МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра теоретических основ компьютерной безопасности и криптографии

Распознавание голоса нейронными сетями

РЕФЕРАТ

студентки 5 курса 531 группы
направления 10.05.01—Компьютерная безопасность
факультета КНиИТ

Змеевой Вероники Александровны

Проверил
доцент _____ И. И. Слеповичев
подпись, дата

Саратов 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Распознавание голоса – технология, позволяющая компьютерам понимать и интерпретировать человеческую речь, – переживает бурный рост благодаря быстрому развитию искусственного интеллекта и, в частности, нейронных сетей. Сегодня системы распознавания голоса интегрированы в повседневную жизнь миллионов людей – от виртуальных помощников, таких как Siri и Алиса, до автоматических переводчиков и систем диктовки. Их применение охватывает широкий спектр областей, включая медицину (автоматический анализ медицинских записей), безопасность (голосовая идентификация личности), автомобилестроение (функции автомобиля, управляемые голосом) и многие другие.

Технология распознавания речи началась в 1950-х годах с систем, которые могли распознавать только несколько слов. Первые примеры, такие как Shoebox от IBM, демонстрировали базовые способности, но испытывали трудности с естественным языком. В 1980-х годах скрытые марковские модели улучшили качество распознавания. Но сейчас традиционные методы распознавания речи, основанные на скрытых марковских моделях (HMM) и методах обработки сигналов, сталкиваются с ограничениями в обработке сложных акустических условий и вариативности речи. Поэтому в нынешнее время активно развиваются и используются повсеместно нейронные сети, способные к обучению на больших объёмах данных.

Цель данного реферата – систематизировать знания о применении нейронных сетей в распознавании речи и показать потенциал этой технологии для решения задач, связанных с пониманием и обработкой человеческого языка.

Задачи данного реферата:

1. Изучить этапы распознавания голоса;

2. Рассмотреть различные архитектуры нейронных сетей для распознавания голоса;
3. Изучить возможности распознавания голоса в условиях шума и многоголосия;
4. Проанализировать инструменты и способы обучения нейронных сетей для работы с речью;

1. Основы распознавания речи

Эта глава посвящена изложению основ традиционных подходов к распознаванию речи, которые, хотя и уступают по эффективности современным нейронно-сетевым методам, играют важную роль в контексте исторического развития и понимания современных архитектур.

1.1 Этапы процесса распознавания речи

Основные этапы распознавания речи включают:

Процесс начинается с записи звука с помощью микрофона или других устройств, таких как наушники или интеллектуальные колонки. Затем этот звук преобразуется в цифровой формат для обработки.

Записанный звук разбивается на более мелкие единицы, такие как фонемы (наименьшие звуковые единицы) и спектрограммы (визуальные карты звуковых частот). Эти признаки необходимы для понимания речи. [6]

Гласный звук состоит из одного слога, слог состоит из одной или нескольких фонем.

Фонемы являются наименьшими элементарными составляющими речи человека, синтез которых на основе аналитических моделей и наборов связных параметров позволит генерировать более крупные элементы речи: звуки, буквы и слова. [1]

