Научная статья

DOI: 10.15593/2224-9397/2022.3.08
УДК 004.89

В.В. Бахтин[1,2], И.А. Подлесных[1,3], С.Ф. Тюрин[1,2]

[1]Пермский национальный исследовательский политехнический университет, Пермь, Россия
[2]Пермский государственный национальный исследовательский университет, Пермь, Россия
[3]АО «Новомет-Пермь», Пермь, Россия

РЕШЕНИЕ ЗАДАЧИ МНОГОКРИТЕРИАЛЬНОЙ ОПТИМИЗАЦИИ ВАРИАНТОВ ДЕКОМПОЗИЦИИ НЕЙРОННОЙ СЕТИ И КОМПОНОВКИ КАСКАДА ВЫЧИСЛИТЕЛЬНЫХ УСТРОЙСТВ МЕТОДОМ ПАРЕТО

Повышение функциональности вычислительной системы зачастую необходимо в реальных рабочих задачах на производстве и в цифровой сфере. При этом повышение функциональности зачастую требует приобретения и внедрения в систему дополнительных вычислительных устройств. При этом часто встречаются ситуации, когда вычислительные мощности уже имеющихся устройств используются не полностью. **Целью исследования** является решение многокритериальной задачи оптимизации для поиска наиболее подходящей схемотехнической модели каскада устройств и наиболее оптимальной декомпозиции нейронной сети. **Методика исследования** базируется на схемотехническом моделировании каскадов нейросетевых устройств, программировании декомпозиции блочной нейронной сети, методе Парето-оптимизации, анализе полученных результатов эксперимента. В **результате исследования** планируется получить решение задачи многокритериальной оптимизации вариантов реализации каскада нейросетевых устройств для функционирования блочной нейронной сети. В статье сформулирована и решена задача многокритериальной оптимизации, которая позволяет определить оптимальный способ декомпозиции нейронной сети в рамках метода синтеза устройств. Рассмотрены различные схемы реализации вычислительного каскада с различным числом устройств и различными вычислительными возможностями узлов. Рассмотрены все варианты декомпозиции монолитной нейронной сети, предлагаемой в рамках усовершенствованного метода синтеза, осуществлены экспериментальные запуски каскадов устройств и отобраны наилучшие конфигурации вычислительных каскадов по предложенной совокупности параметров, что подтвердило применимость метода синтеза нейросетевых устройств.

Ключевые слова: схемотехническое моделирование, метод синтеза, распределенная искусственная нейронная сеть, декомпозиция, туманные вычисления, микроконтроллер, многокритериальная оптимизация по Парето.

V.V. Bakhtin[1], I.A. Podlesnykh[1,3], S.F. Tyurin[1,2]

[1]Perm National Research Polytechnic University, Perm, Russian Federation
[2]Perm State University, Perm, Russian Federation
[3]Novomet, Perm, Russian Federation

SOLUTION OF MULTI-CRITERIA OPTIMIZATION PROBLEM OF NEURAL NETWORK DECOMPOSITION OPTIONS AND THE LAYOUT OF A CASCADE OF COMPUTING DEVICES BY THE PARETO METHOD

Improving the functionality of a computing system is often necessary in real work tasks in production and in the digital sphere. At the same time, increasing functionality often requires the acquisition and implementation of additional computing devices into the system. At the same time, there are often situations when the computing power of existing devices is not fully used. **The aim of the study** is to solve a multi-criteria optimization problem for finding the most suitable circuit model of a cascade of devices and finding the most optimal decomposition of a neural network. **The research methodology** is based on circuit modeling of cascades of neural network devices, programming the decomposition of a block neural network, the Pareto optimization method, and analysis of the experimental results. As a **result of the study**, it is planned to obtain a solution to the problem of multi-criteria optimization of options for implementing a cascade of neural network devices for the functioning of a block neural network. The article formulates and solves the problem of multi-criteria optimization, which makes it possible to determine the optimal way to decompose a neural network within the framework of the device synthesis method. Various schemes for implementing a computational cascade with a different number of devices and different computing capabilities of nodes are considered. All variants of the decomposition of a monolithic neural network proposed within the framework of the improved synthesis method are considered, experimental launches of cascades of devices are carried out and the best configurations of computational cascades are selected according to the proposed set of parameters, which confirmed the applicability of the method of synthesis of neural network devices.

Keywords: circuit modeling, synthesis method, distributed artificial neural network, decomposition, fog computing, microcontroller, multi-criteria Pareto optimization.

Введение

Повышение функциональности вычислительной системы зачастую необходимо в реальных рабочих задачах на производстве и в цифровой сфере. При этом повышение функциональности часто требует приобретения и внедрения в систему дополнительных вычислительных устройств, что повлечет выделение дополнительных денежных средств. При этом часто встречаются ситуации, когда вычислительные мощности уже имеющихся устройств используются не полностью. Как следствие возникает противоречие – с одной стороны, тратятся ресурсы на закупку нового оборудования, с другой стороны, имеются не полностью загруженные мощности в рамках вычислительной системы. Поэтому разработка и реализация усовершенствованного метода синтеза нейро-

сетевых устройств для реализации распределенных нейронных сетей, который сможет сбалансировать нагрузку на уже имеющиеся устройства, являются весьма актуальными в настоящее время. Декомпозицию монолитной нейронной сети можно осуществлять различными способами и с различными архитектурами каскада вычислителей. В **результате исследования** планируется получить решение задачи многокритериальной оптимизации вариантов реализации каскада нейросетевых устройств для функционирования блочной нейронной сети.

Искусственные нейронные сети с определенного числа скрытых слоев нейронов могут именоваться глубокими, а их обучение – глубоким обучением. Название и структура нейронных сетей вдохновлены человеческим мозгом и имитируют способ, которым функционируют биологические сети нейронов [1, 2]. Монолитной нейронной сетью называется нейронная сеть, вычисления всех узлов которой производятся на одном устройстве. Блочной нейронной сетью называется нейронная сеть, которая была получена разделением на части монолитной нейронной сети особым образом. Блочная нейронная сеть предназначена для вычисления на физически разделенных устройствах. Так, все узлы одного блока вычисляются на одном устройстве, однако каждый блок целиком выполняется на отдельном вычислительном устройстве. Сеть устройств, на которых вычисляется блочная нейронная сеть, называется каскадом устройств [3].

