Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки: 03.04.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

Регуляризация нейросетевого слоя путем построения фрейма в пространстве параметров

(магистерская работа)

Студент:
Григорьев Алексей Дмитриевич

(подпись студента)

Научный руководитель:
Гнеушев Александр Николаевич,
к-т физ.-мат. наук

(подпись научного руководителя)

Москва 2022

Содержание

Работа посвящена задаче регуляризации параметров нейронной сети для увеличения эффективности избыточного множества параметров и повышения устойчивости модели. Предлагается модель нейросетевого слоя, основанная на представлении системы весов в виде фрейма в пространстве параметров, что делает разложение входного сигнала по весам данного слоя устойчивым и полным. На основе предложенной модели вводится фреймовая регуляризация параметров слоя в виде штрафа за несоблюдение достаточного условия фрейма. В отличие от существующих методов увеличения эффективности параметров модели, основанных на ортогонализации, предложенный подход накладывает более слабые ограничения на веса модели, которые достаточны для сохранения информации о преобразованном входном сигнале, позволяют восстановить значения входа слоя по его выходу. Предложенный метод регуляризации эффективно обобщается на сверточные слои в блочно-теплицевом представлении и применим к сверточным нейронным сетям. Вычислительный эксперимент, проведенный на выборках CIFAR-10, CIFAR-100 и SVHN, показал превосходство предложенного метода регуляризации по качеству обученных моделей из класса сверточных нейронных сетей. Модели, обученные с фреймовой регуляризацией параметров, обладают большей точностью классификации, обобщающей способностью и устойчивостью к состязательным атакам по сравнению с базовыми подходами.

Ключевые слова: регуляризация параметров, нейронная сеть, линейный базис, фрейм, условие устойчивости, жесткий фрейм, границы фрейма, избыточная линейная система, ортогонализация весов.

1 Введение

Информационные интеллектуальные системы, использующие искусственные нейронные сети для анализа данных и принятия решений, активно внедряются и являются неотъемлемой частью современных решений для автоматизации широкого круга задач. В частности, важным направлением является автоматизация задач, связанных с анализом изображений и построением интеллектуальных видео систем как в промышленности, обеспечении безопасности так и в сфере обслуживания повседневной жизни человека. Глубокие нейронные сети с большим количеством параметров способны с высокой точностью описывать сложные нелинейные зависимости путем оптимизации параметров модели с помощью алгоритма обратного распространения ошибки. Высокая размерность пространства параметров делает задачу оптимизации весов модели сложной. В частности, избыточное число параметров таких моделей приводит к корреляции нейронов сети, что снижает обобщающую способность модели и приводит к неэффективному использованию вычислительных ресурсов.

Для решения проблемы избыточности параметров модели ставится задача поиска моделей с эффективной параметризацией. Как правило для решения данной проблемы предлагаются методы, основанные на структурных видоизменениях модели, или оптимизационные методы, нацеленные на повышение разнообразия нейронов каждого слоя. Структурные подходы подразумевают уменьшение числа параметров модели путем либо оптимизации нейросетевой архитектуры модели [19, 20], либо прореживание параметров модели [21, 22], в то время как подходы, основанные на оптимизации, поощряют разнообразие параметров искусственных нейронов при обучении.

Особый интерес представляет задача регуляризации параметров модели глубокого обучения. Такие методы регуляризации зачастую нацелены прежде всего на предотвращение переобучения, повышение обобщающей способности модели [8,10,12]. Классический метод регуляризации на основе введения штрафа на норму весов ограничивает абсолютные значения параметров и, тем самым, предотвращает возможный рост нормы градиентов параметров, что существенно облегчает обучение нейросетевых моделей стандартными градиентными методами [8].

К настоящему моменту предложено множество методов регуляризации параметров, ориентированных на повышение их разнообразия. Один из успешных подходов основан на обеспечении углового разнообразия значений весов нейронов [2] в векторном пространстве параметров. В частности, угловое разнообразие векторов-параметров нейронов предлагается характеризовать гиперсферической потенциальной энергией, определяемой попарным геометрическим соотношением векторов. Предлагаемая регуляризация параметров подразумевает минимизацию энергии для достижения более равномерно распределенной в пространстве и, следовательно, разнообразной конфигурации искусственных нейронов.

Альтернативное группа методов повышения разнообразия параметров нейронов связана с их ортогонализацией [3–5]. В работе [6] предлагается метод ортогональной

инициализации параметров, ускоряющий сходимость на ранних этапах обучения, однако ортогональность весов необязательно сохраняется на протяжении всего процесса обучения без введения специальной регуляризации, поощряющей ортогональность параметров-векторов нейронов. В связи с этим, развитием идеи ортогонализации параметров является введение регуляризации, направленной на построение ортогональной системы нейронов на заданном слое. Ортогональность весов нейросетевого слоя позволяет сохранить энергию входного сигнала на данном слое, что увеличивает эффективность использования всех нейронов слоя, минимизирует потерю информации о характере сигнала, упрощает оптимизацию параметров сети [3,6]. Многие подходы регуляризации, основанные на ортогонализации, показывают многообещающие результаты в задаче классификации на эталонных выборках [3–5]. Однако условие ортогональности вводит довольно жесткие ограничения на параметры нейронного слоя, и в случае избыточности параметров, при котором число нейронов больше размерности входного сигнала, снижает фактическую емкость нейросетевой модели. Жесткость ограничений препятствует оптимизатору эффективно достигать оптимум и, фактически, ортогональность параметров не достигается.

В данной работе обобщается подход ортогональной регуляризации для увеличения эффективности избыточного множества параметров нейронной сети и повышения устойчивости нейросетевой модели в задачах классификации изображений. Параметры слоя нейронной сети предлагается рассматривать как семейство векторов в евклидовом пространстве такое, что проекция входных изображений на эту систему является устойчивой и полной. В таком случае гарантируется сохранение энергии входного сигнала и его информации в условиях переопределенной системы параметров, характерной для нейронной сети. В отличие от методов ортогонализации параметров предлагается более общий подход, а именно, построение фрейма в евклидовом векторном пространстве параметров каждого слоя при обучении. Нейросетевой слой, веса которого образуют фрейм в пространстве параметров, в общем случае ограничивает относительное изменение энергии преобразованного входного сигнала на данном слое, а в частном случае сохраняет полную энергию аналогично ортогональной системе. Избыточность фрейма присуща нейронной сети и позволяет более точно описывать ее веса. Полнота и устойчивость фреймового представления аналогичны базису. Таким образом, в данной работе формулируется новая функция потерь, которая накладывает слабые ограничения на параметры модели, но обеспечивает базисные свойства для нейросетевого слоя, достаточные для восстановления входного сигнала на входе слоя по его выходу. Фреймовое представление нейросетевого слоя позволяет рассматривать нейронную сеть в виде суперпозиции слоев, на которых происходит агрегация информации, выделение признаков без потерь, и функций активации, которые фильтруют входной сигнал, выбирают значимые для решения задачи признаки.

Цель и задачи исследования. Целью исследования является увеличение эффективности избыточного множества параметров нейронной сети и повышения устой-

чивости модели на основе представления параметров линейной части нейросетевого слоя в виде фрейма в евклидовом векторном пространстве. Для достижения этой цели поставлены следующие задачи:

- изучить существующие методы регуляризации, направленные на повышение разнообразия и эффективности параметров модели;
- предложить модель нейросетевого слоя на основе представления параметров нейронов слоя в виде фрейма в евклидовом пространстве;
- построить фреймовый регуляризатор параметров нейросетевого слоя;
- реализовать предложенный метод и провести вычислительные эксперименты для задачи классификации изображений;
- сравнить предложенный метод с существующими на тестовых выборках для задач классификации, увеличения обобщающей способности и устойчивости к состязательным атакам;
- провести анализ полученных результатов.

Научная новизна. Предложена модель нейросетевого линейного слоя на основе представления его весов в виде фрейма в векторном евклидовом пространстве, гарантирующее сохранение информации о входном сигнале после его преобразования, то есть делающая нейросетевое линейное преобразование обратимым. Фреймовое представление нейросетевого слоя позволяет интерпретировать нейросетевую модель как суперпозицию слоев, преобразующих входной сигнал без потерь информации, и функций активаций, фильтрующих сигнал для выбора наиболее значимых признаков для решения поставленной задачи. Построен регуляризатор параметров нейросетевого слоя на основе достаточного условия фреймового представления.

Практическая ценность. Предложенный метод регуляризации обладает значимо лучшей по сравнению с аналогами эффективностью, достигает большей точности при решении задач классификации при том же количестве параметров модели, применим для сверточных сетей. Модели, обученные с предложенной регуляризацией, значительно более устойчивы к состязательным атакам и к смене домена. Предложенная регуляризация позволяет отказаться от стандартной регуляризации на основе штрафа на норму весов (weight decay) путем введения штрафа на соблюдение границ фрейма.

