегти інформацію, необхідну для кінцевого рішення. Остаточний мультимодальний дескриптор емоційних ознак — це вектор $V=\{v_1, v_2, ..., v_t\}$, який розроблений за допомогою регулярної конкатенації текстових H і акустичних S ознак наступним чином

$$V = [H, S], \tag{3.2}$$

Потім вектор ознак злиття V надходить у DNN з трьома щільними шарами з параметрами, рівними 1024, 512, 4, і шаром Softmax, щоб показати взаємозв'язок ознак [21]. Вихід класифікатора Softmax вказує на відносну ймовірність різних класів емоцій.

3.3 Злиття аудіо-, відео- та текстових ознак

На фінальному етапі, після створення об'єднаних виходів A/V та T/V, настав час об'єднати їх і отримати остаточний висновок. У цьому розділі в останньому шарі класифікації нейронної мережі використовується мультиноміальна логістична регресія, також відома як Softmax-регресія. У мультиноміальній логістичній регресії та аудиторському аналізі або лінійному аудиторському аналізі (linear audit analysis) результати обчислень, виконаних за допомогою k різних лінійних функцій, розглядаються як вхідні дані для цієї функції. Із входом X та матрицею ваг W, вихідна ймовірність j-го класу наступна:

$$p(y = j|X) = \frac{e^{x^T wj}}{\sum_{k=1}^{k} e^{x^T wk}}, \tag{3.3}$$

що може бути представлено як комбінація k лінійних функцій і функції плавного максимуму.

$$X \rightarrow X^T W_1, \dots, X \rightarrow X^T W_k, \tag{3.4}$$

де $X^T W_k$ – скалярний добуток векторів X та W.

4 ТЕСТУВАННЯ ТА АНАЛІЗ

4.1 Метод оцінки

Після підготовки сценаріїв методи використовуються для перевірки точності вилучення ознак та результату їх злиття для більш точного розпізнавання емоцій. Для виконання тесту та аналізу результатів використовується база даних IEMOCAP відповідно до описаних вище кроків.

Для оцінки запропонованого методу була використана база даних емоцій IEMOCAP. У цій базі даних майже 12 годин аудіовізуальних даних, таких як аудіо, відео, записи обличчя та текстові транскрипції.

Для вилучення зайвої і шумової інформації використовуємо критерій хі-квадрат. Після передобробки та вибору ознак визначаються унімодальні ознаки. Нарешті, ці два методи поєднуються для розпізнавання мультимодальних емоцій.

4.2 Порівняння ефективності злиття з іншими методами

Проведено серію тестів для порівняння продуктивності моделі з використанням критерію хі-квадрат та без нього, щоб перевірити важливість використання критерію хі-квадрат для видалення зайвої та шумової інформації в методі мультимодального розпізнавання емоцій.

4.2.1 Результати окремих модулів

Шляхом поєднання мереж CNN і Bi-LSTM отримано кращу модель для розпізнавання мовлення та тексту з вищою точністю, що підтверджує вірність методу злиття.

Результати проведених експериментів надано у вигляді таблиць

розбіжностей, які складаються з рядків, що представляють фактичні класи, тобто класи, до яких належать дані, а також стовпців, які представляють прогнозовані класи, тобто класи, до яких модель класифікувала дані. На перетині рядка та стовпця знаходиться значення, яке показує скільки разів модель класифікувала дані з фактичного класу (рядка) до прогнозованого класу (стовпця).

Таблиця 4.1 – Матриця розбіжностей для аудіо та тексту

A/T	Злість	Щасття	Нейтр.	Сум
Злість	74.92	3.58	3.83	16.81
Щасття	1.92	73.25	2.08	22.03
Нейтр.	2.81	2.51	80.85	13.39
Сум	2.41	9.62	6.68	80.82

Загалом, робота з відео показує вищу точність ніж робота з аудіо та текстом. Це збігається з результатами попередніх досліджень і можна пояснювати тим, що основним методом вираження емоцій для людини є міміка.

