

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет информатики, математики и компьютерных наук

**Программа подготовки бакалавров по направлению
01.03.02 Прикладная математика и информатика**

ПРОЕКТНАЯ РАБОТА

Многомодальная система распознавание эмоциональных состояний
человека

Студенты

Сидорова Анна Павловна,
Жиляева Юлия Николаевна,
Дудникова Екатерина Олеговна

Руководители проекта

Абросимов Кирилл Игоревич,
Львутина Татьяна Владимировна

Нижний Новгород, 2022

Содержание

Введение	5
1 Изучение датасета MELD	6
1.1 Анализ части Audio	7
1.1.1 Изменения в датасете	7
1.1.2 Выделение аудиозаписей из видео	7
1.1.3 Анализ	7
1.1.3.1 Краткий анализ длительности аудио в монологах	9
1.1.3.2 Аудио максимальной длины	9
1.2 Анализ части Text	16
1.3 Анализ части Video	18
2 Изучение моделей и задачи анализа эмоций	21
2.1 Классификация Экмана	21
2.2 Классификация Изарда	21
2.3 Модель Плутчика	22
2.4 Arousal и Valence	23
2.5 Сентимент-анализ	24
2.5.1 Методы определения тональности	24
3 Анализирование статей	26
3.1 M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation	26
3.1.1 Структура	27
3.1.1.1 Multi-modal Fusion Network: M2FNet . . .	27
3.1.1.2 Извлечение признаков на уровне высказываний	28
3.1.1.3 Модуль извлечения признаков	29
3.1.2 Результаты	30
3.2 DialogueTRM: Exploring Multi-Modal Emotional Dynamics in a Conversation	31
3.2.1 Модель	32
3.2.1.1 Динамика внутримодальных эмоций . . .	32
3.2.1.2 Контекстно-зависимые условия (текст) . .	32

3.2.1.3	Контекстно-свободные условия (аудио, видео)	33
3.2.1.4	Динамика интермодальных эмоций	33
3.2.1.5	Прочее	33
3.2.2	Результаты	34
3.3	DialogGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation	35
3.3.1	Методология	35
3.3.2	Контекстно независимое извлечение признаков на уровне высказывания	35
3.3.3	Модель	36
3.3.3.1	Последовательный кодировщик контекста	36
3.3.3.2	Кодировщик контекста на уровне говорящего	36
3.3.3.3	Классификатор эмоций	37
3.3.4	Применение датасета MELD	38
3.4	COSMIC: COmmonSense knowledge for eMotion Identification in Conversations	40
3.4.1	Методология	40
3.4.2	Архитектура модели COSMIC	41
3.4.3	Применение датасета MELD, результаты	42
3.5	UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition	44
3.5.1	Формулировка задачи:	44
3.5.2	Результаты	46
3.6	Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning	47
3.6.1	О чем статья:	47
3.6.2	Методы выделения признаков	47
3.6.3	Отбор признаков	50
3.6.4	Результаты	51
	Список использованных источников	52
A	Приложение части Аудио	53

Б Приложение части Текст	54
В Приложение части Видео	55
Г Приложение общей части	56

Введение

Задача распознавания эмоций в разговорах (далее - ERC) является наиболее значимой в развитии симпатического взаимодействия человека и машины. В разговорных видеороликах эмоции могут присутствовать в нескольких модальностях - аудио, видео и транскрипции речи. Однако данная задача распознавания эмоций всегда считалась сложновыполнимой из-за свойственных этим модальностям характеристик.

Далее мы че то тут должны предложить, кратко че юзаем, но пока я ничего не предлагаю и вообще ничего не знаю, потом напишу)

1 Изучение датасета MELD

Мультимодальный набор данных Emotion Lines был создан путем улучшения и расширения набора данных Emotion Lines. MELD содержит те же экземпляры диалогов, что и в Emotion Lines, но он также включает в себя аудио и визуальные модальности наряду с текстом. MELD содержит более 1400 диалогов и 13000 высказываний из сериала "Друзья".

Статистика датасета:

Статистика	Тренировочная	Валидационная	Тестовая
Модальности	{а, в, т}	{а, в, т}	{а, в, т}
# уникальных слов	5783	1486	2529
Средняя длина фраз	8.63	8.61	8.90
Макс длина фраз	72	42	50
Сред. # эмоций диалог	3.30	3.35	3.24
# диалогов	1038	114	280
# фраз	9989	1108	2610
# говорящих	260	47	100
# смен эмоций	5358	605	1353
Сред. длина фраз	3.16s	3.14s	3.12s

Распределение эмоций в датасете:

Эмоции	Тренировочная	Валидационная	Тестовая
Злость	1109	153	345
Отвращение	271	22	68
Страх	268	40	50
Радость	1743	163	402
Нейтральное	4710	469	1256
Грусть	683	111	208
Удивление	1205	150	281

Далее будут проанализированы три модальности - составляющие этого датасета.

1.1 Анализ части Audio

Изначально аудио данных в данном датасете не имеется. Это означает, что сперва нам предстоит их выделить.

1.1.1 Изменения в датасете

В ходе работы было неоднократно замечено наличие видеоданных, которые не совпадают по длине фраз (фраза могла длиться две секунды, а сам видеоролик минут 5). Такие данные были переделаны по возможности.

К тому же, было найдено несоответствие между csv файлами и кол-вом материала. Ролики, индексы которых не удалось найти в "текстовом" варианте, были удалены.

В названии директории *test/output_repeated_splits_test* repeated присутствует не зря. В тестовой выборке содержались "повторы" видео, которые были улучшены (обрезаны в точь по фразе). Старые версии были заменены на новые.

Весь код и подробные действия Вы можете просмотреть в Приложение части Аудио^[1].

1.1.2 Выделение аудиозаписей из видео

Как и было сказано выше, аудио часть отсутствовала в данном датасете, однако сами видео были со звуком. Поэтому было решено выделить аудио из откорректированного датасета с помощью модуля moviepy. Аудио были названы в том же стиле, что и видео вида diaX_uttY.mp3. Для подробностей обратитесь к тому же пункту Приложение части Аудио^[1]. Аудио сет вы можете скачать по ссылке Приложение части Аудио^[2].

1.1.3 Анализ

Определения, которые могут понадобиться при прочтении этой части:

а) Тональность - это закрепление положения музыкального лада за определёнными по высоте звучания музыкальными тонами.

Например - разговор на «повышенных тонах» (крича, человек поменял «тональность»).

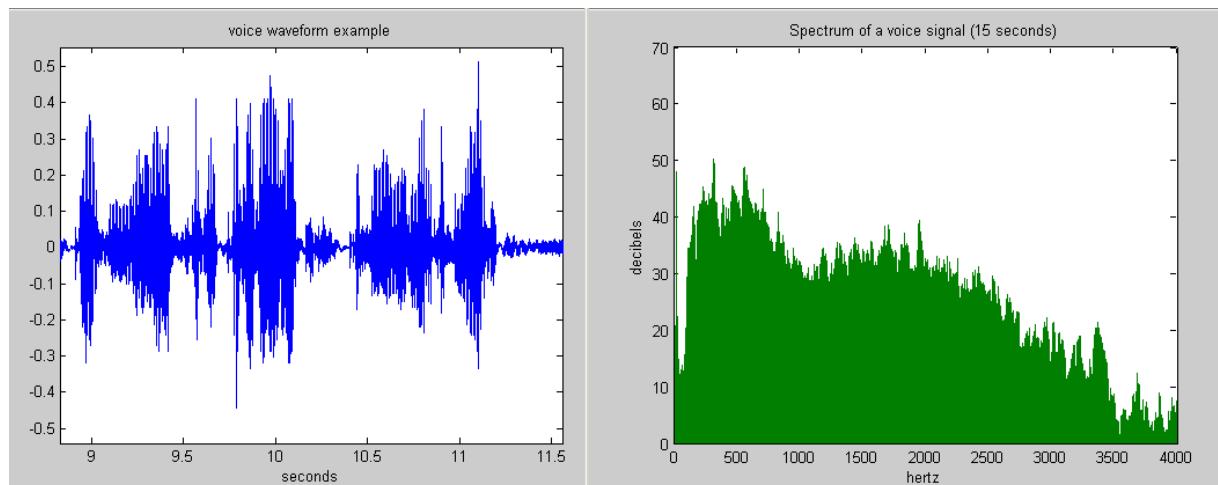
б) Лад – система взаимоотношений звуков, определяющая звукоряд.

Например - мажор и минор.

в) Тембр - яркость звука, его индивидуальность, передаваемая во время звукопроизношения.

г) Спектр - среднее статистическое значение определенного сигнала или типа сигнала (включая шум), проанализированное с точки зрения его частотного содержания.

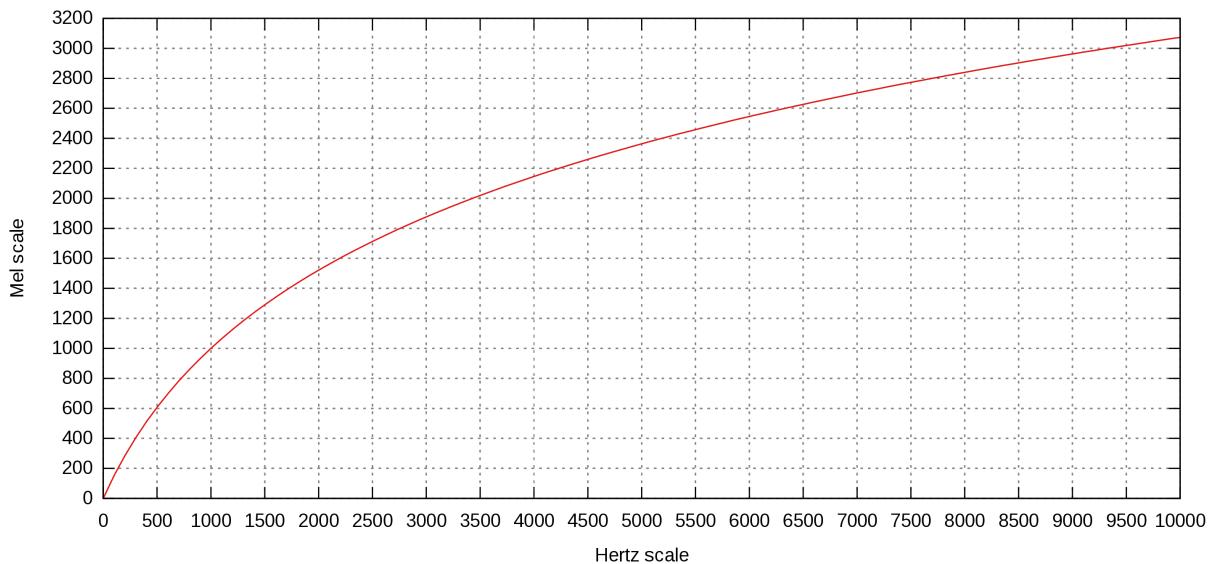
Например форма звуковой волны с течением времени (слева) имеет широкий спектр мощности звука (справа):



д) Частота - количество вхождений повторных событий в единицу времени.

е) Мел — психофизическая единица высоты звука, применяется главным образом в музыкальной акустике. Название происходит от слова «мелодия».

График зависимости высоты звука в мелах от частоты колебаний:



Для анализа я решила взять самую длинную запись от каждой выборки и использовала следующие методы:

- а) Спектограмма;
- б) Выделение гармонических и ударных сигналов;
- в) Мел-спектограммы;
- г) Мел-кепстральные коэффициенты;
- д) Изображение спектрального центроида вместе с формой волны;
- е) Спектограмма логарифма частоты.

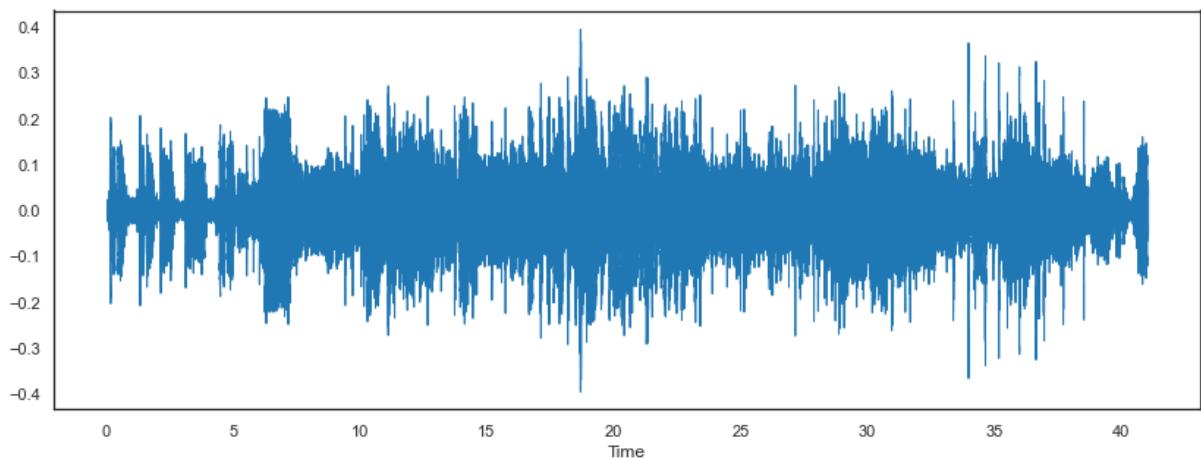
Рассмотрим подробнее аудио из обучающей выборки, остальные результаты вы можете наблюдать в секции "Статистика" Приложение части Аудио^[1].

