

**УО «БРЕСТСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ»**

УДК 004.032.26

**КРОЩЕНКО
Александр Александрович**

**МЕТОДЫ ОБУЧЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ
ДЛЯ ЗАДАЧ КОМПЬЮТЕРНОГО ЗРЕНИЯ**

Диссертация на соискание ученой степени
кандидата технических наук

по специальности 05.13.17 – Теоретические основы информатики

Научный руководитель –
доктор технических наук, профессор
Головко В. А.

Брест 2023

СОДЕРЖАНИЕ

Перечень сокращений и обозначений	4
Введение	6
Общая характеристика работы	7
ГЛАВА 1 Проблемы обучения глубоких нейронных сетей	11
1.1 Общие сведения	11
1.2 Проблемы обучения глубоких моделей	16
1.3 Методы обучения	18
1.4 Критерии минимизации при обучении	24
1.5 Выводы	27
ГЛАВА 2 Метод неконтролируемого предобучения	28
2.1 Обучение RBM	28
2.2 Общее описание предлагаемого метода	34
2.3 Вывод правил предобучения	35
2.3.1 Критерий MSE	36
2.3.2 Критерий СЕ	41
2.4 Предобучение сверточных слоев	45
2.5 Алгоритм редуцирования параметров	49
2.6 Выводы	52
ГЛАВА 3 Экспериментальное обоснование метода неконтролируемого предобучения	55
3.1 Используемое аппаратное и программное обеспечение	55
3.2 Сжатие данных	57
3.3 Классификация	58
3.3.1 Критерий оценки результатов классификации	58
3.3.2 Решение простейшей задачи классификации	59
3.3.3 Описание применяемых выборок	62
3.3.4 Параметры вычислительного эксперимента и результаты	64
3.4 Редуцирование параметров НС	66
3.5 Выводы	71
ГЛАВА 4 Разработка систем компьютерного зрения на основе глубоких нейронных сетей	72
4.1 Архитектура системы компьютерного зрения	72
4.2 Обнаружение солнечных панелей на аэрофотоснимках	74
4.2.1 Постановка задачи и обзор существующих решений	74
4.2.2 Предлагаемый подход	75
4.2.3 Обучающая выборка	78

4.2.4	Оценка качества обученных моделей	79
4.2.5	Результаты обучения и тестирования	81
4.3	Распознавание маркировки продукта	82
4.3.1	Постановка задачи и обзор существующих решений	82
4.3.2	Предлагаемый подход	86
4.3.3	Обучающие выборки	91
4.3.4	Результаты обучения и тестирования	92
4.4	Выводы	94
	Заключение	95
	Список использованных источников	97
	Библиографический список	97
	Список публикаций соискателя ученой степени	103
	ПРИЛОЖЕНИЕ А Исходный код реализации процедур предобучения и редуцирования	108
	ПРИЛОЖЕНИЕ Б Документы об использовании результатов диссертационной работы	119

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

ГНС – глубокая нейронная сеть

ИНС – искусственная нейронная сеть

КЗ – компьютерное зрение

НС – нейронная сеть

СНС – сверточная нейронная сеть

AP (англ. *Average Precision metric*) – метрика, используемая для оценки методов детекции объектов на изображениях, применяется для одноклассовых случаев

CD (англ. *Contrastive Divergence*) – контрастное расхождение

CE (англ. *Cross-Entropy error*) – кросс-энтропийная функция ошибок

CIFAR (англ. *Canadian Institute For Advanced Research*) – Канадский институт перспективных исследований

COCO (англ. *Common Objects in Context*) – выборка размеченных изображений, применяемая для решения задач обнаружения, сегментации и аннотирования объектов на изображениях

CV (англ. *Computer Vision*) – компьютерное зрение

C-RBM (англ. *Classic Restricted Boltzmann Machine Training*) – классический метод обучения ограниченной машины Больцмана

DNN (англ. *Deep Neural Network*) – глубокая нейронная сеть

IoU (англ. *Intersection over Union, мера Жаккара*) – метрика, применяемая для оценки эквивалентности двух множеств, в задачах компьютерного зрения применяется для оценки методов детекции (обнаружения) объектов как величина, описывающая степень перекрытия двух прямоугольных областей, используемых для определения эталонной и предсказываемой моделью локализации объекта

HREBA (англ. *Hybrid Reconstruction Error-Based Approach*) – гибридный вариант предобучения ГНС, сочетающий в себе использование классического (C-RBM) и предлагаемого в данной работе (REBA) методов обучения ограниченной машины Больцмана

mAP (англ. *mean-Average Precision metric*) – метрика, применяемая для оценки качества алгоритмов детекции объектов на изображениях, применяется для многоклассовых случаев детекции

MNIST (англ. *Mixed National Institute of Standards and Technology database*) – смешанная база Национального института стандартов и технологий

MSE (англ. *Mean Squared Error*) – среднеквадратичная ошибка

NLP (англ. *Natural Language Processing*) – обработка естественного языка

Faster-RCNN (англ. *Faster Region-based Convolutional Neural Network*) – ар-

хитектура глубокой сверточной нейронной сети, применяемая для решения задачи детекции (обнаружения) объектов на изображениях

PCA (англ. *Principal Component Analysis*) – метод главных компонент

RBM (англ. *Restricted Boltzmann Machine*) – ограниченная машина Больцмана

REBA (англ. *Reconstruction Error-Based Approach*) – метод обучения ограниченной машины Больцмана, базирующийся на минимизации ошибки восстановления образов на видимом и скрытом слоях

ReLU (англ. *Rectified Linear Unit*) – исправленный линейный элемент

ResNet-50 (англ. *Residual Neural Network*) – архитектура глубокой нейронной сети, которая характеризуется передачей сигналов между отдельными слоями модели, применяется в качестве классификатора

SVM (англ. *Support Vector Machines*) – метод опорных векторов

SVHN (англ. *The Street View House Numbers*) – выборка изображений номеров домов

ВВЕДЕНИЕ

Практические приложения компьютерного зрения с каждым годом становятся все более разнообразными. В современном мире компьютерное зрение используется повсеместно – в сложных производственных и медицинских системах, в интеллектуальных системах интернета вещей и развлекательных приложениях для мобильных устройств.

В качестве основы при разработке таких систем все чаще находят применение глубокие нейросетевые модели. Данные модели показывают впечатляющие результаты при решении самых разнообразных задач компьютерного зрения – распознавания, детекции и сегментации объектов на фото- и видеоизображениях, получения аннотаций для фотографий и генерации изображений по текстовому описанию. Глубокие нейронные сети, применяемые для решения подобных задач, содержат миллионы настраиваемых параметров и, для некоторых архитектур, десятки слоев нейронных элементов.

Обучение подобных «тяжелых» моделей с нуля является нетривиальной задачей. Оно часто сопряжено с риском переобучения, результатом которого является отличная приспособленность сети к данным из обучающей выборки, но плохая обобщающая способность, то есть неэффективность модели для данных, не использовавшихся при обучении. Чаще всего переобученность возникает при применении малой обучающей выборки. Другой проблемой является ресурсоемкость процесса обучения таких моделей, даже при использовании современных технических средств.

Проблемы обучения глубоких нейронных сетей активно изучаются в зарубежных научных школах. В нашей стране такие исследования также проводятся. Однако, нужно отметить, что такие исследования часто носят эмпирический характер, поэтому разработка строгих математически обоснованных методов обучения остается важной задачей теории нейронных сетей.

В диссертационной работе разработаны подходы для неконтролируемого предобучения глубоких нейронных сетей и редуцирования параметров моделей. Предложены алгоритмы для решения практических задач теории компьютерного зрения – обнаружения и локализации солнечных панелей на аэрофотоснимках, обнаружения и распознавания маркировки на поточных производственных линиях. Предложенные методы и алгоритмы позволяют улучшить работу интеллектуальных систем, использующих полно связные и сверточные нейросетевые модели.

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 МОРБ «Алгоритмы интеллектуального анализа и обработки больших объемов данных на основе нейронных сетей глубокого доверия» (ГБ 15/203, № госрегистрации 20150743),
2. ГПНИ «Информатика и космос, научное обеспечение безопасности и защиты от чрезвычайных ситуаций» по заданию «Нейросетевые методы обработки комплексной информации и принятия решений на основе интеллектуальных многоагентных систем» (№ госрегистрации 20140547),
3. ГПНИ «Информатика и космос, научное обеспечение безопасности и защиты от чрезвычайных ситуаций» по заданию «Методы и алгоритмы интеллектуальной обработки и анализа большого объема данных на основе нейронных сетей глубокого доверия» (задание 1.6.05, № госрегистрации 20163595),
4. НИР «Методы и алгоритмы построения интеллектуальных систем анализа и обработки данных», этап «Разработка гибридных интеллектуальных систем на основе нейросимволического подхода» (решение НТС УО «Брестский государственный технический университет» от 12.11.2021, протокол № 6, № 22202052022070),
5. НИР БРФФИ «Модели и исследование 3-Д оцифровки на основе фактических данных и анализа гетерогенных данных» (№ Ф22КИ-046 от 05.11.2021 г., № госрегистрации 20220090).

Цель, задачи, объект и предмет исследования

Целью исследования является разработка эффективных методов и алгоритмов для обучения глубоких нейронных сетей, используемых для решения задач компьютерного зрения, включающих обнаружение солнечных панелей на аэрофотоснимках и распознавание маркировки продукта на конвейерной линии.

Указанная цель определяет следующие задачи исследования:

- 1) Разработать метод неконтролируемого предобучения глубоких нейронных сетей, позволяющий повысить эффективность обучения моделей;
- 2) Разработать алгоритм редуцирования параметров глубоких нейронных сетей, позволяющий упростить структуру моделей;
- 3) Провести сравнительный анализ эффективности разработанных метода обучения и алгоритма редуцирования;
- 4) Разработать нейросетевую систему компьютерного зрения для решения прикладных задач (обнаружения солнечных панелей на аэрофотоснимках и распознавания маркировки продукта на конвейерной линии) с применением предлагаемого метода обучения.

Объектом исследования являются нейросетевые системы компьютерного зрения. Предметом исследования выступают методы и алгоритмы обучения глубоких нейронных сетей и их применение к задачам компьютерного зрения.

Научная новизна

Научная новизна состоит в установлении эквивалентности задач максимизации функции правдоподобия распределения входных данных, минимизации кросс-энтропийной функции ошибки и суммарной квадратичной ошибки при использовании линейных нейронов в пространстве синаптических связей ограниченной машины Больцмана, что позволяет учитывать нелинейную природу нейронных элементов.

Разработан метод обучения ограниченной машины Больцмана на основе доказанных эквивалентностей, применение которого для предобучения ГНС позволяет расширить класс обучаемых моделей и повысить обобщающую способность ГНС.

Разработан алгоритм редуцирования параметров нейросетевой модели, основывающийся на неконтролируемом предобучении, что позволяет уменьшить количество настраиваемых параметров модели без потери обобщающей способности.

Разработана нейросетевая система компьютерного зрения, которая основана на предлагаемом методе предобучения нейросетевых моделей, применение которой позволяет улучшить качество решения задач классификации. Продемонстрирована эффективность системы на примерах решения задач обнаружения солнечных панелей на аэрофотоснимках и распознавания маркировки продуктов на производственной линии.

Положения, выносимые на защиту

1. Установление эквивалентности задач максимизации функции правдоподобия распределения входных данных, минимизации кросс-энтропийной функции ошибки и суммарной квадратичной ошибки при использовании линейных нейронов в пространстве синаптических связей ограниченной машины Больцмана, что позволяет учитывать нелинейную природу нейронных элементов.
2. Метод обучения ограниченной машины Больцмана, основывающийся на установленной эквивалентности, что позволяет расширить класс обучаемых моделей и повысить обобщающую способность глубоких нейронных сетей.
3. Алгоритм редуцирования параметров глубокой нейронной сети, основывающийся на неконтролируемом предобучении сети, что позволяет упростить ее архитектуру (путем сокращения числа настраиваемых параметров модели) без потери обобщающей способности.
4. Нейросетевая система компьютерного зрения, основывающаяся на предлагаемом методе предобучения нейросетевых моделей, применение которой позволяет улучшить качество решения задач классификации (определения наличия солнечных панелей на 7,31% и определения положения продукта на конвейерной линии на 5,24%).

Личный вклад соискателя ученой степени в результаты диссертации

Основные положения диссертации получены соискателем лично. Соавтором основных публикаций автора является научный руководитель д.т.н., профессор В.А. Головко, который осуществлял определение целей и постановку задач исследований, выбор методов исследований, принимал участие в планировании работ и обсуждении результатов. В диссертационную работу не включены результаты, которые были получены другими соавторами или с 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: «8th International Conference on Neural Networks and Artificial Intelligence» (Брест, 3-6 июня 2014 г.); «8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and

Applications» (Варшава, 24-26 сентября 2015 г.); «International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)» (Харьков, 9-12 октября 2018 г.); «14th International Conference on Pattern Recognition and Information Processing (PRIP)» (Минск, 21-23 мая 2019 г.); «Вычислительные методы, модели и образовательные технологии» (Брест, 22-23 октября 2013 г., 15-16 октября 2014 г., 22 октября 2015 г., 21 октября 2016 г., 18 октября 2019 г.); «Современные проблемы математики и вычислительной техники» (Брест, 19-21 ноября 2015 г.); «Открытые семантические технологии проектирования интеллектуальных систем» (Минск, 19-21 февраля 2015 г., 15-17 февраля 2018 г., 21-23 февраля 2019 г., 19-22 февраля 2020 г., 16-18 сентября 2021 г., 20-22 апреля 2023 г.); «11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)» (Краков, 22-25 сентября 2021 г.); «Международный конгресс по информатике: информационные системы и технологии (CSIST'2022)» (Минск, 27-28 октября 2022 г.).

По результатам диссертации получено 3 акта о внедрении.

Опубликованность результатов диссертации

Основные результаты диссертационного исследования опубликованы в 30 научных работах, среди которых: 8 статей в научных изданиях в соответствии с пунктом 19 Положения о присуждении ученых степеней и присвоении ученых званий (3,87 авторских листа), 5 статей в других научных изданиях, 13 статей в сборниках материалов научных конференций и 4 тезисов.

Структура и объем диссертации

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

Полный объем диссертации составляет 117 страниц, из которых 44 рисунка на 32 страницах, 17 таблиц на 9 страницах, 2 приложения на 14 страницах. Список использованных источников состоит из 121 наименования, включая 30 публикаций автора.

ГЛАВА 1

ПРОБЛЕМЫ ОБУЧЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

В данной главе вводится теоретическая основа для понимания основных концепций теории обучения глубоких нейронных сетей. Приводится классификация основных типов слоев и архитектур моделей. Описывается историческая ретроспектива развития теории обучения глубоких моделей, а также основные проблемы, которые возникали и возникают при выполнении обучения.

Осуществляется обзор существующих подходов для обучения глубоких нейронных сетей. Приводится описание методов обучения с использованием стохастического градиентного спуска и функций активации ReLU с ее вариантами. Описываются основные проблемы, присущие подобному типу обучения.

Особое внимание уделяется методам неконтролируемого предобучения, для реализации которых не требуется наличие объемной обучающей выборки. При выполнении предобучения осуществляется начальная инициализация параметров глубокой нейронной сети. Подобная инициализация позволяет уменьшить величину ошибки на этапе дообучения («тонкой настройки») и гарантирует его сходимость.

Приводятся алгоритмы предобучения с использованием RBM и автоэнкодерных моделей, формируемых из слоев ГНС и описывается этап «тонкой настройки» нейронной сети, цель которого состоит в дообучении модели для решения определенной задачи.

1.1 Общие сведения

Вопросы разработки и применения нейросетевых моделей часто рассматриваются в контексте более широкой области машинного обучения, являющейся по сути одним из направлений современной науки об искусственном интеллекте. При этом само машинное обучение сформировалось на стыке нескольких научных дисциплин, включающих вычислительную математику, теорию вероятностей и статистику, линейную алгебру и теорию оптимизации. С помощью алгоритмов машинного обучения создаются модели, которые используют предшествующий опыт для повышения эффективности компьютера в решении определенных классов задач [1, с. 2].

Под понятием предшествующего опыта здесь понимается совокупность статистических данных, собранных при наблюдении и описании некоторого процесса.

Являясь моделями машинного обучения, нейронные сети также способны к решению различных типов задач после реализации итеративного алгоритма настройки параметров, называемого обучением. Само свойство обучаемости приближает искусственные нейросети к их естественному прототипу, а именно к человеческому мозгу, обучение которого является важным процессом и формирует индивидуальность [2, с. 316-317].

В основе теории искусственных нейронных сетей лежит биологически инспирированная модель искусственного нейрона, представляющая собой вычислительный элемент, выполняющий преобразование входных данных (рисунок 1.1). Данное преобразование заключается в вычислении некоторой функции активации (линейной или нелинейной) F от взвешенной суммы S компонент вектора входных данных $X = (x_1, x_2, \dots, x_n)$. Параметры $W = (w_1, w_2, \dots, w_n)$ называются весовыми коэффициентами (весами) нейрона и настраиваются в процессе обучения в соответствии с определенным правилом обучения. Помимо весовых коэффициентов, у нейрона может быть дополнительный параметр T , называемый порогом (или пороговым коэффициентом). Он определяет смещение взвешенной суммы S .

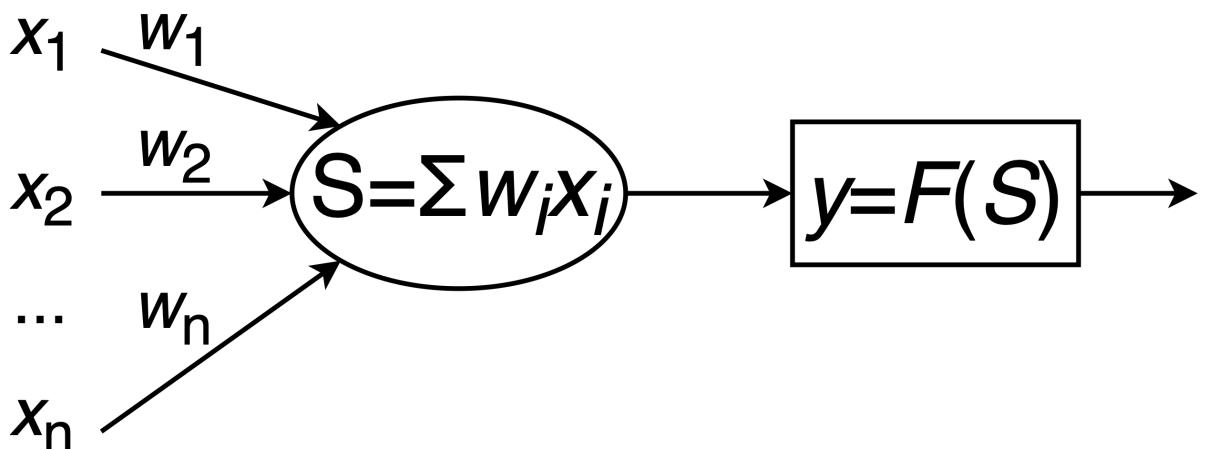


Рисунок 1.1 – Модель искусственного нейрона

Искусственные нейроны могут быть объединены в группы, называемые слоями, по аналогии с биологическими нейронными слоями. Каждый нейрон слоя, обладая независимым набором весовых коэффициентов, получает на вход одинаковый вектор данных X и выполняет его преобразование, формируя отдельную компоненту выходного вектора Y , называемого паттерном выходной активности слоя [3, с. 27] (рисунок 1.2). Весовые ко-

эффициенты отдельных нейронов для удобства объединяются в матрицу, которая называется матрицей весовых коэффициентов слоя (или матрицей весов). Слой, имеющий настраиваемые параметры, называется обрабатывающим (в отличие от распределительных слоев, которые являются предваряющими в любой НС и не имеют параметров).

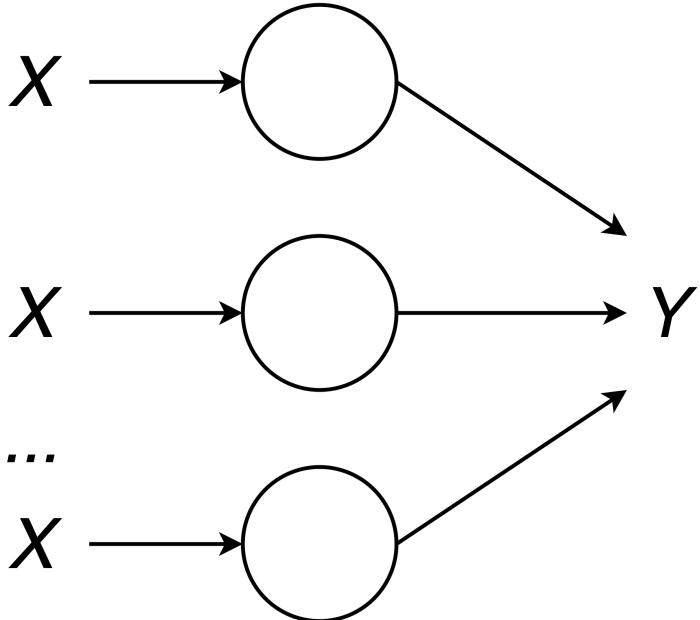


Рисунок 1.2 – Слой нейронных элементов

Наконец, слои нейронных элементов объединяются в последовательности, формируя общую архитектуру нейронной сети (последовательность слоев, их тип, количество нейронов в каждом слое, функцию активации и другие характеристики). Итоговая модель способна выполнять аппроксимацию любой непрерывной функции с любой точностью после выполнения обучения [4, с. 312]. В процессе обучения данные из некоторой выборки по-даются на вход сети с целью определения градиентов изменения параметров нейронных элементов. Одна итерация, в ходе которой модели предъявляются все обучающие примеры, называется эпохой. Данная процедура выполняется до тех пор, пока не будет достигнут ответ сети, полученный с желаемой точностью или пока весовые коэффициенты не войдут в некоторое стационарное состояние.

Искусственные нейронные сети могут быть классифицированы по различным признакам (например, по характеру обучения, по направлению связей, по архитектуре и так далее). В литературе часто используется классификация по архитектурному признаку. Согласно ей, выделяют [3, с. 24-25]:

- 1) сети персептронного типа (многослойные персептроны [5, с. 92-94], рекуррентные [6, с. 182-185; 7, с. 5-6], рециркуляционные [8, с. 360-361],

сверточные [9, с. 194-197; 10, с. 398-402], глубокие);

2) самоорганизующиеся НС (НС Кохонена [11, с. 105-106], сети адаптивного резонанса [12, с. 33-34]);

3) релаксационные НС (сети Хопфилда [13, с. 3089], Хэмминга [14, с. 8-10] и двунаправленная ассоциативная память [15, с. 52-54]);

4) гибридные НС (сети встречного распространения [16], RBF-сети [17, с. 326-327], нечеткие сети [18]);

5) нейронные иммунные сети [19, с. 485].

В данной работе исследуются глубокие нейронные сети, являющиеся развитием архитектур полносвязных сетей (персепtronов) и сверточных нейронных сетей.

В настоящий момент сверточные нейронные сети и глубокие архитектуры на их основе являются доминирующими в области компьютерного зрения [20, с. 436]. В основном это связано с тем, что в подобных задачах используются данные с высокой корреляцией между соседними элементами (для цифровых фотографий и видео в качестве этих значений выступают массивы пикселей), а сверточные слои особенно эффективны при обработке именно таких данных [21, с. 8].

Сверточные нейронные сети базируются на моделях неокогнитрона, предложенной К. Фукусимой [9, с. 193], и сетях с разделяемыми весами. Первая классическая модель сверточной нейронной сети была разработана ЛеКуном и получила название LeNet-5 [22, с. 7]. Сверточные нейросети интегрировали три концепции, называемые локальным рецептивным полем, разделяемыми весовыми коэффициентами и пространственной подвыборкой. Используя локальное рецептивное поле, нейронные элементы первого сверточного слоя могут извлекать простейшие признаки, такие как края, углы и так далее. Первые слои сверточной нейронной сети представляют собой комбинацию сверточных (convolutional) и подвыборочных (pooling) слоев (рисунок 1.3), благодаря чему осуществляется нелинейное иерархическое преобразование входной информации. При решении задачи классификации в качестве последнего слоя могут использоваться SVM-классификатор, многослойный персептрон или любой другой классификатор (на рисунке обозначен как classifier).

Таким образом, сверточные модели могут включать как сверточные слои, так и некоторое количество полносвязных слоев (завершающие полносвязные слои часто используются как классификаторы признаков, извлеченных на начальных сверточных слоях).

Несмотря на лидирующее положение сверточных сетей, полносвязные архитектуры также успешно применяются для некоторых задач ком-

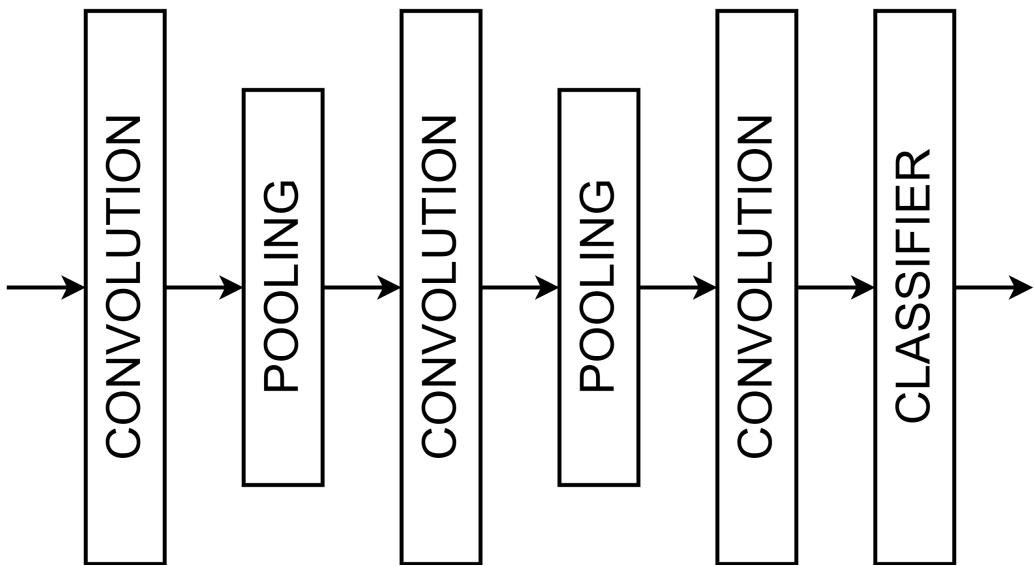


Рисунок 1.3 – Общий вид сверточной нейронной сети

пьютерного зрения (например, при решении задачи трехмерной реконструкции [23, с. 10]). Далее в работе будет показано, что сверточный слой может быть представлен с помощью разряженного слоя. Таким образом, исследование и обучение полносвязных архитектур остается актуальной задачей.

Глубокую нейронную сеть можно определить как искусственную нейронную сеть персептронного типа, имеющую более двух скрытых слоев нейронных элементов.

В общем случае глубокая нейронная сеть (deep neural network – DNN) представляет собой нейросетевую модель с множеством слоев нейронных элементов [24, с. 59]. Рассмотрим многослойный персептрон – полносвязную нейросетевую модель, представленную на рисунке 1.4). Известно, что такая сеть способна к выделению глубокой иерархии признаков [25, с. 3].

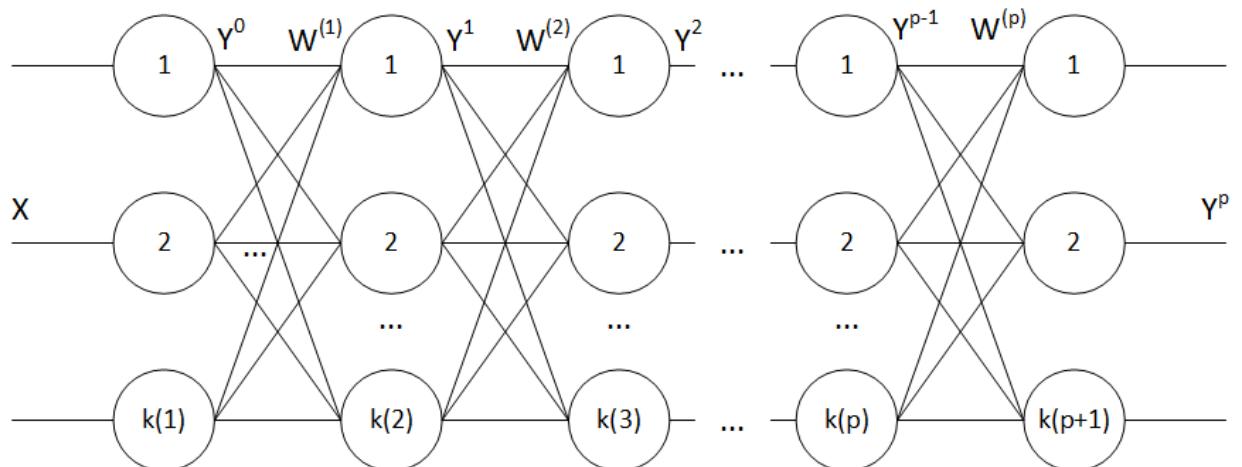


Рисунок 1.4 – Глубокая нейронная сеть

На первом скрытом слое выделяются признаки низкого уровня абстракции, на втором слое – более высокого уровня и так далее [26].

Выходное значение j -го нейрона k -го слоя определяется следующим образом:

$$y_j^k = F(S_j^k), \quad (1.1)$$

$$S_j^k = \sum_{i=1} w_{ij}^k y_i^{k-1} + T_j^k, \quad (1.2)$$

где $F(\cdot)$ – функция активации нейронного элемента k -го слоя,
 S_j^k – взвешенная сумма j -го нейрона k -слоя,
 w_{ij}^k – весовой коэффициент между i -ым нейроном $(k-1)$ -го слоя и j -м нейроном k -го слоя,
 T_j^k – пороговое значение j -го нейрона k -го слоя.

Для первого слоя НС, называемого распределительным, справедливо

$$y_j^0 = x_j. \quad (1.3)$$

В матричном виде выходной вектор k -го слоя

$$Y^k = F(S^k) = F((Y^{k-1})^T W^k + T^k), \quad (1.4)$$

где W^k – матрица весовых коэффициентов k -го слоя,
 Y^{k-1} – выходной вектор-столбец $(k-1)$ -го слоя,
 T^k – вектор пороговых значений нейронов k -го слоя.

Если глубокая нейронная сеть используется для классификации образов, то выходные значения сети часто определяются на основе функции активации **softmax** [3, с. 34]:

$$y_j^F = softmax(S_j) = \frac{e^{S_j}}{\sum_i e^{S_i}}. \quad (1.5)$$

В этом случае значения, получаемые на последнем слое нейронной сети, отражают вероятность принадлежности данного образа к определенному классу.

1.2 Проблемы обучения глубоких моделей

Теория нейронных сетей не развивалась равномерно. Нередко причинами приостановок в исследованиях являлось отсутствие доступных на определенный момент времени эффективных методов обучения моделей.

Начиная с первых моделей ИНС, описанных в работе У. Мак-Каллока и У. Питтса [27, с. 105] в 1943 году, искусственным нейронным сетям были посвящены концептуальные и важные работы того времени (например, правила обучения Д. Хебба [28, с. 62], В. Видроу и М. Хоффа [29, с. 99]). В 1959 году Ф. Розенблatt предложил модель персептрона [30, с. 99]. Однако, после опубликования монографии М. Минского и С. Пайперта «Персептроны» в 1969 году [31, с. 249] крупные исследования в этой области были приостановлены. В этой работе был проведен анализ персептронной модели и выявлены ее основные ограничения, в том числе касающиеся невозможности эффективного решения существующими на тот момент моделями некоторых задач (например, задач инвариантного представления образов).

В конце 1970-х годов произошел очередной всплеск работ, посвященных нейросетевым моделям (например, сетям адаптивного резонанса [32, с. 187], ассоциативной памяти [33, с. 69]). В 1986 году Д. Румельхартом, Дж. Хинтоном и Р. Вильямсом в работе [34, с. 4-8] был предложен метод обратного распространения ошибки для обучения многослойных моделей. Результаты, изложенные в данной работе, открыли теоретические возможности для обучения моделей с большим количеством слоев. Однако, вплоть до середины 2000-х годов сети с количеством слоев три и более широко не применялись, так как считалось, что использование глубоких моделей не дает существенного прироста в эффективности по сравнению с другими подходами. Неэффективность была связана с проблемой «исчезающего» градиента. Она проявляется в том, что при обучении ГНС методом обратного распространения ошибки значения градиентов весовых коэффициентов первых слоев сети быстро становятся близкими к нулю, в результате весовые коэффициенты таких слоев практически не изменяются [25, с. 38]. Это существенно замедляет процесс обучения, делая невозможным практическое применение подобных моделей по сравнению с другими существовавшими на тот момент времени методами (например, [35, с. 274]).

Тем не менее, проводя аналогии с многоуровневыми нейронными архитектурами в человеческом мозге (в частности, строением вентральной зрительной системы), было понятно, что подобный тип искусственных нейронных сетей обладает полезными свойствами, позволяя формировать многоуровневую иерархию признаков [36, с. 66].