Таблиця 4.2 – Матриця розбіжностей для аудіо та відео

A/V	Злість	Щасття	Нейтр.	Сум
Злість	71.89	2.02	11.55	14.17
Щасття	4.73	77.92	2.74	14.29
Нейтр.	4.48	2.22	80.93	12.05
Сум	2.91	9.87	12.16	74.62

Аналіз по аудіо і відео також дозволяє краще розрізняти вираз щасття і сум. Найвищий рівень помилки у розрізенні суму та нейтрального виразу. Впізнавання злості погіршелося порівнянно з аудіо-текстовою моделлю, а розрізнення щасття та суму поліпшелося.

4.2.2 Результати злиття аудіо, відео та текстових

Після злиття аудіо та тексту, аудіо та відео моделей, для остаточної оцінки використовується регресія, і результати зазначених моделей порівнюються з виходом відео, аудіо та текстової моделей за допомогою матриці невідповідностей (таблиця 3).

Як показано в таблицях 4.1 - 4.3, існує невідповідність у емоційних результатів унімодальної моделі. Таблиця 3 показує, що більшість типів емоційної точності покращуються, а невідповідність зменшується за рахунок злиття емоційних ознак аудіо і тексту. Це свідчить про вірність злиття моделей, яке може ефективно зменшити невідповідність та збільшити рівень впізнавання емоційних станів.

Таблиця 4.3 – Матриця розбіжності для методу злиття

A/V/T	Злість	Щасття	Нейтр.	Сум
Злість	79.62	1.26	2.1	16.51
Щасття	1.62	82.9	0.3	14.9
Нейтр.	4.07	1.48	80.94	13.22
Сум	1.91	9.81	7.03	80.88

Після злиття з аудіо-текстовою моделлю, бачимо збільшення точності впізнавання всіх емоцій, особливо злості. Бачимо незначне збільшення помилки класифікації суму як злості. Точність зросла з 76.34% до 82.9% після включення в аналіз тексту отриманого з аудіо, радше за використання лише аудіо та відео-даних. Комбінована модель використовує лексичні значення вимовлених слів як важливі маркери емоцій.

4.3 Аналіз результатів

Деякі емоції краще розпізнаються за допомогою мовлення, такі як

страх і сум, тоді як інші краще розпізнаються за допомогою відео, наприклад, щастя та гнів. У класифікації емоцій голосом, гнів та сум неправильно розпізнаються як вирази щастя та нейтральності, тоді як у класифікації виразів обличчя гнів може бути неправильно розпізнаний як сум, а щастя як нейтральний вираз.

З аудіо-моделлю спостерігалося вищий рівень точності для класів нейтральності та суму порівняно з текстом, але різниця не була такою великою, як для класів гніву та щастя. Декілька випадків щастя були неправильно класифіковані класифікатором. Однак ефективність класифікатора була дуже хорошою для розрізнення між гнівом та сумом. Крім того, деякі випадки щастя були неправильно класифіковані як нейтральні.

Хоча сумні та сердиті обличчя можна було ефективно класифікувати, класифікатор продемонстрував деяку невідповідність. Відокремлення нейтральних класів було точніше в порівнянні з іншими класами. Однак там були випадки помилкової класифікації як сумних і щасливих обличч.