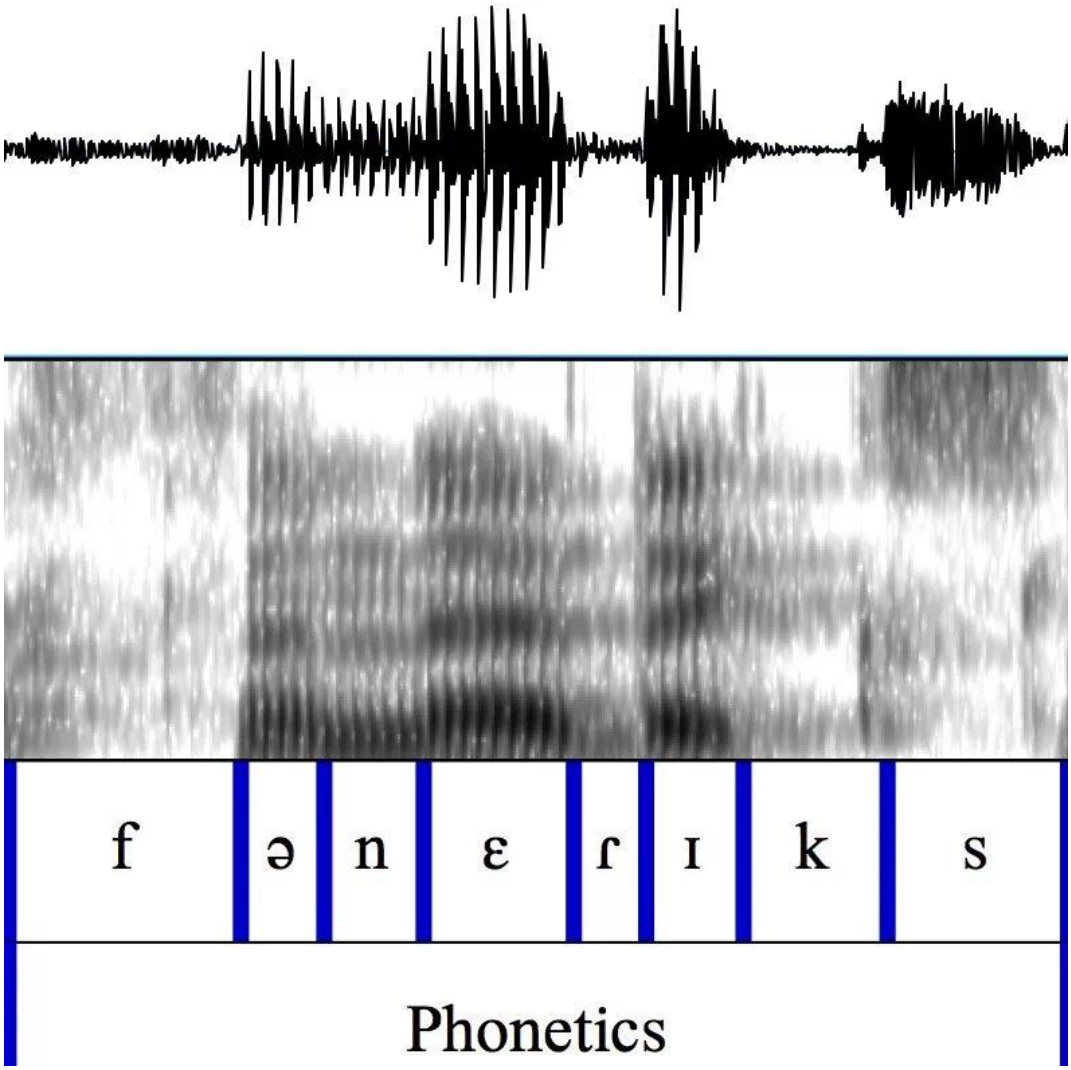

Рисунок 1 – акустическая запись, спектрограмма и фонемы этой записи [5]

Акустические модели обучаются с использованием больших наборов речевых данных, чтобы узнать, как звуковые характеристики связаны с фонемами. Это помогает системе распознавать различные звуки с учетом таких факторов, как голос говорящего, акцент и фоновый шум.

Языковые модели используют грамматические правила и словарный запас для прогнозирования наиболее вероятных слов или фраз на основе контекста. Этот шаг повышает точность, гарантируя, что распознанные слова будут иметь смысл в данном контексте.

На заключительном этапе происходит поиск наиболее вероятной последовательности фонем или слов, которая соответствует извлеченным признакам, с учетом акустической и языковой модели, система объединяет

данные акустической и языковой моделей для создания текста, преобразуя распознанные звуки в грамматически правильные предложения. [6]

1.2 Обзор традиционных методов распознавания речи

Традиционная архитектура системы распознавания речи. включая линейный дискриминантный анализ (LDA), линейное преобразование с максимальным правдоподобием (MLLT), i-вектор, линейную регрессию с максимальным правдоподобием (fMLLR) и другие методы извлечения речевых признаков. Прорыв в области акустических моделей пришелся на 1980-е годы.

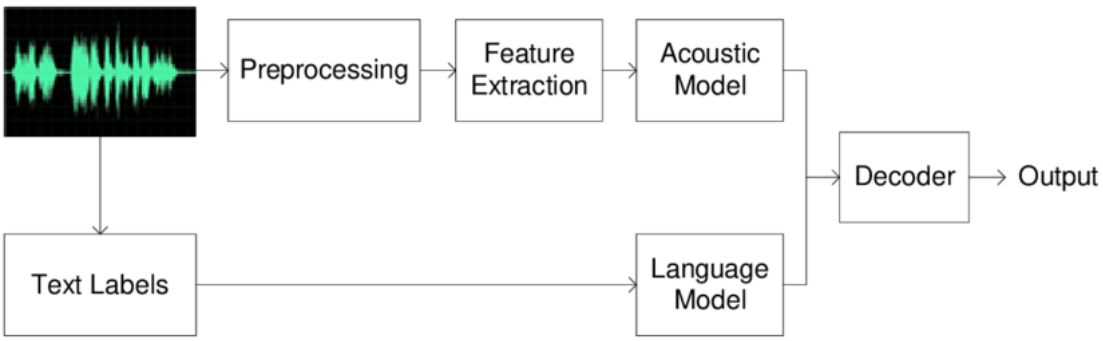

Рисунок 2 – традиционная архитектура системы распознавания речи

Метод, основанный на статистических моделях, которые были представлены методом гауссовой смеси – скрытой марковской модели (GMM-HMM), постепенно стал доминирующим в исследованиях распознавания речи. Также был разработан ряд других связанных технологий, основанных на методе HMM, таких как использование линейной регрессии максимального правдоподобия (MLLR) и критерия максимальной апостериорной вероятности для преодоления проблемы адаптации параметров в процессе обучения HMM. Кроме того, идея объединения состояний была использована для достижения привязки состояний дерева принятия решений, когда имеется много параметров обучения при меньшем количестве обучающих данных.

Искусственная нейронная сеть (ANN) также впоследствии дала новую исследовательскую идею для распознавания речи. В 2006 году Хинтон и др. использовали ограниченную машину Больцмана (RBM) для инициализации узлов нейронной сети, и появилась сеть глубоких убеждений (DBN). Сеть использовала неконтролируемый жадный послойный метод, чтобы максимально сохранить вес моделируемого объекта, и постоянно подгоняла его для получения этого веса. С тех пор сочетание глубокого обучения и традиционных методов вошло в моду, и глубокая нейронная сеть (DNN) продемонстрировала тенденцию к вытеснению модели GMM вместо использования традиционного метода GMM для моделирования состояний HMM.

