УДК 303.732.4 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2022/4/146-155
Поступила в редакцию 01.10.2022
Подписана в печать 05.12.2022

К ВОПРОСУ ОБ УВЕЛИЧЕНИИ ПРОИЗВОДИТЕЛЬНОСТИ МАШИННОГО ОБУЧЕНИЯ НА ЭТАПЕ ВЫБОРКИ ДАННЫХ ПРИ РЕШЕНИИ ЗАДАЧ КЛАССИФИКАЦИИ

© 2022 Р. А. Дьяченко[1], П. А. Косолапов[1], Д. А. Гура[1,2]✉

[1]Кубанский государственный технологический университет
ул. Московская, 2, 350072 Краснодар, Российская Федерация
[2]Кубанский государственный аграрный университет им. И. Т. Трубилина
ул. Калинина, 13, 350044 Краснодар, Российская Федерация

Аннотация. Целью исследования является определение метода хранения данных для задач машинного обучения нейронных сетей и семантической сегментации облаков точек. Рассмотрены существующие способы хранения массивов данных большого размера, проведены экспериментальные исследования для определения быстродействия операции чтения данных. Во время проведения эксперимента была осуществлена подготовка данных, заключающаяся в отборе информации из общей выборки. В качестве критериев отбора выделяются координаты точек, метка класса и количество записей в исходном дата сете. Все необходимые параметры и их структура приведены и описаны в работе. Метки класса, в силу представления исходного дата сета, претерпели некоторое преобразование. После отбора информации была произведена ее конвертация в исследуемые форматы файлов с последующим сохранением для проведения экспериментов. Для проведения исследований были взяты наиболее распространенные форматы файлов, используемые для хранения информации *.csv, *.npy и *.h5. После получения данных для эксперимента последовал этап непосредственно проведения эксперимента. Эксперимент заключался в воспроизведении процесса доступа к информации из предварительно полученных файлов и последующей загрузкой информации на входной слой нейронной сети без процесса обучения. Результатом эксперимента стала статистическая информация о времени чтения файла в зависимости от выбранной структуры и объема хранимой в нем информации. Кроме этого, был подведен итог о целесообразности использования того или иного способа хранения информации в условиях предметной области работы, исходя из принципа работы того или иного формата файла.
Ключевые слова: выборка данных, нейронные сети, Point Cloud, ЦУР.

ВВЕДЕНИЕ

В настоящее время исследуется множество вопросов машинного обучения, которые используют облака точек лазерного отражения в качестве входных данных [1–4]. Увеличения производительности при машинном обучении трехмерных облаков точек лазерного отражения на сегодняшний день являются достаточно актуальными [5–7].

Одной из проблем, связанной с подготовкой данных для машинного обучения является временные задержки, которые возникают при подготовке информации для загрузки на входной слой нейронной сети. В крайних случаях, подготовка информации для обучения может происходить по несколько суток. Это довольно известная проблема, для любого процесса машинного обучения необходимо каким-то образом хранить и обращаться к исходным данным. Данная работа ставит перед собой цель систематизировать уже имеющиеся знания путем проведения анализа скорости доступа к одним и тем же данным,

✉ Гура Дмитрий Андреевич
e-mail: gda-kuban@mail.ru

Контент доступен под лицензией Creative Commons Attribution 4.0 License.
The content is available under Creative Commons Attribution 4.0 License.

хранимым в разных форматах. Существует множество способов сохранения точек лазерного отражения. Задача данного исследования заключается в выборе нескольких популярных форматов и измерение скорости доступа к хранящейся в ней информации с целью выбора одного или нескольких оптимальных вариантов [8, 9].

В качестве источника данных чаще всего используется локальный файл или группа файлов различных форматов, содержащих точки лазерного отражения. Среди используемых форматов файлов могут быть как форматы общего назначения (*.txt, *.csv, *.h5), так и специфичные для хранения геопространственных данных (*.las, *.pcd и другие). В процессе машинного обучения специалист по анализу данных взаимодействует с другими форматами файлов (*.npy, *.pth и другие).

Задачей исследования является установление оптимального выбора для хранения больших данных, в данном конкретном случае — облаков точек. Решение поставленной задачи будет достигаться путем проведения серии экспериментов, в ходе которых для заранее известного объема данных с определенной структурой (объем каждого файла и их структура будет описана в публикации в разделе экспериментов) будет замеряться время чтения данных. Для исследуемых данных в ходе проведения эксперимента не будут производиться никакие дополнительные трансформации. После проведения эксперимента будет получена статистическая информация, на основе которой сделаны выводы о целесообразности использования того или иного типа файлов [10–12].