1.1.3.1 Краткий анализ длительности аудио в модулях

Модуль	# аудио	Средняя длина	Макс длина	Мин длина
Обучающая	9989	3.2	41.1	0.13
Валидационная	1108	3.17	28.6	0.13
Тестовая	2610	3.17	16.8	0.18

1.1.3.2 Аудио максимальной длины

- а) Визуализация звука:



По визуализации много об аудио не скажешь, поэтому посмотрим на другие методы.

6) Выделение гармонических и ударных сигналов:

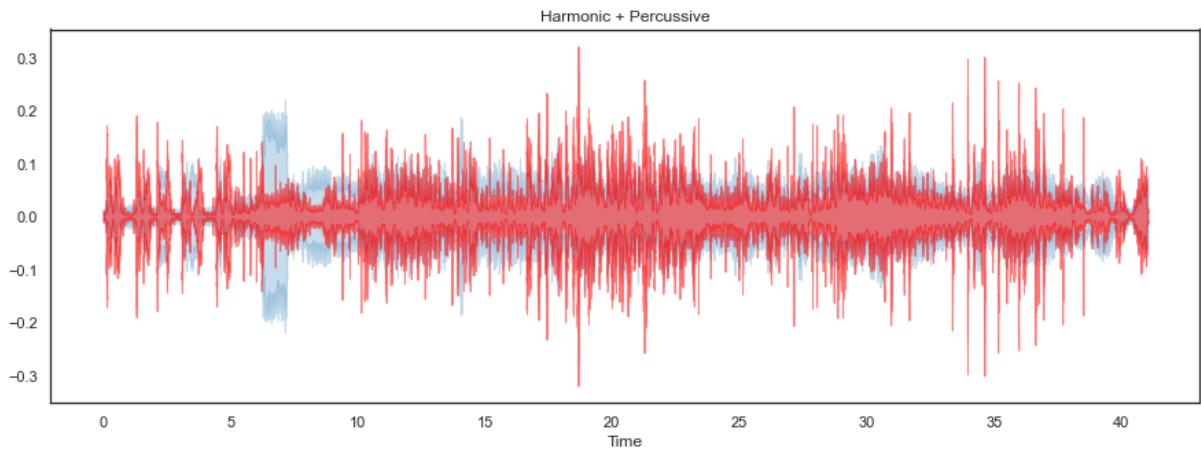
Сначала обратимся к теории, ведь важно понять что это и что эти сигналы могут дать. Подробнее в Приложение части Аудио^[3].

В общем и целом звук разделяется на две подкатегории - гармонический и ударный. Говоря более простым языком **гармонический звук** - высокотональный звук, распознаваемый человеком как мелодия или аккорды. Прототипом гармонического звука является акустическая реализация синусоиды - горизонтальная линия на спектрограмме.

Ударный звук - это столкновение, стук, хлопок или щелчок. Прототип ударного звука - это акустическая реализация импульса, которая соответствует вертикальной линии на спектрограмме.

Ударный звук - красный, гармонический - синий. Можно сравнить с визуализацией выше или прослушать разницу в блокноте.

Результат:

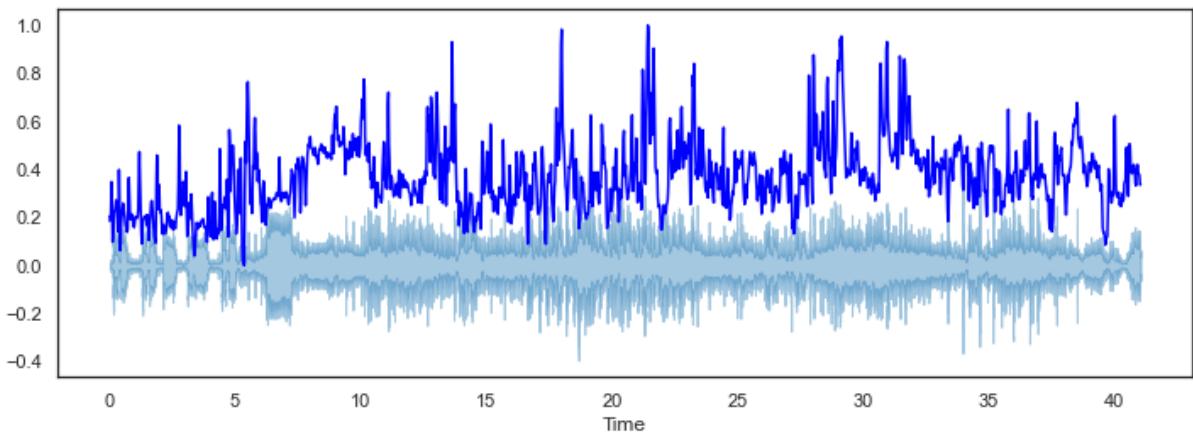


Зачем нам понадобилось выделять эти звуки? В дальнейшем мел-спектограммы данных аудиозаписей планируется использовать в обучении модели. В гармоническом случае выделяется музыка, которая обычно соответствует настроению, если мы говорим о сериалах. В других случаях она может мешать. В данном случае эти звуки будут иметь меньший приоритет по моему мнению. Ударные звуки наоборот выделяют речь среди других звуков при отсутствии ярких ударов, которые могут заглушить голос.

в) Изображение спектрального центроида вместе с формой волны.
Спектральный центроид указывает, на какой частоте сосредоточена энергия спектра или, другими словами, указывает, где расположен "центр масс" для звука. Смещение центроида в сторону высоких частот ощущается как повышение «яркости» тембра звучания.

Данный метод представлен по большей части как ознакомительный, чтобы посмотреть и изучить другие способы анализа и расширить свой кругозор. Дополнительно ниже указана форма волны, чтобы предположительно проследить "зависимость".

Результат:



г) Спектограмма.

Спектограмма - это визуальное представление спектра из частот сигнала, изменяющегося в зависимости от времени. Применяется для идентификации речи, анализа звуков животных, в различных областях музыки и т.д.

Общий формат - это график с двумя геометрическими измерениями: одна ось представляет время, а другая частоту; третье измерение, показывающее амплитуду конкретной частоты при конкретном t , отличается интенсивностью или цветом каждой точки изображения.

Результаты представлены в 1.1.

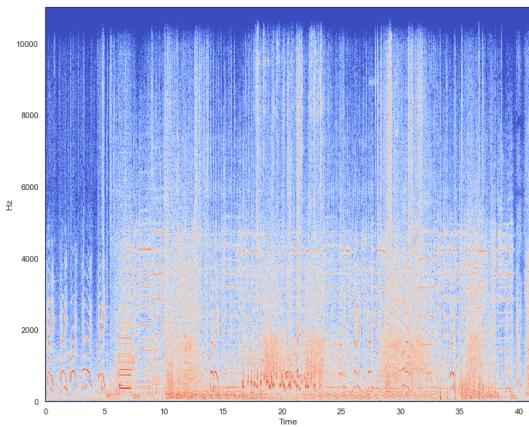


Рисунок 1.1 — Спектограмма

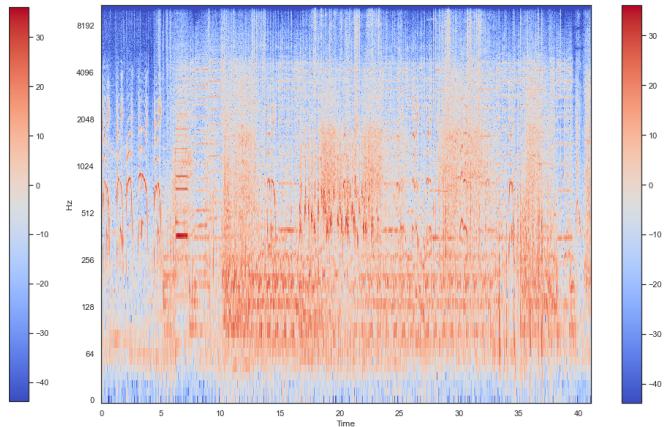


Рисунок 1.2 — log-Спектограмма

Заметно, что большая часть действий происходит в нижней части спектра, мы можем преобразовать ось частот в логарифмическую [1.2].

Для чего и почему это делается подробнее будет описано в пункте про мел-спектограммы.

Если вы прослушали аудио в блокноте или в файле, то могли заметить, что первые 5 секунд будто бы обособляются. В этот момент в аудио почти отсутствуют посторонние звуки, такие как музыка или удары, поэтому различимы моменты, когда один из героев зовет некую "Kathy" (причем похожие дуги можно заметить и на 15ых или 25ых секундах, например, что совпадает с выкриком). Далее заметно увеличение количества светлых (или красных) вертикальных и горизонтальных "полос" при появлении музыки и сопровождающих ударов.

д) Мел-спектограмма.

Мел-спектrogramma — это обычная спектrogramma, где частота выражена не в Гц, а в мелах. Переход к мелам осуществляется с помощью применения мел-фильтров к исходной спектrogramме. Мел-фильтры представляют из себя треугольные функции, равномерно распределенные на мел-шкале.

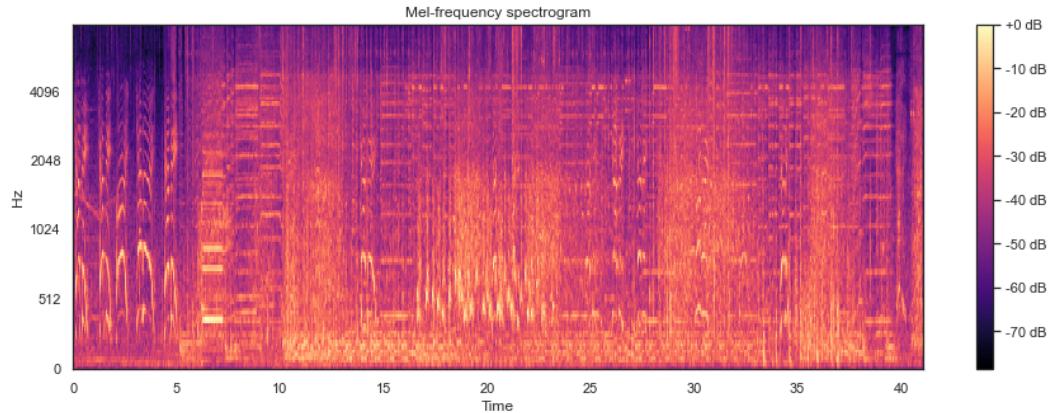
Эксперименты ученых показали, что человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких. То есть, если частота звука изменится со 100 Гц на 120 Гц, человек с очень высокой вероятностью заметит это изменение. А вот если частота изменится с 10000 Гц на 10020 Гц, это изменение мы вряд ли сможем уловить.

В связи с этим была введена новая единица измерения высоты звука — мел. Она основана на психо-физиологическом восприятии звука человеком, и логарифмически зависит от частоты.

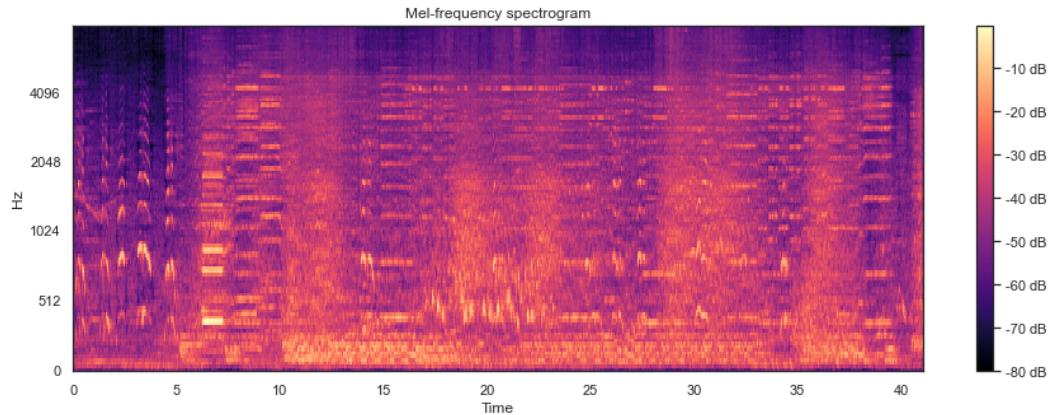
Подробнее с изображениями мел-фильтров можно почитать в Приложение части Аудио^[4].