Первым прорывом стала акустическая модель DNN-HMM, которая в значительной степени способствовала применению глубокого обучения в распознавании речи. Этого достаточно, чтобы продемонстрировать мощь глубокого обучения. Самое главное, что мощная функция извлечения признаков сверточной нейронной сети (CNN) позволяет лучше понимать сложные речевые особенности. Рекуррентная нейронная сеть (RNN), которая подходит для моделирования последовательности, может лучше использовать характеристики взаимосвязей временных рядов для создания контекстно-зависимых моделей.

В частности, новейшая сквозная система распознавания речи устраняет проблему принудительного выравнивания традиционного HMM и обеспечивает общую оптимизацию последовательности предложений. [4]

2. Нейронные сети для распознавания речи

В данной главе будут рассмотрены основные принципы работы нейронных сетей в контексте распознавания речи, включая архитектуры, алгоритмы обучения и подходы к обработке аудиосигналов. А также проанализировано, как нейронные сети обрабатывают различные характеристики речи, такие как фонемы, интонацию и темп, а также обсудим текущие достижения и вызовы, стоящие перед исследователями в этой области. Понимание этих аспектов является ключевым для разработки эффективных и надежных систем распознавания речи, способных интегрироваться в повседневную жизнь пользователей и решать широкий спектр задач.

2.1 Архитектуры нейронных сетей, используемые для распознавания речи

2.1.1 Рекуррентные нейронные сети (RNN)

В наше время лучшим средством для создания движка распознавания речи стала рекуррентная нейросеть (RNN), на которой построены все современные сервисы распознавания голоса, музыки, изображений, лиц, объектов, текста. RNN позволяет с высочайшей точностью понимать слова, а также предсказывать наиболее вероятное слово в рамках контекста, если оно не было распознано.

Что касается приложений распознавания речи, то входной сигнал $x = (x_1, x_2, \dots x_T)$ передается через RNN для вычисления скрытых последовательностей $h = (h_1, h_2, \dots h_N)$ и выходных последовательностей $y = (y_1, y_2, \dots y_N)$, соответственно. Одним из основных недостатков простой формы RNN является то, что она генерирует следующий вывод только на основе предыдущего контекста.

RNN вычисляют последовательность скрытых векторов h как:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

$$y_t = W_{hy}h_t + b_y,$$

где W – веса, b – векторы смещения и h – нелинейная функция.

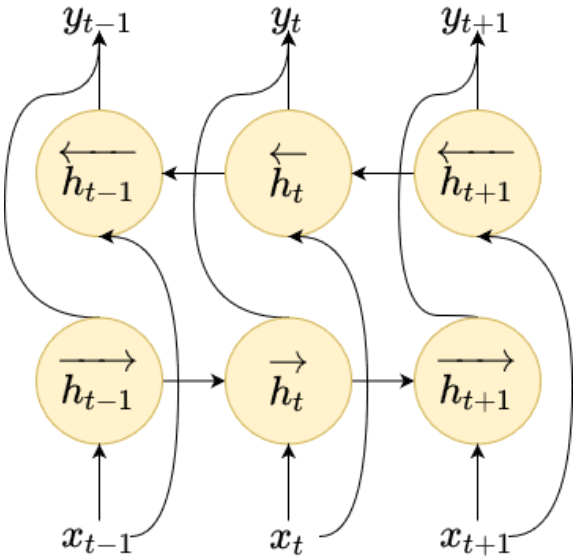

Рисунок 3 - Двунаправленная RNN

Однако при распознавании речи обычно информация о будущем контексте имеет такое же значение, как и о прошлом контексте. Вот почему вместо использования однонаправленной RNN обычно выбираются двунаправленные RNN (BiRNN), чтобы устранить этот недостаток. BiRNN обрабатывает входные векторы в обоих направлениях, т.е. вперед и назад, и сохраняет скрытые векторы состояния для каждого направления, как показано на рисунке выше. [7]

Помимо этого, используются долгая краткосрочная память (Long short-term memory; LSTM) – особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям. Они были представлены Зеппом Хохрайтер и Юргеном Шмидхубером (Jürgen Schmidhuber) в 1997 году, а затем усовершенствованы и популярно изложены

в работах многих других исследователей. Они прекрасно решают целый ряд разнообразных задач и в настоящее время широко используются. [12]

LSTM разработаны специально, чтобы избежать проблемы долговременной зависимости. Запоминание информации на долгие периоды времени – это их обычное поведение, а не что-то, чему они с трудом пытаются обучиться.

Для решения задач машинного перевода долгое время использовались рекуррентные нейронные сети, которые умеют обрабатывать потоковые данные. Однако они плохо работали с длинными зависимостями: для получения полноценного и связного итогового текста недостаточно перевода отдельных предложений, нужно учитывать общий контекст. Чтобы исправить проблему был разработан «механизм внимания» (англ. attention), позволяющий концентрироваться на важных частях текста. С его помощью нейросеть оценивает, какая позиция входящей последовательности важна для конкретной позиции последовательности на выходе. Также рекуррентные сети требовали последовательных вычислений, что ограничивало возможность эффективно применять современные графические процессоры для обучения моделей.

2.1.2 Глубокая нейронная сеть (DNN)

Глубокая нейронная сеть (DNN) – это искусственная нейронная сеть с прямой связью, которая содержит более одного скрытого слоя нейронов.