1. ИНСТРУМЕНТЫ И МЕТОДЫ

Для построения нейронных сетей используется язык *Python* (https://www.python.org) и фреймворк *PyTorch* (https://pytorch.org). Для загрузки дата сетов из внешних хранилищ информации в данном фреймворке существует класс *torch.utils.data.Dataset*. Наследники *torch.utils.data.Dataset* используются классом *torch.utils.data.DataLoader*, который передает информацию на входной слой нейронных сетей.

Для замеров выполнения операций и сбора информации о времени, предлагается использовать стандартный модуль *time*. Для работы с числовыми массивами используется библиотека *NumPy* (https://pytorch.org). Для работы с файлами *HDF5* применяется библиотека *h5py* (https://www.h5py.org). Построения графиков выполняется при помощи библиотеки *matplotlib* (https://matplotlib.org).

Вся исследовательская работа производилась на платформе *Google Collaboratory* (https://colab.research.google.com), а все исходные файлы были размещены на облачном хранилище Google Drive (https://www.google.com/intl/ru/drive/).

В качестве исходных данных использован дата сет под названием *Stanford 3D Indoor Scene Dataset* (*S3DIS*) [13], из которого был выбран ряд файлов, в последствии конвертированные в форматы *csv*, *npy*, *h5*. Дата сет состоит из облаков точек, представляющих внутренние помещения офисного или образовательного назначения [14].

Дата сет разделен на 6 частей, каждая из которых является отдельным зданием, либо его частью. Каждая часть хранится в каталоге от «*Area_1*» до «*Area_6*». В каждом каталоге находится множество подкаталогов, которые представляют отдельно взятое помещение, расположенное внутри здания [15]. О характере помещения говорит название каталога, например «*conferenceRoom*», «*office*», «*hallway*» и т. д. Число в названии каталога используется для уточнения помещения в случае, если подобных помещений в здании несколько [16].

Каждый подкаталог, описывающий отдельное помещение, имеет следующую структуру: в нем располагается подкаталог *Annotations* и файл с расширением *.txt*, повторяющий название каталога [17]. Текстовый файл представляет из себя облако точек, являющееся результатом сканирования всего помещения целиком [18]. Подкаталог *Annotations* содержит в себе набор текстовых файлов, которые представляют из себя сегментированные участки исходного облака [19, 20]. Например, в файле *.\Area_1\conferenceRoom_1\Annotations\table_1.txt* будут содержаться все точки, отраженные от

стола в помещении *conferenceRoom_1*. Название каждого файла в папке *Annotations* говорит о принадлежности данного сегмента к той или иной категории. Всего авторы выделяют 13 категорий, которые сведены в табл. 1.

Таблица 1. Категории разметки дата сета S3DIS
[Table 1. Categories of the S3DIS dataset markup]

Название	Примеры объектов
Ceiling	Потолки
Floor	Полы
Wall	Стены
Beam	Балки
Column	Столбы
Window	Окна
Door	Двери
Table	Столы
Chair	Стулья
Sofa	Диваны
Bookcase	Книжные шкафы и полки
Board	Офисные доски
Clutter	Все остальные элементы

Файлы, расположенные в подкаталоге *Annotations* и файлы, в которых хранится облако точек помещения целиком имеют одинаковую структуру. Каждый файл представляет из себя таблицу, размерами N*6, где N — количество точек в облаке. Следовательно, отдельно взятая строка описывает отдельно взятую точку в облаке, а отдельно взятый столбец описывает ровно один параметр, который имеется у каждой точки в облаке. Среди таких параметров авторы дата сета выделяют следующие (в порядке следования этих параметров в текстовых файлах слева направо):

1. Координата *X*;
2. Координата *Y*;
3. Координата *Z*;
4. Интенсивность компонента *R*;
5. Интенсивность компонента *G*;
6. Интенсивность компонента *B*.

Столбцы 1–3 отвечают за геометрическое положение точки, а столбцы 4–6 отвечают за цвет точки в палитре *RGB*.

