Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н. Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ ФМОП «Факультет Международных Образовательных Программ»

КАФЕДРА ИУ-7 «Программное обеспечение эвм и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Классификация методов обнаружения образцов голоса, синтезированных с помощью нейронных сетей»

Студент ИУ7И-74Б _____ Ахмад Халид Каримзай

(Подпись, дата)

Руководитель _____ А.С. Кострицкий

(Подпись, дата)

2023 г.

РЕФЕРАТ

Расчетно–пояснительная записка **??** с., **??** рис., **??** табл., 0 ист, 1 прил.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей расчетно-пояснительной записке применяют следующие термины с соответствующими определениями.

1) GMM – Gaussian Mixture Model.

2) MFCC – Mel-frequency cepstral coefficients.

3) LFCC – Linear Frequency Cepstral Coefficients.

4) FFT – Fast Fourier Transform.

5) DCT – Discrete cosine transform.

6) CNN – Convolutional Neural Network.

7) DeepFake – Фейковый контент, созданный с помощью нейронных сетей.

ВВЕДЕНИЕ

Аудио-дипфейки представляют собой категорию звуковых файлов, созданных при помощи глубоких нейронных сетей, способных анализировать и воспроизводить звуковые характеристики настолько реалистично, что созданный контент может звучать естественно и непринужденно. Чаще всего эти технологии применяются для имитации голосов людей, но, кроме того, могут вызывать веселье и забаву. Тем не менее, с увеличением популярности аудио-дипфейков возникают вопросы относительно их злоупотребления с целью распространения дезинформации.

Хотя синтетический или фейковый контент существует уже много лет, внимание к контенту, созданному с использованием нейронных сетей, также известного как дипфейк, стало значительным лишь в последние несколько лет. В то время как синтезированные фотографии и видео, порожденные нейронным сетям, привлекли большое внимание, синтетические человеческие голоса тоже достигли выдающегося качества и эффективности. Но, несмотря на их улучшенную реалистичность и доступность, синтетические голоса также несут в себе существенные риски.

В рамках данной научной-исследовательской работы рассмотрим следующие цели и задачи:

1. Синтезирование аудио: Понятие и Типы;
2. Характеристики и особенности аудиоматериала для изучения;
3. Понятие и схема работы системы обнаружения синтетического звука;
4. Классификация и обзор известных методов для обнаружения синтезированного звука.

1 Синтезирование голос

Под термином ”Синтезирование голоса”обычно понимается любой аудиосигнал, важные характеристики которого были изменены при помощи технологий нейронных сетей, сохраняя при этом воспринимаемую естественность. Ранее проведенные исследования в основном выделяли пять видов дипфейкового звука:

1. преобразование текста в речь;
2. преобразование голоса;
3. подделка эмоций;
4. подделка сцен;
5. частично подделка.

также в таблице (**??**), проведено классификация аудио дипфейков по способу генерации, где в первой столбце таблице представлены поддельные типы дипфейков, во второй столбце - поддельные черты, в третей столбце - поддельные продолжительности то есть частично или полностью синтезирован, в четвертом столбце - с помощью нейронной сети указывается что применяемый метод реализован ли с помощью нейронных сетей.

1.1. Преобразование текста в речь

Преобразование текста в речь (TTS) представляет собой широко применяемую технологию, ориентированную на синтез четкой и естественной речи из произвольного текста с использованием моделей, основанных на методах машинного обучения. Современные модели TTS в основном используют глубокие нейронные сети для генерации реалистичной речи, которая максимально приближена к человеческой.

Системы TTS обычно включают в себя два основных модуля: модуль анализа текста и модуль генерации речевых сигналов. Модуль анализа текста разбирает входной текст, определяя тон, интонацию и другие лингвистические

аспекты, необходимые для правильной передачи смысла. Затем модуль генерации речевых сигналов создает звуковую последовательность, соответствующую заданному тексту.

1.2. Преобразование голоса

Преобразование голоса (VC) [**wu2015spoofing**], или клонирование голоса в цифровой форме, фокусируется на изменении звучания речи одного говорящего, подражая тембру и просодии другого говорящего, при этом сохраняя содержание оригинального высказывания. Процесс введения в систему VC обычно включает в себя использование естественных высказываний данного говорящего как входных данных.

