

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Объект авторского права

УДК 004.032.26

МАЦКЕВИЧ

Вадим Владимирович

**МЕТОДЫ ПОСТРОЕНИЯ И ОБУЧЕНИЯ ГЛУБОКИХ
ДОВЕРИТЕЛЬНЫХ НЕЙРОННЫХ СЕТЕЙ, ОСНОВАННЫХ НА
АНСАМБЛЯХ ОГРАНИЧЕННЫХ МАШИН БОЛЬЦМАНА**

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

по специальности 05.13.17 – Теоретические основы информатики

Минск 2025

Научная работа выполнена в Белорусском государственном университете

Научный руководитель **Краснопрошин Виктор Владимирович**,
доктор технических наук, профессор, профессор
кафедры информационных систем управления
Белорусского государственного университета

Официальные оппоненты: **Головко Владимир Адамович**,
доктор технических наук, профессор,
заведующий кафедрой интеллектуальных
информационных технологий УО «Брестский
государственный технический университет»

Белоцерковский Алексей Маратович,
кандидат технических наук, заведующий
отделом интеллектуальных информационных
систем ГНУ «Объединенный институт проблем
информатики НАН Беларуси»

Оппонирующая **УО «Полоцкий государственный**
организация **университет имени Евфросинии Полоцкой»**

Защита состоится 7 мая 2025 года в 14:30 часов на заседании совета по защите диссертаций Д 02.01.02 при Белорусском государственном университете по адресу: 220030 г. Минск, ул. Ленинградская 8, (корпус юридического факультета), ауд. 407. Телефон ученого секретаря 209-55-58, e-mail: Kochyn@bsu.by.

С диссертацией можно ознакомиться в Фундаментальной библиотеке Белорусского государственного университета.

Автореферат разослан «18» марта 2025 года.

Ученый секретарь совета
по защите диссертаций
кандидат технических наук, доцент

 **В.П. Кочин**

ВВЕДЕНИЕ

Интенсивное развитие цифровой экономики приводит к накоплению огромных массивов цифровых данных, что, в свою очередь, стимулирует развитие информационно-компьютерных технологий. Для извлечения из данных полезной для принятия решений информации стали активно применяться нейросетевые технологии. Объемы цифровых данных постоянно растут, а спектр прикладных задач стремительно расширяется. Поэтому исследования, связанные с разработкой моделей и алгоритмов, направленных на развитие нейросетевых технологий, являются актуальными.

Проблемы, связанные с теорией и практикой применения нейронных сетей, активно исследуются в работах зарубежных (Hinton G., LeCun Y., Haykin S., Zheng Sh., Decelle A.) и отечественных (Абламейко С.В., Головкин В.А., Старовойтов В.В., Ковалёв В. А., Лукашевич М. М. и др.) ученых. На международных научных конференциях (International Conference on Artificial Neural Networks (ICANN), International Conference on Artificial Intelligence and Big Data (AIBD), International Conference on Pattern Recognition and Information Processing (PRIP) и др.) широко обсуждаются проблемы развития архитектур нейронных сетей и алгоритмов обучения. Многие ученые отмечают, что большая размерность входных данных, при решении узкоспециализированных задач, порождает избыточность архитектур универсальных нейронных сетей, а разработка специальной сети является достаточно затратной. С другой стороны, с ростом размерности данных усложняется их внутренняя структура, что требует разработки более эффективных алгоритмов обучения.

В диссертационной работе решаются некоторые из указанных выше проблем. Предложена модель нейронной сети в виде ансамбля из ограниченных машин Больцмана, что позволяет понизить параметрическую сложность модели с сохранением качества решения. Разработан оригинальный алгоритм обучения на основе метода отжига, учитывающий свойства пространства решений, что позволяет за счет снижения числа итераций ускорить процесс обучения. Предложена методика обучения алгоритмов случайного поиска, обеспечивающая их инвариантность к объему обучающей выборки. На основе теоретических результатов разработана программная технология (в виде фреймворка) для конструирования и обучения нейронных сетей. За счет использования языка программирования C++ и библиотек OpenMP, OpenCL обеспечена высокая производительность и кросс-платформенность. Все это способствует повышению эффективности и качества решения прикладных задач в целом.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с научными программами (проектами), темами

Тема диссертации соответствует приоритетному направлению научно-технической деятельности согласно пункту 1 перечня приоритетных направлений научной, научно-технической и инновационной деятельности на 2021-2025 годы (Указ Президента Республики Беларусь от 07 мая 2020 г. № 156).

Исследования по теме диссертационной работы проводились в рамках научных программ:

1. НИР БРФФИ «Разработка моделей и технологий оценки состояния компонентов крупномасштабных социально-экономических и организационно-технических систем на основе методов искусственного интеллекта» (№ Ф20Р-177 от 04.05.2020 г., № гос. рег. 20201303).

2. НИР БРФФИ «Многоуровневые модели распознавания в задачах динамического поиска оптимальных решений» (№ Ф21АРМ-005 от 01.11.2021 г., № гос. рег. 20213919).

3. НИР БГУ «Разработать теоретические основы и технологию построения интеллектуальных компьютерных систем для решения практических задач» (от 28.07.2021 г., № гос. рег. 20212834).

4. НИР международный контракт «Договор о развитии технологий разработки алгоритмов обработки изображений дистанционного зондирования Земли» (№: 22СЕТС19-ICN1785 от 22.11.2022 г.).

5. НИР грант Министерства образования «Эффективные методы обучения нейронных сетей на основе случайного поиска» (от 19.02.2024 г., № гос. рег. 20241180).

Цель, задачи, объект и предмет исследования

Целью диссертационной работы является разработка модели глубокой доверительной сети на основе ансамбля из ограниченных машин Больцмана и алгоритма обучения на основе случайного поиска, направленных на эффективную обработку и анализ изображений.

Для достижения цели были решены следующие задачи:

1) Провести анализ современного состояния нейросетевых технологий и существующих подходов к обучению нейронных сетей на основе алгоритмов случайного поиска;

2) Разработать модель глубокой доверительной сети на основе ансамбля из ограниченных машин Больцмана, позволяющую снизить количество настраиваемых параметров нейронной сети и время ее обучения;

3) Разработать алгоритм обучения на основе случайного поиска, учитывающий (для снижения времени обучения) неоднородность пространства решений;

4) Разработать методику обучения нейронной сети алгоритмами случайного поиска, обеспечивающую их инвариантность к объему обучающей выборки;

5) Разработать программную технологию для конструирования и обучения глубоких нейронных сетей с применением модели из ансамбля ограниченных машин Больцмана и алгоритмов обучения на основе случайного поиска.

Объектом исследования являются нейросетевые технологии обработки данных. Предметом исследования выступают архитектуры нейронных сетей и алгоритмы обучения.

Научная новизна

Разработана модель глубокой доверительной нейронной сети, основанная на ансамбле ограниченных машин Больцмана. В сжимающих слоях сети вместо одной используются n машин небольшой размерности, что позволяет снизить количество настраиваемых параметров сети и время ее обучения с сохранением качества решения.