1.1. Выбор файлов для исследования

Изначально предполагалось выбрать файлы объемом в 10000, 50000, 100000, 500000, 1000000 и 5000000 точек. Но так как дата сет был составлен при помощи реальных измерений и добиться необходимых круглых значений без потери исходной информации не представляется возможным, были выбраны дата сеты объемом, близким по значению к предложенным круглым значениям.

На табл. 2 представлены файлы, которые были выбраны для проведения исследования.

Таблица 2. Выбранные для исследования файлы
[Table 2. Files selected for the study]

Название файла	Путь к файлу	Количество точек
clutter_32.txt	*.\Area_1\ hallway_6\ Annotations*	11423
beam_3.txt	*.\Area_1\ hallway_6\ Annotations*	43116
bookcase_8.txt	*.\Area_1\ hallway_6\ Annotations*	103833
ceiling_4.txt	*.\Area_2\ auditorium_2\ Annotations*	498167
ceiling_1.txt	*.\Area_1\ hallway_6\ Annotations*	990521
floor_1.txt	*.\Area_2\ auditorium_2\ Annotations*	2233811

Так как в дата сете файлы представлены только форматом *txt*, то предлагается на их основе генерировать все необходимые форматы файлов. Для работы с облаками точек необходимым минимумом является информация о геометрическом положении точки, следовательно предлагается убрать всю информацию о цвете точки. Также предлагается добавить метку для каждой точки, основываясь на информации, представленной структурой названий файлов. Метка используется

для проверки корректности обучения нейронной сети при помощи сравнения самой метки с результатом, полученным в ходе обучения. В табл. 3 представлены метки для той или иной категории, точки которых содержатся в файлах. Так как используется только малая часть дата сета, то определять метку для каждой категории в условиях данного исследования не является обязательным требованием.

Таблица 3. Список используемых категорий и метки, которые им присвоены
[Table 3. The list of categories used and the labels assigned to them]

Категория	Метка
Beam	1
Bookcase	2
Ceiling	3
Clutter	4
Floor	5

1.2. Порядок проведения эксперимента

Эксперимент проводился по следующему плану:

Шаг 1. Подготовка данных к исследованию.

На данном шаге необходимо преобразовать каждое облако точек в форматы *csv*, *npy*, *h5*. Итогом выполнения шага 1 является подготовленная выборка данных, которая позволит провести исследования.

Предлагается для эксперимента следующие требования к структуре файлов csv: в качестве десятичного разделителя в числах оставить точку, между параметрами в качестве разделителя оставить пробел, первые 3 столбца также оставить без изменений, а в качестве 4 столбца добавить цифровую метку, обозначающую принадлежность к той или иной категории.

Теперь посмотрим на выбранные файлы. Исходя из информации, представленной в табл. 2, все файлы, являются сегментами других облаков, поэтому, все точки, описанные в данных файлах, будут иметь лишь одну категорию. Иными словами, в отдельно взятом файле метки будут совпадать для всех точек. Числовые значения для меток описаны в табл. 3.

Для создания каждого из типов файлов достаточно подготовить данные один раз и трижды сохранить. Это будет лучшим вариантом, по сравнению со сбором информации каждый раз для каждого отдельного примера.

В табл. 4 представлено сопоставление названия исходных файлов с полученными.

Таблица 4. Сопоставление файлов
[Table 4. File Mapping]

Исходный файл	Результирующий файл
clutter_32.txt	*output_1*
beam_3.txt	*output_2*
bookcase_8.txt	*output_3*
ceiling_4.txt	*output_4*
ceiling_1.txt	*output_5*
floor_1.txt	*output_6*

Шаг 2. Создание классов-наследников от *torch.utils.data.Dataset*:

Создание классов, возвращающих объекты-наследники *torch.utils.data.Dataset* для каждого из исследуемых форматов файлов.

Для процесса обучения нейронной сети необходимо каким-то образом передать информацию на входной слой нейросети. Для этого процесса во фреймворке *PyTorch* определен класс *torch.utils.data.DataLoader*. Однако, информация, может быть получена из разных источников, например, из того или иного вида файлов, базы данных, по запросу к *api* сервера и т. д. Для того, чтобы *torch.utils.data.DataLoader* понимал, что он должен передавать в нейросеть, необходим класс *torch.utils.data.Dataset*.

