МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

РАСПОЗНАВАНИЕ И КЛАССИФИКАЦИЯ ЗВУКОВ ПРИ ПОМОЩИ НЕЙРОННЫХ СЕТЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Пелипца Владислава Сергеевича

Научный руководитель
зав. каф. техн. пр., к. ф.-м. н. _____ И. А. Батраева

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы. Понятие «нейронная сеть» возникло в середине прошлого века. Мак-Каллок и Питтс в 1943 году разработали компьютерную модель нейронной сети, которая базировалась на работе головного мозга и математических алгоритмах. По их мнению, нейроны можно представить в виде устройств, работающих с бинарными числами. Ими была представлена структура сети из электронных нейронов, способная осуществлять различные числовые или логические операции, которые только можно представить. Подобная сеть, как они считали, способна даже к обучению, распознаванию образов и т. д., а это значит, что она имеет все черты интеллекта.

Были выявлены два подхода в работе с нейронными сетями. Первый нацелен на изучение биологических процессов в головном мозге, а второй — на применение нейронных сетей для решения разнообразных задач.

Одной из таких задач является распознавание, а именно распознавание звуков. Нейросети считаются довольно действенным способом в решении задач распознавания звуков. Например, распознавание голоса используется для обеспечения безопасности, для определения голоса конкретного человека, для голосового набора текста и для многого другого. Распознавание речи и других звуков очень нужная функция, популярность которой растёт с каждым днём.

Цель бакалаврской работы. Целью бакалаврской работы является разработка нейронной сети для распознавания и классификации звуков, сравнение результатов её работы с другими современными методами схожего назначения.

Поставленная цель определила **следующие задачи**:

1. Изучение сферы распознавания голоса и прочих звуковых сигналов;
2. Изучение структуры аудиофайла формата WAV;
3. Написание программы для получения и преобразования числовых характеристик WAV-файла;
4. Получение навыков для написания и обучения нейронной сети;
5. Выбор инструментов и модели обучения нейронной сети;
6. Получение навыков для обработки звуковых сигналов;
7. Изучение принципов работы различных методов классификации звуков;
8. Получение результатов работы этих методов.

Методологические основы. Методологическими основами для исследований в областях распознавания и классификации звуков при помощи ней-

ронных сетей послужили работы В. Н. Сорокина и А. И. Цыплихина, Е. А. Первушина, Н. С. Клименко, К. Л. Тассова и Р. А. Дятлова.

Теоретическая значимость бакалаврской работы. Теоретическая значимость бакалаврской работы заключается в наблюдениях и выводах, сформированных по результатам работы реализованной нейронной сети и других методов классификации звуков, а также в рассмотрении сферы распознавания звуков, теоретических основ для написания нейронной сети и процесса выделения критериев звукового сигнала.

Практическая значимость бакалаврской работы. Практическая значимость бакалаврской работы заключается в том, что реализованная нейронная сеть представляет собой достаточно точный метод классификации звуков, который не уступает другим существующим методам, и может иметь довольно обширный спектр применения.

Структура и объём работы. Бакалаврская работа состоит из введения, 2х разделов, заключения, списка использованных источников и 6и приложений. Общий объем работы — 53 страницы, из них 34 страницы — основное содержание, включая 22 рисунка и 4 таблицы, цифровой носитель в качестве приложения, список использованных источников информации — 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические вопросы использования нейронных сетей для распознавания звуков» посвящён сфере распознавания звуков, теоретическим основам для написания нейронной сети и процессу выделения критериев звукового сигнала.

Распознавание звуков. Физической основой распознания по голосу служит анатомия речевого тракта, свойства системы управления артикуляцией и особенности голосового источника. Анатомия тракта определяет спектральные характеристики звуков речи, система управления артикуляцией влияет на темп речи, скорость переходных процессов и длительность речевых сегментов, а голосовой источник определяет частоту основного тона и тембральные характеристики речевого сигнала. Мозг человека может не только понять смысл услышанного, но и идентифицировать человека по его речи. У каждого голоса есть индивидуальные особенности и характеристики, по которым можно распознать личность говорящего. Речь является разновидностью звуков, поэтому многие аспекты распознавания речи в частности относятся и к распознаванию звуков в целом.

Обработки речи состоит из:

1. анализа;
2. распознавания;
3. кодирования.

Распознавание, в свою очередь, делится на подвиды:

1. распознавание речи;
2. распознавание дикторов;
3. распознавание языка.