Разработан алгоритм обучения, основанный на модификации оптимизационной схемы Больцмановского отжига. В отличие от классического варианта в рамках итерационного процесса учитывается неоднородность пространства решений, что позволяет снизить время обучения нейронных сетей.

Предложена методика обучения для алгоритмов случайного поиска. В отличие от традиционного подхода на каждой итерации алгоритма используются фиксированный фрагмент обучающей выборки, что обеспечивает инвариантность алгоритмов к объему выборки с сохранением качества обучения.

Спроектирован и реализован кросс-платформенный программный комплекс (фреймворк) для конструирования и обучения глубоких нейронных сетей на гетерогенных системах с поддержкой модели в виде ансамблей из ограниченных машин Больцмана и алгоритмов случайного поиска.

Положения, выносимые на защиту

1. Модель глубокой доверительной нейронной сети, основанная на ансамбле ограниченных машин Больцмана (из n машин небольшой размерности), что позволяет снизить количество параметров и время обучения

с сохранением качества решения.

2. Алгоритм обучения, основанный на модификации оптимизационной схемы Больцмановского отжига и учитывающий неоднородность пространства решений, что позволяет снизить время обучения нейронных сетей.

3. Методика обучения алгоритмами случайного поиска, обеспечивающая их инвариантность к объему обучающей выборки.

4. Программная технология (в виде кросс-платформенного фреймворка) для конструирования и обучения глубоких нейронных сетей на гетерогенных системах с поддержкой модели в виде ансамбля из ограниченных машин Больцмана и алгоритмов обучения на основе случайного поиска.

Личный вклад соискателя ученой степени в результаты диссертации

Основные положения диссертации получены соискателем лично. Соавтором основных публикаций является научный руководитель д.т.н., профессор В.В. Краснопрошин, который принимал участие в определении целей, задач и методов исследований, в планировании работ и обсуждении результатов. Соавторами отдельных публикаций являются А.М. Недзьведь и В.А. Стасюк, которые принимали участие в организации и проведении некоторых экспериментов и обсуждении полученных результатов.

В диссертационную работу не включены результаты, которые были получены другими соавторами. Материалы совместных публикаций использованы соискателем в объеме авторского вклада.

Апробация диссертации и информация об использовании ее результатов

Основные положения и полученные результаты диссертационной работы докладывались и обсуждались на следующих конференциях: "12-th International Workshop on Mathematical Models and their Applications" (Красноярск, 2023); "International Conference Open Semantic Technology for Intelligent Systems" (Минск, 2019, 2020, 2023); "International Conference Computer Data Analysis and Modeling (CDAM)" (Минск, 2019, 2022); "International Conference on Pattern Recognition and Information Processing (PRIP)" (Минск, 2019, 2023); "Международный конгресс по информатике: информационные системы и технологии (CSIST'2022)" (Минск, 2022); "II Международная научно-техническая конференция "Актуальные проблемы науки и техники" (Сарапул, 2022); "II Международная научно-практическая конференция "Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2022)" (Полоцк, 2022); "14-th International

Conference "Intellectualization of information processing" (Москва, 2022); "Международный научно-технический семинар: технологии передачи и обработки информации" (Минск, 2022-23); "X Международная научно-практическая конференция "Информационные технологии в промышленности, логистике и социальной сфере" (Минск, 2019).

По результатам диссертации получено 3 акта о внедрении и 2 Свидетельства о регистрации программных продуктов в Национальном центре интеллектуальной собственности.

Опубликованность результатов диссертации

Основные результаты диссертационного исследования опубликованы в 27 научных работах, среди которых: 8 статей в научных изданиях в соответствии с пунктом 19 Положения о присуждении ученых степеней и присвоении ученых званий (общим объемом 5,16 авторского листа), 2 статьи в других научных изданиях, 15 статей в сборниках материалов научных конференций и 2 тезиса.

Структура и объем диссертации

Диссертация состоит из перечня сокращений и обозначений, введения, общей характеристики работы, четырех глав, заключения, списка использованных источников и одного приложения.

Полный объем диссертации составляет 105 страниц, из которых 10 рисунков на 5 страницах, 6 таблиц на 2 страницах, 1 приложения на 7 страницах. Список использованных источников состоит из 118 наименований, включая 27 публикаций автора (на 4 страницах).

ОСНОВНАЯ ЧАСТЬ

В **первой главе** рассмотрены проблемы нейросетевой обработки данных. Проведен обзор основных архитектур нейронных сетей, алгоритмов обучения и программных средств. Было, в частности, отмечено, что большая размерность входных данных, при решении узкоспециализированных задач, порождает избыточность архитектур универсальных нейронных сетей, а разработка специальной сети является достаточно затратной. Кроме того, с расширением спектра прикладных задач изменяются требования к процессу обучения. Все это приводит к необходимости разработки соответствующих программно-алгоритмических средств. Поэтому проблема развития нейросетевой технологии, и ее отдельных компонент, по-прежнему является актуальной.

На основе проведенного анализа, в конце главы, была сформулирована цель диссертационного исследования. Были определены ее теоретическая,

технологическая и практическая составляющие, в рамках которых поставлены задачи на разработку соответствующих моделей нейронных сетей, алгоритмов обучения и программных средств.

Во **второй главе** представлены результаты теоретических исследований, связанных с проблемой обработки изображений высокого разрешения. Рассмотрены наиболее распространенные виды моделей нейронных сетей, которые используются для решения данной проблемы.

Показано, что глубокие доверительные сети являются универсальным инструментом для обработки данных. Благодаря сжимающим слоям, в качестве которых используются ограниченные машины Больцмана (ОМБ) и автоэнкодеры, потери при сжатии не велики, что обеспечивает высокое качество получаемого решения. Однако указанные сети имеют существенный недостаток. ОМБ и автоэнкодеры являются полносвязными двудольными графами с соответствующим набором настраиваемых параметров. Это делает их весьма затратными с точки зрения вычислительных ресурсов. Поэтому для обработки данных большой размерности в настоящее время используют, как правило, глубокие сверточные нейронные сети.

Одним из достоинств таких сетей является сравнительно небольшое количество параметров, что позволяет производить их обучение гораздо быстрее доверительных.

Однако и эти сети имеют определенные недостатки, одним из которых, в частности, является возможная потеря части данных. Выбор между уровнем потерь и размерностью выходных данных породил большое разнообразие архитектур в целом.

Отмечается, что построение узкоспециализированных сетей является достаточно сложной задачей, поэтому на практике, как правило, применяются универсальные (предобученные) сверточные сети. Это, в свою очередь, приводит к построению больших и сложных архитектур и, следовательно, к некоторому понижению точности и производительности прикладных систем, построенных на их основе.

Таким образом, сверточные сети позволяют обрабатывать данные большой размерности за приемлемое время, в то время как доверительные сети обеспечивают более высокое качество.

Проведен анализ вычислительной сложности указанных архитектур (по количеству настраиваемых параметров и объему вычислений) и построены соответствующие им оценки. Анализ полученных оценок подтвердил эффективность использования сверточных сетей. Вместе с тем указаны ситуации, когда на практике могут возникнуть проблемы для их применения.

