УДК 004.852
DOI: 10.26102/2310-6018/2025.51.4.054

Разработка гибридной модели глубокого обучения для прогнозирования рабочей нагрузки в виртуализированных центрах обработки данных

Б.В. Мартыненков✉

МИРЭА – Российский технологический университет, Москва, Российская Федерация

Резюме. Актуальность исследования обусловлена необходимостью решения задачи проактивного управления рабочей нагрузкой центров обработки данных, базирующихся на технологиях виртуальных машин и контейнеризации приложений. Решение подобной задачи связано с анализом ретроспективных данных рабочей нагрузки, консолидированных в виде временных рядов по используемым за заданный период времени вычислительным ресурсам, таким как загруженность пулов процессоров и оперативной памяти и сохраняемых подсистемой мониторинга ресурсов службы администрирования центра обработки данных. В связи с этим работа направлена на исследование методов и технологий машинного обучения, поддерживающих решение задачи прогнозирования временных рядов. В статье делается обобщение особенностей моделей машинного обучения, основанных на статистических подходах и принципах глубокого обучения. Рассматриваются структурные и функциональные компоненты вариантов нейронных сетей, специализированных на анализе зависимостей во временных рядах и решении задач их прогнозирования. В качестве предлагаемого решения представлена гибридная схема системы глубокого обучения, основанная на последовательном применении каскадов одномерных сверточных нейронных сетей и двунаправленных сетей с долгой краткосрочной памятью. Предлагаются подходы к выбору их структурно-параметрических характеристик. Приводятся результаты сравнительной экспериментальной оценки предлагаемого решения с реализацией системы прогнозирования рабочей нагрузки, основанной на методах статистического прогнозирования.

Ключевые слова: виртуализированный центр обработки данных, рабочая нагрузка, прогнозирование временных рядов, машинное обучение, глубокое обучение, одномерные сверточные нейронные сети, двунаправленные сети с долгой краткосрочной памятью.

Для цитирования: Мартыненков Б.В. Разработка гибридной модели глубокого обучения для прогнозирования рабочей нагрузки в виртуализированных центрах обработки данных. *Моделирование, оптимизация и информационные технологии.* 2025;13(4). URL: https://moitvivt.ru/ru/journal/pdf?id=2097 DOI: 10.26102/2310-6018/2025.51.4.054

Developing a hybrid deep learning model for workload prediction in virtualized data centers

B.V. Martynenkov✉

MIREA – Russian Technological University, Moscow, the Russian Federation

Abstract. The relevance of this research stems from the need to proactively manage the workload of data centers based on virtual machine technologies and application containerization. This task requires analyzing historical workload data consolidated as time series for computing resources used over a given period of time, such as CPU and RAM pool utilization and stored by the resource monitoring subsystem of the data center administration service. Therefore, this article aims to explore machine learning methods and technologies that support time series forecasting. The article summarizes the features of machine learning models based on statistical approaches and deep learning principles. It examines the structural and functional components of neural network variants specialized in analyzing time series

dependencies and solving forecasting problems. The proposed solution is a hybrid deep learning system based on the sequential application of cascades of one-dimensional convolutional neural networks and bidirectional long short-term memory networks. Approaches to the selection of their structural and parametric characteristics are proposed. The results of a comparative experimental evaluation of the proposed solution with the implementation of a workload prediction system, based on statistical prediction methods are presented.

Keywords: virtualized data center, workload, time series forecasting, machine learning, deep learning, one-dimensional convolutional neural networks, bidirectional long short-term memory networks.

For citation: Martynenkov B.V. Developing a hybrid deep learning model for workload prediction in virtualized data centers. *Modeling, Optimization and Information Technology*. 2025;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=2097 DOI: 10.26102/2310-6018/2025.51.4.054

Введение

Современные центры обработки данных (ЦОД) являются сложными, распределенными, многопроцессорными и многокластерными вычислительными структурами с широкими областями применения: от решения узкоспециализированных задач до применения в качестве систем общего назначения. В последнем случае современные ЦОД все чаще применяются для реализации таких моделей по запросу (on demand model), как SaaS (Software as a Service) и PaaS (Platform as a Service) [1], которые поддерживают гибкие технологические процессы, а также служат базой для использования набирающих популярность систем искусственного интеллекта разного уровня специализации.

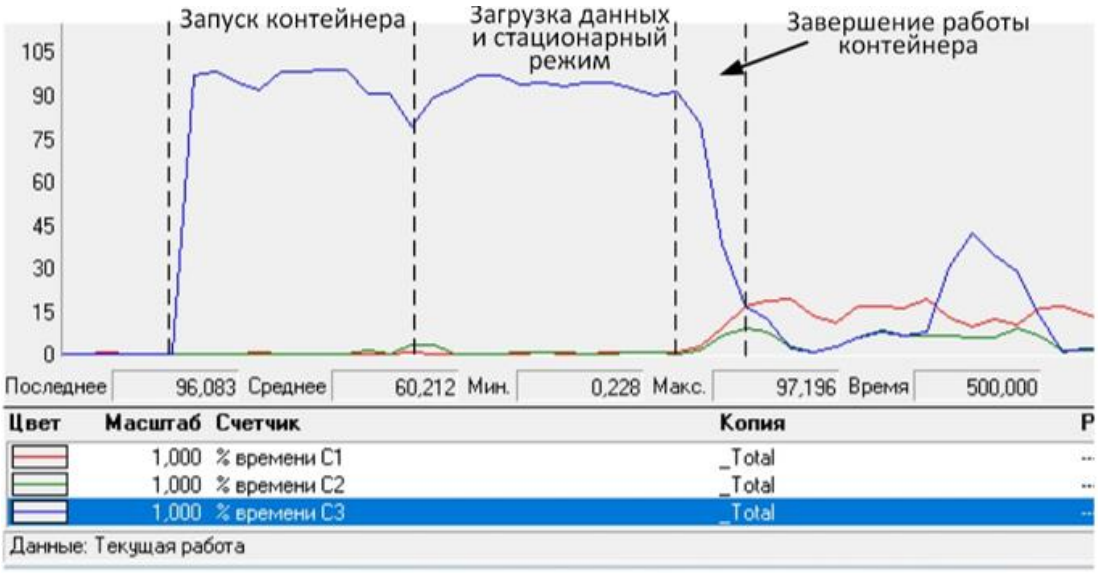
Поскольку функциональной основой моделей SaaS и PaaS является адаптивность поддерживающих их вычислительных структур к потребительской запросной нагрузке, наиболее востребованной схемой реализации современных ЦОД является виртуализация и/или контейнеризация их вычислительных ресурсов (пулы процессорных ядер, оперативной памяти, систем хранения данных, сетевых и иных ресурсов). ЦОД, поддерживающие подобные решения обычно именуются виртуализированными (ВЦОД) [2]. В качестве наиболее известных российских и зарубежных ВЦОД можно представить Yandex Cloud, SberCloud, Amazon Web Services и Microsoft Azure.

