Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)

Факультет Инноваций и Высоких Технологий
Кафедра банковских информационных технологий

Выпускная квалификационная работа бакалавра по направлению
01.03.02 «Прикладная математика и информатика»

Исследование и разработка realtime-моделей распознавания эмоций по голосовым аудиодорожкам с использованием Deep Learning

Студент 798а группы
Содиков М. М.

Научный руководитель
Нейчев Р. Г.

Долгопрудный
2021

Содержание

Аннотация

Задача распознавания естественной человеческой речи на сегодняшний день является одной из самых актуальных задач глубокого обучения. Помимо этой задачи, существует активно решаемая задача распознавания эмоционального тона аудиозаписи человеческого голоса.

В последнее время все чаще исследуются новые подходы к улучшению классификации аудио, уделяется отдельное внимание проблеме дефицита данных.

Целью данной работы является комплексное изучение подходов для Speech Emotion Recognition (англ. 'распознавание эмоций речи'), предлагается новый подход для улучшения робастности модели на основе спектральных аугментаций. Также, исследуется и описывается совершенно новый алгоритм Auto Data Core для разметки данных на основе распознавания эмоций видеодорожек.

Глава 1

Введение

1.1 Описание

Задача моделирования человеческого голоса и распознавания речи по аудиозаписи успешно решается в наше время с помощью методов глубокого обучения (англ. 'speech recognition', 'voice recognition'). Хорошо изучены и известны модели, позволяющие распознавать речь в режиме реального времени ('real time models') [**1**].

Помимо задачи обработки естественного языка по голосовой дорожке, существует актуальная задача распознавания эмоционального окраса аудиозаписи. Важно отметить, что распознавание должно происходить помимо речевого контекста и базироваться исключительно на речевых особенностях произношения и звукоизвлечения (так называемый 'голосовой пульс' от англ. 'glottal pulse'). Конструируемая модель должна быть устойчива к различным выбросам и шумам окружающей среды, при этом уметь тонко улавливать голосовые пульсы, которые и формируют идентификатор эмоционального окраса исследуемой аудиозаписи.

Всего существует 8 классов основных эмоций, считающихся фундаментальными, на которые и происходит классификация [**2**].

Обычно, для экстракции голосовых пульсов и других признаков исходной сырой аудиозаписи голоса человека, используются фильтры и спектрограммы, которые однозначно кодируют запись голоса человека в виде действительной (или в некоторых случаях комплексной) матрицы. Эта матрица в последствии и используется для последующего анализа и обучения моделей любой сложности.

Существуют подходы, обратные описанному выше. Так как задача распознавания и восстановления текста по записи голоса на данный момент решается лучше, чем задача предсказания эмоций напрямую, через обучение на сырых размеченных аудиодорожках, то распознают сначала текстовое представление аудиозаписи, а затем классифицируют на эмоции уже текстовую репрезентацию голоса человека.

В силу сложности разметки и сбора датасета для данной задачи, лучшие sota-алгоритмы, на данный момент используемые в реальных задачах сильно уступают в качестве стандартным методам NLP. Однако, гибридный метод перевода аудио в текст с последующей классификацией текста имеет ряд недостатков. Во-первых, повышается потенциальная погрешность прогноза, т.к. абсолютная ошибка предсказания модели формируется из ошибки модели перевода аудио в текст и самого классификатора. Во-вторых, такой подход не позволит правильно классифицировать сингулярные случаи, в которых может подразумеваться невербальный контекст, который не будет распознан моделью. Например, в случае, когда исходный текст может содержать ироничный посыл или подсмысл: 'Я отправляюсь на встречу. Мне в этом лице идти?'. Данное предложение может быть прочитано как и радостным тоном, так и с отвращением, – обе эти эмоции являются одними из основных классифируемых эмоций.

Данная НИР иллюстрирует новый подход к решению этой задачи, предлагается новый подход к аугментированию данных, а также генерации нового датасета автоматическим образом в unsupervised-режиме (в режиме "без учителя"). Исследуются и обозреваются стандартные подходы к задаче SER (англ. 'Speech Emotion Recognition'), как нейросетевых, так и моделей классического машинного обучения.

В данной главе описывается классическая постановка задачи глубокого обучения, в частности, задача распознавания эмоций человека по аудиозаписи его голоса, описывается проблема разметки данных и приводится обзор существующих моделей, успешно используемых для прикладного анализа эмоционального окраса аудио.

1.2 Формальная постановка задачи

В нашем случае рассматривается стандартная для машинного обучения задача классификации одного метрического пространства в другое.

Первое пространство – пространство сырых аудиосигналов – представляет из себя аудиодорожки в формате .mp3 или .ogg, или любых других доступных цифровых аудиоформатов. Аудиодорожки могут быть как монофоническими, так и стереофоническими – то есть, содержать в себе одно- или двухканальную запись звука. В нашем случае, для упрощения математической базы, мы будем представлять аудиодорожку в виде временного ряда действительных чисел. Это происходит с помощью библиотеки обработки аудиозаписей librosa на языке программирования Python 3.7. Функция librosa.load() умеет в автоматическом режиме преобразовывать сырую аудиодорожку в численное представление.

Обозначим аудиодорожку в виде вектора $\overrightarrow{x}$размерностиn. Тогда $\overrightarrow{x} \in X$, где X - пространство всех аудиодорожек, размерности $n * m$, где m - количество экземпляров в датасете.

Тогда, множество в которое происходит отображение, иначе говоря, множество классов, обозначим за Y, причем один конкретный класс будет являться скаляром $\overrightarrow{y} \in Y$, причем, в нашем случае, $Y \in [0, 7]$, т.к. значимыми являются именно восемь основных эмоций, включая нейтральное состояние, причем более комплексные эмоциональные окраски можно получать посредством смешивания основных базисных эмоций.

Некоторый функционал $\mathcal{F} : X \to Y$ будем называть решающим алгоритмом, или нейросетевой глубокой моделью. Процесс вычисления класса с помощью этой функции называется классификацией на соответствующий класс.

Стоит отметить, что в большинстве случаев мы будем заниматься классификацией не самих сырых аудиодорожек, как отмечалось в главе 1.1, а значащих репрезентаций аудиодорожек – фильтров над аудиодорожкой, спектрограмм аудиосигналов и сигнальной огибающей. Они сформируют из себя другое представление – матрицу размера $k * l$, а сам тензор S будет иметь размерность $m * k * l$.

Обладая данным математическим аппаратом, можно явно описать используемые модели.

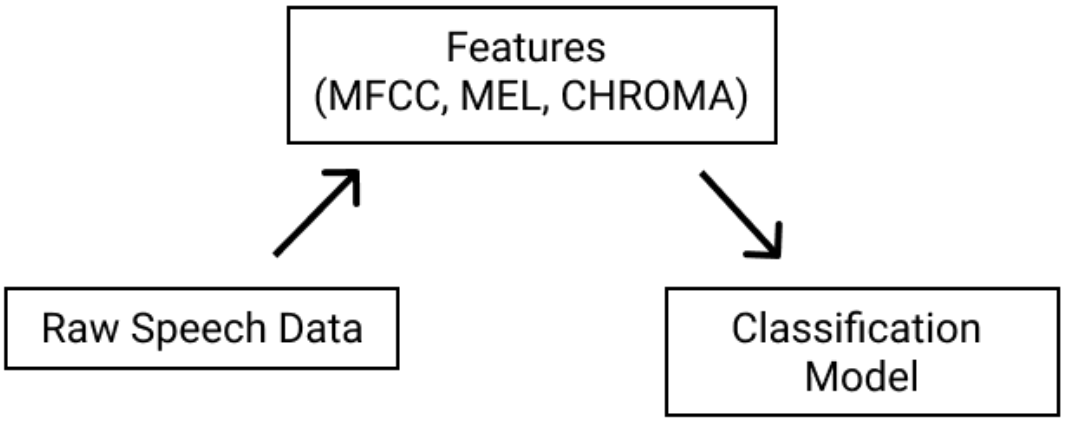

Рис. 1.1. Принципиальная схема pipeline для SER

1.2.1 Линейная регрессия

Линейная регрессия позволяет смоделировать взаимосвязь между двумя переменными, подгоняя линейное уравнение к наблюдаемым данным. Одна переменная считается аргументной переменной, а другая - зависимой переменной. Например, разработчик моделей может захотеть связать вес людей с их ростом, используя модель линейной регрессии. Прежде чем пытаться подогнать линейную модель к наблюдаемым данным, разработчик модели должен сначала определить, существует ли связь между интересующими переменными. Это не обязательно означает, что одна переменная влечет за собой другую (например, более высокие баллы по ЕГЭ не приводят к более высоким оценкам в университете), но говорит о том, что между двумя переменными существует некоторая статистически значимая взаимосвязь.

