РЕФЕРАТ

Расчетно-пояснительная записка к выпускной квалификационной работе «Метод замещения страниц в разделяемом кэш буфере Postgres с использованием нейронных сетей» содержит 89 страниц, 4 части, 27 рисунков, 0 таблиц и список используемых источников из 27 наименований.

Ключевые слова: замещение страниц, рекуррентные сети, Postgres.

Объект разработки – метод замещения страниц в разделяемом кэш буфере Postgres.

Цель работы: разработка метода замещения страниц в разделяемом кэш буфере Postgres с использованием нейронных сетей.

В первой части работы выполнен анализ существующих методов замещения страниц. Изучены принципы работы разделяемого кэш буфера PostgreSQL. Проведен сравнительный анализ нейронных сетей, подходящих для решения задачи. Сформулирована цель и формализована постановка задачи в виде IDEF0-диаграммы.

Во второй части разработан метод замещения страниц в разделяемом кэш буфере PostgreSQL. Описаны основные особенности предлагаемого метода. Сформулированы ограничения предметной области. Изложены основные этапы разрабатываемого метода в виде детализированной диаграммы IDEF0 и схем алгоритмов. Выделены функции и структуры СУБД PostgreSQL, которые используются при работе с кэш буфером.

В третьей части обоснован выбор программных средств реализации метода замещения страниц в разделяемом кэш буфере. Создана обучающая выборка с помощью логирования обращений к буферу. Приведены примеры работы программы. Описаны использованные методы тестирования программного обеспечения и приведены его результаты.

В четвертой части проведено исследование разработанного метода и выявлена зависимость коэффициентов попадания и совпадения от количества обращений к страницам на тестовой выборке. Проведено сравнение полученных результатов и значений этих метрик для существующих аналогов.

Разработанный метод замещения страниц может быть использован в СУБД Postgres. Использование метода позволит повысить коэффициент попадания в разделяемом кэш буфере, что должно привести к уменьшению времени отклика системы.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современные базы данных, такие как PostgreSQL, сталкиваются с постоянно растущими требованиями к производительности и эффективности управления ресурсами. Одним из ключевых аспектов работы базы данных является управление памятью, включая работу с разделяемым кэш буфером [1].

Загрузка данных с диска занимает гораздо больше времени, чем из оперативной памяти, поэтому современные системы управления базами данных используют область в оперативной памяти в качестве буфера для кэширования недавно просмотренных страниц, чтобы в будущем запросы к страницам в буфере выполнялись быстрее. Обычно буфер делится на части одинакового размера, где каждая часть может содержать страницу. Когда транзакция базы данных запрашивает страницу, которая в данный момент не хранится в буфере, она должна быть загружена в буфер. Если для кэширования этой страницы больше нет места, то одна из страниц в буфере должна быть вытеснена, чтобы освободить место для новой запрашиваемой страницы. Выбор такой страницы важен для уменьшения задержки доступа. Если все время для вытеснения будет выбираться страница, к которой в скором времени опять произойдет обращение, то производительность может ухудшиться до случая, когда данные в основном берутся с диска.

Для выбора страницы, которую надо исключить из буфера, применяются различные эвристические алгоритмы. Все эти алгоритмы являются приближением оптимального алгоритма и не учитывают структуру конкретной рабочей нагрузки. Если алгоритм замещения страниц будет учитывать особенности рабочей нагрузки, то число операций чтения и записи на диск может быть снижено, что приведет к повышению производительности системы.

Помимо систем управления базами данных методы замещения страниц используются в операционных системах, аппаратном и программным кэше, а также в других местах, где присутствует два типа памяти, один из которых меньше по объему и быстрее по скорости доступа.

Целью данной работы является разработка метода замещения страниц в разделяемом кэш буфере postgres с использованием нейронных сетей.

Для достижения поставленной цели требуется выполнить следующие задачи:

— сравнить существующие методы замещения страниц;

— описать и спроектировать метод замещения страниц с использованием нейронных сетей;

— разработать программное обеспечение для предложенного метода;

— провести сравнение разработанного метода с существующими аналогами по коэффициентам совпадения и попадания.

1 Аналитический раздел

1.1 Особенности управление памятью

1.1.1 Управление памятью в операционных системах

Виртуальная память представляет собой ключевую концепцию в управлении памятью современных компьютерных систем [2]. Она позволяет программам использовать объем оперативной памяти, превышающий физически доступный, за счет автоматического перемещения данных между основной памятью и вторичным хранилищем. Это достигается благодаря использованию виртуальных адресов, которые транслируются в физические адреса с помощью аппаратных средств.

Каждая программа работает с собственным адресным пространством, которое разбивается на страницы, представляющие собой непрерывные диапазоны адресов. Эти страницы не обязательно должны все одновременно находиться в оперативной памяти для выполнения программы, что позволяет эффективно использовать доступную память.

Когда программа обращается к данным, которые уже находятся в физической памяти, аппаратное обеспечение обеспечивает необходимое отображение адресов. Когда программа пытается получить доступ к странице, которая присутствует в виртуальном адресном пространстве, но отсутствует в физической памяти возникает системное прерывания отсутствия страницы после чего управление передается операционной системе.

Операционная система реагирует на ошибку отсутствия страницы, выбирая страницу при помощи алгоритма замещения и сбрасывая её содержимое на диск, если оно уже не находится там. Затем система извлекает нужную страницу с диска и помещает её в освободившееся место в памяти. После этого в таблицы вносятся соответствующие изменения, и прерванная команда выполняется заново.

Таблица страниц содержит сведения о каждой странице, включая номер страничного блока, который является ключевым элементом страничного отображения. Также в информации содержится бит присутствия-отсутствия, если он равен 1, запись активна и может быть использована, иначе соответствую-

щая виртуальная страница в данный момент отсутствует в памяти, и любое обращение к такой записи вызывает ошибку отсутствия страницы. Биты защиты указывают на тип доступа, который разрешен для страницы. В самом простом случае это один бит, который равен 0 для чтения-записи и 1 для только чтения. В более сложных системах могут быть использованы три бита, каждый из которых разрешает чтение, запись или исполнение страницы. Биты модификации и ссылки служат для отслеживания использования страницы. Бит модификации автоматически устанавливается при записи в страницу и помогает операционной системе определить, нужно ли сохранять страницу на диск при ее выгрузке из памяти. Бит ссылки устанавливается при любом обращении к странице и помогает операционной системе определить, какую страницу следует выгрузить при возникновении ошибки отсутствия страницы.

Большинство программ часто обращаются к ограниченному набору страниц. Из-за этого только небольшая часть записей в таблице страниц активно используется, а остальная практически не задействуется. Основываясь на этом наблюдении, для повышения производительности системы было предложены добавить в аппаратуру специальное устройство, которое называется TLB, и отвечает за трансляцию виртуальных адресов в физические для самых используемых страниц.

При использовании TLB существует 2 типа ошибок: программные и аппаратные. Программная ошибка возникает, когда страница отсутствует в TLB, но есть в памяти, и ее можно исправить простым обновлением TLB без обращения к диску. Это занимает 10-20 машинных команд и несколько наносекунд. Аппаратная ошибка возникает, когда страница отсутствует в памяти и требуется обращение к диску, что занимает несколько миллисекунд. Она обрабатывается значительно медленнее программной ошибки.

При возникновении ошибки отсутствия страницы, операционная система должна определить, какую страницу из памяти исключить, чтобы освободить место для загружаемой страницы. Если страница, которую нужно заместить, была изменена с момента загрузки в память, то ее содержимое должно быть обновлено на диске. Если страница не подвергалась изменениям и дисковая копия актуальна, то перезапись не требуется. В этом случае новая страница просто замещает старую.

1.1.2 Управление памятью в PostgreSQL

Разделяемый кэш буфер сохраняет страницы в оперативной памяти, доступ к которой в сотни тысяч раз быстрее, чем к дисковому хранилищу, где содержится вся информация о состоянии базы данных [3].

В операционной системе также есть дисковый кэш, который решает ту же проблему, поэтому системы управления базами данных обычно стараются избежать двойного кэширования, обращаясь к дискам напрямую, а не через кэш ОС. В случае с PostgreSQL это не так: все данные читаются и записываются с помощью обычных файловых операций [4]. Схема взаимодействия разделяемого кэш буфера и кэша уровня операционной системы представлена на рисунке 1.1.

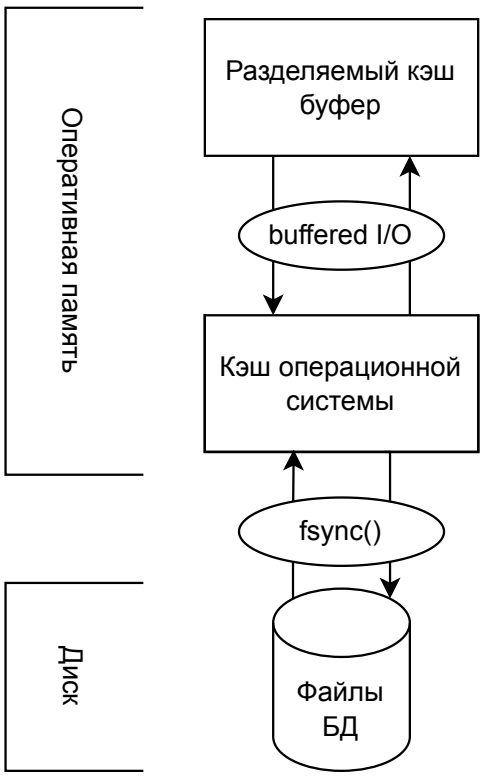

Рисунок 1.1 – Схема взаимодействия разделяемого кэш буфера с ОС

Пои чтении страница сначала ищется в кэш буфере, если она там не

находится, то отправляется запрос операционной системе на чтение страницы. Если операционная система находит ее в своем кэше, то данные сразу копируются в кэш буфер PostgreSQL, иначе операционная система загружает данные с диска в свой кэш и затем они копируются и попадает в разделяемый кэш буфер.

Для записи содержимого буфера на диск используется три процесса:

— user backend;

— bgwriter;

— checkpointer.

Процесс user backend обрабатывает пользовательские соединения. В случае запросов, которые изменяют данные, процесс user backend помечает измененные страницы для дальнейшей записи на диск. В случае, когда процессу надо исключить одну страницу из буфера и записать другую и страница для замещения помечена как измененная, user backend сам инициирует процесс копирования в кэш операционной системы для дальнейшей записи на диск.

Процесс bgwriter работает в фоновом режиме и нужен для снижения нагрузки на user backend и checkpointer. Он запускается раз в некоторый промежуток времени, который устанавливается через конфигурационный файл, просматривает, какие страницы были изменены и записывает их в кэш операционной системы для дальнейшей записи на диск.

Процесс checkpointer отвечает за создание контрольных точек, когда данные гарантированно записываются на диск. В момент вызова контрольной точки все измененные страницы записываются на диск и в лог файл добавляет специальная запись о контрольной точке. Все изменения, сделанные до этой записи, гарантировано записаны на диск.

Для восстановления после аварийного завершения в PostgreSQL используется механизм Write ahead logging. Все изменения перед внесением в страницы в разделяемом кэш буфере записываются в специальный лог файл. После прохождения очередной контрольной точки все изменения гарантировано синхронизированы с диском и записи, предшествующие этой точке, могут быть удалены из лог файла.

Контрольная точка создается в следующих случаях:

— явный вызов команды checkpoint;

— по истечению таймера, который задается в конфигурационном файле, по умолчанию – 300 секунд;

— размер write ahead log файла достиг максимального размера, заданного в конфигурационном файле;

— вызов функции pg_start_backup;

— вызов команды pg_basebackup;

— вызов команды завершения работы СУБД;

— при вызове команд создания и удаления базы данных.

Проблема двойного кэширования отчасти решается тем, что пока страница находится в разделяемом кэш буфере все обращения к ней идут через него и кэш уровня операционной системы задействован не будет. В следствии этого, рано или поздно страница будет исключена из кэша уровня операционной системы, так как к ней не будет обращений.

Буферный кэш является списком блоков. Каждый блок содержит страницу с данными и заголовок. Заголовок содержит:

— номер блока страницы;

— индикатор того, что страница была изменена, но еще не записана на диск;

— число обращений к странице;

— число активных операций или транзакций, которые используют страницу.

При старте все блоки в буферном кэше помещаются в список свободных. Для поиска нужной страницы используется хэш таблица. В качестве ключа используются номер файла и номер страницы в файле.

При обращении к памяти процесс сначала пытается найти страницу в кэше. Если она уже загружена, то счетчик обращений в заголовке соответствующего буфера увеличивается на единицу. До тех пор, пока это счетчик не равен нулю, страница не может быть выгружена из кэша.

Если страница не была найдена в кэше, то она должна быть прочитана с диска в какой-то блок. Если список свободных блоков не пуст, то будет взят первый из него, иначе требуется выбрать страницу, которая будет вытеснена из кэша.

В PostgreSQL для выбора кандидата на замещение анализируются два счетчика – число обращений и количество использований. Алгоритм часы поочередно проходит по всем страницам в кэше и, если оба счетчика равны нулю, то текущая страница будет замещена, иначе оба счетчика уменьшаются на единицу. Для избежания большого числа проходов по всем страницам в поисках кандидата на замещение по умолчанию счетчики не могут быть больше пяти.

Когда кандидат на замещение найден, счетчик использований ассоциированного с ним блока устанавливается в 1, чтобы другие процессы не могли его использовать. Если страница содержит измененную информацию, то запускается процесс переписывания в кэш ОС для дальнейшей записи на диск. После этого новая страница загружается в буфер и для нее выставляется счетчик обращений в единицу.

1.2 Методы замещения страниц

1.2.1 Оптимальный алгоритм

Оптимальной алгоритм предлагает вытеснять страницу, которая будет без ссылок в течение самого длительного времени. Этот алгоритм может быть реализован только во втором идентичном прогоне при условии использования истории страниц полученной во время первого запуска. Когда система сталкивается с нагрузкой в режиме реального времени у нее этой истории нет, поэтому этот алгоритм не может быть реализован на практике. Оптимальный алгоритм может быть использован для оценки других алгоритмов замещения страниц, которые могут быть применены и при первом прогоне.

1.2.2 Алгоритм NRU

Для того чтобы собирать статистику использования страниц виртуальной памяти, большинство компьютеров используют два бита состояния для каждой страницы. Бит R устанавливается при обращении к странице, а бит M устанавливается, когда страница изменяется.

Если аппаратура не поддерживает эти биты, то они могут быть созданы с помощью механизмов операционной системы. При запуске процесса все записи в его таблице страниц помечаются как отсутствующие в памяти. Когда происходит обращение к странице, возникает ошибка отсутствия страницы, и операционная система устанавливает бит R, изменяет запись в таблице страниц, устанавливая режим доступа только для чтения, и перезапускает команду. Если страница впоследствии изменяется, возникает другая ошибка страницы, позволяющая операционной системе установить бит M и изменить режим доступа к странице на чтение-запись.

Аналогичные биты хранятся в заголовке каждого блока в разделяемом кэш буфере PostgreSQL. В качестве бита R можно использовать счетчик обращений, который будет сбрасываться по истечению определенного времени. Для того, чтобы отметить страницы, которые были изменены в заголовке каждого блока присутствует специальный бит модификации.

Идея алгоритма Not Recently Used заключается в следующем: при запуске процесса оба этих бита для всех страниц устанавливаются в 0. При каждом прерывании от таймера бит R сбрасывается, чтобы отличить страницы, к которым не было обращений в последнее время, от тех, к которым были такие обращения.

При возникновении ошибки отсутствия страницы операционная система анализирует все страницы и на основе текущих значений битов R и M разделяет их на четыре категории:

1. К которым не было ни обращений, ни модификаций в последнее время.
2. К которым не было обращений в последнее время, но были модификации.
3. К которым были обращения в последнее время, но не было модификаций.
4. К которым были и обращения, и модификации в последнее время.

Для замещения выбирается произвольная страница из самого низкого непустого класса.

1.2.3 Алгоритм FIFO и его модификации

Система ведет список всех страниц, находящихся в памяти в данный момент. Недавно поступившие страницы находятся в конце списка, а те, что поступили раньше всех, находятся в начале. Если возникает ошибка отсутствия страницы, удаляется страница из начала списка, и в конец добавляется новая страница.

Алгоритм второй шанс является простой модификацией алгоритма FIFO и решает проблему удаления часто востребуемой страницы. Для этого используется проверка бита R самой старой страницы. Если значение этого бита равно нулю, то это означает, что страница не только старая, но и невостребованная, поэтому она сразу же удаляется. Если бит R имеет значение 1, то он сбрасывается, а страница помещается в конец списка страниц, а время ее загрузки обновляется, как будто она только что поступила в память. Затем поиск продолжается.

Алгоритм часы является улучшением алгоритма второй шанс. Он основан на идее использования циклического списка страниц, представленного в виде часов, где стрелка указывает на самую старую страницу.

Принцип работы алгоритма часы следующий:

1. В начале работы алгоритма все страницы помещаются в циклический список в виде часов, где каждая страница имеет бит R, который указывает на ее актуальность.

2. При возникновении ошибки отсутствия страницы проверяется страница, на которую указывает стрелка в циклическом списке.

3. Если бит R этой страницы равен 0, она удаляется из памяти, на ее место загружается новая страница, и стрелка сдвигается вперед на одну позицию.

4. Если бит R равен 1, он сбрасывается, и стрелка перемещается на следующую страницу в списке.

5. Этот процесс повторяется до тех пор, пока не будет найдена страница с битом R = 0.

1.2.4 LRU

Алгоритм замещения наименее востребованной страницы основан на идее, что страницы, которые долгое время не были востребованы, скорее всего, останутся невостребованными, в то время как страницы, которые интенсивно использовались в последнее время, вероятно будут снова востребованы. Поэтому стратегия замещения страниц в этом алгоритме основана на выборе наименее востребованной страницы для удаления.

Для реализации алгоритма Least Recently Used каждая страница в памяти связывается с программным счетчиком, который имеет начальное значение 0. При каждом прерывании от таймера операционная система сканирует все страницы в памяти. Для каждой страницы к счетчику добавляется значение бита R, который равен 0 или 1. Таким образом, счетчики позволяют приблизительно отслеживать частоту обращений к каждой странице.

При возникновении ошибки отсутствия страницы для замещения выбирается та страница, у которой счетчик имеет наименьшее значение, то есть та страница, которая дольше всего не была востребована.

Основная проблема этого алгоритма заключается в том, что он никогда не сбрасывает счетчики и страницы, которые активно использовались в прошлом, и сейчас не востребованы все равное будут оставаться в памяти.

Для борьбы с этой проблемой существует алгоритм старения, который предлагает при каждом прерывании таймера не прибавлять 1 к счетчику, а делать сдвиг вправо и прибавлять 1 к левому биту счетчика.

1.2.5 Алгоритм рабочий набор

Процессы начинают работу без каких-либо страниц в памяти, что приводит к ошибкам отсутствия страниц при первом обращении к данным. Система загружает страницы по мере необходимости. Постепенно процесс получает

большинство необходимых ему страниц и начинает работу более стабильно. Рабочий набор страниц, используемых процессом в данный момент, важен для эффективной работы. Многие системы замещения страниц стремятся отслеживать рабочий набор каждого процесса и обеспечивать его присутствие в памяти, перед перезапуском процесса.