В общем случае структурно-параметрические характеристики ВЦОД, соответствующие текущей потребительской запросной нагрузке, именуются рабочей нагрузкой (workload). Очевидно, что рабочая нагрузка ВЦОД, в общем случае, должна быть адекватна запросной нагрузке, показатели которой могут носить, как периодический (сезонный: время суток, время года), так и случайный (реакция на события и т. д.) характер. Кроме того, на характеристики рабочей нагрузки оказывают влияние как функциональные аспекты ВЦОД, например, латентность механизмов живой миграции виртуальных машин/контейнеров, так и их технологические аспекты, такие как сбои и отказы аппаратного и программного обеспечения. В общем случае можно говорить о наличии некоторых типовых и нетиповых шаблонов рабочей нагрузки (workload patterns). Примеры периодических шаблонов рабочей нагрузки по показателю «% использования времени процессора» представлены на Рисунке 1.

Управление рабочей нагрузкой осуществляется службой администрирования ВЦОД и основано на решении двух классов задач: мониторинга текущего состояния вычислительных ресурсов по показателям их производительности, надежности и т. д.; реконфигурации виртуализированной инфраструктуры ВЦОД путем миграции и/или приостановки функционирования виртуальных машин/контейнеров с целью повышения качества обслуживания потребителей. Таким образом, задача текущей реконфигурации

инфраструктуры ВЦОД связана с анализом результатов мониторинга его вычислительных ресурсов.

Сохраненные результаты мониторинга вычислительных ресурсов ВЦОД в предыдущие периоды его функционирования именуются ретроспективными данными рабочей нагрузки (workload historical data) [3]. Их анализ позволяет реализовывать как функции текущего (реактивного) управления рабочей нагрузкой, так и ее проактивного управления. Последняя базируется на решении класса задач, именуемых прогнозирование рабочей нагрузки (workload prediction). Целью задачи прогнозирования рабочей нагрузки является предсказание ее шаблонов в будущие моменты времени функционирования ВЦОД.



а)

б)

Рисунок 1 – Пример периодических шаблонов рабочей нагрузки:
а – выполнение контейнера Kubernetes; *б* – ежесуточное функционирование виртуализированного веб-сервера
Figure 1 – Example of periodic workload patterns: *a* – running a Kubernetes container; *b* – running a virtualized web server on a daily basis

Из Рисунка 1 видно, что подсистема мониторинга службы администрирования ВЦОД сохраняет ретроспективные данные рабочей нагрузки в виде временного ряда (time series) выбранных показателей. Таким образом, задача прогнозирования рабочей нагрузки относится к классу задач анализа временных рядов.

Рассмотрение исследований в предметной области анализа временных рядов различной природы показало, что в настоящее время наиболее актуальным подходом решения этой задачи является использование методов и моделей машинного обучения (МО). В частности, для решения задачи прогнозирования временных рядов рассматриваются два класса моделей МО: статистические вероятностные модели (СВМО) и МГО – модели на основе принципов глубокого обучения (deep learning).

В общем виде этапы использования СВМО и МГО в процессе решения задачи прогнозирования временных рядов в сравнительном виде представлены на Рисунке 2. Из рисунка видно, что основой СВМО являются хорошо зарекомендовавшие себя в задачах кластеризации и классификации МО, такие как скрытые марковские модели (HMM), машины опорных векторов (SVM), метод K-ближайших соседей (K-means) и др., адаптированные к решению задач прогнозирования путем их дополнения моделями авторегрессии, такими, например, как интегрированное скользящее среднее (ARIMA) и др. [4, 5].

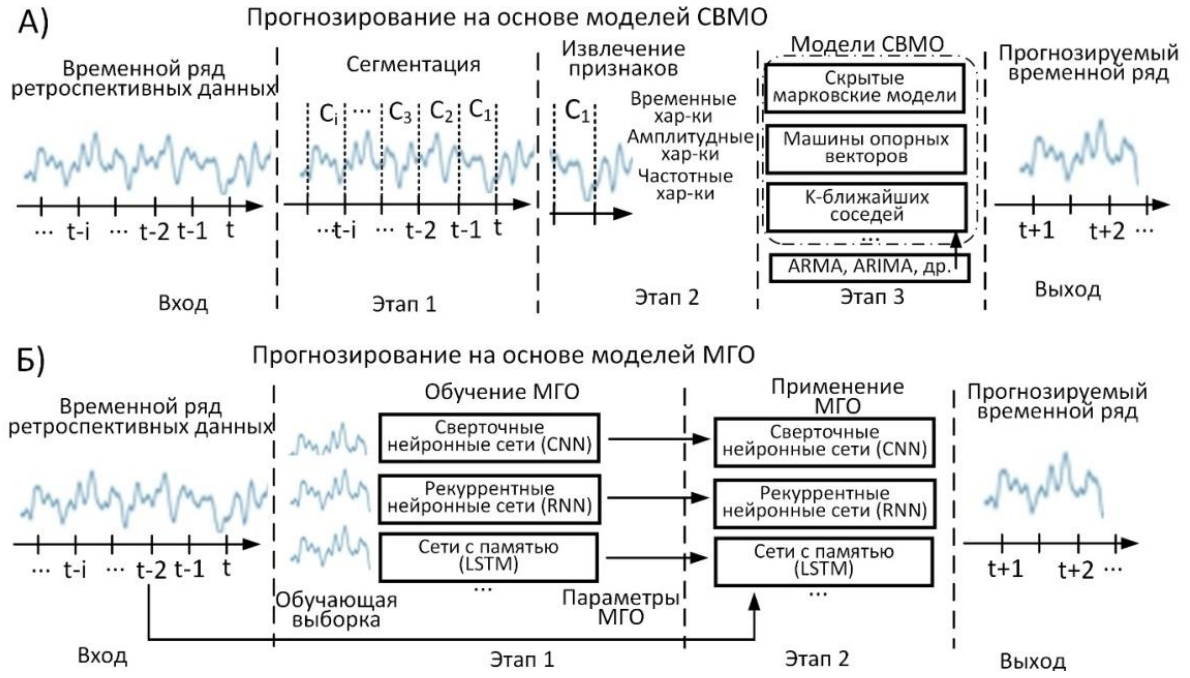


Рисунок 2 – Сравнительное представление этапов функционирования статистических вероятностных моделей и моделей глубокого обучения при решении задачи прогнозирования временных рядов

Figure 2 – A comparative presentation stages of operation for statistical probabilistic models and deep learning models in solving the problem of time series prediction

Примерами моделей МГО, используемых в задачах прогнозирования временных рядов, являются: специальные виды сверточных нейронных сетей (СНН, англ. CNN), рекуррентные нейронные сети (РНН, англ. RNN), сети с долгой краткосрочной памятью (СДКП, англ. LSTM) [6, 7]. При этом, в зависимости от вида МГО, они используются для различных задач анализа временных рядов. Так, для выявления значимых признаков в требуемых участках временного ряда используются одномерные СНН (СНН-ОМ, англ. 1D-CNN) [8], а для задач прогнозирования используются варианты сетей СДКП, такие

как двунаправленные СДКП (СДКП-ДН) [9]. То есть, в зависимости от особенностей решения задачи прогнозирования временных рядов, возможно использование как СНН-ОМ или СДКП-ДН по-отдельности, так и их применение в гибридных вариантах МГО. Целью проводимого исследования является разработка структурно-параметрических характеристик указанных видов сетей, удовлетворяющих условиям решения задачи прогнозирования рабочей нагрузки ВЦОД, разработка схемы их взаимодействия в гибридном варианте системы прогнозирования, а также сравнительное оценивание полученного решения с реализованной в современных крупномасштабных ВЦОД подсистемой прогнозирования рабочей нагрузки, основанной на методах статистического прогнозирования временных рядов.