Предложен подход к построению архитектуры сети, основанный

на комбинации положительных свойств, характерных для доверительных и сверточных нейронных сетей. В сжимающих слоях глубокой доверительной сети вместо одной ограниченной машины Больцмана (ОМБ) предлагается использовать ансамбль ОМБ небольшой размерности. Показано, что эффективность при этом достигается за счет снижения количества настраиваемых параметров и использования естественного параллелизма.

Суть предлагаемого подхода можно описать следующим образом. Исходное растровое изображение разбивается на прямоугольные блоки-фрагменты одинакового размера. При этом разрешение исходных изображений должно быть кратно по ширине и высоте размерам блоков. Архитектуру входного слоя сети предлагается проектировать в виде ансамбля из n ОМБ небольшой размерности. В данном случае фрагментам изображения соответствуют отдельные машины, размеры и наборы параметров которых совпадают (рисунок 1).

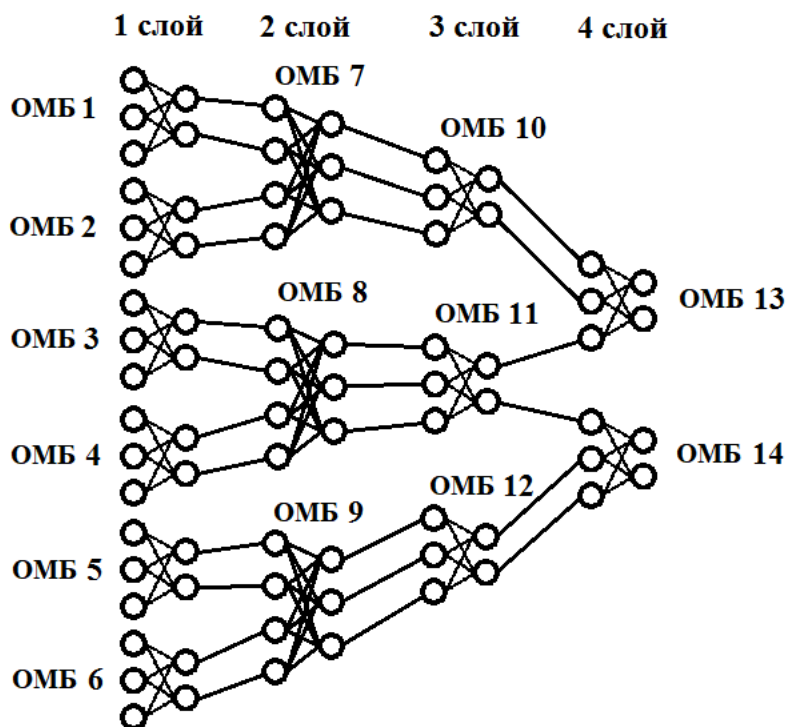


Рисунок 1 – Предлагаемая архитектура нейронной сети

Был проведен сравнительный анализ вычислительной сложности предложенной ансамблевой архитектуры по количеству настраиваемых параметров и объему требуемых вычислений с архитектурами глубоких доверительных и сверточных сетей (таблица 1).

Примечание. В силу большого разнообразия используемых на практике архитектур при вычислении оценок рассматривались в некотором смысле «усредненные» их варианты в рамках каждого типа. Поэтому приведенные в таблице результаты являются приближенными. В большей степени это

касается моделей сверточных сетей. Кроме того, для простоты оценивания и без потери общности в качестве исходных данных рассматривались квадратные цветные растровые изображения из миллиона пикселей. Также при вычислении оценок не учитывались сложностные характеристики одинаковых для всех сетей и расположенных в конечных слоях классификаторов.

Таблица 1 – Сравнительные оценки сложности нейронных сетей

Тип нейронной сети	Количество настраиваемых параметров	Объем вычислений на одно изображение
Глубокая доверительная сеть	$1,5 \cdot 10^{13}$	$1,5 \cdot 10^{13}$
Глубокая сверточная сеть	$5,6 \cdot 10^4$	$1,9 \cdot 10^{10}$
Архитектура сети из ансамбля ОМБ	$9 \cdot 10^3$	$2,4 \cdot 10^8$

Следует отметить, что предлагаемая архитектура сети допускает построение ансамблей машин для любого количества сжимающих слоев. Поэтому при ее проектировании необходимо определять степени сжатия во всех сжимающих слоях, а ансамбли слоев должны удовлетворять следующим двум условиям:

- 1) количество нейронов каждого слоя должно быть кратно числу ОМБ ансамблей текущего и предыдущего слоев;
- 2) необходимо соблюдать степень сжатия ОМБ, формирующих ансамбль машин слоя.

Проведено обоснование корректности предложенной архитектуры. Показано, что потери при сжатии не превышает уровень потерь классического варианта глубокой доверительной сети и семантика сжимаемых изображений при этом сохраняется. Предложен эффективный способ организации процесса обучения, учитывающий специфику ансамблевой архитектуры и ориентированный на параллельную обработку данных.

При анализе эффективности ансамблевой архитектуры было установлено, что качество сжатия и скорость функционирования сети зависит от способа разбиения изображений, то есть от размеров блока. Использование небольших размеров приводит к повышению уровня потерь, а больших – к увеличению времени обучения. Были проведены эксперименты, в результате которых установлено, что оптимальным при построении ансамбля является разбиение изображений на блоки размером 4×4 (таблица 2). Процесс сжатия в экспериментах осуществлялся с помощью ансамбля из ограниченных машин Больцмана типа Гаусс-Бернулли, а для оценки качества использовались известные функционалы.

Таблица 2 – Результаты сжатия изображений

Размер декомпозиции	Функционал MSE	Функционал PSNR	Функционал PSNR_HVS	Функционал SSIM	Время обучения (ч)
2x2	167	26,2	25,8	0,860	0,075
2x4	124	27,4	27,0	0,887	0,25
4x4	116	27,7	27,3	0,895	1,5
5x5	114	27,7	27,3	0,896	2,5

В ходе экспериментов установлено, что увеличение размерности блока выше 4x4 приводит лишь к небольшому росту качества сжатия, которое полностью нивелируется квадратичным ростом времени обучения.

Третья глава посвящена проблемам построения и исследования эффективности алгоритмов обучения, основанных на идеологии случайного поиска.

В первой части главы рассмотрены особенности процесса обучения нейронных сетей. Было показано, что пространство решений в оптимизационной задаче обучения обладает свойством своеобразной неоднородности. В частности, показано, что в диапазоне изменения любого параметра существует относительно небольшой отрезок значений, вероятность нахождения оптимального решения в котором, существенно выше, чем на оставшемся диапазоне. Это означает, что мощность множества, в котором находится оптимальное решение, намного меньше мощности всего пространства решений. Было также показано, что учет указанного свойства в алгоритме случайного поиска позволяет существенно сократить время обучения.