Результаты исследования данной работы доложены на 20-й Всероссийской конференции с международным участием "Математические методы распознавания образов"(г. Москва, 2021 г.) [1].

Работа состоит из 4 разделов, заключения и списка использованных источников. Содержание изложено на 30 страницах. Список литературы включает 31 наименование.

Во введении обосновывается тема исследования, ее актуальность, сформулированы цель и задачи исследования, изложены полученные результаты и их значимость.

В первом разделе приводится описание существующих методов повышения эффективности параметров нейросетевого слоя.

Во втором разделе формулируется постановка задачи.

В третьем разделе приводится решение задачи уменьшения избыточности параметров нейронной сети и повышения устойчивости модели на основе фреймового представления нейросетевого слоя.

В четвёртом разделе описывается вычислительный эксперимент, анализ результатов предложенного подхода регуляризации параметров модели и сравнение с существующими решениями в задаче классификации изображений.

В заключении сформулированы основные результаты работы.

2 Обзор литературы

2.1 Методы регуляризации

Одной из ключевых проблем машинного обучения является сложность обучения модели, которая бы одинаково точно описывала данные, представленные в обучающей выборке, и совершенно новые объекты, которые не встречались модели ранее. В силу чрезвычайно высокого числа параметров многие модели из семейства глубоких нейронных сетей подвержены переобучению, что существенно уменьшает обобщающую способность таких моделей, тем самым ограничивая их применимость вне обучающего домена. К настоящему времени предложено множество подходов к задаче регуляризации, которые накладывают дополнительные ограничения на модель. Данные ограничения повышают устойчивость модели к выбросам, представленным в обучающей выборке, позволяя точнее описывать истинную целевую зависимость в данных (рисунок 1).

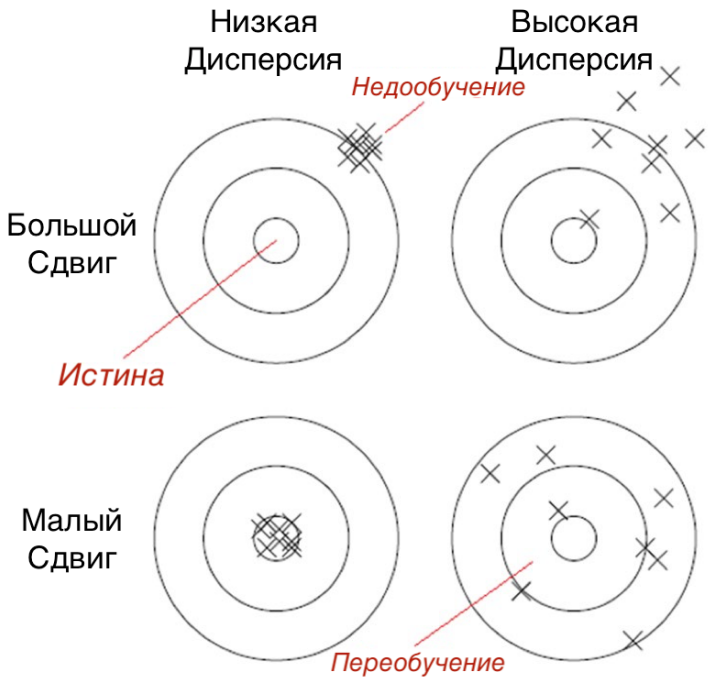

Рис. 1. Взаимосвязь переобучения с ошибкой предсказания [7]

Подходы к регуляризации могут заключаться в произвольном изменении процесса обучения, структуры модели или входных данных, которые направлены на упрощение и ограничение модели путем введения априорных допущений для компенсации ограниченности обучающей выборки. Структурные методы видоизменяют архитектуру модели на основе предположений о ее виде, наиболее успешными структурными походами являются DropOut, DropBlock, Batch Normalization [10–12]. Методы регуляризации, связанные с изменением входных данных, заключаются в аугментации данных для искусственного повышения их разнообразия. Примерами таких подходов являются Cutout [13], Mixup [14], CutMix [15], RandAug [16] (рисунок 2).

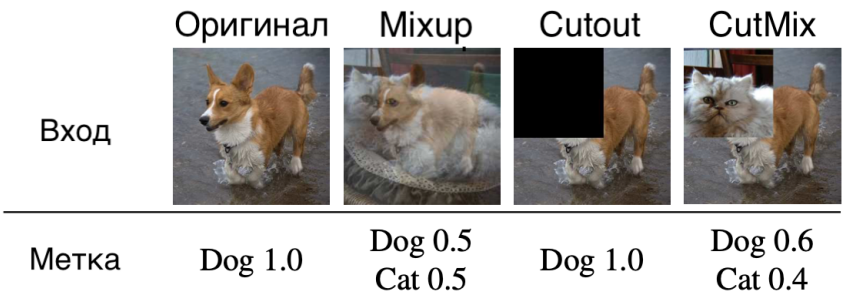

Рис. 2. Примеры аугментаций входных данных [15]

Особый интерес представляют методы регуляризации параметров модели, которые не вносят изменений в архитектурные особенности модели или входные данные, что делает их универсально применимыми для произвольных моделей и задач машинного обучения. Классической регуляризацией подобного типа является штраф на норму весов (weight decay) нейросетевого слоя [8]. Выбор нормы обуславливается целью применения данной аугментации (Рисунок 3). Так, в силу свойств L_1 нормы ее применение в данной аугментации приводит к обнулению части весов, система весов становится более разреженной, что приводит к более эффективному использованию оставшихся параметров модели. Использование L_2 нормы при регуляризации параметров приводит к более однородно распределенным нормам весов и ограничивает их амплитуду, что позволяет избежать переобучения модели и упрощает оптимизацию параметров. Штраф на норму весов является в некоторой степени стандартом в методах регуляризации и применяется при обучении многих глубоких нейросетевых моделей.

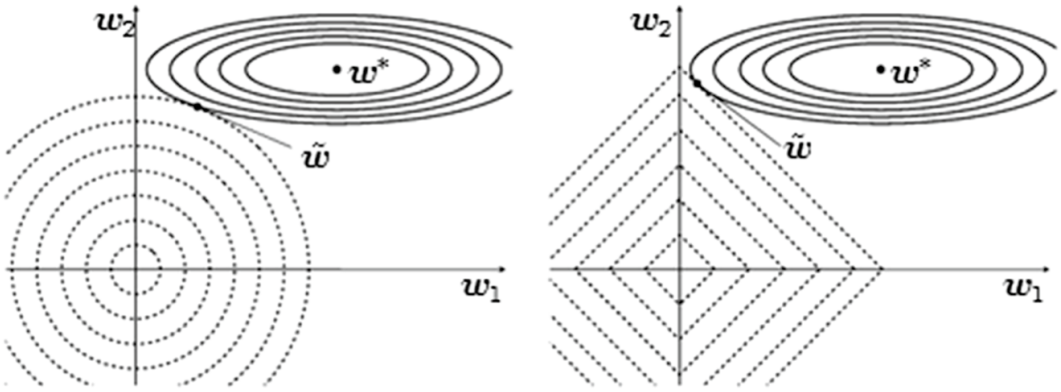

Рис. 3. Пространство параметров при использовании L_2-нормы (слева) и L_1-нормы (справа) для оптимизации двумерного вектора [9]

2.2 Методы увеличения разнообразия параметров

Среди подходов к регуляризации параметров модели отдельная группа методов основывается на повышении разнообразия параметров. Предполагается, что разнообразие параметров слоя позволит исключить коррелированность весов данного слоя и сделает использование параметров более эффективным.

Одним из успешных методов в данной области является метод повышения углового разнообразия векторов весов нейросетевого слоя [2]. Задача увеличения углового разнообразия сводится к задаче Томсона в физике, состоящей в поиске конфигурации электростатических зарядов на единичной сфере, минимизирующей их потенциальную энергию (MHE). В данном случае пара нейронов слоя, нормированных на единичную сферу, соответствуют паре элементарных зарядов, потенциальная энергия взаимодействия которых обратно пропорциональна расстоянию между ними согласно закону Кулона. Регуляризатор для системы нейронов $\{\mathbf{w}\}_{i=1}^{m} \subset \mathbb{R}^n$, являющихся строками матрицы $\mathbf{W} \in \mathbb{R}^{m \times n}$ линейной записи нейросетевого слоя, вводится через полную потенциальную энергию данной системы:

$$E(\mathbf{W}) = \sum_{i \neq j} \left\| \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} - \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \right\|^{-1}. \tag{1}$$