Распределенные нейронные сети изучают исследователи по всему миру. В частности, такие вопросы поднимаются в работах Pierre-Emmanuel Novac и др. [4], Tu Y и др. [5], Руднева В.А. [6], Nicholas J. Cotton и др. [7] и в ряде других работ ученых, работающих над этой темой. Для нейронных сетей, которые вычисляются на распределенных и(или) беспроводных вычислительных системах [8–10], в настоящее время находится множество видов применения: домашняя автоматизация [11], медицина, лингвистика [12–14], экономика [15], безопасность, управление предприятиями, искусство.

Целью представленного исследования является решение многокритериальной задачи оптимизации, поэтому в данной работе была проведена Парето-оптимизация для поиска наиболее подходящей схемотехнической модели каскада устройств для нейросетевого вычисления блочной нейронной сети и поиска наиболее оптимальной декомпозиции монолитной нейронной сети на блоки. В случае многокритериальной задачи оптимизации применяется так называемая Парето-оптимизация [16]. Смысл данной оптимизации состоит в том, чтобы

найти такую конфигурацию параметров, при которой дальнейшее улучшение какого-либо параметра будет приводить к ухудшению какого-либо другого параметра. Следовательно, изначально необходимо задать иерархию параметров, а затем улучшать каждый из них так, чтобы предыдущий параметр в иерархии не ухудшался.

В работе «Математическая модель искусственной нейронной сети для устройств на ПЛИС и микроконтроллерах, ориентированных на туманные вычисления» [17] была выведена математическая формула, которая описывает разделение нейронных сетей на блоки. Также были описаны различные возможные варианты разделения монолитной нейронной сети на блоки в зависимости от таких параметров, как число устройств в каскаде и вычислительная мощность этих устройств.

В работе «Алгоритм разделения монолитной нейронной сети для реализации туманных вычислений в устройствах на программируемой логике» [18] был описан код программы на языке Java, которая выполняет следующие задачи: разделение монолитной нейронной сети на блочные, запуск и вычисление результатов работы полученной блочной нейронной сети. Кроме того, в работе был представлен новый специально разработанный формат файла, который содержит всю необходимую информацию о любой монолитной или блочной нейронной сети.

В работе «Метод синтеза устройств нейросетевого распознавания на программируемой логике для реализации режима fog computing» [19] была получена методика создания устройств на программируемых логических интегральных схемах, способных проводить туманные вычисления для обеспечения работы нейронных сетей. В статье рассмотрена и сформулирована методика создания устройств и реализованы электрические схемы функциональной работы блочных нейронных сетей на каскадах вычислительных устройств различной размерности. Кроме того, в данной работе были осуществлены экспериментальные запуски каскадов устройств и получены подтверждения работоспособности и эффективности метода синтеза нейросетевых устройств.

1. Существующие методы реализации распределенных нейронных сетей

Первым рассмотрим метод, предлагаемый в работе «Распределенный запуск нейронных сетей на множестве вычислительных узлов» [20], предлагающий рассмотреть реализацию нейронных сетей на множестве вычислительных узлов в кластерной среде.

В данной статье авторы представили воспроизводящуюся искусственную нейронную сеть (ВИНС), которая позволяет создавать модель ассоциативной ячейки памяти. Данная ячейка может увеличивать свою емкость при поступлении новых сигналов на вход. Кроме того, можно менять структуру внутреннего состояния связей ВИНС с помощью особых нейронов роста и нейронов модификации связей [20].

В работе рассматриваются два графа: граф узлов нейронной сети и граф физически связанных вычислительных устройств. Представлен оригинальный алгоритм декомпозиции нейронной сети на кластеры нейронов. Рассмотрено практическое применение алгоритма – через реализацию сервера, который управляет распределением задач, входные и выходные данные доступны на основном узле, который выделен в рамках архитектуры системы, остальные узлы доступны для взаимодействия только основному узлу, узлы в рассматриваемой работе называются серверами. Представленная в данной работе архитектура сети серверов изображена на рис. 1.

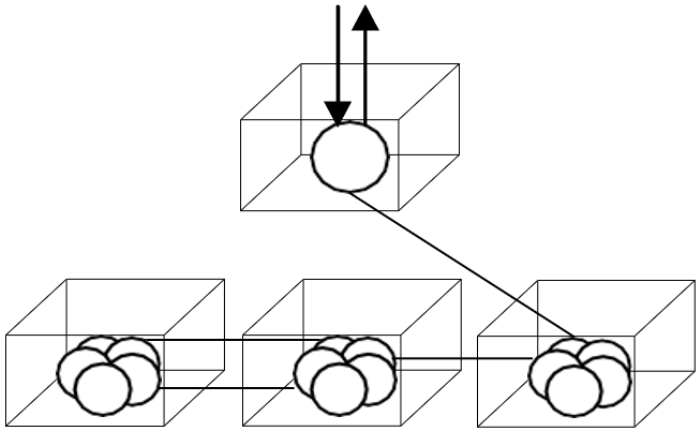

Рис. 1. Архитектура сети серверов

Поскольку нейронная сеть распределяется по вычислительным узлам и данные вычислительные узлы имеют различную вычислительную производительность, появилась проблема синхронизации узлов при работе нейронной сети. Основной вопрос, который решался в данной работе, – поиск оптимального времени ожидания получения сигналов нейроном по его входам. Проблема была решена путем выставления допустимых задержек при передаче данных на основе анализа вычислительных мощностей устройств и пропускной способности каналов связи.