Результат:

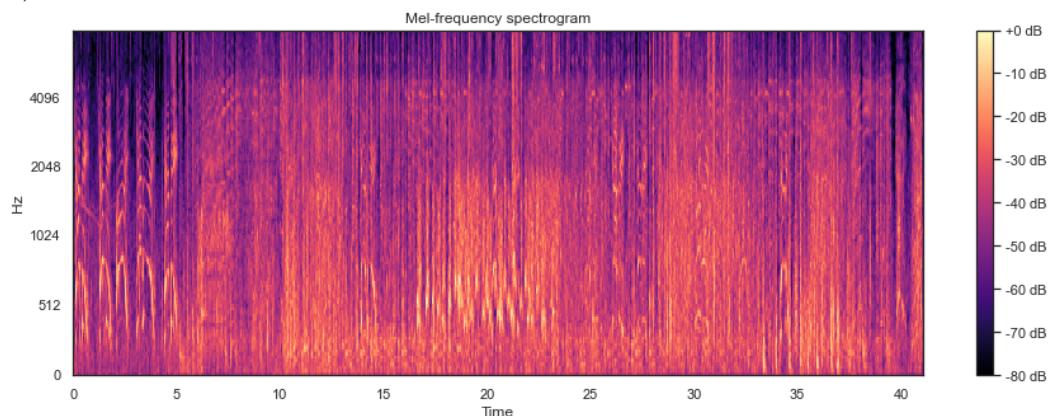
1) Мел-спектограмма основного аудио:



2) Мел-спектограмма гармоничного сигнала:



3) Мел-спектограмма ударного сигнала:



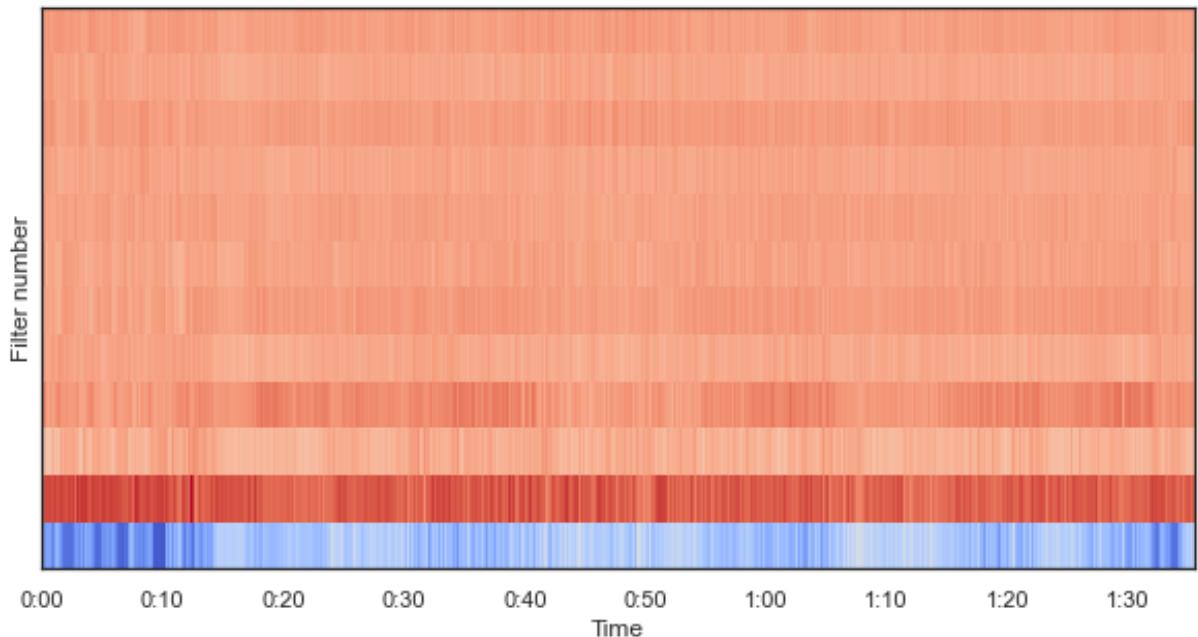
Как и говорилось ранее, в гармоническом звуке наблюдается преобладание горизонтальных полос, в сравнении с основным аудио. В то время как в ударном вертикальные.

е) Мел-кепстральные коэффициенты.

Первый шаг в анализе речевых данных – это выделение признаков, которые являются ”хорошими” для идентификации лингвистического содержания и отбрасыванием всех остальных признаков, отвечающих за шум.

Подробнее о мел-кепстральных коэффициентах вы можете почитать в Приложение части Аудио^[5] или более подробно о реализации в Приложение части Аудио^[6].

Результат:



Однако чаще советуют работать с мел-спектрограммами, особенно в нашей задаче. Значение эмоций у мел-кепстральных коэффициентов занижено.

1.2 Анализ части Text

Текстовая модальность датасета представляет из себя доработанный датасет EmotionLines, поделенный на обучающую (9989 записей, или 73%), валидационную (1109, или 8%) и тестовую выборки (2610, или 19%). Данные разбиты на следующие признаки:

1. Sr No. (порядковый номер высказывания, в основном для ссылки на высказывания в случае нескольких копий в выборке)
2. Utterance (высказывание из EmotionLines в формате строки)
3. Speaker (имя персонажа, которому принадлежит высказывание)
4. Emotion (эмоция высказывания)
5. Sentiment (общее настроение высказывания)
6. Dialogue_ID (номер диалога в выборке)
7. Utterance_ID (номер конкретного высказывания в диалоге)
8. Season
9. Episode
10. StartTime (время начала высказывания в эпизоде, в формате 'hh:mm:ss,ms')
11. EndTime (время окончания высказывания)

Каждое высказывание классифицировано одной из 6 базовых эмоций Экмана: радость, грусть, страх, злость, удивление, отвращение, также была добавлена метка “нейтральная”. Кроме деления на эти 7 классов эмоций есть более грубое деление высказываний на негативные (к ним относятся гнев, отвращение, страх, грусть), позитивные (радость) и нейтральные. Записи, классифицированные как удивление, могут быть с негативным или с позитивным настроением (рис. 1.3). Таким образом, удивление — пример сложной эмоции, которая в разных случаях может быть отнесена к негативным или позитивным эмоциям.

Распределение классов в датасете ожидаемо является несбалансированным: преобладают нейтральные высказывания во всех трёх выборках (рис. 1.4, 1.5). В целом, распределение классов в обучающей, валидационной и тестовой выборках примерно одинаково.

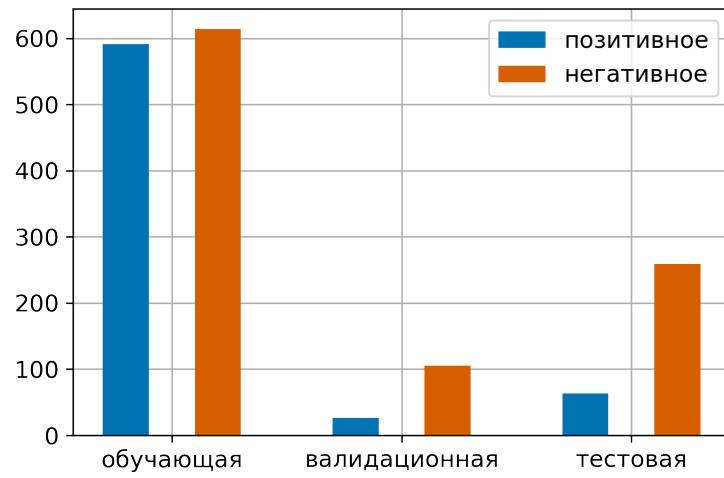


Рисунок 1.3



Рисунок 1.4



Рисунок 1.5 — Распределение классов в датасете

1.3 Анализ части Video

Размер картинки в видео: 720x1280

FPS(Фреймов в секунду): $\approx 24(23,976)$

Статистика	Тренировочная	Валидационная	Тестовая
# видео	9989	1108	2611
Средняя длина видео(с)	3.155	3.135	3.1198
Средняя длина видео(фреймы)	75.18	74.81	74.36
Макс длина видео(с)	41.06	28.55	16.75
Макс длина видео(фреймы)	984.0	684.0	401.0
Мин длина видео(с)	0.08	0.08	0.14
Мин длина видео(фреймы)	2.0	2.0	2.0

Распределение длины видео:

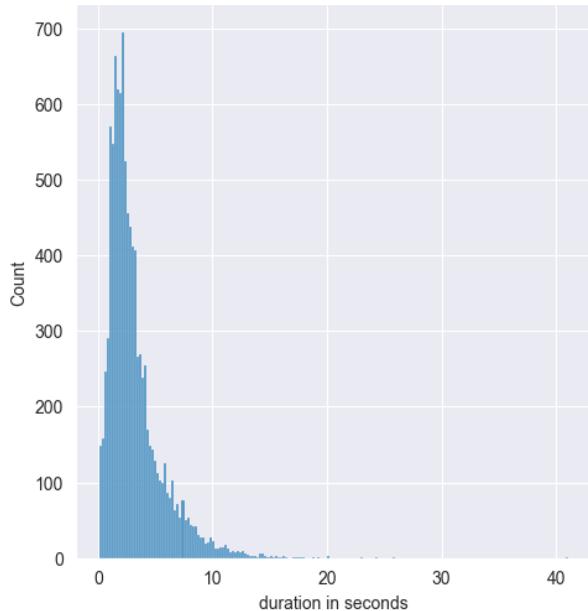


Рисунок 1.6 — В обучающей
выборке

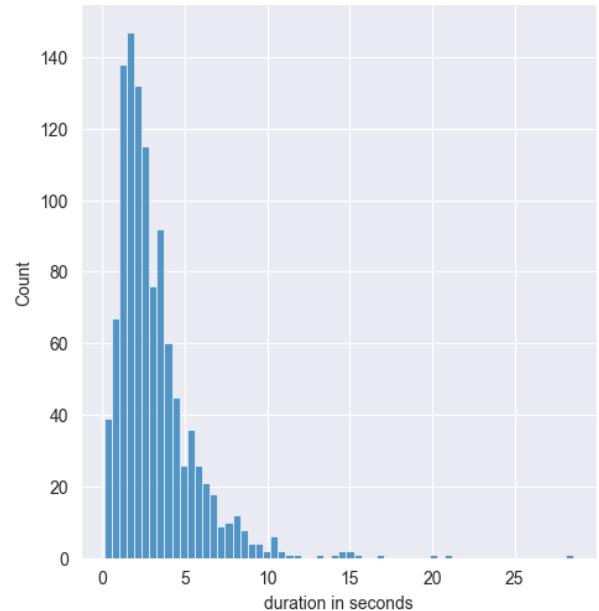


Рисунок 1.7 — В валидационной
выборке

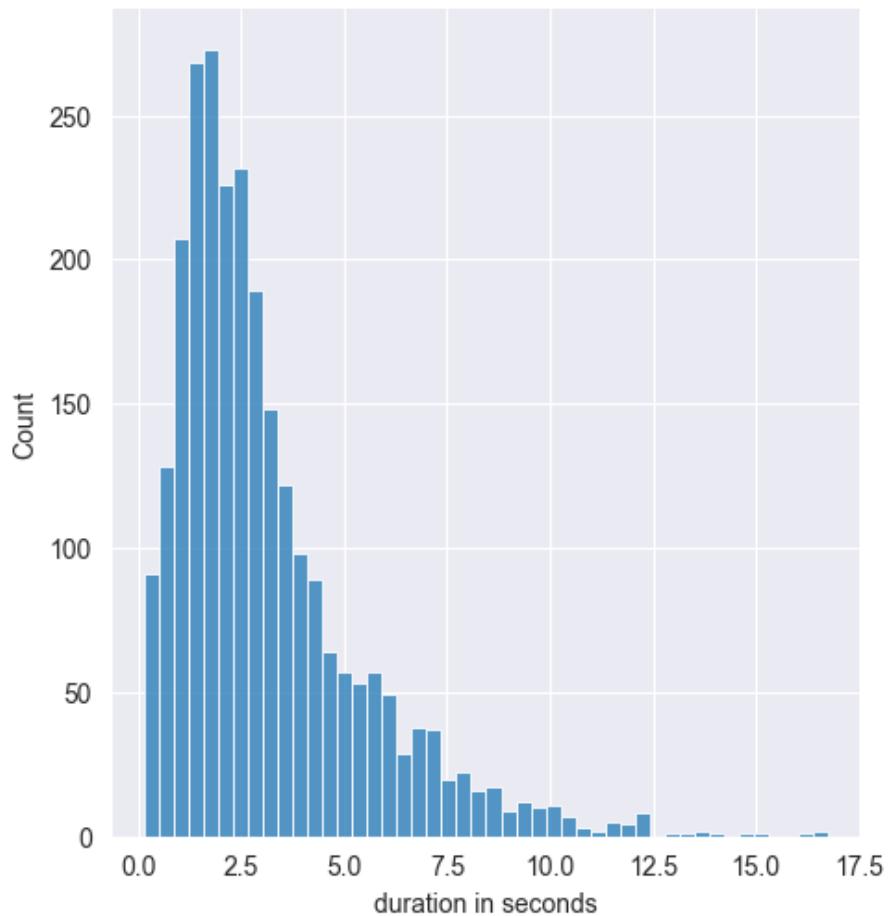


Рисунок 1.8 — В тестовой выборке

Среднее значение считалось с помощью функции *mean* из библиотеки *statistics*.

Для того, чтобы изучить видео, необходимо разделить его на кадры (фреймы), именно их количество и считается в таблице выше. Распределение их количества аналогично распределению длины видео, так как в датасете MELD находятся видео с фиксированным fps (frames per second или количеством фреймов в секунду).

Длину видео можно посчитать с помощью параметра *duration*, предварительно обработанного функцией *VideoFileClip* из библиотеки *moviepy.editor*.

Fps, разрешение, количество фреймов можно извлечь из видео с помощью библиотеки *openCV* (*cv2* в коде).

Для изучения эмоций по видео, нам будет необходимо выделить лицо на кадре. С этим поможет справиться класс *RetinaFace* из библиотеки *batch_face*.

Пример картинки с нарисованным дополнительно прямоугольником по данным полученным детектором лиц.



2 Изучение моделей и задачи анализа эмоций

В данной части мы изучим несколько моделей классификаций эмоций, которые были нам предложены, а именно:

- а) классификации Экмана;
- б) классификации Изарда;
- в) модель Плутчика;
- г) arousal;
- д) valence.

А также рассмотрим неизученный нами ранее сентимент-анализ.