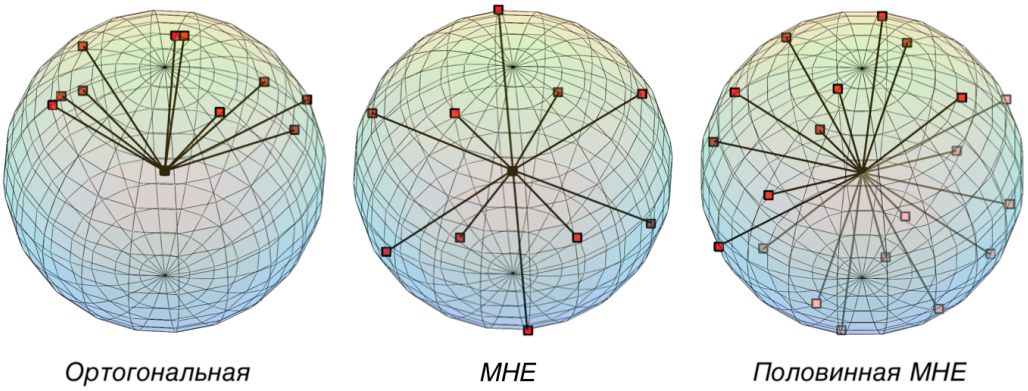

Рис. 4. Конфигурации нейронов для разных типов регуляризации [2]

Несмотря на то что предложенный аналог задачи из физики не имеет аналитического решения для произвольного числа зарядов, задача минимизации регуляризатора (1), построенного на основе полной потенциальной энергии взаимодействия нейронов, успешно решается градиентными методами. В результате данной регуляризации нормированные нейроны слоя становятся более равномерно распределенными на единичной сфере, что соответствует большему угловому разнообразию векторов весов слоя. Предложенная авторами регуляризация применима как для нейронов полносвязного слоя, так и для ядер сверточного слоя. Особую ценность данная регуляризация имеет в задачах, где угловое разнообразие системы векторов полносвязного слоя необходимо для получения качественного решения, например, в случае центроидов классов в задаче распознавания (recognition) для повышения межклассового расстояния. Однако применение данного подхода к сверточным ядрам плохо реализуется, учитывая механизм операции свертки и возможности пересечения ядер при их наложении в различных пространственных позициях.

Ряд подходов к задаче регуляризации с целью повышения разнообразия парамет-

ров прибегают к ортогонализации весов нейросетевого слоя [3–5]. Ввиду того, что не любая система ортогонализуема, в данных работах в зависимости от размерности матрицы $\mathbf{W} \in \mathbb{R}^{m \times n}$ весов слоя предлагается строить ортогональную систему строк или столбцов данной матрицы. Ортогональность весов исключает их коррелированность, более того, система нейронов слоя, обладающая свойством ортогональности, приводит к равенству норм входа и выхода для данного слоя, что упрощает оптимизацию параметров нейросети [3, 6].

В работах [3, 4] предлагаются разные подходы к ортогонализации параметров нейросетевого слоя. Одним из тривиальных подходов является регуляризатор в виде фробениусовой нормы разности матрицы Грама системы весов и единичной матрицы, основанный на определении ортонормированной системы векторов [3]:

$$R(\mathbf{W}) = \begin{cases} \|\mathbf{W}^T\mathbf{W} - \mathbb{I}\|, \ m \geq n \\ \|\mathbf{W}\mathbf{W}^T - \mathbb{I}\|, \ m < n \end{cases}, \tag{2}$$

где $\mathbf{W} \in \mathbb{R}^{m \times n}$ – матрица линейного оператора, задающего нейросетевой слой, $\mathbb{I}$ - единичная матрица. Таким образом, при $m \geq n$ ортогонализуются столбцы матрицы $\mathbf{W}$, а при $m < n$ – строки.

Альтернативный метод ортогональной регуляризации основан на свойстве спектральной ограниченной изометрии и заключается в минимизации спектральной нормы матрицы, определенной как разность матрицы Грама системы весов и единичной матрицы [4]:

$$R(\mathbf{W}) = \sigma(\mathbf{W}^T\mathbf{W} - \mathbb{I}). \tag{3}$$

Данный подход позволяет добиться ортогональности за счет наложения штрафа в случае отличия сингулярных чисел матрицы весов от единицы. На практике ортогональность весов нейросетевого слоя является избыточно строгим условием, которое затрудняет оптимизацию при обучении нейронной сети, в связи с чем оптимальное с точки зрения качества модели применение данной регуляризации приводит к системам весов далеким от ортогональных.

2.3 Регуляризация параметров сверточного слоя

В последнее десятилетие применение сверточных нейронных сетей привело к существенному прогрессу в решении многих задач обработки изображений, отчего регуляризация сверточных слоев является важной задачей глубокого обучения. Описанные в предыдущем подразделе подходы к регуляризации параметров допускают применение к сверточным слоям в случае их линейного представления.

В рассмотренных выше работах линейное представление сверточного слоя задается через преобразование ядра свертки в матрицу, произведение которой с матрицей, полученной объединением локальных областей входа (патчей) для всех позиций ядра,

задает исходную свертку (рисунок 5). Данное представление задается im2col операцией, используемой на практике для эффективного вычисления сверток [17]. Описанная запись сверточного слоя не тождественна исходной ввиду того, что в общем случае размерность входного вектора в данной записи отличается от исходной в силу возможности пересечения ядер при их наложении на различные пространственные позиции входного тензора. В случае отсутствия самопересечения ядер данное представление является точной линейной записью сверточного слоя.

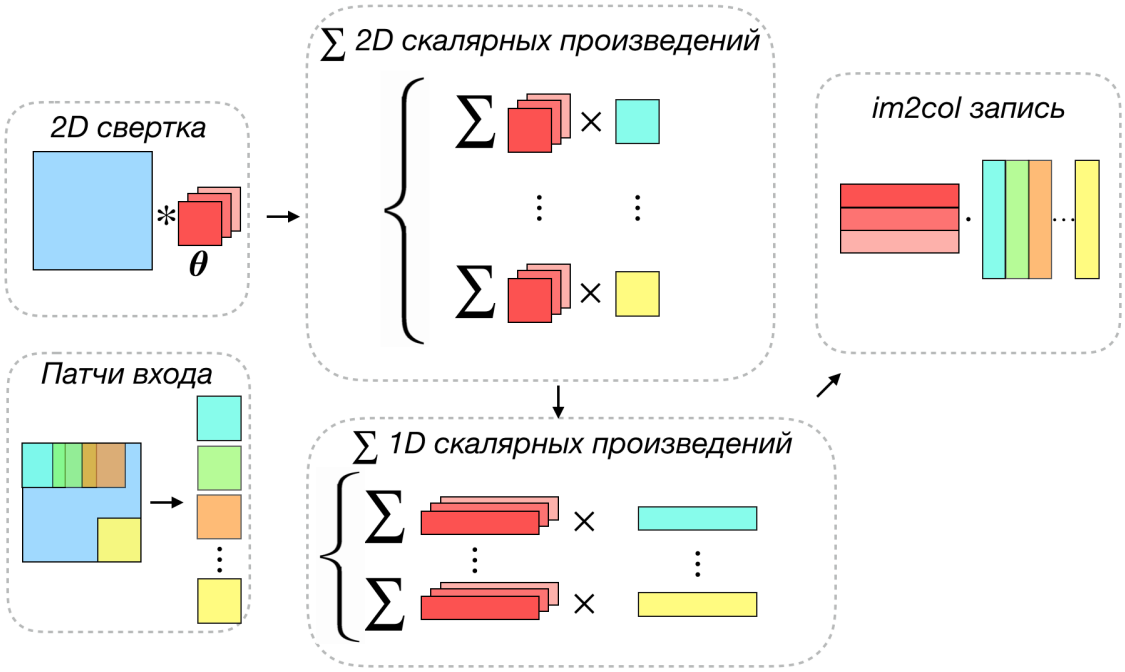

Рис. 5. im2col представление свертки

Произвольный сверточный слой представим в виде линейного оператора, матрица которого задается блочно-теплицевой матрицей, составленной из параметров свертки (рисунок 6). Число блоков по каждой из осей матрицы задается числом входных и выходных каналов в исходной свертке. Каждый из блоков представлен разреженной матрицей, размерность которой задается произведением пространственных размерностей входной и выходной карт признаков соответственно. Высокая размерность матрицы блочно-теплицевой записи свертки осложняет эффективное вычисление данного представления и операций над ним.

Для обобщения методов ортогонализации параметров на сверточные слои в блочно-теплицевом представлении требуется алгоритм эффективного вычисления матрицы Грама для такой системы. Подобный алгоритм, предложенный в работе [5], основан на свойстве разреженности блочно-теплицевой матрицы свертки, что позволяет эффективно вычислять ненулевые элементы каждой из строк матрицы Грама путем свертки ядра с самим собой. На основе предложенного способа вычисления матрицы Грама авторами вводится обобщение метода регуляризации на блочно-теплицево представление свертки, позволяющее существенно повысить эффективность подхода регуляризации без использования аппроксимации свертки через im2col представление.