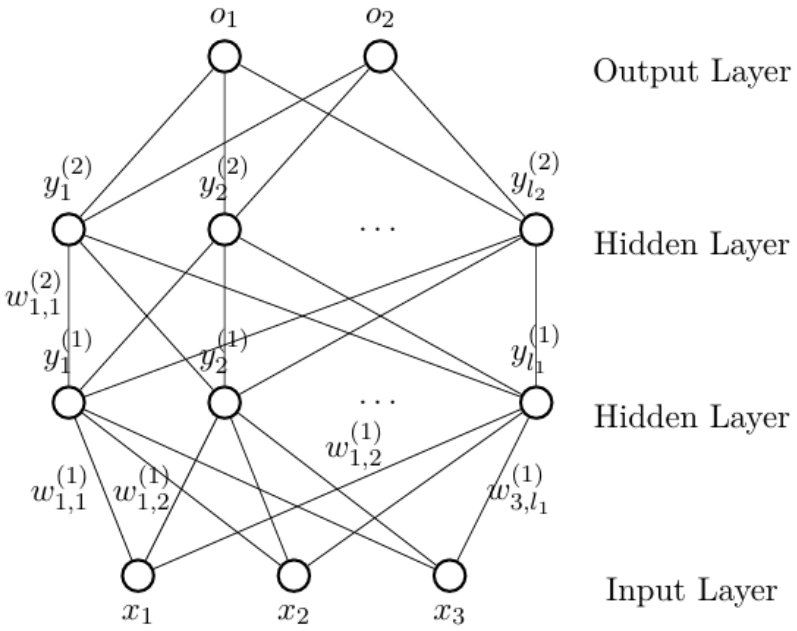

Рисунок 4 – DNN с 2-мя скрытыми слоями

Как показано ранее в работе, распознавание речи является последовательной задачей. Входные данные представляют собой типичные характеристики нескольких временных шагов. Обычная DNN будет полностью подключена между слоями и, таким образом, научится аффинному преобразованию всего временного контекста. Но интуитивно понятно, что существуют преобразования, которые можно изучить в узком контексте. Более глубокие слои могут затем изучить более широкий контекст из активаций скрытых активаций. Для достижения этой цели Вайбель и др. представили нейронную сеть с задержкой во времени (TDNN). Слои TDNN не связаны полностью, но каждый нейрон получает входные данные только из небольшого временного контекста. [11]

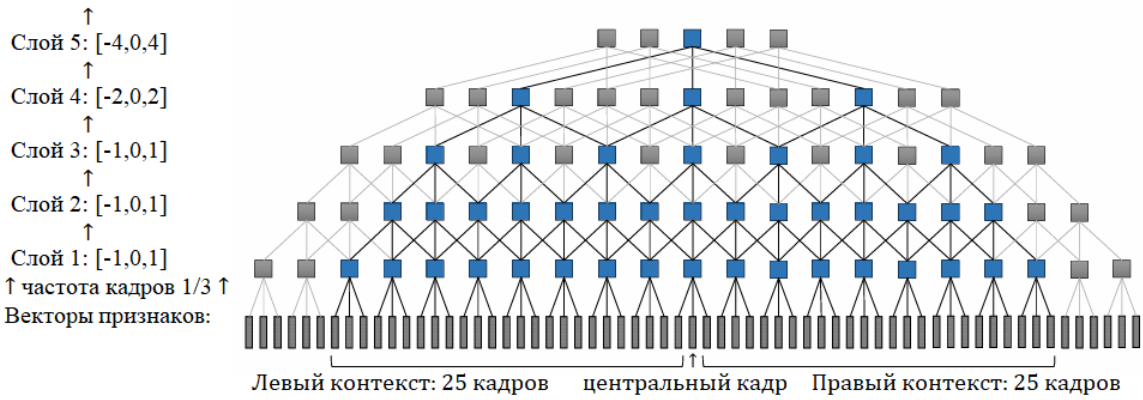

Рисунок 5 – Архитектура нейронной сети с глубокой временной задержкой (Deep TDNN). Цифры в квадратных скобках обозначают задержки кадров на входе для каждого слоя. Каждый квадратный блок представляет собой структурный блок нейронной сети. Сплошные черные линии и синие квадраты указывают путь для создания одного выходного кадра. Пунктирные линии и графические квадраты обозначают соединения и строительные блоки прошлых и будущих рамок. [9]

Кроме того, предполагается, что узкие объекты не зависят от времени, а веса нейронов в одном и том же слое, но с разными временными шагами, являются общими. Сходство со сверткой также дало сети название сверточная нейронная сеть (CNN).

2.1.3 Свёрточные нейронные сети (CNN)

Для повышения скорости распознавания речи необходимо преодолеть разнообразие речевых сигналов, в том числе разнообразие говорящих и разнообразие окружающей среды. Сверточная нейронная сеть обеспечивает трансляционно-инвариантную свертку во времени и пространстве. Если идея сверточной нейронной сети применяется к акустическому моделированию распознавания речи, инвариантность свертки может быть использована для преодоления разнообразия самого речевого сигнала.

Сверточная нейронная сеть сопоставляет части сигнала вместо того, чтобы рассматривать весь сигнал в виде пикселей, поскольку компьютеру становится трудно идентифицировать сигнал, когда рассматривается весь

набор пикселей. Математическая основа сопоставления – фильтрация. Это делается путем рассмотрения объекта, который связан с этим сигналом исправления, а затем один за другим сравниваются пиксели и умножаются друг на друга, а затем суммируются и делятся на общее количество пикселей. Этот шаг повторяется для всех рассматриваемых пикселей.