2.1 Классификация Экмана

Базовые эмоции – эмоции, которые присущи всем здоровым людям и которые одинаково проявляются у представителей самых разных культур, проживающих на разных континентах. Эмоции – общие для всех.

Экман считает, что существует всего семь базовых эмоций: гнев (злость), печаль (грусть), презрение, отвращение, страх, удивление, радость. В обычной жизни они встречаются как в чистом виде, так и в смешанном. Подробнее можно прочитать в общем приложении^[1].

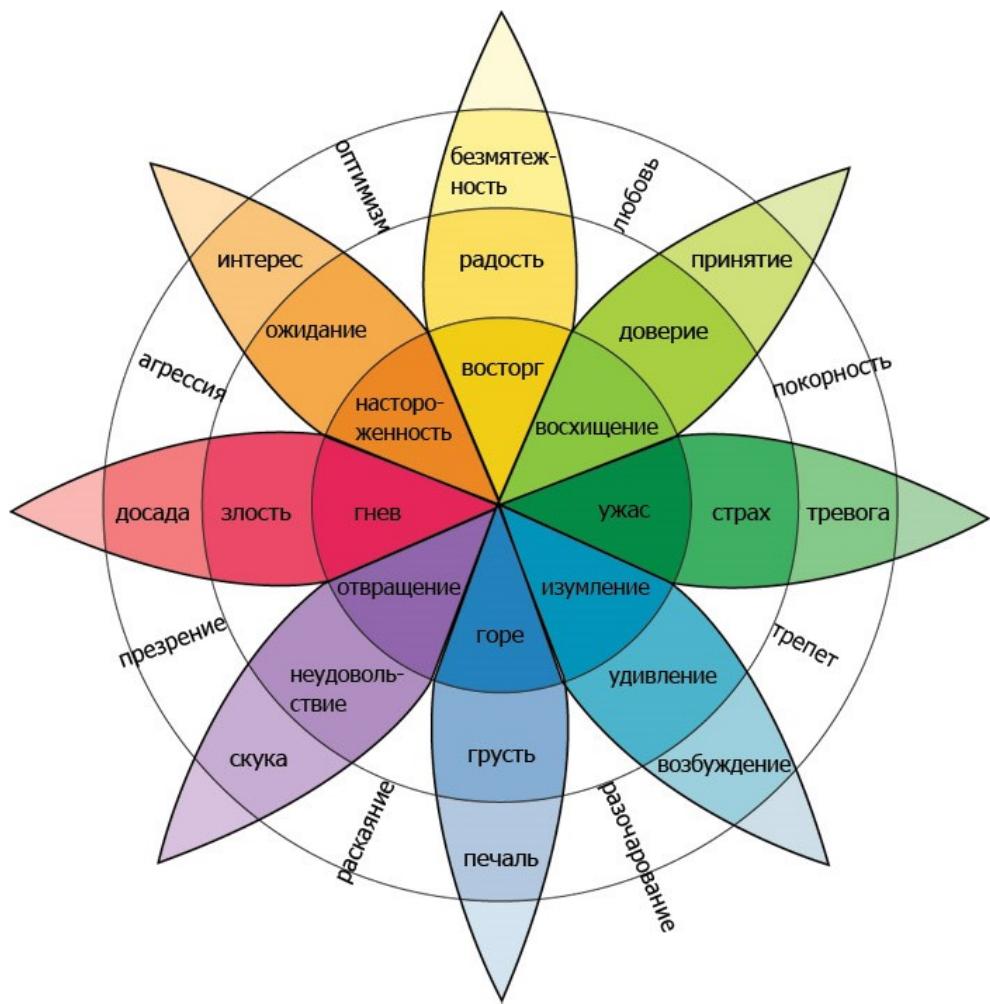
2.2 Классификация Изарда

По Изарду выделяют 11 фундаментальных (базовых) эмоций: Радость, Удивление, Печаль, Гнев, Отвращение, Презрение, Горе-страдание, Стыд, Интерес-волнение, Вина, Смущение.

Все остальные эмоциональные состояния, по версии Изарда, являются производными или составными, т.е. возникают на основе нескольких фундаментальных. Подробнее можно прочитать в общем приложении^[2].

2.3 Модель Плутчика

Колесо эмоций, созданное Плутчиком в 1980 году, представляет эмоции в виде некоего цветка с восемью лепестками. Каждый «лепесток» символизирует одну из прототипных эмоций, причем противоположные эмоции располагаются напротив друг друга. Получается, что полярные эмоции противостоят и уравновешивают одна другую:



Различных эмоций и чувств может быть огромное множество, но все они являются либо базовыми эмоциями, либо их смесью и производными. Так, восемь эмоций, расположенных вне колеса, в пространстве между лепестками, являются сочетанием эмоций смежных лепестков.

Колесо эмоций выполнено в цвете, что обладает большим смыслом. Каждая эмоция обладает своим цветом, а его насыщенность отражает интенсивность переживания, яркость эмоции. Получается, что в самом сердце схемы располагаются наиболее яркие по силе,

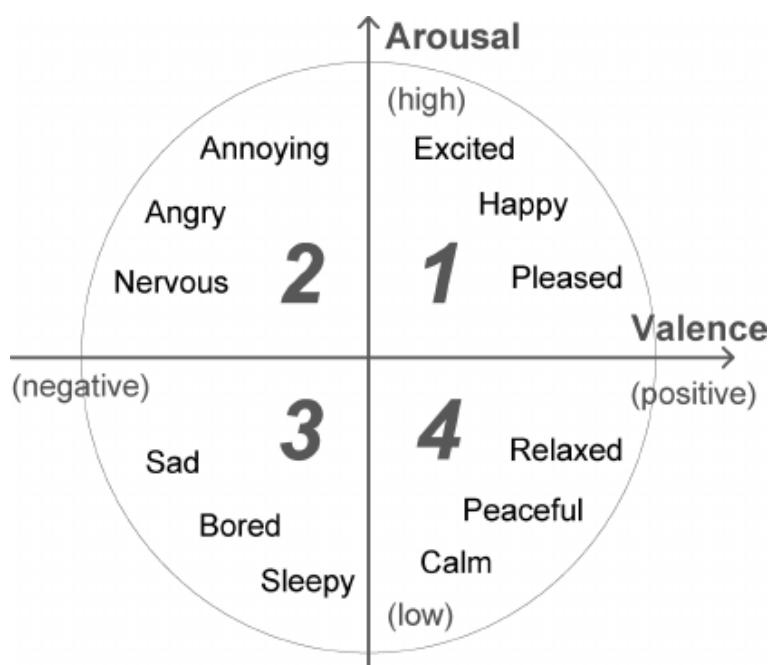
насыщенные эмоции, во втором ряду (втором сегменте лепестков) собственно основные (или базовые) эмоции, и далее – менее выраженный вариант каждой из эмоций.

2.4 Arousal и Valence

Valence - аффективное качество, заключающееся в субъективной привлекательности (положительная валентность) или непривлекательности (отрицательная валентность) для человека предметов, событий или ситуаций.

Arousal - это физиологическое и психологическое состояние пробуждения или органов чувств, стимулированных до точки восприятия.

Люди способны выражать широкий спектр эмоций. Считается, что их можно описать как комбинацию базовых эмоций (например, ностальгия – это что-то среднее между печалью и радостью). Но такой категориальный подход не всегда удобен, т.к. не позволяет количественно охарактеризовать силу эмоции. Поэтому наряду с дискретными моделями эмоций, был разработан ряд непрерывных. В модели Дж. Рассела водится двумерный базис, в котором каждая эмоция характеризуется знаком (valence) и интенсивностью (arousal).



2.5 Сентимент-анализ

Анализ тональности текста (или сентимент-анализ) – одна из задач, с которыми работают специалисты Data Science. С помощью такого анализа можно изучить массив сообщений и иных данных и определить, как они эмоционально окрашены – позитивно, негативно или нейтрально. Тональность – это эмоциональное отношение автора высказывания к некоторому объекту (объекту реального мира, событию, процессу или их свойствам/атрибутам), выраженное в тексте. Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента, называется лексической тональностью (или лексическим сентиментом). Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений) и правил их сочетания.

Исторически сложилось так, что традиционный подход к сентимент анализу представляет собой задачу классификации текста (части текста) на две-три категории (негативный, позитивный, нейтральный или просто: негативный или позитивный). Именно с такой задачи начал свое развитие анализ тональности: оценить сентимент оценочных отзывов по какой-либо тематике (кино, рестораны, электроника и пр.).

Тональность высказывания определяется тремя компонентами: субъектом тональности (кто высказал оценку), объектом тональности (о ком или о чём высказана оценка) и собственно тональной оценкой (как оценили). Примеры в статье [5].

2.5.1 Методы определения тональности

Существует два основных метода решения задачи автоматического определения тональности:

- a) Статистический метод. Для него нужны заранее размеченные по тональности коллекции (корпуса) текстов, на которых происходит

обучение модели, с помощью которой и происходит определение тональности текста или фразы;

б) Метод, основанный на словарях и правилах. Для этого заранее составляются словари позитивных и негативных слов и выражений. Этот метод может использовать как списки шаблонов, так и правила соединения тональной лексики внутри предложения, основанные на грамматическом и синтаксическом разборе.

Кроме того, иногда используют смешанный метод (комбинацию первого и второго подходов).

При статистическом подходе для решения задачи общей классификации текстов на классы тональности широко используют метод опорных векторов (SVM), Байесовы модели, различного рода регрессии.

3 Анализирование статей

В данной главе будет рассмотрен ряд статей для анализа методов с использованием датасета MELD задачи распознавания эмоций. В итоге данный анализ поспособствует в отборе методов для построения лучшей модели.

3.1 M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation

Мультимодальная сеть слияния для распознавания эмоций в разговоре [1] - исследование, в котором предлагается мультимодальная сеть слияния M2FNet, которая извлекает относящиеся к эмоциям функции из визуальной, звуковой и текстовой модальности.

В данной сети используется многоступенчатый механизм слияния внимания (attention fusion mechanism), основанный на сочетании богатых эмоциями скрытых представлений входных данных. Авторы представляют новый feature extractor (соковыжималка признаков) для извлечения скрытых признаков из аудио и видео модальности.

Предлагаемый экстрактор обучается с помощью новой адаптивной функции потери триплетов (triplet loss function) на основе полей для извлечения релевантных (относящихся, связанных) эмоциям признаков из аудио и визуальных данных.

Также авторы приводят сравнение результатов работы метода M2FNet и работы других способов решения задачи распознавания эмоций путем взвешенного среднего значения F1-score для двух датасетов - искомого MELD и IEMOCAP:

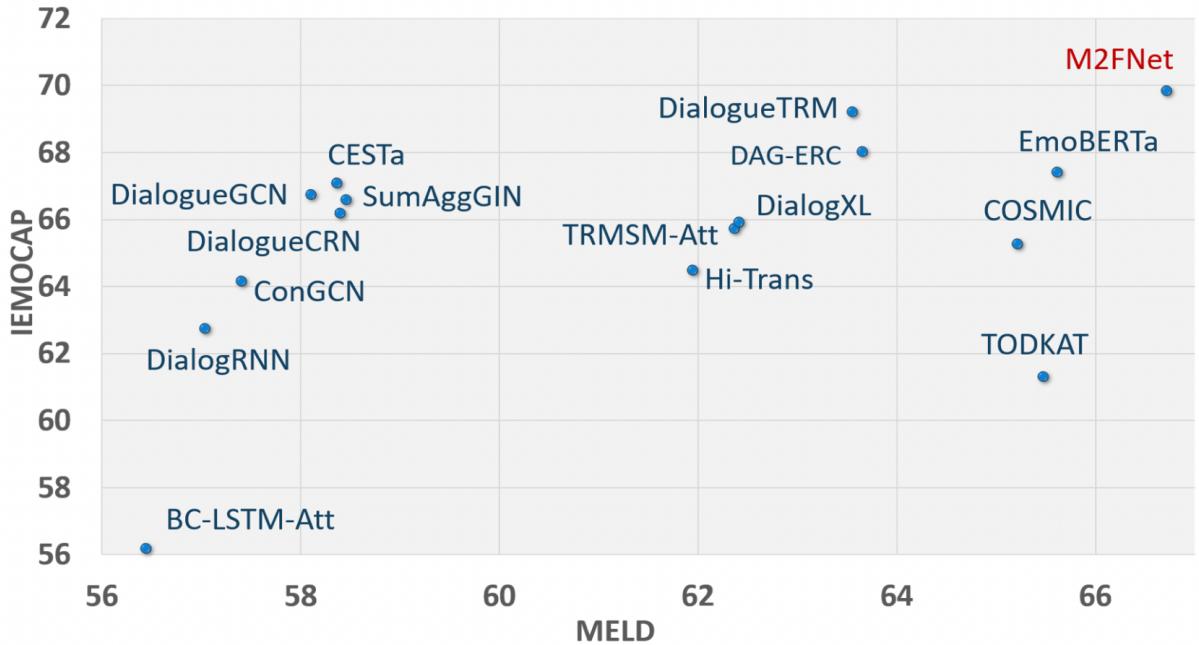


Рисунок 3.1 — Анализ данных MELD и IEMOCAP в терминах взвешенного среднего значения F1

3.1.1 Структура

Здесь будут кратко изложены идеи авторов в решении данной задачи.

3.1.1.1 Multi-modal Fusion Network: M2FNet

Предложенная авторами сеть спроектирована на основе двух уровней извлечения признаков:

- Извлечение признаков уровня высказывания;
- Извлечение объектов на уровне диалога.