Отмечается, что в настоящее время одним из основных инструментов в задачах обучения нейронных сетей являются градиентные алгоритмы. Они имеют ряд неоспоримых положительных свойств, благодаря которым широко применяются на практике. Вместе с тем выявлены и некоторые особенности алгоритмов, которые могут ограничить их использование на практике. В условиях расширения спектра прикладных задач, это может создать определенные проблемы. Поэтому появляется необходимость в поиске альтернативных подходов к обучению. По мере развития вычислительных средств появилась возможность использования алгоритмов на основе случайного поиска. Однако алгоритмы данного типа не учитывали специфику свойства пространства поиска в нейронных сетях. И так как оно, как правило, достаточно велико, то сходимость таких алгоритмов в данном случае крайне низкая. Вместе с тем учет свойства неоднородности пространства решений может повысить скорость работы таких алгоритмов.

Предлагается АСП-алгоритм, основанный на модификации оптимизационной схемы Больцмановского отжига и учитывающий свойство неоднородности пространства решений.

АСП - алгоритм

Предварительный этап.

Задаются начальное решение $X^0(X_1^0, X_2^0, \dots, X_h^0)$, где $X_1^0 = (x_{11}^0, x_{12}^0, \dots, x_{1n_1}^0)$, $X_2^0 = (x_{21}^0, x_{22}^0, \dots, x_{2n_2}^0)$, ..., $X_h^0 = (x_{h1}^0, x_{h2}^0, \dots, x_{hn_h}^0)$ и последовательности температур $T_0, T_1, \dots, T_k$, связанных соотношением:

$$T_k = \frac{T_0}{\ln(k+2)}, k > 0,$$

где T_0 – начальное значение.

Общая k-я итерация.

Пусть X^{k-1} решение, полученное на $k-1$ итерации. Строится окрестность $N(X^{k-1})$.

Процедуру построения окрестности $N(X^{k-1})$ и вероятность перехода к новому решению формально можно записать в виде последовательности следующих шагов.

Шаг 1. Генерируется h случайных перестановок длиной, равной количеству параметров $n_1, n_2, \dots, n_h$ в векторах $X_1^{k-1}, X_2^{k-1}, \dots, X_h^{k-1}$, в соответствии с которыми производятся перестановки их координат.

Генерируется h равномерно распределенных на отрезке, от нуля до величины равной количеству в векторе параметров, дискретных случайных величин $a_1, a_2, \dots, a_h$. Данные величины определяют количество первых элементов каждого вектора, значения координат которого будут изменяться.

Примечание. Для генерации случайных перестановок применяется алгоритм тасования Фишера-Йетса. В то время как для генерации случайных величин может применяться любой датчик случайных чисел (например, программно реализованный в библиотеках современных языков программирования).

Шаг 2. Для каждой изменяемой j координаты i -ого вектора x^{k-1}_{ij} генерируется случайная величина b_{ij} , равномерно распределенная на отрезке $[-l_i/2; l_i/2]$. Данные величины являются параметрами алгоритма и задают мощность окрестности поиска $N(X^{k-1})$ с центром в текущем решении. Мощность $N(X^{k-1})$ определяется выражением

$$|N(X^{k-1})| = \prod_{k=1}^h l_k^{n_k}$$

В соответствии с проведенными выше манипуляциями строится решение $Y=(Y_1,\dots,Y_h)=(y_{11},y_{12},\dots,y_{1n_1},y_{21},y_{22},\dots,y_{2n_2},\dots,y_{h1},y_{h2},\dots,y_{hn_h})$, где значение изменяемых координат y_{ij} вычисляется по формуле $y_{ij} = x_{ij}^{k-1} + b_{ij}$

Далее вычисляются значение целевой функции для Y и вероятность перехода к новому решению.

Пусть X_{k-1} и Y – текущее и сгенерированное решения соответственно. Вероятность перехода к новому решению определяется по формуле:

$$P(X^k = Y | X^{k-1}) = \min \left\{ 1, \exp \left(\frac{F(X^{k-1}) - F(Y)}{T_k} \right) \right\}$$

Генерируется базовая случайная величина ξ , равномерно распределенная на отрезке $[0;1]$. Тогда

$$X^k = \begin{cases} Y, & \xi \leq P(X^k = Y | X^{k-1}) \\ X^{k-1}, & \xi > P(X^k = Y | X^{k-1}) \end{cases}$$

Принцип останова. Если время на обучение алгоритма истекло, то останов, в противном случае – переход на следующую $k+1$ итерацию.

Модификация оптимизационной схемы Больцмановского отжига непосредственно связана со способом выбора начального решения и процедурой построения окрестность текущего решения.

Для АСП-алгоритма была доказана его корректность и показано, что предложенная модификация сохраняет свойство сходимости классической схемы.

Наиболее трудоемким в АСП-алгоритме этапом является оценка решений, для вычисления на каждой итерации которой обрабатывается вся обучающая выборка. Поэтому с увеличением ее размера происходит линейный рост объема вычислений на каждой итерации алгоритма. Это важное его отличие от алгоритмов градиентного типа, у которых на итерациях используются лишь небольшой фрагмент обучающей выборки. Поэтому объем вычислений при обучении у них не зависит от размера выборки.

Предлагается методика, в которой при обучении АСП-алгоритма предлагается использовать описанную выше идеологию градиентных алгоритмов.

Формально данную методику можно описать следующим образом.

Этап 1. Процедура преобразования обучающей выборки.

Пусть задана обучающая выборка S^0 . Процесс ее преобразования можно

описать в виде последовательности следующих шагов.

Шаг 1. Обучающую выборку S^0 дублируют q раз, в результате получается выборка, состоящую из объединения q ее копий.

$$\begin{cases} S^1 = \bigcup_{i=1}^q S_i^1 \\ S_i^1 = S^0, i = \overline{1, q} \end{cases}$$

Шаг 2. Внутри каждой, из полученных копий, случайным способом перемешивают содержащиеся внутри ее объекты, получают выборку

$$S^2 = \bigcup_{i=1}^q S_i^2$$

Шаг 3. Последовательно выбирая полученные копии $S_i^2, i = \overline{1, q}$ разбивают их на m равных частей (предполагается, что выборка разбивается на части-фрагменты без остатка). В результате получают выборку, состоящую объединения qm фрагментов $S_r^3, r = \overline{1, qm}$

$$S^3 = \bigcup_{r=1}^{qm} S_r^3,$$

В результате финальный вариант преобразованной выборки представляет собой упорядоченный массив фрагментов фиксированного размера, которые в некотором порядке используются на втором этапе.

Этап 2. Механизм оценки решения.

Пусть на k -ой итерации алгоритма в окрестности текущего решения $N(X^{k-1})$ сгенерировано случайное решение Y .

Вычисляем индекс j -ого фрагмента $j = [k/w] \bmod (mq) + 1$, где k – номер итерации, а w – параметр, определяющий частоту смены фрагментов выборки. С использованием фрагмента S_j^3 для решения Y вычисляют значение $F(Y, S_j^3) = v_j$, на основе которого получают оценку решения по формуле:

$$F(Y) = \frac{v_j}{\sqrt[m]{m}}$$

Доказана корректность представленной методики: что ее применение не ухудшает качество решения и обеспечивает свойство инвариантности АСП-алгоритма относительно объема обучающей выборки.