Диаграмма рассеяния (или же график линейной регрессии) может быть полезным инструментом для определения силы взаимосвязи между двумя переменными. Если кажется, что нет никакой связи между предложенными объясняющими и зависимыми переменными (т. е. диаграмма рассеяния не указывает на какие-либо тенденции к увеличению или уменьшению), то подгонка модели линейной регрессии к данным, вероятно, не даст полезной модели. Ценной численной мерой связи между

двумя переменными является коэффициент корреляции, который представляет собой значение от -1 до 1, указывающее на силу связи наблюдаемых данных для двух переменных.

Линия линейной регрессии имеет уравнение вида $\mathbf{Y} = a + b\mathbf{X}$, где X - независимая переменная (множество аудиодорожек, а точнее – их численных представлений), а Y - зависимая переменная (количество классов, на которые происходит классификация аудиодорожек). Наклон линии равен b, а a - точка пересечения.

В нашей задаче линейная классификация является так называемым бейзлайном – самой простой базисной моделью, которая справляется с задачей не хуже, чем случайный классификатор.

Также, к задаче SER (англ. 'Speech Emotion Recognition') применяется полиномиальная регрессия [**5**], представляющая собой более усовершенствованный бейзлайн и содержащая в себе более комплексные зависимости.

1.2.2 Решающие деревья и случайный лес

Решающее дерево в самом простом определении представляет из себя бинарное дерево, каждое звено (вершина) которого содержит в себе некоторую логическую предпосылку, а каждое из ребер является дихотомическим дискриминатором для некоторого множества – то есть, по сути, содержит в себе сценарии поиска нужного подмножества исходного классифицируемого пространства.

Более формально, в бинарном решающем дереве каждой внутренней вершине приписана функция $\beta_V : X \to 0, 1$, а листовой некоторое значение или вероятность. Пусть наш алгоритм стартует из корневой вершины и вычисляет значение функции β_{v_0}. В зависимости от результата алгоритм переходит в левую или правую дочернюю вершину, после чего процесс повторяется до тех пор, пока не будет достигнут лист дерева. Такова основная идея бинарного решающего дерева. $\beta_v(x, j, t) = [x_j < t]$— одномерный случай, где t — некоторый порог.

Отметим также, что решающее дерево разбивает всё признаковое пространство на некоторое количество непересекающихся подмножеств

и в каждом выдаёт константный прогноз.

Идея бэггинга (bagging, bootstrap aggregation) основана на обучении некоторого числа одинаковых алгоритмов и построении итоговой композиции как среднее данных алгоритмов.

Алгоритм случайного леса основан на бэггинге над решающими деревьями. В нем мы выполняем следующую последовательность действий: Для $n = 1 \dots N$

- Генерируем бутстрепную подвыборку (bootstrap);
- Формируется по ней решающее дерево $b_n(x)$:
- Дерево растет, пока в каждом листе не окажется не более n_{min} объектов, где n_{min} является гиперпараметром модели;
- При каждом разбиении сначала выбирается k случайных признаков из p и оптимальное разделение ищется только среди них;
- В конце возвращается $a_n(x) = \frac{1}{N}\sum_{n=1}^{N} b_n(x)$, которое и будет являться выходом нашей модели.

1.2.3 Глубокие нейронные сети

Нейронные сети — это семейство алгоритмов, основная идея которых заключается в использовании искусственных нейронов и связей между ними. Под нейроном понимается некоторая функция с множеством входом и одним выходом. Связи между нейронами имеют свои веса. Нейрон считает взвешенную сумму весов на своих входах, добавляет смещение и либо исключает это значение, либо использует дальше. Для последнего используется функция активации, причем ее вид должен быть дифференцируем. Среди наиболее популярных функций следующие:

- Сигмоида — $f(x) = \frac{1}{1+e^{-x}}$
- Гиперболический тангенс — $f(x) = \frac{2}{1+e^{-2x}} - 1$
- ReLu — $f(x) = max(0, x)$
- Softmax — $f_i(\overrightarrow{x}) = \frac{e_i^x}{\sum_{j=1}^{N} e_j^x}$, где $i = 1, ..., N$

Для обучения нейросетей используют метод обратного распространения ошибки, заключающийся в вычислении градиента и обновлении весов сети. Понятие глубоких нейронных сетей является довольно размытым в наше время. Одним из определений является разделение нейросетей на интерпретируемые человеком, то есть можно объяснить, почему связи между нейронами имеют тот или иной вес, и не интерпретируемые. Также глубокие нейросети имеют несколько слоев между входным и выходным слоями. В основном, нейронные сети можно разделить на несколько категорий:

- Сверточные нейросети (англ. Convolutional Neural Network, CNN);
- Рекуррентные нейросети (англ. Recurrent Neural Network, RNN);
- Генеративно-состязательные сети (англ. Generative Adversarial Network, GAN);
- Трансформер (англ. Transformer).

1.3 Датасет

В качестве материала для исследования использовался набор данных The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) с сайта Райерсонского университета SMART Lab [**19**]. Общий размер датасета 24,8 ГБ и он включает в себя 7546 файлов. Мы использовали 24 записи профессиональных актеров (12 мужчин и 12 женщин), озвучивающих два лингвистически идентичных предложения, что необходимо как раз для того, чтобы модель оставалась инвариантной относительно языковой модели. Данные высказывания читаются всеми основными эмоциями: спокойствия, счастья, грусти, злости, страха, удивления и отвращения.

Помимо этого, используется разделение датасета на обучающую, тестирующую и валидационную выборки, в соотношении 70%, 20% и 10% соответственно.

Также, в глубоком обучении повсеместно используется k-fold кросс-валидация. Весь датасет делится на k равных частей и каждая часть

является валидационной выборкой в рамках серии испытаний из k запусков процесса обучения.

Такой подход позволяет валидировать значения гиперпараметров релевантно структуре и природе датасета и избежать обусловленности процесса валидации на одну конкретную часть датасета.

Однако, данный процесс достаточно ресурсоёмок и требует много времени и вычислительных мощностей для обучения, в связи с чем в данной работе не был использован в силу ограничений.

1.4 Фильтры и спектрограммы

Мел (англ. mel)— это психофизическая единица высоты звука, логарифмическое преобразование частоты сигнала. Основная идея этого преобразования заключается в том, что звуки, находящиеся на одинаковом расстоянии по шкале мела, воспринимаются как находящиеся на одинаковом расстоянии от людей. Известно, что людям намного легче различать низкие частоты, чем высокие. В таком случае, даже если расстояние между двумя наборами звуковых сигналов (низкочастотным и высокочастотным) одинаково, наше восприятие этого расстояния — нет. Это делает понятие мела фундаментальным в приложении машинного обучения к задачам обработки звука, поскольку он имитирует наше собственное восприятие последнего.

Перевод из частоты (Гц) в высоту звука в мелах выглядит следующим образом:

$$m = 1127 \cdot ln\left(1 + \frac{f}{700}\right) = 2595 \cdot log_{10}\left(1 + \frac{f}{700}\right)$$

Обратное преобразование:

$$f = 700 \cdot \left(e^{\frac{m}{1127}} - 1\right) = 700 \cdot \left(10^{\frac{m}{2595}}\right)$$

Спектрограммы позволяют нам визуализировать звук и давление, создаваемое звуковыми волнами, что позволяет нам видеть форму записанного звука. Спектрограммы мела — это спектрограммы, визуализирующие звуки по шкале мела, а не по частотной [**13**].

Мел-коэффициенты (Mel Frequency Cepstral Coefficients, MFCCs) используются в задачах распознавания речи и поиска музыкальной информации, в частности, они хорошо представляют тембр. Основной алгоритм вывода MFCCs обычно включает следующие этапы:

- Преобразование из частот в мелы;
- Взятие логарифма от полученной величины;
- Использование дискретного косинусного преобразования;
- Получение в результате спектра по мел-частотам, а не по времени.