Есть большая разница между определением источника звука и самого звука. Это можно сравнить с распознаванием музыкального инструмента по звучанию и определением информации о композиции (название, музыкант, исполнитель и т. д.). Из-за возникшего интереса к распознаванию музыки были созданы сервисы, которые могут найти песню по её звучанию или даже ритму (Shazam, SoundHound, TrackID и т. д.). С распространением смартфонов голосовое управление получило полноценное применение. Были разработаны виртуальные ассистенты, которые могут помочь использовать возможности смартфона с помощью голосовых команд. Компания Google внедрила своего голосового

помощника Voice Search на Android устройства. Apple создала Siri — первого помощника с узнаваемым и запоминающимся голосом. Через некоторое время голосовые помощники появились и у других IT-гигантов: Microsoft представила Cortana, а Amazon — Alexa. Подобные технологии есть и у российских компаний: Алиса (Яндекс), Маруся (VK) и Олег (Тинькофф Банк).

Мировой рынок распознавания речи растёт с каждым днём. Это происходит из-за увеличения спроса на услуги голосовой биометрии. Также растёт частота нарушений безопасности, в то время как безопасность является основным требованием для многих компаний, и для государственных в первую очередь. Высокий спрос голосовой биометрии, обуславливается тем, что голос является уникальным для каждого человека, а это имеет большое значение в установлении личности.

Основными факторами рынка распознавания голоса являются:

1. спрос на услуги голосовой биометрии;
2. спрос на распознавание голоса в судебно-медицинских целях;
3. спрос на распознавание голоса в военных целях;
4. спрос на распознавание голоса в сфере здравоохранения.

Одной из проблем на рынке распознавания голоса является подавление внешних шумов. На этом рынке было много технологических достижений, но решение проблемы удаления посторонних помех всё ещё актуально для разработчиков. Высокая стоимость приложений с технологией распознавания речи также является большой проблемой.

Основными проблемами на рынке распознавания голоса являются:

1. трудность подавления посторонних шумов;
2. высокая стоимость приложений с распознаванием голоса;
3. возникающие иногда проблемы с точностью распознавания голоса;
4. малый уровень безопасности в установлении личности по голосу.

Распознавание звуков может помочь сотрудникам служб безопасности и правоохранительных органов определить, какие звуки представляют важность, а какие можно игнорировать. Если звук признан системой, как требующий внимания, то сотрудникам службы безопасности отправляется оповещение. Далее проводится дополнительный анализ звука на случай возможного ложного срабатывания. В случае подтверждения прогноза системы на место, где был зафиксирован звук, выезжает патруль.

Существует четыре категории звуков, которые наиболее важно уметь распознавать для обеспечения безопасности:

1. звуки выражения агрессии;
2. звуки разбитого стекла;
3. грохот взрыва;
4. звуки огнестрельного оружия.

Нейронные сети. Перед тем как создать нейронную сеть, нужно точно знать для решения каких задач она будет использоваться. Сейчас нейросети используются для многих целей, например, прогнозирование исходов, распознавание изображений и звука и т. д. Построение нейронной сети можно разбить на 5 этапов.

Первый этап — отбор входных данных. Оттуда нужно убрать всю лишнюю информацию. Нужно иметь довольно большое количество входных данных для обучения нейросети. По эмпирическому правилу, отношение количества обучающих примеров к числу соединений в нейросети равно $X < 10$. Перед тем как внести фактор в обучающее множество, нужно оценить его значимость и проанализировать возможный диапазон изменений.

Второй этап — преобразование входных данных с учётом проблем нейросетевой модели. Результативность нейросетевой модели будет лучше, если диапазоны изменения исходных данных соответствуют исходящим данным.

Третий этап — конструирование нейронной сети или же определение её архитектуры, например, необходимое количество слоёв и нейронов в них. Структура нейросети должна быть определена ещё до её обучения, из этого можно понять, что с этим лучше справится опытный разработчик.

Четвертый этап — обучение нейронной сети. Оно может происходить с помощью конструктивного или деструктивного подхода. При конструктивном подходе обучение нейросети проводится на малых нейросетях с постепенным увеличением, пока не будет достигнута нужная точность. При деструктивном подходе за основу берут принцип «прореживания», т. е. из заведомо большой нейросети убирают ненужные нейроны и их связи. Обучение нейронной сети состоит из изменения значений весов при помощи роста объёма информации, подаваемой на вход и получаемой на выходе.