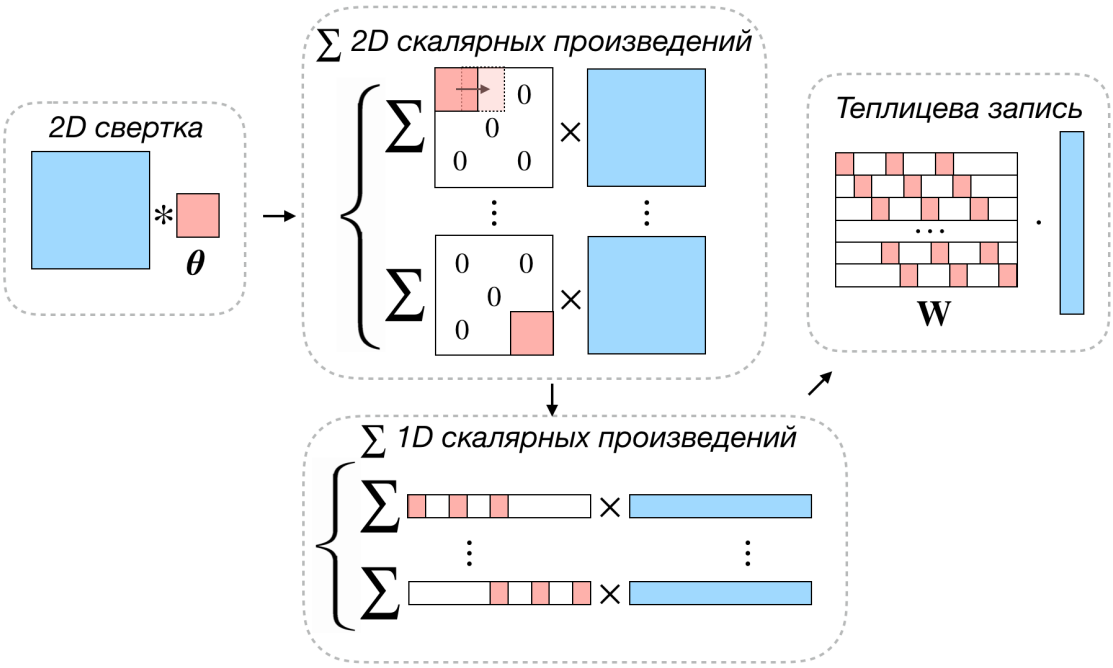

Рис. 6. Теплицево представление одноканальной свертки

3 Постановка задачи

Дана выборка $\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, где $\mathbf{x}_i \in \mathbb{R}^n$ – объект, $y_i \in \{1, \dots C\}$ – метка класса данного объекта, C – число классов в выборке, N – размер выборки.

Задана параметрическая модель $\varphi(\cdot|\mathbf{\Theta})$ из семейства Φ_L глубоких нейронных сетей следующего вида. Семейство Φ_L включает нейросетевые модели, состоящие из L слоев, каждый из которых представим в виде линейного оператора $\mathcal{F}_j : \mathbb{R}^{n_j} \to \mathbb{R}^{m_j}$ и нелинейной функции активации $h_j : \mathbb{R}^{m_j} \to \mathbb{R}^{m_j}$:

$$h_j\left(\mathcal{F}_j(\mathbf{z})\right) = h_j\left(\mathbf{W}_j\mathbf{z}\right),\ \forall \mathbf{z} \in \mathbb{R}^{n_j},\ j = 1, \dots, L, \tag{4}$$

где $\mathbf{W}_j \in \mathbb{R}^{m_j \times n_j}$, $\mathbf{W}_j \subseteq \mathbf{\Theta}$ – матрица линейного оператора $\mathcal{F}_j$, составленная из параметров данного слоя; $n_j, m_j : n_j \leq m_j$ – размерности входа и выхода слоя соответственно. Таким образом, нейросетевые слои моделей из указанного семейства Φ_L не понижают размерность входных данных.

Ставится задача мультиклассовой классификации, решение которой ищется методом минимизации эмпирического риска по заданной выборке $\mathfrak{D}$:

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \ell(\varphi(\mathbf{x}_i|\mathbf{\Theta}), y_i) + \gamma \widetilde{R}(\mathbf{\Theta}), \tag{5}$$

где ℓ – кросс-энтропийная функция потерь:

$$\ell(\varphi(\mathbf{x}|\mathbf{\Theta}), y) = -\sum_{c=1}^{C} \mathbb{1}(y = c) \log \left(\frac{\exp(\varphi(\mathbf{x}|\mathbf{\Theta})_c)}{\sum\limits_{j=1}^{C} \exp(\varphi(\mathbf{x}|\mathbf{\Theta})_j)} \right); \tag{6}$$

$\widetilde{R}(\mathbf{\Theta}) = \sum_{i=j}^{L} R(\mathbf{W}_j)$ – регуляризация параметров нейросетевой модели, представляющая из себя наложение ограничений на параметры каждого слоя в отдельности, γ – коэффициент регуляризации.

Регуляризация параметров $\mathbf{W}_j^T = \left[\mathbf{w}_1^j \dots \mathbf{w}_{m_j}^j\right]$ каждого из слоев $j = 1, \dots, L$ направлена на минимизацию потерь информации на линейной части слоя $\mathcal{F}_j(\mathbf{z}) = \mathbf{W}_j\mathbf{z}$ путем построения системы весов $\{\mathbf{w}_k^j\}_{k=1}^{m_j}$, линейно восстанавливающих вход $\mathbf{z}$ по выходу $\mathcal{F}_j(\mathbf{z})$:

$$\forall \mathbf{z} \in \mathbb{R}^{n_j}\ \exists \tilde{\mathbf{c}} = \tilde{\mathbf{c}}(\mathcal{F}_j(\mathbf{z}), \mathbf{W}_j) :\ \mathbf{z} \approx \hat{\mathbf{z}} = \sum_{k=1}^{m} \tilde{c}_k \mathbf{w}_k^j. \tag{7}$$

В виду ограничения на рассматриваемое семейство моделей Φ_L, не допускающее снижение размерности на линейной части слоя $\mathcal{F}_j(\mathbf{z}) = \mathbf{W}_j\mathbf{z}\ \ \forall j = 1, \dots, L$, такая система весов существует, но не является единственной. Таким образом, отсутствие потерь информации на линейной части нейросетевого слоя $\mathcal{F}_j(\mathbf{z}) = \mathbf{W}_j\mathbf{z}$ соответствует существованию системы весов указанного вида, приводящей к обратимости данной части слоя.

4 Методы

4.1 Фреймовое представление нейросетевого слоя

В работе предлагается рассматривать задачу регуляризации параметров модели с точки зрения увеличения их разнообразия, тем самым минимизируя потери информации на заданном нейросетевом слое. В случае возможности представления слоя в виде линейного оператора (4) с квадратной матрицей весов, ортогональность данной матрицы означает существование обратного оператора для данного слоя. В частности, в этом случае веса слоя образуют базис, и каждый вход раскладывается по базисной системе векторов с уникальными коэффициентами. В предложенной модели нейросетевого слоя предлагается сохранить свойство обратимости, соответствующее отсутствию потерь информации на данном слое, но при этом модель слоя должна включать определенную избыточность для смягчения ограничений, накладываемых на оптимизацию в случае построения ортогональной системы весов.

Предлагается построение полной системы для разложения входных векторов в избыточном пространстве весов каждого слоя, проекция входа на которую устойчива. Построение данной системы весов производится на основе фреймов [23].

Определение *(фрейм)*: $\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$ – *фрейм* в $\mathbb{R}^n$, если $\exists\, A, B:\ 0 < A \leq B < \infty:\ \forall \mathbf{z} \in \mathbb{R}^n$ выполнено *неравенство фрейма*:

$$A\|\mathbf{z}\|^2 \leq \sum_{i=1}^{m} |\langle \mathbf{z}, \mathbf{w}_i \rangle|^2 \leq B\|\mathbf{z}\|^2 \tag{8}$$

где A, B – *границы фрейма*. Если $A = B$, то фрейм называется *жестким*.

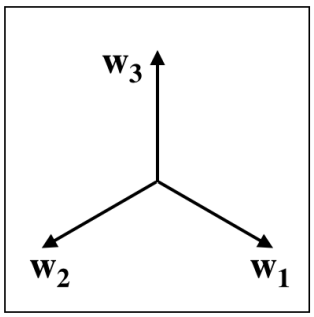

Рис. 7. Пример жесткого фрейма с границей $A = \frac{3}{2}$ в $\mathbb{R}^2$

Правая часть фреймового неравенства (8) выражает ограниченность разложения по фрейму. В конечномерном пространстве данная часть неравенства фрейма всегда выполнена в силу неравенства Коши-Буняковского:

$$\sum_{i=1}^{m} |\langle \mathbf{z}, \mathbf{w}_i \rangle|^2 \leq \left(\sum_{i=1}^{m} \|\mathbf{w}_i\|^2 \right) \|\mathbf{z}\|^2 = B\|\mathbf{z}\|^2, \tag{9}$$

где верхняя граница фрейма $B = \sum\limits_{i=1}^{m} \|\mathbf{w}_i\|^2$. Данная граница может отличаться от опти-

мальной. Левая часть неравенства фрейма (8) соответствует полноте системы векторов, образующей фрейм, и описывает устойчивость фреймового представления.

