DOI: https://doi.org/10.36910/6775-2524-0560-2021-42-11
УДК 0048:681
Логвін Антон Олексійович, аспірант
https://orcid.org/0000-0002-5913-9395
Харківський національний університет радіоелектроніки

ГЛИБИННЕ НАВЧАННЯ ДЛЯ АУДІО-ДОДАТКІВ

Логвін А. О. Глибинне навчання для аудіо-додатків. Розкрито принципи застосування глибокого навчання для нейронних мереж щодо розпізнавання аудіо-сигналів. Відокремлено області подання звуку. Підкреслено, що дослідження буде обмежено аудіо-сигналами. Описано принципи розбиття сигналу на складові елементи та їх вилучення із аудіо запису. Наведено схему формування розподілу аудіо-сигналу та запропоновано загальний підхід до задачі розпізнавання аудіо-сигналів. Він умовно поділений на три окремі етапи: обробка аудіо-запису та його перетворення у частотно-часову область, побудова спектрограми та її перетворення на формат з подальшим виведенням послідовності ознак у вигляді векторів. Визначений коефіцієнт накладання та середньозважений коефіцієнт перекриття (частковий збіг). Сформовано низку значень на основі проведеного експерименту, які показали, що на характеристики / параметри аудіо-додатків, сформовані за допомогою нейронної мережі з глибоким навчанням, має вплив метод підготовки даних, додавання шарів та формування спектру одиниць, що покращує результат за рахунок помноженого часу навчання, те саме стосується і періодичних з'єднань.
Ключові слова: машинне навчання, глибоке навчання, нейронна мережа, аудіо-сигнал, аудіо-додаток, розпізнавання, акорд, музика.

Логвин А. А. Глубинное обучения для аудио-приложений. Раскрыты принципы применения глубокого обучения для нейронных сетей по распознаванию аудио-сигналов. Обособленно области представление звука. Подчеркнуто, что исследование будет ограничено аудио-сигналами. Описаны принципы разбиения сигнала на составляющие элементы и их изъятия из аудио записи. Приведена схема формирования распределения аудио-сигнала и предложен общий подход к задаче распознавания аудио-сигналов. Он условно разделен на три отдельные этапы: обработка аудио-записи и его преобразования в частотно-временной область, построение спектрограммы и ее преобразования в формат с последующим выводом последовательности признаков в виде векторов. Определенный коэффициент наложения и средневзвешенный коэффициент перекрытия (частичное совпадение). Сформирован ряд значений на основе проведенного эксперимента, которые показали, что на характеристики / параметры аудио-приложений, сформированные с помощью нейронной сети с глубоким обучением, влияет метод подготовки данных, добавления слоев и формирования спектра единиц, улучшает результат за счет умноженного времени обучения , то же касается и периодических соединений.
Ключевые слова: машинное обучение, глубокое обучение, нейронная сеть, аудио-сигнал, аудио-приложение, распознавания, аккорд, музыка.

Lohvin Anton. **Deep machine learning for audio applications.** The principles of application of deep learning for neural networks for the recognition of audio signals are disclosed. Apart from the area of sound presentation. It is emphasized that the study will be limited to audio signals. The principles of signal splitting into constituent elements and their removal from audio recording are described. A diagram of the formation of the distribution of an audio signal is given and a general approach to the problem of recognizing audio signals is described. It is conventionally divided into three separate stages: processing of audio recording and its transformation in the time-frequency domain, construction of a spectrogram and its transformation into audio format, followed by outputting a sequence of features in the form of vectors. The overlap ratio and the weighted average overlap ratio (overlap) have been determined. A number of values were formed based on the experiment, which showed that the characteristics / parameters of audio applications formed using a neural network with deep learning are affected by the data preparation method, adding layers and forming a spectrum of units improves the result due to the multiplied training time, the same also applies to periodic connections.
Keywords: machine learning, deep learning, neural network, audio signal, audio application, recognition, chord, music.