Сверточная нейронная сеть (CNN) – это тип NN, в котором входные переменные пространственно связаны друг с другом. Для учета очень важных пространственных положений были разработаны CNN. Они способны не только обнаруживать общие пространственные зависимости, но и распознавать конкретные закономерности. Общие веса, представляющие различные закономерности, улучшают сходимость за счет значительного сокращения числа параметров. CNN распознает небольшие закономерности на каждом уровне, обобщая их (обнаруживая более сложные закономерности более высокого порядка) на последующих уровнях. Это позволяет обнаруживать различные закономерности и сокращает количество весов, которые необходимо изучить, до минимума. [10]

2.1.4 Гибридные модели и трансформеры

Гибридные модели в распознавании речи представляют собой архитектуры, которые эффективно комбинируют преимущества рекуррентных нейронных сетей (RNN), сверточных нейронных сетей (CNN) и трансформеров. Ни одна из этих архитектур не является универсально лучшей для всех задач распознавания речи; каждая имеет свои сильные и слабые стороны. Гибридные модели стремятся преодолеть ограничения отдельных архитектур, комбинируя их способности для достижения более высокой точности, эффективности и устойчивости к шуму.

Типичные архитектуры гибридных моделей:

1. CNN + RNN/LSTM: CNN используется для извлечения локальных признаков из спектрограммы, а затем RNN/LSTM обрабатывает результирующие последовательности признаков, улавливая долгосрочные зависимости.
2. CNN + Transformer: CNN извлекает признаки, а трансформер обрабатывает эти признаки, используя механизм внимания для моделирования зависимостей.
3. RNN/LSTM + Transformer: RNN/LSTM обрабатывают входную последовательность, а затем трансформер использует эти промежуточные представления для моделирования глобальных зависимостей.
4. CNN + RNN/LSTM + Transformer: Сочетание всех трех архитектур, где CNN извлекает локальные признаки, RNN/LSTM обрабатывают временные зависимости, а трансформер моделирует глобальные зависимости.

2.2 Обучение нейронных сетей для распознавания речи

Обучение нейронных сетей для распознавания речи — сложный и ресурсоемкий процесс, требующий больших объемов данных и вычислительных мощностей. В основе обучения лежит принцип обратного распространения ошибки, который позволяет корректировать веса нейронной сети на основе разницы между её предсказаниями и истинными значениями.

Важным этапом обучения является подготовка данных. Необходимо собрать большой корпус аудиоданных, транскрибировать их (преобразовать в текстовую форму) и разделить на обучающую, валидационную и тестовую выборки. Качество транскрипции имеет первостепенное значение; ошибки в транскрипции могут привести к обучению некорректной модели. Данные должны быть разнообразными, чтобы модель могла обобщать на различные голоса, акценты, условия записи и стили речи. Предобработка аудиоданных

также играет ключевую роль: фильтрация шума, нормализация и извлечение признаков (MFCC, спектрограммы и др.) – всё это влияет на качество обучения.

2.2.1 Функции активации

Очень важной для процесса обучения является функция активации $\varphi(x)$ нейрона. Обычно она выглядит как пошаговая функция Хевисайда и выдает значения в диапазоне [-1,1] или [0,1]. Если $\varphi(x)$ нелинейно, то можно показать, что сети, использующие его, по крайней мере, с одним скрытым слоем, могут аппроксимировать любую функцию, что делает их очень мощными. Дополнительные слои с линейной функцией активации могут быть оптимизированы путем изменения веса и смещения соответственно, и поэтому они малопригодны.

Сигмовидная функция является очень распространенной функцией активации. Выходные данные находятся в диапазоне от 0 до 1, что позволяет интерпретировать их как вероятность. Однако использование сигмовидной активации в глубоких сетях может усугубить проблему исчезающего градиента.

$$\sigma(x) = \frac{1}{1 + e^{-\beta x}}$$

Гиперболический тангенс выдает выходные данные в интервале [-1,1]. Хотя это, по сути, просто масштабированная версия сигмовидной функции, ее производная также больше примерно при $x = 0$.

$$tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = 1 - \frac{2}{e^{2x} + 1}$$

Rectified Linear Unit (ReLU) более правдоподобна, чем две предыдущие. Ее вычисление также очень простое, что немного ускоряет обучение. Линейная часть всегда имеет производную от 1 и не усиливает проблему

исчезающего градиента. При отрицательных входных данных выходные данные и производная всегда равны 0. Неудачные изменения веса приводят к этому для всех выборок в обучающих данных. В этом случае нейрон фактически не работает.

$$relu(x) = max(0, x)$$

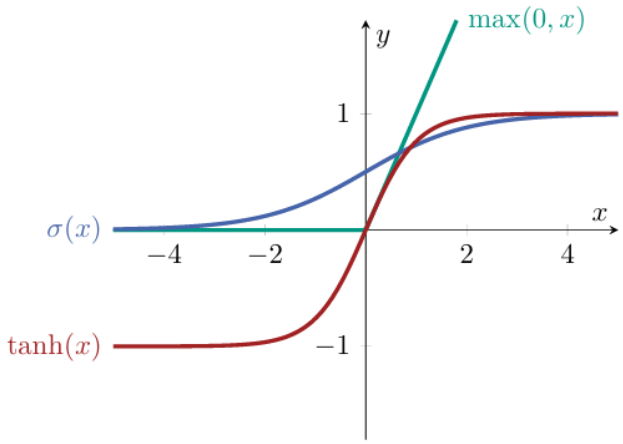

Рисунок 6 –Сигмовидная функция σ(x), гиперболический тангенс $tanh(x)$ и Rectified Linear Unit $relu(x)$.