Итогом выполнения шага 2 являются созданные экспериментальные классы, которые без трансформации могут подготовить данные из исследуемых файлов для передачи сначала в классы-наследники *torch.utils.data.DataLoader*, которые в свою очередь будут передавать их на входной слой нейронной сети.

Шаг 3. Проведение эксперимента.

На данном шаге реализовывается создание объектов соответствующих классов для проведения эксперимента. Замер времени

должен производится только на чтение данных, никаких дополнительных преобразований с данными не производится. Информация о времени получения данных сводится в словарь, где ключом выступает название файла, а значением — время получения информации.

Для чистоты эксперимента, создание объекта и чтение информации производится трижды для каждой группы файлов с одинаковым расширением, после чего производится расчет среднего значения.

Шаг 4. Визуализация результатов.

Визуализация полученных структур данных в виде столбчатых графиков, показывающих зависимость скорости чтения информации из файла от количества точек и формата файлов, в которых они хранятся.

Для создания графиков сначала необходимо распарсить полученные результаты в вид, удобный для построения графиков, после чего построить сами графики. Каждый график содержит информацию о всех трех экспериментах и среднее значение по одному типу файлов.

2. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В ходе проведения эксперимента были получены следующие графики: на рис. 1 представлен график, иллюстрирующий зависимость скорости чтения файлов из формата *csv* от количества точек, на рис. 2 представлен график, иллюстрирующий зависимость скорости чтения файлов из формата *npy* от количества точек, на рис. 3 представлен график, иллюстрирующий зависимость скорости чтения файлов из формата *h5* от количества точек.

Исходя из информации на графиках следует, что формат *csv*, хоть и является популярным, но для работы с облаками точек плохо подходит, ввиду большого увеличения скорости чтения в зависимости от количества точек.

Форматы *npy* и *h5* показали примерно одинаковые результаты и оба могут использоваться для работы с облаками точек, отклонения на графиках могут объяснятся конкретным состоянием платформы *Google*

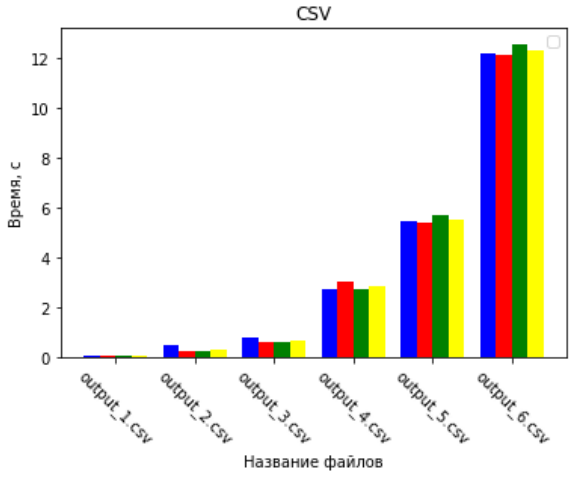

Рис. 1 Результаты экспериментов с файлами в формате csv
[Fig. 1. Results of experiments with csv files]

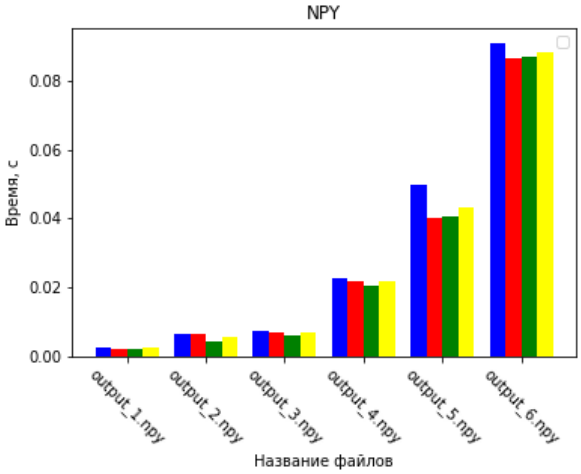

Рис. 2. Результаты экспериментов с файлами в формате npy
[Fig. 2. Results of experiments with files in the npy format]