Вступ та постановка проблеми дослідження. Методи глибокого навчання – це група методів машинного навчання, які можуть вивчати функції ієрархічно від нижнього рівня до вищого, створюючи глибоку архітектуру [1]. Методи глибокого навчання мають можливість автоматично вивчати функції на декількох рівнях, що дозволяє системі вивчати складні функції зіставлення безпосередньо з даних, без допомоги функцій, створених людиною. Ця здатність має вирішальне значення для абстракції високорівневих функцій, оскільки останні важко описати безпосередньо з необроблених навчальних даних. Більш того, з різким зростанням обсягів даних можливість автоматичного вивчення високорівневих функцій стає ще більш важливою$f: X \to Y$.Найхарактернішою рисою методів глибокого навчання є те, що всі їх моделі мають глибоку архітектуру. Цей факт означає, що в мережі є кілька прихованих рівнів.

Обробка звуку охоплює безліч різних областей, всі з яких пов'язані з поданням звуку слухачам. Найбільш важливими є три області: відтворення музики з високою точністю, наприклад, на аудіокомпакт-дисках; голосовий зв'язок; синтетична мова, при якій комп'ютери генерують і розпізнають зразки людського голосу.

У рамках даної статті дослідження буде обмежено аудіо-сигналами. Одним з важливих напрямків пошуку музичної інформації є оцінка елементів, пов'язаних з музичними концептами. Такі

поняття, як нота, акорд, мелодія, тональність, можна визначити по нотній партитурі, але їх вилучення зі звукового сигналу – дуже складне завдання навіть для людини.

Кінцева мета автоматичної оцінки всіх музичних понять аудіо звуку або отримання його транскрипції є проблематичною в умовах сьогодення. Дослідники працюють над різними аспектами цього завдання, такими як витяг звукової мелодії, виявлення звуків або оцінка звукових акордів. Інформація про акорди, що міститься в записі, корисна для структурної сегментації. Ця інформація цінна сама по собі, так як вона може бути використана для індексації музичних записів по їх утриманню, щоб допомогти музикознавцям проаналізувати гармонію пісні або створити унікальну мелодію.

Аналіз останніх досліджень і публікацій. В останні кілька років, чимало, як зарубіжних так і вітчизняних вчених здійснило відкриття у сфері глибокого навчання для аудіо-додатків.

Алгоритми машинного та глибокого навчання та їх використання в прикладних додатках, дослідили В.М. Бродкевич та В.Я. Ремесло [1]. Авторами розглянуто основний смисл машинного навчання (МН), основні визначення терміну машинне навчання та подаються різні точки зору дослідників і розробників експертних організацій; моделі МН та наочне уявлення про них; процес навчання машин; глибоке навчання (ГН) та базові алгоритми глибоких нейронних мереж (згорткових нейронних мереж) як основи ГН; алгоритми машинного навчання в прикладних додатках.

Стосовно акустичних даних варто відмітити роботу А.Г. Кривохати, О.В. Кудіна та А.О. Лісняка [2]. Автори здійснили огляд методів машинного навчання для класифікації акустичних даних. Для підвищення якості розпізнавання аудіо-сигналів, авторами пропонується використання ансамблевого навчання із застосуванням класифікаторів на основі ознак та глибинних нейронних мереж. Науковці наголошують, що різні класифікатори, на вхід яких подаються різні вектори ознак або дані без попередньої обробки, можуть бути відносно ефективними на різних даних, але об'єднуватися в один ефективний класифікатор-ансамбль. Перевагою такого підходу може бути його адаптивність з точки зору вимогливості до обчислювальних ресурсів, оскільки, за необхідністю, можна коректувати кількість класифікаторів, які беруть участь в аналізі.

А.Г. Кривохата [3] присвятила свою роботу розв'язанню задач класифікації звукових сигналів засобами гібридних нейронних мереж із шарами згортки та автокодувальників з оптимізацією їх структури генетичними алгоритмами.

М. М. Глибовець, А. П. Жиркова [4] розглянули особливості використання методів машинного навчання (МН) для класифікації звукової інформації на прикладі розв'язку задачі класифікації міських звуків (МЗ).