Другой важной функцией активации является функция Softmax. Она является обобщением сигмовидной функции и работает с вектором реальных значений, а не с единичным масштабом. Входным вектором могут быть взвешенные активации предыдущего уровня. Результатом будет вектор, который можно интерпретировать как распределение вероятностей.

$$softmax(x_i) = \frac{e_i^x}{\sum_j e_j^x}.$$

2.2.2 Оценка модели

Процесс обучения включает в себя многократное представление обучающих данных модели и корректировку весов с помощью выбранного оптимизатора и функции потерь. Обычно обучение выполняется на мощных графических процессорах (GPU) или специализированных ускорителях (TPU)

из-за высокой вычислительной сложности. Мониторинг процесса обучения с помощью валидационной выборки помогает избежать переобучения (overfitting), когда модель хорошо работает на обучающих данных, но плохо обобщает на новые данные.

После обучения модель оценивается на тестовой выборке, чтобы оценить её обобщающую способность. В распознавании речи часто используется Word Error Rate (WER) — процент ошибок в распознанном тексте по сравнению с эталонным.

2.2.3 Пример обучения сверточной нейронной сети

В статье [2] была разработана нейросеть, различающая произносимые слова. На основе базы данных речевых команд из Google было создано хранилище, которое содержит около 71000 звуковых дорожек с проиндексированными метками, распределенные по папкам.

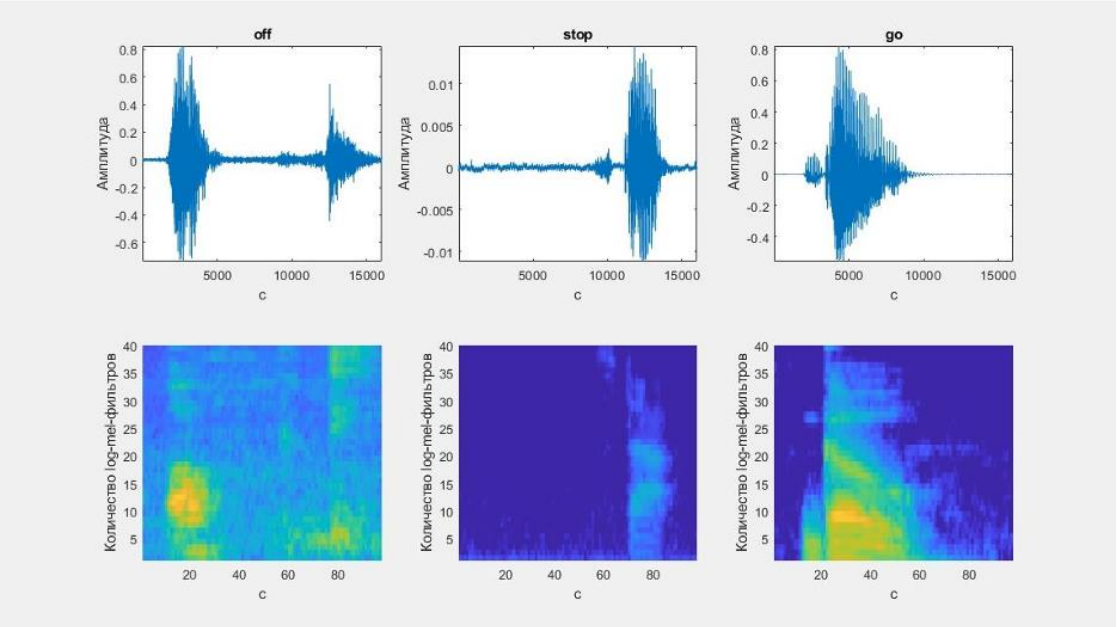

Рисунок 7 – Формы волны и спектрограммы нескольких обучающих примеров

В качестве варианта обучения был использован оптимизатор Adam (adaptive moment estimation - оптимизационный алгоритм) с размером мини пакета 128. Adam – это метод адаптивной скорости обучения, то есть он

рассчитывает индивидуальные скорости обучения для различных параметров. Его название происходит от адаптивной оценки моментов, и причина, по которой его так называют, заключается в том, что Адам использует оценки первого и второго моментов градиента, чтобы адаптировать скорость обучения для каждого веса нейронной сети. Тренировочный цикл длится в течение 25 полных проходов через весь набор тренировок и скорость обучения уменьшится в 10 раз после 20 прохода (Рисунки 8, 9).