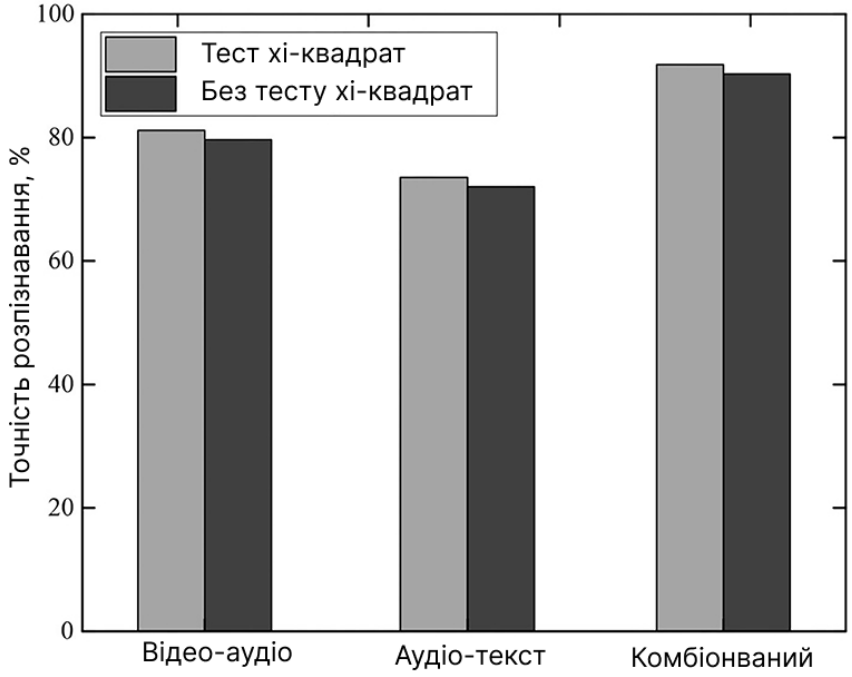

Рисунок 4.1 – Порівняння точності розпізнання методами з (а) вибором значущих ознак з допомогою тесту хі-квадрат, (б) без

У разі злиття мовлення і тексту, а також мовлення і зображення, точність розпізнавання вище, ніж з унімодальними моделями для всіх ознак (рисунок 4.1), і невідповідність зменшується в випадках помилкової класифікації.

У кількох тестах тримодальні та бімодальні моделі працюють краще, ніж унімодальні моделі. Відбір більш значущіх ознак з допомогою тесту хі-квадрат також показав підвищену ефективність на кожній бі-модальній моделі.

4.4 Порівняння зі схожими методами

У [22] досліджено різні підходи до класифікації емоцій за мовленням, використовуючи акустичні та текстові ознаки. Запропоновано отримувати контекстуалізовані векторні представлення слів за допомогою BERT для відображення інформації, що міститься в транскрипціях мовлення, і показуємо, що це призводить до кращої продуктивності порівняно з використанням векторних представлень Glove.

Виявлено, що об'єднання акустичних і текстових систем є корисним для обох наборів даних, хоча спостерігаються лише незначні відмінності між оціненими підходами до об'єднання. Продемонстровано значний вплив критеріїв, використаних для визначення складів перехресної валідації, на результати для IEMOCAP. Зокрема, стандартний спосіб створення складів для цього набору даних призводить до дуже оптимістичної оцінки продуктивності текстової системи, що натякає на те, що деякі попередні роботи можуть переоцінювати перевагу використання транскрипцій.

У [23] Запропоновано метод розпізнавання емоцій у мовленнєвих виразах на основі глибокого навчання з використанням багатомодальної ф'юзії. По-перше, налаштовуються відповідні методи екстракції ознак для різних одиничних модальностей. Серед них голос використовує мережу згорткових нейронних мереж-довготривалої та короткотривалої пам'яті

(CNN-LSTM), а для екстракції ознак обличчя на відео використовується мережа Inception-ResNet-v2. Потім для захоплення кореляції між різними модальностями та в межах модальностей використовується довготривала та короткотривала пам'ять (LSTM). Після процесу відбору ознак за допомогою критерію хі-квадрат одиничні модальності об'єднуються для отримання єдиної злитої ознаки.