Материалы и методы

Как было рассмотрено выше, основой процесса выявления значимых признаков шаблонов рабочей нагрузки ВЦОД может выступать одномерный вид сверточной нейронной сети – СНН-ОМ. Структура СНН-ОМ специализирована для получения входных данных в виде одномерного массива, к варианту которого относятся ретроспективные данные рабочей нагрузки, представленные временными рядами заданных параметров производительности вычислительных ресурсов ВЦОД.

При этом, как и в случае традиционных двумерных вариантов СНН, адаптированных, например, для решения задач классификации изображений, СНН-ОМ содержит чередование сверточных слоев и слоев субдискретизации (pooling-слоев), которые формируют входной вектор признаков для полносвязного слоя, формирующего выход СНН-ОМ. В отличие от двумерных СНН, в которых фильтр, именуемый сканирующим ядром (С – от англ. Core), представлен двумерным массивом, ядро СНН-ОМ является одномерным. Это обеспечивает его «скольжение» вдоль вектора входных значений временного ряда $[x_1,x_2,...,x_n]$ для нахождения требуемых признаков, характеризующих те или иные шаблоны рабочей нагрузки.

При этом операция свертки вектора входных значений временного ряда и ядра C может быть представлена выражением:

$$(x \cdot C)(t) = \sum_{i=0}^{r-1} x(t+i) \cdot C(i), \quad (1)$$

где произведение $(x \cdot C)(t)$ – свертка значений x и C в t-й позиции ряда, r – размерность ядра, $x(t+i)$ – элемент входной последовательности в $t+i$ позиции ряда, $C(i)$ – элемент ядра в i-й позиции ряда.

Очевидно, что от размерности ядра C зависит масштаб охвата значений временного ряда, в разных масштабах выборки его значений.

Прямое распространение (Forward Propagation) в СНН-ОМ включает в себя прохождение входных данных через один или несколько сверточных слоев, слои субдискретизации и полносвязные слои, так что карта признаков Z_c задается как:

$$Z_c = f_c(x \cdot w_c + b_c), \quad (2)$$

где x – входные данные, w_c – веса ядра C, b_c – смещение, f_c – функция активации свертки.

Пространственная размерность карты признаков Z_c уменьшается путем агрегирования информации из соседних значений за счет операции субдискретизации (polling), которая определяется как:

$$A_p = P(Z_c). \quad (3)$$

Далее полносвязный слой объединяет признаки, полученные в результате операций свертки и субдискретизации, и для получения выходных данных используется итоговая функция активации Y.

На Рисунке 3 представлена схема трех последовательных слоев k-го нейрона предлагаемой структуры СНН-ОМ.

Из рисунка видно, что процесс прямого распространения от предыдущего слоя $l-1$ для создания входа k-го нейрона следующего слоя l можно выразить как:

$$x_k^l = b_k^l + \sum_{i=1}^{N_{l-1}} \mathrm{conv1}D(C_{ik}^{l-1}, \mathrm{s}_i^{l-1}), \quad (4)$$

где x_k^l – вход, b_k^l – смещение k-го нейрона в слое l, s_i^{l-1} – выход i-го нейрона в слое $l-1$, C_{ik}^{l-1} – одномерное ядро от i-го нейрона в слое $l-1$ до k-го нейрона в слое l.

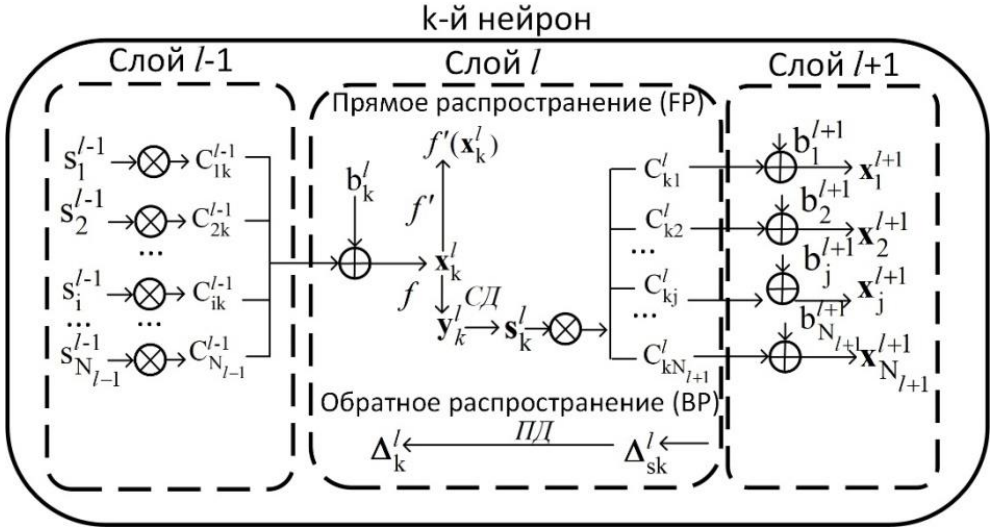


Рисунок 3 – Структурная схема взаимодействия слоев k-го нейрона одномерной сверточной нейронной сети

Figure 3 – Structural diagram of the interaction of layers of the k-th neuron of a one-dimensional convolutional neural network

Для возможности определения произвольного количества скрытых слоев, предлагается рассматривать подход адаптивной СНН [10]. В ней коэффициент субдискретизации СД (Рисунок 3) выходного слоя назначается адаптивно в зависимости от размеров его карты признаков Z_c, то есть, размерность входа карты признаков текущего слоя уменьшается на $r-1$, где r – размерность ядра C_{ik}^{l-1}.

Рассмотрим процесс обучения СНН-ОМ методом обратного распространения ошибки (Back Propagation – BP). Определим $l=1$ и $l=L$ как входной и выходной слои СНН-ОМ соответственно.