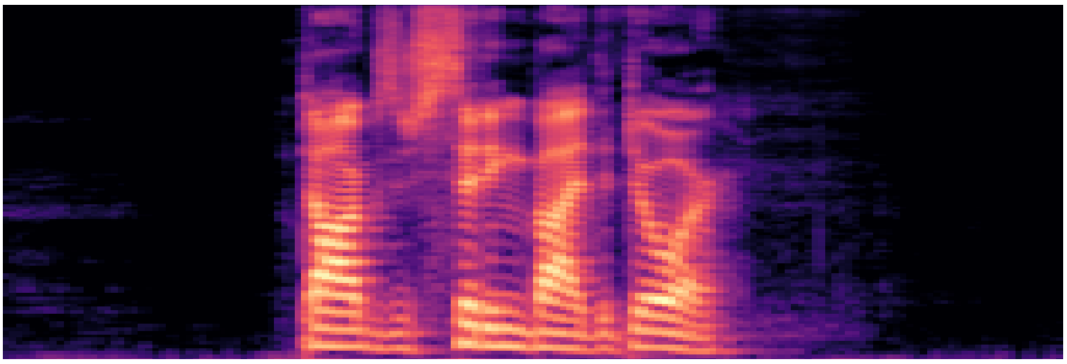

Рис. 1.2. График спектрограммы сигнала семпла из используемого датасета RAVDESS

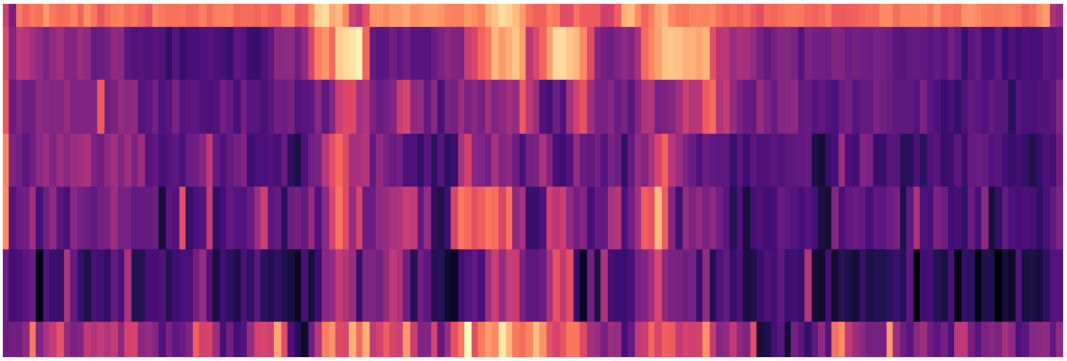

Рис. 1.3. График MFCC-фильтра сигнала семпла из используемого датасета RAVDESS

Рис. 1.4. График Chroma Feature сигнала семпла из используемого датасета RAVDESS

Если поставленная задача машинного обучения, например, автоматическое распознавание речи или шумоподавление, требует использования MFCC, то количество используемых коэффициентов является гиперпараметром модели. Из-за этого количество MFCC будет варьироваться в зависимости от задачи.

Хроматическая характеристика (Chroma feature) тесно связана с двенадцатью различными классами высоты звука. Признаки на основе хроматики являются мощным инструментом для анализа музыки, высоту тона которой можно разделить на категории и чье регулирование приближается к шкале с равномерным тепловым распределением. Одним из основных свойств характеристик хроматики является то, что они улавливают гармонические и мелодические характеристики музыки, при этом будучи устойчивыми к изменениям тембра и инструментального оснащения. Основное наблюдение состоит в том, что люди воспринимают две музыкальные ноты как похожие по цвету, если они отличаются на одну октаву. Основываясь на этом факте, высоту тона можно разделить на две составляющие: высота тона и хроматика. Основная идея хроматики в том, чтобы агрегировать ее для некоторого локального временного окна.

В нашем исследовании chroma feature используется в качестве одного из признаков, наряду со спектрограммой и MFCC.

Более неформально, принято считать, что MFCC содержит в себе главную (направляющую) огибающую человеческого голоса, который в

себе концентрирует информацию об метазвучании аудио, и низкочастотной компоненты.

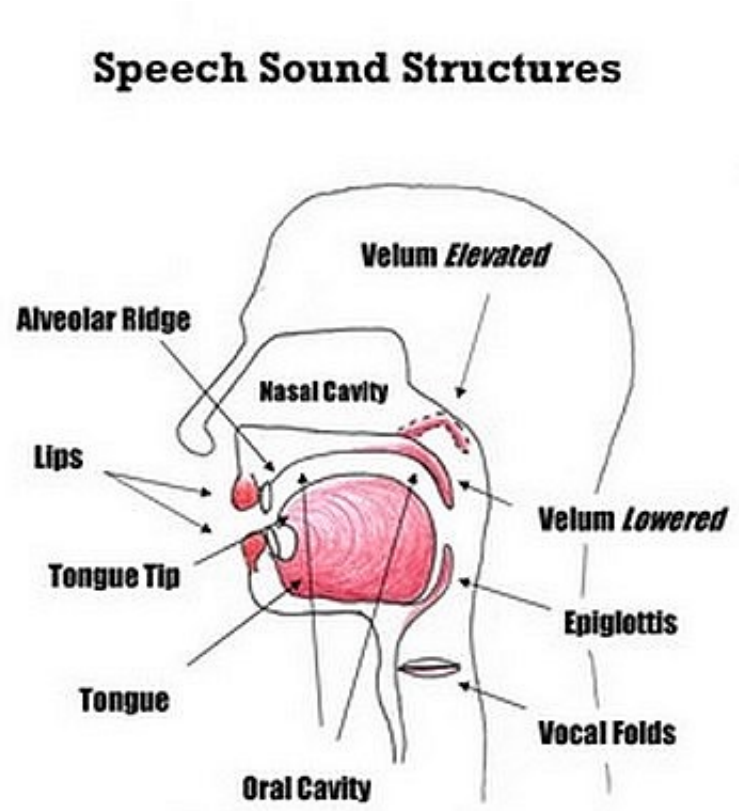

Рис. 1.5. Иллюстрация строения человеческого лица и базисные части носоглотки, используемые для извлечения звука. Считается, что метаинформация о звуковой дорожке концентрируется в коэффициентах MFCC, при использовании спектральной фильтрации. [**22**]

1.5 Проблема разметки данных

Задача SER решается давно, но существует характерная для области обработки аудиосигналов проблема недостатка данных, по крайней мере из открытых источников или в научно-исследовательских целях.

Однако, эту проблему решают с помощью различных методик. Например, широко применяется инструмент аугментации данных, о кото-

ром пойдет речь в 2.4-2.5 блоках. Также, существуют подходы, позволяющие в автоматическом или полуавтоматическом режиме размечать данные, на основе знаний или моделей, перенесенных из другого домена. Подобный подход подробно описан в [**12**].

В частности, в данной работе представлено два подхода к решению этой проблемы, путем расширения датасета в полуавтоматическом и автоматическом режимах, без участия человека или какого-либо дополнительного коллекционирования данных специальным образом с привлечением актеров озвучания, как это классически принято.

1.6 Обзор литературы

На данный момент существует несколько исследований на тему SER, в том числе, исследования возможности работы в условиях дефицита данных.

Как отмечалось выше, в [**12**] рассматривается исследование переноса с домена распознавания речи (NLP) в домен голосовой классификации. Данный подход дает хороший выигрыш, 71% по метрике Accuracy на 4-х классах, однако при расширении количества классов качество модели существенно падает.

В [**13**] предлагается использование спектральных коэффициентов Мела для задачи синтезирования речевого звука. Показывается нейросетевой подход, идейно используемый в настоящем исследовании.

В [**11**] описывается подход обучения классификации голоса человека "без учителяnsupervised learning. Для этого используется фреймворк так называемого контрастного обучения, при котором функционал ошибки является мерой различия звучания двух семплов из датасета.

В [**10**] предлагается метод, с использованием вложений (embeddings) Wav2Vec. Wav2Vec – это алгоритм векторного отображения аудиосигналов в сжатое векторное пространство меньшей размерности. В работе изучается подход классификации не спектрограмм или иных фильтров, а непосредственно сжатых представлений Wav2Vec.

В [**21**] применяются классические алгоритмы машинного обучения, описанные в разделе 1.2.1 - 1.2.2. Они показывают неплохое качество

работы, соразмерно своей легковесности и удобству обучения подобных моделей, однако предложенные

Глава 2

Предлагаемый подход

В данной главе описывается предлагаемый подход к решению задачи SER, давший значительный прирост относительно известных алгоритмов и существующих sota-алгоритмов и моделей машинного и глубокого обучения, применяемых для классификации на эмоции речи человека.