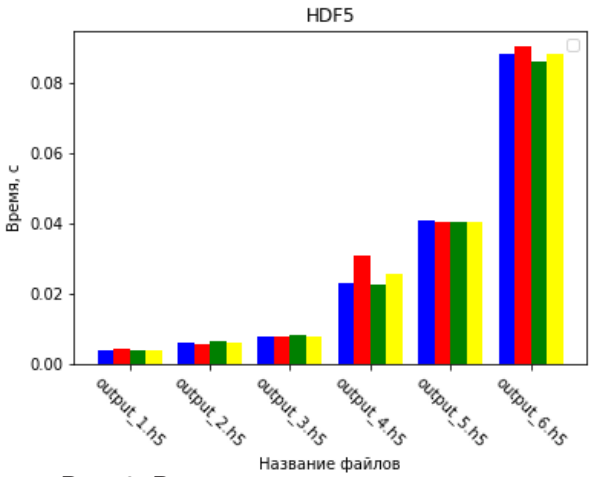

Рис. 3. Результаты экспериментов с файлами в формате h5
[Fig. 3. Results of experiments with files in h5 format]

Collaboratory в момент проведения эксперимента.

Использование того или иного формата из пары *npy* и *h5* могут быть обусловлены и внешними факторами. Например, при работе с фреймворком *PyTorch* будет уместнее использовать формат *npy*, так как дополнительные матричные вычисления можно производить средствами библиотеки *NumPy*, в то же время, формат *HDF5* лучше подходит для долгосрочного хранения информации. Реальные облака точек могут хранится в виде фрагментов в разных файлах. *HDF5* позволяет свести их в один файл и работать с единым источником информации, что может быть полезно при первичной обработке информации и в случаях, когда информация используется для других задач, а не только для обучения нейронной сети, например при визуализации исходных данных.

ЗАКЛЮЧЕНИЕ

В ходе проведенных экспериментов можно сделать следующие выводы о применении исследованных форматов файлов:

1. Формат *.csv не подходит для взаимодействия с файлами, хранящими точки лазерного отражения. С высокой степенью вероятности это связано с тем, что из трех представленных форматов данные являются единственным текстовым, а не бинарным. Следовательно, на хранение одной записи требуется больше ресурсов для чтения, т. к. объем самой записи больше, чем у бинарных файлов.

2. Форматы *.npy и *.h5 примерно равны по производительности. Разницу во времени чтения в пользу того или иного формата можно объяснить спецификой работы платформы Google Collaboratory. Можно предложить сценарий использования для каждого из форматов. *.npy корректнее использовать для взаимодействия с нейронными сетями (загрузка данных, сохранение промежуточных результатов и т. д.) и статистическая обработка информации. *.h5 корректнее использовать в качестве универсального хранилища (объединение информации из нескольких источников, использование в качестве источников для смежных задач, например, визуализация облака, построение полигональной или воксельной модели).

Дальнейшим направлением исследования можно предложить провести аналогичное исследование, где в качестве источника данных будет выступать база данных или выполнение запроса к удаленному серверу. Вторым возможным направлением может быть проведение аналогичного эксперимента со специализированными форматами файлов, направленное на адаптацию результатов под специализированное программное обеспечение.

БЛАГОДАРНОСТИ

Исследование проводилось с использованием оборудования Научно-исследовательского центра пищевых и химических технологий КубГТУ (СКР_3111), разработка которого поддерживается Министерством науки и высшего образования Российской Федерации (Соглашение № 075-15-2021-679)

Исследование выполнено за счет гранта Российского научного фонда № 22-29-00849 «Разработка интеллектуальной информационной системы поддержки принятия решений для решения сложных задач территориального планирования с применением сильного искусственного интеллекта»

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Lean Yu, Xiaoming Zhang, Hang Yin* An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data // Expert Systems with Applications, 2 May 2022. doi.org/10.1016/j.eswa.2022.117363

2. *Dawei Zhao, Qingwei Gao, Dong Sun* Learning view-specific labels and label-feature dependence maximization for multi-view multi-label

classification // Applied Soft Computing, 31 May 2022, doi.org/10.1016/j.cor.2022.105769

3. *Risto Kaijaluoto, Antero Kukko, Harri Kaartinen* Semantic segmentation of point cloud data using raw laser scanner measurements and deep neural networks // ISPRS Open Journal of Photogrammetry and Remote Sensing 16 December 2021 Volume 3 (Cover date: January 2022) Article 100011. doi.org/10.1016/j.ophoto.2021.100011

4. *Di Wang, Lulu Tang, Zhi-Xin Yang* Improving deep learning on point cloud by maximizing mutual information across layers // Pattern Recognition8 July 2022Volume 131 (Cover date: November 2022) Article 108892. doi.org/10.1016/j.patcog.2022.108892

5. *Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, Matt J. Kusner.* Pre-training by completing point clouds// ICLR 2021 Conference Blind Submission, P. 1-20, 2020

6. *Gura D. A. Markovskii I. G. Pshidatok S. K.* Methods of monitoring real estate objects using three-dimensional laser scanning in the specifics of urban lands // Geodesy and cartography = Geodezia i Kartografia. – 2021 . – 82. – P. 45–53. DOI: 10.22389/0016-7126-2021-970-4-45-53.