Примечание. Методика содержит набор параметров q , t и w . Показано, что с помощью экспериментального выбора значений этих параметров можно управлять процессом обучения.

В заключение главы приведены результаты экспериментов по проверке эффективности алгоритмов, основанных на случайном поиске. На примере решения задачи сжатия проведено сравнение АСП алгоритма с популярными градиентными алгоритмами обучения ADAM, FTML (таблица 3). В качестве тестового материала использовались известные выборки CIFAR-10 и STL-10.

Таблица 3 – Результаты обучения

Выборка	Алгоритм обучения	Степень сжатия, бит/пиксель	MSE	PSNR	PSNR_HVS	SSIM	Время обучения, ч
CIFAR-10	ADAM	1,5	435	21,9	22,1	0,765	0,40
		0,75	918	18,6	18,8	0,600	0,52
	FTML	1,5	408	22,1	22,3	0,769	0,38
		0,75	821	19,1	19,2	0,623	0,44
	АСП	1,5	390	22,3	22,5	0,775	0,75
		0,75	645	20,1	20,4	0,686	0,88
STL-10	ADAM	1,5	433	21,9	22,1	0,663	5,85
		0,75	836	19,0	19,2	0,502	6,85
	FTML	1,5	397	22,1	22,3	0,661	5,65
		0,75	756	19,4	19,5	0,509	6,60
	АСП	1,5	384	22,4	22,6	0,671	11,3
		0,75	632	20,2	20,4	0,557	13,1

В четвертой главе описана программная технология для конструирования и обучения глубоких нейронных сетей с поддержкой моделей и алгоритмов, разработанных в диссертации.

Программная технология была реализована в виде кросс-платформенного фреймворка. Отмечено, что современные фреймворки, в большинстве случаев, спроектированы с упором на использование CUDA, а многие из них реализованы на языке Python. Однако, в этом случае, их применение ограничивается видеокартами производителя nvidia, а использование данного языка программирования – существенно снижает производительность. Поэтому предлагаемый фреймворк был разработан на языке C++ с использованием библиотек OpenMP и OpenCL. Библиотеки OpenMP обеспечивают эффективную организацию параллельных вычислений в рамках одного процессора. А библиотеки OpenCL – совместимость и параллельные вычисления на широком классе, отличных от процессора, вычислительных устройств.

В начале главы приводится описание фреймворка и особенности его реализации. Показано, что он состоит из шести основных блоков (рисунок 2). Библиотека нейронных сетей содержит модели: ограниченных машин Больцмана (Гаусс-Бернулли и Бернулли-Бернулли), автоэнкодера, многослойного персептрона, слоев свертки и выборки. На основе указанных моделей заложена возможность проектирования глубоких доверительных сетей для сжатия и предварительной обработки данных. А наличие слоев и многослойного персептронов – нейросетевых классификаторов, ориентированных на решение широкого класса прикладных задач.

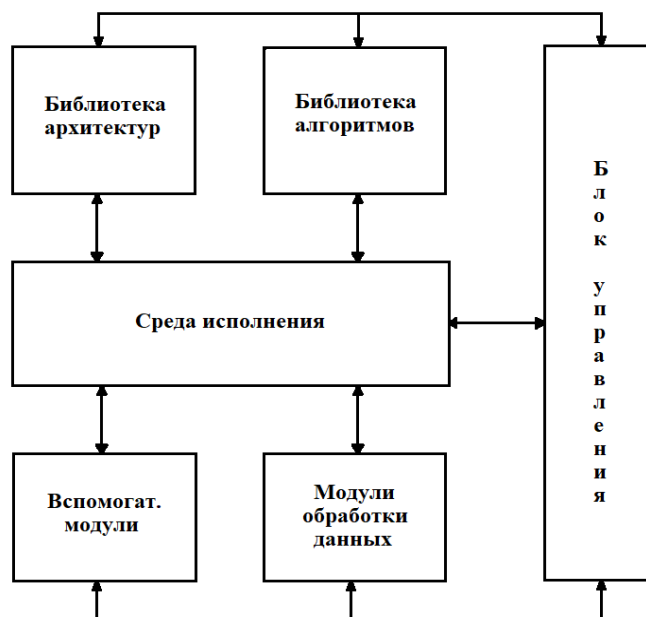


Рисунок 2 – Структура фреймворка

В библиотеке алгоритмов представлен широкий спектр оптимизаторов. Модуль обработки данных содержит необходимый для решения прикладных задач функционал, а вспомогательные модули загружаются по мере их обнаружения.

Фреймворк допускает 4 режима функционирования: подготовительный, режимы конструирования и обучения нейронной сети, а также режим решения прикладной задачи.

При реализации фреймворка использовалась стратегия параллельной обработки, направленная на эффективное его функционирование при решении прикладных задач. В частности, для ускорения работы нейронных сетей реализован блочный алгоритм умножения матриц, что позволяет улучшить кэширование данных и существенно ускоряет их обработку, как на процессоре, так и на видеокарте. Операция матричного умножения на видеокарте реализована в виде процедуры двухуровневого

распараллеливания, которая циклически распределяет процесс вычисления блоков результирующей матрицы. Для сокращения времени на обучение алгоритмов случайного поиска реализована процедура распараллеливания, основанная на сокращении вычислений.

Эффективность полученных в диссертации результатов тестировалось на примере решения хорошо известной прикладной задачи – обнаружение изменений по данным дистанционного зондирования Земли (спутниковые снимки с разрешением 4000x4000 пикселей). В качестве наблюдения рассматривались различные виды объектов: построение или снос капитальных зданий, перестройка или визуальное их изменение, установка или удаление теплиц и т.п.

Для решения задачи в среде фреймворка была спроектирована и обучена глубокая доверительная сеть. Сжимающий слой сети был построен в виде ансамбля из ограниченных машин Больцмана типа Гаусс-Бернулли. В качестве классификатора был выбран многослойный персептрон, состоящий из трех слоев: входного, скрытого и выходного. Входной слой содержал 2400 нейронов, скрытый – 32 нейрона с сигмоидной функцией активации, а выходной – 2 нейрона с функцией активации софтмакс.

Для вычислений использовался компьютер 2x8 RAM DDR3 1600 MHz, с процессором intel i7-4770k и видеокартой nvidia rtx 3070 и операционной системой Ubuntu 20.04.

Качество полученных решений проверялось с использованием двух функционалов, соответственно, для оценки точности и полноты обнаружения.

Сравнение полученных в экспериментах результатов с аналогичными параметрами из литературных источников приведено в таблице 3. В данных работах задача обнаружения изменений решалась с помощью глубоких сверточных нейронных сетей AERNet и D2ANet, содержащих большое количество настраиваемых параметров. В качестве данных в них использовались аналогичные космические снимки, которые были сняты в той же широте, имели те же объекты и параметры съемки.