Существует несколько основных подходов к технологиям VC [**sisman2020overview**], включая статистический параметрический, частотное искажение и выбор единиц измерения. В частности, статистическая параметрическая модель включает в себя вокодер, аналогичный тому, который используется в статистических параметрических системах синтеза речи (TTS).

1.3. Подделка эмоций

Подделка эмоций [**zhao2022emofake**], также известная как модификация эмоционального тонуса, представляет собой технологию, направленную на изменение акустических характеристик звука с целью создания впечатления изменения эмоционального состояния говорящего. Эта методика фокусируется на манипуляции параметрами, такими как тембр, интонация и темп речи, сохраняя при этом остальные аспекты звуковой информации, такие как личность говорящего и содержание высказывания.

1.4. Подделка сцен

Модификация сцены звучания, более известная как подделка сцены [**yi2022scenefake**], представляет собой метод, направленный на сопоставление акустической обстановки оригинального высказывания с другой звуковой сценой, используя технологии улучшения речи. В этом процессе сохраняются как личность говорящего, так и содержание высказывания, при этом происходит изменение окружающей аудиообстановки.

1.5. Частично подделка

Частичная подделка [**yi2021half**], также известная как модификация части высказывания, представляет собой технику, прицельно изменяющую всего лишь несколько слов в оригинальном высказывании. Этот метод осуществляется путем манипулирования исходными аудиоклипами с использованием подлинных или созданных синтезом звуковых фрагментов. Однако при этом ключевым аспектом является сохранение неизменной личности говорящего.

Таблица 1.1. Класификации аудио дипфейков по способу генерации

Поддельный тип	Поддельная черта	Поддельная продолжительность	С помощью нейронной сети
Преобразование текста в речь	Личность спикера, Речевое содержание	полностью	да
Преобразование голоса	Личность спикера	полностью	да
Подделка эмоций	эмоция спикера	полностью	да
Подделка сцен	Акустическая сцена	полностью	да
Частично подделка	Речевое содержание	частично	да

2 Отличительные признаки аудио для изучения

Извлечение признаков представляет собой ключевой модуль классификатора аудиодипфейков. Основной целью этого процесса является изучение характерных особенностей путем выделения акустических артефактов из речевых сигналов, которые могут свидетельствовать о наличии поддельных атак. Большое количество исследований подчеркнуло важность определения полезных признаков для эффективного обнаружения дипфейков.

В данной области уделено значительное внимание выявлению полезных функций, способных обнаруживать характерные аспекты поддельных атак. Признаки, использованные в проведенных исследованиях, условно могут быть разделены на три основные категории [**sahidullah2015comparison**]:

– Спектральные характеристики;

– Просодические характеристики;

– Глубокие характеристики.

2.1. Спектральные характеристики

Спектральные характеристики в анализе звука относятся к характеристикам, которые отражают распределение энергии по различным частотам в сигнале. Эти характеристики вычисляются с использованием математических преобразований, таких как быстрое преобразование Фурье (FFT), и имеют критическое значение для извлечения существенной информации из аудиосигналов для различных применений. Спектральные характеристики могут быть классифицированы на краткосрочные и долгосрочные в зависимости от временного масштаба, в течение которого они вычисляются.

2.1.1. Кратковременные спектральные особенности

Кратковременные спектральные характеристики, извлеченные из коротких кадров обычно длительностью 20-30 мс, описывают кратковременную спектральную огибающую, которая включает в себя акустический коррелят тембра голоса. Кратковременные спектральные характеристики вычисляются, главным образом, путем применения кратковременного преобразования Фурье (STFT)

к речевому сигналу [**xiao2015spoofing**]. При предположении, что речевой сигнал x(t) квазистационарен в течение короткого периода, STFT формулируется следующим образом:

$$X(t,\omega) = |X(t,\omega)|e^{j\phi(\omega)} \tag{2.1}$$

где $|X(t,\omega)|$, это спектр магнитуд а $\phi(\omega)$ представляет собой фазовый спектр в кадре t и частотный диапазон ω. Спектр мощности определяется как $|X(t,\omega)|^2$.