Следующей рассматриваемой работой является «Data Analytics with Deep Neural Networks in Fog Computing Using TensorFlow and Google Cloud Platform» [21], в ней исследователи описывают алгоритм уменьшения количества данных, которые поступают с конечных устройств, перед размещением их на центральный сервер. В данной работе исследователи предлагают заменить узлы тумана на туманные станции (или туманные серверы). Данные туманные станции представляют собой уменьшенную версию облачных серверов. Данные, поступающие на туманные станции, могут быть получены с различных сенсоров и различных устройств, поэтому требуется фильтрация данных, чтобы избежать их смешивания. В данной статье авторы утверждают, что используют библиотеку TensorFlow, чтобы фильтровать данные и сортировать данные, а затем анализировать, нужны ли данные в дальнейшей работе. Таким образом ненужные данные отбрасываются, и тем самым снижается нагрузка на центральный сервер. После того как данные прошли фильтрацию, фильтрующая программа из библиотеки TensorFlow решает, отправить данные прямо на облачный сервер или сначала отправить их на анализ. Существует несколько программ из библиотеки TensorFlow для анализа данных в реальном времени. Данные программы обучены на подходящих наборах данных с использованием машинного обучения. Для обучения модели используется сервис Google Cloud Platform (GCP), который называется Cloud Datalab [22]. Это один из самых мощных инструментов, созданных для анализа данных и построения моделей машинного обучения на GCP.

Авторы приходят к выводу, что их работа принесет значительную пользу, поскольку они фокусируются на применении туманных устройств и предобработке данных перед передачей их на сервер. Они считают, что это может помочь идентифицировать проблемы на ранних стадиях, кроме того, анализ производительности туманных станций позволит обеспечить эффективное распределение ресурсов.

Следующая работа, которую следует рассмотреть, называется «An Efficient Binary Convolutional Neural Network with Numerous Skip Connections for Fog Computing» [23]. В данной работе авторы предлагают систему, в которой общий процесс передачи данных аналогичен системе DeepIns [24]. Однако в данной работе есть два отличия. Во-первых, модель, работающая на туманных узлах в этой системе, представляет собой бинарную глубокую нейронную сеть, которая превосходит про-

стую нейронную сеть в DeepIns. На рис. 2. можно видеть общую структуру предложенной системы для туманных вычислений. Видно, что при заданных ограничениях, чем выше производительность классификации, тем меньше данных требуется передать в облако, что значительно сократит задержку передачи и улучшит коммуникационный трафик. Во-вторых, вместо промежуточных данных в облако передаются необработанные данные из узлов тумана, что может помочь разработчику оптимизировать всю систему.

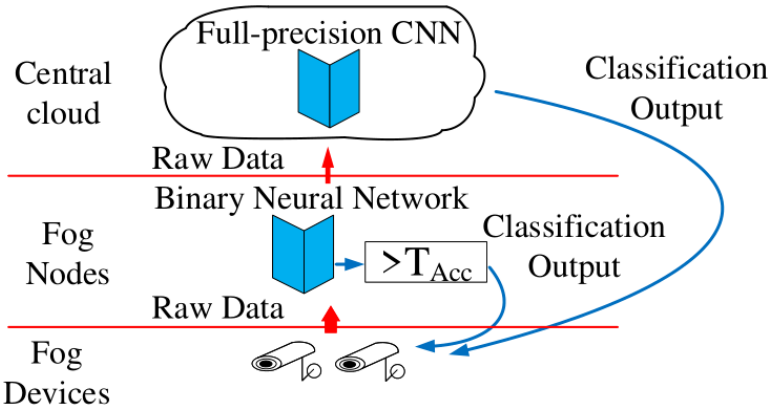

Рис. 2. Общая структура предложенной системы для туманных вычислений

В данном исследовании авторы предлагают решить проблемы с задержкой передачи данных в приложениях глубокого обучения в интеллектуальных системах путем внедрения бинарных нейронных сетей (BNN) в туманные вычисления. Новый BNN под названием BNSC-Net был разработан для повышения точности. Исследователи обнаружили, что точность BNN пропорциональна количеству «пропущенных соединений» в сети. Повторное использование операции конкатенации также может повысить производительность BNN. Кроме того, исследователи обнаружили взаимосвязь между интервалом обновления Straigt-Through-Estimator (STE) и точностью BNN и предложили оптимальный интервал обновления. Чтобы выявить эффективность BNSC-Net, были проведены сравнительные эксперименты между BNSC-Net и современной системой DeepIns. Результаты показывают, что BNSC-Net превосходит DeepIns и эффективен для развертывания в туманных вычислениях.

2. Постановка задачи многокритериальной оптимизации

Предлагаемый метод обеспечивает поиск оптимального по стоимости, тепловыделению и дополнительным затратам электроэнергии решения для декомпозиции нейронной сети на каскад распределенных вычислителей при заданных ограничениях на выбранную архитектуру вычислительной системы (число доступных вычислительных устройств n), их вычислительные мощности $\bar{p}$ и пропускной способности каналов связи q:

1) оценкой стоимости ***C*:**

$$C = \{C_{\tau_1}(n, \bar{p}, q), C_{\tau_2}(n, \bar{p}, q), \dots . C_{\tau_e}(n, \bar{p}, q)\}. \quad (1)$$

2) оценкой времени выполнения нейросетевых вычислений T^1:

$$T^1 = \{T^1_{\tau_1}(n, \bar{p}, q), T^1_{\tau_2}(n, \bar{p}, q), \dots, T^1_{\tau_e}(n, \bar{p}, q)\}. \quad (2)$$

3) оценкой среднего времени отклика пульта на входные сигналы T^2:

$$T^2 = \{T^2_{\tau_1}(n, \bar{p}, q), T^2_{\tau_2}(n, \bar{p}, q), \dots, T^2_{\tau_e}(n, \bar{p}, q)\}. \quad (3)$$

4) оценкой итогового энергопотребления (мощности) *W*:

$$W = \{W_{\tau_1}(n, \bar{p}, q), W_{\tau_2}(n, \bar{p}, q), \dots . W_{\tau_e}(n, \bar{p}, q)\}. \quad (4)$$

При получении оценок необходимо учесть существующие ограничения современных вычислительных устройств. Для подтверждения работоспособности методов выполнить схемотехническое и физическое моделирование решений (декомпозиций) $\Omega = \{\omega_1, \omega_2, \dots \omega_u\}$.