Таблица 3 – Результаты обнаружения изменений

Архитектура сети	Точность, %	Полнота, %
Предлагаемая	76,9	79,5
AERNet	77,2	77,1
D2ANet	81,2	71,4

Полученные в ходе экспериментов результаты подтвердили в целом работоспособность предложенного в диссертации подхода. Следует отметить, что значения точности и полноты для аналогов из литературных источников

вычислялись попиксельно. В то время как для сконструированной сети – дискретно для блоков 20x20 пикселей, поэтому они обладают повышенной погрешностью. Однако незначительное отставание по точности полностью компенсируется превосходством в более важном, в соответствии требованием задачи, параметре – полноте обнаружения.

Пример работы спроектированной в рамках эксперимента сети показан на рисунке 3.

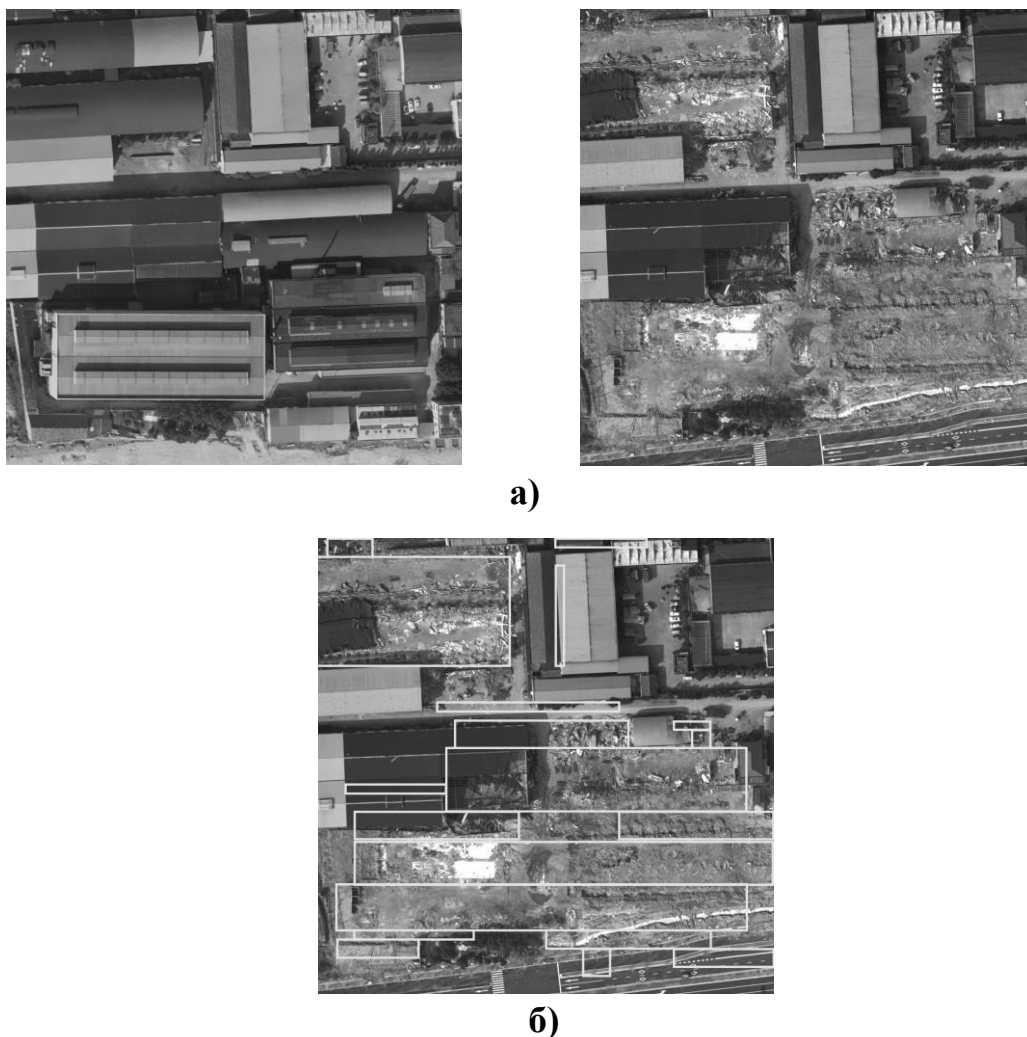


Рисунок 3 – Пример работы нейронной сети: а) пара фрагментов исходного изображения б) выявленные сетью изменения

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Построена модель глубокой доверительной нейронной сети, основанная на ансамбле ограниченных машин Больцмана. В сжимающих слоях сети вместо одной используются n машин небольшой размерности, что позволяет снизить количество настраиваемых параметров сети и время ее

обучения с сохранением качества решения [2-А; 3-А, с. 238-239; 10-А, с. 259; 13-А; 21-А, с. 179].

2. Разработан алгоритм обучения на основе модификации оптимизационной схемы Больцмановского отжига, учитывающей неоднородность пространства решений. Показано, что учет свойства неоднородности повышает скорость сходимости алгоритма. Доказана корректность алгоритма и его гарантированная сходимость (из любого начального) к оптимальному решению по вероятности [2-А; 3-А, с. 237; 4-А, с. 26; 5-А, с. 301; 6-А, с. 136; 7-А, с. 313; 10-А, с. 260; 13-А; 14-А; 15-А, с. 210-211; 16-А; 17-А; 18-А; 19-А; 20-А; 21-А, с. 180-181; 22-А, с. 140-141; 23-А; 24-А; 27-А].

3. Предложена методика для алгоритмов случайного поиска, в соответствии с которой на каждой итерации процесса обучения используется фиксированный фрагмент выборки. Показано, что это обеспечивает инвариантность алгоритмов к изменениям размеров обучающей выборки с сохранением качества решения [7-А, с. 313; 24-А; 25-А].

4. Спроектирован и реализован кросс-платформенный программный комплекс (фреймворк) для конструирования и обучения глубоких нейронных сетей на гетерогенных системах с поддержкой модели в виде ансамблей из ограниченных машин Больцмана и алгоритмов случайного поиска [1-А; 2-А; 3-А, с. 237-238; 6-А, с. 134; 8-А, с. 373; 9-А; 10-А, с. 258; 11-А; 12-А; 25-А; 26-А].

Рекомендации по практическому использованию результатов

Результаты диссертационных исследований были использованы при реализации прикладных систем обнаружения, в учебном процессе и при выполнении ряда НИР. В частности, предложенные модель глубокой доверительной сети и алгоритм обучения использовались для построения системы обнаружения изменений в рамках выполнения международного контракта: НИР «Договор о развитии технологий разработки алгоритмов обработки изображений дистанционного зондирования Земли» (№: 22СЕТС19-ICN1785 от 22.11.2022 г.). Полученные результаты также использовались в модуле интеллектуального анализа данных в рамках системы обнаружения изменений компании «Байлекс Малтимедиа».

Реализованные программные продукты были зарегистрированы в Государственном реестре компьютерных программ. Получены Свидетельства о регистрации в Национальном центре интеллектуальной собственности № 1744-КП для программы «Система обнаружения непогодных изменений» и № 1743-КП для программы «Библиотека алгоритмов глубокого обучения».