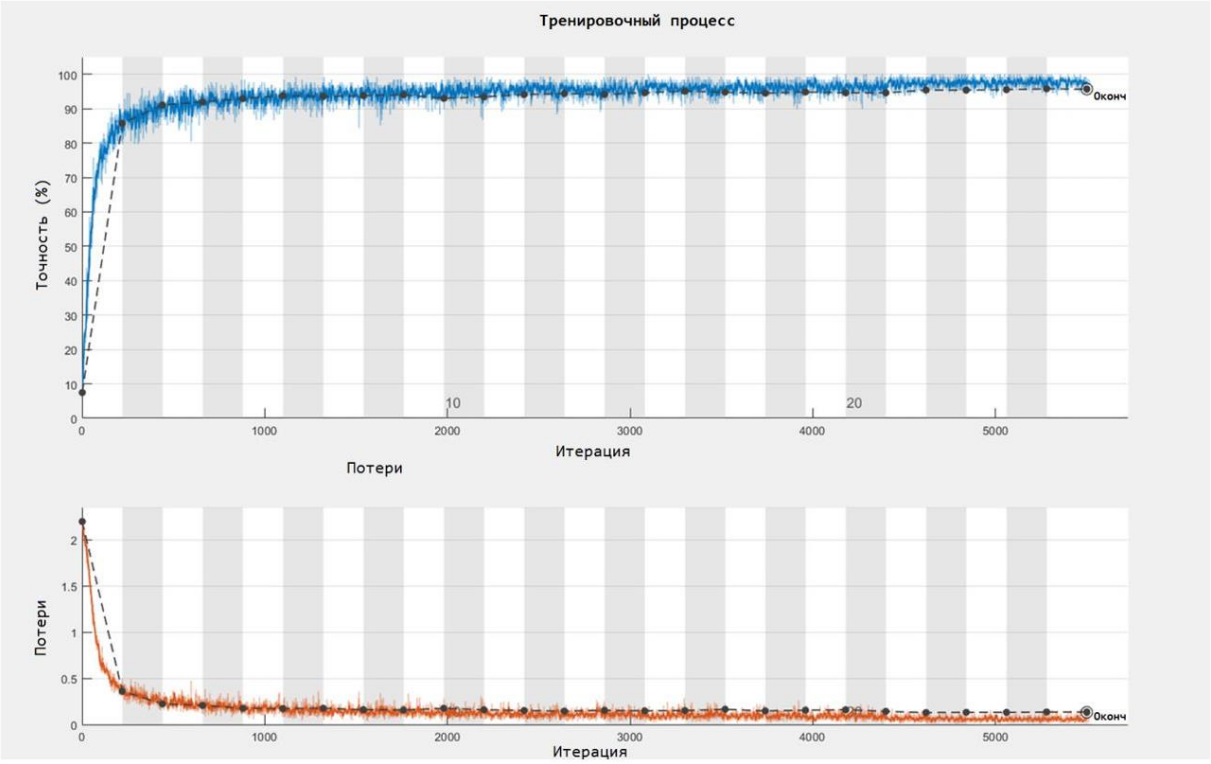

Рисунок 8 – Процесс обучения

Рисунок 9 – Обозначения и параметры обучения

После обучения слои нейросети выглядят следующим образом:

```
21x1 Массив слоев со слоями:

  1   'imageinput'    Image Input            40x98x1 изображения с нормализацией
  2   'conv_1'        Convolution            12 3x3x1 свертки с шагом [1 1] и дополнением "то же самое"
  3   'batchnorm_1'   Batch Normalization    Пакетная нормализация с 12 каналами
  4   'relu_1'        ReLU                   ReLU
  5   'maxpool_1'     Max Pooling            2x2 максимальное объединение с шагом [2 2] и дополнением 'то же самое'
  6   'conv_2'        Convolution            24 3x3x12 свертки с шагом [1 1] и дополнением "то же самое"
  7   'batchnorm_2'   Batch Normalization    Пакетная нормализация с 24 каналами
  8   'relu_2'        ReLU                   ReLU
  9   'maxpool_2'     Max Pooling            2x2 максимальное объединение с шагом [2 2] и дополнением 'то же самое'
 10   'conv_3'        Convolution            48 3x3x24 свертки с шагом [1 1] и дополнением "то же самое"
 11   'batchnorm_3'   Batch Normalization    Пакетная нормализация с 48 каналами
 12   'relu_3'        ReLU                   ReLU
 13   'maxpool_3'     Max Pooling            2x2 максимальное объединение с шагом [2 2] и дополнением 'то же самое'
 14   'conv_4'        Convolution            48 3x3x48 свертки с шагом [1 1] и дополнением "то же самое"
 15   'batchnorm_4'   Batch Normalization    Пакетная нормализация с 48 каналами
 16   'relu_4'        ReLU                   ReLU
 17   'maxpool_4'     Max Pooling            1x13 максимальное объединение с шагом [1 1] и заполнением [0 0 0 0]
 18   'dropout'       Dropout                исключение 20%
 19   'fc'            Fully Connected        12 полностью связанных слоев
 20   'softmax'       Softmax                softmax
 21   'classoutput'   Classification Output   Взвешенная перекрестная энтропия
```

Рисунок 10 – Описание слоев нейросети

Ошибка сети на тренировочном наборе (без дополнения данных) составляет 1.3806% и на проверочном наборе 4.2999%. Таким образом, точность нейросети составляет 95.70%. Точность распознавания может быть улучшена за счет увеличения словарного запаса, подразумевающего предоставления большего количества аудиофайлов для тренировки.

3. Современные достижения и тенденции

Сейчас активно исследуются новые модели и гибридные варианты для извлечения голоса. Две модели однонаправленных нейронных сетей – LSTM с временной задержкой (TDL STM) и параллельной архитектуры LSTM с временной задержкой (PTDLSTM). Обе модели нейронной сети кодера продемонстрировали улучшенные результаты ASR по сравнению с глубокими LSTM и TDNN LST-модель, аналогичная реализации Google и Kaldi, использует три задачи ASR разного объема (от 80 до 960 часов для обработки

данных) и языка (английский и китайский). Среднее относительное увеличение частоты ошибок в слове/символе в модели PTDLSTM составляет 13,2% и 11,0% соответственно, по сравнению с базовыми моделями LSTM и TDNN-LSTM. PTDLSTM также продемонстрировал более высокую частоту ошибок по сравнению с BLSTM с регулируемой задержкой аналогичного размера и с задержкой, которая в данном случае составляет 250 мс, а также улучшил скорость обучения и вывода за счет исключения обратной связи. LSTM-вычисление из перекрывающихся фрагментов.