Получить оптимальный набор H элементов можно, используя метод Парето-оптимизации:

$$H = \langle \langle \omega_1(n, \bar{p}, q) \rangle, \langle \omega_2(n, \bar{p}, q) \rangle, \dots \langle \omega_u(n, \bar{p}, q) \rangle \rangle,$$

такой, что $C(H) \to \min$, $T^2(H) \to \min$ без ухудшения $C(H)$,

$T^1(H) \to \min$ без ухудшения $C(H)$ *и* $T^2(H)$,

$W(H) \to \min$ без ухудшения $C(H)$, $T^2(H)$ *и* $T^1(H)$.

В данном исследовании осуществлялась декомпозиция нейронной сети на несколько устройств с поиском оптимума по параметрам: стоимость, тепловыделение, энергопотребление (при прочих равных приоритет будет отдан варианту, в котором время выполнения будет лучшим).

Находить искомые параметры мы будем по следующим формулам:

1) оценкой стоимости *C*:

$$C = \sum_{i=0}^{n-1} C_i; \quad (5)$$

2) оценкой времени выполнения нейросетевых вычислений T^1:

$$T^1 = \sum_{i=0}^{n-1} T_{i_i}^{\text{exec}} + \sum_{j=1}^{n-1} T_j^{\text{send}}, \quad T_{i_i}^{\text{send}} = \frac{I}{q};$$

$$T^1 = \sum_{i=0}^{n-1} T_{i_i}^{\text{exec}} + \sum_{j=1}^{n-1} \frac{I_j}{q_j}, \quad T_{i_i}^{\text{exec}} = \frac{N}{\nu}, T^1 = \sum_{i=0}^{n-1} \frac{N_i}{\nu_i} + \sum_{j=1}^{n-1} \frac{I_j}{q_j}; \quad (6)$$

3) оценкой среднего времени отклика пульта на входные сигналы T^2: $T^2 = \overline{T_{\text{pult}}} + T_{\text{pult}}^{\text{exec}}$;

4) оценкой итогового энергопотребления (мощности) W:

$$W = \sum_{i=0}^{n-1} W_i. \quad (7)$$

Оценка проводилась на декомпозиции конкретной нейронной сети пятью способами для трех различных конфигураций оборудования: одного без изменений и двух с установкой дополнительного вычислительного устройства. В текущей конфигурации вычислительная система состоит из следующих элементов: микроконтроллер Atmel AT91SAM7X256-AU, коммутатор D-Link DGS-1100-05PDV2, одноплатный микрокомпьютер Raspberry PI Zero, 3-портовый управляемый коммутатор 10/100 Ethernet KSZ8993M. В одном случае предлагается добавить микроконтроллер ATMega32, во втором – еще один одноплатный компьютер Raspberry PI Model 4 B.

Ниже в табл. 1. представлены исходные значения параметра «стоимость» C.

Таблица 1

Исходные значения параметра «стоимость» *C*

Стоимость	C (руб.)	ΔC (руб.)	ΔC (%)
Схема 1 (Исходная)	10601	0	0
Схема 2 (Добавлено у-во 1)	12211	1610	15,18725
Схема 3 (Добавлено у-во 2)	15232	4631	43,68456
Схема 4 (Добавлено у-во 3)	21501	10900	102,8205

Исходные значения параметра «энергопотребление» *W* в табл. 2.

Таблица 2

Исходные значения параметра «энергопотребление» *W*

Энергопотребление	*W* (Вт)	Δ*W* (Вт)	Δ*W* (%)
Схема 1 (Исходная)	6,676	0	0
Схема 2 (Добавлено у-во 1)	6,811	0,135	2,022169
Схема 3 (Добавлено у-во 2)	7,526	0,85	12,73217
Схема 4 (Добавлено у-во 3)	10,176	3,5	52,4266

Полученные значения параметра T^1 представлены в табл. 3.

Таблица 3

Значения параметра T^1

	$T_{исх}$ (мс)	T1 (мс)	T2 (мс)	T3 (мс)
Монолитная сеть на устройстве	–	139	57	22
Способ 1 (Одинаково слоев)	104	125	95	82
Способ 2 (Пропорционально слоев)	86	107	85	53
Способ 3 (Одинаково нейронов)	103	123	94	82
Способ 4 (Пропорционально нейронов)	84	104	85	52
Способ 5 (Минимум отправки данных)	89	102	91	80

Значения параметра T^2 представлены в табл. 4.

Таблица 4

Значения параметра T^2

	$T_{исх}$ (мс)	T1 (мс)	T2 (мс)	T3 (мс)
Монолитная сеть на устройстве	65	–	–	–
Способ 1 (Одинаково слоев)	134	111	111	111
Способ 2 (Пропорционально слоев)	105	96	89	80
Способ 3 (Одинаково нейронов)	134	109	109	109
Способ 4 (Пропорционально нейронов)	104	95	89	80
Способ 5 (Минимум отправки данных)	110	98	98	98

3. Результаты эксперимента

Было проведено несколько запусков различных схемотехнических моделей и физических стендов. Выполнено разделение монолитной нейронной сети на блочную для вычисления ее работы разными способами на сети устройств, полученной в результате схемотехнического моделирования. Входные данные, подаваемые на каждую сеть, были одинаковыми и получены путем генерации слов. Были произведены замеры показателей времени выполнения нейросетевых вычислений и времени отклика пультов голосования. Исходя из представленных данных, можно приступить к Парето-оптимизации и выбрать наиболее оптимальный вариант для его последующего использования. На рис. 3 представлена точечная диаграмма, каждый из узлов – это один из экспериментальных случаев по времени работы, представленных в табл. 3 и 4. Цветовая легенда показывает, к какой из четырех архитектур каскада относится каждый случай.