Кратковременные спектральные характеристики в основном включают кратковременные характеристики, основанные на магнитуде и фазе. Обычно несколько характеристик, базирующихся на магнитуде, напрямую производятся из спектра магнитуды, но большинство из них вычисляются из спектра мощности. Характеристики, основанные на фазе, извлекаются из фазового спектра.

Спектр магнитуды представляет собой график амплитуд сигнала в зависимости от его частоты, тогда как спектр мощности включает в себя квадрат амплитуд, отражающий распределение энергии. Кратковременные характеристики, основанные на магнитуде, могут включать в себя параметры, такие как частоты формант, амплитудные спектральные величины и мел-частотные кепстральные коэффициенты (MFCC).

Характеристики, вычисленные на основе фазы, включают в себя информацию о временных задержках и относительных фазовых сдвигах между различными частотами. Эти параметры могут быть полезными при анализе изменений во времени и при восстановлении оригинального сигнала.

2.1.2. Долгосрочные спектральные особенности

Кратковременные спектральные признаки не очень хорошо передают временные характеристики траекторий речевых признаков из-за того, что они вычисляются покадрово [**wu2013synthetic**]. Поэтому были предложены долгосрочные спектральные характеристики для получения информации на более широких временных интервалах из речевых сигналов, и исследования показали, что они имеют решающее значение для обнаружения поддельной речи.

Долгосрочные спектральные характеристики охватывают более продолжительные временные участки речевых сигналов и могут лучше отражать долгосрочные изменения в акустических свойствах речи. Это позволяет системе

обнаружения более эффективно анализировать речевые траектории и выявлять особенности, связанные с поддельной речью. Такие характеристики могут включать в себя долгосрочные форманты, изменения в спектральной энергии на протяжении времени и другие параметры, охватывающие более широкие аспекты речевого сигнала [**das2019long**].

2.2. Просодические характеристики

Просодия относится к несегментарной информации в речевых сигналах, включая ударение на слоге, интонационные паттерны, темп речи и ритм [**kinnunen2010overview**]. В отличие от кратковременных спектральных характеристик с короткой продолжительностью, обычно составляющей 20-30 мс, просодические особенности охватывают более длинные сегменты, такие как фоны, слоги, слова, высказывания и т.д. Важные просодические параметры включают основную частоту (F0), длительность, распределение энергии, скорость разговора и т.д. Предыдущие исследования [**kinnunen2010overview**] по обнаружению поддельного звука в основном рассматривали три основные просодические характеристики:

- Основная частота (F0): Показатель высоты голоса, который варьируется в зависимости от интонации и эмоционального состояния говорящего;
- Длительность: Временной параметр, отражающий продолжительность звуковых сегментов, таких как слоги и слова, в речевом потоке;
- Распределение энергии: Характеризует энергетические аспекты речи и может отражать эмфатическое выделение или особенности интонации.

Эти просодические характеристики могут быть использованы для выделения особенностей в произношении, что делает их важными для обнаружения поддельного звука, где подделка может влиять на натуральные просодические паттерны голоса. Однако они менее чувствительны к канальным эффектам по сравнению со спектральными функциями [**wu2015spoofing**]. Они могут предоставлять дополнительную информацию к спектральным функциям для повышения эффективности обнаружения поддельного звука.

2.3. Глубокие характеристики

Упомянутые выше спектральные характеристики и просодические характеристики представляют собой большую часть ручных функций с сильными и желательными репрезентативными способностями. Однако их конструирование подвержено предвзятостям из-за ограничений представлений, внесенных человеческими разработчиками [**zeghidour2021leaf**]. Таким образом, глубокие функции вдохновлены для заполнения этого пробела. Глубокие функции изучаются с использованием глубоких нейронных сетей, которые условно можно разделить на три категории: обучаемые спектральные функции, контролируемые функции встраивания и самоконтролируемые функции встраивания.