Первоначально признаки извлекаются на уровне модуля высказываний независимо. Затем, в сети извлечения диалогового уровня, модель учится предсказывать правильную эмоцию для каждого высказывания, используя контекстуальную информацию из диалога в целом.

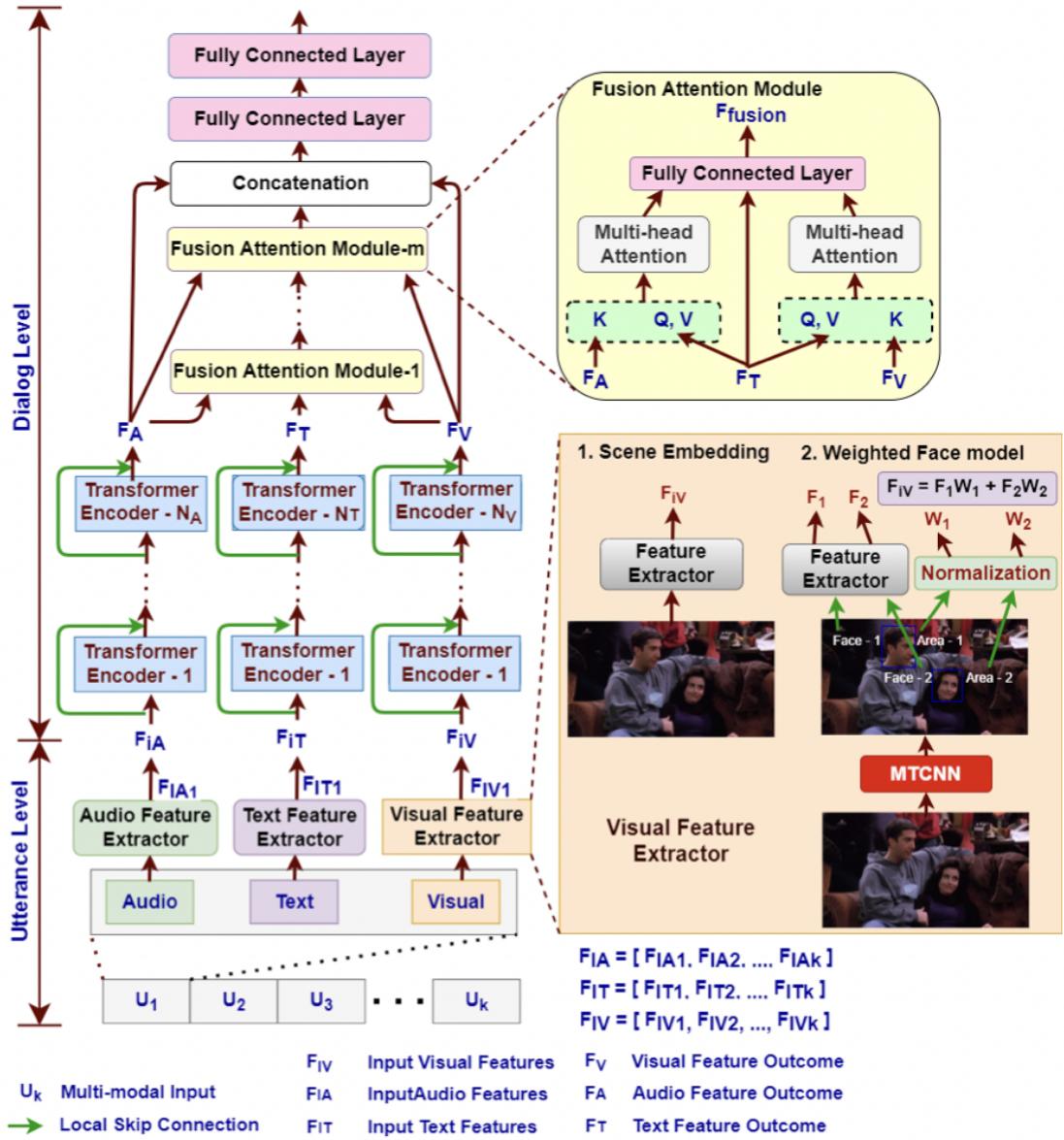


Рисунок 3.2 — M2FNet

3.1.1.2 Извлечение признаков на уровне высказываний

Текст: Чтобы обеспечить глубокое взаимодействие между высказываниями, данные о текстовой модальности передаются через Text Feature Extractor, в котором используется модифицированная модель RoBERTa (ϕ_M -RoBERTa).

Транскрипция каждого высказывания сопровождается текстом предыдущего и следующего высказываний, разделенных символом $< S >$.

Аудио: Первоначально аудио преобразуется в 2D Mel-спектрограмму в формате RGB, а затем передается через модель извлечения признаков. Здесь аудиосигнал сначала обрабатывается с помощью различных методов усиления, таких как time warping и Additive White Gaussian Noise. Затем дополненные сигналы преобразуются в соответствующие Mel-спектрограммы. Для её вычисления используется кратковременное преобразование Фурье с длиной кадра 400 отсчетов (25 мс) и длиной окна в 160 отсчетов (10 мс). Авторы также используют 128 блоков Mel-фильтров для генерации Mel-спектрограммы.

Предлагаемый экстрактор принимает Mel-спектрограммы в качестве входных данных и генерирует соответствующие вложения признаков.

Видео: Чтобы извлечь из визуального сигнала богатые эмоционально значимые характеристики, предлагается двойная сеть, которая использует не только выражение человеческого лица, но и контекстную информацию совместным и стимулирующим образом.

Для обеих задач используется модель извлечения (описанная в пункте 3.1.1.3) и тренируется на базе данных CASIA webface для извлечения более глубоких признаков из визуального изображения (подробнее о шагах двойной сети читать в приведенной статье п.3.2.1 Video часть).

3.1.1.3 Модуль извлечения признаков

Предлагаемый экстрактор разработан на основе триплетной сети, чтобы усилить важность функции триплетных потерь. Первоначально были сгенерированы якорные, положительные и отрицательные образцы для аудио и видео модальностей. Затем эти образцы передаются через сеть кодировщика, за которой следует модуль проектора. Здесь авторы используют стандартную ResNet18 в качестве основы сети кодировщика, в то время как проектор содержит линейный полностью подключенный слой, который проецирует содержимое сети кодировщика на желаемые представления, состоящие из N представлений размерности d.

Предлагаемая модель экстрактора обучается с использованием взвешенной комбинации трех функций потерь, т.е. функций потери адаптивного запаса в триплете (т.е. L_{AMT}), потери ковариации (т.е. L_{Cov}) и потери дисперсии (т.е. L_{Var}). Математически это выражается следующим образом:

$$L_{FE} = \lambda_1 L_{AMT} + \lambda_2 L_{Cov} + \lambda_3 L_{Var},$$

где λ_1 , λ_2 и λ_3 - весовые коэффициенты, которые управляют распределением различных функций потерь.

Подробнее про функции потерь читайте в приведенной статье п.3.3.

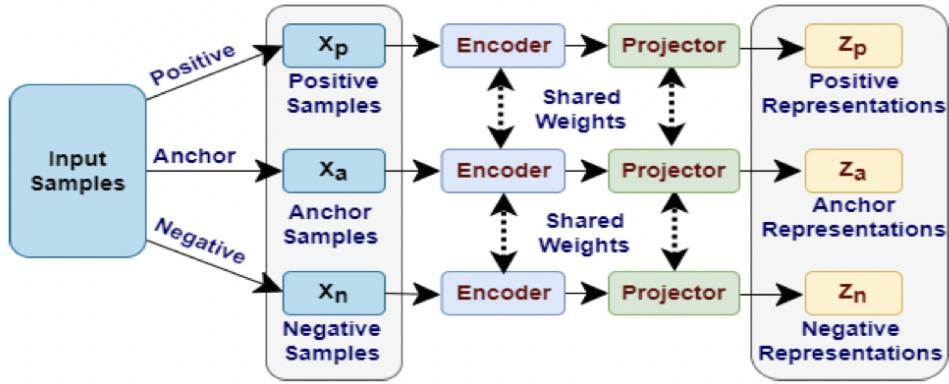


Рисунок 3.3 — Модуль извлечения признаков

Name of Model	MELD	
	Accuracy	Weighted Average F1
BC-LSTM-Att [21]	57.50	56.44
DialogRNN [19]	59.54	57.03
ConGCN [36]	—	57.40
Xie et al. [34]	65.00	64.00
DialogueTRM [20]	65.66	63.55
M2FNet	67.85	66.71

Рисунок 3.4 — Результаты

3.1.2 Результаты

В 3.4 представлены результаты, которых добились авторы данной статьи. Другие строки в таблице - показатели средневзвешенной оценки F1 и точности у современных методов, основанные на мультимодальности.

3.2 DialogueTRM: Exploring Multi-Modal Emotional Dynamics in a Conversation

В данной статье [2] авторы расширяют концепцию динамики эмоций до мультимодальных установок, которые учитывают внутримодальную и интермодальную динамику эмоций.

Динамика внутримодальных эмоций - это эмоциональные факторы, которые модальность получила от самой же себя во время диалога. Данная задача требует временного моделирования в каждой модальности. **Динамика интермодальных эмоций** - это еще одни эмоциональные факторы, которые модальность получала от других модальностей при каждом диалоге. Это задача требует пространственного моделирования во всех модальностях.

Взаимодействие между внутримодальной и интермодальной динамикой эмоций приводит к окончательным эмоциональным прогнозам.

Для того, чтобы распознавать эмоции по тексту, нужно знать контекст. Однако, эмоции отчетливее могут проскакивать на лице (видео) или в голосе (аудио) во время разговора, поэтому моделирование внутримодальной динамики эмоций должно удовлетворять контекстным предпочтениям различных модальностей.

В этой статье авторы предлагают диалоговый трансформатор, который моделирует внутримодальную и интермодальную динамику эмоций одновременно. Для внутримодальной динамики эмоций трансформаторы для временного моделирования упрощаются. Для интермодальной динамики эмоций разработано многоуровневое интерактивное объединение. Благодаря включению внутримодальной и интермодальной динамики эмоций, DialogueTRM обеспечивает более точные эмоциональные прогнозы.

3.2.1 Модель

Здесь будут кратко изложены идеи авторов в решении данной задачи.

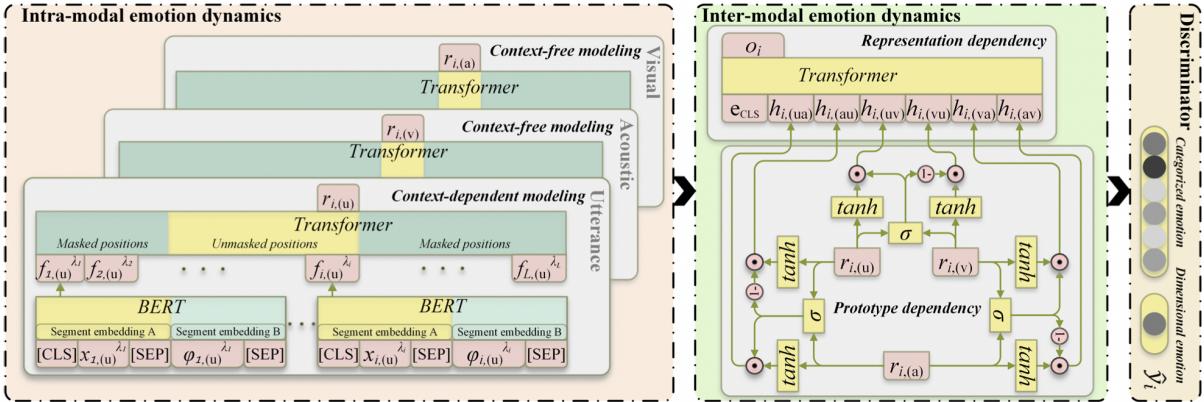


Рисунок 3.5 — Архитектура модели

3.2.1.1 Динамика внутримодальных эмоций

Динамика внутримодальных эмоций должна не только отражать временную зависимость, но и удовлетворять контекстным предпочтениям различных модальностей. Transformer может быть легко переключен на последовательную структуру для контекстно-зависимого моделирования или структуру прямой связи для контекстно-свободного моделирования. Таким образом, авторы используют трансформатор в качестве основы.

3.2.1.2 Контекстно-зависимые условия (текст)

Транскрипции предлагается моделировать в контекстно-зависимых условиях.

В BERT авторы объединяют сразу два процесса - CNN для кодирования высказываний, RNN для изучения зависимости между высказываниями. BERT кодирует каждое высказывание, получая последовательность необработанных лексических входных данных, содержащих информацию не только из самого высказывания, но и из контекста. Поскольку пары высказывание-контекст произносятся одним и тем же говорящим, информация, относящаяся к самостоятельной

зависимости, естественным образом сохраняется в выходных представлениях BERT.

Поскольку информация о говорящем сохраняется, межличностная зависимость может быть смоделирована с помощью взаимодействий в рамках признаков говорящего, полученных на последнем этапе. Вместо того, чтобы использовать сверточные сети графов для соединения этих функций авторы развернули multi-head attention в трансформаторе для вычисления взаимодействий.

3.2.1.3 Контекстно-свободные условия (аудио, видео)

Эмоции, выраженные в акустических и визуальных модальностях, лучше моделировать в контекстно-свободных условиях. Авторы используют openSMILE и 3D-CNN для извлечения акустических и визуальных характеристик. Оба типа признаков извлекаются из видеороликов на уровне высказываний.