Для реализации модели рабочего набора необходимо, чтобы система отслеживала, какие страницы именно входят в рабочий набор. Имея эту информацию, можно использовать следующий алгоритм замещения страниц: при возникновении ошибки отсутствия страницы следует выселить ту страницу, которая не принадлежит рабочему набору.

Рабочий набор представляет собой набор страниц, используемых в k последних обращениях к памяти. Для реализации алгоритма можно отслеживать страницы, использованные в k последних миллисекундах выполнения, вместо поиска страниц, используемых в k последних обращениях. Для получения этой информации можно добавить специальное поле в таблицу страниц и обновлять его на основе бита R по тику таймера.

Если возраст страницы превышает заранее выбранное значение на момент возникновения ошибки, она становится кандидатом на замену. В противном случае удаляется страница с самым большим возрастом или случайная, если у всех страниц одинаковый параметр.

1.2.6 Алгоритм WSClock

Алгоритм WSClock (Working Set Clock) является модификацией алгоритма рабочего набора и базируется на структуре данных, аналогичной циклическому списку страничных блоков, используемой в алгоритме часы.

Основные принципы работы данного алгоритма следующие:

1. Создается пустой циклический список страничных блоков.

2. При загрузке первой страницы она добавляется список. По мере загрузки следующих страниц они также попадают в список, формируя замкнутое кольцо.

3. В каждой записи списка содержится поле времени последнего использования из базового алгоритма рабочего набора, а также биты R и M.

4. При возникновении ошибки отсутствия страницы сначала проверяется страница, на которую указывает стрелка в списке. Если бит R установлен в 1, это означает, что страница была использована в течение текущего такта и не является идеальным кандидатом на удаление.

5. Затем бит R устанавливается в 0, стрелка перемещается на следующую страницу в списке, и процесс повторяется уже для нее.

6. После того, как бит R у страницы, на которую указывает стрелка, равен 0 и ее возраст превышает заданное значение, а также страница не изменена, происходит замещение этой страницы.

7. Если страница была изменена, то планируется запись на диск.

Если стрелка проходит полный круг и хотя бы одна запись на диск запланирована, поиск может продолжаться до тех пор, пока не будет найдена неизмененная страница. В противном случае все страницы считаются частью рабочего набора, и замещается любая страница, которая не была изменена. Если такой страницы нет, то замещается текущая страница.

1.2.7 Сравнительный анализ методов замещения страниц

Оптимальный алгоритм удаляет страницу с самым отдаленным предстоящим обращением. На практике реализовать такой алгоритм невозможно, но его можно использовать в качестве оценочного критерия.

Алгоритм исключения недавно использовавшейся страницы проводит разбиение всех страниц, основываясь на состоянии битов M и R, на 4 класса и проводит замещение произвольной страницы наименьшего непустого класса.

Алгоритм FIFO работает по принципу очереди и удаляет самую старую страницу. Алгоритм второй шанс борется с недостатками FIFO и перед удалением страницы проверяет не используется ли она в данный момент. Алгоритм

часы является разновидностью алгоритма второй шанс, но требует меньше времени на выполнение.

Алгоритм замещения наименее востребованной страницы стремится удалять страницы, которые не были востребованы долгое время. У этого алгоритма есть недостаток, связанный с тем, что страница, которая активно использовалась в прошлом, не обязательно будет востребована сейчас. Для борьбы с этим недостатком был разработан алгоритм старения.

Алгоритм рабочего набора отслеживает набор страниц, используемых за определенный промежуток времени и замещает страницу, которая не относится к рабочему набору. Алгоритм WSClock является оптимизацией алгоритма рабочего набора.

На практике чаще всего используются алгоритм старения и WSClock. Оба обеспечивают неплохую производительность страничной организации памяти и могут быть эффективно реализованы, но не лишены недостатков на определенном наборе задач.

1.3 Нейронные сети

Модель нейронной сети основана на биологическом нейроне. У нейрона есть ядро, которое называется телом. В теле накапливается электрический заряд. С телом соединены отростки. Отростки, по которым сигнал поступает в тело, называются дендритами. Отросток, по которому сигнал передается другим нейронам, называется аксоном. Место, где аксон соединяется с дендритами, называется синапсом. Синапс отвечает за количество заряда, которое перейдет от аксона к дендриту. Синапс может изменяться со временем. Именно с настройкой синапса и связана тренировка биологической нейронной сети.

Математическая модель МакКаллока-Питтса. В математической модели МакКаллока-Питтса, тело нейрона, где накапливается заряд, заменяется на сумматор. Дендриты являются входами сумматора, а выходом – аксоном. Биологический нейрон накапливает заряд до тех пор, пока этот заряд не достигнет какого-то значения, и только после этого этот заряд уходит по аксону к другим нейронам. В математической модели к сигналу после выхода из сумматора применяется функция активации и только после этого

сигнал попадает на дендрит следующего нейрона. Синапсы в математической модели заменяются на веса входов нейрона. Математическая модель нейрона выражается зависимостью 1.1

$$y = f\left(\sum_{i=1}^{n}(w_{\text{i}}x_{\text{i}}) + b\right), \tag{1.1}$$

где y – сигнал на выходе из нейрона, f – функция активации, w_{i} – вес i входа, x_{i} – сигнал этого входа, b – некоторое значение смещения, которое задается отдельно для каждого нейрона. Обучение нейронной сети происходит за счет настройки синаптических весов w_{i} и смещения b.

Функции активации. Существует много различных функций активации. Наиболее популярными считаются логистическую функцию, гиперболический тангенс, ReLU [5]. Важной особенностью функций активации является их дифференцируемость, поскольку при обратном распространении ошибки необходимо вычислять градиенты, использующие производную функции активации.

Логистическая функция преобразовывает поступающие в неё значения в вещественный диапазон [0, 1]. Это означает, что при x>0 выходное значение будет примерно равно единице, а при x<0 будет близким к нулю. Данная функция часто используется в задачах классификации [5]. Логистическая функция определяется зависимостью 1.2.

$$y = \frac{1}{1 + e^{-x}}. \tag{1.2}$$

Гиперболический тангенс схож с логистической функцией, но в отличии от нее может принимать отрицательные значения. Гиперболический тангенс определяется зависимостью 1.3.

$$y = \frac{e^{2x} - 1}{e^{2x} + 1}. \tag{1.3}$$

Функция ReLU возвращает 0, если принимает отрицательный аргумент, в случае же положительного аргумента, функция возвращает само число.

Функция ReLU определяется зависимостью 1.4.

$$\mathrm{ReLU}(x) = \begin{cases} x, & \text{если } x > 0, \\ 0, & \text{иначе,} \end{cases} \tag{1.4}$$

ReLU решает проблему обнуления градиента (ситуация, при которой во время обучения градиенты по всем весам становятся близкими или равными нулю) для положительных чисел, также она вычисляется гораздо проще, чем сигмоидальные функции (логистическая функция, гиперболический тангенс) [5].

Составляющие нейронной сети. При обучении нейронной сети используются две подвыборки обучающего множества. Вся обучающая выборка состоит из какого-то количества объектов, для которых известны признаки, на которые должна обучиться нейронная сеть. Первая подвыборка называется тренировочной и используется для итеративного обучения нейронной сети. Вторая называется тестовой и используется для оценки того, насколько хорошо обучена нейронная сеть.

Нейронную сеть определяют следующие параметры:

— архитектура нейронной сети – отвечает за то, как нейроны связаны между собой;

— функция потерь – определяет насколько точно работает модель [6];

— метод оптимизации – определяет способ уменьшения функции потерь на каждой итерации обучения.

Нейроны делятся на три типа: входной, скрытый и выходной. В том случае, когда нейросеть состоит из большого количества нейронов, вводят термин слоя. Соответственно, есть входной слой, который получает информацию, некоторое количество скрытых, которые ее обрабатывают и выходной слой, который выводит результат [7]. Количество скрытых слоев и число нейронов в каждом из них задают архитектуру нейронной сети.

Методы оптимизации. Один из методов оптимизации – градиентный спуск [8]. Градиентный спуск основан на пошаговом приближении функции к локальному минимуму. На каждой итерации алгоритма новые значения

получаются по формуле 1.5

$$w_1 = w_0 - \alpha \Delta f(w_0), \tag{1.5}$$

где w_1 – вектор новых значений, которые подбираются алгоритмом, w_0 – значения параметров на текущем шаге, $\Delta f(w_0)$ – вектор градиентов функции потерь по каждому из параметров на текущем шаге, α – скорость обучения.

На каждой итерации градиентного спуска требуется считать градиент функции потерь, которая зависит от функций активации каждого из нейронов сети. В связи с этим к фукнциям потерь и активации применяются требования по дифференцируемости.

В связи с тем, что градиентный спуск находит только локальный минимум, не всегда полученный результат будет оптимальным. Результат работы алгоритма зависит от изначальных настроек параметров нейронной сети.

Выделяют три основных типа градиентного спуска [8]:

— мини-пакетный градиентный спуск – в этом случае обучающий набор данных разбивается на небольшие партии, которые используются для расчета ошибки модели и обновления коэффициентов модели;

— стохастический градиентный спуск – в этом случае градиент оптимизируемой функции считается на каждом шаге не как сумма градиентов от каждого элемента выборки, а как градиент от одного, случайно выбранного элемента;

— пакетный градиентный спуск – это разновидность алгоритма градиентного спуска, который вычисляет ошибку для каждого примера в наборе обучающих данных, но обновляет модель только после того, как все обучающие примеры были оценены.

Одной из проблем градиентного спуска является неизменяемая во время обучения скорость обучения. Постоянная скорость обучения может привести к следующим проблемам: если ее значение будет выбрано слишком низким, то модель будет дольше сходиться и требовать большего числа итераций для достижения оптимального решения, если ее значение будет слишком большим, то модель может расходиться и на каждой итерации проходить мимо глобального минимума.

Для решения этой проблемы нужно использовать адаптивно настраивающуюся скорость обучения. Значение скорости обучения для каждого параметра должно настраиваться адаптивно, исходя из правила, что чем больше значение ошибки, тем больше должна быть скорость обучения. Увеличение скорости обучения при больших значениях ошибки дает возможность перескочить через локальные минимумы, а ее уменьшение при малых значениях не дает модели на каждой итерации перескакивать через глобальный минимум.

Для адаптивного изменения весов модели можно использовать алгоритм RMSProps. Этот алгоритм работает по следующим правилам:

— на каждой итерации для каждого параметра считается экспоненциальное скользящее среднее градиента с учетом всей истории обучения;

— при помощи полученных значений для каждого параметра вычисляется скорость обучения и производится обновление весов модели.

Экспоненциальное скользящее среднее на очередной итерации высчитывается по формуле 1.6

$$E_{\mathrm{t}} = \beta * g_{\mathrm{t}}^2 + (1 - \beta) * E_{\mathrm{t}-1}, \tag{1.6}$$

где E_{t} – новое полученное значение экспоненциального скользящего среднего, $E_{\mathrm{t}-1}$ – значение, полученное на предыдущей итерации, β – настраиваемый коэффициент, g_{t} – градиент функции потерь по соответствующему параметру.

После этого веса обновляются с использованием соотношения 1.7

$$w_{\mathrm{t}} = w_{\mathrm{t}-1} - \frac{\eta}{\sqrt{E_{\mathrm{t}}}} g, \tag{1.7}$$

где w_{t} – новое полученное значение параметра модели, $w_{\mathrm{t}-1}$ – предыдущее значение этого параметра, η – скорость обучения, E_{t} – значение экспоненциального скользящего среднего для этого параметра.

Для улучшения сходимости модели при адаптивном обновлении скорости обучения можно считать экспоненциальное скользящее среднее не только по квадрату градиента, но и по самому значению и использовать оба полученных значения при подсчете новой скорости обучения на каждой итерации. Такой подход реализован в алгоритме Adam [9].

Первый и второй моменты высчитываются по формулам 1.8 и 1.9 соответственно

$$m_{\text{t}} = \beta_1 * g_{\text{t}} + (1 - \beta_1) * m_{\text{t}-1}, \tag{1.8}$$

$$v_{\text{t}} = \beta_2 * g_{\text{t}}^2 + (1 - \beta_2) * v_{\text{t}-1}, \tag{1.9}$$

где m_{t} и v_{t} – первый и второй моменты в соответствующий момент времени, β_1 и β_2 – настраиваемые коэффициента, g – градиент функции потерь по соответствующему параметру.

Для увеличения влияния истинных значений градиента на начальных этапах к моментам применяется корректировка по формулам 1.10 и 1.11.

$$\hat{m}_{\text{t}} = \frac{m_{\text{t}}}{1 - \beta_1^t}, \tag{1.10}$$

$$\hat{v}_{\text{t}} = \frac{v_{\text{t}}}{1 - \beta_2^t}, \tag{1.11}$$

Итоговое обновление весов осуществляется по формуле 1.12

$$w_{\text{t}} = w_{\text{t}-1} - \eta \frac{\hat{m}_{\text{t}}}{\sqrt{\hat{v}_{\text{t}}} + \epsilon}, \tag{1.12}$$

где w_{t} – новое полученной значение параметра модели, $w_{\text{t}-1}$ – предыдущее значение этого параметра, η – скорость обучения, ϵ – поправка, защищающая от деления на ноль.

Функции потерь. Согласно исследованиям [10] для задачи классификации самой эффективной функцией потерь являются категориальная перекрестная энтропия, которая определяется выражением 1.13

$$CM = -\sum_{i=1}^{N} t_{\text{i}} \log p_{\text{i}}, \tag{1.13}$$

где N – число классов классификации, t_{i} – 0 или 1 в зависимости от того принадлежит ли изображение на входе нейронной сети классу, за который отвечает i нейрон выходного слоя, p_{i} – результат на выходе из нейрона.

В задачах классификации используют категориальную перекрестную

энтропию в качестве функции потерь. В таком случае на выходном слое нейронной сети создается столько нейронов, сколько возможных классов может иметь объект на входе. В качестве функции активации для каждого из таких нейронов используют софт макс. Софт макс определяется выражением 1.14

$$SM_{\text{i}} = \frac{e^{y_{\text{i}}}}{\sum\limits_{i=1}^{N} e^{y_{\text{j}}}}, \tag{1.14}$$

где y_{i} – результат на выходе из нейрона, к которому применяется функция активации, N – число нейронов в выходном слое, y_{j} – результат на выходе из j нейрона выходного слоя.

Знаменатель в выражении 1.14 отвечает за нормировку. Таким образом, каждый из нейронов выходного слоя показывает вероятность принадлежности объекта на входе нейронной сети к некоторому классу, а сумма всех этих вероятностей будет равна 1.

1.4 Многослойные сети

1.4.1 Перцептрон

Перцептрон – это математическая модель, воспроизводящая принципы обработки информации, схожие с работой человеческого головного мозга. Архитектура перецептрона состоит из трех ключевых элементов [11]:

— сенсоры – получают входные сигналы;

— ассоциативные элементы – обрабатывают данные;

— реагирующие элементы – формирует итоговый отклик.

Эти компоненты организованы в слои нейронной сети: входной, один или несколько скрытых и выходной. Схема трехслойного перцептрона приведена на рисунке 1.2

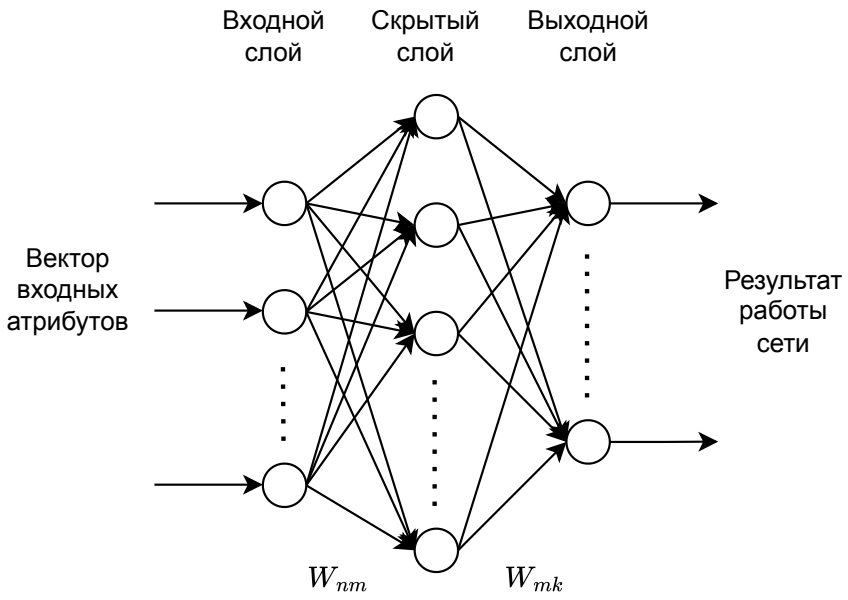

Рисунок 1.2 – Схема перцептрона

W_{nm} и W_{mk} на схеме – матрицы обучаемых весов. n, m, k – число нейронов на входном, скрытом и выходном слоях соответственно.

Результат на выходе из k-го нейрона выходного слоя может быть посчитан по следующей формуле: 1.15

$$Y_k(x) = Y_k(x_1, x_2, ..., x_n) = f_2(\sum_{j=0}^{m}(w_{jk} f_1(\sum_{i=0}^{n}(w_{ij} x_i))), \tag{1.15}$$

где f_1 и f_2 – функции активации на скрытом и выходном слоях соответственно, w_{ij} и w_{jk} – обучаемые веса из матриц W_{nm} и W_{mk}.

В многослойном перцептроне данные последовательно проходят через входной слой, скрытый и выходной. На каждом этапе нейроны слоя выполняют следующие действия:

— линейная комбинация – нейрон умножает сигналы на соответствующие веса и суммирует результаты;

— нелинейное преобразование – полученная сумма пропускается через активационную функцию, которая усиливает характеристики, за которые отвечает соответствующий узел.

Марвин Минский и Сеймур Паперт в своей работе [12] показали, что

однослойные нейронные сети, включающие только входной и выходной слои, способны работать исключительно с линейно разделимыми задачами и обеспечивают лишь линейную аппроксимацию. Этот барьер преодолевается за счёт внедрения скрытых слоёв.

В таких сетях скрытый слой играет ключевую роль: входные данные трансформируются в новое пространство, где выходной слой строит разделяющие поверхности для классификации. Таким образом, модель не только анализирует исходные признаки, но и выявляет признаки признаков, формируемые скрытыми нейронами. Это позволяет сети обучаться сложным нелинейным закономерностям.

Число нейронов на входном слое напрямую зависит от числа входных признаков. Число нейронов на выходном слое зависит от решаемой задачи. Для задач классификации это число равно числу возможных классов и каждый нейрон на выходном слое отвечает за вероятность принадлежности входного параметра одному из классов.

Число нейронов на скрытом слое, как правило, выбирается опытным путем. Слишком большое число нейронов может привести к переобучению модели, а слишком маленькое может не хватить для решения поставленной задачи.

Хехт-Нильсен в своей работе [13], основанной на работе Колмогорова [14] предлагает использовать $2N + 1$ нейронов на скрытом слое, где N – число входов сети.

Баум и Хесслер с своей работе [15] вводят эмпирическое правило для числа весов на скрытом слое для борьбы с переобучением 1.16

$$p > \frac{w}{\epsilon}, \tag{1.16}$$

где p – размер обучающей выборки, w – число весов, ϵ – допустимый уровень ошибки.

Для обучения перцептрона могут быть использованы следующие методы:

— стохастические методы обучения;

— обратное распространение ошибки.