Фрейм обладает рядом свойств, которые делают естественным его использование для описания нейросетевого слоя. Полнота фрейма $\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$ в $\mathbb{R}^n$ и его избыточность при $m > n$ во многом характерны для слоя нейросети и позволяют точнее его описывать. Фреймовая система гарантированно содержит в подсистему, образующую базис в R^n. В частности если вектора фрейма линейно независимы, то сам фрейм является базисом. Фрейм обобщает понятие полной ортогональной системы векторов, образующей ортонормированный базис $\{\mathbf{w}_k\}_{k=1}^n \subset \mathbb{R}^n$. Для ортонормированного базиса выполняется равенство Парсеваля:

$$\|\mathbf{z}\|^2 = \sum_{i=1}^{n} |\langle \mathbf{z}, \mathbf{w}_i \rangle|^2, \quad \forall \mathbf{z} \in \mathbb{R}^n, \tag{10}$$

что соответствует выполнению неравенства фрейма (8) с границами фрейма $A = B = 1$. Таким образом, полная ортогональная система является частным случаем фрейма. В связи с этим ортогональная регуляризация (2) параметров нейросетевого слоя вида (4) соответствует частному случаю построения жесткого фрейма с границами $A = B = 1$ (фрейм Парсеваля-Стеклова), в действительности равенство $\mathbf{W}^T\mathbf{W} = \mathbb{I}$ необходимо и достаточно для того, чтобы строки матрицы $\mathbf{W}$ образовывали фрейм Парсеваля-Стеклова [23]. Более того, фрейм является естественным обобщением полных ортогональных систем с точки зрения сингулярных чисел матрицы $\mathbf{W}$, спектра матрицы $\mathbf{W}^T\mathbf{W}$. Для ортогональной системы $\mathbf{W}^T\mathbf{W} = \mathbb{I}$, все собственные значения равны 1. В то время как если строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ образуют фрейм, то собственные числа $\lambda_1, \ldots, \lambda_n$ матрицы $\mathbf{W}^T\mathbf{W}$ ограничены границами фрейма:

$$A \leq \lambda_i \leq B, \ \forall i = 1, \ldots, n. \tag{11}$$

В случае жесткого фрейма с границами $A = B = 1$ спектры для фрейма и ортогональной системы совпадают в случае одинаковой мощности систем. Для жесткого фрейма $A = B$ все сингулярные числа матрицы $\mathbf{W}$ одинаковы, в данном случае множество векторов пространства наиболее равномерно представляется векторами фреймовой системы, то есть элементы фрейма наиболее разнообразны.

Ограниченность спектра для фреймового представления линейного оператора с матрицей $\mathbf{W}$ позволяет оценить константу Липшица $\mathcal{L}$ данного оператора, равную спектральной норме матрицы $\mathbf{W}$:

$$\mathcal{L} = \|\mathbf{W}\|_2 = \sigma_{\max}(\mathbf{W}) = \sqrt{\lambda_{\max}(\mathbf{W}^T\mathbf{W})} \leq \sqrt{B}. \tag{12}$$

Одним из важнейших свойств фрейма $\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$ является возможность разложения произвольно элемента пространства по дуальному фрейму $\{\widetilde{\mathbf{w}}_i\}_{i=1}^m$, элементы

которого определяются как $\widetilde{\mathbf{w}}_i = \left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{w}_i$, $\mathbf{W}^T = [\mathbf{w}_1 \dots \mathbf{w}_m]$, $i = 1, \dots, m$:

$$\mathbf{z} = \sum_{i=1}^{m} \langle \mathbf{z}, \mathbf{w}_i \rangle \widetilde{\mathbf{w}}_i, \ \forall \mathbf{z} \in \mathbb{R}^n. \tag{13}$$

Разложение по дуальному фрейму задает решение переопределенной системы линейных алгебраических уравнений (СЛАУ) $\mathbf{u} = \mathbf{W}\mathbf{z}$, где $\mathbf{W}^T = [\mathbf{w}_1 \dots \mathbf{w}_m]$, строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ образуют фрейм. Причем фрейм дает устойчивое решение задачи восстановления входа: $\mathbf{z} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{u}$. Обусловленность задачи ограничена отношением границ фрейма:

$$\kappa(\mathbf{W}) = \|\mathbf{W}^T\mathbf{W}\| \, \|(\mathbf{W}^T\mathbf{W})^{-1}\| = \frac{|\lambda_{\max}(\mathbf{W}^T\mathbf{W})|}{|\lambda_{\min}(\mathbf{W}^T\mathbf{W})|} \leq \frac{B}{A}. \tag{14}$$

В случае жесткого фрейма $A = B$ обусловленность задачи оптимальна $\kappa(\mathbf{W}) = 1$.

Приведенные свойства делают до некоторой степени естественным использование фрейма в качестве модели слоя нейросети. Пусть нейросетевой слой, подобно (4), задан линейным оператором $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$ с матрицей $\mathbf{W} \in \mathbb{R}^{m \times n} : \ m \geq n$, $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$, $\forall \mathbf{z} \in \mathbb{R}^n$; $\mathbf{W}^T = [\mathbf{w}_1 \dots \mathbf{w}_m]$. Предлагается рассматривать строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ в качестве фреймовой системы векторов. Наличие разложения по дуальному фрейму (13) позволяет сформулировать следующую теорему.

Теорема (*Григорьев-Гнеушев, 2022*): Для обратимости линейной части нейросетевого слоя $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ необходимо и достаточно, чтобы строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ образовывали фрейм в $\mathbb{R}^n$.

Доказательство:

- **Необходимость:** пусть линейная часть слоя $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ обратима, покажем, что строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ образуют фрейм в $\mathbb{R}^n$.

  Обратимость соответствует линейной восстановимости (7) произвольного входа $\mathbf{z} \in \mathbb{R}^n$ по системе $\{\mathbf{w}_k\}_{k=1}^m$, то есть данная система полна в $\mathbb{R}^n$. Полная система в $\mathbb{R}^n$ является фреймом в данном пространстве [23].

- **Достаточность:** пусть строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ – фрейм в $\mathbb{R}^n$, покажем, что линейный оператор $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ обратим.

  Воспользуемся свойством (13) разложения входа по дуальному фрейму:

$$\mathbf{z} = \sum_{i=1}^{m} \langle \mathbf{z}, \mathbf{w}_i \rangle \widetilde{\mathbf{w}}_i = \sum_{i=1}^{m} (\mathbf{W}\mathbf{z})_i \, \widetilde{\mathbf{w}}_i = \sum_{i=1}^{m} (\mathcal{F}(\mathbf{z}))_i \, \widetilde{\mathbf{w}}_i, \ \forall \mathbf{z} \in \mathbb{R}^n, \tag{15}$$

  где $\{\widetilde{\mathbf{w}}_i\}_{i=1}^m$ – канонический дуальный фрейм. Поскольку элементы дуального фрейма определяются как $\widetilde{\mathbf{w}}_i = \left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{w}_i \ \ i = 1, \dots, m$, выражение (15) пере-

писывается в виде:

$$\mathbf{z} = \sum_{i=1}^{m} \underbrace{\left(\mathcal{F}(\mathbf{z})\right)_i \left(\mathbf{W}^T\mathbf{W}\right)^{-1}}_{\tilde{c}_i(\mathcal{F}(\mathbf{z}),\mathbf{W})} \mathbf{w}_i = \sum_{i=1}^{m} \tilde{c}_i \mathbf{w}_i, \ \forall \mathbf{z} \in \mathbb{R}^n, \tag{16}$$

то есть имеет место линейное восстановление входа $\mathbf{z}$ по выходу $\mathcal{F}(\mathbf{z})$, подобно (7). Таким образом, линейная часть слоя $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ обратима.

4.2 Фреймовая регуляризация

Ввиду того что явная параметризация нейросетевого слоя в качестве фрейма не представляется возможной, предлагается построение регуляризатора, накладывающего штраф за несоблюдение достаточного условия фрейма.