Пятый этап — тестирование полученной модели нейронной сети на независимой выборке входных данных.

Одним из наиболее интересных умений искусственных нейросетей является их способность к обучению. Оно в какой-то степени похоже на развитие человеческого интеллекта. Может возникнуть мысль, что у человека всё так и работает. Но на самом деле это не совсем так, ведь возможности обучения искусственных нейросетей довольно ограничены по сравнению с человеческим мозгом.

Выделение критериев звукового сигнала. Процесс всей обработки звукового сигнала делится на следующие этапы:

1. предобработка сигнала;
2. выделение критериев;
3. распознавание звука.

В результате оцифровки аналогового сигнала, в нём будут посторонние шумы, которые затруднят последующую обработку. Громкость звука зависит от окружающей среды и не является одинаковой для двух разных звуков. Исходя из этого, кроме подавления шума нужно также наладить характеристику амплитуды входного сигнала.

На этапе предобработки сигнала данные фильтруются, а области, в которых нет полезного сигнала, убираются. Чтобы устранить такие области, используется алгоритм билатерального фильтрования.

Следующий шаг заключается в ликвидации областей без полезного сигнала из отфильтрованного сигнала. Для этого все значения амплитудно-временного спектра переносятся в положительную область оси амплитуд, после чего выполняется усреднение амплитудного спектра сигнал окнами в 25 мс. Исходя из вида этого спектра, можно делать выводы о наличии полезного сигнала.

Последним шагом предобработки сигнала является нахождение верхней границы для величин, в которых нет полезного сигнала. Для этого все значения спектра сортируются по возрастанию и с помощью метода золотого сечения находятся два пороговых значения с минимальной ошибкой относительно исходного спектра для кусочно-заданной линейной функции.

Входные векторы в ходе обучения сравниваются с соответствующими выходными векторами, которые могут быть двоичными или непрерывными. После обучения сети приложение входного вектора приводит к нужному выходному вектору. Из-за обобщающей способности сети можно получить правильный выход, даже если входной вектор был немного неверным или неполным.

Второй раздел «Реализация нейронной сети для классификации звуков» посвящён написанию программы для чтения данных WAV-файла, разработке нейронной сети для распознавания и классификации звуков и сравнению результатов её работы с другими современными методами схожего назначения.

Написание программы для чтения данных WAV-файла. Аудиоформат WAV является подвидом RIFF и его основной концепцией является chunk (кусок). WAV-файл состоит из двух раздельных областей: заголовок файла и область данных. В заголовке файла хранится информация о:

1. размере файла;
2. количестве каналов;
3. частоте дискретизации;
4. количестве бит в сэмпле (глубина звучания).

Чтобы лучше понять зачем нужна та или иная величина из заголовка, нужно сперва объяснить, что такое область данных и оцифровка звука. Звуковые колебания при оцифровке принимают ступенчатый вид. Это объясняется тем, что компьютер способен воспроизвести звук определённой громкости (или же амплитуды) в очень короткий промежуток времени, но он не бесконечно короткий и он зависит от частоты дискретизации. Например, у файла с частотой дискретизации 44.1 кГц этот промежуток времени равен 1/44100 секунды. Сейчас звуковые карты могут поддерживать частоту дискретизации до 192 кГц.

Амплитуда (громкость звука в коротком промежутке) выражается числом, которое занимает в файле 8, 16, 24, 32 бита и более. Точность звука зависит как раз от точности амплитуды. Чем больше места в файле занимает число, которым выражается амплитуда, тем больше диапазон значений этого числа и, следовательно, больше точность амплитуды. Амплитуды в моно варианте расположены последовательно, а в стерео сначала идёт значение амплитуды для левого канала, затем для правого, потом снова для левого и так по кругу. Сэмплом называется совокупность амплитуды и короткого промежутка времени.

Код программы для чтения данных WAV-файла был написан на языке C++ в программе Visual Studio 2019. В ходе работы было реализовано чтение всех данных аудиофайла формата WAV, который подаётся на вход. С этими данными также были проделаны некоторые вычисления. Эти числовые характеристики могут быть проанализированы. Фрагмент кода представлен ниже. На рисунке 1 показан результат работы программы.