Стохастические методы обучения предполагают обучение по следующему алгоритму:

1. Выбрать значения весов случайным образом.

2. Подсчитать значение функции потерь.

3. Подкорректировать значение случайного весового коэффициента на небольшую величину. Если коррекция уменьшает значение функции потерь, то оставить ее, иначе вернуться к изначальному состоянию.

4. Повторять шаг 3 до тех пор, пока сеть не будет достаточно обучена.

Для реализации обратного распространения ошибки необходимо осуществить минимизацию функции потерь с помощью изменения значений обучаемых весов в направлении обратном градиенту. Для расчета частной производной по каждому из весов необходимо выразить функцию потерь через значения весов и входных аргументов. Для того, чтобы выразить значение на выходе из k-го нейрона через обучаемые веса и входные параметры можно использовать формулу 1.15.

1.4.2 RBF сеть

RBF сеть является многослойным перцептроном, на скрытом слое которого используются RBF нейроны [16]. Потенциал такого нейрона рассчитывается как евклидово расстояние между векторами весовых коэффициентов и входных величин 1.17:

$$V = ||w - x|| = \sqrt{\sum_{i=1}^{N}(w_i - x_i)^2}, \tag{1.17}$$

где w – вектор весовых коэффициентов, x – вектор входных величин, а N – размер этих векторов.

Для RBF нейронов используется следующая функция активации 1.18:

$$f(V) = e^{-(\frac{V}{b})^2}, \tag{1.18}$$

где V – потенциал нейрона, b – коэффициент разброса, который отвечает за плавность функции.

Схема RBF сети приведена на рисунке 1.3

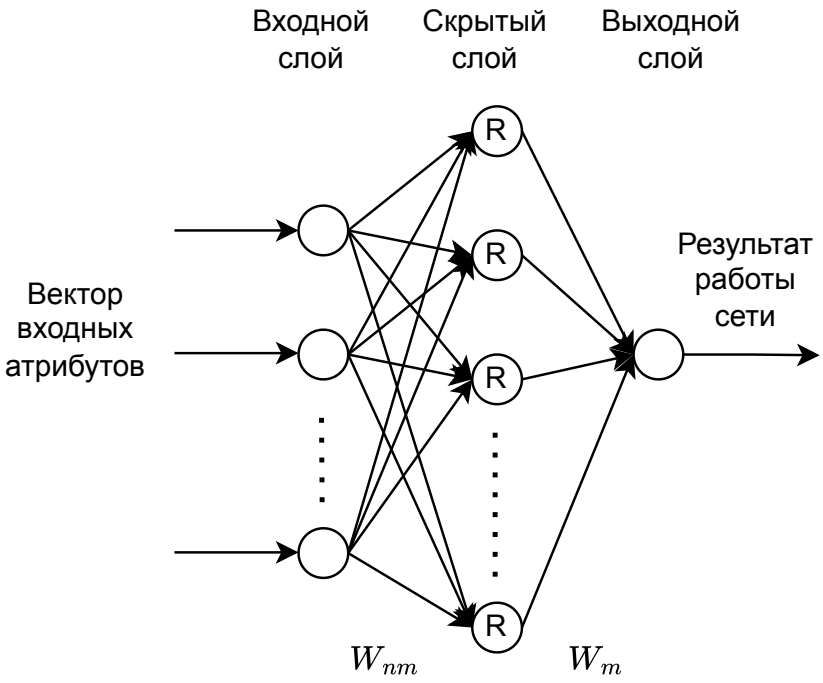

Рисунок 1.3 – Схема RBF сети

На схеме буквой R обозначены RBF нейроны. W_{nm} – матрица обучаемых весов скрытого слоя, W_m – вектор обучаемых весов выходного слоя, n – число нейронов на входном слое, m – число RBF нейронов. Выходной слой является линейной комбинацией активаций RBF нейронов скрытого слоя.

Особенностью такой сети является то, что каждый RBF нейрон активируется только тогда, когда входные данные близки к его эталонному вектору весов w. Таким образом, каждый нейрон скрытого слоя представляет собой центр кластера в пространстве данных, а параметр b отвечает за область охвата этого нейрона. Чем больше значение b, тем большое отклонение от вектора эталонных весов будет приводить к активации нейрона.

Обучение RBF сети состоит из трех этапов:

— выбор центров кластеров;

— расчет параметра разброса b;

— обучения весов выходного слоя.

Для выбора центров кластеров могут быть использованы методы кластеризации, к примеру, k-средний.

Параметр разброса может быть как настроен вручную для каждого кластера, так и задан константой для всех нейронов. К примеру, параметр b может быть задан как среднее расстояние между центрами кластеров.

Для обучения весов выходного слоя может быть использована линейная регрессия.

1.4.3 Вероятностная сеть

Вероятностная модель является развитием RBF сети. Схема такой модели представлена на рисунке 1.4

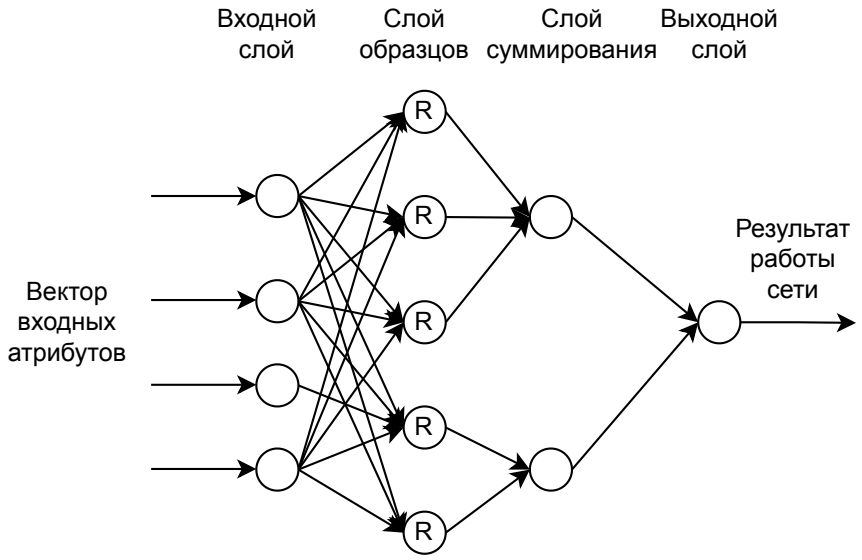

Рисунок 1.4 – Схема вероятностной сети

Вероятностная сеть состоит из следующих слоев:

— входной слой;

— слой образцов;

— суммирующий слой;

— выходной слой.

Входной слой – первый этап работы сети, который принимает входные данные. Каждый нейрон на этом слое соответствует одному входному признаку.

Каждый нейрон на слое образцов соответствует одному образцу из обучающей выборки. Нейроны этого слоя используют RBF функцию активации для вычисления сходства между входным набором данных и образцом, которому они соответствуют.

Каждый нейрон на слое суммирования представляет класс. Нейроны этого слоя суммируют активации RBF нейронов с предыдущего слоя, которые относятся к одному классу.

Выходной слой принимает решение о классификации входных данных, выбирая класс с наибольшей вероятностью.

1.5 Рекуррентные сети

Рекуррентные нейронные сети – нейронные сети с обратной связью между различными слоями нейронов. Их характерная особенность – передача сигналов с выходного или скрытого слоя во входной слой. Рекуррентная нейронная сеть может состоять из любого числа слоев.

Рекуррентные нейронные сети хорошо подходят для обработки последовательностей, например, временные ряды (изменения цен акций, показания датчиков), последовательности с зависимыми элементами (предложения естественного языка), то есть любые данные, где соседние экземпляры (точки выборки) зависят друг от друга и эту зависимость нельзя игнорировать [17].

Общая схема рекуррентной сети приведена на рисунке 1.5.

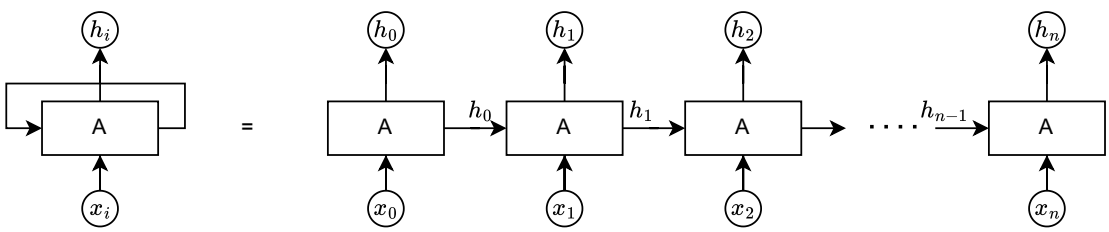

Рисунок 1.5 – Схема рекуррентной сети

Фрагмент нейронной сети A принимает входное значение x_i и возвращает значение h_i. Наличие обратной связи позволяет передавать информацию

от одного шага сети к другому. Рекуррентную сеть можно рассматривать, как несколько копий одной и той же сети, каждая из которых передает информацию последующей копии.

1.5.1 Нейронная сеть Хопфилда

Схема нейронной сети Хопфилда из трех нейронов изображена на рисунке 1.6.

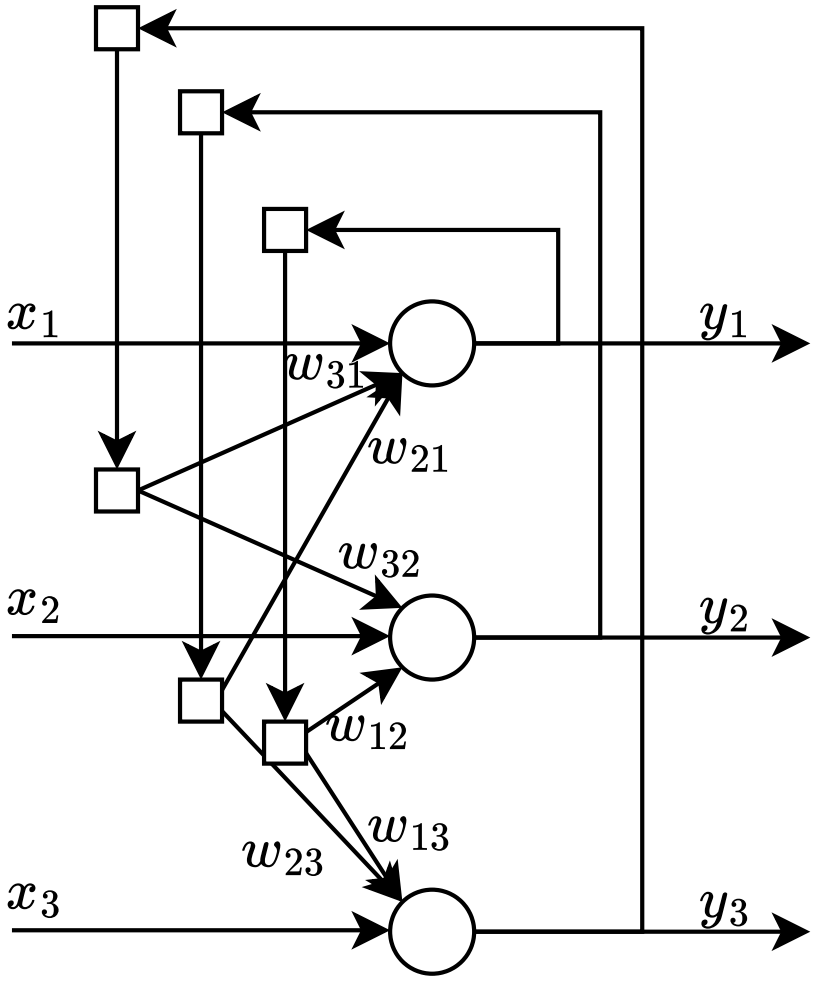

Рисунок 1.6 – Схема нейронной сети Хопфилда

Нейронная сеть Хопфилда – полносвязная однослойная нейронная сеть с симметричной матрицей связей [18]. Функционирование сети продолжается до тех пор, пока не будет достигнуто состояние равновесия, то есть до тех пор, пока новый выход из сети не будет равен предыдущему. Входной образ является начальным состоянием сети, а при равновесии получается выходной.

Сеть состоит из N нейронов, где N – размерность входного и выходного векторов. Каждый нейрон на входе и выходе может принимать одно из двух состояний 1.19:

$$x_i^{(t)}, y_i^{(t)} \in \{-1; +1\}, \tag{1.19}$$

где $x_i^{(t)}$ и $y_i^{(t)}$ – значение на входе и выходе i-го нейрона соответственно в момент времени t.

Работа сети описывается функцией энергии 1.20:

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} wijx_i x_j, \tag{1.20}$$

где w_{ij} – элемент матрицы взаимодействия W, которая состоит из весовых коэффициентов. В процессе функционирования сети функция энергии должна монотонно уменьшаться.

Состояние сети определяется множеством текущих выходных сигналов y_i от всех нейронов. Таким образом, состояние сети является двоичным числом, так как на выходе нейрона может быть только 2 значения. Каждый бит соответствует значению на выходе конкретного нейрона.

Процесс обучения сети заключается в составлении матрицы взаимодействия W. Матрица строится из m эталонных образов. Каждый образ является бинарным вектором размерности N.

Для расчета весовых коэффициентов применяется следующее выражение 1.21:

$$w_{ij} = \frac{1}{N} \sum_{d=1}^{m} X_{id} X_{jd}, \tag{1.21}$$

где N – размерность векторов, m – число запоминаемых выходных векторов, d – индекс запоминаемого выходного вектора, X_{ij} – i-я компонента запоминаемого выходного j-го вектора. Матрица взаимодействий является симметричной. Элементы на главной диагонали матрицы равны нулю.

Значение выхода i-го нейрона в текущий момент рассчитывается по следующей формуле 1.22:

$$y_i^{(t)} = sign(\sum_{j=1, j\neq i}^{N} w_{ji} y_j^{(t-1)}), \tag{1.22}$$

где $sign$ – функция, возвращающая знак аргумента.

Существует два режима работы сети Хопфилда:

1. Синхронный режим. При таком подходе все нейроны просматриваются последовательно, их состояния запоминаются и не меняются до тех пор, пока не будут обработаны все нейроны. Затем состояние всех нейронов синхронно обновляется.

2. Асинхронный режим. При таком режиме работы состояния нейронов обновляются последовательно, то есть для каждого нейрона поочередно вычисляется новое состояние, для каждого следующего нейрона новое состояние вычисляется с учетом всех изменений состояний рассмотренных ранее нейронов.

В асинхронном режиме работы невозможен динамический аттрактор, то есть вне зависимости от количества запомненных образов и начального состояния сеть придет к устойчивому состоянию.

На число образов, которые может запомнить сеть Хопфилда, накладывается ограничение 1.23:

$$M < \frac{N}{2 \log_2 N}, \tag{1.23}$$

где M – максимально число эталонов, которое может запомнить сеть, N – число нейронов.

Одним из недостатков сети является проблема ложных аттракторов: достижение устойчивого состояния сети не гарантирует правильный ответ.

1.5.2 Двунаправленная ассоциативная память

Двунаправленная ассоциативная память является расширением сети Хопфилда [19]. Эта сеть позволяет ассоциировать пары векторов. Схема сети

представлена на рисунке 1.7.

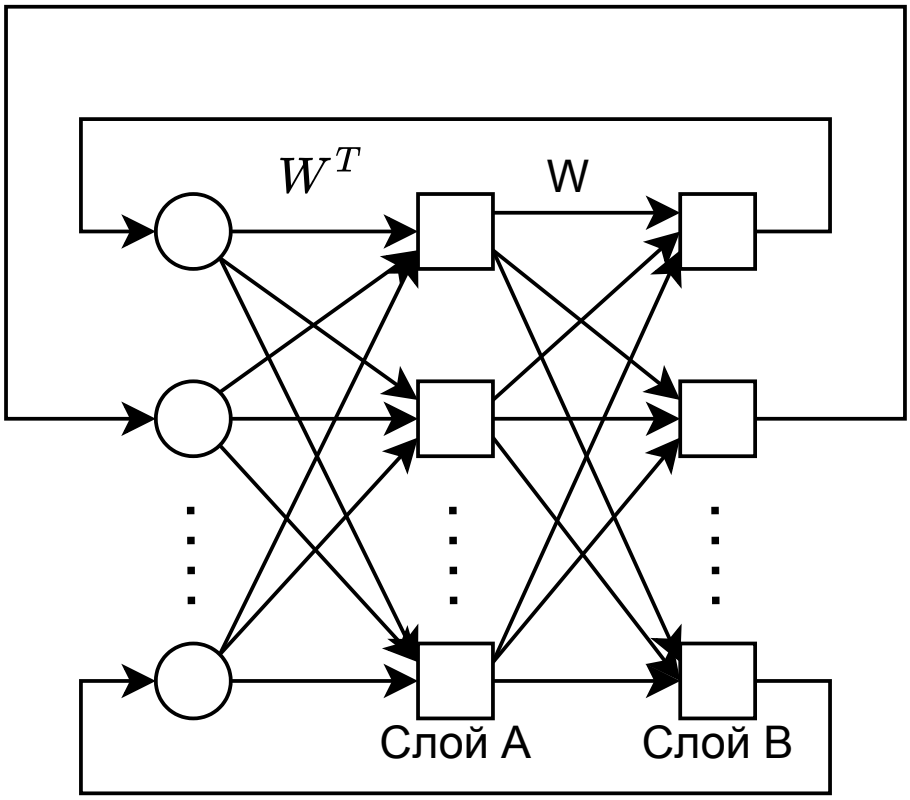

Рисунок 1.7 – Двунаправленная ассоциативная память

Сеть состоит из двух слоев. Входной вектор A поступает на слой A и обрабатывается матрицей весов W сети. В результате вырабатывается выходной вектор B, который поступает на слой B. Вектор B обрабатывается транспонированной матрицей W^T. В результате работы этого слоя получается новый входной вектор A. Такой процесс функционирования сети продолжается до тех пор пока, не будет достигнуто стабильное состояние, при котором ни вектор A, ни вектор B не изменяются.

Обучение сети происходит с помощью обучающего набора, состоящего из пар векторов A и B. Матрица весов вычисляется как сумма произведений

всех векторных пар обучающего набора 1.24:

$$W = \sum_{i=1}^{M} {A_i}^T B_i, \tag{1.24}$$

где W – матрица весов, M – число обучающих пар, A_i и B_i – вектора из обучающего набора.

1.5.3 Сеть LSTM

Сеть LSTM [20] состоит из четырех компонентов:

— состояние ячейки – память сети, которая передается по всей цепочке модулей;

— фильтр забывания – контролирует меру сохранения информации в ячейке;

— входной фильтр – контролирует меру вхождения нового значения в память;

— выходной фильтр – отвечает за меру того, как будет использовано значение из ячейки памяти при расчете выходной функции активации.

Схема LSTM изображена на рисунке 1.8.