7. *Dyachenko R. Gura D. Samarin S. Bespyatchuk D. Solodunov A.* Analysis of algorithms for terrestrial recognition of woody vegetation using 3D-laser scanning technology // IOP Conference Series: Earth and Environmental Science (867) 2021, 012166. DOI:10.1088/1755-1315/867/1/012166

8. *Gura D. A. Bespyatchuk D. A. Samarin S. V. Kiryunikova N. M. Lesovaya E. D.* Technology of three-dimensional laser scanning as a tool to provide safety for sport facilities // Nanotechnologies in construction (13), 2021, P. 259–263. DOI: 10.15828/2075-8545-2021-13-4-259-263

9. *Gura D. A., Gribkova I. S., Khusht N. I., Pshidatok S. K.* Knowledge Base as a Part of Intelligent System for Security Monitoring of Infrastructure Objects // Industry Competitiveness: Digitalization, Management, and Integration. Lecture Notes in Networks and Systems (280), 2021, P. 46–52. DOI:10.1007/978-3-030-80485-5_7

10. *Gura D. A., Dubenko Y. V., Shevchenko G. G., Dyshkant E. E., Khusht N. I.* Three-dimensional laser scanning for safety of transport infrastructure with application of neural network algorithms and methods of artificial intelligence // Lecture Notes in Civil Engineering (50), 2020, P. 185-190. DOI: 10.1007/978-981-15-0454-9_17

11. *Можаев А. Н.* Сегментация облаков точек с помощью средств библиотеки point cloud library // Экстремальная робототехника. – 2018. – Т. 1, № 1. – С. 301–308.

12. *Беляевский, К. О.* Применение динамической аллокации на отображаемой памяти для обработки больших облаков точек в библиотеке PCL // Известия Самарского научного центра Российской академии наук. – 2020. – Т. 22, № 1 (93). – С. 56–64.

13. Stanford 2D-3D-Semantics Dataset (2D-3D-S). – Режим доступа: http://buildingparser.stanford.edu/dataset.html – (дата обращения: 04.10.2022).

14. *Arakelov M. S., Lipilin D. A., Dolgova-Shkhalakhova A. V.* Influence of quarantine measures against the new coronavirus infection covid-19 on the state of black sea coastal waters // Geography, Environment, Sustainability. – 2021. – Т. 14, № 4. – С. 199–204.

15. *Дьяченко Р. А., Гура Д. А., Степаненко В. Е., Самарин С. В., Беспятчук Д. А.* К вопросу о принятии решений о выборе оптимального маршрута при размещении оборудования для статических измерений // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2022. – № 3. – С. 63–72.

16. *Гордеев В. А., Шевченко Г. Г.* Статистические процедуры при обработке малых выборок // Известия высших учебных заведений. Геодезия и аэрофотосъемка. – 2021. – Т. 65, № 2. – С. 152–157.

17. *Zhampeissova K., Gura A., Vanina E., Egorova Z.* Academic performance and cognitive load in mobile learning // International Journal of Interactive Mobile Technologies. – 2020. – Т. 14, № 21. – С. 78–91.

18. *Sakka F., Gura A., Latysheva V., Mamlenkova E., Kolosova O.* Solving technological, pedagogical, and psychological problems in mobile learning // International Journal of Interactive Mobile Technologies. – 2022. – Т. 16, № 2. – С. 144–158.

19. *Shestak V., Gura A., Borisova U., Kozlovskaya D.* International Journal of The role of social networks in the organization of the educational process and learning // Interactive Mobile Technologies. – 2021. – Т. 15, № 11. – С. 96–112.

20. *Дьяченко Р. А., Частикова В. А., Лях А. Р.* Реализация атак уклонением на нейронные сети и методы их предотвращения // Электронный сетевой политематический журнал «Научные труды КубГТУ». – 2022. – № 5. – С. 68–77.