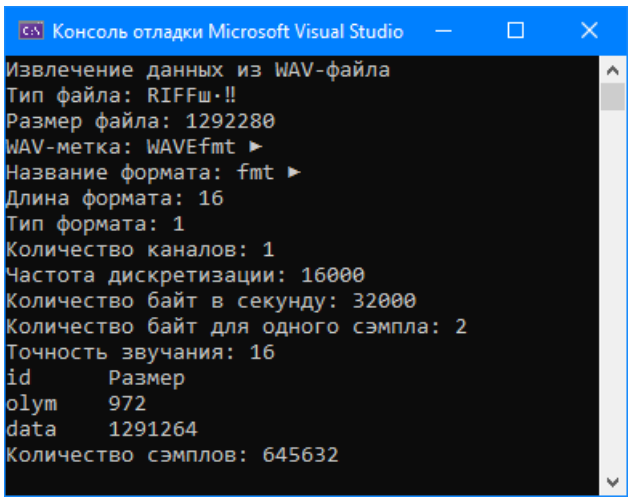

Рисунок 1 – Результат работы программы

Разработка нейронной сети для классификации звуков. Данная нейронная сеть была построена и обучена с помощью открытой программной библиотеки для машинного обучения TensorFlow. Код этой нейросети был написан на языке Python. Целью обучения нейронной сети являлась возможность классифицировать звуки, которые содержатся в аудиофайле. В качестве ПО для работы с нейросетями была выбрана библиотека Python Audio Analysis.

Одним из главных требований для машинного обучения является хороший набор данных. Таких наборов довольно много для распознавания музыки и речи, но не для случайных звуков. Однако был найден набор данных с городскими звуками. Ещё одним набором данных стал Google AudioSet, состоящий из размеченных видеофрагментов YouTube.

Новой целью стало выяснение принципа работы интерфейса YouTube-8M. К счастью, он может работать не только с видео-, но и с аудиоформатом. Из-за ограниченного количества классов в этой библиотеке, были добавлены изменения для того, чтобы число классов можно было задавать. В YouTube-8M можно работать с двумя типами данных: агрегированными и фрагментированными фичами (именно в таком виде Google AudioSet выдаёт данные).

Для машинного обучения гораздо лучше использовать графический процессор, нежели центральный. В процессе работы использовался компьютер с видеокартой NVIDIA GTX 970 4GB. Время обучения не имеет большого значения и пары часов обучения стало достаточно для принятия решения о выборе модели и её точности. В идеале, чем больше точность, тем лучше, но при обучении слишком сложной модели нужно было бы иметь больший объём видеопамяти.

Архитектура разработанной глубокой свёрточной нейронной сети состоит из 3 свёрточных слоёв, чередующихся с 2 операциями объединения, за которыми следуют 2 полносвязных слоя.

Обучение проводилось в течение 100 эпох с фиксированной скоростью обучения 0,001, размером батча равным 512, импульсом 0,9, размером фичи 128 и количеством классов 527.

В качестве инструмента извлечения фич используется модель TensorFlow VGGish. Сначала аудио преобразуется к формату 16 кГц моно. Далее происходит рассчёт спектрограммы с помощью преобразования Фурье с размером окна 25 мс, шагом в 10 мс и периодическим окном Ханна. Затем высчитывается мел-спектрограмма, которая приводит текущую спектрограмму к 64-разрядному мел-диапазону. Потом с помощью формулы $\log(\text{mel-спектр} + 0,01)$ рассчитывается стабилизированная логарифмическая спектрограмма со смещением (это нужно для того, чтобы не получился логарифм нуля). В конце всё это преобразуется в непересекающиеся фрагменты в 0,96 секунды, каждый из которых имеет размерность 64 мел-диапазона на 96 фреймов по 10 мс. Потом полученные данные поступают в модель VGGish для их дальнейшего приведения к векторному виду. Для передачи данных в нейронную сеть и получения результатов нужен интерфейс, за основу которого был взят изменённый интерфейс YouTube-8M.

При запуске программы выводится терминал (как на рисунке 2). Результат работы программы зависит от данных, которые подаются на вход. На терминале вывода информации выводятся данные, полученные на основе прогноза нейронной сети. Чем больше выведенное значение, тем выше вероятность принадлежности входных данных к тому или иному классу звуков.

```
2022-05-20 14:29:38   Music: 0.26
2022-05-20 14:29:43   Music: 0.56, Wind chime: 0.20, Speech: 0.13, Chime: 0.12
2022-05-20 14:29:48   Music: 0.51, Wind chime: 0.10
2022-05-20 14:29:53   Music: 0.68, Spray: 0.15, Grunge: 0.12, Speech: 0.11
2022-05-20 14:29:58   Music: 0.56, Speech: 0.13
2022-05-20 14:30:03   Music: 0.56, Speech: 0.15, Snort: 0.13, Crunch: 0.10
2022-05-20 14:30:08   Speech: 0.28, Music: 0.23, Fly, housefly: 0.23, Bee, wasp, etc.: 0.14
2022-05-20 14:30:13   Speech: 0.33, Music: 0.24, Fly, housefly: 0.18, Bee, wasp, etc.: 0.11
2022-05-20 14:30:18   Fly, housefly: 0.28, Speech: 0.26, Music: 0.19, Bee, wasp, etc.: 0.17
2022-05-20 14:30:23   Speech: 0.21, Music: 0.15
```