Також варто відмітити роботи таких вчених, як: F. Alias, J.C. Socoro, X. Sevillano [5], Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P.J.B. Jackson, M.D. Plumbley[6], F. Camastra, A. Vinciarelli[7], B.L. Sturm, A. Nürnberger, S. Stober, B. Larsen, M. Detyniecki[8], CiresanDan, UeliMeier, JonathanMasci, Luca M. Gambardella, Jurgen Schmidhuber [9], J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter [10], Q. Kong, Y. Xu, W. Wang, M.D. Plumbley[11], J. Stastný, V. Skorpil, J. Fejfar[12], G. Wichern, M. Yamada, H. Thornburg, M. Sugiyama, A. Spanias[13], G. Zaccone, Md. R. Krim [14], H. Purwins, B. Li, T. Virtanen, J. Schlüter, SY. Chang, T. Sainath [15], та інші.

Проте, враховуючи описані наукові набутки, за темою, питання дослідження принципів глибокого навчання для аудіо-додатків залишається відкритим та потребує детального опрацювання.

Мета статті. Розкрити принципи глибокого навчання для аудіо-додатків.

Викладення основного матеріалу дослідження. Побудова моделей машинного навчання для класифікації, опису або генерації аудіо-сигналів зазвичай пов'язана із завданнями моделювання, в яких вхідними даними є аудіосемпли. Аудіосемпли зазвичай представлені у вигляді часових рядів, де вимір по осі Y являє собою амплітуду сигналу. Амплітуда зазвичай вимірюється як функція зміни тиску навколо мікрофону або приймального пристрою, який спочатку приймав звук, а по осі X – час. Попередня обробка набору даних, витяг функцій і розробка функцій – це кроки, які необхідні для витягу інформації з базових даних, інформації, яка в контексті машинного навчання повинна бути корисною для прогнозування класу вибірки або значення деякої цільової змінної. На тлі аудіо-аналізу цей процес, значною мірою, побудовано на пошуку компонентів аудіо-сигналу, які відрізняють його від інших сигналів.

Музичні аудіо-додатки – це тимчасові ряди, в яких події організовані в музичному, а не в реальному часі, які змінюється в залежності від ритму і виразів. Вимірювані сигнали зазвичай об'єднують кілька голосів, які синхронізовані за часом і перекриваються по частоті, змішуючи як короткострокові, так і довгострокові тимчасові залежності. Фактори впливу включають музичну

традицію, стиль, композитора та інтерпретацію. Висока складність і різноманітність породжують проблеми подання сигналів, що добре підходять для високих рівнів абстракції, які забезпечуються перцептивно і біологічно мотивованими методами обробки глибокого навчання.

Типовий аудіо-сигнал можна виразити як функцію амплітуди і часу.

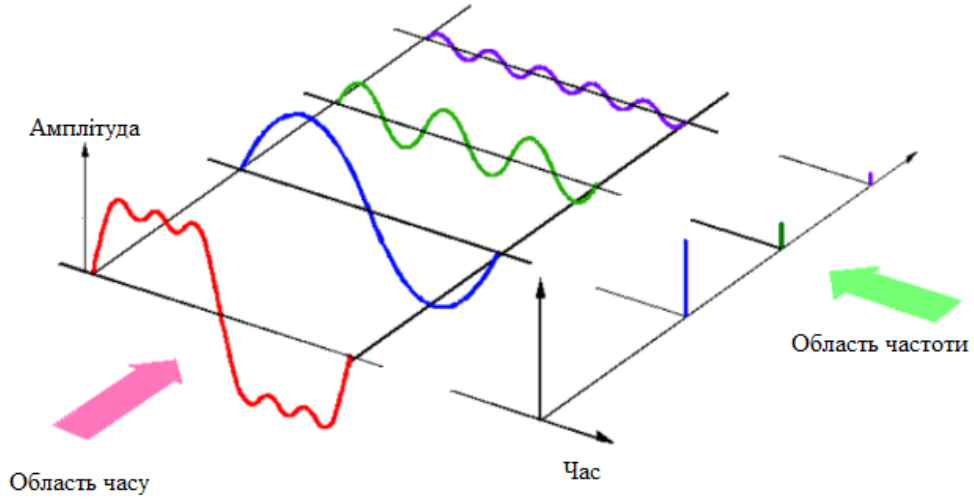

Рис. 1. Схема формування розподілу аудіо-сигналу [16]