Дьяченко Роман Александрович — д-р техн. наук, проф., профессор кафедры Информатики и вычислительной техники Кубанского государственного технологического университета.
E-mail: djachenko.roman@gmail.com
ORCID iD: https://orcid.org/0000-0003-1244-1228

Косолапов Павел Александрович — аспирант кафедры информатики и вычислительной техники Кубанского государственного технологического университета.
E-mail: pawel.kosolapoff@gmail.com
ORCID iD: https://orcid.org/0000-0003-2149-6167

Гура Дмитрий Андреевич — канд. техн. наук, доц., доцент кафедры Кадастра и геоинженерии Кубанского государственного технологического университета, доцент кафедры Геодезии Кубанского государственного аграрного университета им. И. Т. Трубилина.
E-mail: gda-kuban@mail.ru
ORCID iD: https://orcid.org/0000-0002-2748-9622

DOI: https://doi.org/10.17308/sait/1995-5499/2022/4/146-155 ISSN 1995-5499
Received 01.10.2022
Accepted 05.12.2022

ON THE ISSUE OF INCREASING MACHINE LEARNING PERFORMANCE AT THE DATA SAMPLING STAGE WHEN SOLVING CLASSIFICATION PROBLEMS

© 2022 R. A. Dyachenko[1], P. A. Kosolapov[1], D. A. Gura[1,2]✉

[1]Kuban State Technological University
2, Moskovskaya Street, 350072 Krasnodar, Russian Federation
[2]Kuban State Agrarian University named after I. T. Trubilin
13, Kalinina Street, 350044 Krasnodar, Russian Federation

Annotation. The purpose of this study is to determine a data storage method for machine learning tasks of neural networks and semantic segmentation of point clouds. The existing methods of working with large files are considered, experimental studies are carried out to determine the speed of the data reading operation. The experiment consisted in reproducing the process of accessing information from files for which the volume and structure of stored information with time measurement were described. For the research, the most common file formats used for storing information were taken *.csv, *.npy and *.h5. The result of the experiment was statistical information about the file reading time depending on the selected structure and the amount of information stored in it, as well as recommendations for choosing a storage method.
Keywords: data sampling, neural networks, point cloud, SDGs.

✉ Gura Dmitry A.
e-mail: gda-kuban@mail.ru

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Lean Yu, Xiaoming Zhang and Hang Yin* (2022) An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data. *Expert Systems with Applications*, 2 May 2022. doi.org/10.1016/j.eswa.2022.117363

2. *Dawei Zhao, Qingwei Gao and Dong Sun* (2022) Learning view-specific labels and label-feature dependence maximization for multi-view multi-label classification. *Applied Soft Computing*, 31 May 2022, doi.org/10.1016/j.cor.2022.105769

3. *Risto Kaijaluoto, Antero Kukko and Harri Kaartinen* (2022) Semantic segmentation of point cloud data using raw laser scanner measurements and deep neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing,* 16 December 2021. Vol. 3 (Cover date: January 2022) Article 100011. doi.org/10.1016/j.ophoto.2021.100011

4. *Di Wang, Lulu Tang and Zhi-Xin Yang* (2022) Improving deep learning on point cloud by maximizing mutual information across layers. *Pattern Recognition*. 8 July 2022. Vol. 131 (Cover date: November 2022) Article 108892. doi.org/10.1016/j.patcog.2022.108892

5. *Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby and Matt J. Kusner* (202) Pre-training by completing point clouds. *ICLR 2021 Conference Blind Submission*. P. 1–20.

6. *Gura D. A., Markovskii I. G. and Pshidatok S. K.* (2021) Methods of monitoring real estate objects using three-dimensional laser scanning in the specifics of urban lands. *Geodesy and cartography = Geodezia i Kartografia*. (82). P. 45–53. DOI: 10.22389/0016-7126-2021-970-4-45-53.