На рисунке (**??**) представлено, класификации аудио по признаками для обучение:

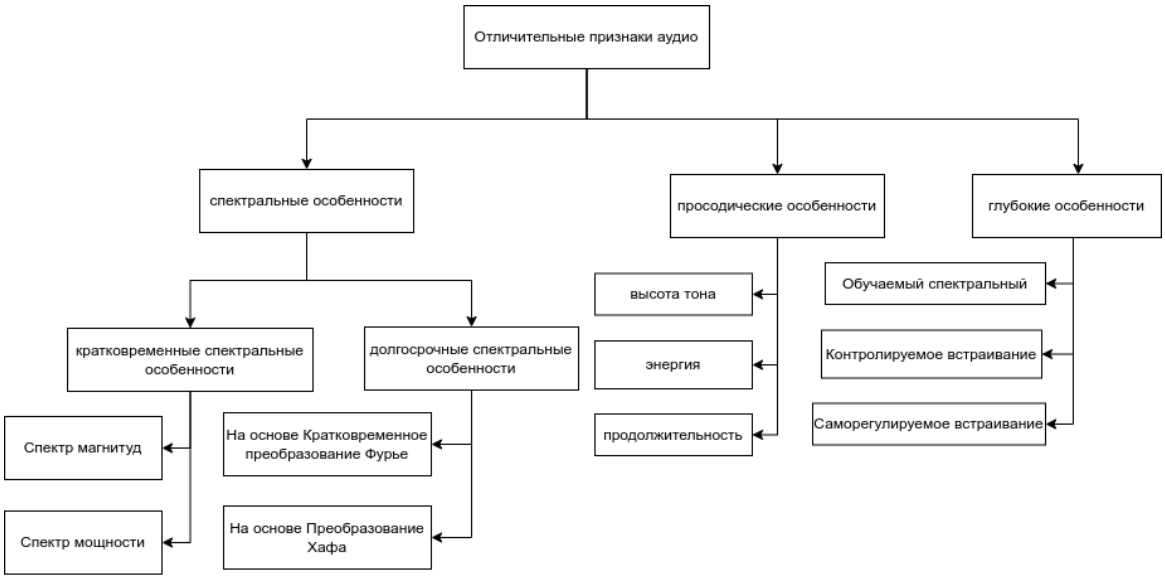

Рис. 2.1. Классификации аудио по признаками

3 Система обнаружения поддельного звука (англ. Audio deepfake detector)

В системах, используемых для обнаружения поддельного звука, важным фактором являются аудио-особенности для изучения, и внутренний классификатор играет ключевую роль в глубоком распознавании аудио. Целью является изучение высокоуровневого представления функций входного интерфейса и моделирование превосходных возможностей обнаружения. Внутренние классификаторы, которые часто используются для обнаружения аудио-дипфейков, в основном делятся на две категории:

– Статистические методы;

– С использованием глубоких нейронных сетей.

3.1. Статистические методы обнаружения Аудио Дипфейк

Под термином ”статистические методы”в статьях часто подразумевают алгоритмы машинного обучения, которые изучают характеристики аудио для решения задачи обнаружения аудиодипфейков. Для этого используются алгоритмы бинарной классификации.

В связи с этим для обнаружения фальшивой речи было использовано множество классических подходов к классификации паттернов. Самые популярные методы обнаружения аудиодипфейков с применением машинного обучения включают:

– Машина опорных векторов (англ. Support Vector Machine);

– Гауссовые модели смеси (англ. Gaussian mixture model).

3.1.1. Машина опорных векторов (SVM)

SVM - это контролируемый метод обучения, который основывается в основном на двух предположениях [**hamza2022deepfake**]:

1. Преобразование данных в многомерное пространство может свести сложные проблемы классификации со сложными поверхностями принятия решений к более мелким проблемам, которые могут быть решены путем их линейного разделения;

2. Только обучающие шаблоны вблизи поверхности принятия решений обеспечивают наиболее чувствительную детали для классификации.