Среднеквадратическая ошибка (MSE) в выходном слое может быть выражена как:

$$\mathrm{MSE}\left(y_1^L,\ldots,y_{N_L}^L\right) = E = \sum_{i=1}^{N_L}\left(y_i^L - t_i\right)^2. \quad (5)$$

Очевидно, что для вектора входных значений временного ряда $[x_1,x_2,\ldots,x_n]$ и соответствующего ему выходного вектора $\left[y_1^L,\ldots,y_{N_L}^L\right]$ требуется найти производную MSE по каждому индивидуальному весу C_{ik}^{l-1}, связанному с k-м нейроном, и смещению b_k^l для последующего применения метода градиентного спуска, обеспечивающего минимизацию MSE. Так, Δ_k^l – дельта k-го нейрона в слое l будет использоваться для обновления значения b_k^l этого нейрона и всех весов нейронов в предыдущем слое, как:

$$\frac{\partial E}{\partial C_{ik}^{l-1}} = \Delta_k^l y_i^{l-1}, \quad (6)$$

$$\frac{\partial E}{\partial b_k^l} = \Delta_k^l.$$

Таким образом, от входного полносвязного слоя до выходного слоя СНН-ОМ скалярная функция обратного распространения Δs_k^l задается как:

$$\frac{\partial E}{\partial s_k^l} = \Delta s_k^l = \sum_{i=1}^{N_{l+1}} \frac{\partial E}{\partial x_i^{l+1}} \cdot \frac{\partial x_i^{l+1}}{\partial s_k^l} = \sum_{i=1}^{N_{l+1}} \Delta_i^{l+1} C_{ki}^l. \quad (7)$$

После того, как функция Δs_k^l выполнена от слоя $l+1$ до слоя l, она распространяется на вход Δ_k^l (Рисунок 3). Определим это преобразование как повышение дискретизации (ПД), как ПД(s_k^l). Тогда Δ_k^l определяется как:

$$\Delta_k^l = \frac{\partial E}{\partial y_k^l} \cdot \frac{\partial y_k^l}{\partial x_k^l} = \frac{\partial E}{\partial \text{ПД}_k^l} \cdot \frac{\partial \text{ПД}_k^l}{\partial y_k^l} f(x_k^l) = \text{ПД}(\Delta s_k^l) \beta f'(x_k^l), \quad (8)$$

где $\beta = (\text{СД})^{-1}$ – операция, обратная субдискретизации, поскольку каждый элемент s_k^l был получен путем усреднения количества элементов СД промежуточного выхода y_k^l.

Тогда Δs_k^l – дельта ошибки при выполнении функции обратного распространения между слоями может быть выражена как:

$$\Delta s_k^l = \sum_{i=1}^{N_{l+1}} \text{conv1D}z(\Delta_k^{l+1}, \text{rev}(C_{ki}^l)), \quad (9)$$

где $\text{rev}(C_{ki}^l)$ – операция реверса массива весов C_{ki}^l, а $\text{conv1D}z()$ – операция свертки в одномерном пространстве с добавлением $r-1$ нулей. При этом чувствительность к весу и смещению определяются выражениями (10) и (11) соответственно:

$$\frac{\partial E}{\partial C_{ki}^l} = \text{convD1}(s_k^l, \Delta_i^{l+1}), \quad (10)$$

$$\frac{\partial E}{\partial b_k^l} = \sum \Delta_k^l (n). \quad (11)$$

Таким образом, процесс обучения СНН-ОМ в целом соответствует таковому для двумерных вариантов СНН, однако существенно зависит от размерности r ядра сверточного слоя. Эта особенность позволяет реализовать обучение варианта СНН-ОМ для поиска требуемых закономерностей временного ряда в определенном временном масштабе.

Для решения задачи охвата различных временных масштабов предлагается использование ансамбля из нескольких СНН-ОМ, отличающихся размерностью r ядра С.

На Рисунке 4 размерность ядер в сверточных слоях СНН-ОМ, входящих в ансамбль, условно представлена как «малая» – min, «средняя» – mid и «большая» – max. Таким образом, ядра разной размерности позволяют охватить разные по масштабу участки временного ряда, что обеспечивает выявление как краткосрочных по времени шаблонов рабочей нагрузки (например, резкое повышение использования вычислительного ресурса в момент запуска приложения), так и шаблонов, продолжительных по времени (суточный рост или спад использования вычислительного ресурса). Выходом каждой СНН-ОМ ансамбля являются вероятности шаблонов рабочей нагрузки на анализируемых участках ее временного ряда. Процесс ансамблирования подразумевает их конкатенацию для формирования входной последовательности, используемой для решения задачи прогнозирования.

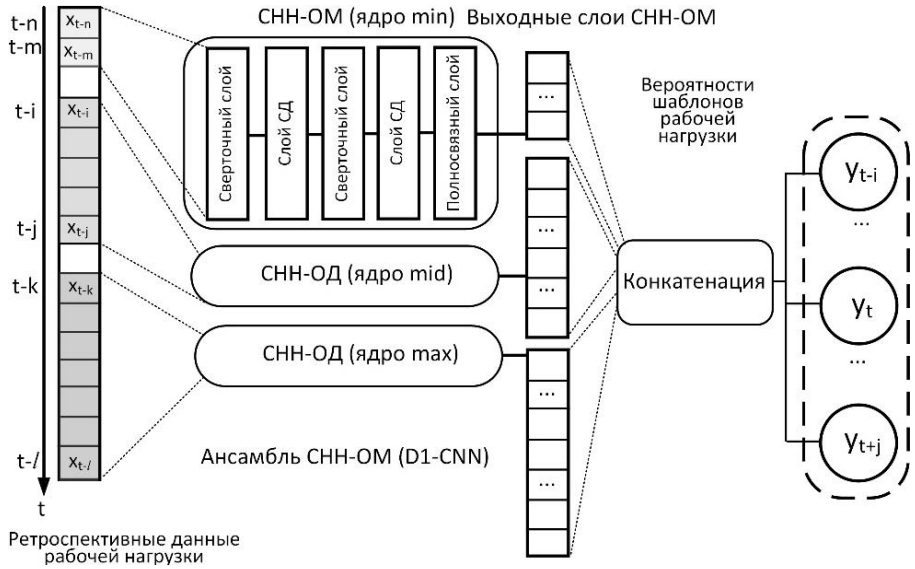


Рисунок 4 – Обобщенная схема ансамбля одномерных сверточных нейронных сетей для решения задачи выявления значимых признаков шаблонов ретроспективных данных рабочей нагрузки
Figure 4 – A generalized scheme an ensemble of one-dimensional convolutional neural networks for solving the problem of identifying significant features of patterns in historical workload data

Сеть СДКП-ДН является модифицированным вариантом классической СДКП сети [9], узлы которой (ячейки, cell) помимо функции активации содержат элементы памяти и систему переключателей (шлюзов, gate), обеспечивающих реализацию функции запоминания и удаления данных.

Таким образом, каждая ячейка СДКП получает входные данные, зависящие от вычислений, выполненных в ячейках на предыдущих временных шагах. В отличие от СДКП архитектура сети СДКП-ДН характеризуется двунаправленным потоком информации, фактически агрегируя две СДКП, каждая из которых обрабатывает данные об одном из направлений их распространения. При этом в СДКП-ДН выходы обеих сетей объединяются на выходном слое.