В 2006 году Джекфри Хинтоном был предложен подход к обучению глубоких моделей, который ознаменовал начало новой эпохи в развитии теоретических работ в данной области [37, с. 1533-1538].

В предложенном методе использовался «жадный» алгоритм послойного предобучения глубоких нейронных сетей (greedy layer-wise algorithm)

[38, с. 3-4], основанный на применении ограниченных машин Больцмана, последовательно формируемых и обучаемых из слоев глубокой нейронной сети.

С этой работы фактически начался новый подъем в исследованиях нейронных сетей, который длится до настоящего времени. За последние годы в области компьютерного зрения благодаря применению глубоких нейронных сетей удалось значительно улучшить результаты для различных задач, включая классификацию, детекцию, сегментацию объектов на изображении, генерацию текстовых описаний к изображениям, генерацию изображений по текстовым описаниям и т.д.

В течение следующих лет сместился основной акцент в исследованиях – вместо послойного обучения стали применяться методы, позволяющие обучать глубокие нейронные сети, начиная с произвольной начальной инициализации, без использования предобучения (например, стохастический градиентный спуск с использованием функции активации ReLU и ее вариантов [39, с. 318]). Однако, применение таких подходов не снижает требований к объему обучающей выборки, которая в идеальных теоретических условиях должна быть сравнима с числом настраиваемых параметров модели, поскольку большие модели, обучаемые на малых выборках, с большей вероятностью будут переобучаться. Результатом этого является отличная приспособленность модели к обучающей выборке, но плохая обобщающая способность, т.е. неэффективность сети на тестовых данных. Таким образом, объем используемой обучающей выборки для глубоких нейронных сетей остается по-прежнему критичным фактором для обучения. Даже наличие подходящей по объему выборки не гарантирует успешное завершение обучения, так как подготовка больших моделей может повлечь неприемлемые аппаратные издержки [40; 41; 42, с. 1].

1.3 Методы обучения

Рассмотрим основные методы, применяемые для обучения ГНС.

Существуют два базовых подхода к обучению глубоких нейронных сетей, оба из которых включают этап предобучения:

- 1) обучение с использованием предобучения на большой обучающей выборке и любого оптимизирующего метода для настройки параметров нейронной сети (I тип);
- 2) обучение с использованием неконтролируемого предобучения (II тип).

Предобучение I типа представляет собой обучение модели с исполь-

зованием метода обратного распространения ошибки, некоторого оптимизирующего метода и активационных функций ReLU и ее вариантов в качестве функций активации нейронов [20, с. 436-438], а затем дообучение полученной модели для новой выборки (transfer learning).

Среди основных оптимизирующих методов, применяемых для предобучения I типа, выделяют [43, с. 263-266]:

1) SGD (стохастический градиентный спуск). В данном методе корректировка настраиваемых параметров ИНС выполняется в направлении, противоположном вектору градиента функции потерь [44, с. 225-237];

2) метод Нестерова. Обучение методом стохастического градиентного спуска нередко происходит очень медленно. Метод Нестерова является вариантом импульсного алгоритма, в котором градиент вычисляется после применения текущей скорости [43, с. 258]. Применение импульсного метода помогает ускорить процесс обучения в случаях зашумленных или небольших по величине градиентов;

3) AdaGrad. В данном методе осуществляется независимая адаптация скоростей обучения всех настраиваемых параметров ИНС [45, с. 2124-2125];

4) RMSProp. Данный метод является модификацией AdaGrad, которая позволяет улучшить его поведение в случае невыпуклой целевой функции [46, с. 1];

5) Adam. Данный метод можно рассматривать как комбинацию RMSProp и AdaGrad [47, с. 5]. Помимо усредненного первого момента, данный метод использует усредненное значение вторых моментов градиентов.

I тип фактически представляет собой обучение с нуля, которое осуществляется на больших выборках данных с использованием подходящих аппаратных средств. Предобученные таким образом модели могут быть использованы для решения других задач. Таким образом, современный инженер получает в свое распоряжение репозиторий готовых предобученных моделей, которые могут быть дообучены на небольших выборках данных, предназначенных для решения конкретной задачи.

Рассмотрим подробно реализацию предобучения II типа.

Исследование методов предобучения II типа остается актуальной и перспективной задачей, поскольку такие методы позволяют существенно сократить размер применяемой обучающей выборки, при этом не переобучая модель [20, с. 439]. Эффектом, достигаемым за счет данного свойства, является снижение требований к используемому для обучения оборудованию.

Обучение нейронной сети с предобучением II типа можно разделить на два этапа [26; 37, с. 1538; 48, с. 18]:

- 1) этап предобучения ГС методом послойного обучения (pre-training).
Осуществляется неконтролируемо;
- 2) дообучение (fine-tuning) при помощи алгоритмов «бодрствования и сна» (wake-sleep algorithm) [49, с. 2] или обратного распространения ошибки.

В ходе **первого этапа** осуществляется инициализация параметров нейросетевой модели. Данный этап является определяющим для последующего дообучения модели. Для реализации первого этапа используются два основных подхода (рисунок 1.5). Формирование вспомогательных моделей (ограниченных машин Больцмана, автоэнкодеров) в обоих подходах происходит из структуры исходной ГС.



Рисунок 1.5 – Методы предварительного обучения ГС

Первый подход основывается на представлении каждого слоя нейронной сети в виде ограниченной машины Больцмана (RBM). При использовании автоэнкодерного подхода, каждый слой представляется автоассоциативной нейронной сетью.

Рассмотрим каждый из этих методов подробнее.

При предобучении ГС в соответствии с **первым подходом** осуществляется построение последовательности RBM-сетей из параметров скрытых слоев ГС и их обучение.

Ограниченнная машина Больцмана состоит из двух слоев стохастических бинарных нейронных элементов, соединенных между собой дву направленными симметричными связями (рисунок 1.6). Входной слой нейронов называется видимым (слой X), а выходной – скрытым (слой Y). Ограниченнная машина Больцмана может генерировать любое дискретное распределение, если используется достаточное число нейронов скрытого слоя [25, с. 57].

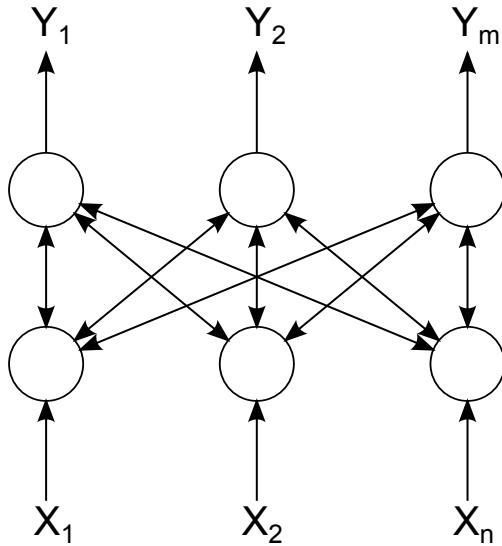


Рисунок 1.6 – Ограниченнная машина Больцмана

В данной модели состояния видимых и скрытых нейронов меняются в соответствии с вероятностной версией сигмоидной функции активации:

$$p(y_j|x) = \frac{1}{1 + e^{-S_j}}, \quad S_j = \sum_i^n w_{ij}x_i + T_j, \quad (1.6)$$

$$p(x_i|y) = \frac{1}{1 + e^{-S_i}}, \quad S_i = \sum_j^m w_{ij}y_j + T_i, \quad (1.7)$$

где S_i, S_j – взвешенные суммы видимого и скрытого слоев,

T_i, T_j – пороговые элементы видимого и скрытого слоев соответственно.

Состояния видимых и скрытых нейронных элементов принимаются независимыми:

$$P(x|y) = \prod_{i=1}^n P(x_i|y),$$

$$P(y|x) = \prod_{j=1}^m P(y_j|x).$$

Таким образом, состояния всех нейронов ограниченной машины Больцмана определяются через распределение вероятностей. В RBM нейроны скрытого слоя являются детекторами признаков, которые сохраняют закономерности входных данных. Основная задача обучения состоит в воспроизведении распределения входных данных на основе состояний нейронов скрытого слоя как можно точнее. Для достижения этого применяется процедура CD- k . В главе 2 будет подробно описана данная процедура и приведен вывод классических правил обучения RBM.

При предобучении ГНС на базе RBM (в процессе первого этапа обучения) вначале создается RBM-модель из первого скрытого слоя ГНС. Для данной RBM входные данные поступают на видимый слой нейронных элементов и используя CD- k процедуру вычисляются состояния скрытых $P(y|x)$ и видимых нейронов $P(x|y)$. В процессе выполнения данной процедуры в течение заданного количества эпох изменяются весовые коэффициенты и пороговые значения сети RBM, которые затем фиксируются. Следующим берется второй скрытый слой глубокой нейронной сети и конструируется новая RBM. Входными данными для нее являются данные с предыдущего слоя (выходные данные, полученные первой RBM). Выполняется обучение новой модели и процесс продолжается для всех последующих слоев нейронной сети, за исключением, возможно, самого последнего (включение последнего слоя в процесс предобучения зависит от решаемой задачи – например, для задачи классификации предобучение последнего слоя не требуется, так как он должен обучаться с учителем), как показано на рисунке 1.7.

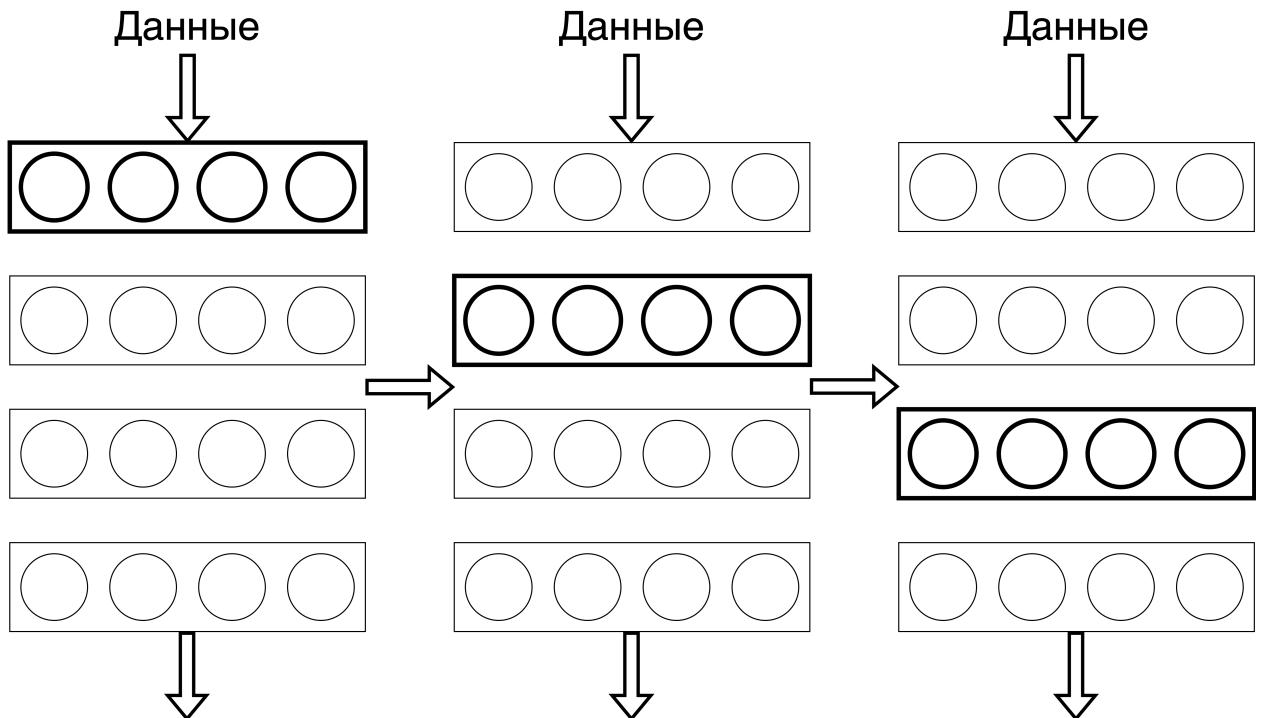


Рисунок 1.7 – «Жадный» алгоритм послойного обучения

При предобучении ГНС в соответствии со **вторым подходом** [50, с. 158], вначале обучается первый слой ГНС как автоассоциативная нейронная сеть с целью минимизации суммарной ошибки реконструкции данных, затем обучается второй слой ГНС и так далее. Для обучения каждого слоя можно использовать алгоритм обратного распространения ошибки.

Рассмотрим персепtron с тремя скрытыми слоями (рисунок 1.8). То-

гда в соответствии с автоэнкодерным методом, прежде всего, берутся первые два слоя нейронной сети (1 и 2) и на базе их конструируется автоассоциативная (рециркуляционная) нейронная сеть (1-2-1), то есть добавляется восстановливающий слой (рисунок 1.9).

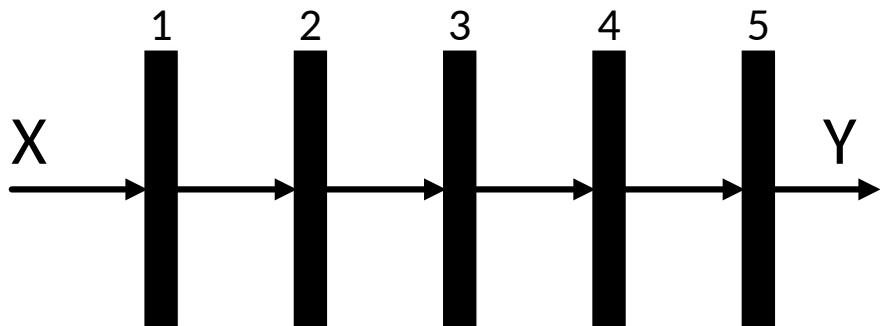


Рисунок 1.8 – Персепtron с тремя скрытыми слоями

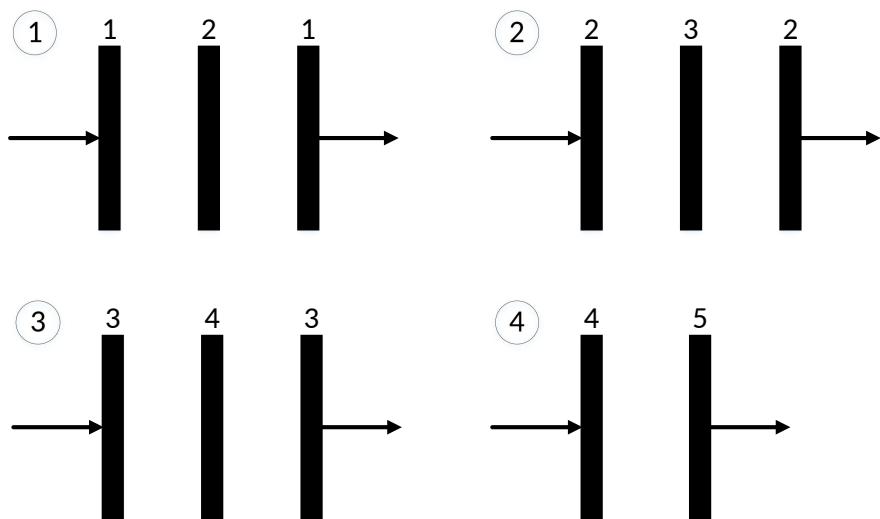


Рисунок 1.9 – Автоэнкодерный метод обучения

Затем, происходит обучение такой сети, например, при помощи алгоритма обратного распространения ошибки с целью минимизации ошибки реконструкции данных.

После этого отбрасывается восстановливающий слой (последний слой) автоассоциативной сети, фиксируются веса скрытого слоя, и конструируется автоассоциативная сеть из следующих двух слоев нейронной сети (2-3-2), которая обучается на основе данных, поступающих с предыдущего (2-го слоя). Процесс продолжается до последнего или предпоследнего слоя, как это схематично изображено на рисунке 1.9.

Данный процесс можно представить в виде следующего алгоритма:

- 1) конструируется автоассоциативная сеть с входным слоем X , скрытым Y и выходным слоем X ;

2) обучается автоассоциативная сеть, например при помощи алгоритма обратного распространения ошибки и фиксируются синаптические связи первого слоя W_1 ;

3) берется следующий слой глубокой нейронной сети и снова формируется автоассоциативная сеть аналогичным образом;

4) используя настроенные синаптические связи предыдущего слоя W_1 , подаются входные данные на вторую автоассоциативную сеть и она обучается аналогичным образом. В результате получаются весовые коэффициенты второго слоя W_2 ;

5) процесс продолжается до последнего или предпоследнего слоя нейронной сети.

Для обоих подходов (RBM и автоэнкодерный), с помощью реализации такого неконтролируемого предобучения можно получить приемлемую начальную инициализацию настраиваемых параметров глубокой нейронной сети.

На **втором этапе**, иначе называемом этапом «тонкой настройки» нейронной сети или fine-tuning, выполняется дообучение всей глубокой модели с использованием, например, алгоритма «бодрствования и сна» или обратного распространения ошибки.

Критически важным для этого этапа является начальная инициализация весов и порогов. Если такая инициализация выполнена в ходе предобучения, то сеть начинает процесс «тонкой настройки» с «хороших» начальных значений, что обуславливает быструю сходимость.

Если же предобучение отсутствует и используется традиционная «поверхностная» схема обучения со случайной инициализацией параметров, обучение сети может быстро остановиться и, как результат, приемлемая обобщающая способность не будет достигнута.

В дальнейшем под предобучением будем понимать предобучение II типа, проводимое с использованием RBM-сетей, если не указано иного.

1.4 Критерии минимизации при обучении

При обучении ГНС (например, на этапе «тонкой настройки», если сеть предобучалась) важно определить критерий минимизации. Наиболее известными и используемыми в настоящий момент функциями ошибок являются функция квадратичной ошибки (SE)

$$E_{SE} = \frac{1}{2} \sum_{j=1}^n (y_j - t_j)^2 \quad (1.8)$$

и кросс-энтропийная функция (CE) – случай мультиклассовой классификации

$$E_{CE_{mult}} = - \sum_{j=1}^n t_j \log(y_j), \quad (1.9)$$

где n – количество нейронных элементов в последнем слое,
 y_j – выход нейрона j ,
 t_j – эталонное значение этого нейрона,
множитель $\frac{1}{2}$ введен для удобства вычислений.

Формула 1.8 справедлива для одного обучающего примера. Для L обучающих примеров она запишется в следующем виде (MSE – среднеквадратичная ошибка):

$$E_{MSE} = \frac{1}{2L} \sum_{k=1}^L \sum_{j=1}^n (y_j^k - t_j^k)^2. \quad (1.10)$$

При обучении автоэнкодеров часто применяется кросс-энтропийная функция, обобщающая двухклассовый случай [51, с. 115]:

$$E_{CE} = - \sum_{j=1}^n (t_j \log(y_j) + (1 - t_j) \log(1 - y_j)). \quad (1.11)$$

Согласно методу обратного распространения ошибки, для настройки весовых коэффициентов и порогов применяются следующие правила, получаемые в соответствии с методом градиентного спуска:

$$w_{i-1,i} = w_{i-1,i} - \alpha \frac{\partial E}{\partial w_{i-1,i}}, \quad (1.12)$$

$$T_i = T_i - \alpha \frac{\partial E}{\partial T_i}, \quad (1.13)$$

где $w_{i-1,i}$ – весовые коэффициенты для i -го обрабатывающего слоя нейронной сети, i изменяется от 1 до N ,

N – количество обрабатывающих слоев нейронной сети,

α – скорость обучения, а соответствующие градиенты находятся по формулам:

$$\frac{\partial E}{\partial w_{i-1,i}} = \frac{\partial E}{\partial S_i} y_{i-1}, \quad (1.14)$$

$$\frac{\partial E}{\partial T_i} = \frac{\partial E}{\partial S_i}, \quad (1.15)$$

где $\frac{\partial E}{\partial S_i}$ – частная производная по взвешенной сумме для i -го слоя НС, y_{i-1} – выход $i - 1$ -го слоя НС.

Частные производные по взвешенным суммам S_i получаются в соответствии со следующими выражениями:

$$\frac{\partial E}{\partial S_i} = (y_i - t_i)F'(S_i), \quad (1.16)$$

$$\frac{\partial E}{\partial S_{i-1}} = \left(\sum_i \frac{\partial E}{\partial S_i} w_{i-1,i} \right) F'(S_{i-1}), \quad (1.17)$$

где F – функция активации нейронов.

Формула 1.16 используется для последнего слоя НС, а формула 1.17 – для других слоев, кроме последнего.

Формула 1.16 актуальна для случая, если используется целевая функция E_{MSE} . Для случая E_{CE} она будет иметь следующий вид:

$$\frac{\partial E}{\partial S_i} = \frac{y_i - t_i}{y_i(1 - y_i)} F'(S_i). \quad (1.18)$$

В случае использования логистической функции, формула 1.18 упрощается, поскольку в этом случае $F'(S_i) = y_i(1 - y_i)$:

$$\frac{\partial E}{\partial S_i} = y_i - t_i. \quad (1.19)$$

Формулы 1.14, 1.15 и 1.17 справедливы как для случая E_{MSE} , так и для E_{CE} .

Для $E_{CE_{mult}}$ эти формулы справедливы для функции активации **softmax** на последнем слое нейронной сети. Покажем это.

Имеем

$$\frac{\partial E_{CE_{mult}}}{\partial S_i} = \sum_{k=1}^n \frac{\partial E_{CE_{mult}}}{\partial y_k} \frac{\partial y_k}{\partial S_i} = \frac{\partial E_{CE_{mult}}}{\partial y_i} \frac{\partial y_i}{\partial S_i} + \sum_{k \neq i} \frac{\partial E_{CE_{mult}}}{\partial y_k} \frac{\partial y_k}{\partial S_i}. \quad (1.20)$$

Очевидно,

$$\frac{\partial E_{CE_{mult}}}{\partial y_i} = -\frac{t_i}{y_i}. \quad (1.21)$$

Учитывая, что

$$y_i = \frac{e^{S_i}}{\sum_{k=1}^n e^{S_k}}, \quad (1.22)$$

где i – индекс последнего слоя сети,

S_i – соответствующая взвешенная сумма, найдем значения частной производной $\frac{\partial y_i}{\partial S_k}$.

1) Случай $i = k$:

$$\begin{aligned}\frac{\partial y_i}{\partial S_i} &= \frac{e^{S_i} \sum e^{S_k} - e^{2S_i}}{(\sum e^{S_k})^2} = \\ &= \frac{e^{S_i} \sum e^{S_k}}{(\sum e^{S_k})^2} - \frac{e^{2S_i}}{(\sum e^{S_k})^2} = y_i - y_i^2 = y_i(1 - y_i);\end{aligned}\quad (1.23)$$

2) $i \neq k$:

$$\frac{\partial y_k}{\partial S_i} = -e^{S_k} \left(\sum e^{S_i} \right)^{-2} e^{S_i} = -y_k y_i. \quad (1.24)$$

Подставляя полученные выражения 1.23, 1.24 и 1.21 в 1.20, получим

$$\frac{\partial E_{CE_{mult}}}{\partial S_i} = -t_i(1 - y_i) + \sum_{k \neq i} t_k y_i = -t_i + y_i \sum_{k=1}^n t_k = y_i - t_i. \quad (1.25)$$

1.5 Выводы

1) Сформулирована проблема обучения глубоких нейронных сетей и обоснована ее актуальность. Приведены предпосылки развития методов обучения глубоких нейронных сетей [1–А, с. 1; 3–А, с. 1].

2) Рассмотрены существующие методы обучения глубоких нейронных сетей [2–А, с. 129; 3–А, с. 8; 18–А].

3) Описаны основные этапы обучения ГНС с использованием неконтролируемого предобучения, включающие в себя предобучение ГНС и «тонкую настройку» модели [1–А, с. 8; 3–А, с. 7-8; 27–А; 28–А].

4) Описаны основные подходы к осуществлению неконтролируемого предобучения ГНС, базирующиеся на использовании ограниченных машин Больцмана и автоассоциативных нейронных сетей [1–А, с. 6; 3–А, с. 8-10].

ГЛАВА 2

МЕТОД НЕКОНТРОЛИРУЕМОГО ПРЕДОБУЧЕНИЯ

В данной главе в контексте проблемы обучения ГНС приводится вывод классических правил обучения ограниченной машины Больцмана (restricted Boltzmann Machine), применяемой на этапе неконтролируемого предобучения ГНС.

Предложен альтернативный подход к обучению ограниченной машины Больцмана [2–A, с. 131–137; 3–A, с. 11–14; 14–A, с. 140–143], базирующийся на минимизации ошибок восстановления образов на его видимом и скрытом слоях, используя итерации сэмплирования Гиббса. В качестве критериев минимизации используются MSE (mean squared error – среднеквадратичная ошибка) и CE (cross-entropy loss – кросс-энтропийная функция потерь).

Для предлагаемого подхода рассмотрен вывод правил обучения RBM. Показано, что полученные правила обобщают классические правила обучения, полученные в работах Дж. Хинтона.

Для основных случаев предлагаемого подхода приводятся правила обучения CRBM, которые могут быть использованы для предобучения глубоких сверточных нейронных сетей. Также показано, что сверточный слой может быть представлен в виде полносвязного слоя.

Предлагается подход к редуцированию размерности нейросетевых моделей за счет уменьшения числа настраиваемых параметров после выполнения процедуры предварительного обучения с использованием сетей RBM.

2.1 Обучение RBM

В главе 1 было описано, что одним из методов обучения ГНС является метод предобучения, который основывается на использовании ограниченных машин Больцмана.

Также, ранее было упомянуто, что основная задача обучения ограниченной машины Больцмана состоит в воспроизведении распределения входных данных на основе состояний нейронов скрытого слоя как можно точнее. Это эквивалентно максимизации функции правдоподобия путем модификации синаптических связей нейронной сети. Покажем основные

этапы вывода классических правил обучения для RBM.

Вероятность нахождения видимого и скрытого нейрона в состоянии (x, y) определяется на основе распределения Гиббса [52, с. 33]:

$$P(x, y) = \frac{e^{-E(x,y)}}{Z},$$

где $E(x, y)$ – энергия системы в состоянии (x, y) ,

Z – параметр, который определяет условие нормализации вероятностей, то

есть, условие, при котором сумма вероятностей $P(x, y)$ равняется единице.

Данный параметр определяется следующим образом:

$$Z = \sum_{(x,y)} e^{-E(x,y)}.$$

Вероятность нахождения видимых нейронов в определенном состоянии равняется сумме вероятностей конфигураций $P(x, y)$ по состояниям скрытых нейронов:

$$P(x) = \sum_y P(x, y) = \sum_y \frac{e^{-E(x,y)}}{Z} = \frac{\sum_y e^{-E(x,y)}}{\sum_{(x,y)} e^{-E(x,y)}}.$$

Для нахождения правил модификации параметров модели необходимо максимизировать вероятность воспроизведения состояний видимых нейронов $P(x)$ ограниченной машиной Больцмана. Для того, чтобы определить максимум функции правдоподобия распределения данных $P(x)$ будем использовать метод градиентного подъема в параметрическом пространстве сети, где в качестве целевой функции используем функцию логарифмического правдоподобия:

$$\ln P(x) = \ln \sum_y e^{-E(x,y)} - \ln \sum_{(x,y)} e^{-E(x,y)}.$$

Тогда градиент равен

$$\frac{\partial \ln P(x)}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \ln \sum_y e^{-E(x,y)} - \frac{\partial}{\partial w_{ij}} \ln \sum_{(x,y)} e^{-E(x,y)}.$$

Преобразуя последнее выражение, получим

$$\begin{aligned}\frac{\partial \ln P(x)}{\partial w_{ij}} = & -\frac{1}{\sum_y e^{-E(x,y)}} \sum_y e^{-E(x,y)} \frac{\partial E(x,y)}{\partial w_{ij}} + \\ & + \frac{1}{\sum_{(x,y)} e^{-E(x,y)}} \sum_{(x,y)} e^{-E(x,y)} \frac{\partial E(x,y)}{\partial w_{ij}}.\end{aligned}$$

Так как

$$P(x, y) = P(y|x)P(x),$$

то

$$P(y|x) = \frac{P(x, y)}{P(x)} = \frac{1/Z e^{-E(x,y)}}{1/Z \sum_y e^{-E(x,y)}} = \frac{e^{-E(x,y)}}{\sum_y e^{-E(x,y)}}.$$

В результате можно получить следующее выражение:

$$\frac{\partial \ln P(x)}{\partial w_{ij}} = - \sum_y P(y|x) \frac{\partial E(x,y)}{\partial w_{ij}} + \sum_{x,y} P(x, y) \frac{\partial E(x,y)}{\partial w_{ij}}. \quad (2.1)$$

В данном выражении первое слагаемое определяет позитивную фазу работы машины Больцмана, когда сеть работает на основе образов из обучающей выборки. Второе слагаемое характеризует негативную фазу функционирования, когда сеть работает в свободном режиме независимо от окружающей среды.

Если рассматривать энергетическую функцию сети РВМ, то в данном случае задача обучения состоит в том, чтобы на основе входных данных найти конфигурацию нейронов скрытого слоя с минимальной энергией.

В результате на обучающем множестве сеть будет иметь меньшую энергию по сравнению с другими состояниями. Функция энергии бинарного состояния (x, y) определяется аналогично энергетической функции сети Хопфилда:

$$E(x, y) = - \sum_i x_i T_i - \sum_j y_j T_j - \sum_{i,j} x_i y_j w_{ij}. \quad (2.2)$$

В этом случае

$$\frac{\partial E(x, y)}{\partial w_{ij}} = -x_i y_j.$$

Подставляя полученное значение в формулу 2.1, получим

$$\frac{\partial \ln P(x)}{\partial w_{ij}} = \sum_y P(y|x)x_iy_j - \sum_{x,y} P(x,y)x_iy_j.$$

Так как математическое ожидание входных данных равняется:

$$E(x) = \sum_i x_i P_i, \quad (2.3)$$

то

$$\frac{\partial \ln P(x)}{\partial w_{ij}} = E[x_iy_j]_{\text{data}} - E[x_iy_j]_{\text{model}}. \quad (2.4)$$

Рассуждая аналогичным образом, находим

$$\frac{\partial \ln P(x)}{\partial T_i} = - \sum_y P(y|x) \frac{\partial E(x,y)}{\partial T_i} + \sum_{x,y} P(x,y) \frac{\partial E(x,y)}{\partial T_i}$$

и

$$\frac{\partial \ln P(x)}{\partial T_j} = - \sum_y P(y|x) \frac{\partial E(x,y)}{\partial T_j} + \sum_{x,y} P(x,y) \frac{\partial E(x,y)}{\partial T_j}.$$

Так как

$$\frac{\partial E(x,y)}{\partial T_i} = -x_i$$

и

$$\frac{\partial E(x,y)}{\partial T_j} = -y_j,$$

получим градиенты для пороговых значений, учитывая формулу 2.3:

$$\begin{aligned} \frac{\partial \ln P(x)}{\partial T_i} &= E[x_i]_{\text{data}} - E[x_i]_{\text{model}}; \\ \frac{\partial \ln P(x)}{\partial T_j} &= E[y_j]_{\text{data}} - E[y_j]_{\text{model}}. \end{aligned} \quad (2.5)$$

Как следует из последних выражений, первое слагаемое характеризует работу сети на основе данных из обучающей выборки, а второе слагаемое характеризует работу сети на основе данных модели (данные генерируемые сетью), то есть в свободном режиме независимо от окружающей среды. Так как вычисление математического ожидания в формулах 2.4 и 2.5 является

сложной задачей, Хинтон предложил использовать аппроксимацию данных слагаемых, получаемую применением процедуры, которую он назвал контрастным расхождением (Contrastive Divergence (CD)) [37, с. 1533].

Такая аппроксимация основывается на дискретизаторе Гиббса (Gibbs sampling). В этом случае первые слагаемые в выражениях 2.4 и 2.5 характеризуют распределение данных в момент времени $t = 0$, а вторые слагаемые характеризуют реконструированные или генерируемые моделью данные в момент времени $t = k$. Исходя из этого, CD- k процедура может быть представлена следующим образом:

$$x(0) \rightarrow y(0) \rightarrow x(1) \rightarrow y(1) \rightarrow \dots \rightarrow x(k) \rightarrow y(k). \quad (2.6)$$

В результате можно сформулировать следующие правила для обучения сети RBM.

В случае применения CD-1 (случай $k = 1$) и учитывая, что в соответствии с методом градиентного подъема

$$w_{ij}(t+1) = w_{ij}(t) + \alpha \frac{\partial \ln P(x)}{\partial w_{ij}(t)},$$

можно получить, что

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) + \alpha(x_i(0)y_j(0) - x_i(1)y_j(1)), \\ T_i(t+1) &= T_i(t) + \alpha(x_i(0) - x_i(1)), \\ T_j(t+1) &= T_j(t) + \alpha(y_j(0) - y_j(1)). \end{aligned} \quad (2.7)$$

Аналогичным образом, для алгоритма CD- k

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) + \alpha(x_i(0)y_j(0) - x_i(k)y_j(k)), \\ T_i(t+1) &= T_i(t) + \alpha(x_i(0) - x_i(k)), \\ T_j(t+1) &= T_j(t) + \alpha(y_j(0) - y_j(k)). \end{aligned} \quad (2.8)$$

Согласно последним выражениям (2.7 и 2.8) легко увидеть, что в процессе обучения RBM минимизируется разница между исходными (оригинальными) данными и данными, генерируемыми моделью.