Так как проблема обнаружения аудиодипфейков представляет собой бинарную классификацию с линейно разделяемыми векторами $x_i \in R^n$, в качестве поверхности принятия решения, используемой для классификации паттерна как принадлежащего к одному из двух классов, используется гиперплоскость H_0. Если x это случайный вектор $n \times R$, тогда мы определяем:

$$f_{(x)} = w.x + b$$

В формуле . это скалярное произведение, набор всех x-векторов, удовлетворяющих уравнению $f_{(x)} = 0$, обозначается как H_0. Предполагая две гиперплоскости, H_1 и H_2, расстояние между ними называется их границей, которую можно представить следующим образом:

$$\begin{cases} H_1 = \{x \in R^n | f_{(x)} > 0\} \\ H_2 = \{x \in R^n | f_{(x)} < 0\} \end{cases}$$

Гиперплоскость решения H_0 зависит от векторов, ближайших к двум параллельным гиперплоскостям, называемым опорными векторами. Запас должен быть максимальным, чтобы получить классификатор, который не очень адаптирован к обучающим данным.

На рисунке (**??**) представлено, машина опорных векторов:

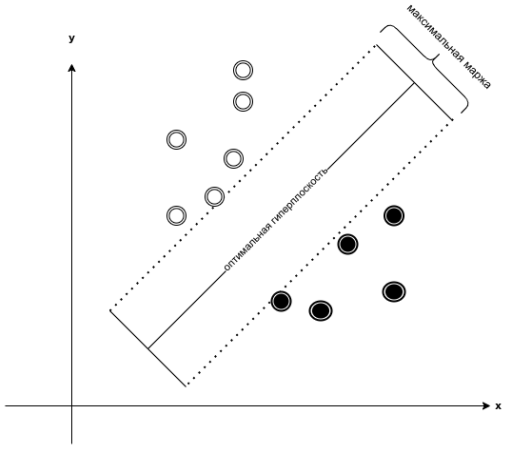

Рис. 3.1. Машина опорных векторов (SVM)

В многих статьях по классификации аудио-дипфейков где классификатор реализован на базе SVM. SVM отлично справляется с четким разделением

выборок и эффективен в средах с высокой размерностью. SVM использует подмножество точек обучения в функции принятия решения, что делает его эффективным с точки зрения использования памяти.

3.1.2. Гауссовые моделей смеси (GMM)

Модель смеси гауссовых распределений (GMM), как следует из названия, представляет собой смесь нескольких гауссовских распределений. Речевые признаки представлены в виде векторов в n-мерном пространстве. Распределение этих векторов признаков представлено смесью гауссовских плотностей.

Для n-мерного вектора признаков x функция плотности смеси для класса s с параметром модели λ^s определяется как [**JOTHILAKSHMI2016301**]:

$$p(x|\lambda^s) = \sum_{i=1}^{\mathcal{M}} \alpha_i^s f_i^s(x)$$

Функция плотности смеси представляет собой взвешенную линейную комбинацию $\mathcal{M}$ унимодальных гауссовских плотностей компонентов $f_i^s(.)$. Каждая функция гауссовской плотности $f_i^s(.)$ параметризуется вектором среднего μ_i^s и ковариационной матрицей Σ_i^s с использованием:

$$f_i^s(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_i^s|}} \exp\left(-\frac{1}{2}(x-\mu_i^s)^T \Sigma_i^{s-1}(x-\mu_i^s)\right)$$

где Σ_i^s является ковариационной матрицей, а $(\Sigma_i^s)^{-1}$ называется обратной ковариационной матрицей.

3.2. Методы с применением глубоких нейронных сетей

Классификаторы современных систем обнаружения аудиодипфейков в основном основаны на методах глубокого обучения, которые значительно превосходят классификаторы, основанные на SVM и GMM, благодаря их мощным возможностям моделирования [**godoy2015using**].

Каждый аудиосигнал может быть представлен на двумерном графике, построенном с использованием математических расчетов. Обработка аудиозаписей в нейронной сети требует большого объема вычислений для создания системы, способной обнаруживать аудиодипфейки с гораздо меньшими затратами вычислений. Это достигается путем преобразования аудиозаписей в изображения

звуковых объектов и последующего получения значений массива в числовом формате, наилучшим образом подходящих для передачи в качестве входных данных в нейронную сеть.