3.2.1.4 Динамика интермодальных эмоций

Интермодальная динамика эмоций должна учитывать разностороннее взаимодействие признаков, чтобы объединить больше прогностических характеристик из разных модальностей. Зависимости прототипа и представления - это две степени детализации для объединения мультимодальных функций.

3.2.1.5 Прочее

В двух словах о других модулях:

a) Зависимость от прототипа

Зависимость прототипа может быть изучена с помощью позиционных взаимодействий между нейронами двух векторов одинаковой размерности. Авторы разработали мультимодальный 'шлюз' для изучения зависимости прототипа, присваивая нейронам в каждом векторе разные веса.

б) Зависимость от представления

Зависимость представления моделируется посредством взаимодействий в предложении из шести закрытых представлений, присваивающих один вес одному представлению. Взаимодействия рассчитываются с помощью глубоких слоев множественного внимания в Трансформаторе.

в) Дискриминатор

Дискриминатор использует двухслойный персепtron со скрытым слоем, активируемым \tanh .

3.2.2 Результаты

Models	MM	IEMOCAP		MELD	
		ACC	F1	ACC	F1
c-LSTM-U	×	56.3	56.1	-	56.7
AGHMN	×	63.5	63.5	59.5	57.5
DGCN	×	65.3	64.2	-	58.1
BiERU	×	66.1	64.7	-	60.8
DialogueTRM-U	×	<u>68.2</u>	<u>68.1</u>	<u>64.6</u>	<u>63.2</u>
c-LSTM-M	✓	59.8	59.0	-	-
CMN	✓	61.9	61.4	-	-
DRNN	✓	63.4	62.7	56.1	57.0
ICON	✓	64.0	63.5	-	-
DialogueTRM-M	✓	69.5	69.7	65.7	63.5

3.3 DialogGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation

В данном исследовании [5] авторы разработали модель на основе графовой нейронной сети, позволяющую лучше классифицировать эмоции в тексте, опираясь на контекст высказывания.

3.3.1 Методология

Одной из наиболее важных частей в распознавании эмоций является контекст. Авторы выделяют два основных типа контекста для распознавания эмоций в диалогах: *последовательный контекст* и *контекст на уровне говорящего*, и выделяют эти два вида контекста для каждого классифицируемого высказывания.

Контекст также учитывает динамику эмоций собеседников в разговоре. Она зависит от двух факторов: влияние других собеседников и влияние говорящего самого на себя. Влияние собеседников обусловлено тем, что говорящие, как правило, копируют своих собеседников, чтобы установить взаимопонимание в ходе разговора. Однако не все собеседники влияют друг на друга в равной степени. Влияние на самого себя возникает из-за того, что участники разговора будут придерживаться своего эмоционального состояния из-за своей эмоциональной инерции, если только собеседники не призывают к изменению.

3.3.2 Контекстно независимое извлечение признаков на уровне высказывания

Авторы использовали свёрточную нейронную сеть для извлечения признаков из высказываний. Используется один свёрточный слой, затем операция подвыборки (max-pooling) и полно связанный слой. Входными данными для этой сети являются 300-мерные предобученные векторы 840B GloVe. Использовались фильтры размера 3, 4 и 5 с 50 картами признаков в каждом. Затем делается max-pooling с размером окна 2 с последующей активацией ReLU. Затем признаки

объединяются и передаются в 100-мерный полно связанный слой, активация которого формирует представление высказывания. Эта сеть обучена на высказываниях с метками эмоций.

3.3.3 Модель

DialogueGCN состоит из трех основных компонентов — последовательного кодировщика контекста, кодировщика контекста на уровне говорящего и классификатора эмоций.

3.3.3.1 Последовательный кодировщик контекста

Диалог передаётся в управляющий рекуррентный блок (Gated recurrent unit, GRU), чтобы зафиксировать информацию о контексте. На этом этапе высказывания кодируются независимо от говорящего.

3.3.3.2 Кодировщик контекста на уровне говорящего

Представляет из себя ориентированный граф из последовательно закодированных высказываний. Разговор, содержащий N высказываний, представляется в виде ориентированного графа $G = (V, E, R, W)$ с вершинами $v_i \in V$, помеченными ребрами (отношениями) $r_{ij} \in E$, где $r \in R$ — тип отношения ребра между v_i и v_j , α_{ij} — вес помеченного ребра r_{ij} , $0 \leq \alpha_{ij} \leq 1$, $\alpha_{ij} \in W$, $i, j \in [1, 2, \dots, N]$.

Граф строится следующим образом. Каждое высказывание в разговоре представлено как вершина $v_i \in V$. Каждая вершина v_i инициализируется соответствующим закодированным последовательным кодировщиком вектором признаков g_i . Рёбра строятся между контекстуально зависимыми высказываниями внутри окна размерами $(10, 10)$ (т.е. рассматриваются только 10 ближайших прошлых высказываний и 10 ближайших следующих). Окно используется для ограничения числа рёбер. Веса рёбер устанавливаются с помощью модуля внимания. Функция внимания вычисляется таким образом, что для каждой вершины входящий набор рёбер имеет суммарный вес

равный 1. С учётом окна W веса вычисляются по формуле:

$$\alpha_{ij} = \text{softmax}(g_i^T W_e [g_{i-10}, \dots, g_{i+10}]), j = i - 10, \dots, i + 10.$$

Отношение r ребра r_{ij} устанавливается в зависимости от двух аспектов: участников диалога обеих вершин и относительной позиции высказываний u_i и u_j в диалоге.

Далее следует преобразование признаков. На первом этапе новый вектор признаков h_i вычисляется для вершины v_i по формуле:

$$h_i^{(1)} = \sigma \left(\sum_{r \in R} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r^{(1)} g_j + \alpha_{ii} W_0^{(1)} g_i \right), i = 1, 2, \dots, N,$$

где N_i^r - соседние индексы вершины i при отношении r ; $c_{i,r}$ - константа нормализации для конкретной задачи (например, она может быть равна $|N_i^r|$); σ - функция активации (например, ReLU); $W_r^{(1)}, W_0^{(1)}$ - обучаемые параметры. Второй этап преобразования применяется к выходным данным первого шага:

$$h_i^{(2)} = \sigma \left(\sum_{j \in N_i^r} W^{(2)} h_j^{(1)} + W_0^{(2)} h_i^{(1)} \right), i = 1, 2, \dots, N,$$

где $W^{(2)}$ и $W_0^{(2)}$ — параметры этих преобразований, σ — функция активации. Этот стек преобразований накапливает нормализованную сумму локального соседства (признаков соседей), т. е. информацию о соседних высказываниях в графе.

3.3.3.3 Классификатор эмоций

Вектора g_i признаков, закодированных последовательным кодировщиком, и вектора $h_i^{(2)}$ признаков, закодированных кодировщиком на уровне говорящего, объединяются, и применяется механизм внимания:

$$h_i = [g_i, h_i^{(2)}],$$

$$\beta_i = \text{softmax} (h_i^T W_\beta [h_1, h_2, \dots, h_N]),$$

$$\tilde{h}_i = \beta_i [h_1, h_2, \dots, h_N]^T.$$

Наконец, высказывание классифицируется с использованием полносвязной сети:

$$l_i = \text{ReLU}(W_l \tilde{h}_i + b_l),$$

$$P_i = \text{softmax}(W_{smax} l_i + b_{smax}),$$

$$\hat{y}_i = \text{argmax}_k(P_i[k]).$$

Для обучения использовалась категориальная кросс-энтропия с l_2 -регуляризацией. Использовалась модификация алгоритма стохастического градиентного спуска Adam. Loss-функция:

$$L = -\frac{1}{\sum_{s=1}^N c(s)} \sum_{i=1}^N \sum_{j=1}^{c(i)} \log P_{i,j}[y_{i,j}] + \lambda \|\theta\|_2,$$

где N - число диалогов, $c(i)$ - число высказываний в диалоге i , $P_{i,j}$ - распределение вероятности меток эмоций для высказывания j в диалоге i , λ - коэффициент регуляризации, θ - обучаемые параметры модели.

3.3.4 Применение датасета MELD

Авторы применяли свою модель на нескольких датасетах: IEMOCAP, AVEC и MELD. Авторы обнаружили, что эмоции в MELD сложнее классифицировать, т.к. высказывания в MELD намного короче и редко содержат выражения, специфичные для эмоций, что означает, что классификация эмоции сильно зависит от контекста. Также, средняя продолжительность разговора составляет 10 высказываний, причём во многих разговорах участвует более 5 человек.

Для сравнения эффективности модели DialogGCN авторы использовали другие модели (см. табл. 3.1). Модель показала результат 58.10% по метрике f1-score.[5]

Авторы предполагают, что увеличение размера привело бы к лучшему результату, но не проводили таких экспериментов из-за вычислительных ограничений. Также, DialogGCN чаще ошибается в классификации высказываний с метками “разочарование”, “злость” и “нейтральное” из-за тонкой разницы между разочарованием и злостью, а также ввиду того, что модель не использует аудио и видео

Методы	MELD
CNN	55.02
bc-LSTM	56.44
bc-LSTM+Att	56.70
DialogueRNN	57.03
DialogGCN	58.10

Таблица 3.1 — Сравнение f1-score для разных моделей

модальности (иначе, в таких высказываниях, например, высокий звук и хмурое выражение лица помогли бы классифицировать запись как “разочарование”).

3.4 COSMIC: COmmonSense knowledge for eMotion Identification in Conversations

Авторы этой статьи [6] в разработке своей модели классификации эмоций в диалогах опираются на так называемый здравый смысл (commonsense knowledge). Он помогает выявить такие элементы разговора, как избегание повторений, задавание вопросов, избегание ответов, не относящихся к делу и т.п. Все эти элементы воздействуют на такие аспекты диалога как беглость речи, заинтересованность, эмпатия. Таким образом, здравый смысл необходим для моделирования характера и течения диалога, а также динамики эмоций говорящих.

3.4.1 Методология

Фреймворк COSMIC состоит из трёх основных частей:

- a) Контекстно независимое извлечение признаков из предобученных преобразующих языковых моделей
- б) Извлечение признаков здравого смысла
- в) Включение знаний здравого смысла для лучшего представления контекста, использование их в классификации эмоций.

Для извлечения независимых от контекста признаков уровня высказывания авторы использовали модель RoBERTa. После настройки в неё передают вектора признаков, закодированные с помощью BPE (byte pair encoding) и со специальным токеном $[CLS]$ в начале, и извлекаются активации последних четырёх слоёв, соответствующих $[CLS]$ -токену. Эти четыре вектора затем усредняются для получения вектора независимых от контекста признаков высказывания размерностью 1024.

На этапе извлечения признаков здравого смысла используется модель COMET. COMET обучается на нескольких графах здравого смысла. Модель получает триплет $\{s, r, o\}$ из графа и учится генерировать объектную фразу o из объединения субъектной фразы s и фразы отношения r . Для выполнения задачи построения конструкций здравого смысла COMET обучается на ATOMIC - наборе

повседневных логических выводов вида «если-то», организованных в виде текстовых описаний. Таким образом авторам удаётся моделировать обстоятельства диалога (событие, персона, психические состояния) и причинно-следственные связи, которые являются важными элементами для понимания разговорного контекста.

3.4.2 Архитектура модели COSMIC

Введём обозначения. Диалог состоит из N высказываний u_1, u_2, \dots, u_N . В нём участвуют M различных спикеров p_1, p_2, \dots, p_M . Контекстно независимые векторы RoBERTa обозначаются x_t , $t = 1, 2, \dots, N$. Векторы, соответствующие намерению X (X - спикер), влиянию на X, реакции X, влиянию на остальных (слушающих) и реакции остальных, обозначаются $IS_{cs}(u_t), ES_{cs}(u_t), RS_{cs}(u_t), EL_{cs}(u_t), RL_{cs}(u_t)$ соответственно. Также формируются вектор состояния контекста c_t и вектор внимания a_t , которые являются общими для всех участников разговора. Внутреннее состояние, внешнее состояние и состояние намерения используются для моделирования различных психических состояний, действий и событий. Они представлены $q_{k,t}, r_{k,t}, i_{k,t}$ для каждого говорящего k . Внутреннее состояние и внешнее состояние можно вместе рассматривать как состояние говорящего. Затем из комбинации трёх состояний и непосредственно предшествующего эмоционального состояния моделируется текущее эмоциональное состояние e_t . Наконец, из e_t выводится класс эмоции.

Моделирование контекста и знаний здравого смысла осуществляется с использованием управляющих рекуррентных блоков (gated recurrent unit, GRU). Используется пять двунаправленных блоков ($GRU_C, GRU_Q, GRU_R, GRU_I$ и GRU_E) для моделирования состояния контекста, внутреннего состояния, внешнего состояния, состояния намерения и эмоционального состояния.

Все высказывания диалога классифицируются на основе e_t полносвязной нейронной сетью.

Иллюстрация работы COSMIC представлена на рис. 3.6. Более подробное описание модели можно найти в п. 3.4 статьи [6].