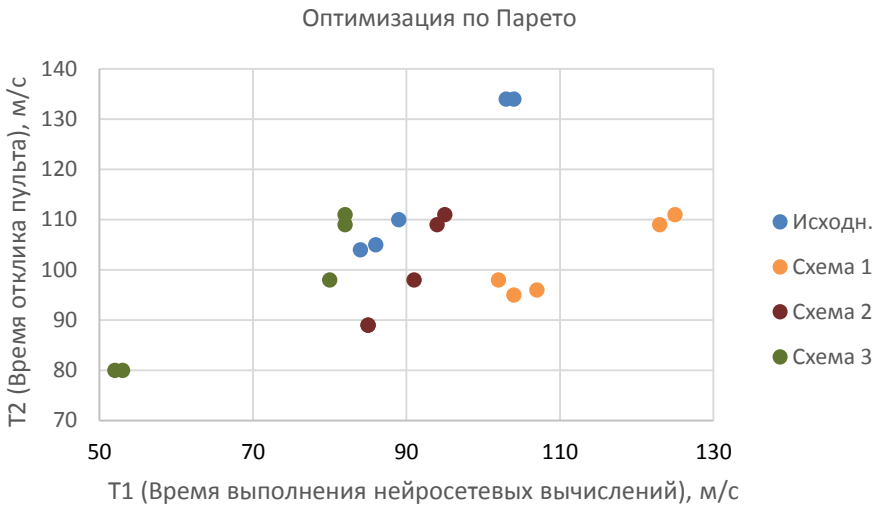

Рис. 3. Точечная диаграмма экспериментов (по схеме сборки каскада)

На рис. 4 представлена та же точечная диаграмма, только цветовая легенда показывает, к какому из пяти способов декомпозиции относится каждый случай.

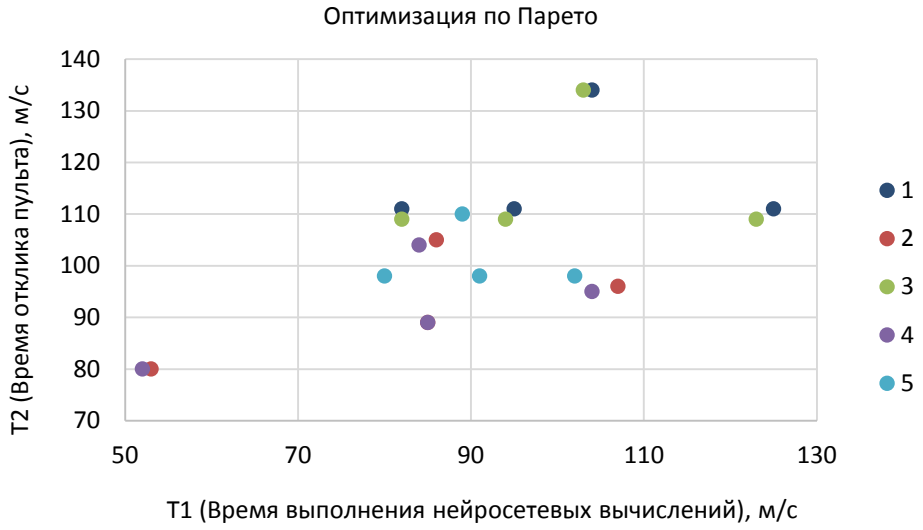

Рис. 4. Точечная диаграмма экспериментов (по способу декомпозиции НС)

Таким образом, самым оптимальным вариантом в двухкритериальной оптимизации будет вариант на архитектурной схеме номер 3, с декомпозицией по 4-му способу. Но у нас есть еще два параметра оп-

тимизации, один из которых является ключевым в рассматриваемой задаче, по параметру стоимости лидирует исходная архитектура вычислительной системы, поскольку для нее нет необходимости закупать дополнительное оборудование. Поэтому в случае с четырьмя критериями оптимизации выигрывает уже вариант с исходной архитектурой, в котором НС также подверглась декомпозиции по 4-му способу. Хотя на рис. 5 видно, что в зону улучшения по двум временным параметрам попадают еще 3 решения.

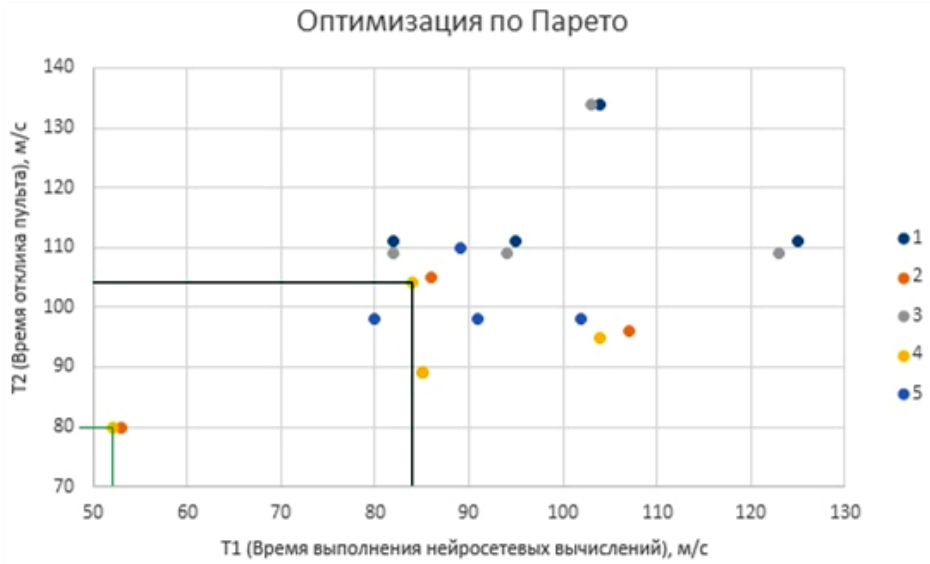

Рис. 5. Парето-оптимальное решение в двухкритериальной задаче
и Парето-оптимальное решение в четырехкритериальной задаче

Наиболее оптимальным вариантом в предлагаемой модели оптимизации по требуемым критериям вариант в схеме с сохранением исходной архитектуры и декомпозицией нейронной сети на блоки способом номер 4, т.е. с разделением на блоки пропорционально мощности с точностью до нейронов на слоях, потому что эта точка принадлежит к исходной схеме физической архитектуры, которая является самой оптимальной по критерию стоимости (табл. 1). Полученные результаты свидетельствуют о том, что в рамках предложенного метода и различных способов его реализации возможно провести оптимизацию и выбрать наиболее удачный сценарий декомпозиции нейронной сети. Полученные результаты подтверждают эффективность предложенного метода.