Существует конференция, где ученые из разных стран представляют свои решения по созданию и обнаружению синтезированного звука, она называется ASVSpoof и проводится каждые три года [**yamagishi2021asvspoof**]. В основном все решения, которые предлагают, основаны на методах глубокого обучения. Рассмотрим два метода, которые показали меньшую ошибку при классификации.

1. Сверточная нейронная сеть;
2. Глубокая нейронная сеть с использованием необработанных сигналов.

3.2.1. общая логика работы нейронных сетей

Первоначальное развитие этих сетей берет свое начало в работе Фрэнка Розенблатта о персептронах и начинается с определения нейрона [**DOU2023484**]. Математически нейрон - это нелинейность, применяемая к аффинной функции. Входные характеристики $x = (x_1, x_2, ..., x_n)$ передаются через аффинную функцию, составленную с нелинейностью ϕ:

$$T(x) = \phi\left(\sum_i W_i * x_i + b\right) = \phi(W * x + b)$$

с заданными весами W и смещением b. Схематично это представлено на рисунке (**??**). Типичной нелинейностью, или функцией активации, является сигмоидная форма, определяемая как:

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

Функция активации не обязательно должна быть только сигмоидной; она выбирается в зависимости от задачи, которую необходимо решить с использованием нейронных сетей.

На рисунке (**??**) представлено, схематическая версия нейрона:

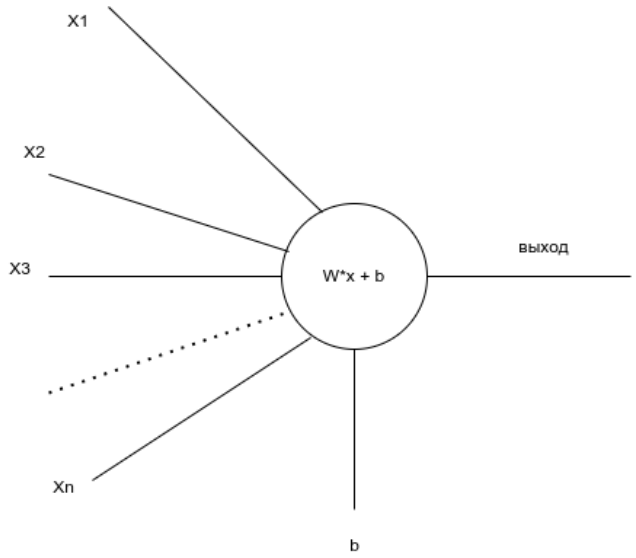

Рис. 3.2. Схематическая версия нейрона

Нейронную сеть можно смоделировать как набор нейронов, которые соединены в ациклический граф. То есть выходные данные некоторых нейронов становятся входными данными для других нейронов, и циклы, в которых выходные данные нейрона отображаются обратно на более ранний промежуточный вход, запрещены. Обычно такие нейроны организованы в слои нейронов. Такая сеть состоит из входного слоя, одного или нескольких скрытых слоев и выходного слоя. В отличие от скрытых слоев, выходной слой обычно не имеет функции активации.

В зависимости от задачи нейронные сети могут обучаться контролируемым или неконтролируемым способом и соответственно в нейронных сетей появится термин функция потерь, также называемая функцией стоимости или целевой функцией, играет ключевую роль в измерении несоответствия между прогнозируемым выходным сигналом сети и фактическими целевыми значениями. Основная цель на этапе обучения — минимизировать эту функцию потерь, поскольку это повышает точность модели при составлении прогнозов и для того что бы уменьшить потери используется обратное распространение ошибки и так представляется:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial T} * \frac{\partial T}{\partial Z} * \frac{\partial Z}{\partial W}$$

Где $\frac{\partial L}{\partial W}$ — это градиент функции потерь по весам W.