В блоке 2.1 дается краткий сводный ретроспективный анализ методов и моделей, описанных в блоке 1.4, но уже применительно к задаче распознавания и классификации эмоционального тона человеческой речи.

В блоке 2.2 предлагается подход, взятый за базисную (основную, 'backbone') модель, которая в дальнейшем используется для улучшения всего pipeline (англ. 'конвейер' – архитектура всей нейросети) глубокого обучения.

В блоке 2.3-2.4 описываются канонические и предлагаемые методы для аугментации данных, которые качественно повышают устойчивость модели к выбросам и в целом позволяют расширить объем данных для обучения, так как в задаче SER вопрос дефицита данных стоит особенно остро, как описывалось в разделе 1.6.

В блоке 2.5 предлагается совершенно новый подход к автогенерации данных, который позволяет генерировать данные на любых языках практически без участия наблюдателя и без наличия вообще какой-либо разметки для дальнейшего обучения предлагаемой модели SER. Описываются технические детали подхода, а также положительные и отрицательные аспекты решения, оценивается качество и ресурсоемкость подобной модели.

В блоке 2.6 подводится небольшой итог всех наблюдений и проделанных экспериментов, а также сводка ключевых метрик и параметров модели, которые важны для прикладного использования изученных методов и подходов.

2.1 Используемые метрики для оценки качества модели

Как и в любой задаче классификации на небинарные классы (multi-class classification), в нашей задаче основной метрикой оценки качества работы алгоритма является точность, или accuracy, которое определяется как отношение количества верно классифицированных семплов к общему количеству представителей этого класса.

$$Accuracy = \sum_{i=0}^{i=7} \frac{|GoodClassified[\text{class}_\text{i}]|}{|Dataset[\text{class}_\text{i}]|}$$

Также, используется матрица ошибок (confusion matrix), которая явно показывает, насколько сильно ошибается модель, какие неверные классы преобладают при предсказании и в целом позволяет на ее основе вычислить другие метрики.

Мы будем пользоваться матрицей ошибок как основным инструментом для оценки качества предлагаемых моделей, а промежуточные гипотезы будем подтверждать лишь относительно метрики Accuracy.

В качестве функционала ошибки для обучения нейросетевых моделей используется стандартный для задачи многоклассовой классификации Cross-Entropy Loss.

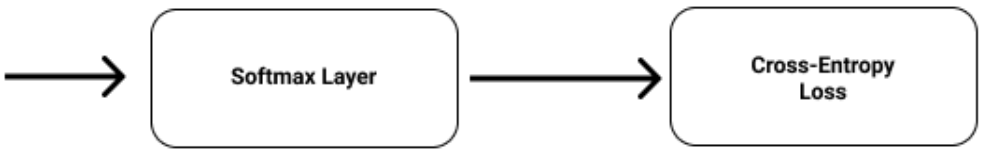

Рис. 2.1. Структура конечного слоя нейросети и функции ошибки при задаче мультиклассификации

2.2 Описание подходов

Канонически, принято использовать для задач, основанных на временных рядах, нейросети с так называемой рекуррентной структурой, к примеру, LSTM, RNN, GRU, причем они показывают хороший результат в различных задачах классификации временных рядов, или предиктивного анализа на основе исторических данных [**6**]. Кроме того, структура рекуррентных нейросетей смоделирована таким образом, чтобы в них накапливался исторический вклад каждой единицы временного ряда, иными словами, чтобы нейросеть "запоминала"предыдущие данные и обосновывала свои целевые предсказания с учетом хронологической значимости данных. Это позволяет, к примеру, хорошо работать с естественным языком, в частности, в приложениях к обработке текста, звука и других данных последовательной (временной) природы.

Как описано в [**7**], RNN также применяется для классической задачи text sentiment analysis (от англ. 'анализ настроения текста') и дает очень хорошие результаты, которые можно использовать в реальных приложениях. Согласно описанному в предыдущей главе подходу, в данной связи применяют гибридный pipeline для задачи SER, который проходит в два этапа: первый этап – перевод звука в текст, второй этап – классификация полученного текста на эмоции. Так как задача классификации текста решается намного лучше, чем задача SER, и не требует высоких затрат ресурсов в силу многих причин, таких как, к примеру, относительная простота разметки данных (в том числе автоматической) и, собственно, самого сбора данных, так и простоты и интерпретируемости подходов машинного обучения, которые используются в области NLP. Аналогично, задача распознавания речи тоже решается на достаточно хорошем уровне, как описано в [**1**].

В ходе данной работы, был исследован подобный гибридный подход и получен результат, который оказался не хуже, чем описанные в оригинальных статьях. Однако, данный подход противоречит в некотором смысле самой сути классификации голоса человека, так как является инвариантной относительно интонации, голосовых импульсов и других невербальных метахарактеристик сырого аудиосигнала, которые неявным образом и формируют истинный эмоциональный контекст речи человека.

Однако был исследован несколько усовершенствованный подход. Текст сам по себе все же несет небольшую часть истинного эмоционального контекста, однако, если утверждается, что основная метаинформация о природе настроеня аудиозаписи содержится как раз в самих особенностях звучания, то можно обучить некоторое вложение пространства аудиодорожек в более сжатое векторное пространство, которое и будет характеризировать эмоциональный контекст самой дорожки. Далее этот эмоциональный контекст передается некоторому классификатору, который, в свою очередь, выдает предположительный класс эмоции этой дорожки. В итоге образуется некоторый ансамбль двух моделей, каждый из которых делает свое собственное независимое предсказание, а итоговый результат определяется вероятностно или эмпирически заданным взвешиванием ответов каждой из нейросетей.

Иными словами, более формально: для модели $\mathcal{F}_1 : X \to Y$ классификации текста, и для двух моделей $\mathcal{F}_2 : X \to \hat{X}$ и $\mathcal{F}_3 : \hat{X} \to Y$, где – пространство эмбеддингов (вложений, от англ. 'embeddings'). Далее, для ответов каждой из моделей y_1 и $y2$, соответственно, можно получить итоговый результат в виде некоторой оценки $\alpha * y_1 + \beta * y_2$, где $\alpha + \beta = 1, \alpha, \beta > 0$. Параметры могут быть заданы как фиксированно, исходя из ряда наблюдений, либо же генерироваться из случаго распределения.

Вопрос получения вложений может быть решен, к примеру, с помощью вариационного автокодировщика, либо других автокодироващиков или GAN [**8**]. То есть, основная цель вложенного пространства – максимально ёмкая репрезентация исходного пространства. Предполагается, что в себе каждый вектор из множества $\hat{X}$ будет содержать информацию

о метаинформации исходной дорожки и представлять из себя выжимку эмоционального тона высказывания.

Относительно метрики, описанной выше, данный подход дает Accuracy 61%. Это очень мало для прикладных задач.

Для повышения качества модели в рамках исследования было решено отказаться от привязки к лингвистическому контексту сказанного человеческой речью, поэтому далее описываемые модели используют для обучения только сами аудиозаписи или некоторые преобразования над этими аудиозаписями – фильтры или спектральные коэффициенты.

Бейзлайном при работе с сырыми аудио будем считать многослойный перцептрон Розенблатта (MLP, Multi-Layer Perceptron) [**9**]. Он представляет из себя один слой нейросети, но не содержит нелинейной части.

В качестве признакового описания сырых аудио будем пользоваться конкатенацией следующих транфсормаций исходного сигнала:

- MFCC;
- Mel-спектрограммы;
- Хроматическое разложение;

Эти сигналы передаются нейросети. В качестве функционала ошибки выступает ранее описанный Cross-Entropy Loss из раздела 2.1.

Ниже, на Рис. 2.2 - 2.3 показаны кривые обучения (график функции ошибки и значение метрики Accuracy) для многослойного перцептрона Розенблатта.

Были выбраны следующие параметры обучения: размер подвыборки для обучения на эпохах, размер скрытых слоев, скорость обучения для планировщика обучения: `batch_size=256`, `hidden_layer_sizes = (300,)`, `learning_rate='adaptive'`.