Кроме того, научные и практические результаты диссертационной работы используются в учебном процессе Белорусского государственного университета.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ УЧЕНОЙ СТЕПЕНИ

Статьи в научных рецензируемых изданиях,

включенных в перечень изданий, и в иностранных научных изданиях

1-А. Краснопрошин, В. В. Эффективная обработка данных на гетерогенных вычислительных устройствах / В. В. Краснопрошин, В. В. Мацкевич // Вестник Брестского государственного технического университета. Серия «Физика математика, информатика». – 2018. – № 5. – С. 15–18.

2-А. Краснопрошин, В. В. Обучение глубоких доверительных сетей на основе метода отжига / В. В. Краснопрошин, В. В. Мацкевич // Вестник Брестского государственного технического университета. Серия «Физика математика, информатика». – 2019. – № 5. – С. 5–8.

3-А. Krasnoproshin, V. V. Neural Network Data Processing Based on Deep Belief Networks / V. V. Krasnoproshin, V. V. Matskevich // Communications in Computer and Information Science. Springer. – 2020. – Vol. 1282. – P. 234–244. – DOI:10.1007/978-3-030-60447-9_15.

4-А. Мацкевич, В. В. Обучение нейронных сетей на основе метода отжига / В. В. Мацкевич // Вестник Полоцкого государственного университета. Серия С. Фундаментальные науки. – 2022. – № 11. – С. 21–29. – DOI: 10.52928/2070-1624-2022-39-11-21-29.

5-А. Matskevich, V. V. An efficient training algorithm of restricted Boltzmann machines / V. V. Matskevich, V. A. Stasiuk // European Proceedings of Computers and Technology. European Publisher. – 2023. – Vol. 1. – P. 296–303. – DOI: 10.15405/epct.23021.36.

6-А. Krasnoproshin, V. V. Neural network software technology trainable on the random search principles / V. V. Krasnoproshin, V. V. Matskevich // Research Papers Collection “Open Semantic Technologies for Intelligent Systems”. – 2023. – Vol. 7. – P. 133–140.

7-А. Krasnoproshin, V. V. Random search in neural networks training / V. V. Krasnoproshin, V. V. Matskevich // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications. – 2024. – Vol. 34, No. 2. – P. 309–316. – DOI: 10.1134/S105466182470010X.

8-А. Zhou, X. Landscape's non-natural changes detection system by satellites images based on local areas / X. Zhou, Q. Bu, V. V. Matskevich, A. M. Nedzved // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications. – 2024. – Vol. 34, No. 2. – P. 365–378. – DOI: 10.1134/S1054661824700159.

Статьи в других научных изданиях

9-А. Krasnoproshin, V. V. Algorithm for fast image compression on

heterogeneous computing devices / V. V. Krasnoproshin, V. V. Matskevich // Research Papers Collection “Open Semantic Technologies for Intelligent Systems”. – 2019. – Vol. 3. – P. 265–268.

10-A. Krasnoproshin, V. V. Software technology for deep learning of belief neural networks / V. V. Krasnoproshin, V. V. Matskevich // Research Papers Collection “Open Semantic Technologies for Intelligent Systems”. – 2020. – Vol. 4. – P. 257–262.

Статьи в сборниках материалов научных конференций

11-A. Мацкевич, В. В. Опыт использования нейронных сетей для распознавания букв / В. В. Мацкевич // Сборник работ 74-й научной конференции студентов и аспирантов Белорусского государственного университета, Т. 1. – Минск: БГУ, 2017. – С. 62–65.

12-A. Мацкевич, В. В. Эффективное обучение нейронных сетей на основе технологии распараллеливания / В. В. Мацкевич // Сборник работ 75-й научной конференции студентов и аспирантов Белорусского государственного университета, Т. 2. – Минск: БГУ, 2018. – С. 245–249.

13-A. Мацкевич, В. В. Алгоритмы обучения ограниченной машины Больцмана / В. В. Мацкевич // Сборник работ 76-й научной конференции студентов и аспирантов Белорусского государственного университета, Т. 1. – Минск: БГУ, 2019. – С. 97–100.

14-A. Krasnoproshin, V. V. Annealing method in training restricted Boltzmann machine / V. V. Krasnoproshin, V. V. Matskevich // Proceedings of the 14-th International Conference PRIP’2019. – Minsk: BSUIR, 2019. – P. 264–268.

15-A. Krasnoproshin, V. V. Statistical approach to image compression based on a restricted Boltzmann machine / V. V. Krasnoproshin, V. V. Matskevich // Proceedings of the 12-th International Conference “Computer Data Analysis and Modeling” – Minsk: BSU, 2019. – P. 207–213.

16-A. Мацкевич, В. В. Возможности метода отжига в задаче обучения нейронных сетей / В. В. Мацкевич // Материалы междунар. науч.-техн. семинара: технологии передачи и обработки информации = Technologies of Information Transmission and Processing. – Минск: БГУИР, 2022. – С. 69–73.

17-A. Мацкевич, В. В. Метод отжига в задаче обучения нейронных сетей / В. В. Мацкевич // Сборник материалов II международной научно-практической конференции «Информационно-коммуникационные технологии: достижения, проблемы, инновации». – Полоцк: ПГУ, 2022. – С. 153–157.

18-A. Мацкевич, В. В. Случайный поиск в задаче обучения нейронных сетей / В. В. Мацкевич // Сборник работ 79-й научной конференции студентов и аспирантов Белорусского государственного университета, Т. 1. – Минск: БГУ, 2022. – С. 521–524.

19-А. Мацкевич, В. В. Эффективный алгоритм обучения нейронных сетей / В. В. Мацкевич // Материалы II Международной научно-технической конференции «актуальные проблемы науки и техники». – РФ, г. Сарепул: ФГБОУ ВО «ИжГТУ им. М.Т. Калашникова», 2022. – С. 422–426.

20-А. Krasnoproshin, V. V. Random search in neural networks training / V. V. Krasnoproshin, V. V. Matskevich // Proceedings of the 13-th International Conference “Computer Data Analysis and Modeling”. – Minsk: BSU, 2022. – P. 96–99.

21-А. Мацкевич, В. В. Эффективный алгоритм обучения на основе случайного поиска / В. В. Мацкевич // Материалы международного научного конгресса по информатике "Информационные системы и технологии", Т. 2. – Минск: БГУ, 2022. – С. 177–182.

22-А. Мацкевич, В. В. Алгоритмы случайного поиска в обучении нейронных сетей / В. В. Мацкевич // Материалы междунар. науч.-техн. семинара: технологии передачи и обработки информации = Technologies of Information Transmission and Processing. – Минск: БГУИР, 2023. – С. 139–144.

23-А. Мацкевич, В. В. Гибридный алгоритм обучения нейронных сетей / В. В. Мацкевич // Сборник работ 80-й научной конференции студентов и аспирантов Белорусского государственного университета, Т. 1. – Минск: БГУ, 2023. – С. 372–375.

24-А. Matskevich, V. V. Fast random search algorithm in neural networks training / V. V. Matskevich // Proceedings of the 16-th International Conference – PRIP’2023. – Minsk: BSU, 2023. – P. 22–24.

25-А. Matskevich, V. V. Neural network software technology trainable on the random search and gradient descent principles / V. V. Matskevich, X. Zhou, Q. Bu // Proceedings of the 16-th International Conference – PRIP’2023. – Minsk: BSU, 2023. – P. 64–67.