Компания Яндекс представила свою модель SpeechKit в 2013 году, это сервис транскрибации от Яндекса, который поддерживает 16 языков для озвучки для создания голосовых помощников и автоматизации колл-центров. Нейронная сеть для акустического моделирования тренируется в несколько этапов. Для инициализации нейросети используется стек из ограниченных машин Больцмана (Restricted Boltzmann Machines, RBM). RBM — это стохастическая нейросеть, которая тренируется без учителя. Хотя выученные ей веса нельзя напрямую использовать для различения между классами акустических событий, они детально отражают структуру речи.

Разрабатываются новые методы оптимизации и регуляризации, позволяющие улучшить обобщающую способность моделей и снизить переобучение. Использование больших объемов данных и методов предобучения (например, transfer learning) значительно повышает точность. Разрабатываются специализированные модули и методы для эффективного подавления шума и артефактов в аудиозаписях, повышая качество распознавания в реальных условиях.

ЗАКЛЮЧЕНИЕ

Автоматическое распознавание речи (АСР) — это одна из самых захватывающих и быстро развивающихся областей искусственного интеллекта. В последние годы мы стали свидетелями значительных изменений в подходах к этой задаче, и основным двигателем этих изменений стали нейронные сети.

Ранее системы АСР базировались на скрытых марковских моделях (HMM), которые использовали статистические методы для обработки звуковых сигналов. Эти модели были достаточно эффективными, но имели свои ограничения. Они плохо справлялись с различными акцентами, диалектами и шумами, а также требовали значительных усилий для настройки и обучения.

С появлением нейронных сетей и, в частности, глубокого обучения, подходы к распознаванию речи претерпели революцию. Нейронные сети, такие как свёрточные нейронные сети (CNN), рекуррентные нейронные сети (RNN) и трансформеры, обеспечили качественный скачок в точности распознавания. Гибридные модели, которые объединяют преимущества этих архитектур, становятся стандартом в современных системах АСР. Благодаря сложным архитектурам, современные системы могут достигать высокой точности распознавания даже в сложных условиях. Современные модели способны распознавать речь на нескольких языках, что делает их универсальными инструментами для глобального использования.

Несмотря на впечатляющие успехи, исследователи сталкиваются с рядом новых проблем. Одной из главных задач остается повышение устойчивости систем к различным типам шума и помех. Например, шумы в общественном транспорте или разговоры нескольких людей одновременно могут значительно ухудшить качество распознавания.

Также существует необходимость в разработке эффективных методов распознавания речи, которые могли бы работать на устройствах с ограниченными ресурсами.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Лань, Г. Синтез фрагментов голоса человека на основе реконструкции частотных спектров / Г. Лань, А. С. Фадеев, А. Н. Моргунов. — Текст : непосредственный // Доклады Томского государственного университета систем управления и радиоэлектроники. — 2021. — № 24. — С. 16-17.

2 Романюк А. Г, Смирнов А.Н., Антонова В.М. Использование глубокого обучения нейросети для распознавания голосовых команд пользователя. Журнал радиоэлектроники [электронный журнал]. 2019. № 11. Режим доступа: http://jre.cplire.ru/jre/nov19/18/text.pdf. DOI 10.30898/1684-1719.2019.11.18

3 Цыбульский А.С. ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ // Международный студенческий научный вестник. – 2017. – № 6.; URL: https://eduherald.ru/ru/article/view?id=17986 (дата обращения: 30.11.2024).

4 Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview / YU CHONGCHONG, KANG MENG, CHEN YUNBING [и др.]. — Текст : непосредственный // IEEE Access. — 2020. — № 8. — С. 163829 - 163843.

5 CLR Phonetics Lab. — Текст : электронный // University of Aizu : [сайт]. — URL: http://clrlab1.u-aizu.ac.jp/index.html (дата обращения: 30.11.2024).

6 Dawood, K. B. How Speech Recognition AI Works and What It Is Used For / K. B. Dawood. — Текст : электронный // folio3.ai : [сайт]. — URL: https://www.folio3.ai/blog/speech-recognition-ai/ (дата обращения: 30.11.2024).

7 Ilias, Papastratis Speech Recognition: a review of the different deep learning approaches / Papastratis Ilias. — Текст : электронный // AI

Summer : [сайт]. — URL: https://theaisummer.com/speech-recognition/#rnns-limitations-and-solutions (дата обращения: 09.12.2024).

8 León González D. J. Deconvolución en audio utilizando modelos basados en Machine Deep Learning. – 2021.

9 Moritz, Niko & Hori, Takaaki & Le Roux, Jonathan. (2019). Unidirectional Neural Network Architectures for End-to-End Automatic Speech Recognition. 76-80. 10.21437/Interspeech.2019-2837.

10 Nagajyothi, D. Speech Recognition Using Convolutional Neural Networks / D. Nagajyothi, P. Siddaiah. — Текст : непосредственный // International Journal of Engineering & Technology. — 2017. — № 7(4). — С. 133-137.

11 Ritter M. Neural Network Architectures for Reverberated Lecture Speech Recognition : дис. – Informatics Institute, 2016.

12 Toshiba Распознавание речи: очень краткий вводный курс / Toshiba. — Текст : электронный // Хабр : [сайт]. — URL: https://habr.com/ru/companies/toshibarus/articles/490732/ (дата обращения: 09.12.2024).