Обобщенно это представляется выражением:

$$\begin{cases} h_f = \mathrm{LSTM}(x_i, h_{f-1}) \\ h_b = \mathrm{LSTM}(x_b, h_{b-1}), \\ h_t = w_t h_f + v_t h_b + b_t \end{cases} \quad (12)$$

где x_i – входные данные, h_f – состояние скрытого слоя при прямом проходе, h_b – состояние скрытого слоя при обратном проходе, h_t – состояние скрытого слоя в t-й позиции ряда, w_t – выходной вес скрытого слоя при прямом проходе, v_t – выходной вес скрытого слоя при обратном проходе, b_t – величина ошибки.

Выходом сети СДКП-ДН является вектор $[y_{t-i},\ldots,y_t,\ldots,y_{t+j}]$, определяющий представление значений временного ряда в моменты времени, предшествующие и последующие значению в t-й позиции ряда. При этом последующие значения $(y_{t+1},\ldots,y_{t+j})$ являются прогнозными.

Использование выходных данных ансамбля сетей СНН-ОМ, содержащих найденные зависимости временного ряда ретроспективных данных, определяющие те или иные шаблоны рабочей нагрузки в качестве входной последовательности сети СДКП-ДН, позволяет определять вероятности появления этих зависимостей на участках временного ряда $t,t+1,\ldots,t+j$. Глубина получения прогнозных значений зависит от структуры скрытых слоев сети СДКП-ДН.

Таким образом, для решения задачи прогнозирования рабочей нагрузки ВЦОД предлагается использовать гибридный вариант МГО, в состав которого входит ансамбль СНН-ОМ, каждая сеть в котором отличается размерностью ядра, а также сеть СДКП-ДН, обобщающая результат конкатенации выходов ансамбля СНН-ОМ. Вариантом такой гибридной модели может выступать каскад сетей СНН-ОМ и СДКП-ДН, каждый уровень в котором настроен на получение прогноза заданного вида зависимостей рабочей нагрузки, которая характерна, например, в ВЦОД общего назначения, обеспечивающих поддержку широкого спектра пользовательских запросов. Пример реализации такой каскадной гибридной модели представлен на Рисунке 5. Из рисунка видно, что каскадными уровнями являются:

– взаимодействующие по выходам ансамбли сетей СНН-ОМ и СНН-ОМ’;

– взаимодействующие по выходу сети СДКП-ДН и СДКП-ДН’, входом которых является конкатенируемый выход каскада СНН-ОМ.

При этом различная размерность ядра в каждой из сетей СНН-ОМ обеспечивает выделение значимых признаков на определенном масштабе временного ряда, а конкатенация выходов этих сетей обеспечивает обобщение одних и тех же зависимостей, найденных разными сетями и выделение зависимостей, специфичных для временного масштаба каждой сети.

Эта особенность позволяет варьировать глубину прогноза в каскаде сетей СДКП-ДН и СДКП-ДН’, которые отличаются количеством скрытых слоев, что обеспечивает различный временной охват прогноза. Естественно, что, в зависимости от задач прогнозирования рабочей нагрузки, количество сетей СДКП-ДН в каскаде может быть больше двух. Однако следует учитывать, что простое масштабирование каскада сетей СДКП-ДН не ведет к увеличению глубины прогноза. Дополнительным условием является подбор соответствующих весов w_t.

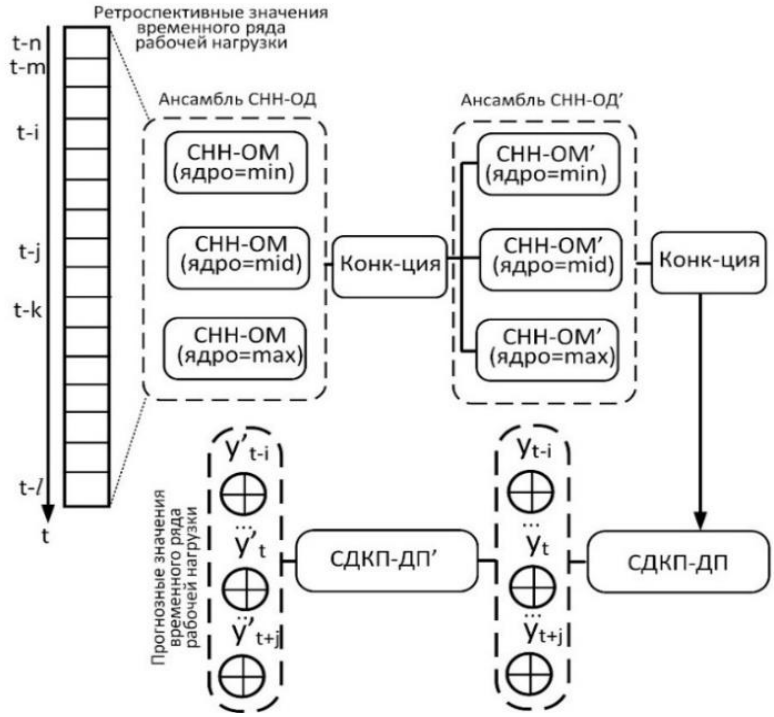


Рисунок 5 – Обобщенная схема каскадной гибридной модели глубокого обучения для прогнозирования рабочей нагрузки ВЦОД на основе ее ретроспективных данных

Figure 5 – A generalized scheme of a cascaded hybrid deep learning model for predicting data center workloads based on its historical data

Результаты

Предлагаемое решение задачи прогнозирования временного ряда ретроспективных данных рабочей нагрузки требует оценивания по критерию пригодности ее применения в подсистеме прогнозирования рабочей нагрузки современных ВЦОД, являющейся составной частью службы их администрирования. В качестве альтернативы предложенной каскадной гибридной модели глубокого обучения была выбрана реализация подсистемы прогнозирования ВЦОД Google Cluster, представленная в [11] и доступная, наряду с базой ретроспективных данных рабочей нагрузки Google Cluster, на открытых платформах хостингов IT-проектов.

Базой подсистемы прогнозирования ВЦОД Google Cluster является комбинированная статистическая модель, основанная на классификаторе шаблонов рабочей нагрузки, использующем вариант машины опорных векторов (SVM), а также модуле прогнозирования временного ряда, основанном на модели авторегрессии, в частности, авторегрессионном интегрированном скользящем среднем (ARIMA) и ее вариантах.

В качестве показателя значений временного ряда рабочей нагрузки использовались сохраненные данные CPU_Usage Rate – коэффициент загрузки процессора: процента времени, в течение которого процессор занят обработкой задач. Этот показатель рассчитывается путем деления времени работы процессора на общее время мониторинга за заданный период. Для снижения влияния факторов зашумления значений CPU_Usage Rate, связанных с проблемой «шумных соседей» (Noisy Neighbours) – взаимного влияния выполняемых потоков разных приложений вследствие особенностей реализации задачи когерентности кэш памяти процессорных ядер, в эксперименте применялся метод модовой декомпозиции сигнала, доказавший свою эффективность в смежных предметных областях [12]. Схема исследовательского стенда для проведения сравнительного эксперимента представлена на Рисунке 6.