Тезисы

26-А. Мацкевич, В. В. Эффективные вычисления на гетерогенных устройствах / В. В. Мацкевич // Тезисы докладов 10-й международной научно-практической конференции “Информационные технологии в промышленности, логистике и социальной сфере”. – Минск: ОИПИ НАН Беларуси, 2019. – С. 143–146.

27-А. Krasnoproshin, V. V. Technology for training neural networks based on the annealing method / V. V. Krasnoproshin, V. V. Matskevich // Book of abstracts of 14-th international conference "Intellectualization of information processing". – Russian Federation, Moscow: Russian Academy of Science, 2022. – P. 161–163.

Мацкевіч Вадзім Уладзіміравіч

**МЕТАДЫ ПАБУДОВЫ І НАВУЧАННЯ ГЛЫБОКІХ ДАВЕРНЫХ
НЕЙРОННЫХ СЕТАК, АСНОВАНЫХ НА АНСАМБЛЯХ
АБМЕЖАВАНЫХ МАШЫН БОЛЬЦМАНА**

Ключавыя словы: глыбокія нейронавыя сеткі, абмежаваная машына Больцмана, алгарытмы навучання, нейросетевая тэхналогія.

Мэтай працы з'яўляецца распрацоўка мадэлі глыбокай давернай сеткі на аснове ансамбля з абмежаваных машын Больцмана і алгарытму навучання на аснове выпадковага пошуку, накіраваных на эфектыўную апрацоўку і аналіз малюнкаў.

Прынцыпы і метады даследавання: у працы выкарыстоўваліся метады тэорыі нейронавых сетак і машыннага навучання, матэматычны аналіз і тэорыя паралельных вылічэнняў.

Атрыманыя вынікі і іх навізна. У дысертацыі распрацаваны мадэлі, алгарытмы і тэхналогія для вырашэння прыкладных задач з выкарыстаннем нейросетевой апрацоўкі дадзеных.

Пабудавана мадэль глыбокай давернай нейронавай сеткі, заснаваная на ансамблі абмежаваных машын Больцмана (з n машын невялікай памернасці), што дазволіла знізіць колькасць параметраў і час навучання з захаваннем якасці рашэння.

Распрацаваны алгарытм навучання, заснаваны на мадыфікацыі аптымізацыйнай схемы Больцманаўскага адпалу і які ўлічвае неаднароднасць прасторы рашэнняў, што дазволіла зменшыць час навучання.

Прапанавана метадыка навучання для алгарытмаў выпадковага пошуку, якая забяспечвае іх інварыянтнасць да аб'ёму навучальнай выбаркі з захаваннем якасці рашэння.

Распрацавана тэхналогія для канструявання і навучання глыбокіх нейронавых сетак на гетэрагенных сістэмах з падтрымкай мадэлі ў выглядзе ансамбляў з абмежаваных машын Больцмана і алгарытмаў выпадковага пошуку, рэалізаваная ў выглядзе крос-платформавага праграмнага комплексу (фрэймворка).

Рэкамендацыі да выкарыстання і вобласць ужывання: машыннае навучанне, штучны інтэлект, сістэмы маніторынгу на аснове дадзеных дыстанцыйнага задання.

Мацкевич Вадим Владимирович

**МЕТОДЫ ПОСТРОЕНИЯ И ОБУЧЕНИЯ ГЛУБОКИХ
ДОВЕРИТЕЛЬНЫХ НЕЙРОННЫХ СЕТЕЙ, ОСНОВАННЫХ
НА АНСАМБЛЯХ ОГРАНИЧЕННЫХ МАШИН БОЛЬЦМАНА**

Ключевые слова: глубокие нейронные сети, ограниченная машина Больцмана, алгоритмы обучения, нейросетевая технология

Целью работы является разработка модели глубокой доверительной сети на основе ансамбля из ограниченных машин Больцмана и алгоритма обучения на основе случайного поиска, направленных на эффективную обработку и анализ изображений.

Принципы и методы исследования: в работе использовались методы теории нейронных сетей и машинного обучения, математический анализ и теория параллельных вычислений.

Полученные результаты и их новизна. В диссертации разработаны модели, алгоритмы и технология для решения прикладных задач с использованием нейросетевой обработки данных.

Построена модель глубокой доверительной нейронной сети, основанная на ансамбле ограниченных машин Больцмана (из n машин небольшой размерности), что позволило снизить количество параметров и время обучения с сохранением качества решения.

Разработан алгоритм обучения, основанный на модификации оптимизационной схемы Больцмановского отжига и учитывающий неоднородность пространства решений, что позволило снизить время обучения.

Предложена методика обучения для алгоритмов случайного поиска, обеспечивающая их инвариантность к объему обучающей выборки с сохранением качества решения.

Разработана технология для конструирования и обучения глубоких нейронных сетей на гетерогенных системах с поддержкой модели в виде ансамблей из ограниченных машин Больцмана и алгоритмов случайного поиска, реализованная в виде кросс-платформенного программного комплекса (фреймворка).

Рекомендации к использованию и область применения: машинное обучение, искусственный интеллект, системы мониторинга на основе данных дистанционного зондирования.

RESUME

Matskevich Vadim Vladimirovich

METHODS FOR CONSTRUCTING AND TRAINING DEEP BELIEF NEURAL NETWORKS BASED ON ENSEMBLES OF RESTRICTED BOLTZMANN MACHINES

Key words: deep neural networks, restricted Boltzmann machine, training algorithms, neural network technology.

The aim of the work is to develop a deep belief network model based on an ensemble of restricted Boltzmann machines and a training algorithm based on random search aimed at efficient image processing and analysis.

Principles and methods of research: in the work was used methods of neural network theory and machine learning, mathematical analysis and the theory of parallel computing.

Obtained results and their novelty. The thesis develops models, algorithms and technology for solving applied problems using neural network data processing.

A model of a deep belief neural network based on an ensemble of restricted Boltzmann machines (of n small-sized machines) is built, which made it possible to reduce the number of parameters and training time while maintaining the solution quality.

A training algorithm based on a modification of the Boltzmann annealing optimization scheme and taking into account the heterogeneity of the solution space is developed, which made it possible to reduce the training time.

A training technique for random search algorithms is proposed, ensuring their invariance to the training dataset size, while maintaining the solution quality.

A technology has been developed for constructing and training deep neural networks on heterogeneous systems with model support in the form of ensembles of restricted Boltzmann machines and random search algorithms, implemented as a cross-platform software package (framework).

Recommendations for use and application scope: machine learning, artificial intelligence, monitoring systems based on remote sensing data.



Подписано в печать 11.03.2025. Формат 60×84 $\frac{1}{16}$. Бумага офсетная.
Цифровая печать. Усл. печ. л. 1,39. Уч.-изд. л. 1,5.
Тираж 60 экз. Заказ 32.

Отпечатано с оригинала-макета заказчика
в республиканском унитарном предприятии
«Издательский центр Белорусского государственного университета».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 2/63 от 19.03.2014.
Ул. Красноармейская, 6, 220030, Минск.