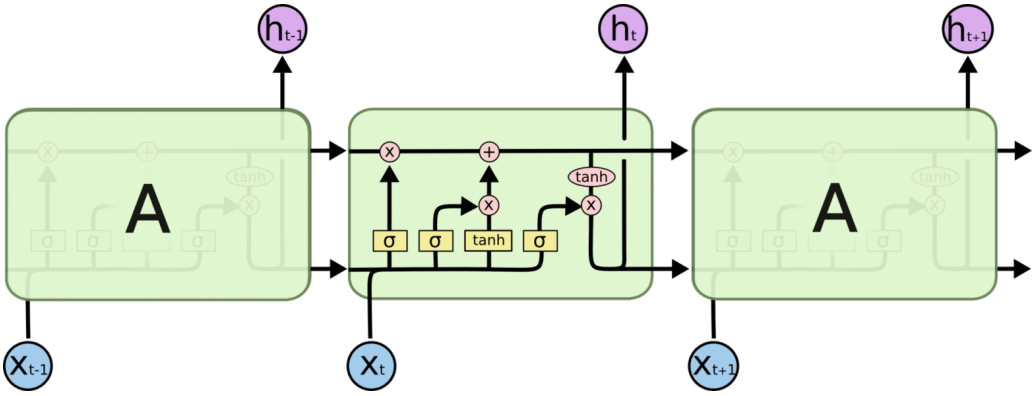

Рисунок 1.8 – Схема LSTM сети

Ключевой из них – состояние ячейки, которая переходит между повторяющимися модулями сети, подвергаясь преобразованиям. Три других

компонента отвечают за забывание прошлого состояния ячейки, обновление состояния на основе входных данных и выхода из прошлого модуля, а также за получение выходного значения из текущего блока [21].

Первый компонент нужен для определения того, какую часть информации можно выбросить из состояния ячейки. На вход к нему поступают входные данные в текущий блок и выходной вектор из прошлого модуля. На выходе при помощи сигмоидного фильтра для каждого значения в состоянии ячейки вычисляется число от 0 до 1. Результат работы этого фильтра описывается следующим выражением 1.25:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{1.25}$$

где $[h_{t-1}, x_t]$ – конкатенация результата работы предыдущего слоя и входного вектора в текущий слой, W_f и b_f – матрица и вектор обучаемых весов, f_t – результат работы фильтра.

Задача следующего компонента – определить какая новая информация будет сохранена в ячейке. Для этого сначала при помощи сигмоидного входного фильтра определяются значения, которые будут сохранены в ячейке, а затем с использованием слоя гиперболического тангенса вычисляются новые значения кандидатов на попадание в ячейку. Работа этого фильтра определяется при помощи выражений 1.26 - 1.27:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{1.26}$$

$$\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C), \tag{1.27}$$

где i_t определяет, какие значения будут сохранены в ячейке, $\hat{C}_t$ – новые значения кандидатов на попадание в ячейку, W_i, W_C, b_i, b_c – матрицы и вектора обучаемых весов.

Обновление состояния ячейки происходит с использованием следующего выражения 1.28:

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t, \tag{1.28}$$

где C_t – новое состояние ячейки, C_{t-1} – состояние ячейки на прошлом шаге.

Задача последнего компонента – определить, какая информация будет

на выходе из текущего модуля. Для этого используется поточечное умножение текущего состояния ячейки, пропущенного через фильтр гиперболического тангенса, и входных данных, объединенных с выходом из прошлого модуля и прошедших через сигмоидный фильтр. Работа этого компонента определяется выражениями 1.29 - 1.30:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{1.29}$$

$$h_t = o_t * \tanh(C_t), \tag{1.30}$$

где h_t – результат работы текущего слоя, C_t – состояние ячейки, W_o и b_o – матрица и вектор обучаемых весов.

1.6 Переобучение нейронной сети

1.6.1 Проблема переобучения нейронной сети

Проблема переобучения в нейронных сетях заключается в том, что модель запоминает данные только из обучающей выборки, не обобщая свои знания на новые, ранее не встречавшиеся данные. Это происходит из-за того, что модель адаптируется к обучающим примерам, вместо того, чтобы учиться классифицировать новые данные [22]. Признаком переобучения модели является существенно большее значение ошибки распознавания на тренировочной выборке, нежели на тестовой. Зачастую переобучение появляется из-за использования слишком сложных моделей, либо наборов данных, в которых вхождения похожи друг на друга [22].

Недообучение - это противоположная проблема переобучения нейронных сетей. Оно характеризуется тем, что алгоритм обучения не достигает удовлетворительной точности на обучающем множестве. Это может быть связано с тем, что выбрана слишком простая модель или недостаточно обучающих примеров. В результате модель не сможет классифицировать данные в более сложных случаях. [22].

Примеры недообученной, переобученной и оптимально обученной ней-

ронной сети приведены на рисунке 1.9.

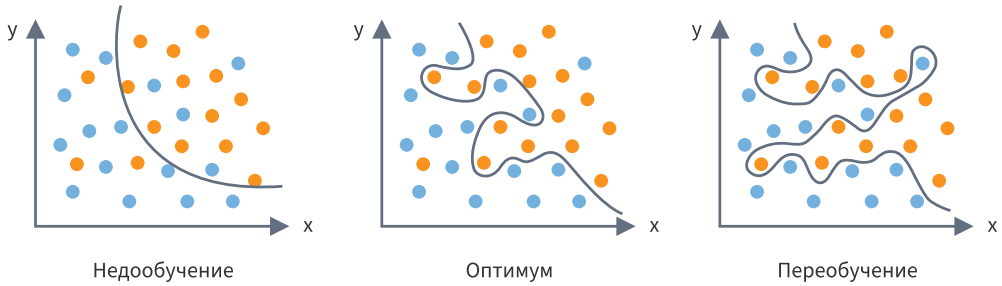

Рисунок 1.9 – Пример переобучения, недообучения и оптимального обучения

В этом примере нейронная сеть используется для разделения входного множества объектов на два класса. В первом случае нейронная сеть является недообученной, так как вероятность ошибки равна 0.22 и может быть еще уменьшена за счет использование более сложной формы зависимости.

На примере по середине нейронная сеть строит общую зависимость для данного набора данных, не подгоняя значения под аномальные элементы для рыжего и голубого класса слева и справа соответственно.

Пример справа показывает переобученную нейронную сеть, которая строит зависимость, подгоняя параметры под аномальные параметры обучающей выборки.

Для борьбы с переобучением можно использовать следующие способы:

— аугментация обучающей выборки;

— метод раннего останова;

— регуляризация;

— батч нормализация.

1.6.2 Аугментация

Первый способ борьбы с переобучением – аугментация обучающей выборки. Аугментацией называется этап обучения нейронных сетей, состоящий в модификации обучающей выборки. В основном этот метод применяется для

изображений (поворот, масштабирование, зеркальное отражение и т. д.) по определенному правилу с целью расширить обучающую выборку и повысить ее разнообразие [23].

Существует три основных вида аугментации:

— геометрическая аугментация – изменение геометрических параметров изображения, таких как поворот, масштабирование, сдвиг и отражение;

— цветовая аугментация – изменение цветовых параметров изображения, таких как яркость, контрастность и насыщенность;

— добавление шума.

Аугментация первого типа обычно улучшает качество работы сверточных нейронных сетей, так как такие сети не инвариантны к масштабу, и изменение масштаба изображения значительно повышает разнообразие данных, позволяя сети обучаться на более разнообразных наборах данных [23]. В статье [24] описывается повышение точности распознавания нейронной сети на 10 процентов за счет использования аугментации масштаба.

Аугментации второго типа предполагают случайное изменение компонент R, G, B цвета пикселей изображения. Это один из самых эффективных методов аугментации данных, потому что нейросети без этой аугментации имеют тенденцию к заучиванию правил вида «сумма цветов пикселей в области». Также такая аугментация может улучшить способность распознавания сети при различных условиях освещенности [23].

Аугментация добавлением шума на изображение повышает устойчивость модели к шуму на реальных изображениях.

1.6.3 Метод раннего останова

Метод раннего останова после каждой эпохи обучения проверяет точность модели на обучающей и тестовой выборках. Обучение нейронной сети начинается при случайных значениях весов, и с каждой эпохой обучения точность на обучающей выборке будет повышаться, а на тестовой точность

сначала будет расти, в момент, когда сеть будет достаточно обучена, зафиксируется на некотором значении, а потом начнет падать из-за переобучения модели.

Суть метода раннего останова заключается в отслеживании точности модели на тестовой выборке и остановке обучения в момент, когда она начинает расти.

К преимуществам данного метода можно отнести:

— сокращение времени обучения, так как нейронная сеть не будет обучаться, когда в этом уже нет необходимости;

— отсутствие дополнительных затрат на дополнение обучающей выборки.

1.6.4 Регуляризация

Метод регуляризации заключается в ограничение значений весовых коэффициентов нейронной сети, что делает их распределение более равномерным. Это достигается за счет добавления некоторого штрафа за увеличение весов нейронной сети в функцию потерь.

Существует три основных вида регуляризации [25]:

— L1 регуляризация, которая также называется Лассо регуляризацией, она добавляет штраф от суммы абсолютных значений весов модели;

— L2 регуляризация, которая также называется регуляризацией Тихонова, она добавляет штраф от суммы квадратов весов модели;

— дропаут, который случайным образом удаляет связи между нейронами.

В пером виде регуляризации новая функция потерь описывается выражением 1.31

$$L_{\text{new}} = L_{\text{old}} + \lambda \sum_{i=1}^{N} |w_{\text{i}}|, \tag{1.31}$$

где L_{new} – новое значение функции потерь, полученное после регуляризации, L_{old} – значение функции потерь до проведения регуляризации, λ – коэффициент штрафования весов, N – число весов в модели, w_{i} – значение i-го веса модели.

При коэффициенте λ равном нулю никакой регуляризации не будет и модель переобучится. При повышении коэффициента штрафования модель будет приближаться к оптимальной, то есть значение ошибки на тестовой выборке будет падать, но чем больше значение этого коэффициента, тем ближе значения всех весов будут к нулю, тем дальше модель отклоняется от локального минимума функции потерь до регуляризации и тем больше растет ошибка модели не тренировочной выборке. Таким образом, значение коэффициента λ должно подбираться экспериментально во время обучения. Чем меньше ошибка на тренировочной выборке, тем лучше подобран коэффициент штрафования.

В регуляризации Тихонова штраф считается не по сумме абсолютных значений, а по сумме квадратов и выражается зависимостью 1.32

$$L_{\text{new}} = L_{\text{old}} + \lambda \sum_{i=1}^{N} w_{\text{i}}^2, \tag{1.32}$$

где все параметры аналогичны параметрам в выражении 1.31.

Главное различие между двумя методам заключается в том, что регуляризация Лассо уменьшает коэффициент менее важной характеристики до нуля, полностью удаляя ее из рассмотрения, а регуляризация Тихонова уменьшает веса, но не делает их равными нулю [25].

Еще одним видом регуляризации является дропаут. Суть метода заключается в том, что на каждой итерации обучения нейронной сети все связи между нейронами удаляются с некоторой вероятностью p. Иными словами это означает, что на каждой итерации обучения модели значение каждого веса w_{i} нейронной сети может быть на одну итерацию приравнено к нулю с некоторой вероятностью p.

Таким образом, регуляризация борется с проблемой переобучения нейронной сети и повышает ее обобщающую способность [25]. К недостатками регуляризации можно отнести то, что:

— добавление штрафа или удаление некоторых связей может привести к ухудшению точности модели на обучающих данных;

— нельзя заранее оптимальным образом определить коэффициент штрафования весов λ и вероятность удаления связи между нейронами p.

1.6.5 Нормализация

Обычно при обучении нейронной сети шаг градиентного спуска делается не по одному конкретному примеру, а сразу по некоторому набору обучающих примеров. Такой подход имеет следующие преимущества [25]:

— усреднение градиента по нескольким примерам представляет собой апроксимацию градиента по всему тренировочному множеству, и чем больше примеров используется в одном мини-батче, тем точнее это приближение, использование всего обучающего множества невозможно в силу ограничений вычислительных ресурсов;

— в глубоких нейронный сетях к каждому примеру в отдельности требуется применить большое число последовательных операций, в случае использования некоторого набора обучающей примеров, можно выполнять эти последовательные операции в параллельном режиме для каждого примера в отдельности.

При таком подходе к обучению и использовании глубоких нейронных сетей возникает проблема, связанная с тем, что изменение распределения активаций выходов первых слоев на очередном шаге градиентного спуска приводит к сдвигу распределения данных во всех последующих слоях, что затрудняет их обучение и может ухудшить результаты. Для борьбы с этой проблемой используется пакетная нормализация, которая позволяет нормализовать выходы каждого слоя в процессе обучения. Это делает распределение данных более стабильным и уменьшает влияние сдвига распределения на последующие слои [26]. Такая проблема получила название внутреннего сдвига переменных.

В исследованиях, приведенных в статье [27], говорится, что процесс обучения сходится быстрее, когда входы нейронной сети нормализованы, то есть их математическое ожидание приведено к нулю, а матрица ковариаций – к единичной. Если применять нормализацию к входам каждого слоя, то удастся избежать проблемы внутреннего сдвига переменных.

Для выполнения нормализации требуется предварительно рассчитать математическое ожидание и дисперсию элементов батча, которые определяются выражениями 1.33 и 1.34 соответственно.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_{\text{i}}, \tag{1.33}$$

где μ – математическое ожидание элементов бача, N – размер бача, x_{i} – i-ый элемент бача.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{\text{i}} - \mu)^2, \tag{1.34}$$

где σ – дисперсия элементов бача, N – размер бача, x_{i} – i-ый элемент бача, μ – значение математического ожидания, посчитанное по формуле 1.33.

Тогда нормализацию входов можно проводить, используя выражение 1.35

$$\hat{x_{\text{i}}} = \frac{x_{\text{i}} - \mu}{\sqrt{\sigma^2 + \epsilon}}, \tag{1.35}$$

где $\hat{x_{\text{i}}}$ – нормализованное значение i-го входа, x_{i} – ненормализованное значение i-го входа, μ и σ – математическое ожидание и дисперсия, посчитанные по формулам 1.33 и 1.34 соответственно, ϵ – некоторая константа, которая нужна для предотвращения деления на ноль.

Такая нормализация имеет существенный недостаток: в случае, если в качестве функции активации слоя используется сигмоидальная функция, например логистическая 1.2, то после нормализации нелинейность, которую давала эта функция активации, пропадет, так как большинство значений будут попадать в область, где эта функция ведет себя линейно, и функция активации фактически станет линейной [28].

Для того, чтобы компенсировать этот недостаток, слой нормализации должен быть способен в некоторых случаях практически никак не менять входные значения. Достигается это при помощи введения двух новый коэффициентов: коэффициент масштарбирования и сдвига нормализации. Итоговое выражение для слоя нормализации определяется зависимостью 1.36

$$y_{\text{i}} = \gamma_{\text{i}}\hat{x_{\text{i}}} + \beta_{\text{i}}, \tag{1.36}$$

где y_{i} – i-ый выход слоя нормализации, $\hat{x_{\text{i}}}$ – величина, полученная из выражения 1.35, γ_{i} и β_{i} – коэффициенты масштабирования и сдвига, которые настраиваются во время обучения модели.

Значения математического ожидания и дисперсии во время обучения от батча к батчу будут изменяться, но на этапе тестирования модели все изменяемые параметры должны быть зафиксированы. Для того, чтобы определить значения математического ожидания и дисперсии на этапе тестирования, эти величины накапливаются во время обучения с использованием экспоненциального скользящего среднего, которое определяется зависимостью 1.37

$$EMA_{\mathrm{t}} = \alpha * x_{\mathrm{t}} + (1 - \alpha) * EMA_{\mathrm{t}-1}, \tag{1.37}$$

где EMA_{t} – значение экспоненциального скользящего среднего в точке t, $EMA_{\mathrm{t}-1}$ – значение экспоненциального скользящего среднего в точке t минус 1, причем значение экспоненциального скользящего среднего в нуле EMA_{t} равно x_0, x_{t} – значение исходной функции, в нашем случае это математическое ожидание или дисперсии, в момент времени t, α – коэффициент характеризующий скорость уменьшения весов, принимает значение от 0 и до 1, чем меньше его значение тем больше влияние предыдущих значений на текущую величину среднего.

В случае, когда входной батч описывается кортежем (N, C, H, W), где N – число элементов в батче, C – число каналов в каждом элементе, H и W – высота и ширина каждого изображения, нормализация считается по всем пикселям, всех изображений по каждому из каналов.

1.7 Ансамблевые методы

Ансамблевые методы классификации основаны на том, что несколько классификаторов обучаются на одном и том же наборе обучающих данных, а затем их прогнозы объединяются для классификации элементов тестового набора данных. Математическим обоснованием этой идеи служит теорема Кондорсье о жюри присяжных [29].

Классификатор называется слабым, если его ошибка на обучающей выборке менее 50 процентов, но больше нуля. Тогда, объединив предсказания нескольких таких классификаторов, можно достичь большей точности классификации на элементах тестовой выборки [29].

Выделяют 3 основных метода ансамблевой классификации [29]:

— бэггинг;

— бустинг;

— стекинг.

Идея бэггинга [30] состоит в том, что если размер обучающей выборки не велик, то можно создать много случайных выборок из исходной путем отбора некоторых элементов, и обучить слабые классификаторы на эти подвыборки. Таким образом, каждая модель имеет свой набор обучающих примеров и старается сделать предсказания на основе своего подмножества данных. Затем результаты всех моделей комбинируются для получения итоговых предсказаний.

Бустинг – это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов [29]. Таким образом, при бустинге каждая последующая модель обучается на ошибках предыдущей и старается их компенсировать, повышая точность классификации общей модели.

Идея стекинга [31] заключается в введении некоторого алгоритма классификации и его обучении. При стекинге, в отличие от бустинга и бэггинга, классификаторы должны быть разной природы [29]. Обучение модели при стекинге можно свести к следующим трем шагам:

— обучающая выборка разбивается на две непересекающихся подвыборки;

— первая подвыборка используется для обучения классификаторов;

— вторая для обучения алгоритма, который на вход принимает выходы со всех классификаторов.

Главным недостатком стекинга является деление обучающей выборки на две части.

1.8 Формализованная постановка задачи

Цель работы – разработать метод замещения страниц в разделяемом кэш буфере Postgres с использованием нейронных сетей.

Для достижения поставленной цели требуется выполнить следующие задачи:

— сравнить существующие методы замещения страниц;

— описать и спроектировать метод замещения страниц с использованием нейронных сетей;

— разработать программное обеспечение для предложенного метода;

— провести сравнение разработанного метода с существующими аналогами по коэффициентам совпадения и попадания.

На вход методу подается атрибуты страницы, к которой происходит обращение, и атрибуты всех страниц, которые уже находятся в буфере. Результатом работы метода является индекс страницы для замещения в буфере.

На входные данные накладываются следующие ограничения:

— размер буфера совпадает с тем, на котором обучалась модель;

— число страниц в буфере не больше 256;

— шаблон обращений к страницам совпадает с тем, на котором обучалась модель.

На рисунке 1.10 приведена IDEF-0 диаграмма уровня A0 метода замещения страниц в разделяемом кэш буфере Postgres с использованием нейронных сетей.

Рисунок 1.10 – IDEF-0 диаграмма метода замещения страниц

1.9 Вывод

Управление памятью в операционных системах и СУБД, таких как PostgreSQL, основано на схожих принципах минимизации задержек при работе с диском. В операционных системах ключевую роль играет виртуальная память, использующая страничную организацию и механизмы трансляции адресов, что позволяет эффективно распределять физическую память между процессами и избегать её переполнения за счёт выгрузки редко используемых страниц на диск.