Схожим образом можно получить правила обучения для группового случая (например, при использовании mini-batch оптимизации, слу-

чай CD-1):

$$w_{ij}(t+1) = w_{ij}(t) + \frac{\alpha}{L} \sum_{l=1}^L (x_i^l(0)y_j^l(0) - x_i^l(1)y_j^l(1)),$$

$$T_i(t+1) = T_i(t) + \frac{\alpha}{L} \sum_{l=1}^L (x_i^l(0) - x_i^l(1)),$$

$$T_j(t+1) = T_j(t) + \frac{\alpha}{L} \sum_{l=1}^L (y_j^l(0) - y_j^l(1)),$$

где L – количество образов в группе.

Суммируя вышесказанное, приведем полный алгоритм обучения RBM для случая CD-1 (алгоритм 1).

Алгоритм: Процедура обучения RBM для случая бинарных данных

Исходные данные: $\mathbf{x}(0)$ – вектор-образ из обучающей выборки,
 α – скорость обучения

Результат: матрица весовых коэффициентов \mathbf{W} , вектор порогов
 нейронов видимого слоя \mathbf{V} , вектор порогов нейронов
 скрытого слоя \mathbf{H}

foreach нейрона скрытого слоя j **do**

Вычислить $P(y_j(0) = 1|x_i(0))$ (для биномиальных нейронов
 равняется $\text{sigmoid}(\sum_i w_{ij}x_i(0) + H_j))$;
 Генерировать $y_j(0) \in \{0, 1\}$ из $P(y_j(0)|x_i(0))$;

foreach нейрона видимого слоя i **do**

Вычислить $P(x_i(1) = 1|y_j(0))$ (для биномиальных нейронов
 равняется $\text{sigmoid}(\sum_j w_{ij}y_j(0) + V_i))$;
 Генерировать $x_i(1) \in \{0, 1\}$ из $P(x_i(1)|y_j(0))$;

foreach скрытых нейронов j **do**

Вычислить $P(y_j(1) = 1|x_i(1))$ (для биномиальных нейронов
 равняется $\text{sigmoid}(\sum_i w_{ij}x_i(1) + H_j))$;

$\mathbf{W} \leftarrow \mathbf{W} + \alpha(\mathbf{x}(0)\mathbf{y}(0)^T - \mathbf{x}(1)\mathbf{P}(\mathbf{y}(1) = 1|\mathbf{x}(1))^T);$

$\mathbf{V} \leftarrow \mathbf{V} + \alpha(\mathbf{x}(0) - \mathbf{x}(1));$

$\mathbf{H} \leftarrow \mathbf{H} + \alpha(\mathbf{y}(0) - \mathbf{P}(\mathbf{y}(1) = 1|\mathbf{x}(1)));$

2.2 Общее описание предлагаемого метода

По сравнению с методом обучения RBM, предложенным Джейфри Хинтоном [53, с. 3], основывающимся на применении энергетической функции (energy-based method) и оперирующим линейным представлением нейронных элементов, предлагаемый метод позволяет учитывать их нелинейную природу.

Рассмотрим ограниченную машину Больцмана, которую будем представлять в виде трех слоев нейронных элементов [2–A, с. 131]: видимого, скрытого и видимого (рисунок 2.1). В приведенном представлении матрица весовых коэффициентов W_{ji} для второго слоя получается транспонированием матрицы W_{ij} первого слоя. Фактически, при выполнении обучения, производится однократное изменение ее элементов, поскольку это одна и та же матрица.

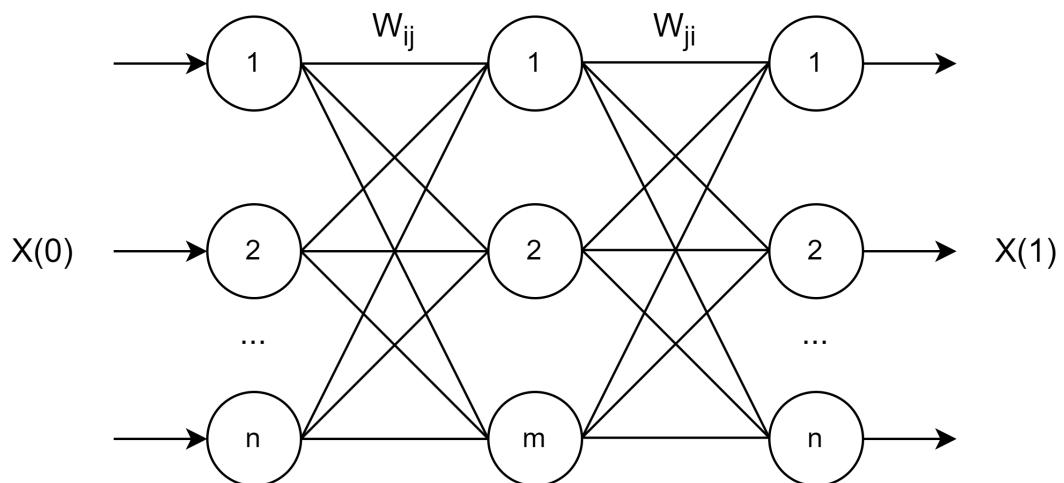


Рисунок 2.1 – Развернутое представление RBM

Представим сэмплирование Гиббса, используя развернутое представление RBM (рисунок 2.2).

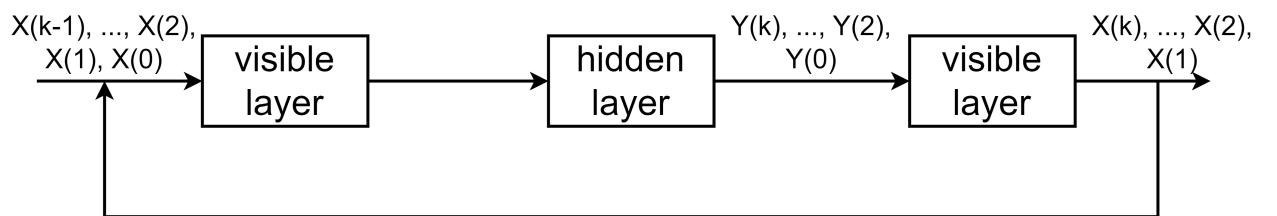


Рисунок 2.2 – Сэмплирование Гиббса

Сэмплирование Гиббса заключается в следующей процедуре. Пусть $x(0)$ – входной вектор, который поступает на видимый слой в момент времени $t = 0$. Тогда выходные значения нейронов скрытого слоя будут опре-

деляться следующим образом:

$$y_j(0) = F(S_j(0)), \quad (2.9)$$

$$S_j(0) = \sum_i w_{ij}x_i(0) + T_j. \quad (2.10)$$

Инверсный (последний) слой реконструирует входной вектор на основе данных со скрытого слоя и текущего значения настраиваемых параметров (порогов видимого слоя и матрицы весов). В результате получается восстановленный вектор $x(1)$ в момент времени $t = 1$:

$$x_i(1) = F(S_i(1)), \quad (2.11)$$

$$S_i(1) = \sum_j w_{ij}y_j(0) + T_i. \quad (2.12)$$

Затем вектор $x(1)$ поступает на видимый слой, и вычисляются выходные значения нейронов скрытого слоя:

$$y_j(1) = F(S_j(1)), \quad (2.13)$$

$$S_j(1) = \sum_i w_{ij}x_i(1) + T_j. \quad (2.14)$$

Продолжая данный процесс, можно получить на шаге k следующие выражения:

$$y_j(k) = F(S_j(k)), \quad S_j(k) = \sum_i w_{ij}x_i(k) + T_j,$$

$$x_i(k) = F(S_i(k)), \quad S_i(k) = \sum_j w_{ij}y_j(k-1) + T_i.$$

2.3 Вывод правил предобучения

Хинтоном была предложена энергетическая модель, базирующаяся на идее максимизации функции правдоподобия распределения входных данных $P(x)$. Вывод классических правил обучения был приведен в разделе 2.1. Основываясь на идее использования ограниченной машины Больцмана в качестве вспомогательной модели для проведения предобучения,

было предложено использование двух разных критериев для ее обучения [2–A, с. 132-137]. Первый критерий основывается на минимизации среднеквадратичной ошибки (MSE), а второй – на минимизации кросс-энтропийной функции ошибки (CE).

Покажем, что применение различных критериев минимизации позволяет, тем не менее, получить одинаковые правила обучения.

2.3.1 Критерий MSE

В случае использования в качестве критерия обучения MSE основной целью обучения ограниченной машины Больцмана является минимизация суммарной среднеквадратичной ошибки реконструкции данных на скрытом и видимом (восстанавливающем) слое, которая в случае CD-k определяется следующим образом:

$$E_s(k) = \frac{1}{2L} \left(\sum_{l=1}^L \sum_{j=1}^m \sum_{p=1}^k (y_j^l(p) - y_j^l(p-1))^2 + \sum_{l=1}^L \sum_{i=1}^n \sum_{p=1}^k (x_i^l(p) - x_i^l(p-1))^2 \right),$$

где k определяет параметр процедуры сэмплирования Гиббса,
 n – количество нейронов в видимом слое,
 m – количество нейронов в скрытом слое,
 L – размерность обучающей выборки.

В случае CD-1 суммарная среднеквадратичная ошибка:

$$E_s(1) = \frac{1}{2L} \left(\sum_{l=1}^L \sum_{j=1}^m (y_j^l(1) - y_j^l(0))^2 + \sum_{l=1}^L \sum_{i=1}^n (x_i^l(1) - x_i^l(0))^2 \right). \quad (2.15)$$

Как следует из приведенных выше выражений, ошибка состоит из двух частей: ошибки восстановления информации на видимом и скрытом слоях, то есть может быть представлена в следующем виде:

$$E_s(k) = E_h(k) + E_v(k), \quad (2.16)$$

где

$$E_h(k) = \frac{1}{2L} \sum_{l=1}^L \sum_{j=1}^m \sum_{p=1}^k (y_j^l(p) - y_j^l(p-1))^2, \quad (2.17)$$

$$E_v(k) = \frac{1}{2L} \sum_{l=1}^L \sum_{i=1}^n \sum_{p=1}^k (x_i^l(p) - x_i^l(p-1))^2. \quad (2.18)$$

Найдем правила обучения, соответствующие критерию 2.16 и докажем их эквивалентность классическим правилам обучения RBM при выполнении некоторых специальных условий.

Теорема 1. Максимизация функции правдоподобия распределения данных $P(x)$ в параметрическом пространстве ограниченной машины Больцмана эквивалентна минимизации суммарной квадратичной ошибки сети в том же пространстве при использовании линейных нейронов.

Доказательство: Рассмотрим последовательное обучение RBM, когда модификация синаптических связей происходит после подачи каждого входного образа на сеть (онлайн-обучение). В соответствии с методом градиентного спуска для минимизации суммарной квадратичной ошибки сети, синаптические связи должны изменяться следующим образом:

$$w_{ij}(t+1) = w_{ij}(t) - \alpha \frac{\partial E}{\partial w_{ij}(t)}, \quad (2.19)$$

$$T_i(t+1) = T_i(t) - \alpha \frac{\partial E}{\partial T_i(t)}, \quad (2.20)$$

$$T_j(t+1) = T_j(t) - \alpha \frac{\partial E}{\partial T_j(t)}. \quad (2.21)$$

В случае CD-k квадратичная ошибка E для одного образа:

$$E = \frac{1}{2} \sum_{j=1}^m \sum_{p=1}^k (y_j(p) - y_j(p-1))^2 + \frac{1}{2} \sum_{i=1}^n \sum_{p=1}^k (x_i(p) - x_i(p-1))^2.$$

Тогда

$$\begin{aligned} \frac{\partial E}{\partial w_{ij}} &= \frac{\partial E}{\partial y_j(p)} \frac{\partial y_j(p)}{\partial S_j(p)} \frac{\partial S_j(p)}{\partial w_{ij}} + \frac{\partial E}{\partial x_i(p)} \frac{\partial x_i(p)}{\partial S_i(p)} \frac{\partial S_i(p)}{\partial w_{ij}} = \\ &= \sum_{p=1}^k (y_j(p) - y_j(p-1)) x_i(p) F'(S_j(p)) + \\ &\quad + \sum_{p=1}^k (x_i(p) - x_i(p-1)) y_j(p-1) F'(S_i(p)). \end{aligned}$$

Если ограниченная машина Больцмана использует линейные нейроны с линейной функцией активации, то

$$\frac{\partial S_i(p)}{\partial w_{ij}} = F'(S_i(p)) = \frac{\partial S_j(p)}{\partial w_{ij}} = F'(S_j(p)) = 1.$$

Тогда

$$\frac{\partial E}{\partial w_{ij}} = \sum_{p=1}^k (y_j(p)x_i(p) - y_j(p-1)x_i(p-1)) = y_j(k)x_i(k) - y_j(0)x_i(0).$$

В результате можно получить CD-k правило обучения RBM:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(x_i(0)y_j(0) - x_i(k)y_j(k)).$$

Аналогичным образом для пороговых значений:

$$\begin{aligned} T_j(t+1) &= T_j(t) + \alpha(y_j(0) - y_j(k)), \\ T_i(t+1) &= T_i(t) + \alpha(x_i(0) - x_i(k)). \end{aligned}$$

Как видно последние выражения совпадают с классическим правилом обучения ограниченной машины Больцмана для CD-k. Отсюда следует, что для линейной RBM максимизация функции правдоподобия распределения данных $P(x)$ эквивалентна минимизации суммарной квадратичной ошибки сети. Теорема доказана.

Следствие 1.1. Линейная ограниченная машина Больцмана с точки зрения обучения эквивалентна автоассоциативной нейронной сети при использовании в ней при обучении сэмплирования Гиббса.

Следствие 1.2. Для нелинейной ограниченной машины Больцмана правило модификации синаптических связей в случае CD-k будет следующим:

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \\ &- \alpha \left(\sum_{p=1}^k (y_j(p) - y_j(p-1))x_i(p)F'(S_j(p)) + (x_i(p) - x_i(p-1))y_j(p-1)F'(S_i(p)) \right), \\ T_i(t+1) &= T_i(t) - \alpha \left(\sum_{p=1}^k (x_i(p) - x_i(p-1))F'(S_i(p)) \right), \\ T_j(t+1) &= T_j(t) - \alpha \left(\sum_{p=1}^k (y_j(p) - y_j(p-1))F'(S_j(p)) \right). \end{aligned}$$

Следствие 1.3. Для нелинейной ограниченной машины Больцмана правило модификации синаптических связей в случае CD-1 будет следующим:

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \\ &- \alpha((y_j(1) - y_j(0))F'(S_j(1))x_i(1) + (x_i(1) - x_i(0))F'(S_i(1))y_j(0)), \end{aligned}$$

$$T_i(t+1) = T_i(t) - \alpha(x_i(1) - x_i(0))F'(S_i(1)),$$

$$T_j(t+1) = T_j(t) - \alpha(y_j(1) - y_j(0))F'(S_j(1)).$$

При использовании группового обучения (batch learning), метод градиентного спуска примет следующий вид:

$$w_{ij}(t+1) = w_{ij}(t) - \alpha \frac{\partial E_s}{\partial w_{ij}(t)}, \quad (2.22)$$

$$T_i(t+1) = T_i(t) - \alpha \frac{\partial E_s}{\partial T_i(t)}, \quad (2.23)$$

$$T_j(t+1) = T_j(t) - \alpha \frac{\partial E_s}{\partial T_j(t)}. \quad (2.24)$$

Теорема 2. При использовании CD-k для нелинейной ограниченной машины Больцмана в случае группового обучения правило модификации синаптических связей определяется на основе следующих выражений:

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \\ &- \frac{\alpha}{L} \left(\sum_{l=1}^L \sum_{p=1}^k \Delta_j^l(p)x_i^l(p)F'(S_j^l(p)) + \Delta_i^l(p)y_j^l(p-1)F'(S_i^l(p)) \right), \\ T_i(t+1) &= T_i(t) - \frac{\alpha}{L} \left(\sum_{l=1}^L \sum_{p=1}^k \Delta_i^l(p)F'(S_i^l(p)) \right), \\ T_j(t+1) &= T_j(t) - \frac{\alpha}{L} \left(\sum_{l=1}^L \sum_{p=1}^k \Delta_j^l(p)F'(S_j^l(p)) \right), \end{aligned}$$

где $\Delta_j^l(p) = y_j^l(p) - y_j^l(p-1)$,
 $\Delta_i^l(p) = x_i^l(p) - x_i^l(p-1)$.

Процесс доказательства данной теоремы является аналогичным доказательству теоремы 1.

Следствие 2.1. При использовании CD-1 для нелинейной ограниченной машины Больцмана в случае группового обучения правило модификации синаптических связей определяется на основе следующих выражений:

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \\ &- \frac{\alpha}{L} \left(\sum_{l=1}^L (y_j^l(1) - y_j^l(0))x_i^l(1)F'(S_j^l(1)) + (x_i^l(1) - x_i^l(0))y_j^l(0)F'(S_i^l(1)) \right), \end{aligned}$$

$$T_i(t+1) = T_i(t) - \frac{\alpha}{L} \left(\sum_{l=1}^L (x_i^l(1) - x_i^l(0)) F'(S_i^l(1)) \right),$$

$$T_j(t+1) = T_j(t) - \frac{\alpha}{L} \left(\sum_{l=1}^L (y_j^l(1) - y_j^l(0)) F'(S_j^l(1)) \right).$$

Следствие 2.2. При использовании CD-k для линейной ограниченной машины Больцмана в случае группового обучения правило модификации синаптических связей определяется на основе следующих выражений:

$$w_{ij}(t+1) = w_{ij}(t) + \frac{\alpha}{L} \sum_{l=1}^L (x_i^l(0)y_j^l(0) - x_i^l(k)y_j^l(k)),$$

$$T_i(t+1) = T_i(t) + \frac{\alpha}{L} \sum_{l=1}^L (x_i^l(0) - x_i^l(k)),$$

$$T_j(t+1) = T_j(t) + \frac{\alpha}{L} \sum_{l=1}^L (y_j^l(0) - y_j^l(k)).$$

Следствие 2.3. При использовании CD-1 для линейной ограниченной машины Больцмана в случае группового обучения правило модификации синаптических связей определяется на основе следующих выражений:

$$w_{ij}(t+1) = w_{ij}(t) + \frac{\alpha}{L} \sum_{l=1}^L (x_i^l(0)y_j^l(0) - x_i^l(1)y_j^l(1)),$$

$$T_i(t+1) = T_i(t) + \frac{\alpha}{L} \sum_{l=1}^L (x_i^l(0) - x_i^l(1)),$$

$$T_j(t+1) = T_j(t) + \frac{\alpha}{L} \sum_{l=1}^L (y_j^l(0) - y_j^l(1)).$$

Таким образом, получены правила обучения для ограниченной машины Больцмана, которые базируются на минимизации квадратичной ошибки восстановления информации на видимом и скрытом слоях. Предложенный метод позволяет учитывать нелинейную природу нейронных элементов. Показано, что классические выражения для обучения ограниченной машины являются частным случаем предложенного метода. Доказана теорема об эквивалентности максимизации функции правдоподобия распределения входных данных $P(x)$ и минимизации суммарной квадратичной ошибки сети в одном и том же пространстве синаптических связей для линейной ограниченной машины Больцмана.

2.3.2 Критерий СЕ

В случае CD-k кросс-энтропийная функция ошибки для видимого слоя определяется следующим образом:

$$CE_v(k) = -\frac{1}{L} \sum_{l=1}^L \sum_{p=1}^k \sum_{i=1}^n x_i^l(p-1) \log(x_i^l(p)) + (1 - x_i^l(p-1)) \log(1 - x_i^l(p)),$$

где L – размер обучающей выборки,
 k – параметр процедуры сэмплирования,
 n – количество нейронов на видимом слое.

Аналогично для скрытого слоя:

$$CE_h(k) = -\frac{1}{L} \sum_{l=1}^L \sum_{p=1}^k \sum_{j=1}^m y_j^l(p-1) \log(y_j^l(p)) + (1 - y_j^l(p-1)) \log(1 - y_j^l(p)),$$

где m – количество нейронов на скрытом слое.

Общая функция определяется как сумма кросс-энтропийных функций ошибки видимого и скрытого слоев:

$$CE_s(k) = CE_h(k) + CE_v(k). \quad (2.25)$$

Докажем следующую теорему.

Теорема 3. Максимизация функции правдоподобия распределения входных данных $P(x)$ эквивалентна минимизации кросс-энтропийной целевой функции $CE_s(k)$ в одном и том же параметрическом пространстве ограниченной машины Больцмана (случай $k = 1$).

Доказательство. В случае CD-1 кросс-энтропийная функция для одного обучающего примера будет иметь следующий вид:

$$\begin{aligned} CE(1) &= - \sum_{i=1}^n (x_i(0) \log(x_i(1)) + (1 - x_i(0)) \log(1 - x_i(1))) - \\ &\quad - \sum_{j=1}^m (y_j(0) \log(y_j(1)) + (1 - y_j(0)) \log(1 - y_j(1))) = \\ &= CE_v(1) + CE_h(1). \end{aligned} \quad (2.26)$$

Найдем частные производные кросс-энтропийной функции по весовым элементам. Получим

$$\frac{\partial CE(1)}{\partial w_{ij}} = \frac{\partial CE_v(1)}{\partial w_{ij}} + \frac{\partial CE_h(1)}{\partial w_{ij}}.$$

Тогда

$$\begin{aligned}
\frac{\partial CE_v(1)}{\partial w_{ij}} &= -\frac{x_i(0)}{x_i(1)}x_i(1)(1-x_i(1))y_j(0) + \frac{1-x_i(0)}{1-x_i(1)}x_i(1)(1-x_i(1))y_j(0) = \\
&= -x_i(0)(1-x_i(1))y_j(0) + (1-x_i(0))x_i(1)y_j(0) = \\
&= -x_i(0)y_j(0) + x_i(0)x_i(1)y_j(0) + x_i(1)y_j(0) - x_i(0)x_i(1)y_j(0) = \\
&= -x_i(0)y_j(0) + x_i(1)y_j(0)
\end{aligned}$$

и

$$\begin{aligned}
\frac{\partial CE_h(1)}{\partial w_{ij}} &= -y_j(0)(1-y_j(1))x_i(1) + (1-y_j(0))y_j(1)x_i(1) = \\
&= -y_j(0)x_i(0) + y_j(0)y_j(1)x_i(1) + y_j(1)x_i(1) - y_j(0)y_j(1)x_i(1) = \\
&= -y_j(0)x_i(1) + y_j(1)x_i(1).
\end{aligned}$$

Окончательно получим

$$\begin{aligned}
\frac{\partial CE(1)}{\partial w_{ij}} &= \frac{\partial CE_v(1)}{\partial w_{ij}} + \frac{\partial CE_h(1)}{\partial w_{ij}} = \\
&= -x_i(0)y_j(0) + x_i(1)y_j(0) - y_j(0)x_i(1) + y_j(1)x_i(1) = x_i(1)y_j(1) - x_i(0)y_j(0).
\end{aligned}$$

Аналогично, для пороговых элементов имеем:

$$\begin{aligned}
\frac{\partial CE(1)}{\partial T_i} &= \frac{\partial CE_v(1)}{\partial T_i} = -\frac{x_i(0)}{x_i(1)}x_i(1)(1-x_i(1)) + \frac{1-x_i(0)}{1-x_i(1)}x_i(1)(1-x_i(1)) = \\
&= -x_i(0) + x_i(0)x_i(1) + x_i(1) - x_i(0)x_i(1) = x_i(1) - x_i(0);
\end{aligned}$$

$$\begin{aligned}
\frac{\partial CE(1)}{\partial T_j} &= \frac{\partial CE_h(1)}{\partial T_j} = -y_j(0)(1-y_j(1)) + (1-y_j(0))y_j(1) = \\
&= -y_j(0) + y_j(0)y_j(1) + y_j(1) - y_j(0)y_j(1) = y_j(1) - y_j(0).
\end{aligned}$$

Теорема доказана.

Рассматривая более общий случай CD-k, можно доказать следующую теорему.

Теорема 4. Максимизация функции правдоподобия распределения входных данных $P(x)$ эквивалентна минимизации кросс-энтропийной целевой функции $CE_s(k)$ в одном и том же параметрическом пространстве ограниченной машины Больцмана (случай произвольного k).

Доказательство. В случае CD-k кросс-энтропийная функция для одного обучающего примера будет иметь следующий вид:

$$\begin{aligned}
CE_s(k) &= - \sum_{p=1}^k \sum_{i=1}^n (x_i(p-1) \log(x_i(p)) + (1 - x_i(p-1)) \log(1 - x_i(p))) - \\
&- \sum_{p=1}^k \sum_{j=1}^m (y_j(p-1) \log(y_j(p)) + (1 - y_j(p-1)) \log(1 - y_j(p))) = \\
&= CE_v(k) + CE_h(k). \quad (2.27)
\end{aligned}$$

Как и в случае CD-1 находим соответствующие частные производные по весовым и пороговым элементам. Имеем:

$$\frac{\partial CE_s(k)}{\partial w_{ij}} = \frac{\partial CE_v(k)}{\partial w_{ij}} + \frac{\partial CE_h(k)}{\partial w_{ij}}.$$

Причем

$$\begin{aligned}
\frac{\partial CE_v(k)}{\partial w_{ij}} &= \\
&= - \sum_{p=1}^k (x_i(p-1)(1 - x_i(p))y_j(p-1) - (1 - x_i(p-1))x_i(p)y_j(p-1)) = \\
&= - \sum_{p=1}^k (x_i(p-1)y_j(p-1) - x_i(p-1)x_i(p)y_j(p-1) - \\
&\quad - x_i(p)y_j(p-1) + x_i(p-1)x_i(p)y_j(p-1)) = \\
&= \sum_{p=1}^k (x_i(p)y_j(p-1) - x_i(p-1)y_j(p-1))
\end{aligned}$$

И

$$\begin{aligned}
\frac{\partial CE_h(k)}{\partial w_{ij}} &= \\
&= - \sum_{p=1}^k (y_j(p-1)(1 - y_j(p))x_i(p) - (1 - y_j(p-1))y_j(p)x_i(p)) = \\
&= - \sum_{p=1}^k (y_j(p-1)x_i(p) - y_j(p-1)y_j(p)x_i(p) - \\
&\quad - y_j(p)x_i(p) + y_j(p-1)y_j(p)x_i(p)) = \\
&= \sum_{p=1}^k (x_i(p)y_j(p) - x_i(p)y_j(p-1)).
\end{aligned}$$

А значит

$$\begin{aligned}
\frac{\partial CE_s(k)}{\partial w_{ij}} &= \sum_{p=1}^k (y_j(p)x_i(p) - x_i(p-1)y_j(p-1)) = \\
&= y_j(1)x_i(1) - x_i(0)y_j(0) + x_i(2)y_j(2) - x_i(1)y_j(1) + \cdots + \\
&\quad + x_i(k)y_j(k) - x_i(k-1)y_j(k-1) = \\
&= x_i(k)y_j(k) - x_i(0)y_j(0).
\end{aligned}$$

Аналогично для пороговых элементов:

$$\begin{aligned}
\frac{\partial CE_s(k)}{\partial T_i} &= \frac{\partial CE_v(k)}{\partial T_i} = \\
&= - \sum_{p=1}^k (x_i(p-1)(1 - x_i(p)) - (1 - x_i(p-1))x_i(p)) = \\
&= \sum_{p=1}^k (x_i(p) - x_i(p-1)) = x_i(1) - x_i(0) + x_i(2) - x_i(1) + \cdots + x_i(k) - x_i(k-1) = \\
&= x_i(k) - x_i(0);
\end{aligned}$$

$$\begin{aligned}
\frac{\partial CE_s(k)}{\partial T_j} &= \frac{\partial CE_h(k)}{\partial T_j} = \\
&= - \sum_{p=1}^k (y_j(p-1)(1 - y_j(p)) - (1 - y_j(p-1))y_j(p)) = \\
&= \sum_{p=1}^k (y_j(p) - y_j(p-1)) = y_j(1) - y_j(0) + y_j(2) - y_j(1) + \cdots + y_j(k) - y_j(k-1) = \\
&= y_j(k) - y_j(0).
\end{aligned}$$

Теорема доказана.

Из доказанных теорем 3 и 4 следует, что правила обучения ограниченной машины Больцмана могут быть получены более простым путем, чем при использовании традиционного подхода, основанного на применении функции энергии. Произведя минимизацию кросс-энтропийной функции и используя сэмплирование Гиббса, мы получили правила обучения RBM, предложенные Дж. Хинтоном.

Таким образом, на основании доказанных теорем 1-4 можно сформулировать общую теорему.

Теорема 5. Максимизация функции правдоподобия распределения входных данных $P(x)$ эквивалентна минимизации кросс-энтропийной функции и специальному случаю минимизации среднеквадратичной ошибки в одном и том же параметрическом пространстве ограниченной машины Больцмана:

$$\max(\ln P(x)) = \min(CE_s) = \min(E_s). \quad (2.28)$$

Данная теорема обобщает полученные результаты. Из нее следует, что использование различных критериев при обучении приводит к эквивалентным правилам. Таким образом, общий принцип обучения ограниченной машины Больцмана одинаков при использовании различных целевых функций. Решение задач максимизации функции правдоподобия и минимизации кросс-энтропийной функции ошибки осуществляется с помощью эквивалентных правил в предположении о линейности используемых нейронных элементов. Важно отметить, что, при использовании критерия среднеквадратичной ошибки, мы учитываем также возможное нелинейное представление нейронных элементов. Следовательно, преимущество в использовании среднеквадратичной ошибки в качестве критерия оптимизации состоит в том, что при использовании MSE могут быть получены правила обучения для обоих (линейного и нелинейного) случаев.

В дальнейшем изложении для идентификации предлагаемого подхода используется имя **Reconstruction Error-Based Approach (REBA)**. Классический метод будем называть C-RBM (Classic Restricted Boltzmann Machine Training).

2.4 Предобучение сверточных слоев

Выполнение предобучения сверточных слоев имеет особое значение при решении задач компьютерного зрения в силу эффективности таких слоев при обработке визуальной информации, представленной с помощью отдельных изображений и видео.

Отметим, что преобразования, выполняемые над сверточными слоями при обучении, идентичны преобразованиям для полносвязных слоев. Отличие заключается в операции «развертки» (deconvolution), которая используется на этапе предобучения при реконструкции активности видимых нейронов и на этапе тонкой настройки для метода обратного распространения при вычислении ошибок на скрытых слоях сети. Эта операция применяется в сверточных нейронных сетях для повышения частоты дискретизации. Рассмотрим визуальное представление операции свертки для случая карт 3Х3 и 2Х2 (рисунок 2.3).

$$\begin{array}{|c|c|c|} \hline X_1 & X_2 & X_3 \\ \hline X_4 & X_5 & X_6 \\ \hline X_7 & X_8 & X_9 \\ \hline \end{array} * \begin{array}{|c|c|} \hline y_1 & y_2 \\ \hline y_3 & y_4 \\ \hline \end{array} = \begin{array}{|c|c|} \hline Z_1 & Z_2 \\ \hline Z_3 & Z_4 \\ \hline \end{array}$$

Рисунок 2.3 – Операция свертки

Сверточный слой с одним ядром может быть представлен в виде разряженного слоя, а операция «развертки» – слоем, полученным применением симметричного преобразования относительно разряженного. При этом полученная сеть представляет собой автоэнкодер с симметричными связями относительно скрытого слоя (рисунок 2.4).

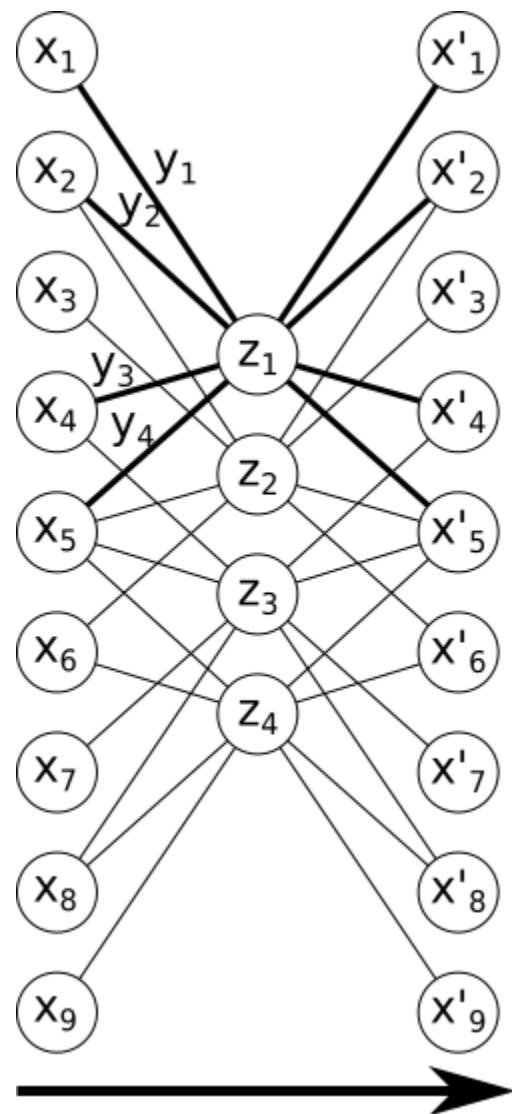


Рисунок 2.4 – Разряженный автоэнкодер операции свертки

На основании данного представления могут быть получены формулы для выполнения «развертки»:

$$x'_1 = z_1 y_1$$

$$\begin{aligned}
x'_2 &= z_1y_2 + z_2y_1 \\
x'_3 &= z_2y_2 \\
x'_4 &= z_1y_3 + z_3y_1 \\
x'_5 &= z_1y_4 + z_2y_3 + z_3y_2 + z_4y_1 \\
x'_6 &= z_2y_4 + z_4y_2 \\
x'_7 &= z_3y_3 \\
x'_8 &= z_3y_4 + z_4y_3 \\
x'_9 &= z_4y_4
\end{aligned}$$

Сама операция «развертки» представлена на рисунке 2.5.