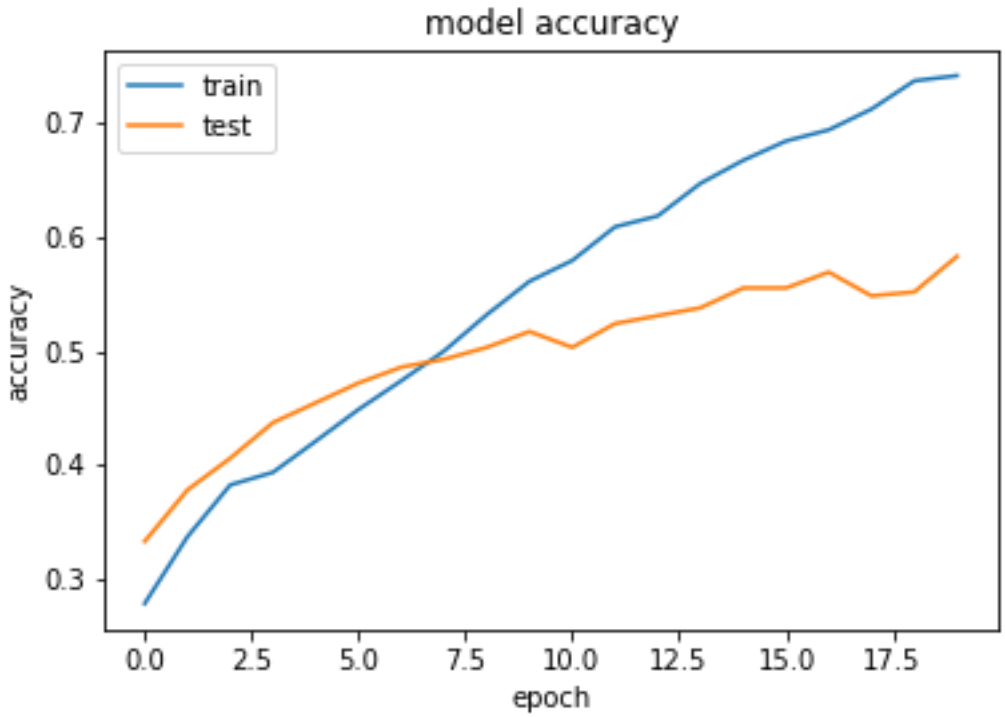

Рис. 2.2. Accuracy модели MLP с приведенными ниже параметрами на датасете RAVDESS. Обучение с использованием в качестве признаков стандартного вектора фильтров и спектральных коэффициентов.

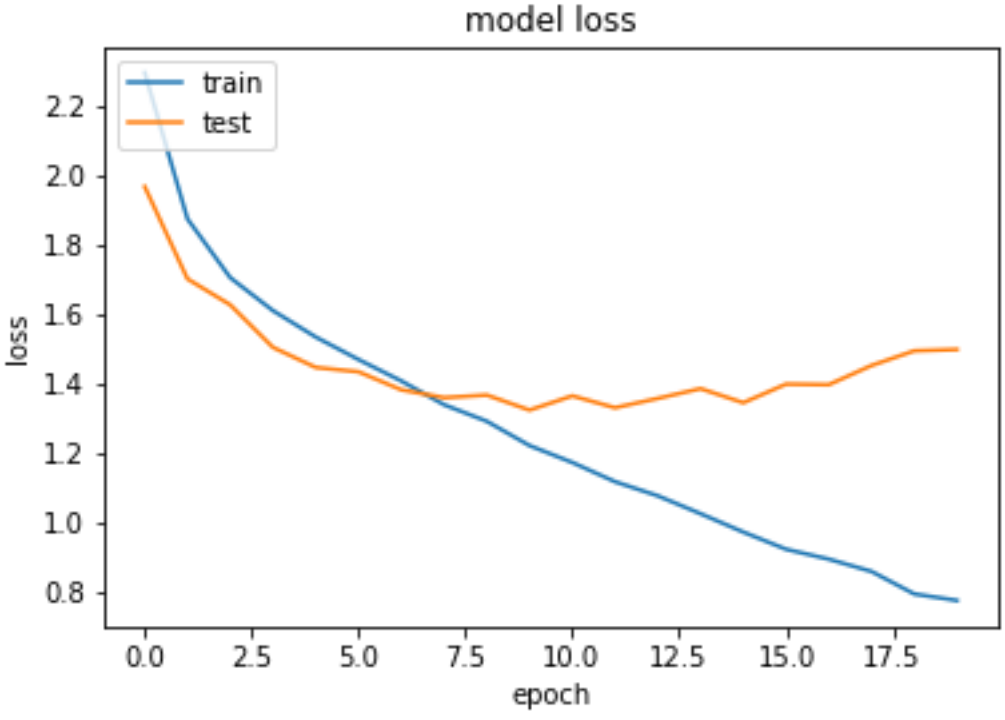

Рис. 2.3. Loss (функционал ошибки) модели MLP с приведенными ниже параметрами на датасете RAVDESS. Обучение с использованием в качестве признаков стандартной матрицы фильтров и спектральных коэффициентов.

Стоит отметить, что на графике функционала ошибки явно виден тренд к переобучению начиная примерно с 10-й эпохи. Данный тренд сохранялся и даже усиливался при изменении параметров. Объясняется это тривиальностью используемой модели, однако все равно показатель метрики Accuracy = 60%, при обучении на 50 эпохах является неплохим для такого простого бейзлайна.

2.3 Модель на основе сверточной нейросети (VGG-16)

В качестве основной модели будем использовать сверточную нейросеть (англ. convolutional neural network, CNN).

В качестве входных данных нейросети подаются те же матрицы спектров, что и для MLP.

В архитектуре самой модели, основной ее частью, выступает широко используемая для задач компьютерного зрения модель VGG-16. Она является предобученной на датасете ImageNet. Так как по сути, матрицы звуковых представляются несколькими двухмерными изображениями, утверждается, что данная модель сможет уловить закономерности, присущие для исходного аудиосигнала.