В PostgreSQL управление памятью адаптировано под специфику работы с базами данных. Разделяемый буферный кэш служит для хранения часто используемых страниц данных в оперативной памяти, что сокращает количество обращений к диску. Однако это порождает проблему двойного кэширования, так как ОС также использует свой дисковый кэш. PostgreSQL частично решает её, минимизируя взаимодействие с кэшем ОС: пока страница находится в буфере СУБД, обращения к ней идут напрямую.

Алгоритмы замещения страниц предлагают различные стратегии баланса между производительностью и ресурсозатратностью. Теоретически оптимальный алгоритм демонстрирует максимальную эффективность, удаляя страницу с самым отдалённым обращением, но его практическая реализация невозможна из-за отсутствия данных о будущих запросах. Более простые методы, такие как NRU и FIFO с модификациями («второй шанс», «часы»), используют биты обращения и циклические списки для минимизации накладных расходов, однако их эффективность ограничена в динамичных сценариях.

Алгоритмы LRU с механизмом старения и WSClock учитывают не только количество обращений, но и время последнего обращения.

Эти методы демонстрируют удовлетворительную производительность в общих сценариях, но их эффективность снижается в условиях динамичных или нестандартных шаблонов доступа к данным. Оптимальный алгоритм является хорошей метрикой для оценки алгоритмов замещения страниц – чем больше различие результатов разработанного метода и оптимального, тем больше возможностей для улучшения разработанного алгоритма. Таким образом, при разработке метода необходимо стараться приблизить его к результатам, которые выдает оптимальный алгоритм.

Нейронные сети предлагают перспективное решение для адаптивного управления кэшем. Их ключевое преимущество – способность обучаться на истории обращений к страницам, выявляя скрытые закономерности и прогнозируя востребованность данных. Это позволяет приблизить стратегию замещения к оптимальному алгоритму.

Для запоминания последовательности обращений к страницам нужно использовать рекуррентные нейронные сети, так как они сохраняют информацию о предыдущих состояниях, что позволяет выявлять скрытые временные зависимости и прогнозировать востребованность страниц на основе контекста.

Для решения проблемы переобучения нейронных сетей нужно использовать: регуляризацию, которая вводит дополнительные ограничения на большие веса модели, нормализацию, которая приводит значения входных признаков к одному диапазону, а также метод раннего останова, который отслеживает момент, когда модель начинает переобучаться, и прекращает обучение в этот момент.

2 Конструкторский раздел

2.1 Входные данные

На вход методу подается атрибуты страницы, к которой происходит обращение, и атрибуты всех страниц, которые уже находятся в буфере.

Для сохранения истории обращений к кэш буферу в функцию ReadBufferExtended был добавлен вызов функции печати в лог файл атрибутов страницы, к которой идет обращение. Эта функция вызывается, каждый раз, когда необходимо прочитать страницу из буфера.

Для каждой страницы извлекается следующий набор атрибутов:

— идентификатор отношения;

— номер страницы в файле;

— наличие индекса;

— позиция в буфере.

Если в момент обращения к странице она не находится в буфере, то позиция задается размером буфера.

2.2 Проектирование метода

Детализированная IDEF0 диаграмма метода замещения страниц уровня A0 приведена на рисунке 2.1.

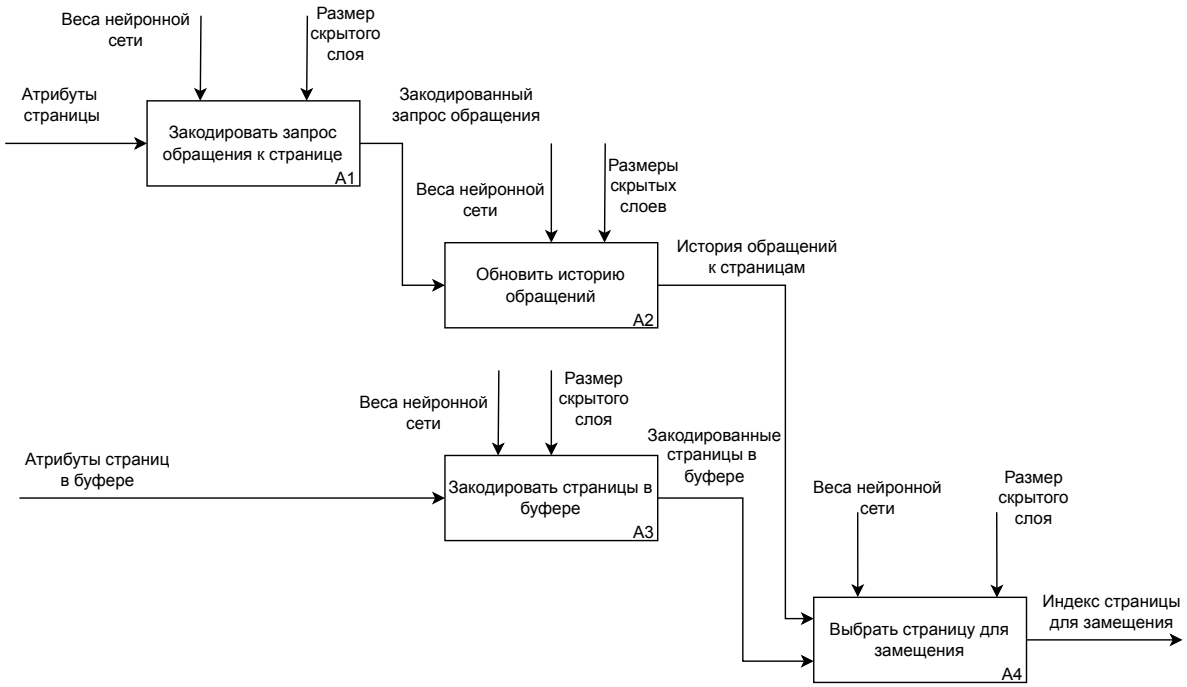

Рисунок 2.1 – Детализированная IDEF0 диаграмма

Кодировщик запроса обращения к странице отвечает за скрытое представление атрибутов страницы, к которой происходит очередное обращение. Схема кодировщика запроса обращения к странице изображена на рисунке 2.2.

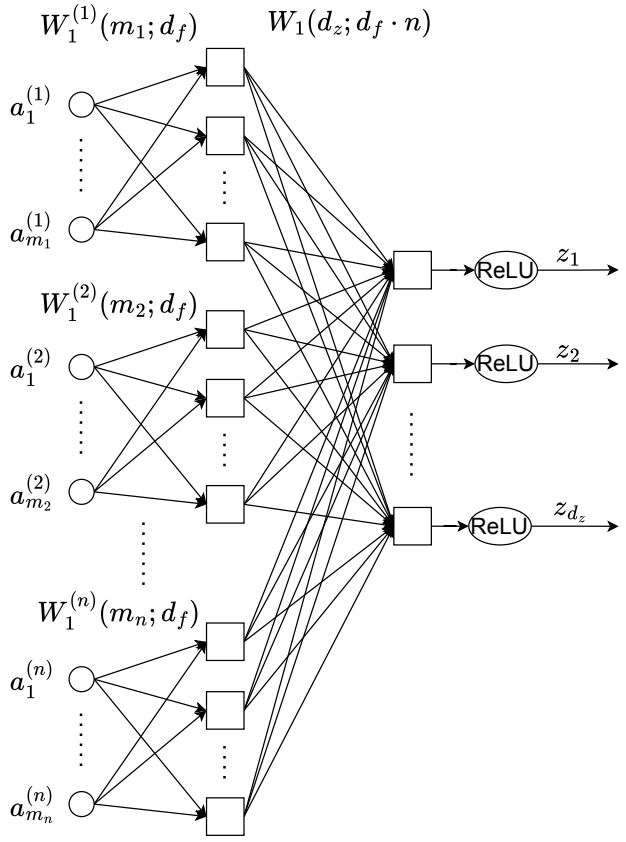

Рисунок 2.2 – Схема кодировщика запроса обращения к странице

На вход кодировщика поступают n атрибутов страницы. Каждый атрибут может иметь m_i возможных значений, где i – индекс атрибута. Каждый атрибут представляется в виде вектора $a^{(i)}$ размерности m_i. Для категориальных данных используется техника однозначного кодирования, а для числовых – применяется хэш функция и к полученному результату применяется техника однозначного кодирования. $W_1^{(i)}$ – матрица обучаемых весов для скрытого представления i-го атрибута. Вектор z – выходной вектор из сети. W_1 – матрица обучаемых весов, при помощи который получается результирующий вектор из скрытых представлений атрибутов сети. В качестве функции активации на последнем слое используется функция Relu. d_f и d_z являются настраиваемыми параметрами, которые отвечают за число нейронов, отвечающий за скрытое представление каждого атрибута, и число нейронов на выходном слое соответственно.

Работу сети можно описать с помощью выражений 2.1 - 2.3:

$$f^{(i)} = a^{(i)} W_1^{(i)} i \in \{1; n\}, \tag{2.1}$$

$$f = [f^{(1)}, f^{(2)}, ..., f^{(n)}], \tag{2.2}$$

$$z = ReLU(W_1 f^T + l_1), \tag{2.3}$$

где f является конкатенацией векторов скрытых состояний атрибутов страницы, а l_1 – обучаемым вектором.

Кодировщик страниц в буфере нужен для скрытого представления каждой страницы в буфере. Схема кодировщика представлена на рисунке 2.3

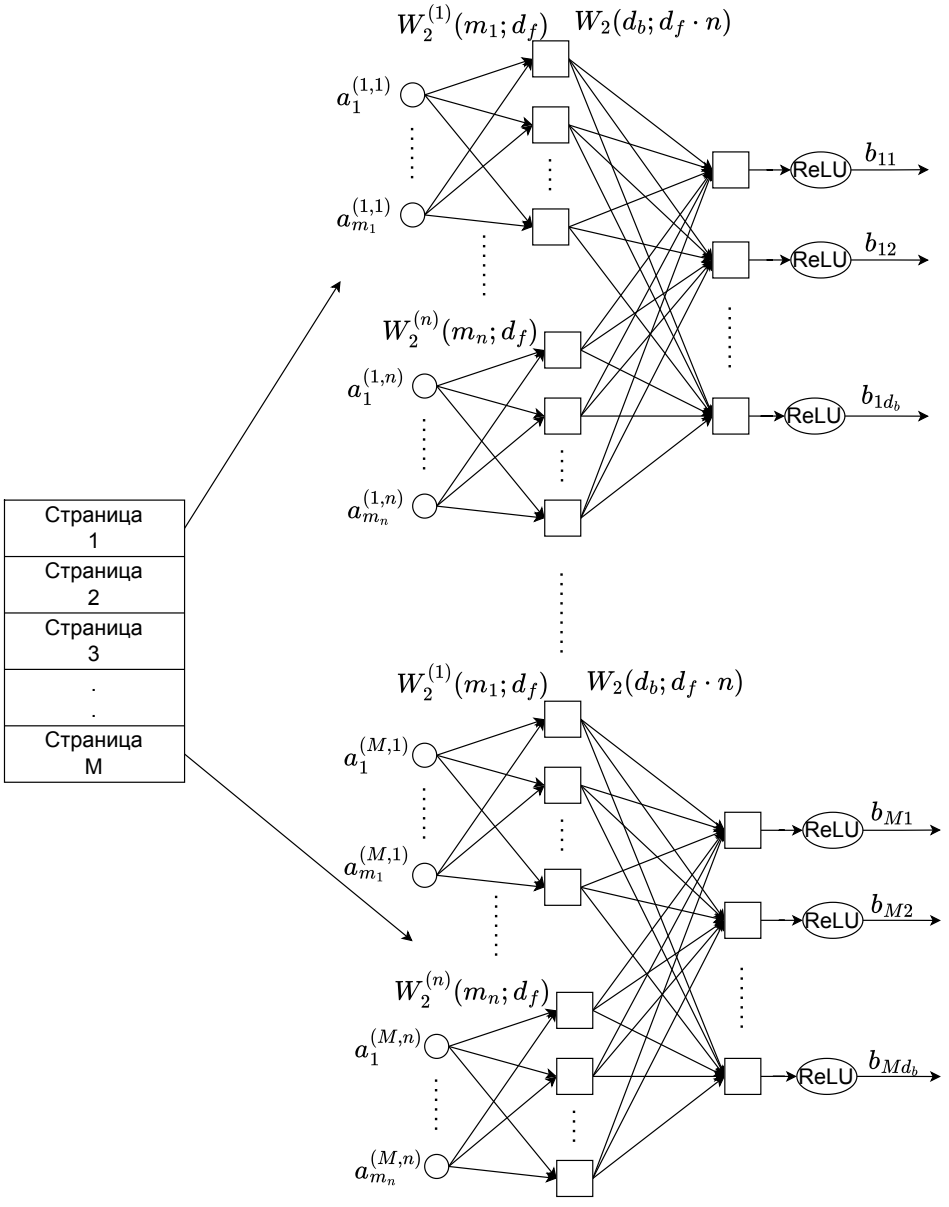

Рисунок 2.3 – Схема кодировщика страниц в буфере

На вход кодировщика поступают M страниц из буфера. Каждая страница представляется в виде n атрибутов. Процесс обработки атрибутов для каждой страницы такой же, как и в кодировщике запроса обращения к странице. Для каждой i-ой страницы в буфере вычисляется вектор b_i. d_b является настраиваемым параметром, который отвечает за размерность векторов b_i.

Для получения скрытого представления каждого атрибута и для вычисления закодированного представления страницы используется одни и те же матрицы весов $W_2^{(i)}$ и W_2 для всех страниц в буфере. За счет этого матрицы весов не привязаны к конкретной позиции страницы в буфере и истории страниц на этой позиции. При обратном распространении ошибки влияние веса из матрицы W_2 будет учитываться для всех векторов b_i.

Обозначим результат работы сумматора нейрона на выходном слое как s_{ij}. Индексация в матрице s совпадает с матрицей b. Тогда для вычисления ошибки по весу w_{ij} из матрицы W_2 на ребре, которое соединяет j-ый нейрон из второго слоя и i-ый нейрон из выходного слоя, используется выражение 2.4:

$$\frac{\delta E}{\delta w_{ij}} = \sum_{k=1}^{M} \frac{\delta E}{\delta b_{ki}} \frac{\delta b_{ki}}{\delta s_{ki}} \frac{\delta s_{ki}}{\delta w_i j}, \tag{2.4}$$

где E – функция ошибки, $\frac{\delta E}{\delta b_{ki}}$ – ошибка полученная со следующего слоя.

Функционирование кодировщика определяется выражениями 2.5 - 2.7:

$$f^{(j,i)} = a^{(j,i)} W_2^{(i)} j \in \{1; M\} i \in \{1; n\}, \tag{2.5}$$

$$f^{(j)} = [f^{(j,1)}, f^{(j,2)}, ..., f^{(j,n)}], \tag{2.6}$$

$$b_j = ReLU(W_2 f^{(j)T} + l_2), \tag{2.7}$$

где $f^{(j)}$ – конкатенация скрытых представлений атрибутов для j-ой страницы в буфере, b_j – скрытое представление этой страницы, l_2 – вектор обучаемых весов.

Кодировщик истории обращений. Для обновления истории обращений используется сеть LSTM. На вход сети поступают результат работы кодировщика обращения к странице, предыдущий результат кодировщика истории обращений и предыдущее состояние ячейки.

Функционирование кодировщика определяется выражениями 2.8 - 2.13:

$$f_t = \sigma(W_f[h_{t-1}, z_t] + b_f), \tag{2.8}$$

$$i_t = \sigma(W_i[h_{t-1}, z_t] + b_i), \tag{2.9}$$

$$\hat{C}_t = \tanh(W_C[h_{t-1}, z_t] + b_C), \tag{2.10}$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t, \tag{2.11}$$

$$o_t = \sigma(W_o[h_{t-1}, z_t] + b_o), \tag{2.12}$$

$$h_t = o_t * \tanh(C_t), \tag{2.13}$$

где $[h_{t-1}, z_t]$ – конкатенация результата работы предыдущего слоя кодировщика истории и скрытого состояния, полученного из кодировщика обращения к странице, W_f и b_f – матрица и вектор обучаемых весов, f_t – результат работы фильтра забывания, i_t определяет, какие значения будут сохранены в ячейке, $\hat{C}_t$ – новые значения кандидатов на попадание в ячейку, W_i, W_C, b_i, b_c – матрицы и вектора обучаемых весов, C_t – новое состояние ячейки, C_{t-1} – состояние ячейки на прошлом шаге, h_t – результат работы текущего слоя, C_t – состояние ячейки, W_o и b_o – матрица и вектор обучаемых весов. Вектора h_t и C_t имеют размерность d_h, где d_h – настраиваемый параметр.

Модуль выбора страниц для замещения. На вход модуля поступают результаты работы кодировщика страниц в буфере и кодировщика истории обращений. Для выбора страницы, которая будет удалена из буфера используется указательная нейронная сеть с механизмом внимания [32].

Нейронные сети с механизмом внимания – это архитектуры, которые позволяют моделям динамически фокусироваться на наиболее релевантных частях входных данных при обработке информации. Этот подход нашел применения в областях обработки естественного языка, компьютерного зрения и других задач, где важно учитывать контекст и зависимости между элементами последовательности. В модуле выбора страниц для замещения

контекстом является результат работы кодировщика истории, а элементами последовательности – результаты работы кодировщика страниц в буфере.

Указательные сети – архитектура сетей с механизмом внимания, предназначенная для решения задач, где выходные элементы представляют собой позиции в входной последовательности. В указательных сетях механизм внимания используется как указатель на один из элементов входной последовательности, а не для создания контекстного вектора, как в классических моделях с механизмом внимания.

Схема модуля выбора страниц для замещения представлена на рисунке 2.4

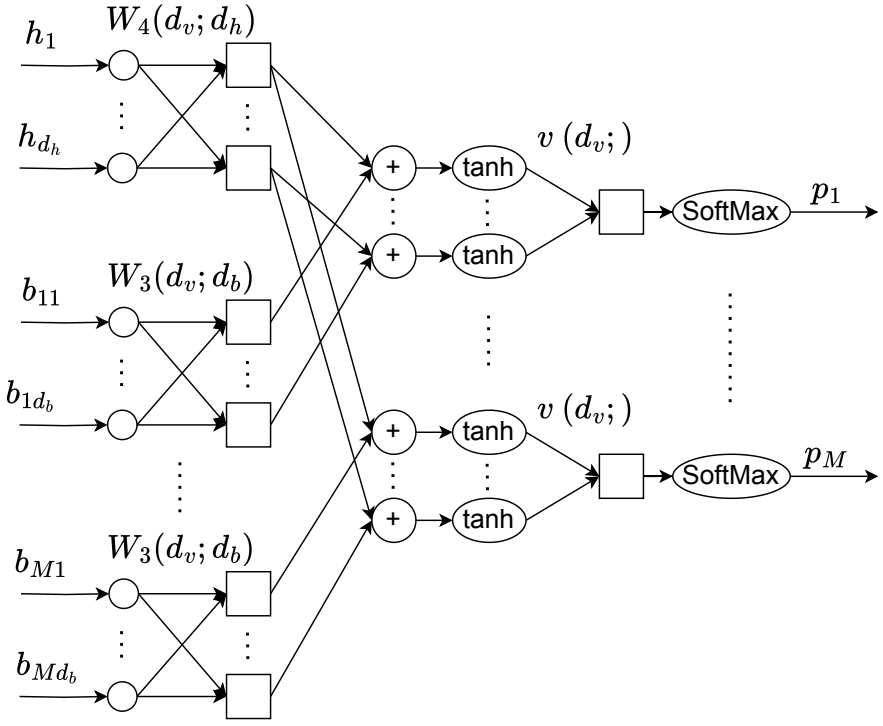

Рисунок 2.4 – Схема модуля выбора страниц для замещения

W_4 – матрица обучаемых весов, которая используется для преобразования вектора h, полученного из кодировщика истории в вектор контекста размерности d_v. d_v является настраиваемым параметром модели.