0	0	0	0
0	Z_1	Z_2	0
0	Z_3	Z_4	0
0	0	0	0

*

y_4	y_3
y_2	y_1

=

X'_1	X'_2	X'_3
X'_4	X'_5	X'_6
X'_7	X'_8	X'_9

Рисунок 2.5 – Операция «развертки»

Введя операции матричного поворота на 180 градусов и заполнения нулями границ матрицы (padding), можно получить следующие общие формулы для выполнения «развертки» (рисунок 2.6).

padding(

Z_1	Z_2
Z_3	Z_4

, 1) * rot180(

y_1	y_2
y_3	y_4

) =

X'_1	X'_2	X'_3
X'_4	X'_5	X'_6
X'_7	X'_8	X'_9

Рисунок 2.6 – Формула «развертки»

Введем условное обозначение для операции «развертки» – \circledast .

Применяя полученную формулу для выполнения «развертки», можно осуществлять процедуру CD на сверточных слоях ГНС, которые в этом случае рассматриваются как сверточные ограниченные машины Больцмана (CRBM, рисунок 2.7).

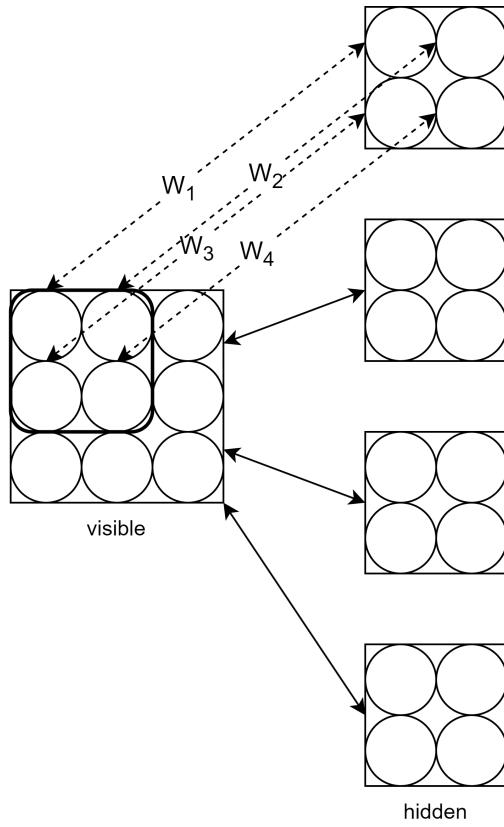


Рисунок 2.7 – Визуализация CRBM

Рассмотрим формулы для получения необходимых карт признаков для расчета градиента изменения параметров слоя в случае использования процедуры CD-1:

$$y_j(0) = F(x_i(0) * w_{ij} + T_j),$$

$$x_i(1) = F(y_j(0) \circledast w_{ij} + T_i),$$

$$y_j(1) = F(x_i(1) * w_{ij} + T_j),$$

где w_{ij} – i -тый компонент j -того ядра свертки,

$x_i(0)$ – i -тый компонент входной карты признаков,

T_j – пороговый элемент j -того нейрона скрытого слоя,

$y_j(0)$ – j -тый компонент выходной карты признаков,

$x_i(1)$ – i -тый компонент реконструированной входной карты признаков,

T_i – пороговый элемент i -того нейрона видимого слоя,

$y_j(1)$ – j -тый компонент реконструированной выходной карты признаков.

Классические правила обучения RBM могут быть переформулированы для случая CRBM (CD-1, последовательное обучение):

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(x_i(0) * y_j(0) - x_i(1) * y_j(1)),$$

$$T_i(t+1) = T_i(t) + \alpha(x_i(0) - x_i(1)),$$

$$T_j(t+1) = T_j(t) + \alpha(y_j(0) - y_j(1)).$$

В соответствии с предлагаемым методом, правила обучения могут быть переформулированы следующим образом (CD-1, последовательное обучение):

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \alpha((y_j(1) - y_j(0))F'(S_j(1)) * x_i(1) + \\ &\quad (x_i(1) - x_i(0))F'(S_i(1)) * y_j(0)), \\ T_i(t+1) &= T_i(t) - \alpha(x_i(1) - x_i(0))F'(S_i(1)), \\ T_j(t+1) &= T_j(t) - \alpha(y_j(1) - y_j(0))F'(S_j(1)). \end{aligned}$$

Аналогичным образом могут быть получены правила для CD- k случая процедуры обучения и группового случая.

Таким образом, для глубокой сверточной нейронной сети, содержащей слои разных типов (полносвязные и сверточные) возможно сочетание нескольких вариантов обучения – с представлением в виде CRBM (для сверточных слоев) и в виде RBM (для полносвязных).

2.5 Алгоритм редуцирования параметров

Известно, что глубокие нейронные сети обладают высокой степенью избыточности. Проведенные исследования и разработанные практические решения позволяют осуществлять обучение редуцированных нейросетевых моделей, при создании и обучении которых в качестве вспомогательной выступает объемная модель, имеющая миллиарды настраиваемых параметров. При этом получаемая упрощенная модель, обладая рядом преимуществ (улучшенная производительность в так называемом inference-режиме, сниженные требования к оперативной памяти, меньший размер модели при выполнении сериализации), характеризуется обобщающей способностью не хуже чем у «большого» варианта модели, решающего ту же задачу. Таким образом, становится возможным формирование модели со схожими характеристиками, но с редуцированным количеством параметров или с их измененным качеством, что отражается на времени отработки сети и ее размерах. Можно предположить, что указанные параметры могут быть исключены без существенного ухудшения в эффективности работы модели. Важный вопрос, возникающий при выполнении такой операции

редуцирования, касается выбора алгоритма отсеивания малоинформативных параметров.

Известно, что полносвязные нейронные сети обладают определенной степенью избыточности. Полносвязный слой в сравнении со сверточным содержит большее количество настраиваемых параметров, однако в задачах компьютерного зрения сверточные нейронные сети показывают существенно лучшие результаты по обобщающей способности, чем полносвязные. Таким образом, очевидно, что в полносвязных сетях при большем количестве настраиваемых параметров, они используются менее оптимально. Можно предположить, что указанные «избыточные» параметры могут быть исключены без существенного ухудшения эффективности работы модели. Важный вопрос, возникающий при выполнении такой операции редуцирования, касается самого алгоритма отсеивания малоинформативных параметров.

К основным методам компрессии (сжатия) нейросетевых моделей на настоящий момент относятся:

- 1) квантизация – quantization [54, с. 3];
- 2) дистилляция – distillation [55, с. 2-3];
- 3) матричная факторизация низкого ранга – low-rank matrix factorization [56, с. 6656-6657];
- 4) прореживание параметров (прунинг) – pruning [57, с. 1].

Квантизация используется для физического уменьшения памяти, занимаемой моделью. В этом подходе осуществляется редуцирование типа данных, используемого для хранения параметров нейросетевой модели (например, осуществляется переход от 32-битного к 16-битному или 8-битному представлению типа).

Дистилляция основывается на возможности использования более «тяжелых» моделей для обучения моделей меньшего размера. В этом случае обученная модель-учитель генерирует примеры, которые используются для обучения модели-ученика. Размер такой модели-ученика может быть существенно меньше первоначальной сети. По разновидностям выделяют следующие типы дистилляции:

- офлайн-дистилляция (применяется последовательное обучение модели-учителя, затем – модели-ученика);
- онлайн-дистилляция (применяется одновременное обучение модели-учителя и модели-ученика);
- самодистилляция (в качестве модели-учителя и модели-ученика выступает одна модель).

Матричная факторизация низкого ранга позволяет осуществить де-

композицию матрицы весовых коэффициентов большого размера на совокупность матриц меньшего размера. Применение данного подхода помогает уменьшить память, занимаемую моделями и ускорить время работы модели. К недостаткам метода относят его вычислительную сложность. *Прореживание параметров (прунинг)* позволяет удалить часть связей и нейронов (в зависимости от выполняемой техники редуцирования), что уменьшает структурную избыточность «тяжелых» глубоких нейросетевых моделей. По типу выполняемой техники прунинга выделяют:

- прунинг связей;
- прунинг нейронов или фильтров (в зависимости от типа слоя – полносвязного или сверточного);
- прунинг слоев.

При осуществлении прунинга необходимо определить критерий прореживания параметров, то есть правило, согласно которому будут определяться удаляемые связи модели. Существует несколько таких критериев, одним из наиболее часто используемых является прореживание по величине (magnitude pruning), согласно которому отсеиваемые связи определяются следующим образом:

$$thresh(w_i) = \begin{cases} w_i, |w_i| > t, \\ 0, |w_i| \leq t, \end{cases}$$

где параметр t определяет степень прореживания.

Таким образом, происходит обнуление тех параметров, влияние которых на выходную активность заданного слоя минимально.

При разработке алгоритма редуцирования, описанного в данной работе, использовался вариант данного критерия, согласно которому

$$t = s * \sigma_l, \quad (2.29)$$

где σ_l определяет стандартное отклонение параметров слоя l , параметр s подбирается экспериментально, например, исходя из чувствительности слоя (первые сверточные слои обладают большей чувствительностью, чем остальные).

Предлагаемый алгоритм редуцирования связей НС основывается на прунинге, но имеет модификацию, которая заключается в наличии этапа неконтролируемого предобучения на основе RBM.

Рассмотрим подход для редуцирования связей полносвязной нейронной сети, основанный на использовании предобучения. Первый и четвертый этапы данной процедуры аналогичны этапам выполнения предобучения типа II, описанного в главе I. В ходе выполнения дополнительных этапов 2

и 3 формируются разряженные связи между входными и выходными нейронными элементами слоя и уменьшается его размерность за счет удаления части нейронных элементов, которые не используются при «тонкой настройке» и дальнейшей эксплуатации нейросетевой модели (рисунок 2.8):

1) неконтролируемое предобучение НС с использованием «жадного» алгоритма, начиная с первого слоя. Параметры каждого слоя, представленные весовыми и пороговыми коэффициентами, настраиваются в соответствии с правилами обучения ограниченной машины Больцмана;

2) «обнуление» весовых коэффициентов слоев нейронной сети, абсолютные значения которых не превышают некоторый заданный порог $t > 0$. Иначе говоря, параметры со значениями, попадающими в интервал $[-t, t]$, не изменяются при дальнейшем обучении;

3) реконфигурация ГНС, заключающаяся в удалении нейронов, не участвующих в формировании выходной активности сети (то есть нейронов, которые имеют нулевые векторы весов или, а, в случае сверточных слоев, нейронов, имеющих нулевые ядра свертки). Реконфигурация выполняется в соответствии со следующими правилами:

для каждого i -того слоя НС, кроме первого и последнего:

- если j -тый вектор-столбец матрицы весовых коэффициентов W_i нулевой, то удалить j -тый вектор-столбец из W_i и удалить j -тую вектор-строку из W_{i+1} ;
- если k -тая вектор-строка матрицы весовых коэффициентов W_i нулевая, то удалить k -тую вектор-строку из W_i и удалить k -тый вектор-столбец из матрицы W_{i-1} ;

В случае применения алгоритма к сверточной нейронной сети осуществляется удаление отдельных ядер сверточного слоя

4) «тонкая настройка» нередуцированных параметров НС.

В ходе реализации второго этапа алгоритма возможно включение дополнительных операций уплотнения разреженных матриц параметров для достижения более компактного представления весовых коэффициентов.

При реализации данного алгоритма исключаются параметры модели, значениями которых можно пренебречь в соответствии с некоторым заданным порогом отсеивания.

2.6 Выводы

1) Предложен альтернативный подход к обучению ограниченной машины Больцмана, базирующийся на идее минимизации суммарной квадратичной ошибки восстановления образов на скрытом и видимом слоях

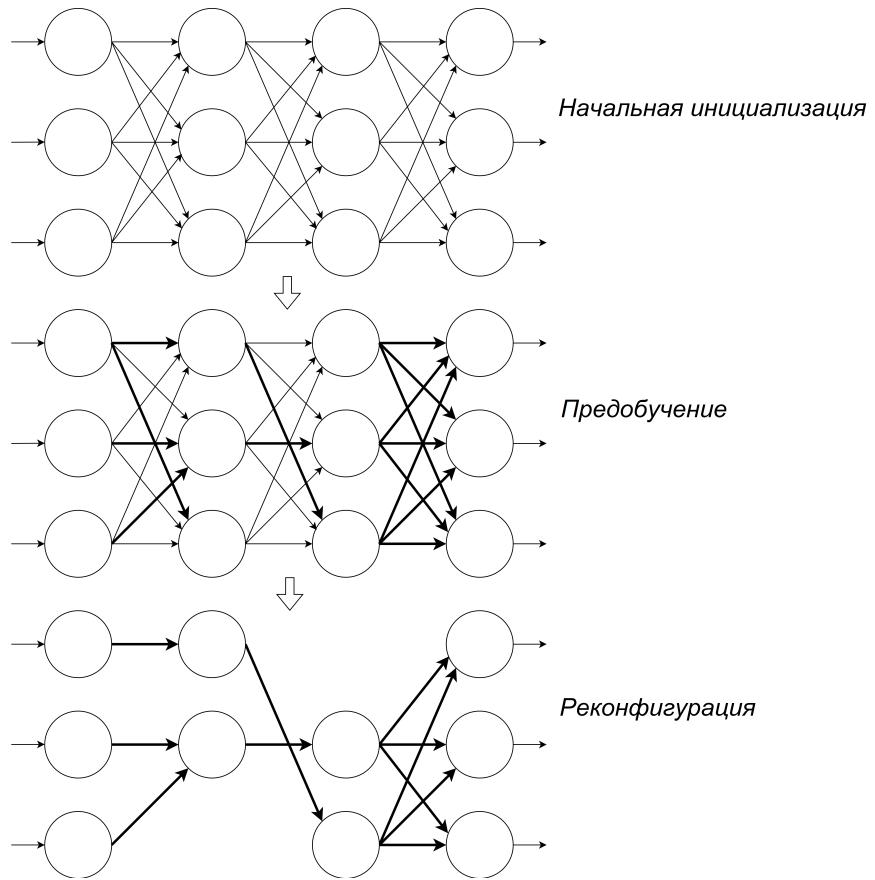


Рисунок 2.8 – Выполнение редуцирования весовых коэффициентов с архитектурной реконфигурацией НС

[1–А, с. 8-9; 2–А, с. 131-132; 3–А, с. 11; 14–А, с. 140-143; 16–А, с. 184; 15–А, с. 484-485; 18–А, с. 93-94].

2) Доказана теорема об эквивалентности максимизации функции правдоподобия распределения входных данных $P(x)$ и минимизации суммарной квадратичной ошибки восстановления образов на скрытых и видимых слоях в одном и том же параметрическом пространстве ограниченной машины Больцмана при использовании линейных нейронов [2–А, с. 133-134; 3–А, с. 12].

3) Доказана теорема об эквивалентности максимизации функции правдоподобия распределения входных данных $P(x)$ минимизации кросс-энтропийной целевой функции $CE_s(k)$ в одном и том же параметрическом пространстве ограниченной машины Больцмана (случай CD-1) [2–А, с. 136; 3–А, с. 13-14].

4) Сформулирована обобщающая теорема об эквивалентности максимизации функции правдоподобия распределения входных данных $P(x)$, минимизации кросс-энтропийной функции, минимизации суммарной квадратичной ошибки (специальный случай) в одном и том же параметрическом пространстве ограниченной машины Больцмана [2–А, с. 136; 3–А, с. 14;

5–А, с. 8].

5) Приведены правила обучения для различных случаев линейной и нелинейной машины Больцмана (RBM) [2–А, с. 134-135; 3–А, с. 12-13; 15–А, с. 485].

6) Приведены правила для обучения сверточных слоев ГНС с использованием модели CRBM [8–А, с. 129-130].

7) Предложен метод редуцирования параметров глубокой полносвязанной нейронной сети, основывающийся на использовании процедуры неконтролируемого предобучения слоев [7–А, с. 296-297; 8–А, с. 130; 26–А, с. 939].

ГЛАВА 3

ЭКСПЕРИМЕНТАЛЬНОЕ ОБОСНОВАНИЕ МЕТОДА НЕКОНТРОЛИРУЕМОГО ПРЕДОБУЧЕНИЯ

В данной главе приводится практическое обоснование теоретических результатов, описанных в главе 2, а именно, результаты сравнительного анализа методов предобучения и алгоритма редуцирования весовых коэффициентов нейросетевых моделей.

Для оценки эффективности предлагаемых методов применяются известные выборки данных, используемые для апробации подходов в области машинного обучения:

1) выборка Iris flower dataset – набор из 150 4-х мерных векторов признаков, представляющих свойства видов ирисов, относящихся к трем различным классам [58, с. 180];

2) выборка MNIST – набор из 70.000 черно-белых изображений рукописных цифр. Каждое изображение имеет размерность 28X28 пикселей и представлено в виде одномерного вектора данных [59];

3) выборки CIFAR-10 и CIFAR-100 – наборы, содержащие по 60.000 цветных изображений объектов, принадлежащих 10 (CIFAR-10) и 100 (CIFAR-100) различным классам. Каждое изображение имеет размерность 32X32 пикселя [60, с. 3].

Приводятся результаты экспериментальных исследований, полученных для задач сжатия данных, классификации и редуцирования параметров ГНС.

Исходный код реализованных методов предобучения и редуцирования ИНС приводится в приложении А.

3.1 Используемое аппаратное и программное обеспечение

Выполнение экспериментов с глубокими нейронными сетями даже при условии использования выборок среднего размера занимает продолжительное время и требует наличия значительных вычислительных ресурсов. Важными аппаратными ресурсами, влияющими на вычислительный процесс и потенциальными источниками аварийного завершения программы, становятся оперативная память и количество обрабатывающих элементов в случае использования графических ускорителей. В данной работе про-

блема наличия данных ресурсов была частично решена использованием сервиса Google Colab [61]. Данный сервис предоставляет доступ к видеоускорителям A100 и V100, производительности которых оказалось достаточно для проведения всех расчетов. Существуют другие способы ускорения расчетов, производимых при обучении нейронной сети, например, использование вычислительных кластеров [62, с. 701].

Также при обучении моделей использовались локальные ресурсы:

- 1) ноутбук N580VD-DM298 (процессор Intel Core i7 7700HQ, 4-х ядерный, с максимальной тактовой частотой 3800 МГц; оперативная память 16 Гб типа DDR4; видеокарта NVIDIA GeForce GTX 1050 4 ГБ);
- 2) десктопный компьютер (процессор Intel Core i7-4790K, 4-х ядерный, с максимальной тактовой частотой 4.00 ГГц; оперативная память 8 Гб типа DDR3; видеокарта NVIDIA GeForce GTX 750 Ti 2 ГБ).

В вычислительных экспериментах использовалось следующее основное программное обеспечение:

- интерпретатор языка программирования Python версии 3.X.;
- интегрированная среда разработки PyCharm;
- и дополнительные библиотеки Python:
 - Neptune – пакет для поддержки эксперимента в области машинного обучения, содержит набор инструментов для поддержки логирования результатов обучения, воспроизводимости и отладки процесса;
 - NumPy – пакет для работы с массивами;
 - Pandas – пакет для обработки и анализа данных;
 - Pillow – пакет для работы с графическими файлами;
 - OpenCV – библиотека для решения различных задач КЗ;
 - scikit-learn – библиотека машинного обучения;
 - SciPy – библиотека с реализацией фундаментальных алгоритмов для научных и технических вычислений;
 - TensorFlow – библиотека машинного обучения с реализацией основных классов для работы с глубокими нейронными сетями;
 - TensorBoard – специализированная библиотека для визуализации процесса обучения НС, а также агрегирования основной информации об обучении;
 - PyTorch – библиотека машинного обучения с акцентом на глубокие нейронные сети, а также задачи CV и NLP;
 - Torchvision – пакет, содержащий реализации основных архитектур моделей, базовые преобразования данных и выборки, применяемые при решении задач компьютерного зрения (входит в инфраструктуру PyTorch).

3.2 Сжатие данных

Для первоначальной оценки эффективности предлагаемого подхода предобучения глубокой нейронной сети, были проведены вычислительные эксперименты на искусственном наборе данных x , лежащем на одномерном многообразии (спиральная петля), погруженном в трехмерное пространство [63, с. 58] и генерируемом равномерно распределенным параметром $t \in [-1, 1]$ [3-A, с. 14-15]:

$$\begin{cases} x_1 = \sin(\pi t) + \mu \\ x_2 = \cos(\pi t) + \mu, \\ x_3 = t + \mu \end{cases}$$

где μ – Гауссов шум со средним 0 и стандартным отклонением 0.05.

Для проверки предложенного метода, был обучен семислойный автоэнкодер, используя выборку из 1000 примеров. Глубокий автоэнкодер показан на рисунке 3.1. Мы использовали логистические функции активации для всех слоев нейронной сети, за исключением среднего слоя («узкого места»). На этом слое использовалась линейная функция активации.

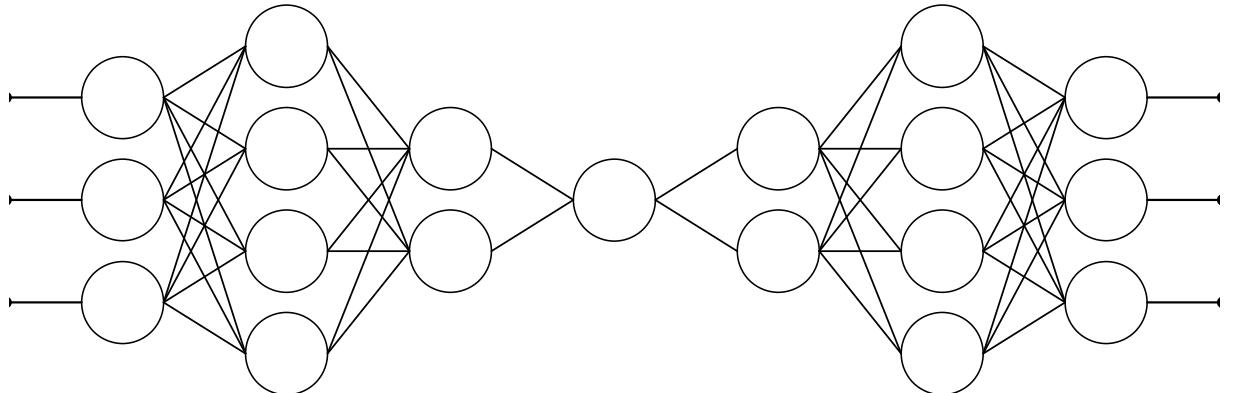


Рисунок 3.1 – Глубокий автоэнкодер

Результаты обучения показаны в таблице 3.1. Здесь MSE – среднеквадратичная ошибка на выборке обучения, MS – среднеквадратичная ошибка на тестовой выборке для проверки обобщающей способности сети. Количество примеров в тестовой выборке – 1000. Скорость обучения α – 0.1 для классического метода обучения и 0.5 для REBA для всех экспериментов.

Число эпох предобучения – 10. Число эпох «тонкой настройки» нейронной сети – 1000. Из полученных результатов видно, что использование метода предобучения REBA позволило улучшить обобщающую способность глубокого автоэнкодера для случаев CD-1 и CD-10. На рисунках 3.2

Таблица 3.1 – Сравнение методов предобучения (сжатие)

Метод	CD-k	MSE	MS
C-RBM	1	0,699	0,886
	5	0,710	0,932
	10	0,689	0,916
	15	0,688	0,873
REBA	1	0,673	0,851
	5	0,719	0,966
	10	0,677	0,907
	15	0,700	0,895

и 3.3 изображены оригинальные данные, на которых производилось обучение, и восстановленные из одного нелинейного компонента, используя тестовые данные. Как можно видеть, автоэнкодер восстанавливает данные из одного нелинейного компонента с хорошей точностью.

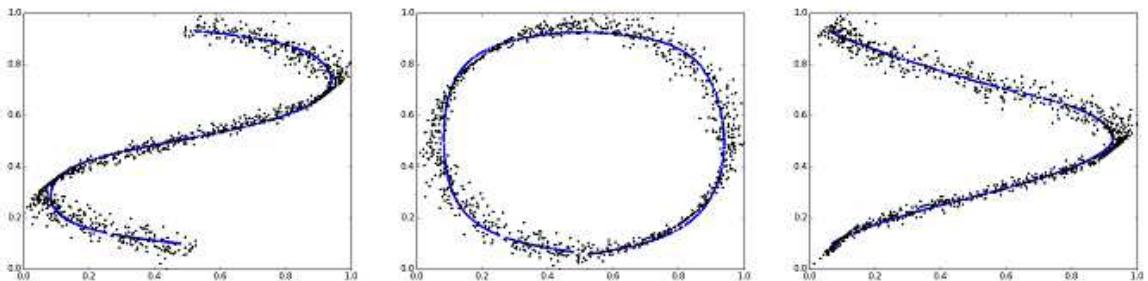


Рисунок 3.2 – 2D-изображения оригинальных и реконструированных данных

3.3 Классификация

3.3.1 Критерий оценки результатов классификации

Для оценки качества решения задачи классификации применялся следующий подход. Вначале определялся k -тый нейрон, выходное значение которого было максимальным для заданного образа s (данное число соответствует метке класса, получаемого моделью):

$$k_s = \arg \max_j y_j^s. \quad (3.1)$$

Затем получившееся значение сравнивалось с эталонными значениями и количество совпадений суммировалось для всех образов:

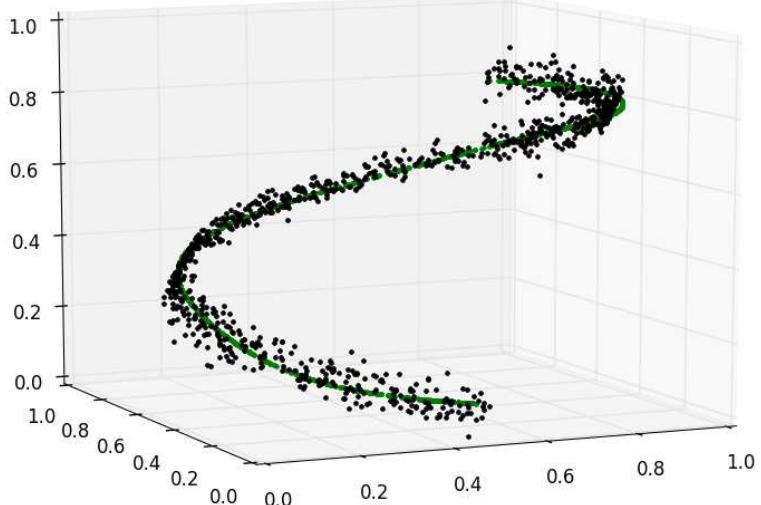


Рисунок 3.3 – 3D-изображение оригинальных и реконструированных данных

$$S = \sum_{s=1}^L [k_s = e_s], \quad (3.2)$$

где $[x]$ – нотация (скобка) Айверсона для высказывания x ,
 L – общее количество образов из оцениваемого множества,
 e_s – эталонное значение, метка класса, соответствующая s -тому образу.

Значения нотации Айверсона могут быть получены по следующей формуле:

$$[x] = \begin{cases} 1, & x \text{ is True} \\ 0, & x \text{ is False.} \end{cases}$$

Таким образом, общая эффективность на тестовой выборке может быть получена по следующей формуле:

$$\text{Efficiency} = \frac{S}{L} * 100\%. \quad (3.3)$$

3.3.2 Решение простейшей задачи классификации

Рассмотрим решение задачи классификации на примере известной выборки Фишера [58, с. 180]. Эта выборка была представлена Рональдом

Фишером в 1936 году в качестве примера для демонстрации разработанного им метода линейного дискриминантного анализа.

Отличительная особенность этой задачи в том, что один класс образов из данной выборки является линейно разделимым, в то время как два других – нет (рисунок 3.4).

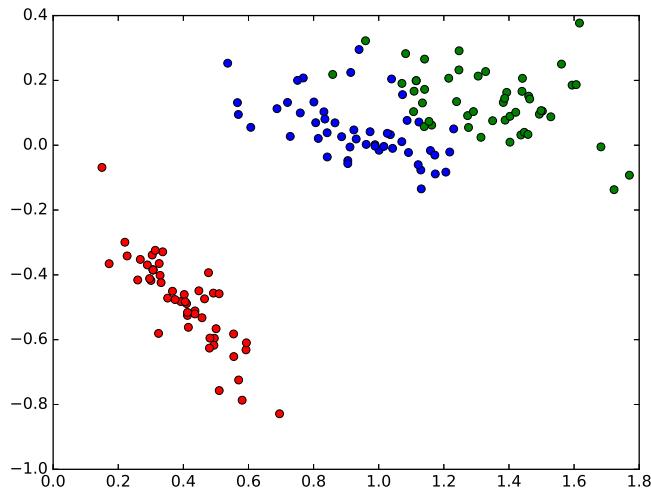


Рисунок 3.4 – Визуализация образов ирисов Фишера (получена применением метода РСА)

Применяя классические подходы кластеризации (например, метод k-средних), можно заметить неприемлемое качество распознавания на границах двух линейно неразделимых классов (рисунок 3.5).

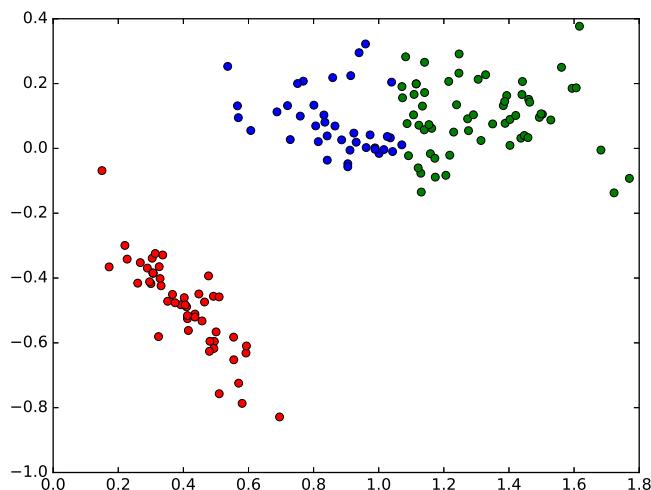


Рисунок 3.5 – Кластеризация образов ирисов Фишера методом k-средних

Для решения этой задачи была использована сеть с архитектурой **4-32-16-8-3** с сигмоидными функциями активации на каждом обрабатывающем слое [30–A]. Другие параметры:

- фаза предобучения: скорость – 0.1 (для REBA 0.4), моментный параметр – переменный (от 0.5 до 0.9), размер мини-батча - 5, количество эпох обучения каждого слоя – 50;
- фаза обучения: скорость – 0.05 (с редуцированием, коэффициент 0.99), моментный параметр – 0.9, размер мини-батча – 5, количество эпох обучения – 2000, параметр L2-регуляризации (weight decay) – 0.00001.

После обучения нейронной сетью была достигнута совокупная ошибка распознавания 99,33%. Таким образом, неправильно распознанным остался только один образ из всей выборки (рисунок 3.6).

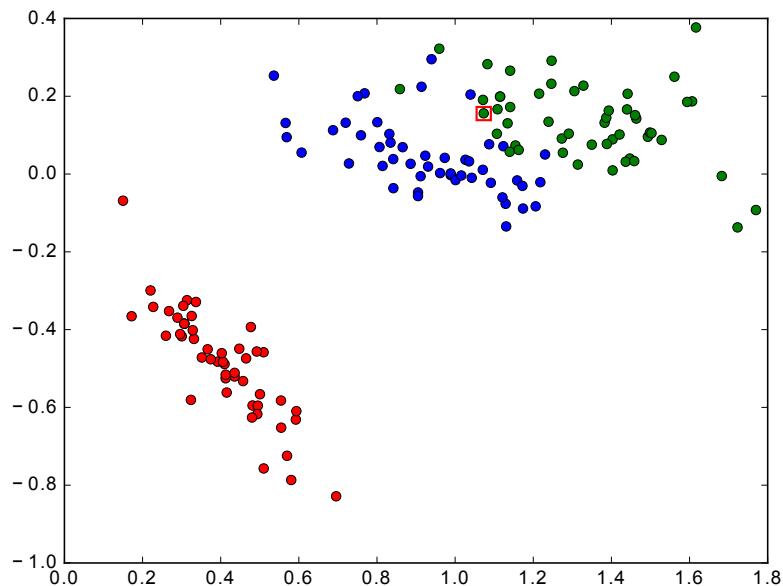


Рисунок 3.6 – Результат работы обученной нейронной сети с отмеченным неправильно классифицированным образом

Эволюция среднеквадратичной ошибки после предобучения разными методами и без предобучения изображена на рисунке 3.7.