Таблиця 4.4 – Огляд тестування методів на датасеті IEMOCAP

Праця	Рік	Модальність	Спосіб злиття	Точність
Mittel et al. [17]	2020	Текст-аудіо	Рівень ознак та рішення	4 класи: 80.51
Liu Dong et al. [23]	2021	Текст-аудіо	Рівень ознак	7 класів: 65.1
Pepino et al. [22]	2020	Відео-текст-аудіо	Рівень ознак	4 класи: 82.7
поточна	2024	Відео-текст-аудіо	Рівень ознак та рішення	4 класи: 82.9

Нарешті, злиті ознаки, що виходять з LSTM, використовуються як вхідні дані для класифікатора LIBSVM для реалізації кінцевого розпізнавання емоцій. Результати експериментів показують, що точність розпізнавання запропонованого методу на наборах даних MOSI та MELD становить відповідно 87,56% і 90,06%, що краще, ніж у інших методів порівняння. Це заклало певний теоретичний фундамент для застосування багатомодальної ф'юзії в розпізнаванні емоцій.

ВИСНОВКИ

В ході робти було розглянуто задачу розпізнавання емоцій спікера у мовленні, що залишається досить складною задачею класифікації через схожість у мовленні проявів різних емоцій і різницю у інтонації різних мовців, на яку впливає біологічний вік і стать людини. Були розглянуті методи машинного навчання, які для цього використовуються, зокрема глибокі нейронні мережі на базі рекурентних мереж, згорткових мереж і трансформерів.

Визначено, що використання спектрограм мовлення як вхідних даних дозволяє моделям машинного навчання точніше виявляти емоційні стани спікера, аніж робота з одновимірним аудіопотоком. Для впізнавання проявів емоцій у мовленні в першу чергу важливі закономірності зміни частоти і гучності голосу протягом часу.

Було розглянуто датасети у відкритому доступі, які використовуються для тренування моделей розпізнавання емоцій. Виявлено, що їх мітки відповідають психологічним моделям з семи-восьми дискретних емоцій та (в деяких) - інтенсивності, подібно до класифікації Плутчика.

Була поставлена мета вдосконалення рішення задачі розпізнавання емоцій у мовленні шляхом підбору моделі машинного навчання і функцій передобробки, які б дозволили підвищити точність розпізнавання емоцій з голосового сигналу.

Були обрані інструменти для розробки: мова програмування Python і фреймворк машинного навчання TensorFlow для побудови нейронної мережі, допоміжні бібліотеки NumPy, Pandas та Scikit-learn.

Була представлена мультимодальна модель розпізнавання емоцій використання аудіо, відео та тексту з бази даних IEMOCAP. Спочатку для визначення акустичних ознак емоцій використовували LSTM та двоканальну CNN. Потім Bi-LSTM використовується для виділення елементів тексту.

Нарешті, було використано DNN для злиття ознак. Модель використовувала мережі глибокого навчання, штучні та вдосконалені ознаки та враховувала час та текстову інформація в даних. Для оптимізації моделі також використовувалася адаптація L2.

Метод бімодального злиття (A/V) використовувався одночасно з використання глибокого навчання LSTM-CNN та Inception-ResNet-v2 для екстракції ознак виразу обличчя в аудіо та відео. Крім того, отримані ознаки були подані в блок LSTM та окремі методи були пов’язані між собою, щоб отримати злиття ознак за допомогою відбору ознак тестом хі-квадрат.

Вихідні ознаки з LSTM були подані в класифікатор Softmax для розпізнавання емоцій. Результати тесту на наборі даних IEMOCAP показали, що ефективність розпізнавання емоцій була найвищою при вазі мережі близько 0,57, та додаванні тесту хі-квадрат. У цьому випадку виділення ознаки познозв’язного шару мережі Inception-Res Net-v2 було найвищим.

Цей метод намагався усунути шум і надмірність внутрішніх ознак. Проте були деякі комплементарні кореляції в інформації модуля.

Нарешті, методи (A/T) і (A/V) були злиті, і результат розпізнавання емоцій було отримано запропонованим методом (A/T/V) з точністю 82,9%.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, Sandra L. Schneider (2023): Speech emotion recognition using machine learning — A systematic review, Intelligent Systems with Applications, Volume 20, 2023, 200266, ISSN 2667-3053, https://doi.org/10.1016/j.iswa.2023.200266.

2. Антон, Смотрицький & Radiuk, Pavlo. (2023). Розпізнавання мімічних проявів емоцій засобами штучного інтелекту. 10.13140/RG.2.2.33509.58085.