W_3 – матрица обучаемых весов, которая используется для преобразования закодированного состояния очередной страницы в буфере в вектор размерности d_v, который будет использован в функции внимания.

v – вектор обучаемых весов, который который используется при вычислении функции внимания.

Функционирование модуля выбора страниц для замещения определяется выражениями 2.14 - 2.16:

$$u_i = v * \tanh(W_3 b_i^T + W_4 h^T), i \in \{1; M\}, \tag{2.14}$$

$$p_i = SoftMax(u_i), \tag{2.15}$$

$$r = \arg\max_i p_i, \tag{2.16}$$

где M – число страниц в буфере, u_i – результат функции внимания для i-ой страницы в буфере, $\arg\max_i p_i$ – функция, которая возвращает индекс максимального элемента в последовательности, r – результат работы спроектированного метода замещения страниц.

Схемы алгоритмов обучения нейронных сетей и прохождения одной эпохи приведены на рисунках 2.5 и 2.6 соответственно.

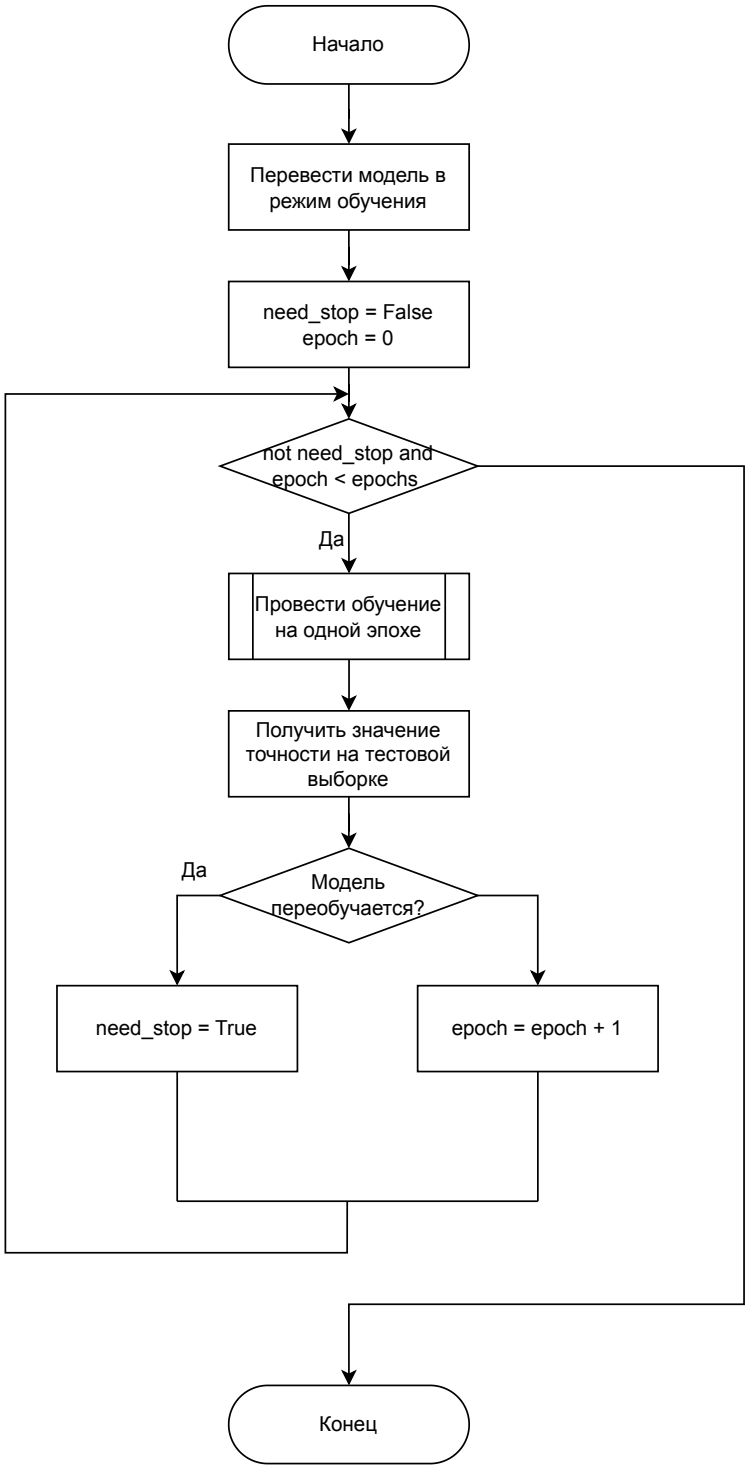

Рисунок 2.5 – Схема алгоритма обучения нейронной сети

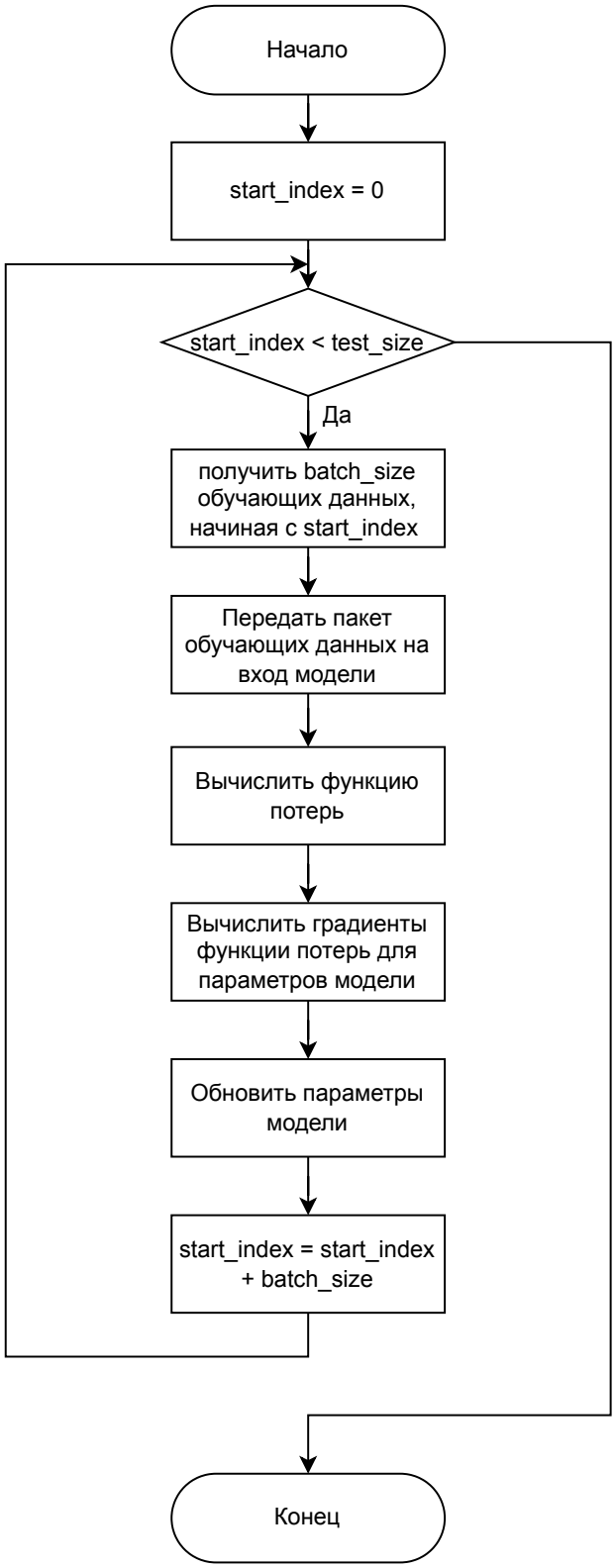

Рисунок 2.6 – Схема алгоритма прохождения одной эпохи

2.3 Структура программного обеспечения

Программное обеспечение состоит из пяти модулей:

— модуль получения обучающей выборки;

— кодировщик запросов обращения к страницам;

— кодировщик истории запросов обращения к страницам;

— кодировщик страниц в буфере;

— модуль выбора страниц для замещения.

Структурная схема взаимодействия модулей разрабатываемого программного обеспечения представлена на рисунке 2.7.

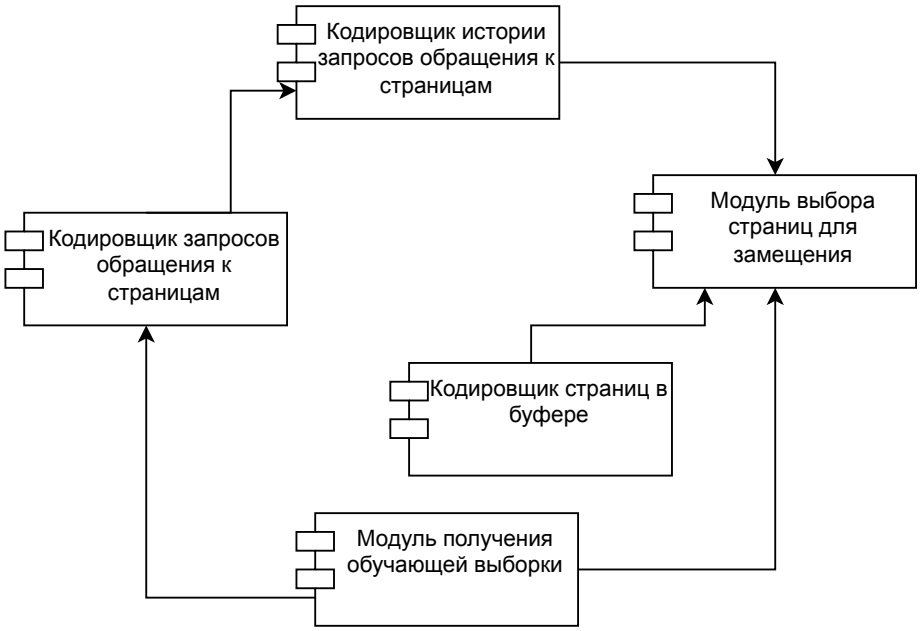

Рисунок 2.7 – Структура программного обеспечения

Модуль получения обучающей выборки нужен для обработки лог файла и создания обучающей и тестовой выборок. Остальные модули предназначены для обучения и взаимодействия с уже обученными моделями.

2.4 Набор обучающих данных

Для имитации нагрузки на СУБД и получения истории обращений к страницам был использован тестовый сценарий TPC-C [33].

TPC-C – это стандартный тест для оценки производительности систем, обрабатывающих транзакции в режиме реального времени. Он имитирует работу оптового поставщика с распределённой сетью складов и предназначен для тестирования СУБД на реалистичных сценариях высокой нагрузки.

Для обработки нагрузки создается база данных, состоящая из следующих таблиц:

— Warehouse (склады);

— District (регионы складов);

— Customer (клиенты);

— Order (заказы);

— Order-Line (позиции заказов);

— Item (товары);

— Stock (запасы на складах);

— History (история платежей).

TPC-C включает 5 типов транзакций, имитирующих реальные бизнес-операции:

1. NewOrder (45%): создание нового заказа (вставка данных в таблицы Order, Order-Line, обновление Stock).

2. Payment (43%): Обработка платежа (обновление Customer, Warehouse, District, вставка в History).

3. Delivery (4%): Доставка заказа (удаление из Order, обновление Customer).

4. OrderStatus (4%): Проверка статуса заказа (выборка из Order, Order-Line, Customer).

5. StockLevel (4%): Проверка уровня запасов (выборка из Stock).

С помощью тестовой нагрузки было получено 6 554 959 обращений к страницам. Выборка была поделена на тренировочную и тестовую в отношении семь к трем.

Корреляция между атрибутами страниц, посчитанная на всей выборке, представлена на рисунке 2.8.

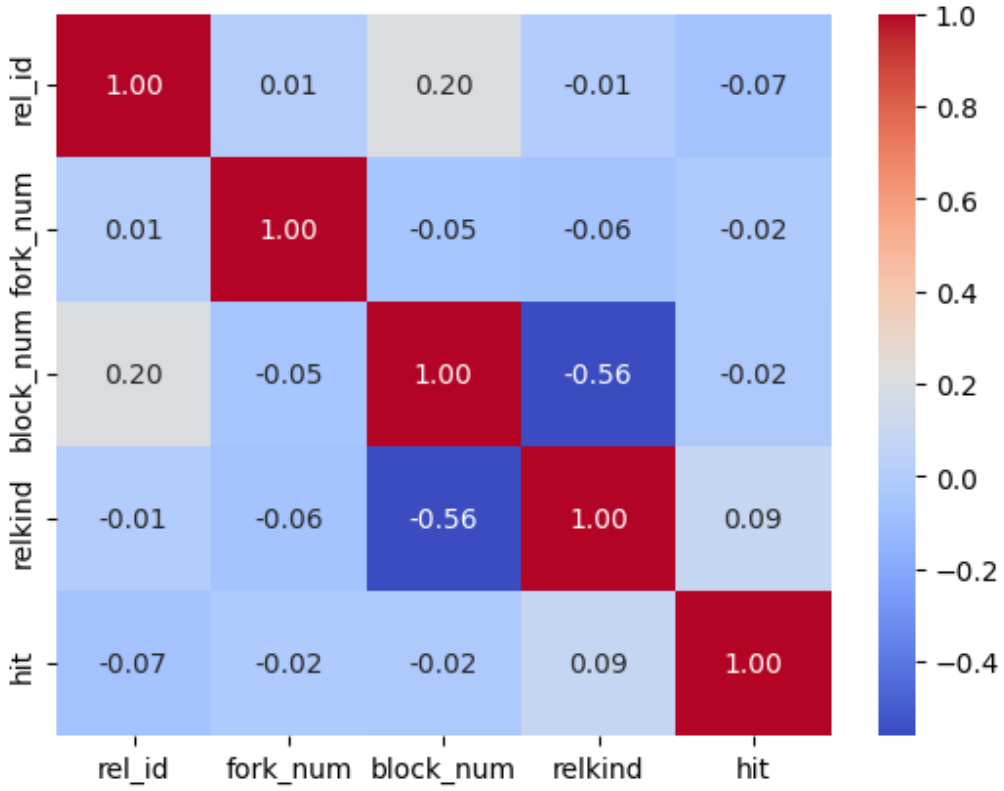

Рисунок 2.8 – Корреляция между атрибутами страниц

Корреляция вычислялась по формуле Пирсона 2.17:

$$c = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{2.17}$$

где n – число элементов в последовательности, x_i и y_i – элементы последовательностей, между которыми считается корреляция, $\overline{x}$ и $\overline{y}$ – средние значения элементов последовательностей.

Количество замещений страницы оптимальным алгоритмом в зависимости от индекса страницы в буфере для тренировочной и тестовой выборок

приведено на рисунках 2.9 и 2.10 соответственно.

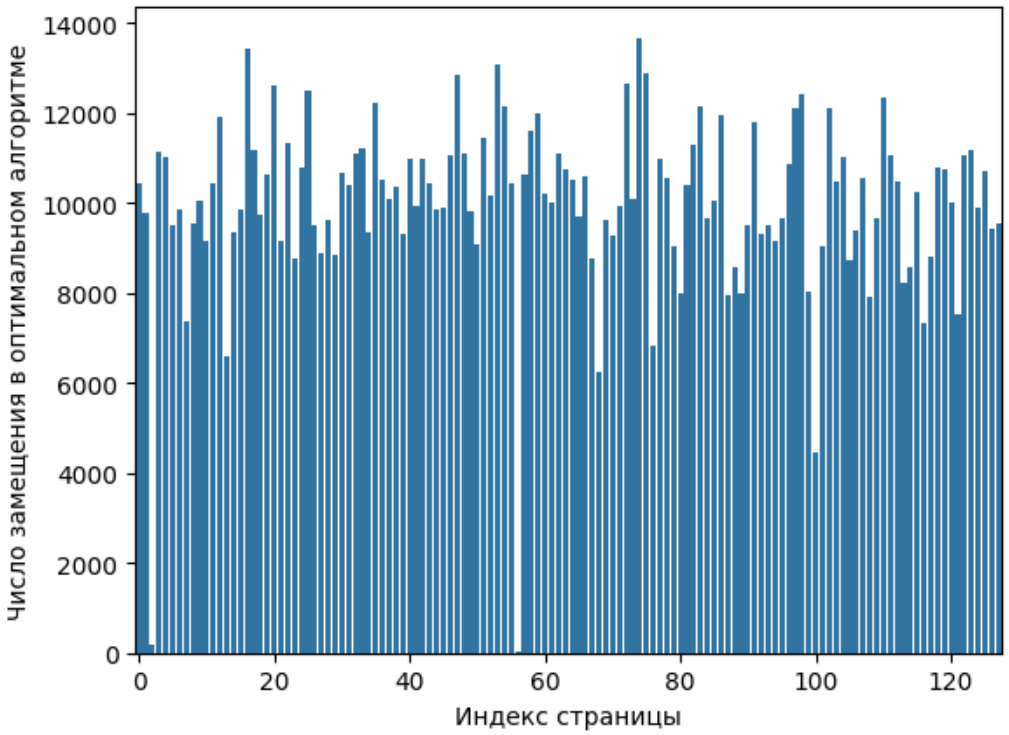

Рисунок 2.9 – Количество замещений страницы по индексу на тренировочной выборке

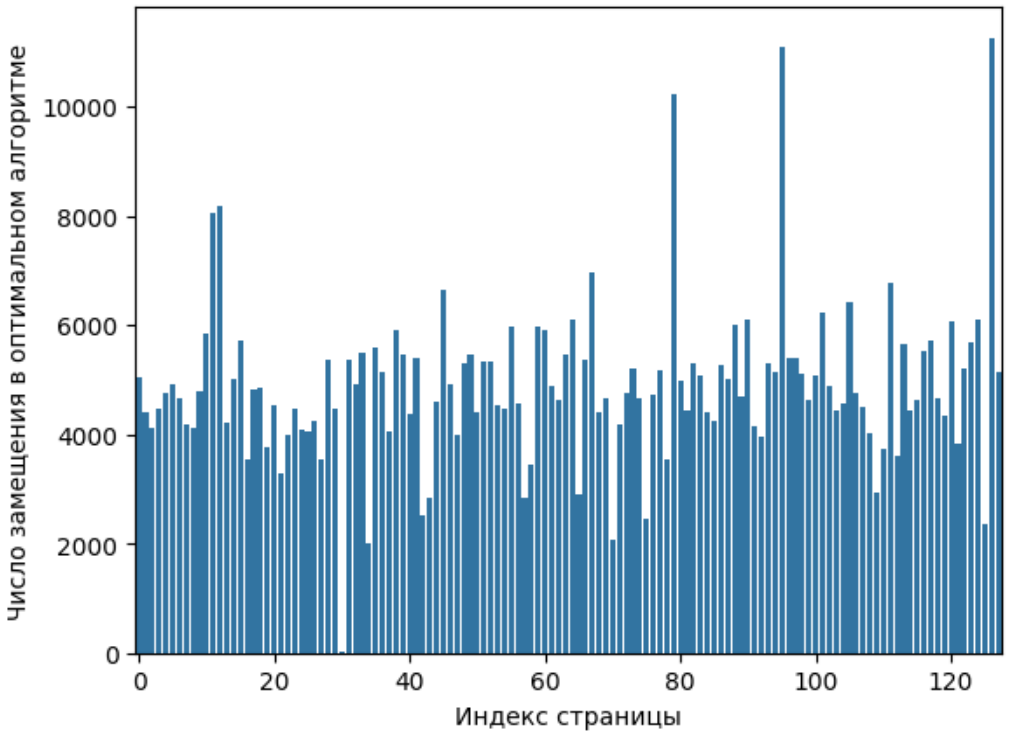

Рисунок 2.10 – Количество замещений страницы по индексу на тестовой выборке

2.5 Вывод

В данном разделе были определены ограничения, которые накладываются на входные данные, и требования, которые предъявляются к разрабатываемому программному обеспечению.