Как видно из представленного графика, предобучение позволяет получить хорошую начальную инициализацию весов и порогов нейронной сети, что делает последующий этап «тонкой настройки» методом обратного распространения ошибки более эффективным.

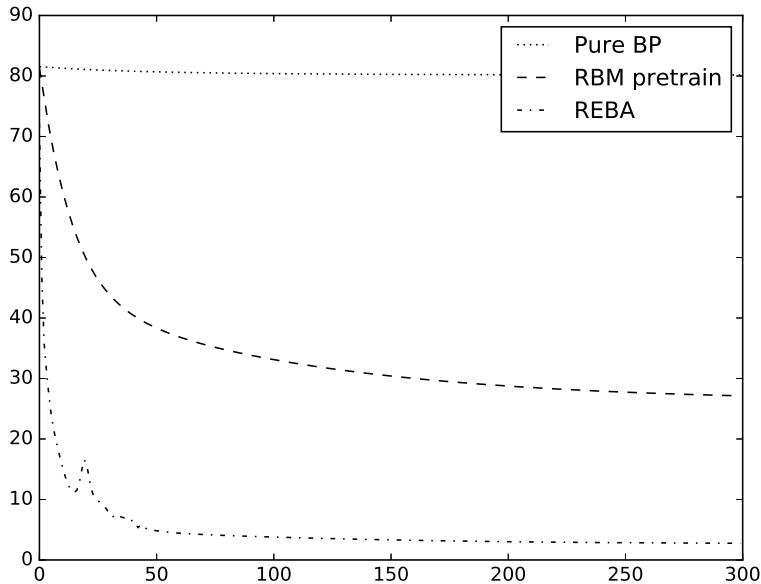


Рисунок 3.7 – Эволюция ошибок обучения разными методами

3.3.3 Описание применяемых выборок

С практической точки зрения (и в соответствии с объектом исследования) наибольший интерес представляют выборки графических образов.

Задача распознавания графических образов является одной из основных в области компьютерного зрения. В настоящий момент «золотым стандартом» выборок, применяемых для оценки эффективности моделей выступают MNIST, CIFAR-10 и CIFAR-100.

Выборка MNIST (Mixed National Institute of Standards and Technology database) является классической при тестировании систем распознавания образов, а также широко используемой для обучения и тестирования алгоритмов машинного обучения. Она сформирована как подмножество более крупной оригинальной выборки NIST [59], изображения из которого были дополнительном предобработаны (путем изменения размера и центрирования).

Выборка MNIST состоит из 60000 образов для обучения и 10000 образов для тестирования. Каждый образ представляет собой изображение цифры размером 28X28 пикселей в градациях серого цвета. На рисунке 3.8 изображен фрагмент базы изображений MNIST с наиболее труднораспознаваемыми цифрами.

Выборка CIFAR-10 [60, с. 3] является подмножеством выборки Tiny Images [64, с. 1958] и включает в себя 60.000 цветных изображений технических средств и живых существ, принадлежащий 10 различным классам



Рисунок 3.8 – Фрагмент базы изображений MNIST

(рисунок 3.9) по 6.000 изображений на каждый класс. Каждое изображение имеет размер 32Х32 пикселя.

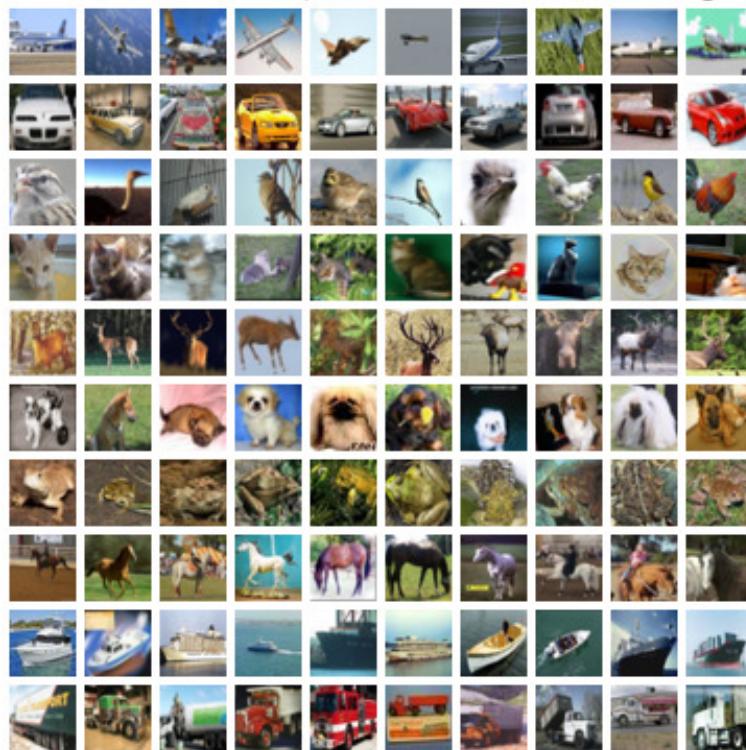


Рисунок 3.9 – Фрагмент базы изображений CIFAR-10

Выборка CIFAR-100 идентична выборке CIFAR-10 по параметрам изображений, которые в нее включены, но отличается количеством представленных классов изображений (в этой выборке общее количество классов составляет 100). Общая размерность выборки составляет также 60.000 изображений, по 600 изображений на каждый класс.

Выборки CIFAR-10 и CIFAR-100 разделены авторами на обучающую и тестовую подвыборки (объемом 50.000 и 10.000 изображений соответственно).

3.3.4 Параметры вычислительного эксперимента и результаты

В качестве целевой модели для решения задачи распознавания изображений из выборки MNIST была выбрана сверточная нейронная сеть с параметрами, представленными в таблице 3.2.

Таблица 3.2 – MNIST: основные параметры используемой модели

Параметр	Значение
Архитектура	40X5X5 – 40X5X5 – 640X320 – 320X160 – 160X10
Функция активации	ReLU
Функция активации на последнем слое	Softmax
Начальная инициализация параметров	Нормальное распределение

Пояснение по описанию используемых архитектур НС. В данной работе используется следующая нотация для описания архитектур НС:

- для полносвязных слоев $N \times M$, где N обозначает количество входных нейронов слоя, а M – количество выходных нейронов;
- для сверточных слоев $K \times S \times S$, где K обозначает количество ядер свертки в соответствующем слое, а $S \times S$ – размерность ядра свертки.

При этом для полносвязных слоев используется сокращенная форма нотации, обозначающая количество нейронов в каждом слое, например, нотация **784-100-100-10**, где 784 – это число нейронов в первом (распределяющем) слое, 100 – число нейронов во втором слое и так далее.

Как видно из представленной выше таблицы, использовалась архитектура с 5 обрабатывающими слоями, 2 из которых сверточные, а 3 – полносвязные. При этом общее число параметров модели составило 299.170.

В качестве функции активации использовалась функция ReLU на всех слоях сети, за исключением последнего слоя, на котором применялась softmax-функция. Использование ReLU позволило варианту обучения без предобучения начать процесс и завершить его с приемлемым значением эффективности.

В таблице 3.3 приведены основные используемые параметры обучения.

Помимо классического и предложенного методов предобучения в ходе проведения экспериментов был протестирован подход, при котором первый слой нейросетевой модели предобучается с использованием классического метода обучения RBM, а все прочие слои, кроме последнего классифицирующего – с использованием предлагаемого подхода REBA. Будем обозначать такой вариант предобучения как гибридный (HREBA).

Эксперименты проводились с одной и той же начальной инициализацией параметров для всех методов серией в 10 попыток, получаемые результаты затем усреднялись. В ходе эксперимента сравнивались четыре основных варианта обучения:

- 1) BP – обучение без предобучения;
- 2) REBA – обучение с предобучением (предлагаемый подход);
- 3) HREBA – обучение с предобучением (гибридный подход);
- 4) C-RBM – обучение с классическим методом предобучения).

Таблица 3.3 – MNIST: основные параметры обучения

Этап	Параметр	Значение
Предобучение	Скорость обучения	0,000125
	Размер мини-батча	128
	Моментный параметр	[0,5; 0,9]
	Количество эпох обучения	30
Обучение	Скорость обучения	0,001
	Размер мини-батча	128
	Моментный параметр	0,9
	Количество эпох обучения	50

В результате были получены показатели эффективности для выше-перечисленных методов, представленные в таблице 3.4.

Таблица 3.4 – MNIST: результаты обучения

Метод обучения	Эффективность, %
BP	99.367
REBA	99.371
HREBA	99.458
C-RBM	99.447

Как видно из представленных результатов, лучший средний показатель был достигнут гибридным методом HREBA, сочетающим в себе предо-

бучение классическим и предложенным подходом, при этом максимальная эффективность была получена этим же методом и составила **99.53 %**.

При проведении экспериментов с выборками CIFAR-10 и CIFAR-100 использовалась модель с архитектурой, представленной в таблице 3.5.

Для выборки CIFAR-100 изменения в архитектуре модели коснулись только последнего слоя (вместо 10 выходных нейронов использовалось 100 – по количеству классов в данной выборке).

Таблица 3.5 – CIFAR-10/CIFAR-100: основные параметры используемых моделей

Параметр	Значение
Архитектура	64X5X5 – 32X5X5 – 800X128 – 128X10/100
Функция активации	ReLU - Tanh - ReLU
Функция активации на последнем слое	Softmax
Начальная инициализация параметров	Нормальное распределение
Общее число параметров модели	159.914

В результате были получены показатели эффективности для вышеперечисленных методов, представленные в таблице 3.6 (для выборки CIFAR-10) и 3.7 (для выборки CIFAR-100).

Таблица 3.6 – CIFAR-10: результаты обучения

Метод обучения	Эффективность, %
BP	69.74
REBA	71.20
HREBA	71.59
C-RBM	71.51

Лучший результат был получен методом HREBA и составил **72.32%**.

Лучший результат также был получен методом HREBA и составил **40.26%**.

3.4 Редуцирование параметров НС

Экспериментальные исследования метода редуцирования параметров проводились на полносвязных и сверточных нейронных сетях.

Таблица 3.7 – CIFAR-100: результаты обучения

Метод обучения	Эффективность, %
BP	36.83
REBA	38.9
HREBA	39.86
C-RBM	39.71

В *первой группе* экспериментов производилось редуцирование различных архитектур полносвязных нейронных сетей, применяемых для классификации изображений из выборок MNIST [59], CIFAR10 и CIFAR100 [60, с. 3]. Были проведены серии экспериментов, включающих различные используемые выборки, архитектуры и варианты предобучения. В рамках одной выборки и архитектуры НС текущая инициализация параметров сохранялась для возможности сравнения эффективности различных вариантов предобучающей процедуры. Ниже для рассматриваемых выборок приведены основные параметры, включающие скорость обучения, размер мини-батча, моментный параметр и количество эпох для предобучения и «тонкой настройки» моделей (таблица 3.8).

Таблица 3.8 – Основные параметры обучения

Этап	Параметр	Значение
Обучение	Скорость обучения	0.05-0.1
	Размер мини-батча	100
	Моментный параметр	0.9
	Количество эпох обучения	50-100
Предобучение	Скорость обучения	0.05-0.2
	Размер мини-батча	32-100
	Моментный параметр	[0.5, 0.9]
	Количество эпох обучения	10

В результате вычислительного эксперимента были получены результаты для различных выборок, архитектур НС и значений параметра редуцирования t (таблицы 3.9-3.13).

Как видно из приведенных результатов, исследуемые архитектуры в целом сохраняют обобщающую способность, будучи редуцированными более чем на 80 процентов.

Также можно заметить, что чем больше настраиваемых параметров в модели, тем эффективней осуществляется редуцирование. Однако с увеличением параметра редуцирования эффективность исходной сети постепенно снижается, так как редуцированию начинают подвергаться параметры, наличие которых влияет на выходы нейронной сети.

Таблица 3.9 – Результаты для сети 784-800-800-10 на выборке MNIST

Тип	Эффективность, %, C-RBM / REBA	Количество параметров, C-RBM / REBA	Редуцировано параметров, %, C-RBM / REBA
без редуц.	98.63 / 98.33	1276810 / 1276810	0/0
t=0.2	98.61 / 98.27	233760 / 279635	81.69 / 78.1
t=0.5	98.03 / 98.05	32524 / 32817	97.45 / 97.43
t=0.8	97.1 / 96.48	17061 / 12217	98.66 / 99.04

Таблица 3.10 – Результаты для сети 784-1600-1600-800-800-10 на выборке MNIST

Тип	Эффективность, %, C-RBM / REBA	Количество параметров, C-RBM / REBA	Редуцировано параметров, %, C-RBM / REBA
wr	98.76 / 98.37	5747210 / 5747210	0/0
t=0.2	98.51 / 98.55	710734 / 781103	87.63 / 86.41
t=0.5	98.01 / 98.03	54709 / 43867	99.05 / 99.24
t=0.8	96.9 / 93.08	25385 / 14914	99.56 / 99.74

Таблица 3.11 – Результаты для сети 3072-1024-512-256-128-64-10 на выборке CIFAR10

Тип	Эффективность, %, C-RBM / REBA	Количество параметров, C-RBM / REBA	Редуцировано параметров, %, C-RBM / REBA
wr	58.56 / 55.85	3844682 / 3844682	0/0
t=0.2	58.69 / 54.37	409211 / 227072	89.36 / 94.09
t=0.5	42.08 / 41.2	29033 / 11320	99.24 / 99.71
t=0.8	23.02 / 10.0	10058 / 4886	99.74 / 99.87

Таблица 3.12 – Результаты для сети 3072-512-256-128-64-10 на выборке CIFAR10

Тип	Эффективность, %, C-RBM / REBA	Количество параметров, C-RBM / REBA	Редуцировано параметров, %, C-RBM / REBA
wr	57.28 / 53.69	1746506 / 1746506	0/0
t=0.2	56.83 / 41.72	220037 / 126846	87.40 / 92.73
t=0.5	45.29 / 44.93	20431 / 11383	98.83 / 99.35
t=0.8	10.0 / 10.0	8599 / 3797	99.51 / 99.78

Полученные результаты обосновывают возможность предобучения глубокой нейронной сети с использованием неконтролируемой процедуры без получения эффекта переобучения и снижения эффективности модели, так как в процессе предобучения фактически снижается влияние определенных параметров модели на итоговую выходную активность сети. Такие

Таблица 3.13 – Результаты для сети 3072-3072-1024-512-256-128-64-100 на выборке CIFAR100

Тип	Эффективность, %, C-RBM / REBA	Количество параметров, C-RBM / REBA	Редуцировано параметров, %, C-RBM / REBA
wr	20.84 / 21.63	13290788 / 13290788	0/0
t=0.2	20.77 / 21.01	1304525 / 703319	90.18 / 94.71
t=0.5	13.4 / 1.0	49847 / 24636	99.62 / 99.81
t=0.8	2.67 / 1.0	21329 / 16977	99.84 / 99.87

параметры имеют “паразитический” характер и фактически являются фактором переобучения модели. На этапе «тонкой настройки» модели они не модифицируются и могут быть удалены после этапа предобучения.

Помимо редуцирования параметров, было выполнено так называемое архитектурное редуцирование (шаг 3 алгоритма), на котором производилось упрощение структуры слоев моделей за счет исключения нейронов с нулевыми векторами связей.

В следующих таблицах (3.14, 3.15) приведены результаты архитектурного редуцирования для двух различных архитектур и обучающей выборки MNIST. Для предобучения использовался метод REBA.

Таблица 3.14 – Результаты для сети 784-800-800-10 на выборке MNIST

Параметр	Исходная	Редуцированная
t=0.2	784-800-800-10	784-800-556-10
t=0.5	-//-	784-710-422-10
t=0.9	-//-	784-91-114-10

Таблица 3.15 – Результаты для сети 784-1600-1600-800-800-10 на выборке MNIST

Параметр	Исходная	Редуцированная
t=0.2	784-1600-1600-800-800-10	784-889-192-686-221-10
t=0.5	-//-	784-464-157-567-182-10
t=0.9	-//-	784-17-101-118-50-10

Из приведенных таблиц можно заметить, что для моделей с большим количеством слоев и нейронов в каждом слое вероятность архитектурного редуцирования выше и оно начинает происходить уже для сравнительно небольших значений параметра редуцирования t , в том время как для более «поверхностных» моделей оно существенное проявляется только для больших значений параметра. Стоит отметить, что архитектуры, которые сохраняют почти все нейроны (например, 3.14, архитектура при $t=0.2$), все

же при этом теряют для каждого нейрона основную часть связей (сохранение нейрона в данном случае будет иметь место, даже при сохранении одной связи нейрона с предыдущим слоем).

Вторая группа экспериментов проводилась с использованием выборок CIFAR-10 и CIFAR-100 над сверточной архитектурой вида $32 \times 3 \times 3 - 32 \times 3 \times 3 - 64 \times 3 \times 3 - 64 \times 3 \times 3 - 128 \times 3 \times 3 - 128 \times 3 \times 3 - 2048 \times 512 - 512 \times N$, где N – это количество классов (для CIFAR-10 – 10 классов, для CIFAR-100 – 100 классов).

В результате вычислительного эксперимента были получены сравнительные результаты для трех основных вариантов обучения — без осуществления редуцирования (стандартный режим обучения), с редуцированием по предлагаемому алгоритму и с редуцированием по существующему алгоритму прунинга [57, с. 1] (таблица 3.16 и 3.17).

Таблица 3.16 – Результаты для сверточной сети на выборке CIFAR10

Тип	Эффективность на тестовой выборке, %	Процент редуцированных параметров по слоям, %	Средний процент редуцирования, %
Нередуцированная сеть	71.79	–	0
Редуцированная сеть (классический прунинг)	71.63	10.76, 9.30, 8.88, 8.01, 8.02, 8.17, 7.97	8.73
Редуцированная сеть (предлагаемый алгоритм)	73.47	8.45, 8.65, 8.55, 8.24, 7.95, 7.97, 7.97	8.25

Таблица 3.17 – Результаты для сверточной сети на выборке CIFAR100

Тип	Эффективность на тестовой выборке, %	Процент редуцированных параметров по слоям, %	Средний процент редуцирования, %
Нередуцированная сеть	35.51	–	0
Редуцированная сеть (классический прунинг)	34.02	10.42, 7.47, 7.92, 8.25, 7.97, 8.00, 7.94	8.28
Редуцированная сеть (предлагаемый алгоритм)	40.41	9.49, 8.18, 8.12, 7.93, 8.13, 8.05, 7.94	8.26

Во всех экспериментах использовался параметр редуцирования $s =$

0.1 (формула 2.29).

Можно заметить, что предлагаемый алгоритм обладает большей однородностью в выполнении редуцирования по слоям, в то время как классический подход осуществляет более заметное редуцирование первых слоев нейронной сети. Несмотря на то, что получившийся средний процент редуцирования для предлагаемого алгоритма ниже, чем у классического, можно отметить, что классический подход уступает предлагаемому в обобщающей способности редуцированной версии.

3.5 Выводы

- 1) Проведен сравнительный анализ методов предобучения ГНС для задачи сжатия данных с использованием автоэнкодерной модели НС [1–A, с. 10; 3–A, с. 14–15; 14–A, с. 143–145; 17–A; 29–A].
- 2) Проведен сравнительный анализ методов предобучения ГНС для задачи классификации с использованием выборок Iris flower, MNIST, CIFAR-10 и CIFAR-100. Предложена модификация метода предобучения [1–A, с. 11; 9–A, с. 11; 17–A; 29–A; 30–A].
- 3) Проведены экспериментальные исследования редуцирования параметров глубокой нейронной сети с использованием различных методов предобучения и архитектур ГНС [7–A, с. 298; 8–A, с. 130–131].

ГЛАВА 4

РАЗРАБОТКА СИСТЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ НА ОСНОВЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

В данной главе рассмотрена архитектура системы компьютерного зрения, базирующаяся на разработанных методах предобучения и редуцирования.

Рассматриваются решения практических задач компьютерного зрения, а именно определения наличия и локализации солнечных панелей на аэрофотоснимках и распознавания маркировки продукта, которые основываются на разработанной системе.

Научные и практические результаты, полученные в рамках данной диссертационной работы, внедрены и используются на следующих предприятиях Республики Беларусь:

- «Intelligent Semantic Systems» (г. Минск) при разработке подсистемы компьютерного зрения для осуществления распознавания объектов;
- ОАО «Савушкин продукт» (г. Брест) при разработке системы автоматического контроля качества нанесения маркировки продукции;
- в учебном процессе учреждения образования «Брестский государственный технический университет».

Сведения об использовании результатов диссертационной работы отражены в соответствующих актах внедрения, приведенных в приложении Б.

4.1 Архитектура системы компьютерного зрения

Для разработки и поддержки решений в области КЗ, была разработана система, интегрирующая в себе средства, позволяющие использовать все предложенные методы при обучении и валидации нейросетевых моделей.

На рисунке 4.1 представлена базовая архитектура предлагаемой модели с основными модулями.

К базовым модулям системы относятся:

1) модуль организации обучения модели КЗ – один из основным модулей, в котором осуществляется контроль за процессом обучения нейросетевых моделей;

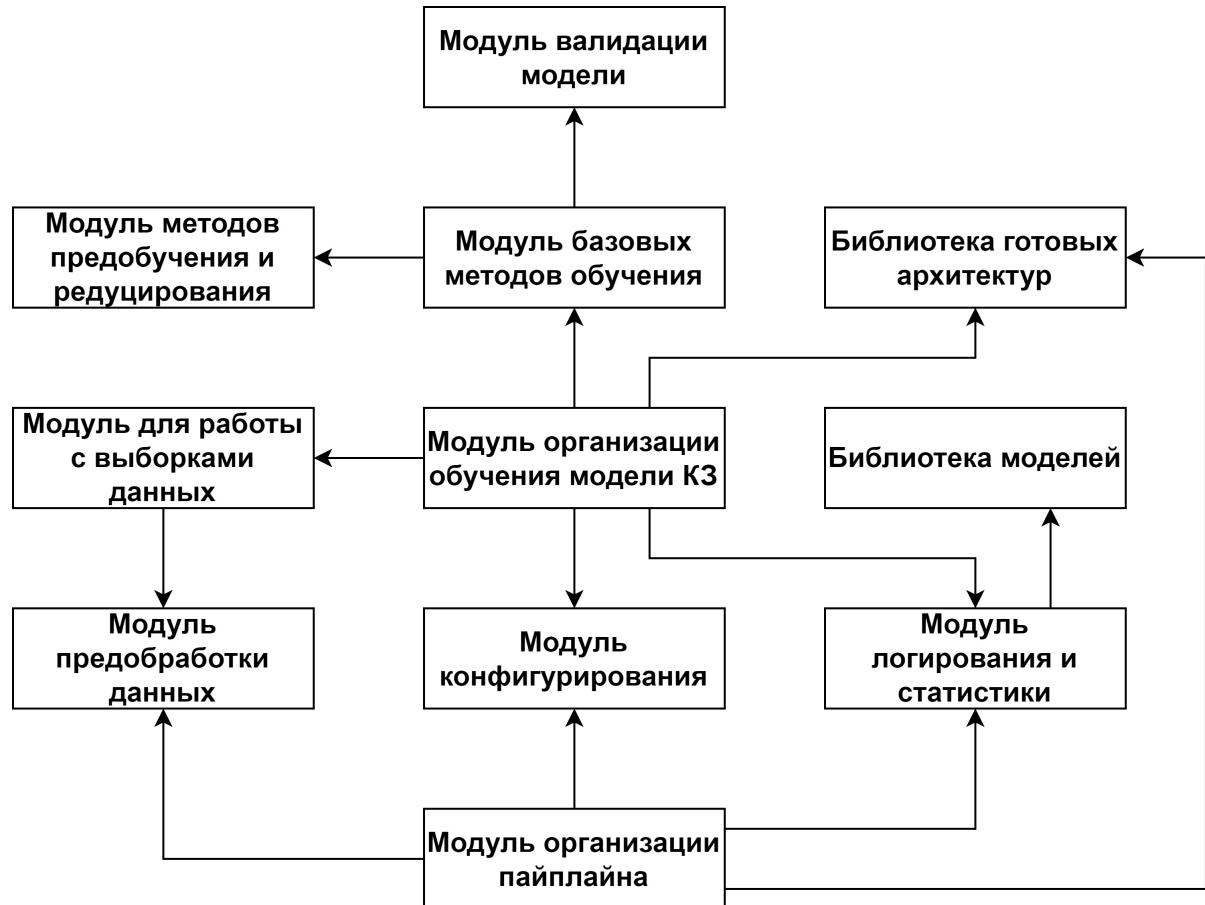


Рисунок 4.1 – Схема разработанной системы КЗ

- 2) модуль конфигурирования – хранит основные настройки, касающиеся процесса обучения и развертывания моделей;
- 3) модуль для работы с выборками данных – осуществляет загрузку данных, используемых для обучения моделей;
- 4) модуль предобработки данных – выполняет предварительную обработку данных (нормализацию, стандартизацию и так далее), зависящую от выборки;
- 5) модуль базовых методов обучения – инкапсулирует функционал, связанный с обучением, выбором оптимизатора, политики изменения скорости и так далее;
- 6) модуль методов предобучения и редуцирования – содержит реализованную логику предлагаемых в данной работе подходов;
- 7) библиотека готовых архитектур – содержит описание НС, которые могут быть использованы как целевые для обучения и конструирования пайплайнов;
- 8) библиотека моделей – содержит средства доступа к бинарным файлам данных обученных НС;
- 9) модуль валидации моделей – средства тестирования обученных НС;

10) модуль организации пайплайна – средства построения пайплайнов из обученных моделей и их развертывания.

Далее опишем задачи, которые были решены с использованием функционала разработанной системы.

4.2 Обнаружение солнечных панелей на аэрофотоснимках

4.2.1 Постановка задачи и обзор существующих решений

По причине увеличивающейся доли использования солнечной энергии, спрос на фотоэлектрические элементы в мире постоянно растет. С начала 2010-х годов индустрия производства и поддержки солнечных панелей переживает стремительный рост. Так, по состоянию на 2022 год только в США в данной отрасли работает более 10000 компаний и 263.000 сотрудников [65].

В связи с растущей популярностью этой технологии проблемы, связанные с обслуживанием солнечных панелей, также становятся актуальными. Многие сервисные компании заинтересованы в получении информации о потенциальных клиентах. Таким образом, анализ фотоснимков с целью обнаружения солнечных панелей и накопления статистической информации является актуальной задачей.

Поставленная задача заключалась в разработке решения для обнаружения солнечных панелей на изображениях с их последующей локализацией.

В статье Дж. Малофа [66] был описан подход, базирующийся на использовании SVM для автоматического обнаружения солнечных панелей, используя фотографии высокого разрешения, полученные с помощью спутника. В этом подходе сначала применяется операция предварительного скрининга, которая идентифицирует регионы, которые затем обрабатываются с целью выделения признаков. В качестве выходной информации, получаемой с помощью данной модели, выступает список регионов и доверительные значения, показывающие, насколько вероятно наличие солнечной панели в заданной области. Общая эффективность подхода определения наличия панелей в [66] составляет порядка 94%. Однако, нужно отметить, что данный метод дает лишь приблизительную информацию о местонахождении панели. Оценки точности локализации панелей в работе не приводится.

В [67, с. 4-5] используется подход, базирующийся на применении деревьев решений. Достигнутый показатель локализации панелей составил

90% в случае использования определенных параметров алгоритма, а сам метод состоит из четырех стадий. В работе в качестве обучающей выборки используется Aerial Imagery Dataset, включающий изображения разрешением до 5000 X 5000 пикселей [68, с. 6].

4.2.2 Предлагаемый подход

Предложенный нейросетевой алгоритм для обнаружения солнечных панелей отличается тем, что в качестве базовой модели используется сверточная нейронная сеть, которая показывает лучшие результаты при решении задач классификации и детекции объектов на изображениях. Другое важное отличие заключается в использовании фотографий низкого разрешения для обучения нашей модели (например, фото из Google Maps, нередко неудовлетворительного качества, рисунок 4.2).

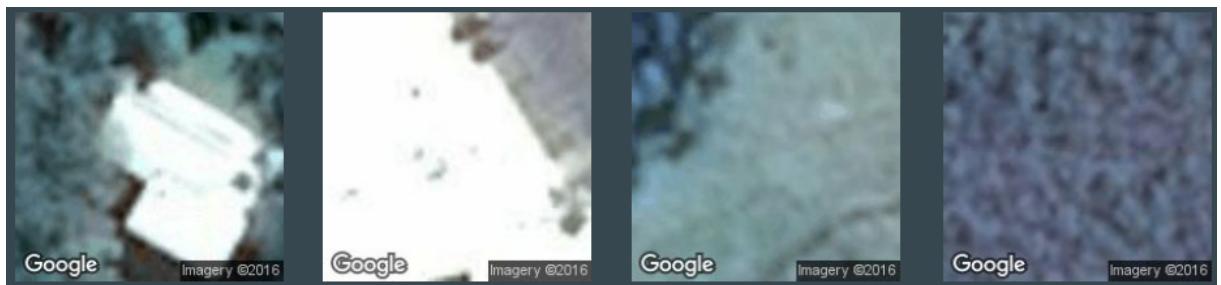


Рисунок 4.2 – Примеры изображений низкого качества из обучающей выборки

Предлагаемое решение для обнаружения солнечных панелей включает два основных компонента, в составе которых используются предобученные глубокие нейронные сети:

- 1) классификатор для оценки наличия солнечной панели на аэрофотоснимке;
- 2) детектор для локализации солнечной панели (рисунок 4.3).

Двухэтапность в решении поставленной задачи позволяет существенно ускорить обработку изображений, большую часть которых занимают объекты, отличные от искомых (солнечных панелей), т.к. нейросетевые модели, используемые на этапе локализации, как правило, более ресурсоемки. Таким образом, модель первого этапа выполняет роль фильтра для последующей обработки. Тем самым предлагаемый алгоритм можно применять для последовательной обработки больших изображений.

Классификатор для оценки наличия солнечной панели на аэрофотоснимке. Для этой подзадачи применялся сверточный классификатор с архитектурой, представленной на рисунке 4.4.