Фреймовое неравенство (8) для нейросетевого слоя вида (4) записывается в виде:

$$A\|\mathbf{z}\|^2 \leq \|\mathbf{W}\mathbf{z}\|^2 \leq B\|\mathbf{z}\|^2, \ \forall \mathbf{z} \in \mathbb{R}^n, \tag{17}$$

фрейм в данном случае образуют строки матрицы $\mathbf{W}$. Фреймовое неравенство (17) может быть переписано в следующем виде:

$$\begin{cases} \mathbf{x}^T \left(\mathbf{W}^T\mathbf{W} - A\,\mathbb{I}\right)\mathbf{x} \geq 0, \ \forall \mathbf{x} \in \mathbb{R}^n, \\ \mathbf{x}^T \left(-\mathbf{W}^T\mathbf{W} + B\,\mathbb{I}\right)\mathbf{x} \geq 0, \ \forall \mathbf{x} \in \mathbb{R}^n. \end{cases} \tag{18}$$

Неравенства данной системы соответствуют положительной полуопределенности матрицы $\left(\mathbf{W}^T\mathbf{W} - A\,\mathbb{I}\right)$ и $\left(-\mathbf{W}^T\mathbf{W} + B\,\mathbb{I}\right)$ соответственно. Регуляризатор для произвольной матрицы $\mathbf{W}$ весов слоя определяется путем введения штрафа за нарушение этих неравенств, пользуясь следующим достаточным условием положительной полуопределенности матрицы, являющимся следствием теоремы Гершгорина [24].

Матрица $\mathbf{V} \in \mathbb{R}^{m\times m}$ положительно полуопределена, если:

1. она обладает свойством диагонального преобладания:

$$|v_{ii}| \geq \sum_{j\neq i} |v_{ij}| \ \forall i = 1, \ldots, m, \tag{19a}$$

2. ее диагональные элементы неотрицательны:

$$v_{ii} \geq 0 \ \forall i = 1, \ldots, m. \tag{19b}$$

Пусть $\mathbf{V} = \mathbf{W}^T\mathbf{W}$, $M(v) = \min(v, 0)$; введем фреймовый регуляризатор для от-

дельного нейросетевого слоя с матрицей весов $\mathbf{W}$:

$$R(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{M\big(v_{ii} - A - \sum_{j=1}^{n}|v_{ij}|\big)^2}_{\text{штраф } i\text{-ой строки } (\mathbf{W}^T\mathbf{W} - A\,\mathbb{I})} + \underbrace{M\big(-v_{ii} + B - \sum_{j=1}^{n}|v_{ij}|\big)^2}_{\text{штраф } i\text{-ой строки } (-\mathbf{W}^T\mathbf{W} + B\,\mathbb{I})}, \tag{20}$$

где $v_{ij} = (\mathbf{W}^T\mathbf{W})_{ij}$.

Отметим, что регуляризация на основе введения штрафа на L_2 норму весов (weight decay) противоречит предложенной фреймовой регуляризации, поскольку приводит к уменьшению диагонально доминирования матрицы $\mathbf{W}^T\mathbf{W}$ путем минимизации диагональных элементов – квадратов весов слоя.

4.3 Фреймовая регуляризация проекционного слоя

Несмотря на то что многие слои наиболее успешных нейросетевых архитектур [18, 20] представимы в виде линейного оператора, не понижающего размерность входа, то есть удовлетворяют представлению (4), во многих архитектурах так же встречаются слои, снижающие размерность. Предлагается обобщение сформулированного ранее метода регуляризации на слои, уменьшающие размерность и представимые в виде линейного оператора с матрицей $\mathbf{W} \in \mathbb{R}^{m\times n} : m < n$.

При $m < n$ вектора $\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$, являющиеся строками матрицы $\mathbf{W}$, не образуют полную систему в $\mathbb{R}^n$, что делает невозможным разложение произвольного вектора пространства по данной системе без потерь. Таким образом, слой представляет собой проекцию на подпространство меньшей размерности. Предлагается построить модель нейросетевого слоя, задающего проекцию с потерей наименьшего количества информации для некоторого подпространства исходного пространства.

Пусть $\mathcal{V} = \{\mathbf{W}^T\boldsymbol{\nu} | \boldsymbol{\nu} \in \mathbb{R}^m\} = \operatorname{span}\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$ – линейная оболочка системы $\{\mathbf{w}_k\}_{k=1}^m$, тогда сужение $\widetilde{\mathcal{F}}$ оператора $\mathcal{F}$ на подпространство $\mathcal{V}$ задается в следующем виде: $\forall \mathbf{z} \in \mathcal{V}\ \widetilde{\mathcal{F}}(\mathbf{z}) = \mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{W}^T\boldsymbol{\nu}$. Оператор $\widetilde{\mathcal{F}}$ не понижает размерность входа. В допущении, что $\{\mathbf{w}_k\}_{k=1}^m$ – линейно независимая система, матрица Грама $\mathbf{W}\mathbf{W}^T$ невырождена. Соответственно, в рамках данного допущения оператор $\widetilde{\mathcal{F}}$ обратим, обратный оператор $\widetilde{\mathcal{F}}^{-1}$ задается матрицей $(\mathbf{W}\mathbf{W}^T)^{-1}$. То есть для рассматриваемого слоя потери информации отсутствуют на множестве входов $\mathcal{V} \subset \mathbb{R}^n$.

Оператор $\widetilde{\mathcal{F}}$ удовлетворяет модели (4), для него применимы рассуждения о построении фреймового регуляризатора. Неравенство фрейма (8) в данном случае имеет вид:

$$\begin{aligned} &A\|\mathbf{z}\|^2 \le \|\mathbf{W}\mathbf{z}\|^2 \le B\|\mathbf{z}\|^2,\ \forall \mathbf{z} \in \mathcal{V} \iff \\ &A\|\mathbf{W}^T\boldsymbol{\nu}\|^2 \le \|\mathbf{W}\mathbf{W}^T\boldsymbol{\nu}\|^2 \le B\|\mathbf{W}^T\boldsymbol{\nu}\|^2,\ \forall \boldsymbol{\nu} \in \mathbb{R}^m \end{aligned} \tag{21}$$

и соответствует положительной полуопределенности двух матриц:

$$\begin{cases} \mathbf{W}\mathbf{W}^T\left(\mathbf{W}\mathbf{W}^T - A\,\mathbb{I}\right) \succeq 0, \\ \mathbf{W}\mathbf{W}^T\left(-\mathbf{W}\mathbf{W}^T + B\,\mathbb{I}\right) \succeq 0. \end{cases} \tag{22}$$

Для выполнения данной системы достаточно, чтобы каждая из матриц $\left(\mathbf{W}\mathbf{W}^T - A\,\mathbb{I}\right)$ и $\left(-\mathbf{W}\mathbf{W}^T + B\,\mathbb{I}\right)$ была положительно полуопредена. Достаточность следует из коммутативности матриц $\mathbf{W}\mathbf{W}^T$, $(\mathbf{W}\mathbf{W}^T - A\,\mathbb{I})$ и $(-\mathbf{W}\mathbf{W}^T + B\,\mathbb{I})$, а так же наблюдения о том, что положительная полуопределенность матрицы $(\mathbf{W}\mathbf{W}^T - A\,\mathbb{I})$, где $A > 0$, влечет положительную полуопределенность матрицы $\mathbf{W}\mathbf{W}^T$.

Таким образом, данное достаточное условие симметрично условию (18) с точностью до порядка транспонирования матриц, соответственно, справедливы рассуждения о достаточном условии положительной полуопределенности через диагональное доминирование (19). Ввиду этого при $\mathbf{V} = \mathbf{W}\mathbf{W}^T$ регуляризация имеет вид (20), как и в случае $m \geq n$. В данном случае $\{\mathbf{w}_k\}_{k=1}^m$, действительно, является линейно независимой системой, а значит образует базис в своей линейной оболочке $\mathcal{V}$.

4.4 Фреймовая регуляризация сверточных слоев

Особый интерес представляет класс нейросетевых моделей, состоящий из сверточных нейронных сетей, в силу успешности их применения во многих задачах обработки изображений и компьютерного зрения. Сверточный слой удовлетворяет линейному представлению (4) в случае, если размерность входа не снижается. В качестве матрицы $\mathbf{W}$ весов линейного представления сверточного слоя выступает блочно-теплицева матрица, составленная из параметров свертки.

На практике размерность матрицы $\mathbf{W}$ оказывается слишком высокой для эффективного вычисления матрицы $\mathbf{W}^T\mathbf{W}$, что накладывает ограничения на применимость предложенного метода в исходном виде. Ввиду разреженности блочно-теплицевой матрицы $\mathbf{W}$ предлагается использовать алгоритм вычисления ненулевых элементов каждой из строк матрицы $\mathbf{W}^T\mathbf{W}$ на основе свертки ядер с самими собой, предложенный в работе [5]. Эффективность алгоритма обусловлена конечным набором возможных пространственных пересечений двух сверточных ядер (рисунок 8).

Отличительной особенностью предложенного регуляризатора 20 является инвариантность его значений к позициям внедиагональных элементов каждой из строк матрицы $\mathbf{W}^T\mathbf{W}$. Данное свойство фреймового регуляризатора позволяет успешно вычислять его значения на основе приведенного выше алгоритма вычисления ненулевых элементов матрицы $\mathbf{W}^T\mathbf{W}$.