7. *Dyachenko R., Gura D., Samarin S., Bespyatchuk D. and Solodunov A*. (2021) Analysis of algorithms for terrestrial recognition of woody vegetation using 3D-laser scanning technology. *IOP Conference Series: Earth and Environmental Science* (867). 012166. DOI:10.1088/1755-1315/867/1/012166

8. *Gura D. A., Bespyatchuk D. A., Samarin S. V., Kiryunikova N. M and Lesovaya E. D.* (2021) Technology of three-dimensional laser scanning as a tool to provide safety for sport facilities. *Nanotechnologies in construction*. (13). P. 259–263. DOI: 10.15828/2075-8545-2021-13-4-259-263

9. *Gura D. A., Gribkova I. S., Khusht N. I. and Pshidatok S. K.* (2021) Knowledge Base as a Part of Intelligent System for Security Monitoring of Infrastructure Objects. I*ndustry Competitiveness: Digitalization, Management, and Integration. Lecture Notes in Networks and Systems*. (280). P. 46–52. DOI:10.1007/978-3-030-80485-5_7

10. *Gura D. A., Dubenko Y. V., Shevchenko G. G., Dyshkant E. E. and Khusht N. I.* (2020) Three-dimensional laser scanning for safety of transport infrastructure with application of neural network algorithms and methods of artificial intelligence. *Lecture Notes in Civil Engineering* (50). P. 185–190. DOI: 10.1007/978-981-15-0454-9_17

11. *Mozhaev A. N.* (2018) Segmentatsiya oblakov tochek s pomoshch'yu sredstv biblioteki point cloud library. Ekstremal'naya robototekhnika. V. 1, No 1. P. 301–308.

12. *Belyaevskiy K. O.* (2020) Primenenie dinamicheskoy allokatsii na otobrazhaemoy pamyati dlya obrabotki bol'shikh oblakov tochek v biblioteke PCL. *Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk*. V. 22, No 1 (93). P. 56–64.

13. Stanford 2D-3D-Semantics Dataset (2D-3D-S). Access mode: http://buildingparser.stanford.edu/dataset.html. (Retrieved date: 04.10.2022).

14. *Arakelov M. S., Lipilin D. A. and Dolgova-Shkhalakhova A. V.* (2021) Influence of quarantine measures against the new coronavirus infection covid-19 on the state of black sea coastal waters. *Geography, Environment, Sustainability*. V. 14, No 4. P. 199–204.

15. *Dyachenko R. A., Gura D. A., Stepanenko V. E., Samarin S. V. and Bespyatchuk D. A.* (2022) On the issue of decision-making on the choice of an optimal route when placing equipment for static measurements. *Bulletin of the Vo-*

ronezh State University. Series: System Analysis and Information Technologies. No. 3. P. 63–72.

16. *Gordeev V. A. and Shevchenko G. G.* (2021) Statistical procedures for processing small samples. *Izvestia of higher educational institutions. Geodesy and aerial photography*. V. 65, No 2. P. 152–157.

17. *Zhampeissova K., Gura A., Vanina E. and Egorova Z.* (2020) Academic performance and cognitive load in mobile learning. *International Journal of Interactive Mobile Technologies*. V. 14, No 21. P. 78–91.

18. *Sakka F., Gura A., Latysheva V., Mamlenkova E. and Kolosova O.* (2022) Solving technological, pedagogical, and psychological problems in mobile learning. *International Journal of Interactive Mobile Technologies*. V. 16, No 2. P. 144–158.

19. *Shestak V., Gura A., Borisova U. and Kozlovskaya D.* (2021) International Journal of The role of social networks in the organization of the educational process and learning. *Interactive Mobile Technologies*. V. 15, No 11. P. 96–112.

20. *Dyachenko R. A., Chastikova V. A. and Lyakh A. R.* (2022) Implementation of evasion attacks on neural networks and methods of their prevention. *Electronic network polythematic journal "Scientific works of KubSTU"*. No 5. P. 68–77.

Dyachenko Roman A. — Doctor of Technical Sciences, Professor, Professor of the Department of Informatics and Computer Engineering of the Kuban State Technological University.
E-mail: djachenko.roman@gmail.com
ORCID iD: https://orcid.org/0000-0003-1244-1228

Kosolapov Pavel A. — Candidate of the Department of Computer Science and Computer Engineering of the Kuban State Technological University.
E-mail: pawel.kosolapoff@gmail.com
ORCID iD: https://orcid.org/0000-0003-2149-6167

Gura Dmitry A. — Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Cadastre and Geo-Engineering of the Kuban State Technological University, Associate Professor of the Department of Geodesy of the Kuban State Agrarian University named after I. T. Trubilin.
E-mail: gda-kuban@mail.ru
ORCID iD: https://orcid.org/0000-0002-2748-9622