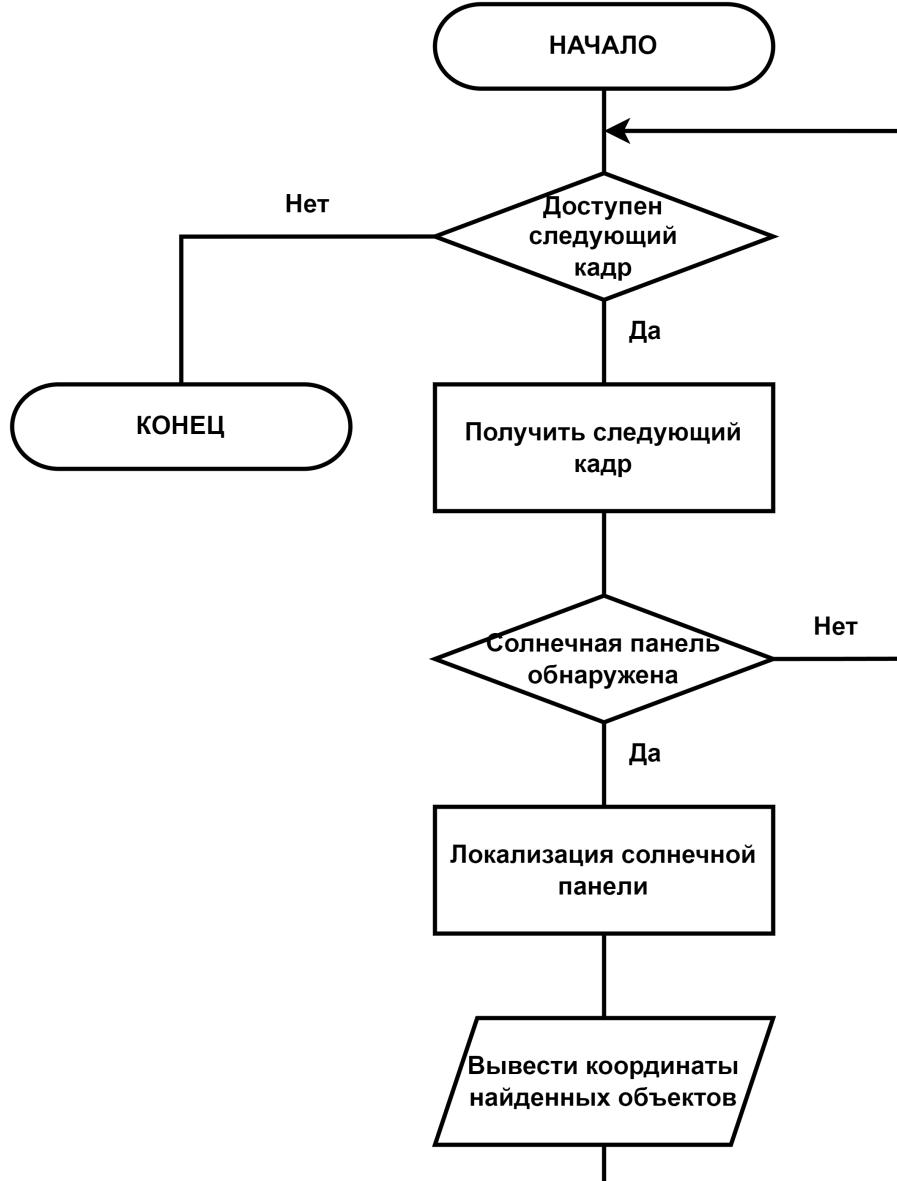


Рисунок 4.3 – Блок-схема алгоритма для обнаружения солнечных панелей

Представленная нейронная сеть состоит из шести слоев. Первые три слоя являются сверточными и выполняют выделение высокоуровневых признаков. Последние три слоя представляют собой полносвязные слои нейронных элементов, решающие задачу классификации. Для всех слоев, кроме последнего, используется функция активации ReLU (4.1):

$$ReLU(x) = \max(0, x). \quad (4.1)$$

В последнем полносвязном слое используется сигмоидная функция активации (4.2):

$$Logistic(x) = \frac{1}{1 + e^{-x}}. \quad (4.2)$$

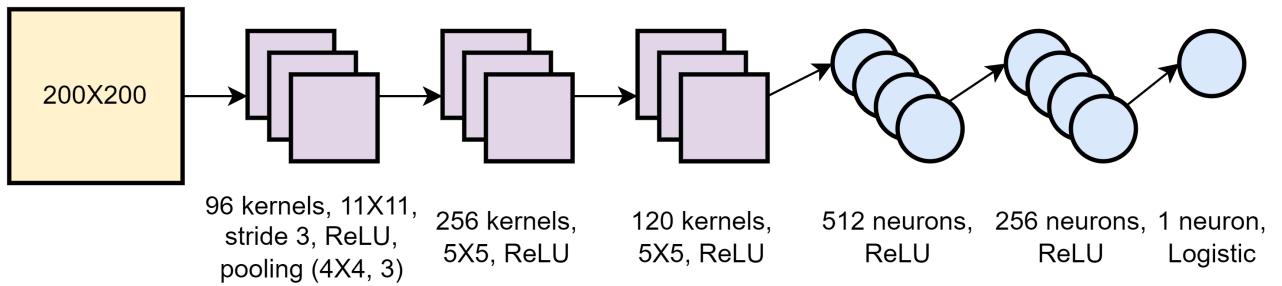


Рисунок 4.4 – Архитектура используемой сверточной нейронной сети

На первом сверточном слое шаг свертки берется равным 3, для уменьшения размерности карт признаков, получаемых с данного слоя.

Последний слой модели содержит один нейрон, по возвращаемому значению которого оценивается вероятность наличия солнечной панели на изображении. Для предобучения сверточной нейронной сети использовался предложенный метод REBA, метод стохастического градиентного спуска применялся на этапе «тонкой настройки».

Детектор для локализации солнечной панели. Для решения подзадачи локализации солнечной панели применялась ГНС Faster-RCNN (Faster Region-based Convolutional Neural Network) с классификатором ResNet-50 (рисунок 4.5), предобученным на выборке COCO [69, с. 2]. Классификатор выделяет карты признаков для поступающих на его вход изображений, передавая их на вход детектирующей части нейронной сети, которая представлена сверточной нейронной сетью, осуществляющей генерацию координат прямоугольных областей (боксов) и меток классов для каждого бокса с оценкой уверенности (confidence score). Внутри боксов с заданной вероятностью находится объект обнаруживаемого детектором класса.

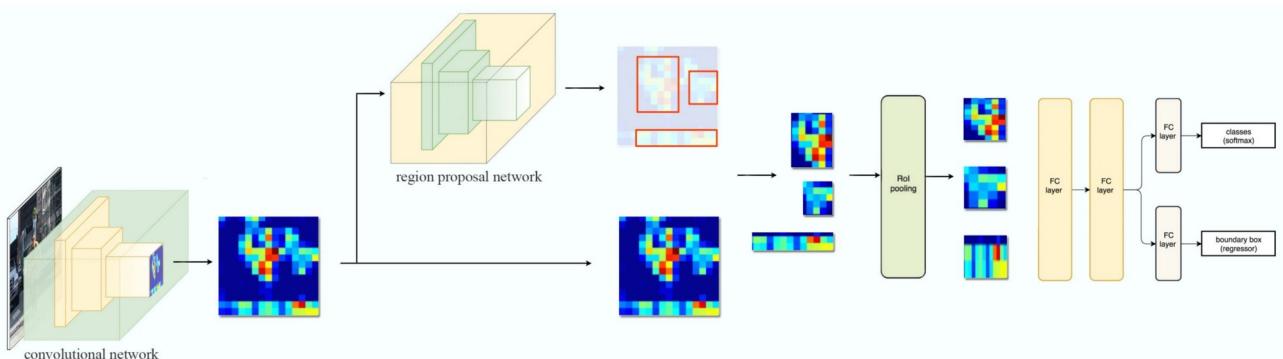


Рисунок 4.5 – Глубокая сверточная сеть Faster-RCNN [70]

Такая архитектура обеспечивает получение высоких показателей эффективности при решении задач обнаружения объектов в системах, где

общее время, затраченное на обработку и вывод результатов, некритично [71, с. 29].

4.2.3 Обучающая выборка

Исходными данными, используемыми для формирования обучающей выборки при решении задачи обнаружения солнечных панелей, являлись цветные изображения, полученные из приложения Google Maps с разрешением 200 X 200 пикселей (рисунок 4.6).



Рисунок 4.6 – Примеры изображений из обучающей выборки

Для обучения модели классификатора использовалась выборка из 3347 фотографий (где 1643 изображения содержали солнечные панели, а 1704 – не содержали), при этом 80% исходной выборки формировали обучающую выборку, а 20% – тестовую. Для обучения модели детектора использовалась выборка из 1000 изображений, 800 из которых было отнесено к обучающей, а 200 – к тестовой выборкам.

Подготовка обучающей выборки для классификатора состояла в ручном переборе изображений и отнесении их к двум группам изображений – содержащих солнечные панели и, соответственно, не содержащих. При подготовке обучающей выборки для детектора панелей также использовался ручной перебор изображений с определением для каждого характеристики прямоугольных областей, включающих солнечные панели (длина, ширина, координаты левого верхнего угла). При этом фиксировались все интересующие объекты (для случая множественных панелей).

Выделение прямоугольных областей для некоторых изображений может не выглядеть целесообразным, например, случаи, когда панели распо-

ложены под углом к горизонтальной оси снимка и имеют вытянутую форму (рисунок 4.7). Однако, для таких фотографий удается получить приемлемые результаты локализации на тестовых данных, несмотря на то, что большая часть области включает фоновое изображение.



Рисунок 4.7 – Пример изображения с преобладающим фоном

4.2.4 Оценка качества обученных моделей

Для оценки качества применялись две основные метрики.

Первая метрика использовалась для оценки классификатора. В режиме тестирования модели применялось простое преобразование выходных данных – пороговая функция вида:

$$b_s = [y_s > 0, 5],$$

где y_s представляет реальное выходное значение CNN-сети, возвращаемое для s -го входного изображения выборки,

b_s – бинаризованная форма, полученная вычислением нотации Айверсона:

$$[x] = \begin{cases} 1, & \text{x is True} \\ 0, & \text{x is False.} \end{cases}$$

Метрика оценки рассчитывалась по формулам:

$$A = \frac{S}{L} * 100\%,$$

$$S = \sum_{s=1}^L I[b_s = e_s],$$

где e_s – эталонное значение (метка),

L – объем выборки.

Таким образом, для оценки эффективности модели подсчитывается общий процент верных ответов, которые были получены с ее помощью на тестовой выборке.

Для оценки эффективности модели локализации использовалась метрика mAP.

Метрика mAP является одной из основных метрик, используемых для оценки качества моделей, применяемых для детекции. Эта метрика используется вместе со своими модификациями, вычисленными для различных значений порога IoU. Так как для рассматриваемой задачи существует только один класс объектов, метрика mAP совпадает с AP [72, с. 11].

Как известно, точность (Precision) вычисляется по формуле:

$$P = \frac{TP}{TP + FP}, \quad (4.3)$$

где TP и FP обозначают соответственно число истинно-положительных и ложно-положительных результатов детекции, и, соответственно, P определяет долю корректных детекций в общем числе детекций, сделанных моделью.

Полнота (Recall) определяется следующим образом:

$$R = \frac{TP}{TP + FN}, \quad (4.4)$$

где FN обозначает количество ложно-отрицательных результатов, R определяет долю положительных случаев в общем количестве.

Величина TP равна числу спрогнозированных моделью боксов, для каждого из которых значение IoU , вычисленное относительно эталонных боксов (*Ground-true box*), превысило некоторый заданный порог (например, 0,5). В этом случае говорят об истинно-положительной детекции. Если было спрогнозировано несколько боксов для одного бокса-эталона, то выбирается один бокс с наибольшим значением IoU , а остальные рассматриваются как FP .

Меры точности и полноты (формулы 4.3 и 4.4) используются при вычислении метрики AP (средняя точность):

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1\}} P_{interp}(R),$$

где $P_{interp}(R) = \max_{\tilde{R}: \tilde{R} \geq R} P(\tilde{R})$ – интерполированное значение точности, вычисленное для каждого уровня полноты.

Среднее AP , вычисленное по всем классам выборки дает величину mAP :

$$mAP = \frac{1}{L} \sum_{i=1}^L AP_i,$$

где L – число классов объектов, представленных в выборке.

4.2.5 Результаты обучения и тестирования

Сверточная сеть для определения наличия солнечных панелей обучалась в течение 70 эпох, используя следующие параметры:

- число эпох предобучения – 10;
- скорость обучения – 0,001;
- моментный параметр – 0,9;
- weight-decay – 0,0005;
- размер мини-батча – 20;
- вероятность применения dropout для полно связанных слоев – 0,5.

Эффективность классификации для данной задачи, составила **87,46%** (таблица 4.1).

Таблица 4.1 – Матрица ошибок

	Спрогнозировано «No»	Спрогнозировано «Yes»	Точность, %
Ожидалось «No»	325	32	91
Ожидалось «Yes»	52	261	83
Итого	377	293	87

На рисунке 4.8 изображена ROC-кривая, построенная для обученного классификатора.

Другие характеристики полученного классификатора:

- полнота = 0,8339,
- специфичность = 0,9104,
- точность = 0,8907,
- F-мера = 0,8614.

Эффективность классификатора, обученного без процедуры предобучения, составила **80,15%** на исследуемой выборке. Таким образом, использование предлагаемого метода предобучения позволило добиться прироста эффективности в **7,31%**.

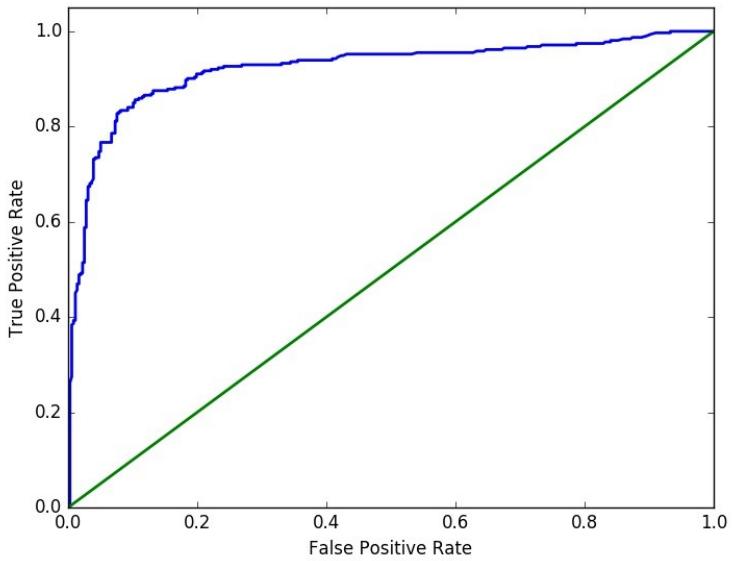


Рисунок 4.8 – ROC-кривая для обученного бинарного классификатора, $AUC = 0,92$

Сеть для локализации объектов обучалась в течение 5000 итераций (под итерацией здесь понимается настройка параметров для одного случайного изображения из обучающей выборки), используя следующие параметры:

- скорость обучения – 0,0003;
- моментный параметр – 0,9.

После выполнения обучения, проводилось исследование обобщающей способности сети с использованием тестовой выборки из 200 изображений. Полученный результат составил **$AP=0,9299$** .

Визуализация выходов первого сверточного слоя, полученная после выполнения обучения, представлена на рисунке 4.9.

На рисунке 4.10 изображены результаты детекции солнечных панелей на изображениях из тестовой выборки.

Как можно видеть, предложенный метод осуществляет достаточно точное обнаружение. На рисунке 4.11 изображены результаты работы метода для произвольных изображений (не из тестовой выборки).

4.3 Распознавание маркировки продукта

4.3.1 Постановка задачи и обзор существующих решений

В основе конкурентности любого производства лежит выпуск разнообразной и качественной продукции. Качество является определяющим

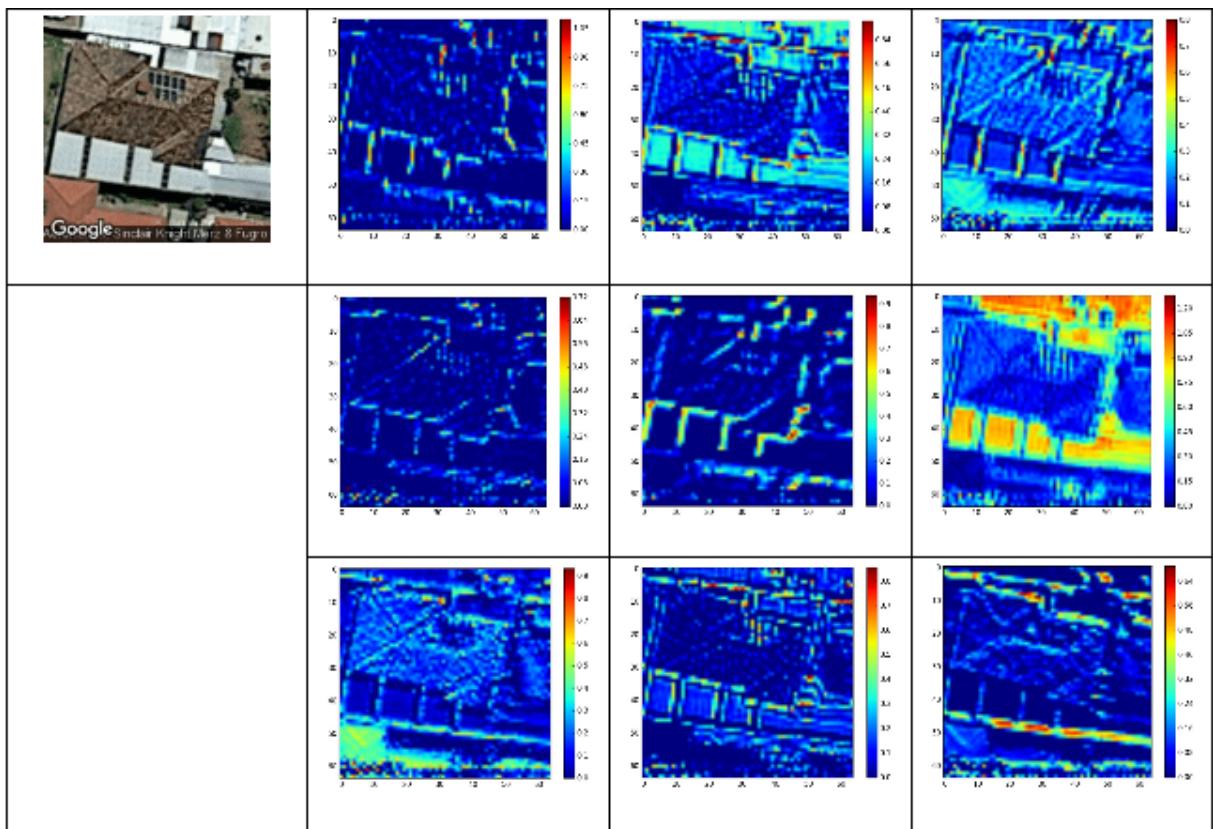


Рисунок 4.9 – Визуализация выходов первого слоя НС

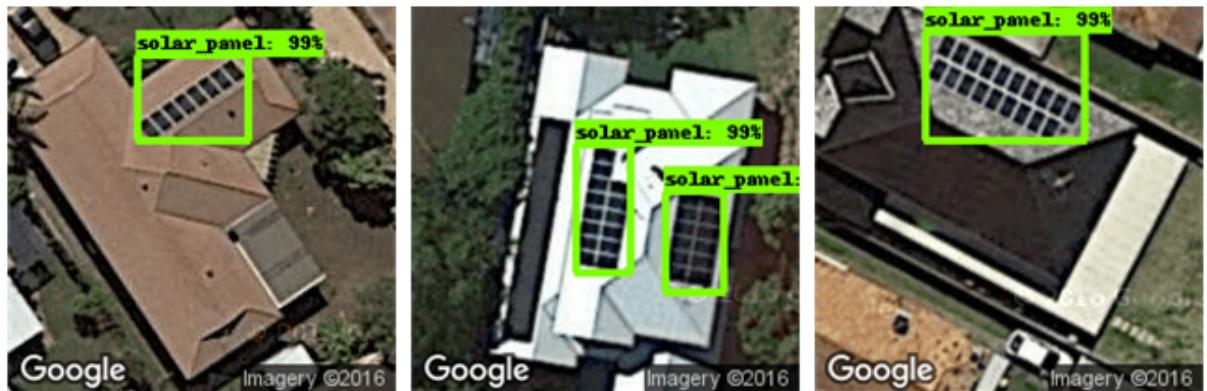


Рисунок 4.10 – Результаты детекции для изображений из тестовой выборки

критерием при выборе клиентом того или иного продукта, а разнообразие позволяет охватить различные группы потенциальных покупателей. Системы, которые автоматизируют процесс проверки качества и при этом поддерживают многообразие продукции, выпускаемой на большом предприятии, имеют особую ценность. При этом процесс проверки качества осуществляется не только для самого продукта, но и для той упаковки и маркировки, которую видит покупатель. Так как упаковка формирует первоначальное впечатление о товаре, ее качество является одной из причин, которая дает покупателю основание для покупки. Маркировка как

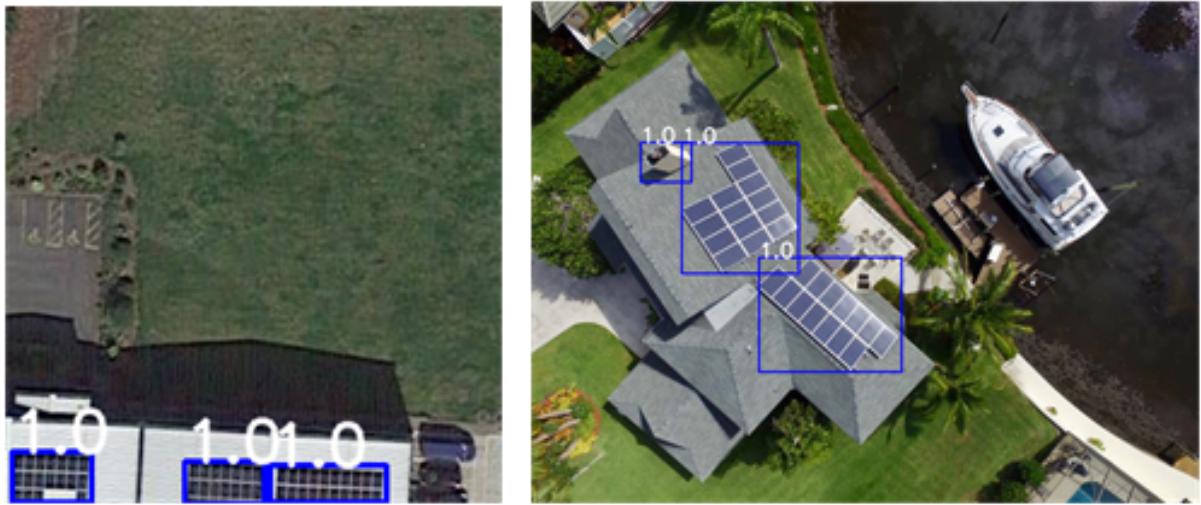


Рисунок 4.11 – Результаты детекции для произвольных изображений

элемент упаковки, гарантирует покупателю сохранность продукции в течение указанного срока при соблюдении условий хранения.

В последнее время методы искусственного интеллекта в целом и машинного обучения в частности широко используются в промышленных системах, устанавливаемых на предприятиях для контроля за процессом производства. Интеллектуальные подсистемы позволяют упростить многие рутинные операции, проводимые для поддержания качества готовой продукции. Например, контроль за правильностью нанесения маркировки ранее производился исключительно оператором-человеком. Сейчас, с развитием теории компьютерного зрения, получившей значительный рывок благодаря постоянно развивающейся области обучения глубоких нейронных сетей, становится особенно актуальной разработка систем, позволяющих осуществлять рутинные операции быстрее и регулярнее, чем это мог бы сделать человек.

Поставленная задача заключалась в разработке решения для распознавания маркировки продуктов, производимых ОАО «Савушкин продукт». Пример продукта с маркировкой, распознавание которой выполняется, представлен на рисунке 4.12.

Важным аспектом является то, что распознавание должно выполняться в реальном времени, основываясь на данных, поступающих с камеры, установленной над производственной линией. Результаты производимого распознавания используются для контроля корректности маркировки.

К основным проблемам, которые могут возникать в процессе маркирования продуктов, относятся [24–A]:



Рисунок 4.12 – Продукт с алфавитно-цифровой маркировкой

1) **отсутствие чернил:** в случае поступления на конвейер продуктов без маркировки, система должна сделать вывод об отсутствии чернил, опционально запросив проверку, обратившись к принтеру;

2) **сдвиг камеры:** если от нейросетевых модулей не поступают данные о результатах распознавания, но системе известно, что движение по конвейеру началось, то она должна сделать вывод о сдвиге камеры;

3) **ошибочная маркировка:** маркировка была обнаружена и распознана, но не совпала с эталонным представлением. В этом случае должен быть сделан вывод о том, что маркировка неверная;

4) **нечитаемая маркировка:** в случае, если маркировка получается смазанной и не может быть распознана, необходимо остановить конвейер и сообщить об ошибке оператору.

Для 1,3 и 4 проблемы необходимо выполнить отсев продуктов, которые имеют “проблемную” маркировку. Возникновение этих проблем предполагает полную остановку движения конвейера и сообщение оператору о возникшей проблеме.

Следует отметить, что разработанная система является частью более общей системы, которая используется для анализа и обработки указанных проблем в процессе нанесения маркировки. Общая система реализована с использованием отечественной технологии OSTIS [73, с. 15].

Несмотря на существующий интерес к автоматизации производственных процессов и неоспоримые преимущества, которые влечет ее внедрение, подобные задачи решаются в большинстве случаев с участием человека. Оператор периодически выборочно проверяет часть продукции. У такого подхода есть недостатки:

– есть вероятность того, что будет пропущен момент, когда появится дефект маркировки;

- скорость реакции человека на возникающую нештатную ситуацию может быть недостаточной;
- человек может не заметить небольшое расхождение проверяемой маркировки с эталонной (например, в случае ошибочной цифры в дате или номере партии);
- работа по ручной проверке является монотонной.

Существующие аппаратные разработки базируются на использовании специальных датчиков [74].

Такие решения осуществляют распознавание маркировки, но имеют ряд важных недостатков:

- нестабильное качество распознавания, зависящее от условий, при которых производится съемка (в частности, от освещенности). Так как производственная линия движется быстро, то необходимые условия для качественного распознавания чаще всего не соблюдаются;
- необходимость покупки специализированного программного обеспечения для настройки датчиков.

Таким образом, появляется необходимость осуществлять контроль за функционированием самой системы распознавания.

В процессе решения указанной задачи были определены факторы, имеющие критическое влияние на процесс решения и подбор алгоритмов:

- **высокая скорость видеопотока.** Так как скорость видеопотока составляет 76 кадров в секунду, то время обработки каждого кадра составляет около 13 миллисекунд. Следует отметить, что этого времени недостаточно для запуска сложной нейросетевой архитектуры, а также для обработки каждого кадра потока;
- **сложность корректной прямой детекции символов (цифр).** Помимо символов, содержащихся непосредственно в маркировке, в кадр могут попадать символы, нанесенные на другие объекты, например, на сам конвейер или его части. Помимо этого, следует отметить, что изображение попадает на нейронную сеть с уменьшенным разрешением, что приводит к сложности распознавания очень мелких объектов.

4.3.2 Предлагаемый подход

Предлагаемый подход состоит в использовании конвейерной структуры из отдельных нейросетевых модулей, каждый из которых решает собственную подзадачу распознавания маркировки.

Задачей данного конвейера является обнаружение маркировки, определение ее типа и ее распознавание.

Остановимся далее подробнее на архитектуре предлагаемого решения.

Осуществляя декомпозицию решаемой задачи можно выделить следующие подзадачи:

- 1) оценка положения товара;
- 2) детекция продукта и маркировки;
- 3) распознавание частей маркировки;
- 4) «сборка» маркировки: формирование выходной информации (даты производства товара, номера в партии и т.д.);
- 5) проверка распознанной маркировки (определение корректности распознанных данных в соответствии с определенным шаблоном).

В процессе распознавания маркировки решаются некоторые дополнительные задачи, заключающиеся в проверке корректности печати:

- 1) определение отсутствия маркировки на продукте;
- 2) определение присутствия искажений маркировки (отсутствия ее частей).

Указанные задачи решаются в процессе выполнения основных этапов распознавания.

Проблемы, указанные в постановке задачи, решаются архитектурно.

1) высокая скорость видеопотока: решается пропуском незначащих кадров, в которых товар находится не в середине кадра. Это позволяет увеличить интервал времени, необходимый для обработки изображения нейросетью. Оценка значимости осуществляется простой моделью-классификатором с малым временем отработки;

2) сложность корректной прямой детекции символов: решается осуществлением декомпозиции задачи обнаружения на отдельные подзадачи. В предлагаемой системе в начале обнаруживается товар, затем маркировка на товаре и, наконец, отдельные символы (цифры).

В предлагаемой нейросетевой системе практически для каждой подзадачи используется отдельная нейросетевая модель, что позволяет легко модифицировать систему, улучшать отдельные модули, изменять их, а также добавлять новые.

Архитектура системы распознавания представлена на рисунке 4.13.

Опишем основные применяемые нейросетевые модели и их роль в общей архитектуре.

Классификатор положения продукта. Используется предобучаемый сверточный классификатор, который определяет значимость текущего кадра для возможности проведения последующего анализа. При этом наиболее значимым является кадр, в котором товар находится ближе все-

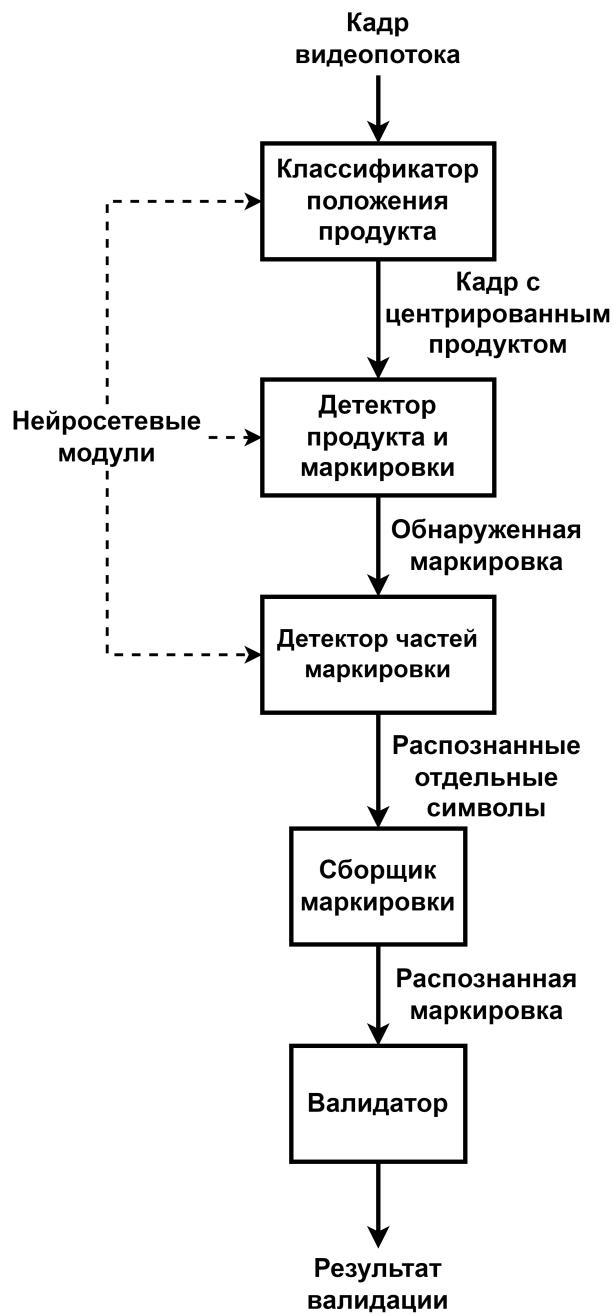


Рисунок 4.13 – Структура системы распознавания маркировки

го к центру кадра (рисунок 4.14). Было выделено 4 класса, представляющих основные позиции продукта в кадре. Класс 1 описывает минимальную удаленность продукта от центра кадра. Кадры, отнесенные к этому классу отбираются для последующих этапов обработки и анализа. Классы 2 и 3 описывают разные степени удаленности (условно среднюю и максимальную). Класс 4 используется для случая отсутствия продукта в кадре (пустая производственная линия). Кадры, которые идентифицированы как принадлежащие классам 2-4, в дальнейшем анализе не участвуют.

Архитектура применяемого классификатора представлена на рисунке 4.15 [4-А]. Он состоит из 5 слоев и имеет 4 выходных нейрона по числу

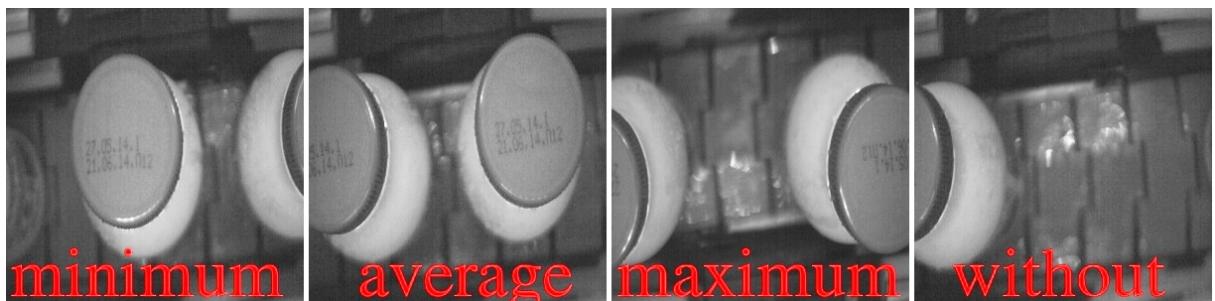


Рисунок 4.14 – Примеры изображений разных классов по степени удаленности объекта от центра кадра

классов, определяющих положение товара в кадре. На всех слоях используется функция активации ReLU за исключением 3-го и последнего слоев. Они используют линейную и softmax-функции активации соответственно. Также применяется max pooling после первого и второго сверточного слоев с параметром stride, равным 2.

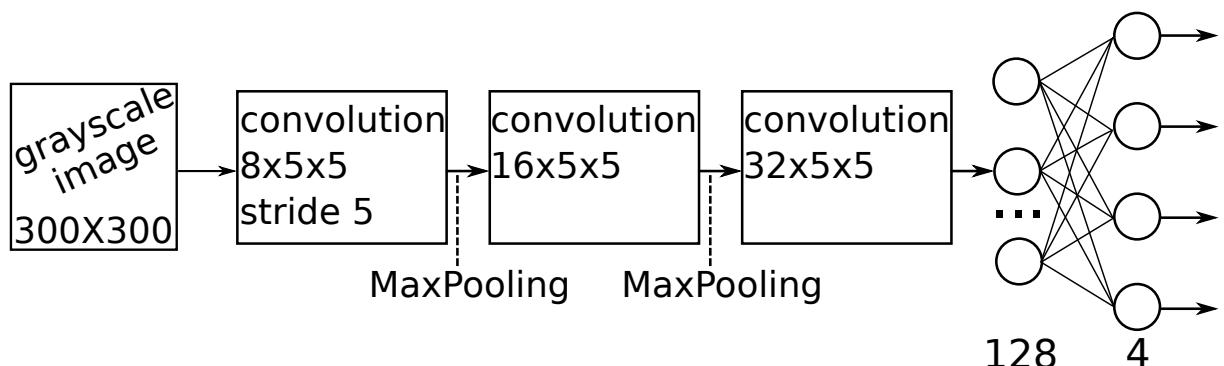


Рисунок 4.15 – Структура классификатора для оценки положения бутылки [24–A]

Детектор продукта и маркировки. Данный детектор осуществляет поиск товара и маркировки в кадре. Здесь в качестве архитектуры была выбрана сеть SSD [75, с. 2-4] на базе классификатора MobileNet v1 [76, с. 4].