Была детализирована IDEF-0 диаграмма уровня A0, описанная в разделе формализованной постановки задачи, а также было проведено разбиение программного обеспечения на модули. Задача выбора страниц для замещения из буфера была разбита на четыре подзадачи: кодирование запроса обращения к странице, обновление истории обращений, кодирование страниц в буфере, выбор страниц для замещения на основе контекста.

Для решения каждой подзадачи была спроектирована архитектура нейронной сети. Для обеспечения временных зависимостей при обновлении истории была выбрана сеть LSTM. Для выбора страницы для замещения была спроектирована указательная сеть с механизмом внимания, которая выбирает одну из страниц в буфере для замещения. Была определена схема алгоритма обучения составленных нейронных сетей.

Для генерации реалистичной нагрузки использован тестовый сценарий TPC-C, что позволило получить шесть с половинной миллионов обращений к страницам. Распределение замещений страниц на тренировочной и тестовой выборках подтвердило сбалансированность данных.

3 Технологический раздел

3.1 Средства реализации программного обеспечения

В качестве языка программирования был выбран Python [34]. Данный выбор обусловлен тем, что Python имеет множество библиотек, таких как TensorFlow, Keras, PyTorch, которые предоставляют множество инструментов для создания и обучения нейронных сетей.

В качестве библиотеки для создания нейронной сети была выбрана библиотека PyTorch [35] версии 2.0.0, так как она имеет следующие возможности и инструменты:

— динамический граф вычислений, использование которого облегчает отладку моделей;

— возможность переноса вычислений на GPU;

— набор инструментов для создания различных слоев, из которых складывается архитектура нейронной сети;

— API на языке C++, что позволяет обучить модель с использованием интерпретируемого языка Python, а использовать ее на компилируемом языке C++.

Для работы с большими данными была выбрана библиотека numpy версии 1.21.0, так как она использует оптимизированный код на C, что позволяет выполнять вычисления быстрее, чем с использованием чистого Python, а также потому что классы этой библиотеки интегрируются с библиотекой PyTorch, которая используется для создания нейронной сети.

Для анализа входных данных использовались библиотеки matplotlib, pandas и seaborn. Pandas поддерживает чтение данных из различных форматов данных в структуру DataFrame – таблицу с индексами и метками, а также совместимость с другими библиотеками, такими как numpy, для передачи данных. Matplotlib имеет возможность создания различных видов графиков: линейные, столбчатые, гистограммы, а также имеет интеграцию с Jupyter Notebook для интерактивной визуализации. Seaborn позволяет строить карту корреляций для элементов Dataframe.

Для создания графического интерфейса был использован фреймворк PyQT [36], так как он является кроссплатформенным, имеет собственную библиотеку стандартных виджетов, имеет документацию по всем структурам, а также поддерживает последние стандарты языка Python и имеет дополнительные утилиты такие как QtDesigner, которые упрощают создание графических интерфейсов.

3.2 Разработка программного комплекса

Для создания обучающей выборки необходимо провести следующую временную модификацию в СУБД PostgreSQL: нужно найти функцию, которая вызывается каждый раз при обращении к страницам в буфере и добавить в эту функцию вызов функции для записи информации о странице в специальный файл.

Для управления разделяемым кэш буфером используются следующие функции:

1. ReadBuffer – захватывает буфер, увеличивает его pin count, и загружает страницу в буфер, если её там нет.
2. ReleaseBuffer – уменьшает pin count буфера, освобождая его для возможного повторного использования.
3. LockBuffer, LockBufferForCleanUp, ConditionalLockBufferForCleanUp – управление блокировками.
4. BgBufferSync – фоновая запись измененных буферов на диск.
5. CheckPoint – управление контрольными точками.
6. MarkBufferDirty – помечает буфер, как требующий записи на диск.
7. FlushBuffer – сбрасывает содержимое буфера на диск.
8. BufferAlloc – выделяет буфер для новой страницы.

Функция ReadBuffer принимает указатель на структуру отношения, для которого читается блок данных, и номер блока для чтения. Внутри функции

ReadBuffer происходит вызов функции ReadBufferExtended, которая дополнительно принимает имя физического хранилища, способ чтения и стратегию управления буфером. Вся логика обращений и изменений разделяемого кэш буфера написана внутри функции ReadBufferExtended. Эта функция вызывается с различными аргументами еще из 66 мест, поэтому именно в нее надо добавлять запись в лог файл с информацией об обращении к странице. Модифицированный код функции ReadBufferExtended приведен в листинге А.1 (приложение А).

Для реализации кодировщика запросов обращения к странице был разработан класс PageAccEncoder. Класс является наследником класса Module из библиотеки PyTorch:

— слои эмбеддиногов, которые преобразуют атрибуты страницы в некоторые скрытые вектора;

— полносвязная сеть, которая объединяет результаты эмбедингов для каждого атрибута страницы и создает скрытое представление страницы.

Для целочисленных атрибутов применяется модель HashEmbedding, которая сначала вычисляет хэш функцию, а затем к полученному значению применяет слой эмбеддинга. Для категориальных атрибутов он применяется сразу.

Полносвязный слой является последовательным применением Linear блока и функции активации ReLU. Реализация класса PageAccEncoder представлена в листинге А.2 (приложение А).

Для реализации кодировщика страниц в буфере был разработан класс PageBufferEncoder. Этот класс является наследником PageAccEncoder, так как с каждой страницей в буфере требуется выполнить те же действия, что и с страницей, к которой идет новое обращение. Для параллелизации и ускорения процессов взаимодействия с моделью перед применением слое модели для всех атрибутов страниц в буфере вызывается функция torch.stack, которая нужная для конкатенации тензоров вдоль новой оси. Реализация класса PageBufferEncoder представлена в листинге А.3.

В качестве кодировщика истории обращений используется библиотечная модель LSTM.

Для реализации модуля выбора страниц для замещения был разработан класс PageEviction, который наследуется от класса Module из библиотеки PyTorch. Модель состоит из трех линейных слоев:

— attention_page – преобразует скрытое представление страницы в буфере в пространство внимания;

— attention_context – преобразует результат работы кодировщика истории в это же пространство;

— attention_v – вычисляет итоговые оценки внимания на основе объединенных признаков из пространства внимания.

Реализация класса PageEviction представлена в листинге А.4.

Класс PageAccModel объединяет в себе все 4 описанных выше модуля. Процесс обработки входных данных и выбора страницы для замещения состоит из четырех шагов:

— кодирование запроса обращения к странице – используется модель PageAccEncoder;

— обновление истории обращений – используется модель LSTM;

— кодирование страниц в буфере – используется модель PageBufferEncoder;

— выбор страницы для замещения – используется модуль PageEviction.

Реализация класса PageAccModel представлена в листинге А.5.

3.3 Обучение и тестирование модели

Обучение модели проводилось на машине с процессором Intel Core i9-10900, 64 гигабайтами оперативной памяти и графической картой NVIDIA GeForce RTX 3080 с 16 гигабайтами памяти типа GDDR6.

В качестве оптимизатора функции потерь был выбран Adam, так как он автоматически адаптирует скорость обучения для каждого параметра в зависимости от его градиента, что позволяет более эффективно использовать скорость обучения и ускоряет сходимость.

Обучение модели проводилось на протяжении 100 эпох. После прохождения каждой эпохи веса модели сохранялись в файл и вычислялась точность модели на тестовой выборке. Была выбрана модель с наивысшей точностью на тестовой выборке.

Графики зависимостей точности модели на тестовой и обучающей выборках от номера эпохи обучения приведены на рисунке 3.1.

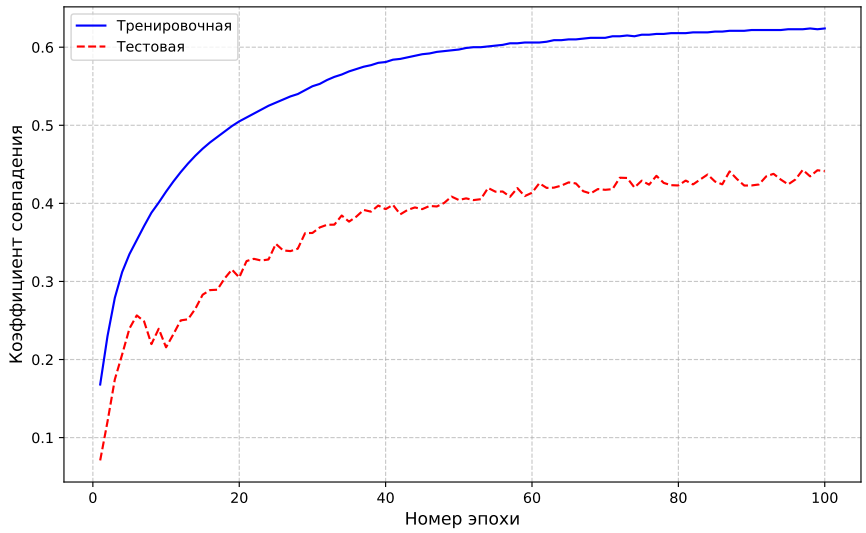

Рисунок 3.1 – Точность при обучении модели на тренировочной и тестовой выборках

Наивысшая точность модели была получена на 97 эпохе – 44.2 процента. Точность на обучающей выборке составила 63 процента.

3.4 Взаимодействие с разработанным ПО

Взаимодействие с разработанным программным обеспечение осуществляется через графический пользовательский интерфейс. Интерфейс приложения представлен на рисунке 3.2.

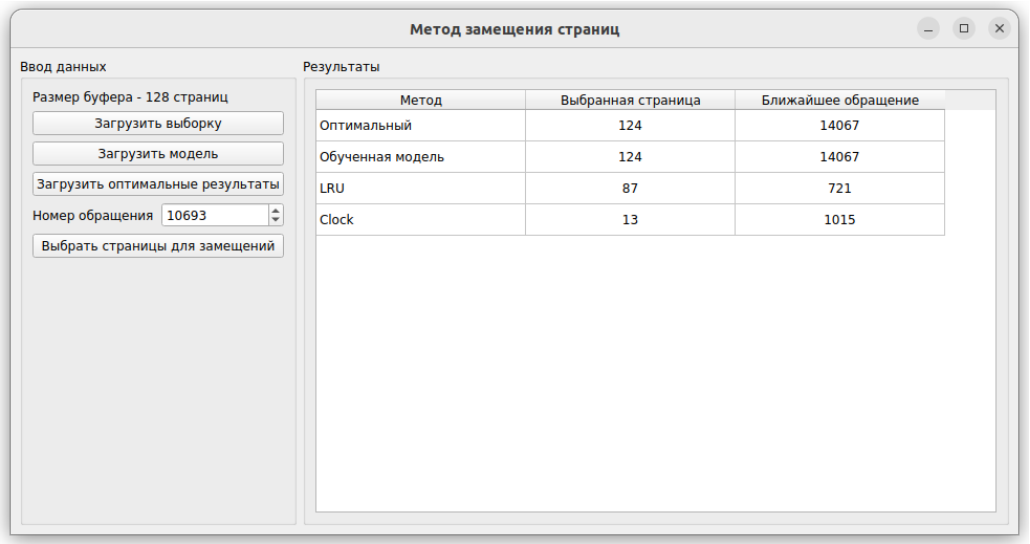

Рисунок 3.2 – Интерфейс разработанного приложения

Графический интерфейс поделен на две части. Первая часть позволяет выбрать и загрузить выборку для тестирования, обученную модель и оптимальные результаты, а также выбрать номер обращения, на котором требуется сравнить алгоритмы замещения. Вторая часть является таблицей, которая отображает результаты работы методов. В первом столбце таблице написано название метода, затем индекс выбранной методом страницы для замещения, затем число, которое показывает через сколько обращений к буферу выбранную страницу необходимо будет загрузить обратно.

3.5 Вывод

В рамках данного раздела были выбраны средства реализации программного обеспечения метода замещения страниц с использованием нейронных сетей. В качестве языка программирования был выбран Python, а для проектирования и обучения нейронных сетей была использована библиотека PyTorch.

Также был проведен анализ функций в СУБД PostgreSQL и проведена модификация одной из функций для созданий обучающей выборки.

Были созданы и обучены модели, которые решают свои подзадачи в рамках метода замещения страниц: кодирование запросов обращения к страницам, кодирование страниц в буфере, кодирование истории обращений, а также модель для выбора страницы для замещения на основе скрытого

представления страниц в буфере и контекста, полученного из кодировщика истории.

Точность совпадений с оптимальным алгоритмом на обучающей выборке составила 63 процента, а на тестовой – 44.2 процента.

4 Исследовательский раздел

4.1 Подбор параметров сети

Для оценки разработанного метода вводятся следующие метрики качества:

— коэффициент попадания – отношение числа обращений к страницам, которые уже загружены в буфер, к общему числу обращений;

— коэффициент совпадения – отношение количества совпавших с оптимальным алгоритмов кандидатов на замещение с общим числом запросов поиска страниц для вытеснения.

Размер скрытых слоев модели подбирался экспериментально. Графики зависимости коэффициента совпадения в зависимости от эпохи обучения для различных размеров скрытых слоев на обучающей и тестовой выборках представлены на рисунках 4.1 и 4.2 соответственно.

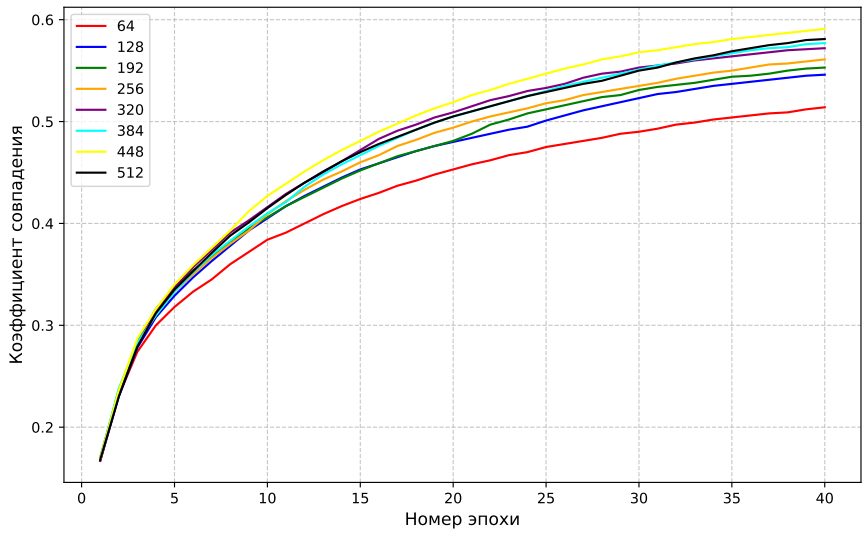

Рисунок 4.1 – Точность модели для различных размеров скрытых слое на тренировочной выборке

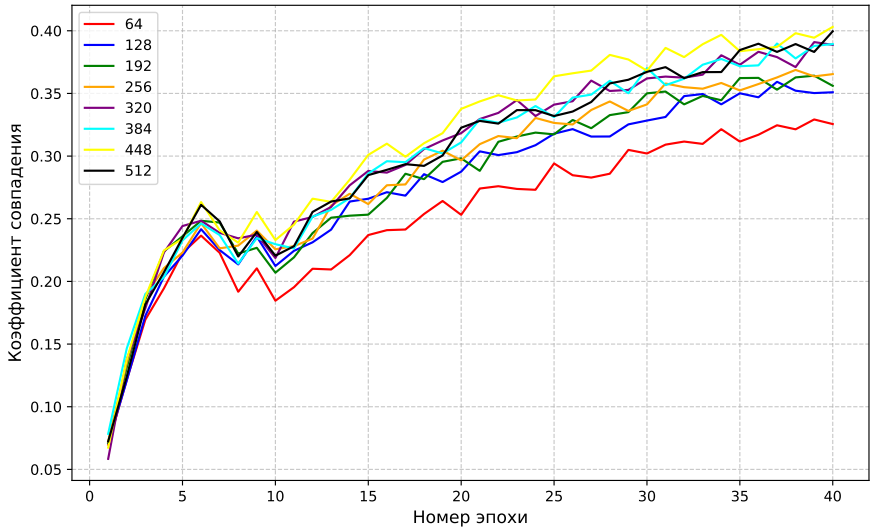

Рисунок 4.2 – Точность модели для различных размеров скрытых слое на тестовой выборке

Исходя из полученных результатов, настраиваемые параметры модели: d_z, d_b, d_h и d_v были выбраны равными 448, а d_f – 32.

4.2 Сравнение с аналогами

Сравнение коэффициентов попадания для разработанного метода и существующих аналогов приведено на рисунке 4.3.

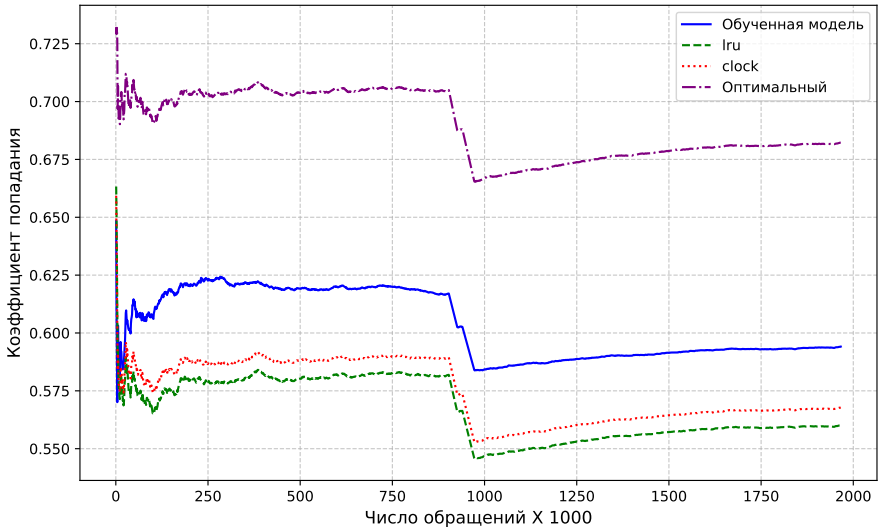

Рисунок 4.3 – Коэффициент попадания в зависимости от числа обращений для различных методов

Из графиков видно, что коэффициент попадания для разработанного метода в среднем на 0.02 выше чем для алгоритма clock, который в настоящее время используется в PostgreSQL. Также коэффициент попадания для разработанного метода на 0.08 ниже, чем у оптимального алгоритма. Таким образом, разработанный метод лучше существующий аналогов, но все еще имеет возможность для улучшения.