3. Опанасенко, Я. П. Система розпізнавання емоцій людини на відео : магістерська дис. : 121 Інженерія програмного забезпечення / Опанасенко Ярослав Павлович. – Київ, 2021. – 75 с.

4. Anvarjon, T.; Mustaqeem; Kwon, S. Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. Sensors 2020, 20, 5212.

5. Опанасенко, Я. П. Система розпізнавання емоцій людини на відео : магістерська дис. : 121 Інженерія програмного забезпечення / Опанасенко Ярослав Павлович. – Київ, 2021. – 75 с.

6. Lei, S., Dong, G., Wang, X., Wang, K., & Wang, S. (2023). InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework. ArXiv, abs/2309.11911.

7. Li, J., Liu, Y., Wang, X., & Zeng, Z. (2023). CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition. ArXiv, abs/2307.15432.

8. Hu, G., Lin, T., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022). UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. Conference on Empirical Methods in Natural Language Processing.

9. Li, J., Wang, X., Lv, G., & Zeng, Z. (2022). GA2MIF: Graph and Attention-based Two-stage Multi-source Information Fusion for Conversational Emotion

Detection. ArXiv, abs/2207.11900.

10.Kim, J., & Lee, S. (2023). CoordViT: A Novel Method of Improve Vision Transformer-Based Speech Emotion Recognition using Coordinate Information Concatenate. 2023 International Conference on Electronics, Information, and Communication (ICEIC), 1-4.

11.Croitoru, F., Ristea, N., Ionescu, R.T., & Sebe, N. (2022). LeRaC: Learning Rate Curriculum. ArXiv, abs/2205.09180.

12.Ristea, N., Ionescu, R.T., & Khan, F.S. (2022). SepTr: Separable Transformer for Audio Spectrogram Processing. ArXiv, abs/2203.09581.

13.Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z., & Li, R. (2020). Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. Interspeech.

14.Sadok, S., Leglaive, S., & S'eguier, R. (2023). A Vector Quantized Masked Autoencoder for Speech Emotion Recognition. 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 1-5.

15.Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., & Fernández-Martínez, F. (2021). A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. Applied Sciences.

16.Lucas Goncalves, Seong-Gyun Leem, Wei-Cheng Lin, Berrak Sisman, Carlos Busso (2023): Versatile Audio-Visual Learning for Handling Single and Multi Modalities in Emotion Regression and Classification Tasks. ArXiv, abs/ 2305.07216.

17.Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. Proceedings of the AAAI Conference on Artificial Intelligence, 34(02), 1359-1367. https://doi.org/10.1609/aaai.v34i02.5492.

18.Jianfeng Zhao, Xia Mao, Lijiang Chen (2019): Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and

Control, Volume 47, 2019, Pages 312-323, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2018.08.035.

19.Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency (2018): Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. ArXiv, abs/1811.09362.

20.Mittal, Trisha & Bhattacharya, Uttaran & Chandra, Rohan & Bera, Aniket & Manocha, Dinesh. (2020): M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. Proceedings of the AAAI Conference on Artificial Intelligence. 34. 1359-1367. 10.1609/aaai.v34i02.5492.

21.Sebastian, Jilt & Pierucci, Piero. (2019): Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. 51-55. 10.21437/Interspeech.2019-3201.

22.Pepino, L., Riera, P., Ferrer, L., Gravano, A.: Fusion approaches foremotion recognition from speech using acoustic and textbased fea-tures. In: ICASSP 2020 - 2020 IEEE International Conference onAcoustics Speech Signal Process. ICASSP, IEEE, pp. 6484–6488. Barcelona, Spain (2020). https://doi.org/10.1109/ICASSP40776.2020.9054709Pub.

23.Liu, D., Wang, Z., Wang, L., Chen, L.: Multi-modal fusion emotionrecognition method of speech expression based on deep learning.Front. Neurorobotics (2021). https://doi.org/10.3389/fnbot.2021.697634.