Независимое обнаружение товара и маркировки позволяет идентифицировать ситуацию с отсутствующей маркировкой автоматически. Для этого проверяется логическое условие отсутствия маркировки при наличии самого товара в кадре.

Следует отметить, что данный детектор может применяться для обнаружения разных типов маркировки.

В итоге, если маркировка была обнаружена, выполняется передача ее изображения (в оригинальном размере) следующей модели для обработки.

Детектор частей маркировки. Для анализа алфавитно-цифровой маркировки также используется детектор SSD-MobileNet v1, который осу-

ществляет обнаружение отдельных цифр маркировки.

После отработки детектора составных частей осуществляется «сборка» распознанной маркировки и обработка результата.

Принцип работы нейросетевого компонента на примере алфавитно-цифровой маркировки продемонстрирован на рисунке 4.16.

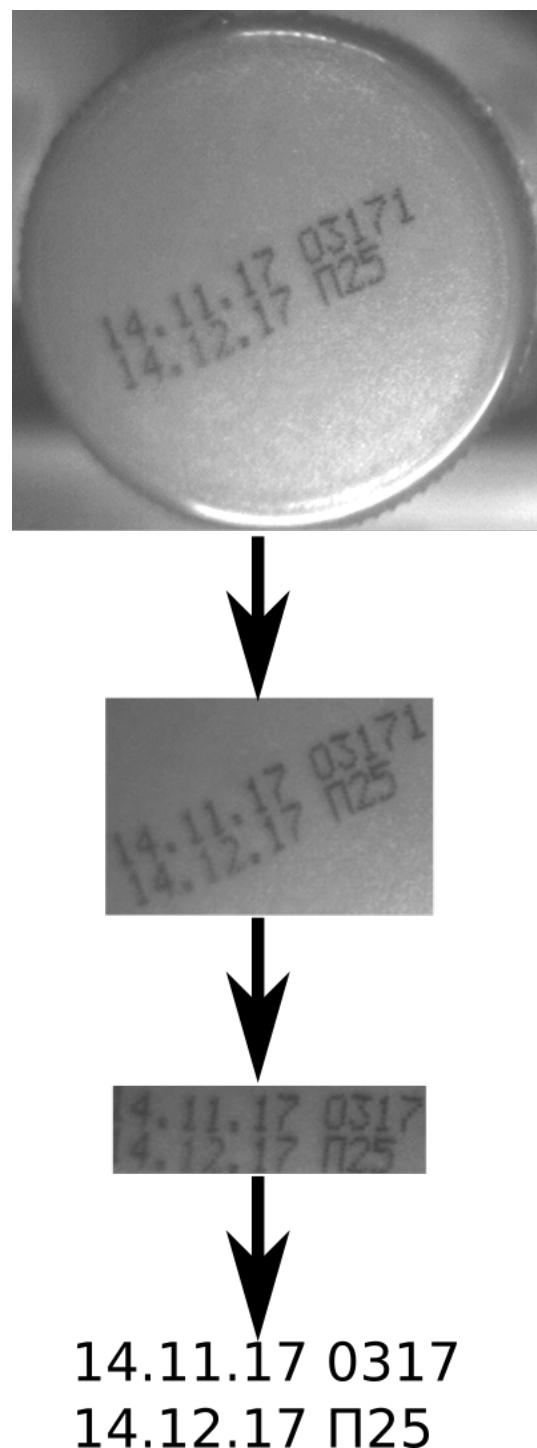


Рисунок 4.16 – Этапы анализа изображения

4.3.3 Обучающие выборки

В процессе подготовки нейросетевых моделей использовались различные обучающие выборки:

- выборка для обучения классификатора положения;
- выборка для обучения детектора маркировок и товаров;
- выборка для обучения детектора частей маркировки.

Выборка для обучения классификатора положения продукта. Для создания выборки использовалась модель Faster R-CNN [70, с. 3] (на базе предобученного классификатора ResNet50 [77, с. 6]). Данная модель обладает лучшими показателями эффективности, чем SSD-MobileNet, но уступает ей в скорости. Она использовалась для автоматической разметки имеющихся данных (главным образом видеофайлов производственного процесса) по степени удаленности товара от центра кадра. В качестве метрики применялось евклидово расстояние от центра продукта до центра кадра. Таким образом были сформированы четыре класса изображений, которые использовались для последующего обучения сверточного классификатора. Общий объем выборки составил 6189 изображений, 1303 из которых составили тестовую выборку.

Выборка для обучения детектора продукта и маркировки. Для формирования этой выборки использовались размеченные вручную изображения. Общий объем выборки составил 815 изображений, 163 из которых составили тестовую выборку.

Выборка для обучения детектора частей маркировки. Для создания этой выборки применялась выборка SVHN [78, с. 3] (выборка номеров домов), а также размеченные цифровые маркировки из уже сформированных выборок. Использовался вариант выборки SVHN, включающий 33402 изображения в обучающей части и 13068 в тестовой (рисунок 4.17). Объем выборки размеченных цифровых маркировок составил 419 изображений.



Рисунок 4.17 – Пример изображений из выборки SVHN

4.3.4 Результаты обучения и тестирования

После обучения классификатора положения продукта итоговая эффективность распознавания составила **93.27%**. Для обучения данной модели применялось предобучение по предлагаемому методу REBA. Для классификатора, обученного без использования предобучения эффективность составила **88,03%**, таким образом, использование предобучения позволило увеличить эффективность распознавания на **5,24%** на исследуемой выборке.

Оба детектора (продуктов/маркировок и частей маркировки) обучались на основе предобученных моделей (использовалось предобучение I типа).

Применение SSD-модели позволило достичь эффективности детекции в **99% ($mAP = 0.99$)** для обнаружения товара и маркировки и **92% ($mAP = 0.92$)** для отдельных цифр. Кроме этого, высокая скорость обработки позволила успешно обнаруживать маркировку в видеопотоке. Результаты эффективности распознавания отдельных цифр представлены в таблице 4.2.

Таблица 4.2 – Эффективность обнаружения отдельных классов цифр

Class label	AP
0	0.9218
1.	0.9107
2.	0.9354
3.	0.9286
4.	0.9265
5.	0.9137
6.	0.9274
7.	0.9167
8.	0.9646
9.	0.8975
mAP	0.92429

Результаты работы детектора продукта и маркировки и детектора частей маркировки изображены на рисунках 4.18 и 4.19.

Помимо алфавитно-цифровой маркировки (рисунок 4.12), начиная с недавнего времени [79], продукция ООО «Савушкин Продукт» выпускается с вариантами маркировки, включающей код Data Matrix (рисунок 4.20) [80]. Данный тип маркировки является удобным и емким представлением специальных и общих данных о продукте.

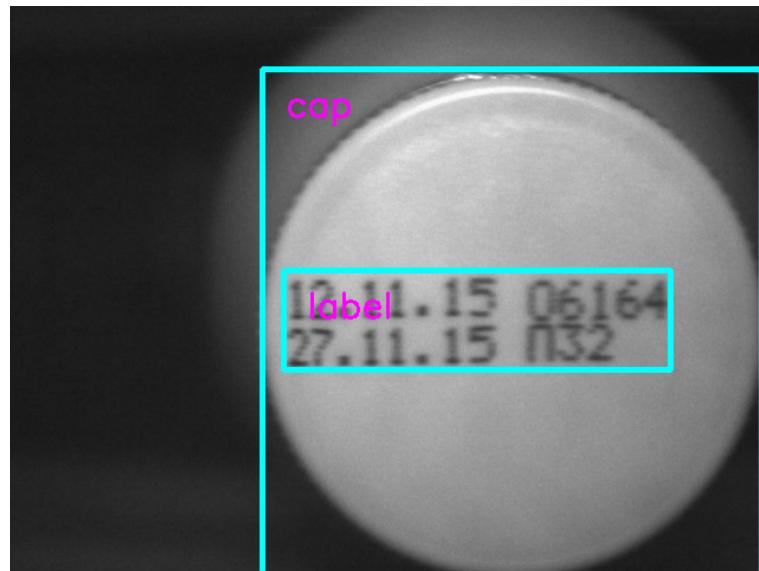


Рисунок 4.18 – Обнаруженный продукт и маркировка

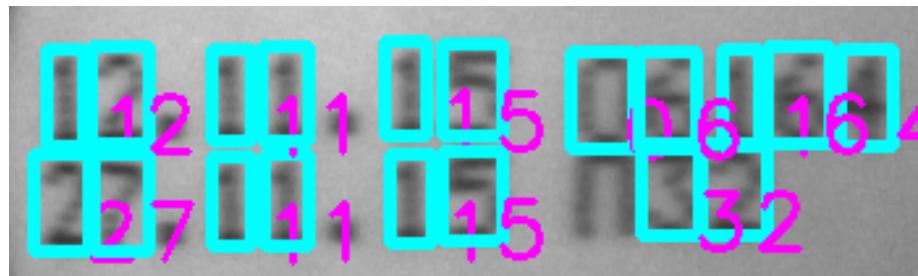


Рисунок 4.19 – Обнаруженные цифры в маркировке



Рисунок 4.20 – Продукт с кодом Data Matrix

Внедрение нового типа маркировки не снижает актуальность предлагаемой разработки, поскольку алфавитно-цифровая маркировка остается единственным вариантом маркировки, который может быть распознан непосредственно человеком. Помимо этого, он продолжает использоваться на всей продукции.

Отметим также, что расширение разработанной системы для обработки новых типов маркировки возможно путем дообучения детектора маркировок на новых данных. Это позволяет гибко расширять систему

без необходимости ее повторной реализации.

4.4 Выводы

1) Предложено решение задачи определения наличия и детекции солнечных панелей на аэрофотоснимках. Достоинством предлагаемого решения является возможность работы с изображениями, имеющими низкое разрешение. Использование предобученных сверточных нейронных сетей позволило увеличить эффективность на 7,31% в определении наличия солнечных панелей на фото [13–A, с. 383-385; 19–A, с. 15-18; 22–A].

2) Предложено решение задачи распознавания маркировки продукта на производственной линии. Достоинствами предлагаемого решения является модульная структура, позволяющая осуществлять независимую поддержку каждого функционального модуля. Использование предобученного нейросетевого классификатора позволило увеличить эффективность на 5,24% для оценки положения продукта на изображении. Помимо этого, локализация маркировки и продукта была выполнена с точностью 99%, а обнаружение отдельных составляющих маркировки (цифр) – с точностью 92% [4–A; 6–A, с. 223; 10–A, с. 369-370; 11–A, с. 94-96; 12–A, с. 178-179; 20–A, с. 135-136; 21–A; 23–A, с. 44-50; 24–A; 25–A, с. 89-94].

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Установлена эквивалентность задач максимизации функции правдоподобия распределения входных данных, минимизации суммарной квадратичной ошибки сети при использовании линейных нейронов и минимизации кросс-энтропийной функции ошибки сети в пространстве синаптических связей ограниченной машины Больцмана. Из полученных теоретических результатов следует, что природа неконтролируемого обучения в RBM-сети является идентичной при использовании различных целевых функций [1–A, 2–A, 3–A, 5–A, 9–A, 16–A, 18–A].

2. Разработан метод обучения ограниченной машины Больцмана, базирующийся на минимизации квадратичной ошибки сети в скрытом и видимом слоях, что позволяет учитывать нелинейную природу нейронных элементов. Разработанный метод применен для обучения глубоких полно связных и сверточных архитектур нейронных сетей и протестирован на выборках MNIST, CIFAR-10, CIFAR-100. Показано, что предложенный метод обладает большей эффективностью, чем классический [1–A, 2–A, 3–A, 5–A, 9–A, 14–A, 15–A, 16–A, 17–A, 18–A, 27–A, 28–A, 29–A, 30–A].

3. Предложен алгоритм редуцирования параметров глубокой нейронной сети, базирующийся на неконтролируемом предобучении сети и позволяющий сократить количество настраиваемых параметров сети и упростить ее архитектуру без потери обобщающей способности. Проведены вычислительные эксперименты, доказывающие эффективность предложенного метода [7–A, 8–A, 26–A].

4. Разработана нейросетевая система компьютерного зрения, основанная на предлагаемых методах предобучения и редуцирования, с помощью которой решены прикладные задачи компьютерного зрения, а именно:

- задача обнаружения солнечных панелей на аэрофотоснимках – предлагаемое решение позволяет повысить эффективность при определении наличия солнечной панели на 7,31% с последующей локализацией солнечной панели [13–A, 19–A, 22–A];
- задача распознавания маркировки продукта на производственной линии – предлагаемое решение позволяет повысить эффективность при оценке положения продукта с маркировкой в кадре на 5,24% с последующей локализацией маркировки с точностью 99% и обнаружением ее отдельных составляющих (цифр) с точностью 92%. Предлагаемое

решение позволяет осуществлять обработку изображений в реальном времени [4–A, 6–A, 10–A, 11–A, 12–A, 20–A, 21–A, 23–A, 24–A, 25–A].

Рекомендации по практическому использованию результатов

Научные и практические результаты диссертационной работы использованы в научно-исследовательских работах, учебном процессе, а также в ряде прикладных систем.

Разработанные алгоритмы обнаружения и локализации солнечных панелей на аэрофотоснимках, методы предобучения нейросетевых моделей и инструментальные средства внедрены и используются в компании ООО «Intelligent Semantic Systems» при разработке интеллектуальных систем в составе подсистем компьютерного зрения.

Предложенный метод предобучения глубоких нейронных сетей применялся при обучении модели классификатора кадров для нейросетевой системы распознавания маркировки продукта на производственной линии ОАО «Савушкин продукт».

Кроме того, научные и практические результаты диссертационной работы используются в учебном процессе учреждения образования «Брестский государственный технический университет».

Реализация программных модулей осуществлялась с использованием свободного ПО и может быть непосредственно применена при разработке отечественных проектов в области компьютерного зрения без необходимости приобретения дорогостоящих программных средств.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

Библиографический список

1. Mitchell, T. M. Machine learning / T. M. Mitchell. — McGraw-hill New York, 1997.
2. Кандель, Э. В поисках памяти: Возникновение новой науки о человеческой психике / Э. Кандель. — Москва: М.: Астрель: CORPUS, 2012.
3. Головко, В. А. Нейросетевые технологии обработки данных / В. А. Головко, В. В. Краснопрошин. — Минск: БГУ, 2017.
4. Cybenko, G. V. Approximation by superpositions of a sigmoidal function / G. V. Cybenko // Mathematics of Control, Signals and Systems. — 1989. — Vol. 2. — P. 303–314.
5. Ivakhnenko, A. G. Cybernetics and Forecasting Techniques / A. G. Ivakhnenko, V. G. Lapa. Modern analytic and computational methods in science and mathematics. — American Elsevier Publishing Company, 1967. <https://books.google.by/books?id=z7hPAQAAIAAJ>.
6. Elman, J. L. Finding structure in time / J. L. Elman // Cogn. Sci. — 1990. — Vol. 14. — P. 179–211.
7. Jordan, M. I. Serial order: A parallel distributed processing approach: Tech. Rep. ICS Report 8604 / M. I. Jordan: Institute for Cognitive Science, University of California, San Diego, 1986.
8. Hinton, G. E. Learning representations by recirculation / G. E. Hinton, J. L. McClelland // NIPS. — 1987.
9. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position / K. Fukushima // Biological Cybernetics. — Vol. 36. — 1980. — P. 193–202.
10. Handwritten digit recognition with a back-propagation network / Y. LeCun, B. E. Boser, J. S. Denker [et all] // NIPS. — 1989.
11. Kohonen, T. Self-organizing maps / T. Kohonen. — Berlin: Springer, 2001.
12. Grossberg, S. Competitive learning: From interactive activation to adaptive resonance / S. Grossberg // Cogn. Sci. — 1987. — Vol. 11. — P. 23–63.
13. Hopfield, J. Neurons with graded response have collective computational properties like those of two-state neurons. / J. Hopfield // Proceedings of the National Academy of Sciences. — 1984. — Vol. 81, № 10. — P. 3088–3092.

14. Lippmann, R. An introduction to computing with neural nets / R. Lippmann // IEEE ASSP Magazine. — 1987. — Vol. 4, № 2. — P. 4–22.
15. Kosko, B. Bidirectional associative memories / B. Kosko // IEEE Trans. Syst. Man Cybern. — 1988. — Vol. 18. — P. 49–60.
16. Hecht-Nielsen, R. Counterpropagation networks / R. Hecht-Nielsen // Appl. Opt. — 1987. — Vol. 26, № 23. — P. 4979–4984.
17. Broomhead, D. Radial basis functions, multi-variable functional interpolation and adaptive networks / D. Broomhead, D. Lowe // Royal Signals and Radar Establishment Malvern (United Kingdom). — 1988. — Vol. RSRE-MEMO-4148.
18. Jang, J.-S. Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [book review] / J.-S. Jang, C.-T. Sun, E. Mizutani // Automatic Control, IEEE Transactions on. — 1997. — Vol. 42. — P. 1482 – 1484.
19. Neural Network and Artificial Immune Systems for Malware and Network Intrusion Detection / Vladimir Golovko, Sergei Bezobrazov, Pavel Kachurka, Leonid Vaitsekhovich // Advances in Machine Learning II: Dedicated to the Memory of Professor Ryszard S.Michalski. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. — P. 485–513.
20. LeCun, Y. Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. — 2015. — № 521 (7553). — P. 436–444.
21. An introductory review of deep learning for prediction models with big data / Frank Emmert-Streib, Zhen Yang, Han Feng [et al] // Frontiers in Artificial Intelligence. — 2020. — Vol. 3.
22. Gradient-based learning applied to document recognition / Y. LeCun, L. Bottou, Y. Bengio, P. Haffner // Proceedings of the IEEE. — Vol. 86. — 1998. — P. 2278–2324.
23. Nerf: Representing scenes as neural radiance fields for view synthesis / B. Mildenhall, P. P. Srinivasan, M. Tancik [et al] // ECCV. — 2020.
24. Головко, В. А. От многослойных персепtronов к нейронным сетям глубокого доверия: парадигмы обучения и применение / В. А. Головко // Лекции по Нейроинформатике. — НИЯУ МИФИ, 2015. — С. 47–84.
25. Bengio, Y. Learning deep architectures for ai / Y. Bengio // Foundations and Trends in Machine Learning. — 2009. — № 2(1). — P. 1–127.
26. Hinton, G. Reducing the dimensionality of data with neural networks / G. Hinton, R. Salakhutdinov // Science. — 2006. — № 313 (5786). — P. 504–507.

27. McCulloch, W. A logical calculus of ideas immanent in nervous activity / W. McCulloch, W. Pitts // Bulletin of Mathematical Biophysics. — 1943. — Vol. 5. — P. 127–147.
28. Hebb, D. O. The organization of behavior: A neuropsychological theory / D. O. Hebb. — New York: Wiley, 1949.
29. Widrow, B. Adaptive switching circuits / B. Widrow, M. E. Hoff // 1960 IRE WESCON Convention Record, Part 4. — New York: IRE, 1960. — P. 96–104.
30. Rosenblatt, F. Principles of neurodynamics: perceptrons and the theory of brain mechanisms / F. Rosenblatt. — Washington: Spartan Books, 1962.
31. Minsky, M. Perceptrons: An Introduction to Computational Geometry / M. Minsky, S. Papert. — Cambridge, MA, USA: MIT Press, 1969.
32. Grossberg, S. Adaptive pattern classification and universal recoding: II. feedback, expectation, olfaction, illusions / S. Grossberg // Biol. Cybern. — 1976. — Vol. 23, № 4. — P. 187–202.
33. Kohonen, T. Associative memory: a system-theoretical approach / T. Kohonen. — Springer-Verlag Berlin ; New York, 1977. — P. 176.
34. Rumelhart, D. E. Learning representations by back-propagating errors / D. E. Rumelhart, G. E. Hinton, R. J. Williams // nature. — 1986. — Vol. 323, № 6088. — P. 533–536.
35. Cortes, C. Support-vector networks / C. Cortes, V. Vapnik // Mach. Learn. — 1995. — Vol. 20, № 3. — P. 273–297.
36. Behnke, S. Hierarchical Neural Networks for Image Interpretation / S. Behnke. Lecture Notes in Computer Science. — Springer Berlin, Heidelberg, 2003.
37. Hinton, G. E. A fast learning algorithm for deep belief nets / G. E. Hinton, S. Osindero, Y. Teh // Neural Computation. — 2006. — № 18. — P. 1527–1554.
38. Hinton, G. Greedy layer-wise algorithm / G. Hinton // Journal of Machine Learning Research. — 2009. — № 1. — P. 1–40.
39. Glorot, X. Deep sparse rectifier networks / X. Glorot, A. Bordes, Y. Bengio // Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. — Vol. 15. — 2011. — P. 315–323.
40. Taori, R. Alpaca: A strong, replicable instruction-following model. — Mode of access: <https://crfm.stanford.edu/2023/03/13/alpaca.html>. — 2023. — Date of access: 05.03.2023.

41. Taori, R. Stanford alpaca: An instruction-following llama model. — Mode of access: https://github.com/tatsu-lab/stanford_alpaca. — 2023. — Date of access: 05.03.2023.
42. Zhang, S. Opt: Open pre-trained transformer language models. — 2022.
43. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль. — Москва : ДМК Пресс, 2017. — 652 с.
44. Хайкин, С. Нейронные сети: полный курс, 2-е издание / С. Хайкин. — Москва: Издательский дом «Вильямс», 2006. — 1104 с.
45. Duchi, J. Adaptive subgradient methods for online learning and stochastic optimization / J. Duchi, E. Hazan, Y. Singer // Journal of Machine Learning Research. — 2011.
46. Hinton, G. Lecture 6d – a separate, adaptive learning rate for each connection. slides of lecture neural networks for machine learning. — Mode of access: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf. — 2012. — Date of access: 2023-03-05.
47. Kingma, D. Adam: A method for stochastic optimization / D. Kingma, J. Ba // 3rd International Conference for Learning Representations. — 2014.
48. Hinton, G. Training products of experts by minimizing contrastive divergence / G. Hinton // Neural Computation. — 2002. — № 14. — P. 1771–1800.
49. The “wake-sleep” algorithm for unsupervised neural networks / G. E. Hinton, P. Dayan, B. J. Frey, R. M. Neal // Science. — 1995. — Vol. 268, № 5214. — P. 1158–1161.
50. Greedy layer-wise training of deep networks / Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle // Advances in neural information processing systems. — MA: MIT Press, Cambridge, 2007. — P. 153–160.
51. Using different cost functions to train stacked auto-encoders / Telmo Amaral, Luís M. Silva, Luís A. Alexandre [et al] // 2013 12th Mexican International Conference on Artificial Intelligence. — 2013. — P. 114–120.
52. Gibbs, J. W. Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics / J. W. Gibbs. — C. Scribner’s sons, 1902.
53. Hinton, G. E. A practical guide to training restricted boltzmann machines: Tech. Rep. 2010-000 / G. E. Hinton: University of Toronto, 2010.

54. Quantized neural networks: Training neural networks with low precision weights and activations / I. Hubara, M. Courbariaux, D. Soudry [et al] // arXiv. — 2016. — Vol. 1609.07061.
55. Hinton, G. E. Distilling the knowledge in a neural network / G. E. Hinton, O. Vinyals, J. Dean // ArXiv. — 2015. — Vol. abs/1503.02531.
56. Low-rank matrix factorization for deep neural network training with high-dimensional output targets / T. N. Sainath, B. Kingsbury, V. Sindhwani [et al] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. — 2013. — P. 6655–6659.
57. Pruning from scratch / Y. Wang, X. Zhang, L. Xie [et al] // arXiv. — 2019. — Vol. 1909.12579.
58. Fisher, R. The use of multiple measurements in taxonomic problems / R. Fisher // Annals of Eugenics. — 1936. — № 7. — P. 179–188.
59. LeCun, Y. The mnist database of handwritten digits. — Mode of access: <http://yann.lecun.com/exdb/mnist/>. — Date of access: 2023-03-05.
60. Krizhevsky, A. Learning multiple layers of features from tiny images. — Mode of access: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. — 2009. — Date of access: 2023-03-05.
61. Perry, C. Colab's 'pay as you go' offers more access to powerful nvidia compute for machine learning. — Mode of access: <https://blog.tensorflow.org/2022/09/colabs-pay-as-you-go-offers-more-access-to-powerful-nvidia-compute-for-machine-learning.html>. — Date of access: 2023-03-05.
62. Turchenko, V. Parallel batch pattern training algorithm for deep neural network / V. Turchenko, V. Golovko // High Performance Computing & Simulation (HPCS). — IEEE, 2014. — P. 697–702.
63. Scholz, M. Nonlinear principal component analysis: neural network models and applications / M. Scholz, M. Fraunholz, J. Selbig. — Springer, 2008. — P. 44–67.
64. Torralba, A. 80 million tiny images: A large data set for nonparametric object and scene recognition / A. Torralba, R. Fergus, W. T. Freeman // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2008. — Vol. 30, № 11. — P. 1958–1970.
65. Solar industry data // seia.org. — Mode of access: <https://www.seia.org/solar-industry-research-data>. — 2020. — Date of access: 05.03.2023.
66. Automatic solar photovoltaic panel detection in satellite imagery / J. Malof, R. Hou, L. Collins [et al] // International Conference on Renewable Energy Research and Applications ICRERA' 2015. — 2015.

67. Automatic detection of solar photovoltaic arrays in high resolution aerial imagery / J. Malof, K. Bradbury, L. Collins, R. Newell // Applied Energy. — Vol. 183. — 2016. — P. 229–240.
68. Bradbury, K. Distributed solar photovoltaic array location and extend data set for remote sensing object identification. — Mode of access: <http://dx.doi.org/10.6084/m9.figshare.3385780.v1>. — 2016. — Date of access: 05.03.2023.
69. Lin, T. Microsoft coco: Common objects in context. — Mode of access: <https://arxiv.org/pdf/1405.0312.pdf>. — 2015. — Date of access: 05.03.2023.
70. Ren, S. Faster r-cnn: Towards real-time object detection with region proposal networks. — Mode of access: <https://arxiv.org/abs/1506.01497>. — 2016. — Date of access: 05.03.2023.
71. Нейросетевые модели детекции товаров на изображении / В. А. Головко, А. А. Крощенко, Е. В. Михно, О. Ю. Войцехович // Вестник Брестского государственного технического университета. Физика, математика, информатика. — 2018. — № 5. — С. 27–30.
72. The pascal visual object classes (voc) challenge / Mark Everingham, Luc Gool, Christopher K. Williams [et al] // Int. J. Comput. Vision. — 2010. — Vol. 88, № 2. — P. 303–338.
73. Голенков, В. В. Технология комплексной поддержки жизненно-го цикла семантически совместимых интеллектуальных компьютерных систем нового поколения / В. В. Голенков. — Беспринт, 2023.
74. Omron documentation. — Mode of access: <https://robotics.omron.com/browse-documents>. — Date of access: 05.03.2023.
75. Liu, W. Ssd: Single shot multibox detector. — Mode of access: <https://arxiv.org/abs/1512.02325>. — 2015. — Date of access: 05.03.2023.
76. Howard, A. Mobilenets: Efficient convolutional neural networks for mobile vision applications. — Mode of access: <https://arxiv.org/abs/1704.04861>. — 2017. — Date of access: 05.03.2023.
77. He, K. Deep residual learning for image recognition. — Mode of access: <https://arxiv.org/abs/1512.03385>. — 2015. — Date of access: 05.03.2023.
78. Reading digits in natural images with unsupervised feature learning / Y. Netzer, T. Wang, A. Coates [et al] // NIPS Workshop on Deep Learning and Unsupervised Feature Learning. — 2011.
79. Milk products – electronic sign. — Режим доступа: <https://datamark.by/markirovka-molochnoy-produktsii>. — Дата доступа: 05.03.2023.
80. Gs1 datamatrix – electronic sign. — Mode of access: <https://datamark.by/voprosy-otvety/gs1-datamatrix>. — Date of access: 05.03.2023.

Список публикаций соискателя ученой степени

Статьи в научных рецензируемых изданиях, включенных в перечень издаий, и в иностранных научных изданиях

1–А. Головко, В.А. Персептроны и нейронные сети глубокого доверия : обучение и применение / В.А. Головко, А.А. Крощенко // Вестник Брестского государственного технического университета. Физика, математика, информатика. — Брест, 2014. — № 5 (89). — С. 2–12.

2–А. Golovko, V. The Nature of Unsupervised Learning in Deep Neural Networks : A New Understanding and Novel Approach / V. Golovko, A. Kroshchanka, D. Treadwell // Optical Memory and Neural Networks. — 2016. — Vol. 25, № 3. — P. 127–141.

3–А. Головко, В.А. Теория глубокого обучения : конвенциальный и новый подход / В.А. Головко, А.А. Крощенко, М.В. Хацкевич // Вестник Брестского государственного технического университета. Физика, математика, информатика. — Брест, 2016. — № 5 (101). — С. 7–16.

4–А. Крощенко, А.А. Реализация нейросетевой системы распознавания маркировки продукции / А.А. Крощенко, В.А. Головко // Вестник Брестского государственного технического университета. Физика, математика, информатика. — Брест, 2019. — № 5 (118). — С. 9–12.

5–А. Golovko, V.A. Deep Neural Networks : Selected Aspects of Learning and Application / V.A. Golovko, A.A. Kroshchanka, E.V. Mikhno // Pattern Recognition and Image Analysis. — 2021. — Vol. 31, № 1. — P. 132–143.

6–А. Kroshchanka, A. Neural network component of the product marking recognition system on the production line / A. Kroshchanka, D. Ivaniuk // Open Semantic Technologies for Intelligent Systems : research papers collection / Belar. State Univ. of Informatics and Radioelectr.; ed. : V.V. Golenkov (ed.-in-chief) [et al.]. — Minsk, 2021. — Iss. 5. — P. 219–224.

7–А. Kroshchanka, A.A. Method for Reducing Neural-Network Models of Computer Vision / A.A. Kroshchanka, V.A. Golovko, M. Chodyka // Pattern Recognition and Image Analysis. — Berlin, Heidelberg : Springer-Verlag, 2022. — Vol. 32, № 2. — P. 294–300.

8–А. Kroshchanka, A.A. Reduction of Neural Network Models in Intelligent Computer Systems of a New Generation / A. Kroshchanka // Open Semantic Technologies for Intelligent Systems (OSTIS-2023) : research papers collection / Belar. State Univ. of Informatics and Radioelectr.; ed. : V.V. Golenkov (ed.-in-chief) [et al.]. — Minsk, 2023. — Iss. 7. — P. 127–132.

Статьи в других научных изданиях

9–А. Головко, В.А. Метод обучения нейронной сети глубокого доверия и применение для визуализации данных / В.А. Головко, А.А. Крощенко // Комп’ютерно-інтегровані технології: освіта, наука, виробництво. — Луцьк, 2015. — № 19. — С. 6–12.

10–А. Головко, В.А. Интеграция искусственных нейронных сетей с базами знаний / В.А. Головко, В.В. Голенков, В.П. Иващенко, В.В. Таберко, Д.С. Иванюк, А.А. Крощенко, М.В. Ковалёв // Онтология проектирования. — 2018. — Т. 8. — № 3 (29). — С. 366–386.

11–A. Golovko, V. Principles of decision-making systems building based on the integration of neural networks and semantic models / V. Golovko, A. Kroshchanka, V. Ivashenko, M. Kovalev, V. Taberko, D. Ivaniuk // Open Semantic Technologies for Intelligent Systems (OSTIS-2019) : research papers collection / Belar. State Univ. of Informatics and Radioelectr.; ed. : V.V. Golenkov (ed.-in-chief) [et al.]. — Minsk, 2019. — Iss. 3. — P. 91–102.

12–A. Golovko, V. Implementation of an intelligent decision support system to accompany the manufacturing process / V. Golovko, A. Kroshchanka, M. Kovalev, V. Taberko, D. Ivaniuk // Open Semantic Technologies for Intelligent Systems (OSTIS-2020) : research papers collection / Belar. State Univ. of Informatics and Radioelectr.; ed. : V.V. Golenkov (ed.-in-chief) [et al.]. — Minsk, 2020. — Iss. 4. — P. 175–182.

13–A. Golovko, V. Deep Convolutional Neural Network for Detection of Solar Panels / V. Golovko, A. Kroshchanka, E. Mikhno, M. Komar, A. Sachenko // Data-Centric Business and Applications. ICT Systems-Theory, Radio-Electronics, Information Technologies and Cybersecurity. — Cham: Springer International Publishing, 2021. — P. 371–389.