На рисунке (**??**) представлено, трехслойная нейронная сеть с тремя входами и двумя скрытыми слоями:

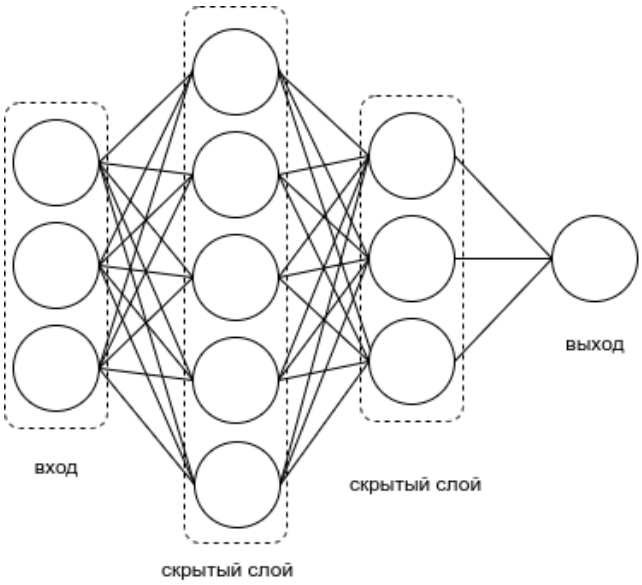

Рис. 3.3. Трехслойная нейронная сеть с тремя входами и двумя скрытыми слоями

3.2.2. Сверточная нейронная сеть

Сверточные нейронные сети (CNN) представляют собой частный случай нейронных сетей с прямой связью. В отличие от обычных нейронных сетей, слои CNN содержат нейроны, расположенные в нескольких измерениях: в каналах, ширине, высоте и количестве фильтров в простейшем двумерном случае.

Наиболее распространенными строительными блоками, с которыми сталкиваются в большинстве архитектур CNN, являются уровень свертки, уровень пула и полностью связанные уровни.

Свертки

Математическая свертка $(x * w)(a)$ функций x и w определяется во всех измерениях как:

$$(x * w)(a) = \int x(t) * w(t - a)\, da$$

где a находится в R^n для любого $n \geq 1$, а интеграл заменяется его многомерным вариантом. В терминологии сверточных нейронных сетей x называется входом, w называется фильтром или ядром, а выход часто называют активацией или картой признаков.

Нелинейности

Нелинейности необходимы для проектирования нейронных сетей: без них нейронная сеть будет вычислять линейную функцию своих входных данных, что является слишком ограничительным. Выбор нелинейности может оказать большое влияние на скорость обучения нейронной сети. Следовательно, часто после того, как мы получаем результаты из каждой свертки, мы применяем функцию активации.

Объединение слоев

Целью слоя объединения является создание сводной статистики его входных данных и уменьшение пространственных размеров карты объектов. Наиболее распространенной формой является максимальный пул, который использует шаг 2 вместе с размером ядра 2. Это соответствует пространственному разбиению карты объектов на регулярную сетку квадратных или кубических блоков со стороной 2 и взятию максимального или среднего значения по таким блокам для каждого входного объекта.

На рисунке (**??**) представлено, максимальный пул с шагом 2 вместе с размером ядра 2:

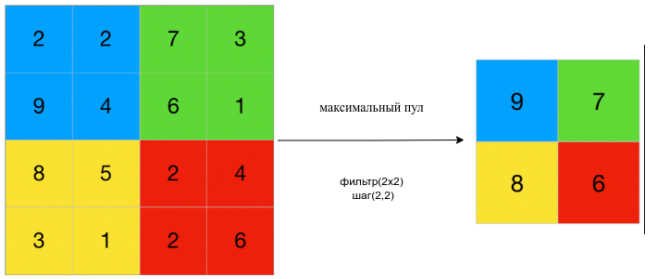

Рис. 3.4. максимальный пул с шагом 2 вместе с размером ядра 2

Полностью связанные слои

Полностью связанный слой с n входными измерениями и m выходными измерениями определяется следующим образом.

Выход слоя определяется параметрами: весовой матрицей $W \in M_{m,n}(R)$, имеющей m строк и n столбцов, и вектором смещения $b \in R^m$. Для данного входного вектора $x \in R^n$, выход полносвязного слоя FC с функцией активации f определяется как:

$$FC(x) := f(W \cdot x + b) \in R^m$$

В приведенной выше формуле $W \cdot x$ — это произведение матриц, а функция f применяется покомпонентно.

ЗАКЛЮЧЕНИЕ

ПРИЛОЖЕНИЕ А