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
image_input (InputLayer)     [(None, 350, 350, 3)]     0
_________________________________________________________________
vgg16 (Functional)           (None, 10, 10, 512)       14714688
_________________________________________________________________
flatten (Flatten)            (None, 51200)             0
_________________________________________________________________
fc1 (Dense)                  (None, 4096)              209719296
_________________________________________________________________
fc2 (Dense)                  (None, 4096)              16781312
_________________________________________________________________
dropout (Dropout)            (None, 4096)              0
_________________________________________________________________
predictions (Dense)          (None, 8)                 32776
=================================================================
Total params: 241,248,072
Trainable params: 241,248,072
Non-trainable params: 0
```

Рис. 2.4. Архитектура нейросети на основе CNN с использованием VGG-16

Обученная нейросеть на основе VGG-16 дает следующие показатели матрицы ошибок (confusion matrix), приведенные на Рис. 2.5.

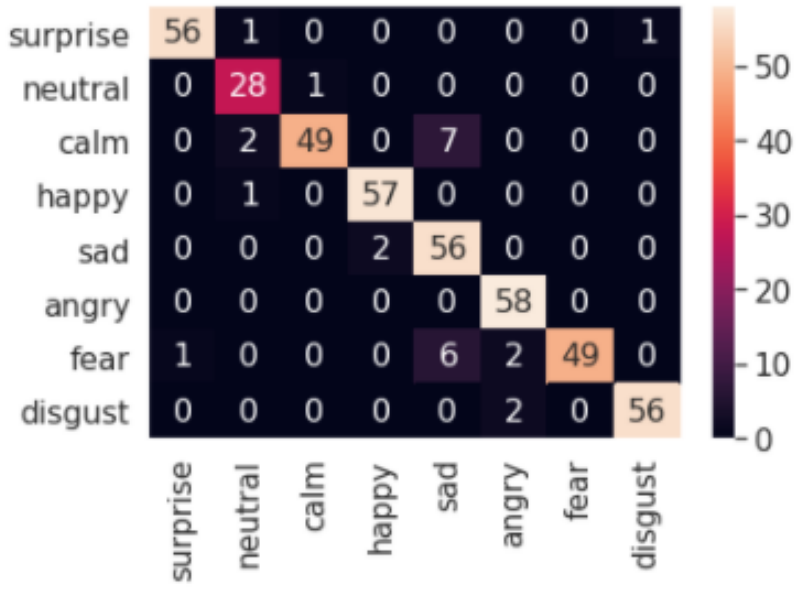

Рис. 2.5. Confusion matrix нейросети на основе CNN с использованием VGG-16

Accuracy модели на валидационной выборке достигает почти 75%.

Однако, если посмотреть на кривые обучения, а именно Loss модели во время обучения на 300 эпохах, то можно увидеть, что модель склонна переобучаться, даже не смотря на то, что она достаточно нетривиально по своей архитектуре. Связано это с тем, что природа данных подразумевает достаточно однотипное звучание каждого семпла, при этом содержит достаточно мало данных, что неизбежно приводит к переобучению модели.

Даже не смотря на то, что показатели целевой confusion matrix выглядит достаточно хорошо, плохая кривая обучения на Рис. 2.6 говорит о том, что модель может сильно ошибаться при классификации данных из другого распределения, нежели тех данных, на которых происходило обучение.

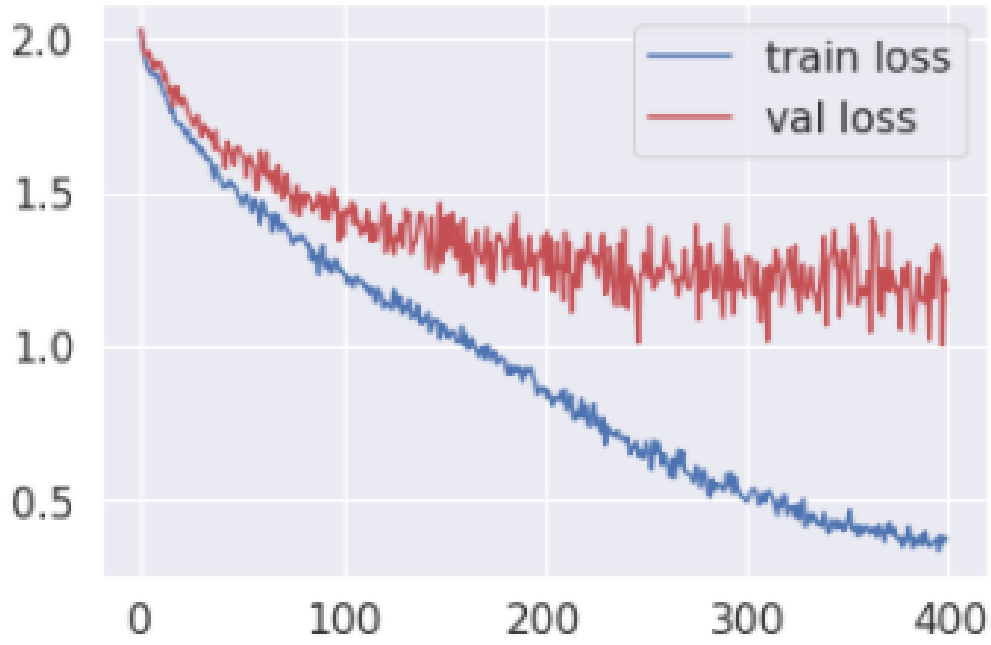

Рис. 2.6. Loss нейросети на основе CNN с использованием VGG-16. Видно сильное переобучение модели почти с самых первых эпох.

Поэтому, с этой целью был разработан бот в мессенджере Telegram для удобного и быстрого тестирования моделей и гипотез в режиме inference. Подробнее о ПО изложено в разделе "Созданное ПО" в главе "Выводы".

При исследовании работы бота в режиме реального времени, были получены эмпирические показатели ниже, чем на отложенной (валидационной выборке).

В связи с этим, возникает необходимость в расширении выборки данных, вариации разнородности исходных аудиозаписей.

В следующих главах описаны существующие и предлагаемые пайлпайны аугментации данных.

2.4 Стандартные аугментации сырых аудиозаписей

При работе с аудиосигналами, или, более того, работе с сигналами любой другой природы, классическим подходом к расширению и аугментированию данных является перечень следующих преобразований:

- Смешивание двух аудиодорожек в разных фазах и с разными амплитудами;
- Добавление шума из некоторого распределения (нормального, Пуассона и т.д.);
- Инвертирование звука;
- Комбинирование всех вышепредложенных аугментаций (вероятностно или фиксированно).

Подробнее данные аугментации описаны в работе [**15**].

Однако, подобные аугментации не сильно позволяют достичь разнородности данных, так как действуют достаточно поверхностным образом на сигналы. Они позволяют добиться увеличения качества, однако не всегда позволяют достичь высокой робастности модели [**16**].

Более глубокое изменения данных, с сохранением изначальной структуры звука – это аугментации спектрального разложения, описанные в следующем разделе.

2.5 Аугментации спектрального разложения

Аугментация спектра хорошо подходит для разнородной трансформации исходной выборки временных рядов или аудиосигналов, в частности, для нашей задачи обработки естественного языка.

Весь процесс аугментирования описывается следующим алгоритмом:

- Исходный сигнал переводится в спектральное представление (к примеру, с помощью быстрого алгоритма преобразования Фурье [**17**];
- К получившемуся спектру добавляется некоторый шум из какого-либо случайного распределения или в виде некоторой двухмерной или одномерной функции (по соответствующим осям коэффициентов);
- Полученный зашумленный спектр переводится обратно с помощью алгоритма обратного быстрого преобразования Фурье (или, например, с помощью Griffin Lim) [**18**];
- Полученная аугментированная аудиодорожка передается на вход классифицирующей нейросетевой модели и происходит ее дальнейшее обучение.

Стоит отметить, что применение аугментаций может быть как рандомизированным (случайным), так и последовательным.

Однако, эксперименты показали, что случайный подход сильно эффективнее последовательного. Происходит это потому, что при последовательном применении аугментаций спектр становится сильно зашумленным, что приводит к потере изначальной природы звуковых данных.