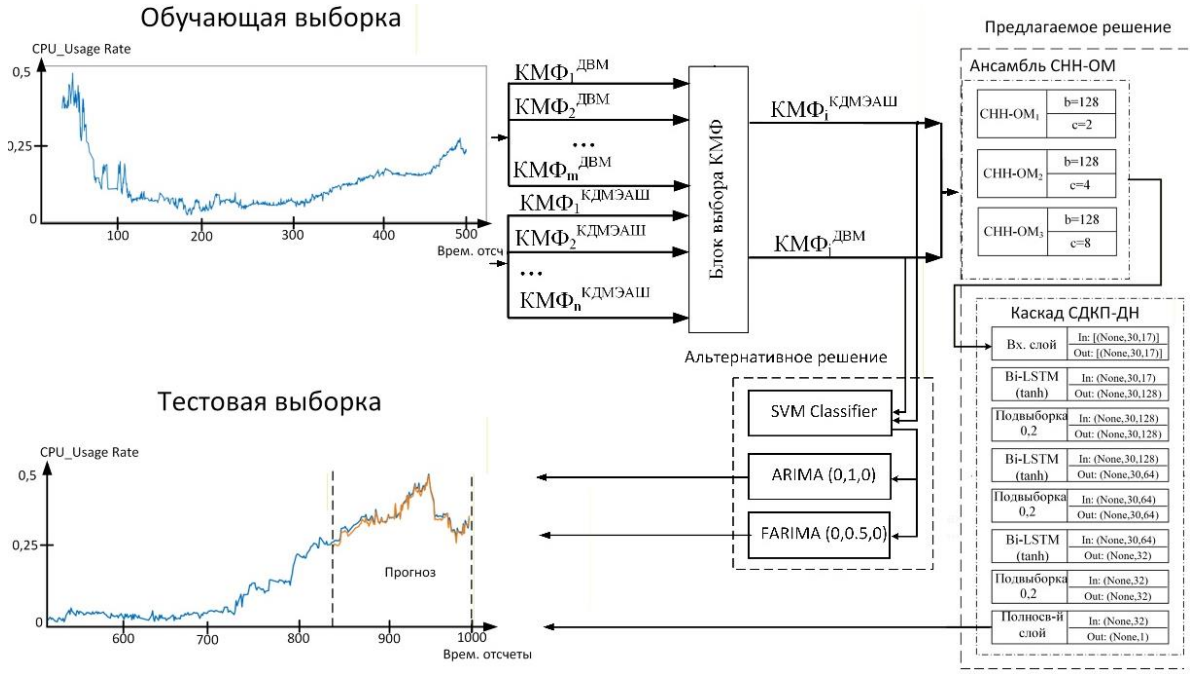
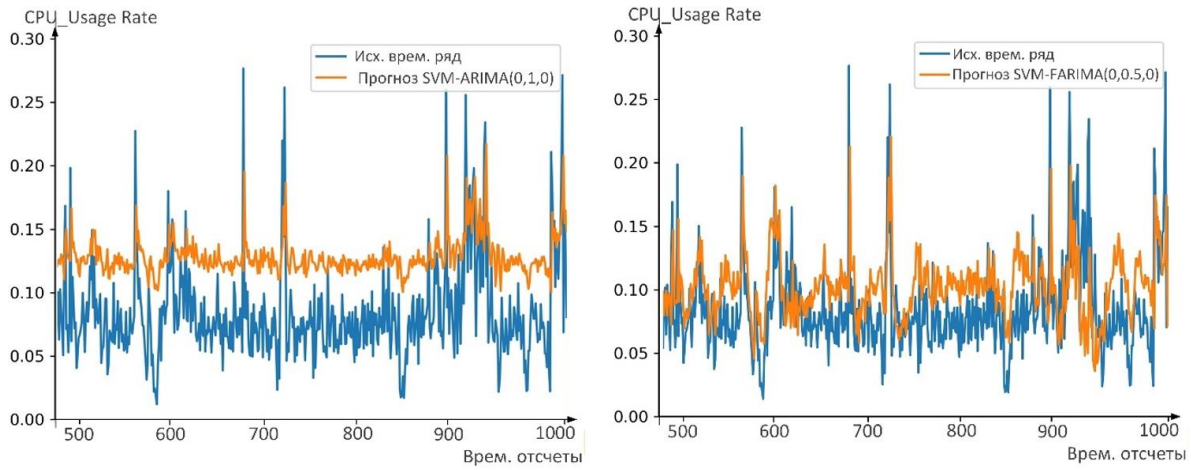


Рисунок 6 – Схема экспериментального стенда для сравнительного оценивания результатов прогнозирования предложенной каскадной гибридной модели глубокого обучения и альтернативного решения (SVM-ARIMA, SVM-FARIMA)

Figure 6 – Experimental setup for comparative evaluation of prediction results of the proposed cascaded hybrid deep learning model and alternative solution (SVM-ARIMA)

Из рисунка видно, что предлагаемое решение (каскадная гибридная модель глубокого обучения) сравнивалось с вариантом подсистемы прогнозирования Google Cluster, основанном на классификаторе SVM (машина опорных векторов), как и сеть СНН-ОМ в предлагаемом решении, выделяющем шаблоны рабочей нагрузки (участки нестационарности временного ряда), и прогнозной модели ARIMA(p,d,q), где d – коэффициент дифференцирования – является целочисленным и равен 1 (линейный тренд значений временного ряда). Дополнительно, с целью получения картины прогнозного тренда на разных масштабах рассмотрения, использовался вариант модели FARIMA(p,d,q) (фрактальный авторегрессионный интегральный процесс скользящего среднего), применяемой для долгосрочных прогнозов, где коэффициент d является дробным в диапазоне $d \in (-0{,}5;0{,}5)$.

Полученное в результате модовой декомпозиции временного ряда множество колебательных модовых функций (КМФ) было предварительно преобразовано в совокупность векторов-скользящих окон фиксированного размера, представленных 500 условными отсчетами. Обучающая выборка охватывала окно 0–500 отсчетов, тестовая выборка охватывала окно 501–1000 отсчетов. При этом прогнозными рассматривались значения последних 100–150 отсчетов. Результаты оценивания процесса прогнозирования способом SVM-ARIMA(0,1,0) представлены на Рисунке 7*а*, а SVM-FARIMA(0,0.5,0) представлены на Рисунке 7*б*.