4.3 Сравнение различных размеров буфера

Было проведено сравнение точности модели на тренировочной и тестовой выборках при различных размерах буфера: 64, 128 и 256 страниц. Результаты для тренировочной и тестовой выборках представлена на рисунках 4.4 и 4.5 соответственно.

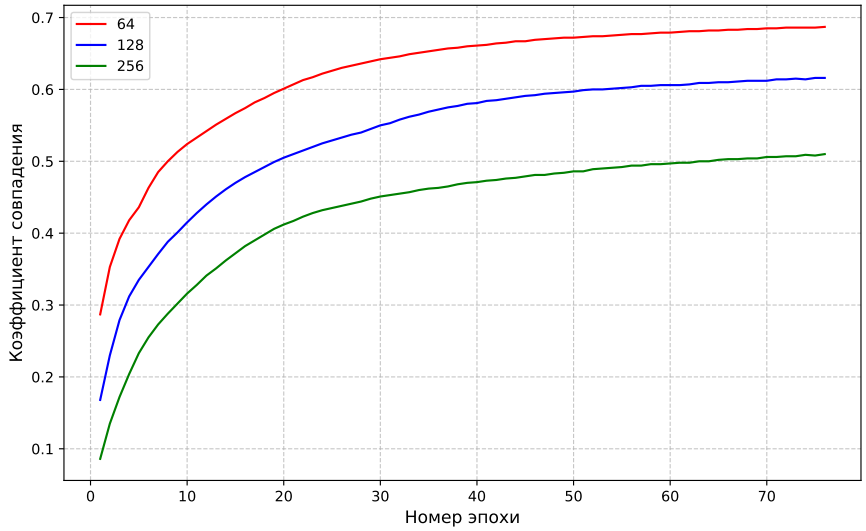

Рисунок 4.4 – Коэффициент совпадения в зависимости от эпохи обучения для различных размеров буфера на тренировочной выборке

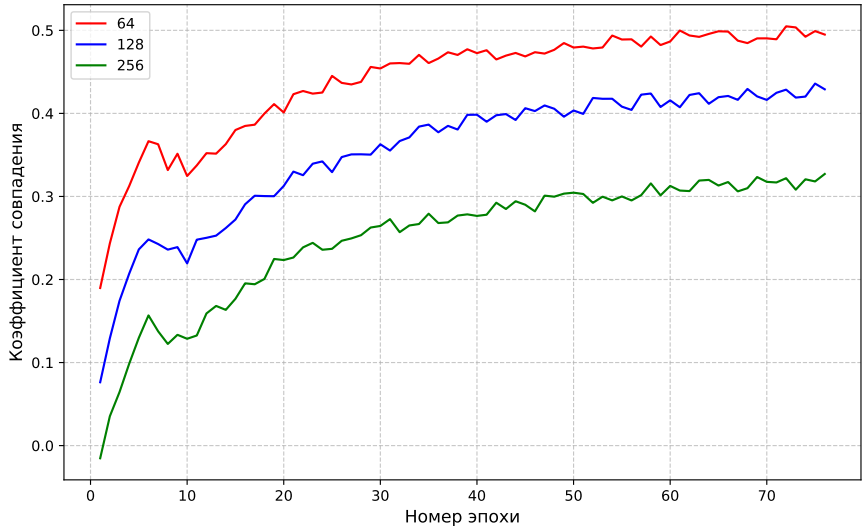

Рисунок 4.5 – Коэффициент совпадения в зависимости от эпохи обучения для различных размеров буфера на тестовой выборке

Из полученных графиков видно, что чем больше размер буфера, тем ниже точность модели как на тренировочной, так и на тестовой выборках. Это связанно с тем, что увеличение размера буфера приводит к увеличению числа обучаемых параметров модели и числа возможных вариантов ответов. Таким образом, метод может оказаться неэффективным при большом размере буфера. По умолчанию разделяемых кэш буфер Postgres содержит 128 страниц.

Оценка коэффициентов попадания для разработанного метода и аналогов на таком размере буфера показала, что разработанный метод в среднем на два процента лучше аналогов по этому показателю.

4.4 Вывод

Проведенные исследования позволили определить настраиваемые параметры модели, такие как размеры скрытых слоев d_z, d_b, d_h, $d_v = 448$, $d_f = 32$, обеспечивающие баланс между точностью и вычислительной сложностью. Разработанный метод продемонстрировал улучшение коэффициента попадания на 0.02 по сравнению с алгоритмом clock, используемым в PostgreSQL, что подтверждает его практическую эффективность. Однако отставание на 0.08 от оптимального алгоритма указывает на потенциал для дальнейшей оптимизации.

Анализ влияния размера буфера показал, что увеличение его объема приводит к снижению точности модели из-за роста числа обучаемых параметров и вариантов вытеснения. Результаты, полученные для размера буфера по умолчанию, показали улучшение по сравнению с существующими аналогами по введенным метрикам качества. Это подтверждает целесообразность внедрения метода в реальные системы с аналогичными настройками.

Таким образом, предложенный метод является перспективным решением для управления замещением страниц в разделяемом кэш буфере PostgreSQL, которое может повысить производительность системы управления базами данных за счет меньшего числа операций, взаимодействующих с диском.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы был спроектирован и разработан метод замещения страниц в разделяемом кэш буфере PostgreSQL с использованием нейронных сетей. В ходе выполнения работы были выполнены следующие задачи:

— проведено сравнение существующих методов замещения страниц;

— описан и спроектирован метод замещения страниц с использованием нейронных сетей;

— разработано программное обеспечение для предложенного метода;

— проведено сравнение разработанного метода с существующими аналогами по коэффициентам совпадения и попадания.

Цель работы достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Peiquan Y.* Learned buffer replacement for database systems // Proceedings of the 2022 5th International Conference on Data Storage and Data Engineering. — 2022. — С. 18—25.
2. *Noor H.* Virtual Memory Management // Securing the Digital Realm: Advances in Hardware and Software Security, Communication, and Forensics. — 2025. — С. 126.
3. *Shaik B.* PostgreSQL Configuration: Best Practices for Performance and Security. — Apress, 2020.
4. *Rogov E.* PostgreSQL 14 Internals // Postgres Professional. — 2023. — С. 471.
5. *Бабушкина Н. Е.* ВЫБОР ФУНКЦИИ АКТИВАЦИИ НЕЙРОННОЙ СЕТИ В ЗАВИСИМОСТИ ОТ УСЛОВИЙ ЗАДАЧИ // Донской государственный технический университет. — 2022. — С. 4.
6. *Антонов Г. В.* ПРОСТАЯ НЕЙРОННАЯ СЕТЬ И ЕЕ ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ // ФГБОУ ВО Великолукская государственная сельскохозяйственная академия. — 2021. — С. 11.
7. *Левченко К. М.* Нейронные сети // Белорусский государственный университет информатики и радиоэлектроники. — 2022. — С. 5.
8. *Барвинский Д. А.* Применение метода градиентного спуска в решении задач оптимизации // Тенденции развития науки и образования. — 2021. — С. 61.
9. *Zhang Y.* Why transformers need adam: A hessian perspective // Advances in Neural Information Processing Systems. — 2024. — Т. 37. — С. 37.
10. *Апарнев А. Н.* Анализ функций потерь при обучении сверточных нейронных сетей с оптимизатором Adam для классификации изображений // ВЕСТНИК МОСКОВСКОГО ЭНЕРГЕТИЧЕСКОГО ИНСТИТУТА. ВЕСТНИК МЭИ. — 2020. — С. 90.
11. *Koumi S.* A multilayer perceptron neural network approach for optimizing solar irradiance forecasting in Central Africa with meteorological insights // Scientific Reports. — 2024. — Т. 14, № 1. — С. 24.

12. *Minsky M.* Perceptrons: An Introduction to Computational Geometry. — MIT Press, 1969.

13. *Hecht-Nielsen R.* Kolmogorov's Mapping Neural Network Existence Theorem. — 1987.

14. *Колмогоров А. Н.* О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного и сложения // Доклады академии наук СССР. — 1957. — Т. 114, № 5. — С. 953—956.

15. *Baum E.* What Size Net Gives Valid Generalization? // Neural Computation. — 1989. — Т. 1, № 1. — С. 151—160.

16. *Jiang Q.* An efficient multilayer RBF neural network and its application to regression problems // Neural computing and Applications. — 2022. — С. 1—18.

17. *Дель И. В.* Прогноз приземной температуры воздуха на основе модели рекуррентной нейронной сети // Сборник статей Всероссийской молодежной научной конференции студентов. — 2021.

18. *Ромасенко А. А.* ЗАПОМИНАНИЕ И ОТОБРАЖЕНИИ ОБРАЗОВ НА ОСНОВЕ НЕЙРОННОЙ СЕТИ ХОПФИЛДА // Всероссийская научно-методическая конференция, посвященная 70-летию Оренбургского государственного университета. — 2022. — С. 1384—1392.

19. *Бахтин А.* Интеллектуальные системы управления технологическими процессами. — 2024.

20. *Zhang J.* A review of recurrent neural networks: LSTM cells and network architectures // Neural computation. — 2019. — Т. 31, № 7. — С. 1235—1270.

21. *Al-Selwi M. S.* RNN-LSTM: From applications to modeling techniques and beyond—Systematic review // Journal of King Saud University-Computer and Information Sciences. — 2024. — С. 34.

22. *Воронецкий Ю. О.* Методы борьбы с переобучением искусственных нейронных сетей // Научный аспект. — 2019. — Т. 13, № 2. — С. 1639—1647.

23. *Парасич А. В.* Формирование обучающей выборки в задачах машинного обучения. Обзор // Информационно-управляющие системы. — 2021. — С. 61—70.

24. *Cai Z.* Cascade R-CNN: High Quality Object Detection and Instance Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2021. — Т. 43, № 5. — С. 1483—1498.

25. *Пырнова О.* МЕТОДЫ И ПРОБЛЕМЫ ПЕРЕОБУЧЕНИЯ МНОГОСЛОЙНОЙ НЕЙРОННОЙ СЕТИ // ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В СТРОИТЕЛЬНЫХ, СОЦИАЛЬНЫХ И ЭКОНОМИЧЕСКИХ СИСТЕМАХ. — 2020. — С. 101—103.

26. *Burkholz R.* Batch normalization is sufficient for universal function approximation in CNNs. — 2024.

27. *Lecun Y.* Efficient Backprop // Neural Networks: Tricks of the Trade. — 2019. — С. 99—48.

28. *Николенко С.* Глубокое обучение. Погружение в мир нейронных сетей // Издательство Питер. — 2018. — С. 477.

29. *Кашницкий Ю. С.* Ансамблевый метод машинного обучения, основанный на рекомендации классификаторов // Интеллектуальные системы. Теория и приложения. — 2015. — Т. 19, № 4. — С. 37—55.

30. *Ntayagabiri J.* OMIC: A Bagging-Based Ensemble Learning Framework for Large-Scale IoT Intrusion Detection // Journal of Future Artificial Intelligence and Technologies. — 2025. — Т. 1, № 4. — С. 401—416.

31. *Daza A.* Stacking ensemble approach to diagnosing the disease of diabetes // Informatics in Medicine Unlocked. — 2024. — Т. 44. — С. 22.

32. *Vinyals O.* Pointer networks // Advances in neural information processing systems. — 2015. — Т. 28.

33. *Leutenegger S.* A modeling study of the TPC-C benchmark, vol. 22. — 1993.

34. Язык программирования Python [Электронный ресурс]. — Режим доступа: `https://www.python.org/`. — (Дата обращения: 28.04.2025).

35. Библиотека PyTorch [Электронный ресурс]. — Режим доступа: `https://pytorch.org/`. — (Дата обращения: 28.04.2025).

36. Библиотека PyQt [Электронный ресурс]. — Режим доступа: https://doc.qt.io/. — (Дата обращения: 28.04.2025).

ПРИЛОЖЕНИЕ А
Разработанный метод

Листинг А.1 – Модификация ReadBufferExtended

```
Buffer
ReadBufferExtended(Relation reln, ForkNumber forkNum,
   BlockNumber blockNum,
                   ReadBufferMode mode, BufferAccessStrategy
                      strategy)
{
    bool        hit;
    Buffer      buf;

    if (RELATION_IS_OTHER_TEMP(reln))
        ereport(ERROR,
                (errcode(ERRCODE_FEATURE_NOT_SUPPORTED),
                 errmsg("cannot access temporary tables of other
                    sessions")));

    pgstat_count_buffer_read(reln);
    buf = ReadBuffer_common(RelationGetSmgr(reln),
      reln->rd_rel->relpersistence,
                            forkNum, blockNum, mode, strategy,
                               &hit);

    const char *forkNumStr = "";
    if (forkNum == MAIN_FORKNUM)
        forkNumStr = "MAIN_FORKNUM";
    else if (forkNum == FSM_FORKNUM)
        forkNumStr = "FSM_FORKNUM";
    else if (forkNum == VISIBILITYMAP_FORKNUM)
        forkNumStr = "VISIBILITYMAP_FORKNUM";
    else if (forkNum == INIT_FORKNUM)
        forkNumStr = "INIT_FORKNUM";

    const char *readBufModeStr = "";
    if (mode == RBM_NORMAL)
        readBufModeStr = "RBM_NORMAL";
    else if (mode == RBM_ZERO_AND_LOCK)
        readBufModeStr = "RBM_ZERO_AND_LOCK";
    else if (mode == RBM_ZERO_AND_CLEANUP_LOCK)
```

```
        readBufModeStr = "RBM_ZERO_AND_CLEANUP_LOCK";
    else if (mode == RBM_ZERO_ON_ERROR)
        readBufModeStr = "RBM_ZERO_ON_ERROR";
    else if (mode == RBM_NORMAL_NO_LOG)
        readBufModeStr = "RBM_NORMAL_NO_LOG";

    elog(WARNING,
       "\n=======================================\nbuffer=%d
       rel_id=%u is_local_temp=%s fork_num=%s
       block_num=%u mode=%s strategy=%s relam=%u
       relfilenode=%u relhasindex=%s relpersistence=%c
       relkind=%c relnatts=%d relfrozenxid=%u
       relminmxid=%u
       hit=%s\n=======================================\n",
          buf,
          reln->rd_rel->oid,
          SmgrIsTemp(RelationGetSmgr(reln)) ? "true" : "false",
          forkNumStr,
          blockNum,
          readBufModeStr,
          // strategyStr,
          reln->rd_rel->relam,
          reln->rd_rel->relfilenode,
          reln->rd_rel->relhasindex ? "true" : "false",
          reln->rd_rel->relpersistence,
          reln->rd_rel->relkind,
          reln->rd_rel->relnatts,
          reln->rd_rel->relfrozenxid,
          reln->rd_rel->relminmxid,
          hit ? "true" : "false");

    if (hit)
        pgstat_count_buffer_hit(reln);
    return buf;
}
```

Листинг А.2 – Класс кодировщика запросов обращения к страницам

```
class PageAccEncoder(nn.Module):
    def __init__(self, hidden_size, embedding_size, buf_size):
        super(PageAccEncoder, self).__init__()

        self._hash_size = 5000
```

```
        page_params = len(fields(PageBatch)) * embedding_size

        assert(len(fields(PageBatch)) == 6)
        self._emb_layer = nn.ModuleDict({
            "rel_id": HashEmbedding(self._hash_size,
               embedding_size),
            "fork_num": HashEmbedding(self._hash_size,
               embedding_size),
            "block_num": HashEmbedding(self._hash_size,
               embedding_size),
            "relfilenode": HashEmbedding(self._hash_size,
               embedding_size),
            "rel_kind": nn.Embedding(10, embedding_size),
            "position": nn.Embedding(buf_size + 1,
               embedding_size)
        })

        self._page_enc = nn.Sequential(
            nn.Linear(page_params, hidden_size * 2),
            nn.ReLU(),
        )

    def forward(self, page_batch: PageBatch):
        embeddings = []
        for field in fields(PageBatch):
            val = getattr(page_batch, field.name)
            emb = self._emb_layer[field.name](val)
            embeddings.append(emb)

        page_acc_input = torch.cat(embeddings, dim=1)
        page_acc_enc = self._page_enc(page_acc_input)

        return page_acc_enc
```

Листинг А.3 – Класс кодировщика страниц в буфере

```
class PageBufferEncoder(PageAccEncoder):
    def __init__(self, hidden_size, embedding_size, buf_size):
        super().__init__(hidden_size, embedding_size, buf_size)

    def forward(self, buffer_batch: list[PageBatch]):
        embeddings = []
        for field in fields(PageBatch):
```

```
            val_stack = torch.stack([getattr(page_in_buf,
               field.name) for page_in_buf in buffer_batch])
            emb = self._emb_layer[field.name](val_stack)
            embeddings.append(emb)

        # (buffer_size, batch_size, num_features)
        buf_input = torch.cat(embeddings, dim=2)
        buf_res = self._page_enc(buf_input)

        return buf_res.permute(1, 0, 2)
```

Листинг А.4 – Модуль выбора страниц для замещения

```
class PageEviction(nn.Module):
    def __init__(self, input_page_size, input_context_size,
       hidden_size):
        super(PageEviction, self).__init__()

        self._hidden_size = hidden_size

        self._attention_page = nn.Linear(input_page_size,
           hidden_size, bias=False)
        self._attention_context = nn.Linear(input_context_size,
           hidden_size, bias=False)
        self._attention_v = nn.Linear(hidden_size, 1, bias=False)

    def forward(self, buffers, context):
        """
        buffers (batch_size, buffer_size, num_features)
        context (batch_size, num_features)
        """

        # (batch_size, 1, hidden_size)
        context_attention =
           self._attention_context(context).unsqueeze(1)
        # (batch_size, buf_size, hidden_size)
        context_attention_expanded =
           context_attention.expand(-1, buffers.size(1), -1)

        buffers_attention = self._attention_page(buffers)

        combined_attention =
           torch.tanh(context_attention_expanded +
```

```
            buffers_attention)
        scores =
            self._attention_v(combined_attention).squeeze(-1)

        return scores
```

Листинг А.5 – Модель, реализующая алгоритм замещения страниц

```
class PageAccModel(nn.Module):
    def __init__(self, hidden_size, lstm_hidden_size, buf_size,
        embedding_size = 32):
        super(PageAccModel, self).__init__()

        self._page_acc_encoder = PageAccEncoder(hidden_size,
            embedding_size, buf_size)
        self._lstm = nn.LSTM(input_size=hidden_size,
            hidden_size=lstm_hidden_size, batch_first=True)
        self._buf_page_encoder = PageBufferEncoder(hidden_size,
            embedding_size, buf_size)
        self._page_evict = PageEviction(hidden_size,
            lstm_hidden_size, hidden_size)


    def forward(self, page_batch: PageBatch, buffer_batch:
        list[PageBatch], h0=None, c0=None):
        page_acc_res = self._page_acc_encoder(page_batch)

        lstm_input = page_acc_res.view(page_acc_res.shape[0], 1,
            page_acc_res.shape[1])
        if h0 is None or c0 is None:
            lstm_out, (hn, cn) = self._lstm(lstm_input)
        else:
            lstm_out, (hn, cn) = self._lstm(lstm_input, (h0, c0))

        lstm_out_flat = lstm_out.view(lstm_out.shape[0],
            lstm_out.shape[2])

        buf_res = self._buf_page_encoder(buffer_batch)

        res = self._page_evict(buf_res, lstm_out_flat)

        return res, hn, cn
```

ПРИЛОЖЕНИЕ Б
Презентация