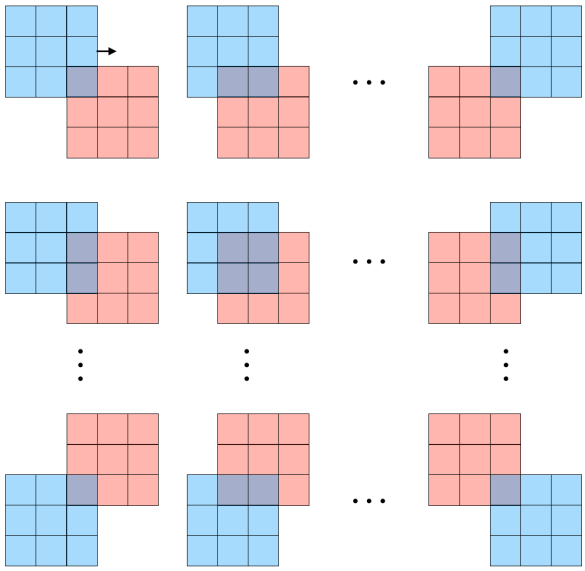

Рис. 8. Возможные пространственные пересечения пары ядер 3×3

5 Вычислительный эксперимент

В данной работе вычислительные эксперименты разделяются на 3 набора по направлению исследований. В рамках первой группы экспериментов проводится сравнение предложенного метода регуляризации с существующими в рамках задачи мультиклассовой классификации. Вторая группа экспериментов посвящена исследованию обобщающей способности моделей, обученных с разными видами регуляризаций, при смене домена. Третье направление экспериментов связано с исследованием устойчивости моделей к состязательным атакам.

5.1 Классификация изображений

В качестве данных при проведении вычислительных экспериментах по мультиклассовой классификации использовались три выборки изображений: CIFAR-10, CIFAR-100 [27] и SVHN [28]. Данные описаны в таблице 1.

Таблица 1. Описание выборок

Выборка	Число изображений	Число классов
CIFAR-10	60000	10
CIFAR-100	60000	100
SVHN	$\sim$100000	10

В качестве модели использовалась сверточная нейронная сеть архитектур ResNet-34 и ResNet-50 соответственно [18]. В качестве функции потерь выступала кросс-энтропия, к которой добавлялся один из регуляризаторов с коэффициентом, приведенным авторами данного метода. Так же при обучении использовалась регуляризация на основе L_2 штрафа на норму весов [8], за исключением экспериментов с методом фреймовой регуляризации по причине негативного влияния на диагональное доминирование. Модели обучены на 200 эпохах с размером батча 128, в качестве оптимизатора использовался Adam с начальным шагом 0.01 и мультипликативным уменьшением шага с коэффициентом 0.1 на 100, 150 и 180 эпохах [25].

Произведено сравнение предложенного метода регуляризации на основе фреймовой модели нейросетевого слоя с существующими методами: минимизацией гиперсферической потенциальной энергии (Minimum Hyperspherical Energy) [2], спектральной ограниченной изометрии (Spectral Restricted Isometry Property) [4] и подходами к ортогонализации (Weights Orthogonalization, Orthogonal Convolutions) [3, 5]. Ортогонализация применялась к сверточным слоям в im2col представлении, для последних двух из указанных методов использовалось блочно-теплицево представление. В качестве показателей качества использовалась стандартная мера качества в задаче классификации – точность (accuracy). В силу сбалансированности классов во всех используемых выборках использование данного показателя качества считается оправданным.

Результаты работы методов на выборках описаны в таблице 2 и таблице 3 соответственно.

Таблица 2. Точность (%) методов регуляризации (ResNet-34)

Метод регуляризации	CIFAR-10	CIFAR-100	SVHN
Без регуляризации	94.53 ± 0.03	75.58 ± 0.08	96.50 ± 0.03
Minimum Hyperspherical Energy	94.58 ± 0.04	75.78 ± 0.08	96.59 ± 0.03
Weights Orthogonalization	94.59 ± 0.04	75.98 ± 0.08	96.51 ± 0.02
Spectral Restricted Isometry Property	94.72 ± 0.03	76.24 ± 0.09	96.57 ± 0.03
Orthogonal Convolutions	95.03 ± 0.04	76.57 ± 0.06	96.66 ± 0.02
Фреймовая регуляризация	**95.17 ± 0.05**	**77.61 ± 0.07**	**96.85 ± 0.02**

Таблица 3. Точность (%) методов регуляризации (ResNet-50)

Метод регуляризации	CIFAR-10	CIFAR-100	SVHN
Без регуляризации	94.83 ± 0.04	77.20 ± 0.07	96.92 ± 0.03
Minimum Hyperspherical Energy	94.88 ± 0.03	77.34 ± 0.06	96.94 ± 0.02
Weights Orthogonalization	94.92 ± 0.04	77.38 ± 0.06	96.91 ± 0.03
Spectral Restricted Isometry Property	95.01 ± 0.03	77.40 ± 0.07	96.95 ± 0.03
Orthogonal Convolutions	**95.29 ± 0.03**	77.77 ± 0.07	97.01 ± 0.02
Фреймовая регуляризация	95.25 ± 0.04	**78.35 ± 0.06**	**97.10 ± 0.02**

Полученные результаты показывают, что предложенный метод значимо превосходит базовые подходы с точки зрения качества классификации на выборках CIFAR-100 и SVHN для обеих моделей. На выборке CIFAR-10 предложенный метод сравним по качеству с лучшим из базовых методов.

В работе так же была исследована зависимость качества классификации от границ фрейма при регуляризации. Данная зависимость для модели ResNet-34 на выборке CIFAR-100 представлена на рисунке 9.

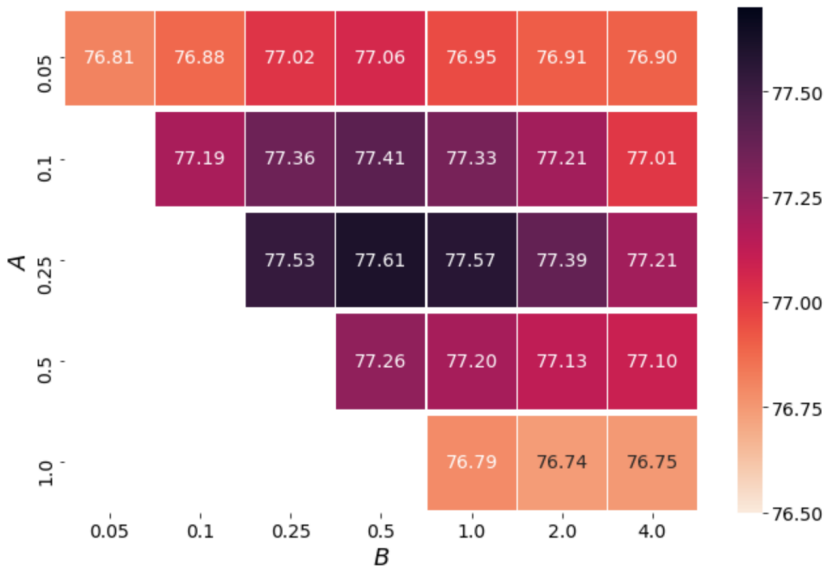

Рис. 9. График зависимость точности (%) классификации от границ фрейма A, B

5.2 Устойчивость к смене домена

Произведено сравнение предложенного метода регуляризации на основе фрейма с базовыми методами с точки зрения устойчивости к смене домена. Модели архитектуры ResNet-34, обученные на выборке CIFAR-10 с разными видами регуляризации, сравнивались в терминах точности на новых доменах, которые задавались выборками CIFAR-10-C и CINIC-10 соответственно [29, 30]. CIFAR-10-C [30] является аугментированной версией выборки CIFAR-10 с 19 разными типами возмущений, среди которых присутствуют нормальный шум, размытие, изменение контрастности и прочее. Примеры аугментированных изображений представлены на рисунке 10. CINIC-10 [29] – подвыборка ImageNet [31], включающая классы из CIFAR-10.

Рис. 10. Примеры аугментированных изображений из выборки CIFAR-10-C

Результаты работы методов при смене домена приведены в таблице 4. Для моделей с регуляризацией выбраны субоптимальные эпохи из обучения для уравнивания качества моделей на исходном домене и независимого сравнения на тестовых доменах.

Таблица 4. Точность (%) методов регуляризации на разных доменах

Метод регуляризации	CIFAR-10 $^{(*)}$	CIFAR-10-C	CINIC-100
Без регуляризации	94.53 ± 0.03	74.77 ± 0.25	67.91 ± 0.35
Orthogonal Convolutions	94.52 ± 0.01	76.27 ± 0.19	69.87 ± 0.29
Фреймовая регуляризация	94.53 ± 0.01	**76.65 ± 0.15**	**71.20 ± 0.32**

$^{(*)}$ – исходный домен

Модель, обученная с предложенным методом регуляризации на основе фреймового представления слоя, обладает существенно более высокой обобщающей способностью по сравнению с моделью, обученной без регуляризации, направленной на повышение разнообразия параметров. Фреймовая регуляризация обладает несколько более высокой устойчивостью к смене домена по сравнению с ортогональной регуляризацией.