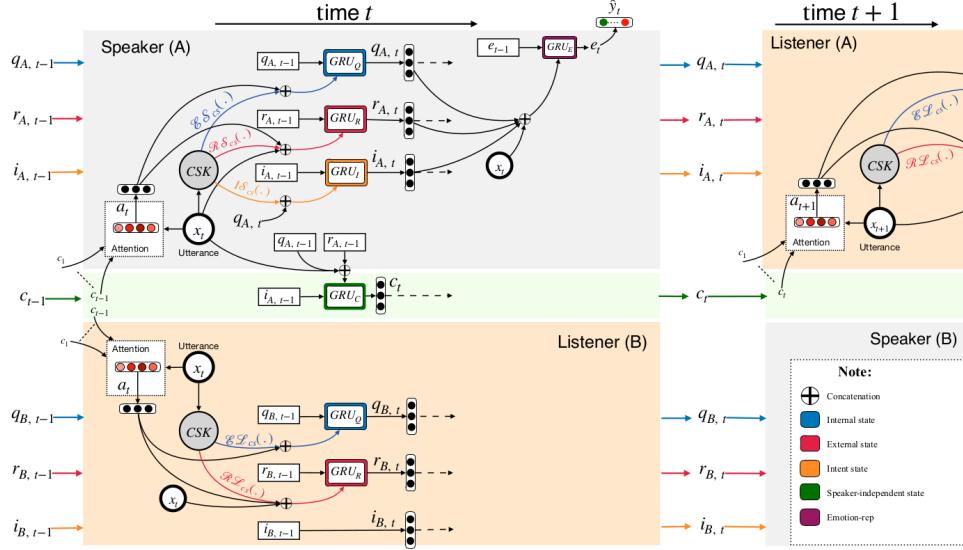


Рисунок 3.6 — Иллюстрация структуры COSMIC

Methods	IEMOCAP		DailyDialog		MELD		EmoryNLP	
	W-Avg F1		Macro F1	Micro F1	W-Avg F1 (3-cl)	W-Avg F1 (7-cl)	W-Avg F1 (3-cl)	W-Avg F1 (7-cl)
GloVe-based	CNN	52.04	36.87	50.32	64.25	55.02	38.05	32.59
	ICON	58.54	-	-	-	-	-	-
	KET	59.56	-	53.37	-	58.18	-	34.39
	ConGCN	-	-	-	-	57.40	-	-
	DialogueRNN	62.57	41.80	55.95	66.10	57.03	48.93	31.70
(Ro)BERT(a)-based	BERT DCR-Net	-	48.90	-	-	-	-	-
	BERT+MTL	-	-	-	-	61.90	-	35.92
	RoBERTa	54.55	48.20	55.16	72.12	62.02	55.28	37.29
	RoBERTa DialogueRNN	64.76	49.65	57.32	72.14	63.61	55.36	37.44
	COSMIC	65.28	51.05	58.48	73.20	65.21	56.51	38.11
w/o Speaker CSK		63.27	50.18	57.45	72.94	64.41	55.46	37.35
w/o Listener CSK		65.05	48.67	58.28	72.90	64.76	56.57	38.15
w/o Speaker, Listener CSK		63.05	48.68	56.16	72.62	64.28	55.34	37.10

Рисунок 3.7 — Сравнение результатов COSMIC и двухих моделей на различных датасетах

3.4.3 Применение датасета MELD, результаты

Авторы датасета отметили те же сложности в классификации высказываний диалогов в текстовом виде, что и авторы предыдущей рассмотренной статьи, а именно: высказывания часто очень короткие, высказывания нечасто содержат слова, характерные для эмоций, в каждом разговоре участвует более двух говорящих, каждый из которых произносит лишь небольшое количество реплик. Тем не менее, модель

COSMIC показывает более высокие значения взвешенной f1-score, чем другие рассмотренные авторами для сравнения модели (рис. 3.7).

3.5 UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition

UniMSE - мультимодальная структура (фреймворк) обмена знаниями о настроениях, которая выполняет задачи объединения (Unify) MSA и ERC. MSA - Multimodal Sentiment Analysis - мультимодальный анализ настроений, ERC - Emotion Recognition in Conversations - Распознавание эмоций в разговорах.

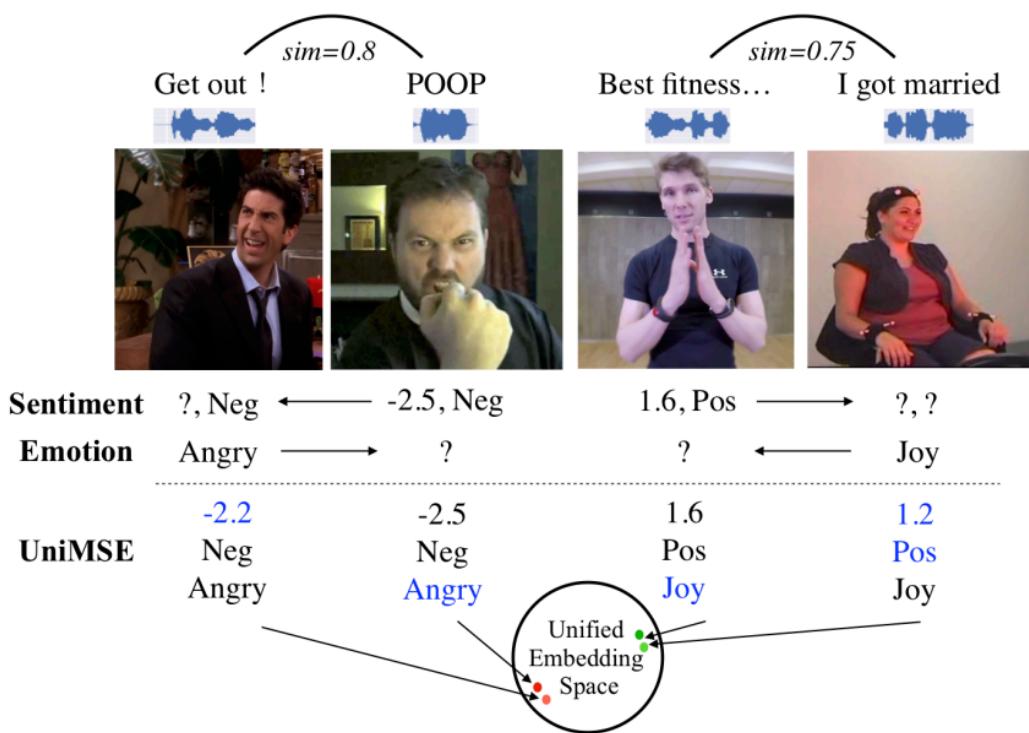


Рисунок 3.8 — Иллюстрация настроений и эмоций, разделяющих единое пространство встраивания. Внизу можно видеть, что на основе высокого *sim* между примерами делается вывод об одинаковой полярности чувств и об одинаковых эмоциях в высказывании.

3.5.1 Формулировка задачи:

Цель MSA: предсказать полярность произнесенной фразы/речи, а целью ERC является предсказывание конкретной эмоции. UniMSE

задается целью объединить решения этих задач в одно на основе взаимодополняемости чувств и эмоций.

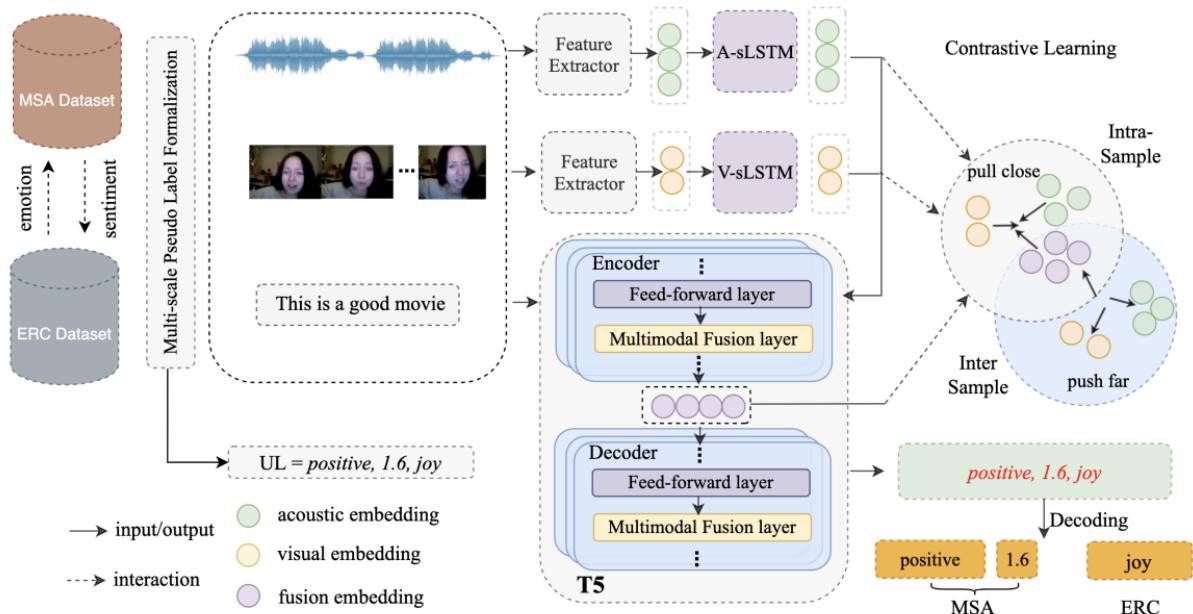


Рисунок 3.9 — UniMSE

Для работы UniMSE включает формализацию задачи, предварительно обученную модель для слияния модальностей и интермодальное контрастивное обучение. Во-первых, в автономном режиме обрабатываются метки задач MSA и ERC в формат универсальной метки (UL). После этого из датасетов извлекаются аудио и видео признаки, используя унифицированные экстракторы признаков. После получения данных они обрабатываются двумя LSTM (Long short-term memory; LSTM – особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям.). Для текстовой модальности авторы используют трансформер T5.

Аудио: Авторы обрабатывают необработанный акустический ввод в числовые последовательные векторы с помощью librosa, чтобы извлечь Mel-спектrogramму в качестве звуковых характеристик.

Видео: Для получения информации из видео авторы статьи извлекают фиксированные Т-кадры из каждого сегмента и используют

efficientNet, предварительно обученную (под наблюдением) на VGGface и наборе данных AFEW, чтобы получить характеристики видео.

Чтобы далее была возможность обрабатывать данные, каждое модальное представление визуальных и акустических данных обрабатывается до одинаковой длины с помощью одномерного временного сверточного слоя (1D temporal convolutional layer).

Для обработки данных используют внутримодальное сравнительное обучение (Inter-modality CL - Contrastive learning. На вход нейросети подаётся пара объектов и она определяет, похожи они или нет).

3.5.2 Результаты

Данный метод показал себя несколько лучше указанных ниже методов определения эмоций.

Method	MOSI					MOSEI					MELD		IEMOCAP	
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	ACC↑	WFI↑	ACC↑	WFI↑
LMF	0.917	0.695	33.20	-/82.5	-/82.4	0.623	0.700	48.00	-/82.0	-/82.1	61.15	58.30	56.50	56.49
TFN	0.901	0.698	34.90	-/80.8	-/80.7	0.593	0.677	50.20	-/82.5	-/82.1	60.70	57.74	55.02	55.13
MFM	0.877	0.706	35.40	-/81.7	-/81.6	0.568	0.703	51.30	-/84.4	-/84.3	60.80	57.80	61.24	61.60
MTAG	0.866	0.722	38.90	-/82.3	-/82.1	-	-	-	-	-	-	-	-	-
SPC	-	-	-	-/82.8	-/82.9	-	-	-	-/82.6	-/82.8	-	-	-	-
ICCN	0.862	0.714	39.00	-/83.0	-/83.0	0.565	0.704	51.60	-/84.2	-/84.2	-	-	64.00	63.50
MuLT	0.861	0.711	-	81.50/84.10	80.60/83.90	0.580	0.713	-	-/82.5	-/82.3	-	-	-	-
MISA	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.717	-	82.59/84.23	82.67/83.97	-	-	-	-
COGMEN	-	-	43.90	-/84.34	-	-	-	-	-	-	-	-	68.20	67.63
Self-MM	0.713	0.798	-	84.00/85.98	84.42/85.95	0.530	0.765	-	82.81/85.17	82.53/85.30	-	-	-	-
MAG-BERT	0.712	0.796	-	84.20/86.10	84.10/86.00	-	-	-	84.70/-	84.50/-	-	-	-	-
MMIM	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94	-	-	-	-
<i>DialogueGCN</i>	-	-	-	-	-	-	-	-	-	-	59.46	58.10	65.25	64.18
<i>DialogueCRN</i>	-	-	-	-	-	-	-	-	-	-	60.73	58.39	66.05	66.20
<i>DAG-ERC</i>	-	-	-	-	-	-	-	-	-	-	63.65	-	68.03	-
<i>ERMC-DisGCN</i>	-	-	-	-	-	-	-	-	-	-	64.22	-	64.10	-
<i>CoG-BART*</i>	-	-	-	-	-	-	-	-	-	-	64.81	-	66.18	-
<i>Psychological</i>	-	-	-	-	-	-	-	-	-	-	65.18	-	66.96	-
<i>COSMIC</i>	-	-	-	-	-	-	-	-	-	-	65.21	-	65.28	-
<i>TODKAT*</i>	-	-	-	-	-	-	-	-	-	-	65.47	-	61.33	-
<i>MMGCN</i>	-	-	-	-	-	-	-	-	-	-	58.65	-	66.22	-
<i>MM-DFN</i>	-	-	-	-	-	-	-	-	-	-	62.49	59.46	68.21	68.18
UniMSE	0.691	0.809	48.68	85.85/86.9	85.83/86.42	0.523	0.773	54.39	85.86/87.50	85.79/87.46	65.09	65.51	70.56	70.66

Рисунок 3.10 — Results

3.6 Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning

3.6.1 О чем статья:

В данной статье предлагается мультимодальный метод распознавания эмоций слияния для речевых выражений, основанный на глубоком обучении. Во-первых, соответствующие методы извлечения признаков настраиваются для разных одиночных модальностей. Среди них голос использует сверточную нейронную сеть с долговременной и краткосрочной памятью (CNN-LSTM), а выражение лица в видео использует сеть Inception-Res Net-v2 для извлечения признаков. Затем используется долговременная и краткосрочная память (LSTM) для фиксации корреляции как между различными модальностями, так и внутри модальностей. После процесса выбора признаков тестом хи-квадрат отдельные модальности объединяются для получения единого признака слияния. Наконец, признаки данных слияния, выдаваемые LSTM, используются в качестве входных данных классификатора LIBSVM для реализации окончательного распознавания эмоций.