4. Результаты исследования

Результатом данной работы является метод создания устройств, на которых возможна реализация искусственных нейронных сетей, ориентированных на туманные вычисления, т.е. таких устройств, которые позволяют обрабатывать процесс работы нейронной сети. В рамках данной статьи были успешно решены задача оптимизации вариантов разделения нейронной сети на блоки с большим числом критериев, а также задача планирования и составления каскада вычислительных устройств методом Парето. Были получены оптимальные варианты реализации каскада нейросетевых устройств в зависимости от иерархии параметров оценки полученных вариантов.

Еще одним важным результатом является успешное тестирование распознавания при помощи нейронной сети, исполняемой на устройствах, а также измерение результирующих параметров устройств. Эти параметры показывают, что полученный в данной работе метод исправно функционирует и способен эффективно решать задачу.

В дальнейшем планируется продолжить данные исследования в направлении модификации предложенного метода синтеза таким образом, чтобы повысить отказоустойчивость синтезированного каскада устройств, например, внедрением дополнительных резервных вычислительных устройств или каналов связи. Кроме того, планируется модификация данного метода для возможности организации рекуррентной нейронной сети.

Заключение

Целью данной работы являлось решение многокритериальной задачи оптимизации, поэтому в данной работе была проведена Парето-оптимизация для поиска наиболее подходящей схемотехнической модели каскада устройств для нейросетевого вычисления блочной нейронной сети и поиска наиболее оптимальной декомпозиции монолитной нейронной сети на блоки. На вход данным моделям подавались одинаковые данные. Оптимизация проводилась по четырем параметрам: суммарной стоимости устройств каскада, энергопотребления каскада устройств, времени выполнения нейросетевых вычислений и времени отклика пульта. В результате данного исследования выяснилось, что оптимальным является разложение на блоки пропорционально мощности с точностью до нейронов на слоях.

Научная новизна данного исследования состоит в том, что появилась возможность определения максимально оптимальной конфигурации каскада устройств и конфигурации блочной нейронной сети для вычисления исходной нейронной сети. Чтобы определить, какое из разложений и на каком именно физическом каскаде вычислительных устройств наиболее эффективно, и потребовалось решение задачи оптимизации с множеством критериев. Пример решения такой задачи был рассмотрен в представленной работе.

Библиографический список

1. Deep learning in the fog / A. Sobecki [et al.] // International Journal of Distributed Sensor Networks. – 2019. – Vol. 15, № 8. DOI: 10.1177/1550147719867072

2. Akhmetzyanov K.R., Yuzhakov A.A., Kokoulin A.N. Neural Network Development Based on Knowledge about Environmental Influence // 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). – 2020. – P. 218–221. DOI: 10.1109/EIConRus49466.2020.9039226

3. Бахтин В.В., Подлесных И.А. Алгоритм построения графа совместной работы каскадов устройств нейросетевого распознавания, реализующих блочные нейронные сети // Сб. материалов IX Междунар. науч. конф., посв. 85-лет. проф. В.И. Потапова. – Омск, 2021. – С. 277–278.

4. Quantization and deployment of deep neural networks on microcontrollers / P.E. Novac [et al.] // Sensors. – 2021. – Vol. 21, № 9. – P. 2984. DOI: 10.3390/s21092984

5. A power efficient neural network implementation on heterogeneous FPGA and GPU devices / Y. Tu [et al.] // 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). – 2019. – P. 193–199. DOI: 10.1109/IRI.2019.00040

6. Руднев В.А. Применение микроконтроллеров для реализации нейронных сетей // Вестник Южно-Урал. гос. ун-та. Сер.: Компьютерные технологии, управление, радиоэлектроника. – 2012. – № 23. – С. 181–183.

7. Cotton N.J., Wilamowski B.M., Dundar G. A neural network implementation on an inexpensive eight bit microcontroller // 2008 International Conference on Intelligent Engineering Systems. – 2008. – P. 109–114. DOI: 10.1109/INES.2008.4481278

8. Борзов Д.Б., Дюбрюкс С.А., Соколова Ю.В. Метод и методика беспроводной передачи данных в мультипроцессорных системах для нестационарных объектов обмена // Труды МАИ. – 2020. – № 114. DOI: 10.34759/trd-2020-114-13

9. Виноградов Г.П., Емцев А.С. Методы локализации и отслеживания целей в беспроводных сенсорных сетях // Перспективные системы и задачи управления: материалы XVII Всерос. науч.-практ. конф.; Управление и обработка информации в технических системах: материалы XIII Молодеж. школы-сем. / Южный федер. ун-т. – Ростов-н/Д, 2022. – С. 52–63.

10. Виноградов Г.П. Отслеживание мобильных объектов средствами беспроводных сенсорных сетей // Информационные и математические технологии в науке и управлении. – 2022. – № 1 (25). – С. 58–69.

11. An application placement technique for concurrent IoT applications in edge and fog computing environments / M. Goudarzi [et al.] // IEEE Transactions on Mobile Computing. – 2020. – Vol. 20, № 4. – P. 1298–1311. DOI: 10.1109/TMC.2020.2967041

12. Бахтин В.В. Модификация алгоритма идентификации и категоризации научных терминов с использованием нейронной сети // Нейрокомпьютеры: разработка, применение. – 2019. – Т. 21, № 3. – С. 14–19.

13. Bakhtin V.V., Isaeva E.V. New TSBuilder: shifting towards cognition // 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). – 2019. – P. 179–181. DOI: 10.1109/EIConRus.2019.8656917

14. Bakhtin V.V., Isaeva E.V., Tararkov A.V. TSMiner: from TSBuilder to Ecosystem // 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). – 2021. – P. 221–224. DOI: 10.1109/ElConRus51938.2021.9396569

15. Yasnitsky L.N., Yasnitsky V.L. Technique of design for integrated economic and mathematical model for mass appraisal of real estate property. Study case of Yekaterinburg housing market // Journal of Applied Economic Sciences. – 2016. – Vol. 11, № 8. – P. 1519–1530.

16. Подиновский В.В., Ногин В.Д. Парето-оптимальные решения многокритериальных задач. – М.: Наука; Гл. редакция физ.-мат. лит-ры, 1982. – 256 с.

17. Бахтин В.В. Математическая модель искусственной нейронной сети для устройств на ПЛИС и микроконтроллерах, ориентированных на туманные вычисления // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2021. – № 40. – С. 109–129.