Статьи в сборниках материалов научных конференций

14–A. Golovko, V. A Learning Technique for Deep Belief Neural Networks / V. Golovko, A. Kroshchanka, U. Rubanau, S. Jankowski // Neural Networks and Artificial Intelligence : proc. of the 8th Internat. Conf. ICNNNAI 2014, Brest, Belarus, June 3-6, 2014 / V. Golovko, A. Imada (eds.). — Springer, 2014. — Vol. 440. Communication in Computer and Information Science. — P. 136–146.

15–А. Головко, В.А. Применение нейронных сетей глубокого доверия для выделения семантически значимых признаков / В.А. Головко, А.А. Крощенко // Открытые семантические технологии проектирования интеллектуальных систем : материалы V Междунар. науч.-технич. конф.

OSTIS-2015, Минск, 19-21 февраля 2015 г. / УО «Бел. гос. ун-т информатики и радиоэлектроники», ГУ «Администрация Парка высоких технологий»; редкол.: В.В. Голенков (отв. ред.) [и др.]. — Минск, 2015. — Р. 481–486.

16—A. Golovko, V. A New Technique for Restricted Boltzmann Machine Learning [Electronic resource] / V. Golovko, A. Kroshchanka, V. Turchenko, S. Jankowski, D. Treadwell // Proceedings of the 8th IEEE International Conference of Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS), Warsaw, Poland, Sept. 24-26, 2015 / Research Institute for Intelligent Computer Systems, Ternopil National Economic University and V.M. Glushkov Inst. of Cybernetics, National Academy for Sciences of Ukraine, Warsaw University of Technology. — P. 182–186. — Mode of access : <https://ieeexplore.ieee.org/document/7340725>. — Date of access : 05.06.2023.

17—A. Крощенко, А.А. Применение нейронных сетей глубокого доверия в интеллектуальном анализе данных / А.А. Крощенко // Современные проблемы математики и вычислительной техники : сб. материалов IX Респ. науч. конф. молодых ученых и студентов, Брест, 19-21 ноября 2015 г. / Мин. обр. Респ. Бел., УО «Брест. гос. технич. ун-т»; редкол.: В.С. Рубанов (гл. ред.) [и др.]. — Брест, 2015. — С. 12–14.

18—A. Golovko, V. Theoretical Notes on Unsupervised Learning in Deep Neural Networks [Electronic resource] / V. Golovko, A. Kroshchanka // Proceedings of the 8th Internat. Joint Conf. on Computational Intelligence (IJCCI 2016), Porto, Portugal, Nov. 9-11 2016. — Р. 91–96. — Mode of access : <https://scitepress.org/papers/2016/60843/60843.pdf>. — Date of access : 05.06.2023.

19—A. Golovko, V. Convolutional Neural Network Based Solar Photovoltaic Panel Detection in Satellite Photos [Electronic resource] / V. Golovko, S. Bezobrazov, A. Kroshchanka, A. Sachenko, M. Komar, A. Karachka // The 9th IEEE Internat. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, Romania, Sept. 21-23 2017. — Р. 14–19. — Mode of access : <https://ieeexplore.ieee.org/document/8094501>. — Date of access : 05.06.2023.

20—A. Golovko, V.A. Integration of artificial neural networks and knowledge bases / V.A. Golovko, A.A. Kroshchanka, V.V. Golenkov, V.P. Ivashenko, M.V. Kovalev, V.V. Taberko, D.S. Ivaniuk // Открытые семантические технологии проектирования интеллектуальных систем : материалы VIII Междунар. науч. технич. конф., Минск, 15-17 февраля 2018 г. / Мин. обр. Респ. Бел., УО «Бел. гос. ун-т информатики и радиоэлектро-

ники»; редкол.: В.В. Голенков (гл. ред.) [и др.]. — Минск, 2018. — Вып. 2. — С. 133–146.

21—А. Головко, В.А. Нейросетевые модели глубокого обучения для решения задач распознавания объектов на изображении / В.А. Головко, А.А. Крощенко // Вычислительные методы, модели и образовательные технологии : сб. материалов VII Междунар. науч.-практич. конф., Брест, 19 октября 2018 г. / Брест. гос. ун-т имени А.С. Пушкина; под общ. ред. А.А. Козинского. — Брест, 2018. — С. 3–5.

22—A. Golovko, V. Development of Solar Panels Detector [Electronic resource] / V. Golovko, A. Kroshchanka, S. Bezobrazov, A. Sachenko, M. Komar, O. Novosad // 2018 International Scientific-Practical Conference «Problems of Infocommunications. Science and Technology (PIC S&T 2018)», Kharkiv, Ukraine, 9-12 October 2018 / Institute of Electrical and Electronics Engineers, Inc. — Mode of access : <https://www.semanticscholar.org/paper/Development-of-Solar-Panels-Detector-Golovko-Kroshchanka/914c0b8c159c64100609fc8455636b1e3f8568cb>. — Date of access : 05.06.2023.

23—A. Golovko, V. Brands and caps labeling recognition in images using deep learning [Electronic resource] / V. Golovko, A. Kroshchanka, E. Mikhno // Pattern Recognition and Information Processing : revised selected papers of the 14th Internat. Conf. PRIP 2019, Minsk, May 21-23 2019. — Eds. : S.V. Ablameyko, V.V. Krasnoproschin, M.M. Lukashevich. — P. 35–51. — Mode of access : https://www.researchgate.net/publication/337459978_Brands_and_Caps_Labeling_Recognition_in_Images_Using_Deep_Learning. — Date of access : 05.06.2023.

24—А. Головко, В.А. Обнаружение и распознавание маркировки продукции с помощью нейросетевых алгоритмов / В.А. Головко, А.А. Крощенко // Вычислительные методы, модели и образовательные технологии: сб. материалов VII Междунар. науч.-практич. конф., Брест, 18 октября 2019 г. / Брест. гос. ун-т имени А.С. Пушкина; под общ. ред. А.А. Козинского. — Брест, 2019. — С. 3–6.

25—A. Golovko, V. Neuro-Symbolic Artificial Intelligence: Application for Control the Quality of Product Labeling [Electronic resource] / V. Golovko, A. Kroshchanka, M. Kovalev, V. Taberko, D. Ivaniuk // Open Semantic Technologies for Intelligent Systems : revised selected papers of the 10th Intern. Conf. (OSTIS 2020), Minsk, Febr. 19-22, 2020 / Eds. : V. Golenkov, V. Krasnoproschin, V. Golovko, E. Azarov. — P. 81–101. — Mode of access : https://libeldoc.bsuir.by/bitstream/123456789/42395/1/Golovko_Neuro_Symbolic.pdf. — Date of access : 05.06.2023.

26—A. Kroshchanka, A. The Reduction of Fully Connected Neural Network Parameters Using the Pre-training Technique / A. Kroshchanka, V. Golovko // Proceedings of the 11th IEEE Intern. Conf. on Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS), Cracow, Poland, September 22-25, 2021 / Cracow University of Technology [et al.]. — Vol. 2. — P. 937–941. — Mode of access : https://www.researchgate.net/publication/357613109_The_Reduction_of_Fully_Connected_Neural_Network_Parameters_Using_the_Pre-training_Technique. — Date of access : 05.06.2023.

Тезисы

27—А. Крощенко, А.А. Методы глубокого обучения нейронных сетей / А.А. Крощенко // Вычислительные методы, модели и образовательные технологии : сб. матер. региональной науч.-практич. конф., Брест, 22-23 окт. 2013 г. / Брест. гос. ун-т им. А.С. Пушкина; под общ. ред. О.В. Матысика. — Брест, 2013. — С. 21–22.

28—А. Головко, В.А. Об одном методе обучения нейронных сетей глубокого доверия / В.А. Головко, А.А. Крощенко // Вычислительные методы, модели и образовательные технологии : сб. матер. Междунар. науч.-практич. конф., Брест, 15-16 окт. 2014 г. / Брест. гос. ун-т им. А.С. Пушкина; под общ. ред. О.В. Матысика. — Брест, 2014. — С. 98–99.

29—А. Головко, В.А. Применение нейронных сетей глубокого доверия в интеллектуальном анализе данных / В.А. Головко, А.А. Крощенко // Вычислительные методы, модели и образовательные технологии : сб. матер. Междунар. науч.-практич. конф., Брест, 22 окт. 2015 г. / Брест. гос. ун-т им. А.С. Пушкина; под общ. ред. О.В. Матысика. — Брест, 2015. — С. 97–98.

30—А. Крощенко, А.А. Применение глубокой нейронной сети для решения задачи распознавания образов / А.А. Крощенко // Вычислительные методы, модели и образовательные технологии : сб. матер. Междунар. науч.-практич. конф., Брест, 21 окт. 2016 г. / Брест. гос. ун-т им. А.С. Пушкина; под общ. ред. О.В. Матысика. — Брест, 2016. — С. 132–133.

ПРИЛОЖЕНИЕ А

Исходный код реализации процедур предобучения и редуцирования

```
import torch
import torch.nn as nn
from common_types import InitTypes

class UnifiedClassifier(nn.Module):
    def __init__(self, layers_config):
        super().__init__()
        self.layers = nn.ModuleList()
        self.a_functions = []
        print(layers_config["architecture"])
        self.layers_config = layers_config["architecture"]
        layer_index = 0
        for layer in self.layers_config:
            if len(layer[0]) == 2:
                new_layer = nn.Linear(*layer[0])
            if len(layer[0]) == 3:
                new_layer = nn.Conv2d(*layer[0])
            self.layers.append(new_layer)
            self.a_functions.append(layer[1][-1])
            layer_index += 1

    def forward(self, x):
        for layer, afunc, layer_config in zip(self.layers, self.a_functions, self.layers_config):
            x = afunc(layer(x))
            if len(layer_config) == 3:
                post_processing_actions = layer_config[2]
                for action in post_processing_actions:
                    x = action(x)
        return x

class RBM(nn.Module):
    def __init__(self, n_vis, n_hid, a_func, init_type,
                 without_sampling):
        super(RBM, self).__init__()
        if init_type == InitTypes.Kaiming:
            W = torch.empty(n_vis, n_hid)
            v = torch.empty(1, n_vis)
            h = torch.empty(1, n_hid)
            torch.nn.init.kaiming_normal_(W, mode='fan_out',
```

```

        nonlinearity="relu")
    torch.nn.init.kaiming_normal_(v, mode='fan_out',
        nonlinearity="relu")
    torch.nn.init.kaiming_normal_(h, mode='fan_out',
        nonlinearity="relu"))
    self.W = nn.Parameter(W)
    self.v = nn.Parameter(v)
    self.h = nn.Parameter(h)
elif init_type == InitTypes.SimpleNormal:
    self.W = nn.Parameter(0.1 * torch.randn(n_vis,
        n_hid))
    self.v = nn.Parameter(torch.zeros(1, n_vis))
    self.h = nn.Parameter(-1 * torch.ones(1, n_hid))
elif init_type == InitTypes.SimpleUniform:
    self.W = nn.Parameter(0.02 * torch.rand(n_vis,
        n_hid)-0.01)
    self.v = nn.Parameter(0.02 * torch.rand(1, n_vis)
        -0.01)
    self.h = nn.Parameter(0.02 * torch.rand(1, n_hid)
        -0.01)
self.a_func = a_func
self.without_sampling = without_sampling

def visible_to_hidden(self, v):
    weighted_sum = torch.mm(v, self.W) + self.h
    output = self.a_func[1](weighted_sum)
    return output, weighted_sum

def hidden_to_visible(self, h):
    weighted_sum = torch.mm(h, self.W.t()) + self.v
    output = self.a_func[0](weighted_sum)
    return output, weighted_sum

def forward(self, v0):
    h0, h0_ws = self.visible_to_hidden(v0)
    if self.without_sampling:
        h_sampled = h0
    else:
        if self.a_func[1] == torch.sigmoid:
            h_sampled = 1. * (h0 > torch.rand(h0.shape))
        elif self.a_func[1] == torch.relu:
            h0_std = torch.std(h0, dim=0, unbiased=False)
            h_sampled = h0 + torch.normal(0, h0_std)
        else:
            h_sampled = h0
    v1, v1_ws = self.hidden_to_visible(h_sampled)
    h1, h1_ws = self.visible_to_hidden(v1)

```

```

        return v0, v1, v1_ws, h0, h0_ws, h1, h1_ws

    def __str__(self) -> str:
        return "vis: " + str(self.v.shape) + \
               "hid: " + str(self.h.shape)

class CRBM(nn.Module):
    def __init__(self, n_vis_channels, n_hid_channels,
                 kernel_size, a_func, init_type, without_sampling):
        super(CRBM, self).__init__()
        if init_type == InitTypes.Kaiming:
            W = torch.empty(n_hid_channels, n_vis_channels,
                            kernel_size, kernel_size)
            torch.nn.init.kaiming_normal_(W, nonlinearity="relu")
            self.W = nn.Parameter(W)
        elif init_type == InitTypes.SimpleNormal:
            self.W = nn.Parameter(0.01 * torch.randn(
                n_hid_channels, n_vis_channels, kernel_size,
                kernel_size))
        elif init_type == InitTypes.SimpleUniform:
            self.W = nn.Parameter(0.02 * torch.randn(
                n_hid_channels, n_vis_channels, kernel_size,
                kernel_size) - 0.01)
        self.a_func = a_func
        self.without_sampling = without_sampling

    def visible_to_hidden(self, v):
        weighted_sum = torch.convolution(v, self.W, None,
                                         stride=[1,1], padding=[0,0], dilation=[1,1],
                                         transposed=False, output_padding=[0,0], groups=1)
        output = self.a_func[1](weighted_sum)
        return output, weighted_sum

    def hidden_to_visible(self, h):
        weighted_sum = torch.conv_transpose2d(h, self.W, None,
                                              stride=[1, 1], padding=[0, 0], output_padding=[0,
                                                0], groups=1,
                                              dilation=[1, 1])
        output = self.a_func[0](weighted_sum)
        return output, weighted_sum

    def forward(self, v0):
        h0, h0_ws = self.visible_to_hidden(v0)
        if self.without_sampling:
            h_sampled = h0
        else:

```

```

        if self.a_func[1] == torch.sigmoid:
            h_sampled = 1. * (h0 > torch.rand(h0.shape))
        elif self.a_func[1] == torch.relu:
            h0_std = torch.std(h0, dim=0, unbiased=False)
            h_sampled = h0 + torch.normal(0, h0_std)
        else:
            h_sampled = h0
    v1, v1_ws = self.hidden_to_visible(h_sampled)
    h1, h1_ws = self.visible_to_hidden(v1)
    return v0, v1, v1_ws, h0, h0_ws, h1, h1_ws

def __str__(self) -> str:
    return "vis: " + str(self.v.shape) + \
           "hid: " + str(self.h.shape)

import torch
from torch import nn
import utilities as utl
import config
from models import RBM, CRBM

class RBMStack:
    def __init__(self, layers_config, device, init_type,
                 without_sampling):
        self.rbm_stack = []
        self.layers = layers_config["architecture"]
        self.input_dim = layers_config["input_dim"]
        self.device = device
        for i in range(0, len(self.layers) - 1):
            layer_params = self.layers[i][0]
            layer_activation_function = self.layers[i][1]
            rbm_constructor = RBM if len(layer_params) == 2
            else CRBM
            rbm = rbm_constructor(*layer_params,
                                  layer_activation_function, init_type,
                                  without_sampling)
            self.rbm_stack.append(rbm.to(self.device))

    def _prepare_train_set(self, train_set, batch_size,
                          input_dim, layer_index=0):
        batches_count = len(train_set) / batch_size
        if isinstance(input_dim, tuple):
            resulted_array = torch.zeros((train_set.shape[0], *
                                         input_dim))
        else:
            resulted_array = torch.zeros((train_set.shape[0],
                                         input_dim))

```

```

i = 0
while i < batches_count:
    _slice = slice(i * batch_size, (i + 1) * batch_size)
    inputs = train_set[_slice]
    if layer_index == 0:
        resulted_array[_slice] = inputs
    else:
        inputs = inputs.to(self.device)
        v0, v1, v1_ws, h0, h0_ws, h1, h1_ws = self.
            rbm_stack[layer_index - 1](inputs)
        resulted_array[_slice] = h0
    i += 1
if len(self.layers[layer_index - 1]) == 3:
    post_processing_actions = self.layers[layer_index -
        1][2]
    for action in post_processing_actions:
        if not isinstance(action, torch.nn.Dropout):
            resulted_array = action(resulted_array)
return resulted_array

def train(self, train_set, pretrain_type):
    layers_losses = {}
    train_set = self._prepare_train_set(train_set, config.
        pretraining_batch_size, self.input_dim)
    batches_count = len(train_set) / config.
        pretraining_batch_size
    layer_index = 0
    with torch.no_grad():
        for rbm, layer in zip(self.rbm_stack, self.layers):
            output_shape = utl.
                train_rbm_with_custom_dataset(
                    train_set, self.device, rbm, pretrain_type,
                    batches_count)
            layer_index += 1
            train_set = self._prepare_train_set(train_set,
                config.pretraining_batch_size, output_shape
                [1:],
                layer_index
            )
    return layers_losses

def torch_model_init_from_weights(self, torch_model):
    with torch.no_grad():
        for i in range(0, len(torch_model.layers) - 1):
            if len(torch_model.layers_config[i][0]) == 2:
                torch_model.layers[i].weight.data = self.

```

```

        rbm_stack[i].W.T
        torch_model.layers[i].bias.data = torch.
            reshape(self.rbm_stack[i].h, (len(self.
                rbm_stack[i].h[0]),))
    if len(torch_model.layers_config[i][0]) == 3:
        torch_model.layers[i].weight.data = self.
            rbm_stack[i].W

def do_reduction(self, layers_config):
    with torch.no_grad():
        for i in range(0, len(self.layers) - 1):
            mask = torch.abs(self.rbm_stack[i].W) > config.
                reduction_param
            self.rbm_stack[i].W *= mask.double()
    condition = None
    for i in range(0, len(self.layers) - 1):
        if condition is not None:
            self.rbm_stack[i].W = nn.Parameter(self.
                rbm_stack[i].W[condition])
        condition = torch.abs(self.rbm_stack[i].W).sum(
            dim=0) != 0
        self.rbm_stack[i].W = nn.Parameter(self.
            rbm_stack[i].W[:, condition])
        self.rbm_stack[i].h = nn.Parameter(self.
            rbm_stack[i].h[:, condition])
        layers_config["architecture"][i][0] = tuple(
            self.rbm_stack[i].W.shape)
    previous_neurons_count = layers_config["
        architecture"][len(self.layers)-2][0][1]
    end_layer_neurons_count = layers_config["
        architecture"][len(self.layers)-1][0][1]
    layers_config["architecture"][len(self.layers)
        -1][0] = (previous_neurons_count,
               end_layer_neurons_count)

import random
import config
import torchvision.datasets as datasets
import torch
from rbm_stack import RBMStack
from models import UnifiedClassifier
from torch import nn
import torch.optim as optim
import data_config
import torch.nn.functional as F
from common_types import PretrainingType, DatasetType,
    Statistics

```

```

from models import RBM

def get_random_seeds(count):
    seeds = []
    for i in range(0, count):
        seeds.append(random.randint(0, config.max_random_seed))
    return seeds

def get_dataset_constructor(dataset_type: DatasetType):
    dataset_selector = {
        DatasetType.MNIST: datasets.MNIST,
        DatasetType.CIFAR10: datasets.CIFAR10,
        DatasetType.CIFAR100: datasets.CIFAR100
    }
    return dataset_selector[dataset_type]

def train_rbm_with_custom_dataset(train_set, device, rbm,
                                  pretrain_type, batches_count):
    train_func = train_rbm if isinstance(rbm, RBM) else
        train_crbm
    losses, output_shape = train_func(rbm, device,
                                      batches_count, train_set, pretrain_type)
    return output_shape

def train_rbm(rbm, device, batches_count, train_set,
             pretrain_type):
    delta_weights = torch.zeros(rbm.W.shape).to(device)
    delta_v_thresholds = torch.zeros(rbm.v.shape).to(device)
    delta_h_thresholds = torch.zeros(rbm.h.shape).to(device)
    losses = []
    act_func = rbm.a_func
    for epoch in range(config.pretraining_epochs):
        loss = 0.
        i = 0
        momentum = config.momentum_beg if epoch < config.
            momentum_change_epoch else config.momentum_end
        while i < batches_count:
            inputs = train_set[i * config.
                pretraining_batch_size:(i + 1) * config.
                pretraining_batch_size].to(device)
            v0, v1, v1_ws, h0, h0_ws, h1, h1_ws = rbm(inputs)
            if pretrain_type == PretrainingType.RBMClassic:
                der_v, der_h = 1, 1

```

```

        rate = config.pretraining_rate
    elif pretrain_type == PretrainingType.REBA:
        der_v = (act_func[0](v1_ws + 0.00001) -
                  act_func[0](v1_ws - 0.00001)) / 0.00002
        der_h = (act_func[1](h1_ws + 0.00001) -
                  act_func[1](h1_ws - 0.00001)) / 0.00002
        rate = config.pretraining_rate_reba
    if config.with_adaptive_rate:
        b_h = h0 * ((v1 * v0).sum() + 1) - h1 * (1 + (
            v1 * v1).sum())
        b_v = v0 * (1 + (h0 * h0).sum()) - v1 * (1 + (
            h0 * h1).sum())
        rate = (((v0-v1) * b_v).sum() + ((h0-h1) * b_h)
                .sum()) / ((b_h*b_h).sum() + (b_v*b_v).sum())
    part_v = (v1 - v0) * der_v
    part_h = (h1 - h0) * der_h
    delta_weights = delta_weights * momentum + rate /
        config.pretraining_batch_size * (
            torch.mm(part_v.T, h0) + torch.mm(v1.T,
                part_h))
    delta_v_thresholds = delta_v_thresholds * momentum
    + rate / config.pretraining_batch_size * part_v.
    sum(0)
    delta_h_thresholds = delta_h_thresholds * momentum
    + rate / config.pretraining_batch_size * part_h.
    sum(0)
    rbm.W -= delta_weights
    rbm.v -= delta_v_thresholds
    rbm.h -= delta_h_thresholds
    part_loss = ((v1 - v0) ** 2).sum()
    loss += part_loss.item()
    i += 1
    print(loss)
    losses.append(loss)
return losses, h0.shape

```

```

def train_crpbm(rbm, device, batches_count, train_set,
                 pretrain_type):
    delta_weights = torch.zeros(rbm.W.shape).to(device)
    losses = []
    act_func = rbm.a_func
    for epoch in range(config.pretraining_epochs):
        loss = 0.0
        i = 0
        momentum = config.momentum_beg if epoch < config.

```

```

        momentum_change_epoch else config.momentum_end
while i < batches_count:
    inputs = train_set[i * config.
                       pretraining_batch_size:(i + 1) * config.
                       pretraining_batch_size].to(device)
    v0, v1, v1_ws, h0, h0_ws, h1, h1_ws = rbm(inputs)
    if pretrain_type == PretrainingType.RBMClassic:
        der_v, der_h = 1, 1
        rate = config.pretraining_rate
    if pretrain_type == PretrainingType.REBA:
        der_v = (act_func[0](v1_ws+0.00001) - act_func
                  [0](v1_ws-0.00001)) / 0.00002
        der_h = (act_func[1](h1_ws+0.00001) - act_func
                  [1](h1_ws-0.00001)) / 0.00002
        rate = config.pretraining_rate_reba
    part_v = (v1 - v0) * der_v
    part_h = (h1 - h0) * der_h
    first_convolution_part = torch.convolution(
        torch.permute(part_v, (1, 0, 2, 3)),
        torch.permute(h0, (1, 0, 2, 3)), None, stride
            =[1,1], padding=[0,0], dilation=[1,1],
            transposed=False, output_padding=[0,0],
            groups=1)
    second_convolution_part = torch.convolution(
        torch.permute(v1, (1, 0, 2, 3)),
        torch.permute(part_h, (1, 0, 2, 3)), None,
        stride=[1,1], padding=[0,0], dilation=[1,1],
        transposed=False, output_padding=[0,0],
        groups=1)
    common_conv_expr = first_convolution_part +
        second_convolution_part

    delta_weights = delta_weights * momentum + rate /
        config.pretraining_batch_size * torch.permute(
            common_conv_expr, (1, 0, 2, 3))
    rbm.W -= delta_weights
    loss += ((v1 - v0) ** 2).sum()
    i += 1
    print(loss.item())
    losses.append(loss.item())
return losses, h0.shape

def train_torch_model(model, train_loader, test_loader,
                      optimizer, criterion, device):
    best_total_accuracy = 0
    losses = []

```

```

for epoch in range(config.finetuning_epochs):
    running_loss = 0.0
    for i, data in enumerate(train_loader, 0):
        inputs, labels = data[0].to(device), data[1].to(
            device)

        optimizer.zero_grad()

        outputs = model(inputs)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()

        running_loss += loss.item()
    if epoch % config.test_every_epochs == 0:
        current_accuracy = test_torch_model(model,
                                              test_loader, device)
        if current_accuracy > best_total_accuracy:
            best_total_accuracy = current_accuracy
    losses.append(running_loss)
    print(running_loss)
return best_total_accuracy, losses

def test_torch_model(model, test_loader, device):
    correct_answers = 0
    with torch.no_grad():
        for data in test_loader:
            images, labels = data[0].to(device), data[1].to(
                device)
            outputs = model(images)
            _, predictions = torch.max(outputs, 1)
            correct_answers += (predictions == labels).sum()
    return 100 * float(correct_answers) / len(test_loader.
                                                dataset)

def run_experiment(layers_config, pretrain_type, meta_data,
                   device, init_type, without_sampling):
    rbm_stack = RBMStack(layers_config, device, init_type,
                         without_sampling)
    layers_losses = None
    train_loader = meta_data[2]
    train_set = data_config.get_tensor_dataset_from_loader(
        train_loader)
    if pretrain_type != PretrainingType.Without:
        layers_losses = rbm_stack.train(train_set,

```

```
    pretrain_type)
if config.with_reduction:
    rbm_stack.do_reduction(layers_config)

classifier = UnifiedClassifier(layers_config).to(device)
rbm_stack.torch_model_init_from_weights(classifier)

criterion = nn.CrossEntropyLoss(reduction="sum")
optimizer = optim.SGD(classifier.parameters(), lr=config.
    finetune_rate, momentum=config.finetuning_momentum,
    weight_decay=1e-6)

test_loader = meta_data[3]
best_total_acc, losses = train_torch_model(classifier,
    train_loader, test_loader, optimizer, criterion, device)

return Statistics.get_train_statistics(layers_losses,
    best_total_acc), losses
```

ПРИЛОЖЕНИЕ Б

Документы об использовании результатов диссертационной работы

Адкрытае акцыянернае
таварыства «Савушкін прадукт»
(ААТ «Савушкін прадукт»)
вул. Янкі Купалы, 118, 224028, г. Брэст,
тэл. +375 (162) 47 06 24, факс 47 10 80,
e-mail: info@savushkin.by
URL: <https://www.savushkin.by/>
р/р BY04 BELB 3012 1500 7502 8022 6000
у ААТ «Банк БелЭБ», код BELBBY2X
УНП 200030514, АКПА 05882743



Открытое акционерное
общество «Савушкін продукт»
(ОАО «Савушкін продукт»)
ул. Янки Купалы, 118, 224028, г. Брест,
тел. +375 (162) 47 06 24, факс 47 10 80,
e-mail: info@savushkin.by
URL: <https://www.savushkin.by/>
р/сч BY04 BELB 3012 1500 7502 8022
6000
в ОАО «Банк БелЭБ», код BELBBY2X
УНП 200030514, ОКПО 05882743

17.05.2023 №
На № от

Учреждение образования "Брестский
государственный технический
университет"

Акт внедрения результатов диссертационного исследования

Мы, нижеподписавшиеся, директор по ИТ ОАО «Савушкін Продукт» Таберко В.В. и начальник БПР ОАО «Савушкін продукт» Иванюк Д.С. с одной стороны, и соискатель Крощенко А.А., с другой стороны, составили настоящий АКТ ВНЕДРЕНИЯ результатов диссертационного исследования на тему «Методы обучения глубоких нейронных сетей для задач компьютерного зрения» (научн. руководитель д.т.н., проф. Головко В.А.).

Разработанная Крощенко А.А. нейросетевая интеллектуальная система, а также соответствующие методы обучения и инструментальные средства внедрены и используются предприятием для контроля качества маркировки продукции на поточной линии.

Интеллектуальная система включает в себя комплекс модулей, формирующий конвейер обработки входного изображения из видеопотока. Разработанные программные модули базируются на применении нейросетевых классификаторов и детекторов (5 моделей). Кроме этого, система включает в себя средства обучения и управления работой моделей.

Достоинством указанной системы, методов обучения и средств в отличие от применявшимся на предприятии ранее, является их гибкость, позволяющая осуществлять дообучение и интеграцию новых модулей для распознавания различных типов маркировок, а также высокие показатели скорости работы и качества распознавания (99 % при обнаружении отдельных товаров и маркировки, 92% при обнаружении отдельных цифр для алфавитно-цифрового типа маркировки).

Соискатель

А.А. Крощенко

Начальник бюро перспективных
разработок

Д.С. Иванюк

Директор по ИТ

В.В. Таберко



МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
"БРЕСТСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ"

УТВЕРЖДАЮ

Проректор по учебной работе



Базенков Т.Н.
"16" марта 2015 г.

М.П.

УТВЕРЖДАЮ

Проректор по научной работе



Рубанов В.С.
"16" марта 2015 г.

М.П.

АКТ ВНЕДРЕНИЯ

результатов НИР в учебный процесс

Мы, нижеподписавшиеся, начальник НИЧ Макарук Д.Г., научный руководитель д.т.н., профессор Головко В.А., исполнитель аспирант кафедры ИИТ 2-го года обучения Кроцленко А.А. с одной стороны и начальник учебно-методического отдела Сторожук Н.Ю., декан ФЭИС к.ф.-м.н., доцент Ракецкий В.М. и заведующий кафедрой ИИТ д.т.н., профессор Головко В.А. с другой стороны составили настоящий *Акт внедрения* результатов научно-исследовательской работы «Нейросетевые методы обработки комплексной информации и принятия решений на основе интеллектуальных многоагентных систем», выполненной в рамках задания ГРНИ «Информатика и космос, научное обеспечение безопасности и защиты от чрезвычайных ситуаций 1.16» , № ГР 20140547, научный руководитель – д.т.н., профессор Головко В.А.

Основные результаты работы:

- разработан альтернативный подход обучения нейронных сетей глубокого доверия, построенной на основе ограниченных машин Больцмана
- для подкрепления теоретических результатов проведены вычислительные эксперименты по исследованию широко известных выборок обучения (MNIST, CIFAR-10, NORB).

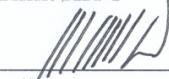
Указанная работа внедрена в учебный процесс в 2014-2015 уч. г. на кафедре интеллектуальных информационных технологий в качестве лекционного и лабораторного курсов для студентов специальностей 1-40 03 01 «Искусственный интеллект» и 1-53 01 02 «Автоматизированные системы обработки информации»

Наименование объекта внедрения и характеристика его преимущества: на простой теоретической базе разработан альтернативный подход к обучению нейронных сетей глубокого доверия, позволяющий активизировать решение совокупности актуальных научно-исследовательских задач.

Учебно-методический эффект:

- освоение современных и востребованных теоретических знаний, нацеленных на решение актуальных научно-технических задач;
- активизация научно-исследовательской инициативы студентов.

Начальник НИЧ



/ Д.Г. Макарук

Научный руководитель

Исполнитель



/ В.А. Головко

/ А.А. Кроцленко

Начальник учебного отдела

Декан факультета

Заведующий кафедрой

/ Н.Ю. Сторожук

/ В.М. Ракецкий

/ В.А. Головко

АКТ ВНЕДРЕНИЯ
результатов диссертационного исследования

г. Минск

31.03.2023 г.

Мы, нижеподписавшиеся, директор ООО «Интелиджент семантик системс» Грюневальд Т. и технический руководитель проектов ООО «Интелиджент семантик системс» Ковалев М.В. с одной стороны, и представитель БГТУ, соискатель Крошенко А.А. с другой стороны, составили настоящий АКТ ВНЕДРЕНИЯ результатов диссертационной работы на тему «Методы обучения глубоких нейронных сетей для задач компьютерного зрения» (научн. руководитель д.т.н., проф. Головко В.А.).

Разработанные алгоритмы классификации и локализации солнечных панелей на аэрофотоснимках, методы обучения нейросетевых моделей и инструментальные средства внедрены и используются на предприятии ООО «Интелиджент семантик системс» при разработке интеллектуальной диалоговой программной системы в составе подсистемы компьютерного зрения.

Применение разработанных соискателем алгоритмов, методов обучения и инструментальных средств обеспечило возможность повышения точности классификации объектов на изображениях, которая при одинаковых условиях оказалась выше, чем при использовании традиционных подходов.

Директор ООО
«Интелиджент семантик системс»



Т. Грюневальд

Соискатель



А.А. Крошенко

Технический руководитель проектов



М.В. Ковалев