3.6.2 Методы выделения признаков

Аудио: В своей работе авторы статьи пишут, что для выделения аудио-признаков из речи недостаточно использовать только LSTM, поэтому предлагаемый метод использует сеть слияния CNN и LSTM для изучения глубоких абстрактных акустических эмоциональных особенностей речи. Процесс извлечения признаков речи на основе CNN-LSTM показан на Рисунке 3.12.

В сети CNN-LSTM объединяются структурные характеристики модели сети CNN, и на ее основе добавляется слой LSTM. Используется сверточный слой для преобразования входных данных, а затем они вводятся в слой LSTM для повторного обучения. Часть между упомянутыми слоями состоит из двух слоев свертки и слоя максимального объединения. Параметры сверточного слоя такие же,

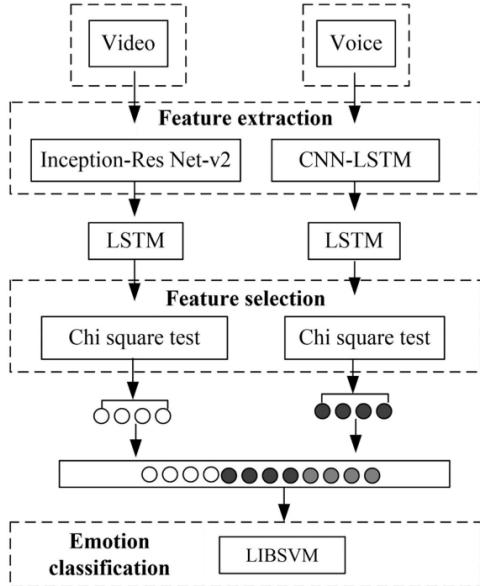


FIGURE 1 | Multimodal emotion recognition model based on deep learning.

Рисунок 3.11 – Общая архитектура модели

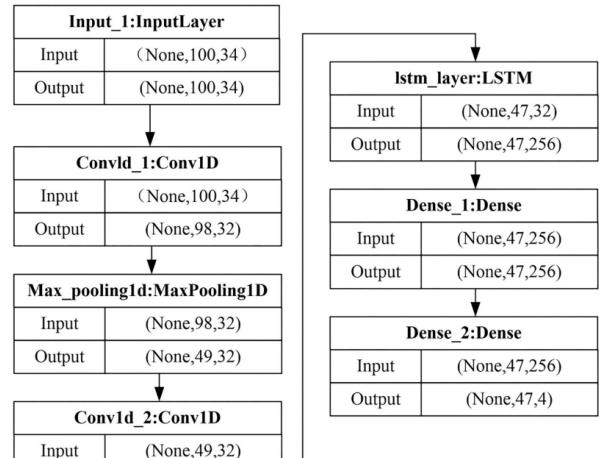


FIGURE 2 | Model flow graph of CNN-LSTM speech feature extraction.

Рисунок 3.12 – CNN-LSTM модель извлечения признаков из речи

как и в сети CNN, то есть 32×3 , а функция активации — Rectified Linear Unit (ReLU). Для выходного размера слоя LSTM установлено значение 256, а для параметра Recurrent Dropout установлено значение 0.2. Этот параметр представляет собой долю отброшенных линейных преобразований в рекурсивном состоянии. Выходным слоем всех моделей является плотный слой, а выходным измерением является количество типов классификации настроений.

Видео: Обычно объем данных в видео очень большой, поэтому для уменьшения количества следует найти на изображениях лица и оставить только их, обрезав лишнее. Предлагаемый метод выражает информацию об эмоциях в видео, извлекая признаки кратковременным преобразованием Фурье (STFT) и используя сеть Inception-Res Net-v2 для извлечения глубоких характеристик. Кратковременное преобразование Фурье является общим инструментом для обработки речевого сигнала. Оно определяет полезный класс распределения времени и частоты, который специфицирует комплексную амплитуду любого сигнала, изменяющегося со временем и частотой. Фактически процесс вычисления кратковременного преобразования Фурье

заключается в разделении более длинного временного сигнала на более короткие отрезки той же длины и вычислении преобразования Фурье (спектра Фурье) на каждом таком коротком отрезке.

Inception-Res Net-v2 — это CNN, созданная Google в 2016 году путем внедрения модели остаточной сети (ResNet) на основе модели Inception. Обладает точным эффектом распознавания очень похожих объектов. Эта сеть является вариантом ранней модели Inception V3. Он опирается на идею остаточного соединения в модели Microsoft ResNet, поэтому нейронную сеть можно обучать глубже. Используя несколько сверток 3×3 вместо сверток 5×5 и 7×7 , модуль Inception может быть чрезвычайно упрощен. В значительной степени снижается вычислительная сложность и размерность параметров, ускоряется обучение сети. Но по сравнению с другими сетями пространственная сложность этого алгоритма выше. Блок-схема сетевого режима Inception-Res Net-v2 показана на Рисунке 3.13.

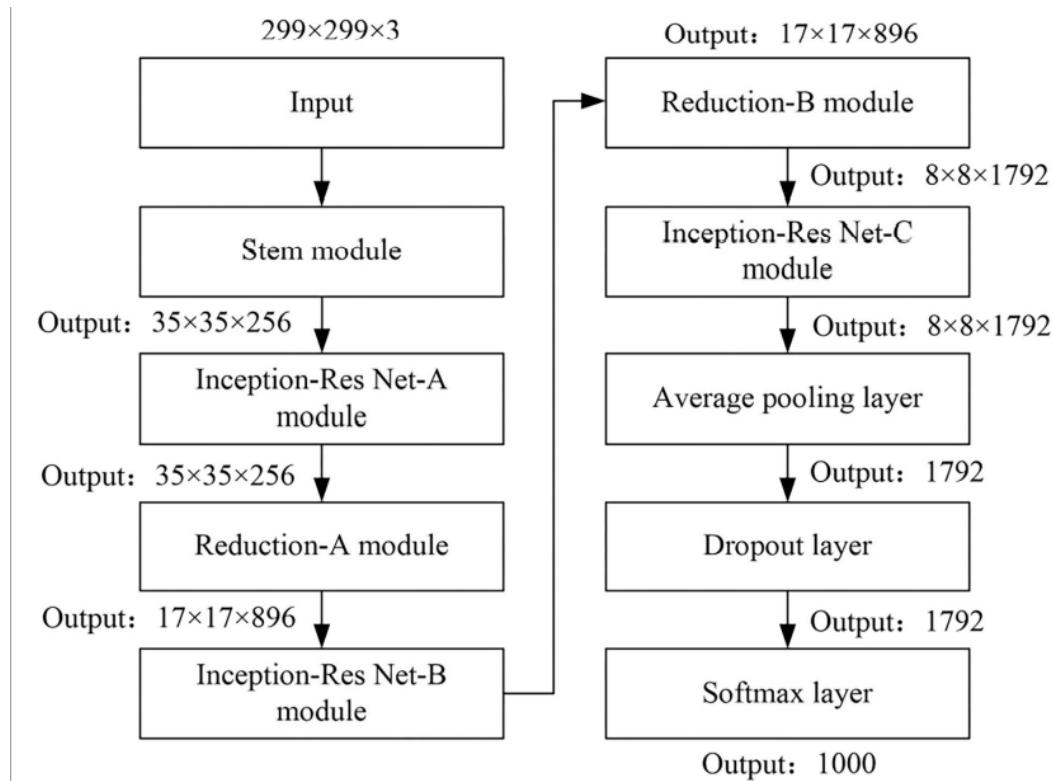


FIGURE 3 | Network mode flow diagram of Inception-Res Net-v2.

Рисунок 3.13 — Блок-схема сетевого режима Inception-Res Net-v2

Сеть Inception-Res Net-v2 в основном состоит из шести модулей, включая stem-модуль, Inception-Res Net-A, Reduction A, Inception-Res Net B, Reduction-B и Inception-Res Net C. Выходные векторы слоя Convolution, слоя Avg Pool, слоя Dropout и слоя Fully Connected, которые окончательно соединены сетью, можно рассматривать как характеристики глубины выборок, полученных после каждого слоя обучения. Путем классификации и сравнения признаков, выдаваемых Inception-ResNet-v2 в слое Convolution, слое Avg Pool, слое Dropout и слое Fully Connected, признаки полностью подключенного слоя с лучшим эффектом распознавания выбираются в качестве извлеченных глубоких признаков из видео.

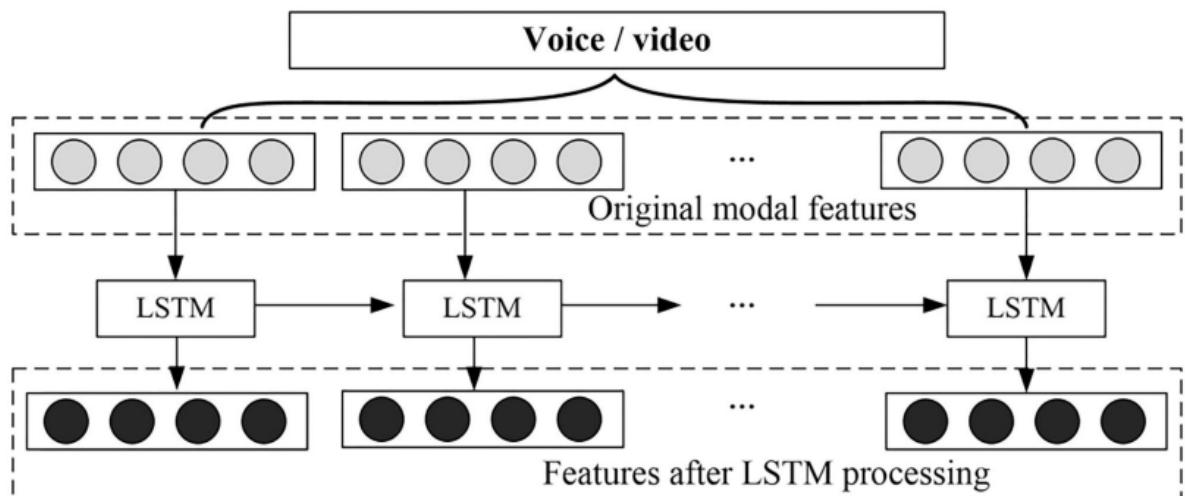


FIGURE 4 | RNN is used to capture modal internal dependencies.

Рисунок 3.14 — Объединение признаков разных модальностей

3.6.3 Отбор признаков

Чтобы уловить зависимость между внутренними характеристиками каждой модальности, используется структура LSTM-RNN. Корреляция существует между внутренними признаками одной модальности, а взаимозависимость существует между признаками разных модальностей. Когда признаки мономодальной модели скучны, информация, содержащаяся в другой модальности, может помочь в принятии решения. В связи с этим предлагаемый метод также

использует структуру LSTM для получения зависимости между различными режимами, и конкретная структура показана на Рисунке 3.14.

3.6.4 Результаты

Экспериментальные результаты показывают, что точность распознавания предлагаемого в статье метода на наборах данных MOSI и MELD составляет 87,56 и 90,06% соответственно, что лучше, чем у других методов.

Список использованных источников

1. V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe. *M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 4652–4661, June 2022.
2. Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. *DialogueTRM: Exploring Multi-Modal Emotional Dynamics in a Conversation*. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2694–2704, Punta Cana, Dominican Republic. Association for Computational Linguistics.
3. Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, Yongbin Li. *UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition*. arXiv preprint arXiv:2211.11256. November 2022.
4. Dong Liu, Zhiyong Wang, Lifeng Wang and Longxi Chen *Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning*. Frontiers in Neurorobotics Volume 15 Article 697634. July 2021
5. Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. *DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 154–164, Hong Kong, China. Association for Computational Linguistics.
6. Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. *COSMIC: COmmonSense knowledge for eMotion Identification in Conversations*. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2470–2481, Online. Association for Computational Linguistics.

Приложение А Приложение части Аудио

1. [Блокнот анализа]
2. [Скачать Аудио]
3. [Гармоничные и ударные звуки]
4. [Мел-спектограмма]
5. [Мел-кепстральные коэффициенты]
6. [Мел-кепстральные коэффициенты (подробно)]

Приложение Б Приложение части Текст

1. [Блокнот анализа]
2. [Скачать текст]

Приложение В Приложение части Видео

1. [Блокнот анализа]

Приложение Г Приложение общей части

1. [Биография Экмана, связанная с его классификацией]
2. [Биография Изарда, связанная с его классификацией]
3. [Сентимент-анализ. Определяем эмоциональные сообщения на Хабре]
4. [Сентимент-анализ. Википедия]
5. [Сентимент-анализ текста.]