18. Бахтин В.В. Алгоритм разделения монолитной нейронной сети для реализации туманных вычислений в устройствах на программируемой логике // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2022. – № 41. – С. 123–145.

19. Тюрин С.Ф., Бахтин В.В., Подлесных И.А. Метод синтеза устройств нейросетевого распознавания на программируемой логике для реализации режима fog computing // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2022. – № 41. – С. 168–188.

20. Ионов С.Д. Распределенный запуск нейронных сетей на множестве вычислительных узлов // Вестник УГАТУ. – 2013. – № 2 (55).

21. Priyabhashana H.M.B., Jayasena K.P.N. Data Analytics with Deep Neural Networks in Fog Computing Using TensorFlow and Google Cloud Platform // 2019 14th Conference on Industrial and Information Systems (ICIIS). – IEEE, 2019. – P. 34–39.

22. Kumar M. Google cloud platform: a powerful big data analytics cloud platform // Int J. Res. Appl. Sci. Eng. Technol. – 2016. – Vol. 4, № 11. – P. 387–392.

23. An efficient binary convolutional neural network with numerous skip connections for fog computing / L. Wu [et al.] // IEEE Internet of Things Journal. – 2021. – Vol. 8, № 14. – P. 11357–11367.

24. Li L., Ota K., Dong M. Deep learning for smart industry: Efficient manufacture inspection system with fog computing // IEEE Transactions on Industrial Informatics. – 2018. – Vol. 14, № 10. – P. 4665–4673.

References

1. Sobecki A. et al. Deep learning in the fog. *International Journal of Distributed Sensor Networks*, 2019, vol. 15, no. 8. DOI: 10.1177/1550147719867072

2. Akhmetzyanov K.R., Yuzhakov A.A., Kokoulin A.N. Neural Network Development Based on Knowledge about Environmental Influence. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2020, pp. 218-221, DOI: 10.1109/EIConRus49466.2020.9039226

3. Bakhtin V.V., Podlesnykh I.A. Algoritm postroeniia grafa sovmestnoi raboty kaskadov ustroistv neirosetevogo raspoznavaniia, realizuiushchikh blochnye neironnye seti [Algorithm for constructing a coworking cascades graph of neural network recognition devices implementing block neural networks]. *Sbornik materialov IX Mezhdunarodnoi nauchnoi konferentsii, posviashchennoi 85-letiiu professora V.I. Potapova*. Омск, 2021, pp. 277-278.

4. Novac P.E. et al. Quantization and deployment of deep neural networks on microcontrollers. Sensors, 2021, vol. 21, no. 9, 2984 p. DOI: 10.3390/s21092984

5. Tu Y. et al. A power efficient neural network implementation on heterogeneous FPGA and GPU devices. *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 2019, pp. 193-199. DOI: 10.1109/IRI.2019.00040

6. Rudnev V.A. Primenenie mikrokontrollerov dlia realizatsii neironnykh setei [The use of microcontrollers for the implementation of neural networks]. *Vestnik Iuzhno-Ural'skogo gosudarstvennogo universiteta. Komp'iuternye tekhnologii, upravlenie, radioelektronika*, 2012, no. 23, pp. 181-183.

7. Cotton N.J., Wilamowski B.M., Dundar G. A neural network implementation on an inexpensive eight bit microcontroller. *2008 International Conference on Intelligent Engineering Systems*, 2008, pp. 109-114. DOI: 10.1109/INES.2008.4481278

8. Borzov D.B., Diubriuks S.A., Sokolova Iu.V. Metod i metodika besprovodnoi peredachi dannykh v mul'tiprotsessornykh sistemakh dlia nestatsionarnykh ob"ektov obmena [Method and methodology of wireless data transmission in multiprocessor systems for non-stationary exchange objects] *Trudy Moskovskii aviatsionnyi institut*, 2020, no. 114. DOI: 10.34759/trd-2020-114-13

9. Vinogradov G.P., Emtsev A.S. Metody lokalizatsii i otslezhivaniia tselei v besprovodnykh sensornykh setiakh [Methods of localization and tracking of targets in wireless sensor networks]. *Perspektivnye sistemy i zadachi upravleniia. Materialy XVII Vserossiiskoi nauchno-prakticheskoi konferentsii; Upravlenie i obrabotka informatsii v tekhnicheskikh sistemakh. Materialy XIII Molodezhnoi shkoly-seminara*. Rostov-n/D: Iuzhnyi federal'nyi universitet, 2022, pp. 52-63.

10. Vinogradov G.P. Otslezhivanie mobil'nykh ob"ektov sredstvami besprovodnykh sensornykh setei [Tracking of mobile objects by means of wireless sensor networks]. *Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii*, 2022, no. 1 (25), pp. 58-69.

11. Goudarzi M. et al. An application placement technique for concurrent IoT applications in edge and fog computing environments. *IEEE Transactions on Mobile Computing*, 2020, Vol. 20, no. 4, pp. 1298-1311. DOI: 10.1109/TMC.2020.2967041

12. Bakhtin V.V. Modifikatsiia algoritma identifikatsii i kategorizatsii nauchnykh terminov s ispol'zovaniem neironnoi seti [Modification of the algorithm for identification and categorization of scientific terms using a neural network]. *Neirokomp'iutery: razrabotka, primenenie*, 2019, vol. 21, no. 3, pp. 14-19.

13. Bakhtin V.V., Isaeva E.V. New TSBuilder: shifting towards cognition. *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2019, pp. 179-181. DOI: 10.1109/EIConRus.2019.8656917

14. Bakhtin V.V., Isaeva E.V., Tararkov A.V. TSMiner: from TSBuilder to Ecosystem // 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). – 2021. – P. 221–224. DOI: 10.1109/ElConRus51938.2021.9396569

15. Yasnitsky L.N., Yasnitsky V.L. Technique of design for integrated economic and mathematical model for mass appraisal of real estate property. Study case of Yekaterinburg housing market. *Journal of Applied Economic Sciences*, 2016, vol. 11, no. 8, pp. 1519-1530.