а) *б*)

Рисунок 7 – Сравнение экспериментальных серий оценки эффективности процесса прогнозирования временного ряда: *а* – способом SVM-ARIMA; *б* – способом SVM-FARIMA
Figure 7 – Comparison of experimental series for evaluating the effectiveness of the time series forecasting process: *a* – using the SVM-ARIMA method; *b* – using the SVM-FARIMA method

Из Рисунка 7 видно, что классификатор SVM достаточно точно выделил шаблоны рабочей нагрузки, а модель прогнозирования ARIMA достаточно точно спрогнозировала линейный тренд на участках стационарности. Его расхождение с трендом исходного временного ряда на 0,1 коэффициента CPU_Usage Rate связано с тем, что процесс сглаживания начался не на участке стационарности, а на первом выделенном SVM шаблоне рабочей нагрузки. В целом, это расхождение не оказывает влияния на принятие решения о реконфигурации. Модель прогнозирования FARIMA (Рисунок 7*б*) продемонстрировала более высокую точность прогноза, в силу ее ориентации на тренд рабочей нагрузки в долгосрочной перспективе (каковым является тестовый временной ряд). При этом, ей, как и модели ARIMA, присуще смещение прогнозных оценок на участках стационарности, правда, в меньшей степени: 0,05 против 0,1 у ARIMA. Также

особенностью модели прогнозирования FARIMA является более высокая вычислительная сложность, в силу не целочисленного значения коэффициента *d*.

Результаты прогнозирования предлагаемым решением представлены на Рисунке 8. При этом для учета влияния на точность прогнозирования числа эпох обучения рассмотрены прогнозные временные ряды, полученные после 10 эпох обучения (Рисунок 8*а*), 20 эпох обучения (Рисунок 8*б*), 30 эпох обучения (Рисунок 8*в*) и 40 эпох обучения (Рисунок 8*г*).

Из Рисунка 8 видно, что увеличение количества эпох обучения (в частности, обучения модели СДКП-ДН, асимптотически увеличивает точность прогноза при нелинейном росте времени, затрачиваемом на обучение.

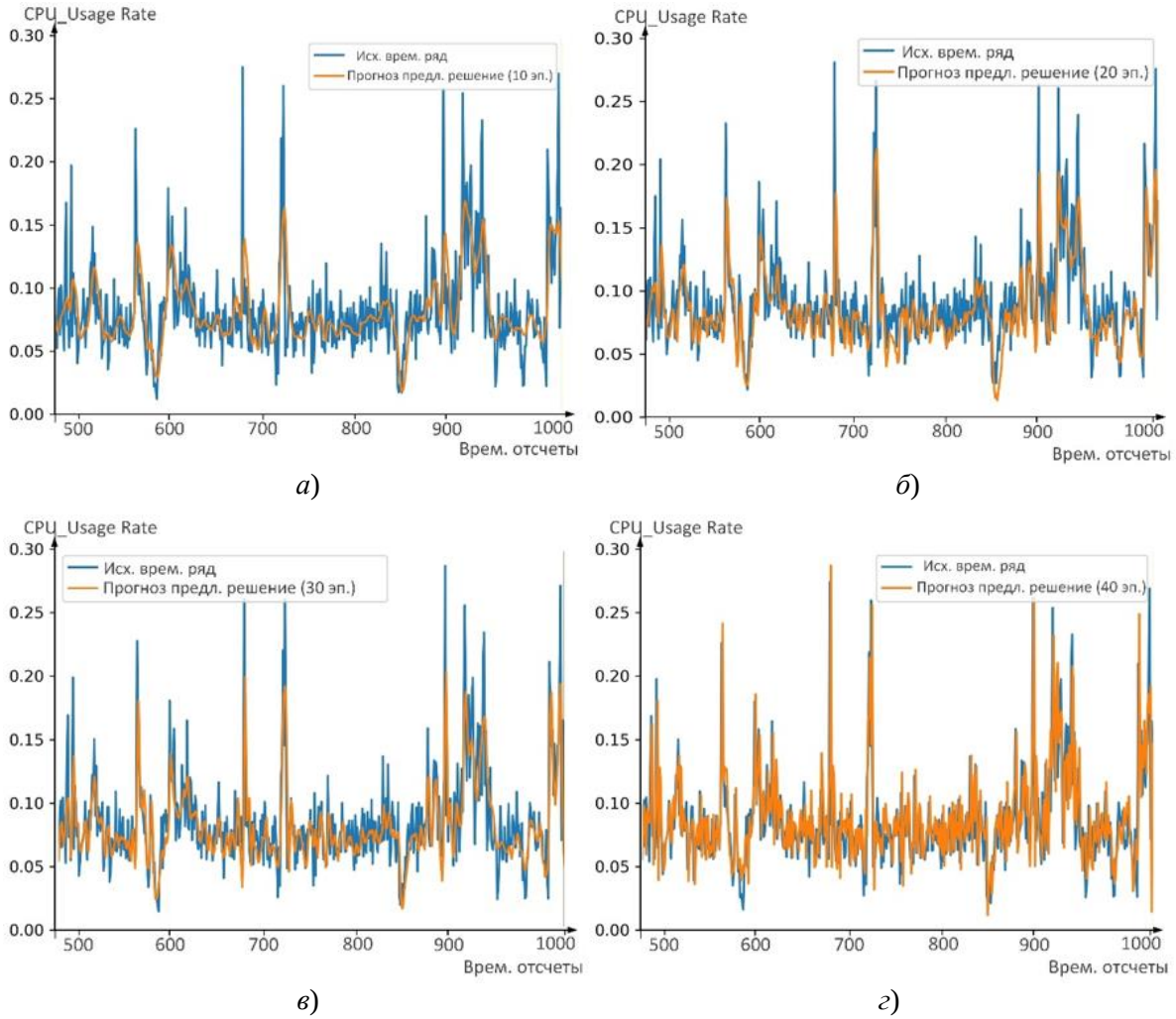


Рисунок 8 – Сравнение экспериментальных серий оценки эффективности процесса прогнозирования временного ряда предлагаемой каскадной гибридной модели глубокого обучения: *а* – для 10 эпох обучения; *б* – для 20 эпох обучения; *в* – для 30 эпох обучения; *г* – для 40 эпох обучения

Figure 8 – Comparison of experimental series for evaluating the performance of the time series forecasting process of the proposed cascade hybrid deep learning model: *a* – for 10 training epochs; *b* – for 20 training epochs; *c* – for 30 training epochs; *d* – for 40 training epochs

Результаты сравнения полученных экспериментальных серий предлагаемого решения и альтернативного решения (SVM-ARIMA и варианты) показывают, что в анализируемом окне временных отсчетов (850–1000) в общем случае предложенное решение дает прогнозные значения, в среднем отклоняющиеся от значений исходного

временного ряда на участках стационарности менее чем на 2 %, в то время как прогнозные значения альтернативного способа имеют отклонение от 5 % (FARIMA) до 15 % (ARIMA). Причины такого расхождения рассмотрены выше. В общем случае способ на основе SVM-ARIMA (или вариантов ARIMA) предпочтительно использовать для получения быстрого результата о тренде рабочей нагрузки, в то время как результаты, полученные предлагаемым решением, целесообразно использовать для развернутого анализа проблемных участков временного ряда с целью более тонкой оптимизации процесса реконфигурации ВЦОД.