5.3 Устойчивость к состязательным атакам

Исследована устойчивость моделей, обученных с разными видами регуляризации, к состязательным атакам. В качестве метода состязательной атаки использовался подход SimBA, представляющий из себя итеративную атаку типа "черный ящик" [26]. Итерация атаки производится путем добавления к входу модели случайного вектора из ортонормированного базиса во входном пространстве с малым весом, знак которого определяется исходя из значений выхода модели для данного видоизмененного входа. Вектора ортонормированного базиса в конечномерном пространстве входов однозначно соответствуют элементам входного тензора: содержат единственную единицу на соответствующей позиции, остальные компоненты – нули. Таким образом, итерация атаки соответствует возмущению строго одного элемента входного тензора.

Эффективность атаки оценивается по доле успешных атак (Attack Success Rate – ASR) при заданном числе итераций. Успешность атаки в задаче классификации соответствует изменению предсказания модели для видоизмененного атакой входа, атака производится только для верно классифицированных объектов.

Результаты успешности атак на модели, обученные с разными регуляризациями, отображены в таблице 5. На рисунке 11 представлена зависимость доли успешных атак от числа итераций метода SimBA.

Таблица 5. Зависимость ASR (%) от числа итераций

Метод регуляризации	# Итераций				
	1	10	50	100	1000
Без регуляризации	52.08	59.37	84.38	92.71	93.75
Orthogonal Convolutions	41.30	57.61	83.69	91.30	92.06
Фреймовая регуляризация	**39.56**	**49.45**	**80.20**	**84.61**	**86.81**

Исходя из полученных результатов, заключается, что модель, обученная с предложенным методом регуляризации, существенно более устойчива к исследованной состязательной атаке по сравнению с моделью, обученной без регуляризации. Фреймовое представление нейросетевых слоев позволяет ограничить сверху константу Липшица (12), что делает модель более устойчивой по отношению к малому возмущению входа. Для модели, обученной с фреймовой регуляризацией, требуется до ~ 1.4 раза большего числа итераций для достижения доли успешных атак ASR $= 70\%$ в сравнении с ортогональной регуляризацией, что может объясняться тем, что фреймовая регуляризация позволяет достигать более ограниченного спектра весов слоя.

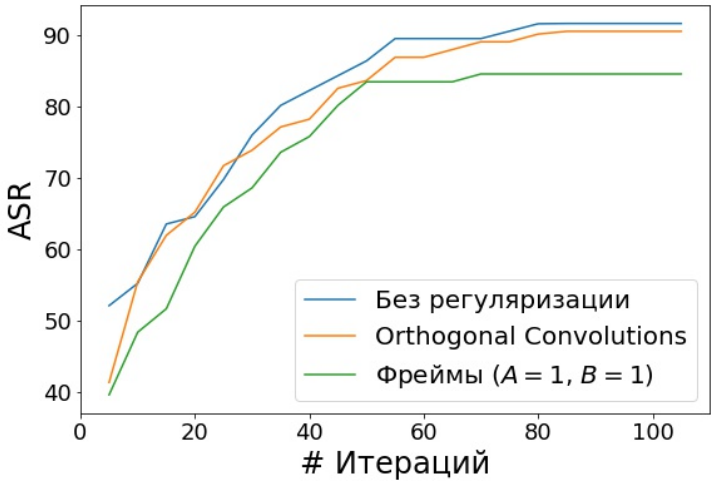

Рис. 11. График зависимости доли успешных атак ASR (%) от числа итераций

6 Заключение

В работе была поставлена задача регуляризации нейронной сети, направленная на увеличение эффективности избыточного множества параметров и повышения устойчивости модели. Были изучены существующие подходы и выявлены их недостатки. В частности, ортогонализация параметров нейросетевого слоя является избыточно жестким ограничением и, фактически, ортогональность весов не достигается. Для увеличения разнообразия параметров нейронной сети был обобщен метод ортогонализации и предложена модель нейросетевого слоя, представляющая параметры слоя в виде фрейма. Представление параметров слоя в виде фреймовой системы делает разложение входного сигнала по весам слоя полным и устойчивым, исключает потерю информации на данном слое. На основе предложенной модели нейросетевого слоя разработан регуляризатор, накладывающий штраф на параметры за несоблюдение достаточного условия фрейма. Проведен вычислительный эксперимент по оценке качества разработанного метода в сравнении с альтернативными подходами в задачах классификации изображений, увеличения обобщающей способности и устойчивости к состязательным атакам. Показано превосходство моделей, обученных с использованием фреймовой регуляризации, с точки зрения точности классификации и устойчивости модели по сравнению с базовыми методами регуляризации параметров.

Список литературы

[1] Григорьев А.Д., Гнеушев А.Н. Регуляризация параметров нейронной сети на основе неравенства Рисса //Математические методы распознавания образов: Тезисы докладов 20-й Всероссийской конференции с международным участием, г. Москва 2021 г. — М.: Российская академия наук, 2021. — С. 121-122.

[2] Liu W. et al. Learning towards minimum hyperspherical energy //Advances in neural information processing systems. – 2018. – Т. 31.

[3] Xie D., Xiong J., Pu S. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2017. – С. 6176-6185.

[4] Bansal N., Chen X., Wang Z. Can we gain more from orthogonality regularizations in training deep networks? //Advances in Neural Information Processing Systems. – 2018. – Т. 31.

[5] Wang J. et al. Orthogonal convolutional neural networks //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2020. – С. 11505-11515.

[6] Andrew M. Saxe et. al. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks //International Conference on Learning Representations (ICLR 2014). – 2014.

[7] Domingos P. A few useful things to know about machine learning //Communications of the ACM. – 2012. – Т. 55. – №. 10. – С. 78-87.

[8] Krogh A., Hertz J. A Simple Weight Decay Can Improve Generalization. //Advances in Neural Information Processing Systems 4. – 1992; – С. 950–957.

[9] Goodfellow I., Bengio Y., Courville A. Deep learning. – MIT press, 2016.

[10] Srivastava N. et al. Dropout: a simple way to prevent neural networks from overfitting //The journal of machine learning research. – 2014. – Т. 15. – №. 1. – С. 1929-1958.

[11] Ghiasi G., Lin T. Y., Le Q. V. Dropblock: A regularization method for convolutional networks //Advances in neural information processing systems. – 2018. – Т. 31.

[12] Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift //International conference on machine learning. – PMLR, 2015. – С. 448-456.

[13] DeVries T., Taylor G. W. Improved regularization of convolutional neural networks with cutout //arXiv preprint arXiv:1708.04552. – 2017.

[14] Zhang H. et al. mixup: Beyond empirical risk minimization //International Conference on Learning Representations (ICLR 2018). – 2018.

[15] Yun S. et al. Cutmix: Regularization strategy to train strong classifiers with localizable features //Proceedings of the IEEE/CVF international conference on computer vision. – 2019. – С. 6023-6032.

[16] Cubuk E. D. et al. Randaugment: Practical automated data augmentation with a reduced search space //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2020. – С. 702-703.

[17] Chellapilla K., Puri S., Simard P. High performance convolutional neural networks for document processing //Tenth international workshop on frontiers in handwriting recognition. – Suvisoft, 2006.

[18] He K. et al. Deep residual learning for image recognition //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 770-778.

[19] Liu C. et al. Progressive neural architecture search //Proceedings of the European conference on computer vision (ECCV). – 2018. – С. 19-34.

[20] Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks //International conference on machine learning. – PMLR, 2019. – С. 6105-6114.

[21] Molchanov P. et al. Pruning convolutional neural networks for resource efficient inference //International Conference on Learning Representations (ICLR 2017). – 2017. – С. 1–17.

[22] Molchanov P. et al. Importance estimation for neural network pruning //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2019. – С. 11264-11272.

[23] Casazza P. G., Kutyniok G. Finite frames: Theory and applications. – Springer Science & Business Media, 2012.

[24] Bell H. E. Gershgorin's theorem and the zeros of polynomials //The American Mathematical Monthly. – 1965. – Т. 72. – №. 3. – С. 292-295.

[25] Kingma D. P., Ba J. Adam: A method for stochastic optimization //International Conference on Learning Representations (ICLR 2015). – 2015.

[26] Guo C. et al. Simple black-box adversarial attacks //International Conference on Machine Learning. – PMLR, 2019. – С. 2484-2493.

[27] Krizhevsky A. et al. Learning multiple layers of features from tiny images. – 2009.

[28] Netzer Y. et al. Reading digits in natural images with unsupervised feature learning. – 2011.

[29] Darlow L. N. et al. Cinic-10 is not imagenet or cifar-10 //arXiv preprint arXiv:1810.03505. – 2018.

[30] Hendrycks D., Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations //International Conference on Learning Representations (ICLR 2019). – 2019.

[31] Deng J.et al. ImageNet: A large-scale hierarchical image database //IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – C. 248-255.