Также, следует уточнить, какие функции были использованы в качестве аугментации.

В рамках данной НИР был запущен эксперимент со следующими функциями шумов:

- Случайный (белый, нормальный) шум;
- Saw Tooth (пилообразная периодическая функция): $\left(\frac{t}{p} - \left\lfloor \frac{1}{2} + \frac{t}{p} \right\rfloor\right)$, p - период;

- Синусоидальная функция $A * sin(\omega * t)$, где A - амплитуда синуса, ω - частота;
- Случайное зануление коэффициентов;
- Умножение k случайных коэффициентов на некоторую случайную величину или константу;

Все параметры являются гиперпараметры и должны подбираться либо эмпирически либо с помощью методов поиска гиперпараметров по сетке.

Ниже видны графики шумов, использованных для аугментации:

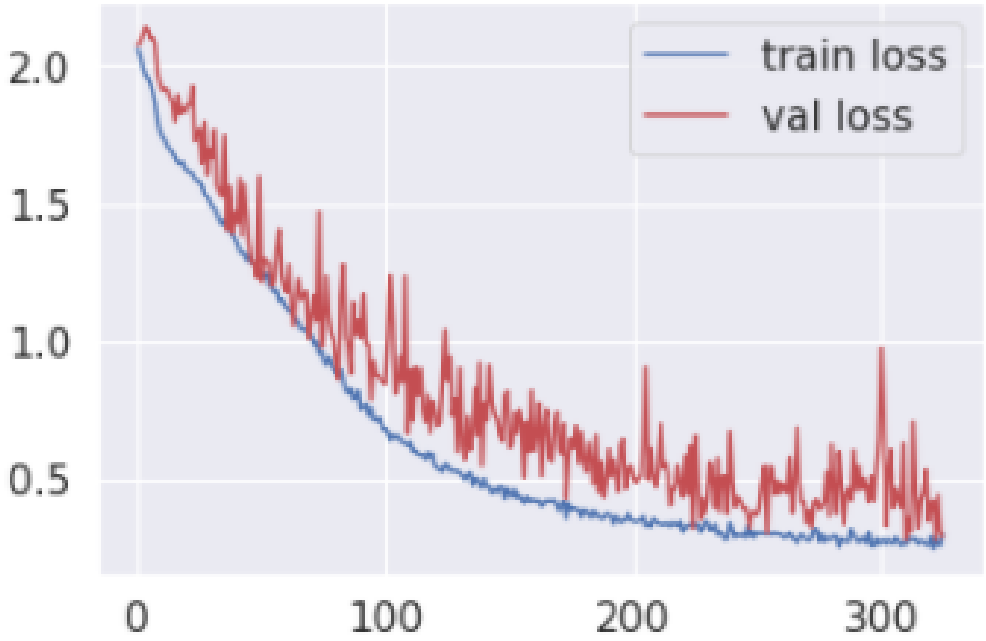

Рис. 2.7. Кривая обучения на валидационной выборке стала значительно более приближенной к кривой на обучающей выборке, а также очевидна хорошая сходимость процесса обучения.

После применения пайплайна аугментации, кривая обучения модели на основе VGG-16 выглядит, как показано на Рис. 2.7. Видно, что обучение модели стало заметно лучше, теперь не происходит переобучения модели. Это объясняется разнородностью обучаемых данных, а также расширением распределения исходной выборки.

На графиках ниже можно проследить, как изменяется состояние исходного сырого сигнала после применения различных спектральных аугментаций:

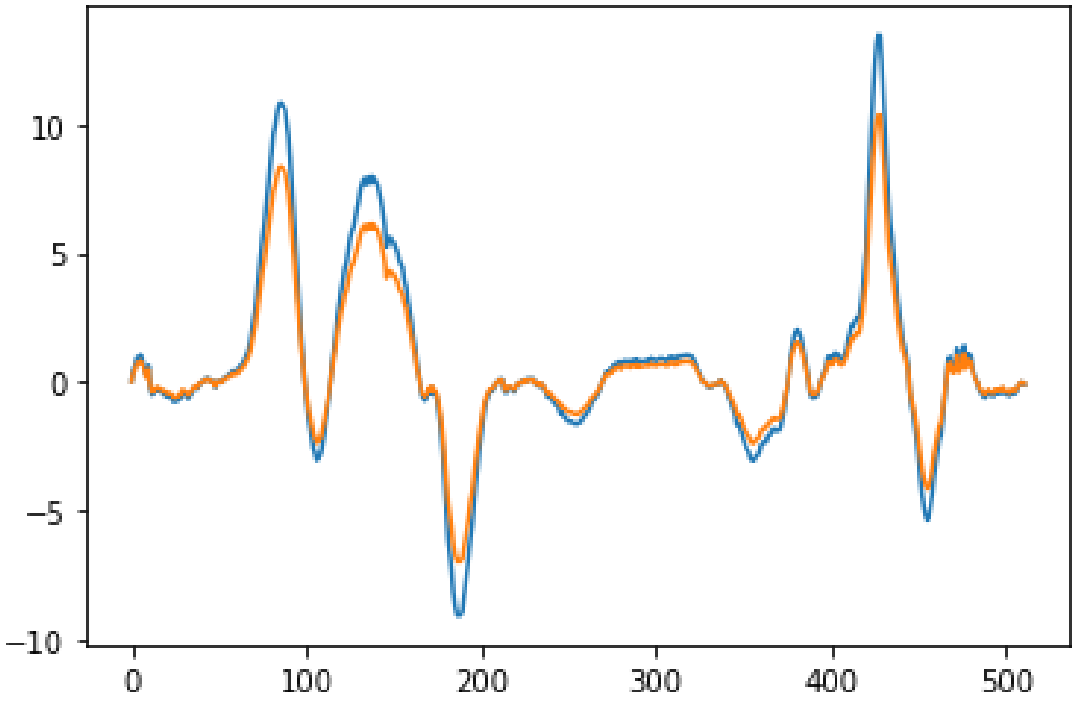

Рис. 2.8. Аугментация 2-го типа, Saw Tooth (пилообразная периодическая функция);

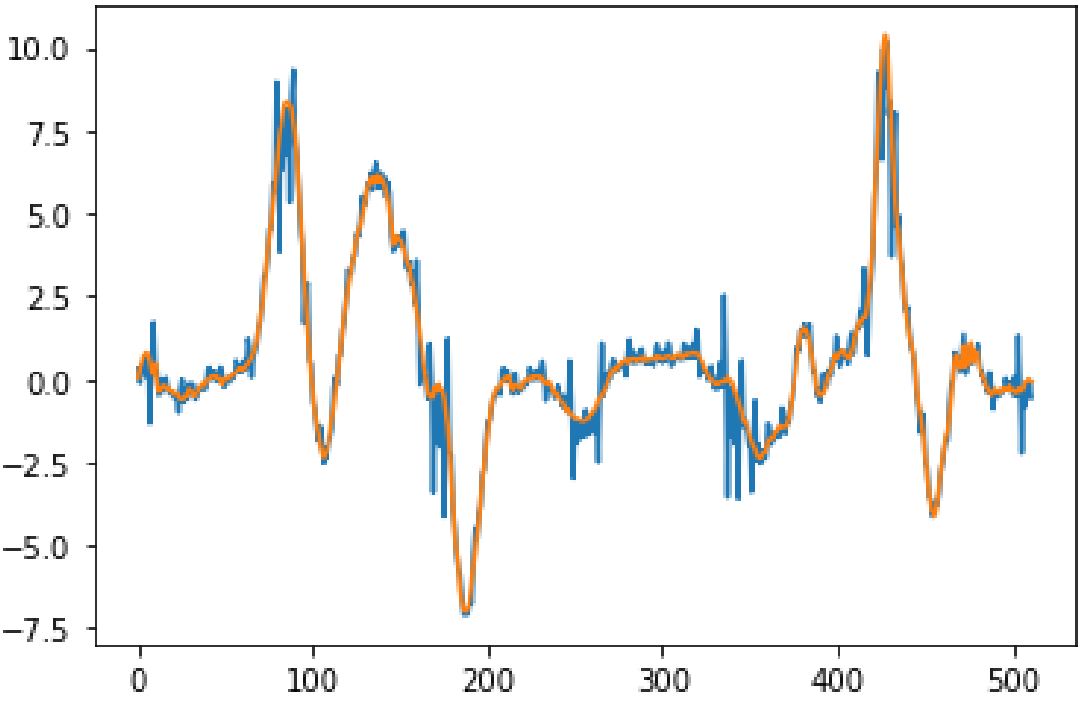

Рис. 2.9. Аугментация 5-го типа, Умножение k случайных коэффициентов на некоторую случайную величину или константу;

2.6 Автогенерация данных: Auto Data Core

В данной главе предлагается совершенно новый подход для автоматической генерации и авторазметки данных. Мы решили назвать ее 'Auto Data Core' (англ. 'ядро автоматической разметки данных').

Суть алгоритма автоматической разметки данных сводится к использованию промежуточного звена в виде нейросети $\mathcal{M}$, которая является предобученной нейросетью для распознавания человеческих эмоций по фотографиям.

Для алгоритма необходим объемный датасет с видеозаписями лиц людей с эмоциональными выражениями. Например, это могут быть различные интервью, выступления, или вечерние телепередачи. Суть алгоритма сводится к следующим шагам:

- Исходный озвученный видеопоток разделяется на аудиопоток A и видеопоток V;

  (1) Нейросеть $\mathcal{M}$ классифицирует V на некоторый класс: $\mathcal{M}(V[i_1 : i_2]) = y_j$, где $i_{1,2}$ – рассматриваемые концы отрезка времени (таймфрейма) в секундах, y_j – соответствующий класс для рассматриваемого таймфрейма;

  (2) Соответствующий таймфрейм ($A[i_1 : i_2]$) аудиопотока помечается как класс эмоции y_j;

- Шаги (1)-(2) алгоритма повторяются, пока не исчерпаются исходные потоки аудио и видео.

В нашем случае, мы пользовались самым простым бейзлайном для модели распознавания в виде ансамбля CNN и RNN моделей. Однако, существуют более качественные модели.

Это было сделано с целью лишь продемонстрировать потенциал данного алгоритма и указать на принцип работы самой схемы. Также, в силу ограниченности ресурсов времени и вычислительных мощностей, оказалось практически невозможным протестировать алгоритм на более крупном датасете и сложной комплексной модели $\mathcal{M}$.

Но даже для подобного минимально возможного набора инструментов, Auto Data Core показал сильный результат.

Финальная точность разметки на 8-и основных классах эмоций достигла значения около 50% (точность разметки была проверена как с помощью предобученных классификаторов, так и чисто эмпирически с помощью случайных выборок и оценки наблюдателем), что является достаточно высоким показателем, учитывая количество классов и пороги возможностей использованных методов.

Это означает, что при применении более совершенных моделей, а также более полноценных и качественных видеопотоков, записанных в максимально доступном разрешении,

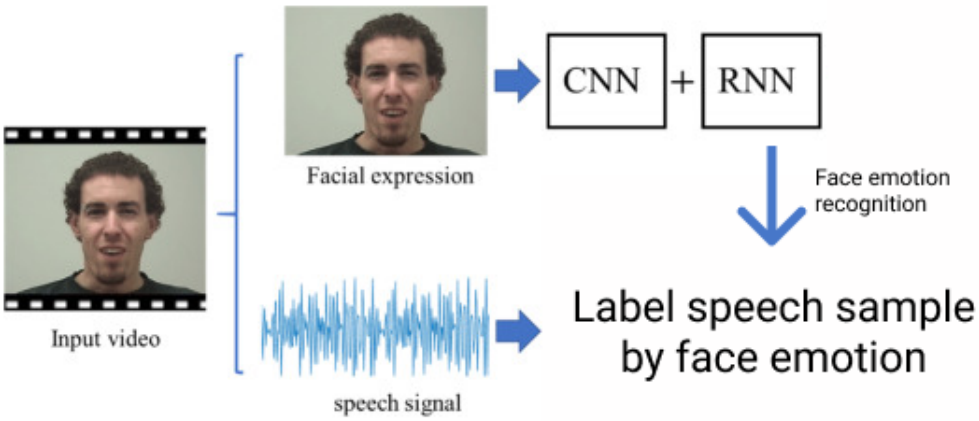

Рис. 2.10. Принципиальная схема алгоритма. Исходный видеопоток разделяется на аудио- и видео- составляющие. Видеопоток обрабатывается нейросетью, которая классифицирует выбранный временной фрейм некоторой эмоцией, которой в дальнейшем помечается звуковая дорожка и помещается в датасет.

На полученном датасете была обучена основная модель, которая дала следующие показатели Accuracy на валидационной и обучающей выборке:

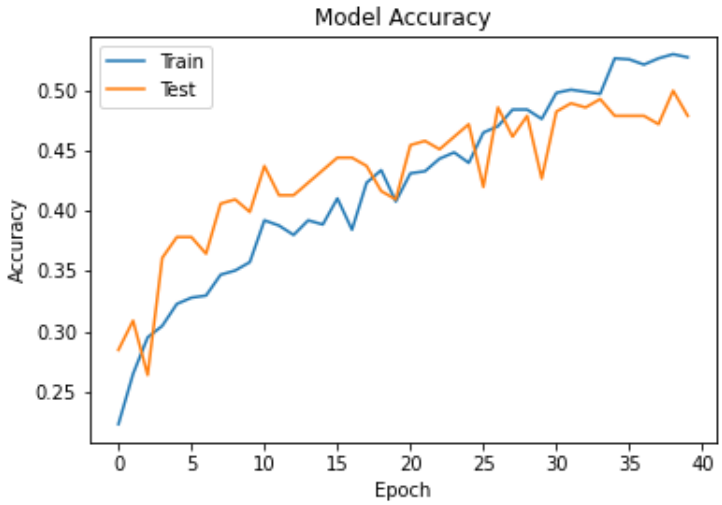

Рис. 2.11. Accuracy модели на основе VGG-16 во время обучения на полученном синтетическом датасете.

2.7 Ключевые метрики, анализ результатов

В этой секции подведем итоге предложенного алгоритма и обозначим основные значения метрик, а также дадим краткий анализ наблюдаемых экспериментов.

Повышение робастности модели с помощью спектральных аугментаций дала прирост качества по метрике Accuracy до 6%: точность предсказаний повысилась с 75% до 81%, что является существенным приростом в контексте задачи SER.