Между тем следует отметить особенности предложенной каскадной гибридной модели глубокого обучения, выявленные в результате проведения эксперимента и косвенно влияющие на эффективность ее использования. Так, общее время обучения предложенной модели, складывающееся из времени обучения ансамбля сетей СНН-ОМ и времени обучения каскада сетей СДКП-ДН, существенно зависит от:

– размера ядер, выбираемых для каждого экземпляра ансамбля сетей СНН-ОМ (в эксперименте это 2, 4 и 8);

– количества скрытых слоев в сетях каскада СДКП-ДН;

– размера вектора-скользящего окна временного ряда, используемого в качестве тестовой выборки.

В ходе эксперимента время обучения предложенного решения было в среднем в 2–4 раза больше времени обучения статистической модели прогнозирования. Это требует в дальнейшем проведения дополнительных исследований, связанных с оптимизацией структурно-параметрических характеристик разработанной каскадной гибридной модели глубокого обучения по комплексному показателю «точность прогноза – время обучения».

Заключение

Статья посвящена решению проблемы прогнозирования рабочей нагрузки виртуализированных центров обработки данных с использованием ретроспективных данных мониторинга показателей производительности их базовых вычислительных ресурсов. Анализ существующих исследований в данной области позволил установить, что одним из вариантов решения этой проблемы является использование моделей глубокого обучения, специализированных на анализ и обобщение данных временных рядов, которыми представляются ретроспективные данные рабочей нагрузки.

Для этого подхода были исследованы возможности одномерных сверточных нейронных сетей, обеспечивающих поиск закономерностей в данных временных рядов, а также предложена ансамблевая модель на основе этого вида сетей, повышающая точность обобщения результатов поиска зависимостей за счет разной размерности ядер (фильтров) в составе каждой сети ансамбля. Также, в качестве модели прогнозирования данных, являющихся выходом этой ансамблевой модели, предложено использование двунаправленной сети с долгой краткосрочной памятью, обеспечивающей формирование прогнозных значений временного ряда рабочей нагрузки. Подобная гибридная модель глубокого обучения может быть каскадирована для случая сложно организованных ретроспективных данных рабочей нагрузки, характерных для центров обработки данных общего назначения.

Представлена структура экспериментального стенда для проведения сравнительного оценивания эффективности процесса прогнозирования предложенной каскадной гибридной модели глубокого обучения с существующим решением на основе модели статистического прогнозирования, а также обобщенные результаты экспериментальных серий.

Дальнейшие направления исследования связаны с решением задачи оптимизации структурно-параметрических характеристик предложенной гибридной модели под конкретные условия ее эксплуатации.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Shen L., Qian Sh., Zhai T., Li L., Li Zh. Research on Cloud Computing High-Density Data Center Infrastructure and Environment Matching Technology. In: *2020 2nd International Conference on Computer Science Communication and Network Security (CSCNS2020): MATEC Web of Conferences: Volume 336, 22–23 December 2020, Sanya, China*. EDP Sciences; 2012. https://doi.org/10.1051/matecconf/202133602028
2. Uddin M., Rahman A.A., Shah A., Memon J. Virtualization Implementation Approach for Data Centers to Maximize Performance. *Asian Journal of Scientific Research*. 2012;5(2):45–57. https://doi.org/10.3923/ajsr.2012.45.57
3. Cox-Fuenzalida L.-E. Effect of Workload History on Task Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 2007;49(2):277–291. https://doi.org/10.1518/001872007X312496
4. Tran V.G., Debusschere V., Bacha S. Hourly Server Workload Forecasting up to 168 Hours Ahead Using Seasonal ARIMA Model. In: *2012 IEEE International Conference on Industrial Technology, 19–21 March 2012, Athens, Greece*. IEEE; 2012. P. 1127–1131. https://doi.org/10.1109/ICIT.2012.6210091
5. Sun Q., Tan Zh., Zhou X. Workload Prediction of Cloud Computing Based on SVM and BP Neural Networks. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*. 2020;39(3):2861–2867. https://doi.org/10.3233/JIFS-191266
6. Nguyen H.M., Kalra G., Kim D. Host Load Prediction in Cloud Computing Using Long Short-Term Memory Encoder-Decoder. *The Journal of Supercomputing*. 2019;75(11):7592–7605. https://doi.org/10.1007/s11227-019-02967-7
7. Пашшоев Б., Петрусевич Д.А. Анализ нейросетевых моделей для прогнозирования временных рядов. *Russian Technological Journal*. 2024;12(4):106–116. https://doi.org/10.32362/2500-316X-2024-12-4-106-116
   Pashshoev B., Petrusevich D.A. Neural Network Analysis in Time Series Forecasting. *Russian Technological Journal*. 2024;12(4):106–116. https://doi.org/10.32362/2500-316X-2024-12-4-106-116
8. Mitiche I., Nesbitt A., Conner S., Boreham Ph., Morison G. 1D-CNN Based Real-Time Fault Detection System for Power Asset Diagnostics. *IET Generation, Transmission & Distribution*. 2020;14(24):5766–5773. https://doi.org/10.1049/iet-gtd.2020.0773
9. Wibawa A.P., Fadhilla A.F., Paramarta A.Kh.I., et al. Bidirectional Long Short-Term Memory (Bi-LSTM) Hourly Energy Forecasting. In: *International Conference on Computer Science Electronics and Information (ICCSEI 2023)*: *E3S Web of Conferences, Volume 501, 12–13 December 2023, Yogyakarta, Indonesia*. EDP Sciences; 2024. https://doi.org/10.1051/e3sconf/202450101023
10. Ban Y., Zhang D., He Q., Shen Q. APSO-CNN-SE: An Adaptive Convolutional Neural Network Approach for IoT Intrusion Detection. *Computers, Materials and Continua*. 2024;81(1):567–601. https://doi.org/10.32604/cmc.2024.055007
11. Rasheduzzaman M., Islam A., Rahman R.M. Workload Prediction on Google Cluster Trace. *International Journal of Grid and High Performance Computing*. 2014;6(3):34–52. https://doi.org/10.4018/ijghpc.2014070103
12. Almalchy M.T., Ciobanu V., Popescu N. Noise Removal from ECG Signal Based on Filtering Techniques. In: *2019 22nd International Conference on Control Systems and Computer Science (CSCS), 28–30 May 2019, Bucharest, Romania*. IEEE; 2019. P. 176–181. https://doi.org/10.1109/CSCS.2019.00037

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Мартыненков Борис Витальевич, преподаватель, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: borikan33@mail.ru

Boris V. Martynenkov, University Teacher, MIREA – Russian Technological University, Moscow, the Russian Federation.

Статья поступила в редакцию 08.10.2025; одобрена после рецензирования 02.12.2025; принята к публикации 09.12.2025.

The article was submitted 08.10.2025; approved after reviewing 02.12.2025; accepted for publication 09.12.2025.