16. Podinovskii V.V., Nogin V.D. Pareto-optimal'nye resheniia mnogokriterial'nykh zadach [Pareto-optimal solutions to multiobjective problems]. Moscow: Nauka; Glavnaia redaktsiia fiziko-matematicheskoi literatury, 1982, 256 p.

17. Bakhtin V.V. Matematicheskaia model' iskusstvennoi neironnoi seti dlia ustroistv na PLIS i mikrokontrollerakh, orientirovannykh na tumannye vychisleniia [Mathematical model of an artificial neural network for FPGA devices and microcontrollers focused on fog computing]. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniia*, 2021, no. 40, pp. 109-129.

18. Bakhtin V.V. Algoritm razdeleniia monolitnoi neironnoi seti dlia realizatsii tumannykh vychislenii v ustroistvakh na programmiruemoi logike [Monolithic neural network separation algorithm for implementing fog computing in programmable logic devices]. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniia*, 2022, no. 41, pp. 123-145.

19. Tiurin S.F., Bakhtin V.V., Podlesnykh I.A. Metod sinteza ustroistv neirosetevogo raspoznavaniia na programmiruemoi logike dlia realizatsii rezhima fog computing [Method of synthesis of neural network recognition devices based on programmable logic for the implementation of fog computing]. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniia*, 2022, no. 41, pp. 168-188.

20. Ionov S.D. Raspredelennyi zapusk neironnykh setei na mnozhestve vychislitel'nykh uzlov [Distributed launch of neural networks on a set of computing nodes]. *Vestnik Ufimskogo gosudarstvennogo aviatsionnogo tekhnicheskogo universiteta*, 2013, no. 2 (55).

21. Priyabhashana H.M.B., Jayasena K.P.N. Data Analytics with Deep Neural Networks in Fog Computing Using TensorFlow and Google Cloud Platform. *2019 14th Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2019, pp. 34-39.

22. Kumar M. Google cloud platform: a powerful big data analytics cloud platform. *Int J Res Appl. Sci. Eng. Technol.*, 2016, vol. 4, no. 11, pp. 387-392.

23. Wu L. et al. An efficient binary convolutional neural network with numerous skip connections for fog computing. *IEEE Internet of Things Journal.*, 2021, vol. 8, no. 14, pp. 11357-11367.

24. Li L., Ota K., Dong M. Deep learning for smart industry: Efficient manufacture inspection system with fog computing. *IEEE Transactions on Industrial Informatics*, 2018, vol. 14, no. 10, pp. 4665-4673.

Сведения об авторах

Бахтин Вадим Вячеславович (Пермь, Россия) – аспирант, младший научный сотрудник кафедры «Автоматика и телемеханика» Пермского национального исследовательского политехнического университета (614990, Пермь, Комсомольский пр., 29, e-mail: bakhtin_94@bk.ru); старший преподаватель кафедры «Информационная безопасность и системы связи» Пермского государственного национального исследовательского университета (614990, Пермь, ул. Букирева, 15).

Подлесных Иван Александрович (Пермь, Россия) – аспирант кафедры «Автоматика и телемеханика» Пермского национального исследовательского политехнического университета (614990, Пермь, Комсомольский пр., 29, e-mail: podlesnihwork@gmail.com); математик ДИР ИТЦ АО «Новомет-Пермь» (614065, Пермь, ш. Космонавтов, 395).

Тюрин Сергей Феофентович (Пермь, Россия) – заслуженный изобретатель Российской Федерации, доктор технических наук, профессор, профессор кафедры «Автоматика и телемеханика» Пермского национального исследовательского политехнического университета (614990, Пермь, Комсомольский пр., 29, e-mail: tyurinsergfeo@yandex.ru); профессор кафедры «Математическое обеспечение вычислительных систем» Пермского государственного национального исследовательского университета (614990, Пермь, ул. Букирева, 15).

About the authors

Vadim V. Bakhtin (Perm, Russian Federation) – Graduate Student, junior researcher of the Department of Automation and Telemechanics Perm National Research Polytechnic University (614990, Perm, 29, Komsomolsky pr., e-mail: bakhtin_94@bk.ru), Senior lecturerat the Department of Information Security and Communication Systems Perm State University (614990, Perm, 15, Bukireva str.).

Ivan A. Podlesnykh (Perm, Russian Federation) – Graduate Student of the Department of Automation and Telemechanics Perm National Research Polytechnic University (614990, Perm, 29, Komsomolsky pr., e-mail: podlesnihwork@gmail.com), Mathematician at Novomet (614065, Perm, 395, shosse Kosmonavtov).

Sergey F. Tyurin (Perm, Russian Federation) – Honored Inventor of the Russian Federation, Doctor of Technical Sciences, Professor, Professor at the Department of Automation and Telemechanics Perm National Research Polytechnic University (614990, Perm, 29, Komsomolsky pr., e-mail: tyurinsergfeo@yandex.ru), Professor at the Department of Software Computing Systems Perm State University (614990, Perm, 15, Bukireva str.).

Поступила: 10.10.2022. Одобрена: 22.10.2022. Принята к публикации: 22.12.2022.

Финансирование. Исследование проводится при поддержке РФФИ на средства гранта № 20-37-90036 Аспиранты «Метод синтеза устройств нейросетевого распознавания для реализации режима Fog computing».

Конфликт интересов. Конфликт интересов по отношению к статье отсутствует.

Вклад авторов. Все авторы внесли равный вклад в написание статьи.

Просьба ссылаться на эту статью в русскоязычных источниках следующим образом:

Бахтин, В.В. Решение задачи многокритериальной оптимизации вариантов декомпозиции нейронной сети и компоновки каскада вычислительных устройств методом Парето / В.В. Бахтин, И.А. Подлесных, С.Ф. Тюрин // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. – 2022. – № 43. – С. 136–156. DOI: 10.15593/2224-9397/2022.3.08

Please cite this article in English as:

Bakhtin V.V., Podlesnykh I.A., Tyurin S.F. Solution of multi-criteria optimization problem of neural network decomposition options and the layout of a cascade of computing devices by the Pareto method. *Perm National Research Polytechnic University Bulletin. Electrotechnics, information technologies, control systems*, 2022, no. 43, pp. 136-156. DOI: 10.15593/2224-9397/2022.3.08