Рисунок 2 – Результат запуска программы

Сравнение различных методов классификации звуков. Было проведено сравнение результатов точности распознавания различных методов классификации звуков между собой и с реализованной нейронной сетью. Для тестирования точности распознавания использовались стандартные общедоступные наборы звуковых данных: ESC-50 и ESC-10. ESC-50 представляет собой набор данных из 2000 коротких (по 5 секунд) звуковых записей окружающей среды, относящихся к 50 одинаково сбалансированным категориям, выбранным из 5 основных групп (животные, естественные звуки природы, человеческие неречевые звуки, домашние звуки и городские шумы). В каждой категории по 40 образцов. Набор данных ESC-10 является подмножеством ESC-50, которое состоит из 10 классов (лай собаки, шум дождя, звук морских волн, плач ребёнка, тиканье часов, чихание человека, крик петуха, треск огня, шум вертолёта и бензопилы).

В ходе сравнения результатов точности распознавания на тестовых наборах данных было выявлено, что наиболее точной является нейросеть SoundNet. Её точность распознавания наиболее близка к точности распознавания человеком. Такие результаты сравнения можно объяснить различными способами обучения, объёмом и качеством набора данных для обучения и т. д. Точности распознавания звуков различными методами представлены в таблице 1.

Таблица 1 – Точность распознавания звуков различными методами

Метод	Точность на наборе данных из	
	50 категорий	10 категорий
SVM-MFCC	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest	44.3%	72.7%
Реализованная нейросеть	51.6%	75.4%
Piczak ConvNet	64.5%	81.0%
SoundNet	74.2%	92.2%
Человеческий слух	81.3%	95.7%

ЗАКЛЮЧЕНИЕ

При работе над нейронной сетью для классификации звуков на практике были изучены многие аспекты нейросети (её построение, архитектура, обучение и т. д.). Сферы распознавания голоса и музыки довольно большие, чего нельзя сказать об остальных звуках. С этой целью и создаются такие инструменты по распознаванию звуков, ведь на такие услуги тоже есть довольно большой спрос.

Целью дипломной работы являлась разработка нейронной сети для распознавания и классификации звуков, сравнение результатов её работы с другими современными методами схожего назначения. В связи с указанной целью были выполнены все поставленные задачи:

1. Изучена сфера распознавания голоса и прочих звуковых сигналов;
2. Изучена структура аудиофайла формата WAV;
3. Написана программы для получения и преобразования числовых характеристик WAV-файла;
4. Получены навыки написания и обучения нейронной сети;
5. Выбраны инструменты и модель обучения нейронной сети;
6. Получены навыки обработки звуковых сигналов;
7. Изучены принципы работы различных методов классификации звуков;
8. Получены результаты работы этих методов.

Основные источники информации

1. *Сорокин, В. Н.* Верификация диктора по спектрально-временным параметрам речевого сигнала / В. Н. Сорокин, А. И. Цыплихин. — Москва: Информационные процессы, 2010. — С. 18. — Яз. рус.
2. *Первушин, Е. А.* Обзор основных методов распознавания дикторов / Е. А. Первушин. — Омск: Омский государственный университет им. Ф.М. Достоевского, 2011. — С. 14. — Яз. рус.
3. *Клименко, Н. С.* Разработка структуры текстонезависимой системы идентификации диктора / Н. С. Клименко. — Донецк: Институт проблем искусственного интеллекта, 2012. — С. 11. — Яз. рус.
4. Global Automatic Speech Recognition Market 2014-2018. — Лондон: Technavio, 2013. — С. 54. — Яз. англ.
5. *Тассов, К. Л.* Метод идентификации человека по голосу / К. Л. Тассов, Р. А. Дятлов. — Москва: МГТУ им. Н.Э. Баумана, 2013. — С. 10. — Яз. рус.