Более того, мы рассмотрели подход для синтетической генерации данных Auto Data Core. Та же модель, обученная на этой нейросети на 40 эпох показала точность работы в 50%, а при валидации на оригинальном датасете RAVDESS – около 42%.

Глава 3

Выводы

3.1 Созданное ПО

В рамках данного исследования был разработан автономный бот в мессенджере Telegram, который может в режиме реального времени анализировать голосовые сообщения, отправляемые боту в виде диалога и называть одну из 8 доступных эмоций.

Также создана библиотека, в которой имплементированы Python-скрипты для запуска обучения описанных моделей как на датасете RAVDESS, так и на любых других кастомизированных датасетах.

Весь код проекта доступен на `https://github.com/MakhmoodSodikov/SER_bachelor_thesis`

3.2 Итоги

В ходе исследования было выявлено, что можно весомо оптимизировать работу классических нейросетевых подходов для домена задачи SER.

Были предложены новые виды аугментаций, отличных от стандартных аугментаций миксования или фазового сдвига сигналов аудио, используемых в других задач. Спектральные аугментации позволили сделать процесс обучения модели значимо более гладкой и сходящейся, избавиться от проблемы переобучения, свойственной задачам с малым количеством данных.

Был предложен пайплайн моделирования шума для аугментации спектральных коэффициентов сигналов.

Помимо этого, был предложен алгоритм автоматической разметки данных на основе модели распознавания эмоций человека по видеозаписи лица (Auto Data Core).

Качество работы данного алгоритма весьма невелико, но для специфических задач, в условиях полуавтоматической разметки с контролем человека, она дает неплохие показатели целевых метрик.

Аугментации увеличивают робастность модели до 15% относительно метрики Accuracy на реальных зашумленных данных, таким образом, в совокупности с классическими аугментациями позволяют побить качество sota-алгоритмов.

Так как пайплайны аугментирования и авторазметки были исследованы вне критической привязки непосредственно к самой структуре и архитектуре модели глубокого обучения, предметом дальнейших исследований может быть предложено изучение применимости данных подходов к другим моделям и архитектурам, которые более релевантны для домена задачи SER и обработки аудиосигналов.

Однако, т.к. подходы, предложенные в работе, смогли показать прирост относительно существующих baseline и sota-алгоритмов, потенциал подходов явно высок и имеет потенциал для совершенствования.

Список литературы

[1] Developing Real-time Streaming Transformer Transducer for Speech Recognition on Large-scale Dataset, Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu, Jinyu Li (2021)

[2] Emotion Recognition from Speech, Kannan Venkataramanan, Haresh Rengaraj Rajamohan (2019)

[3] Text-based emotion detection: Advances, challenges, and opportunities, Francisca Adoma Acheampong, Chen Wenyu, Henry Nunoo-Mensah (2020)

[4] Encyclopedia of Personality and Individual DifferencesEditors: Virgil Zeigler-Hill, Todd K. Shackelford (2017)

[5] Speech Emotion Recognition System Based on L1 Regularized Linear Regression and Decision Fusion, Ling CenZhu Liang YuMing Hui Dong (2011)

[6] Time Series Forecasting Using LSTM Networks: A Symbolic Approach, Steven Elsworth, Stefan Güttel (2020)

[7] Deep Learning Based Text Classification: A Comprehensive Review, Shervin Minaee et al. (2021)

[8] Neighbor Embedding Variational Autoencoder, Renfei Tu, Yang Liu, Yongzeng Xue, Cheng Wang, Maozu Guo (2021)

[9] Multilayer Perceptron Algebra – A Mathematical Theory on the Design of Neural Networks, Zhao Peng

[10] Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings, Leonardo Pepino, Pablo Riera, Luciana Ferrer (2021)

[11] Contrastive Unsupervised Learning for Speech Emotion Recognition, Mao Li, Bo Yang, Joshua Levy (2021)

[12] A Transfer Learning Method for Speech Emotion Recognition from Automatic Speech Recognition, Sitong Zhou, Homayoon Beigi (2020)

[13] HIGH-QUALITY SPEECH SYNTHESIS USING SUPER-RESOLUTION MEL-SPECTROGRAM, Leyuan Sheng, Dong-Yan Huang (2019)

[14] Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi (2010)

[15] Audio augmentation for speech recognition, Tom Ko, Vijayaditya Peddinti, Daniel Povey, S. Khudanpur (2015)

[16] IMPROVING NOISE ROBUSTNESS OF AN END-TO-END NEURAL MODEL FOR AUTOMATIC SPEECH RECOGNITION, Jagadeesh Balam, Jocelyn Huang, Vitaly Lavrukhin, Slyne Deng, Somshubra Majumdar, Boris Ginsburg (2020)

[17] Discrete and Fast Fourier Transform Made Clear, Peter Zeman (2019)

[18] Deep Griffin-Lim Iteration, Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, Noboru Harada (2019)

[19] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Livingstone, Steven R.; Russo, Frank A. (2018)

[20] Modelling speech emotion recognition using logistic regression and decision trees, Agnes Jacob (2017)

[21] Speech emotion recognition based on DNN-decision tree SVM model, Linhui Sun, Bo Zou, Sheng Fu, Jia Chen, Fu Wang (2019)

[22] TECHNIQUES FOR FEATURE EXTRACTION IN SPEECH RECOGNITION SYSTEM : A COMPARATIVE STUDY, Urmila Shrawankar (2013)