Деякі пристрої можуть уловлювати ці звуки і представляти їх у форматі необхідному для обробки на машині. Приклади цих форматів:

– wav (Waveform Audio File);
– WMA (Windows Media Audio);
– mp3 (MPEG-1 Audio Layer 3).

Процес обробки звуку включає витяг акустичних характеристик, що відносяться до поставленого завдання, за якими слідують схеми прийняття рішень, які включають виявлення, класифікацію та об'єднання знань.

Загальний підхід до задачі розпізнавання аудіо-сигналів можна розділити на 3 етапи. Спочатку аудіо-запис попередньо обробляється (тут часто визначається биття) і перетворюється у частотно-часову область наприклад, з використанням швидкого перетворення Фур'є або Q перетворення, в результаті чого вийде так звана спектрограма. Кожен стовпець спектрограми представляє спектр короткого фрагменту оригінального запису.

Потім до спектрограми застосовується серія перетворень для розгляду аудіо. Властивості і фактори, згадані вище, за допомогою яких аудіо-сигнал перетворюється на послідовність ознак у вигляді векторів.

Деякими широко використовуваними типами ознак є вектори профілю класу основного тону і логарифмічний крок кольору [3]. Їх зазвичай називають векторами кольоровості, тому що вони мають 12 компонентів, по одному на кожен клас висоти звуку (клас висоти звуку об'єднує всі частоти, які відповідають нотам з однаковою назвою, що належать різним октавах). Нарешті, послідовність векторів ознак перетворюється на послідовність символів хорди. Межі акордів утворюються через початковий поділ аудіо-запису на набір фрагментів з фіксованими межами. Імовірнісні моделі, такі як приховані марковські моделі і динамічні байєсовські мережі часто використовуються з причини їх здатності моделювати серії послідовних подій і включень різних факторів (басова нота, музична тональність).

Цільова послідовність хордових команд може бути отримана з послідовності спостережуваних векторів ознак шляхом застосування до моделі алгоритму Вітербо.

Прості алгоритми, які порівнюють тільки вектори ознак з визначеним шаблоном вектору для акордів [4] працюють трохи гірше, але не вимагають навчання і тому не можуть бути адаптовані до конкретної послідовності акордів або музичного стилю. Гарна якість розпізнавання акордів може бути досягнута без імовірнісних моделей, використовуючи більш досконалі функції з більш простим класифікатором.

Багатообіцяючі результати в цьому напрямку були отримані в [6] за допомогою глибокої згорткової нейронної мережі. У відповідності до розпізнавання акордів – багатошарові перцептрони, попередньо навчені як складові автокодери шумозаглушення. На відміну від згортальних нейронних

мереж вони обробляють спектрограму, як послідовність стовпців, а не як композицію прямокутників. Мережі цього типу успішно застосовуються для автоматичного розпізнавання мови [5].

Серйозний недолік нейронної мережі – це відсутність наочної інтерпретації значень в прихованих шарах. Таким чином, можна додати додатковий шар логістичної регресії з 12 виходами, щоб зробити мережевий вихід.Цей додатковий шар пов'язаний з прихованим шаром внутрішнього автоенкодера, і його висновок буде розглядатися як вектор кольоровості. Фактично, автоенкодери використовуються для попереднього навчання мережевих шарів і розміщення їх параметрів в області, де, ймовірно, знаходяться їх кращі значення. Потрібне тонке налаштування для настройки параметрів останнього шару логістичної регресії.

Дані для навчання беруться з відомих міток акордів. Кожен з них може бути представлений як ідеальний 12-мірний вектор, який має 1 на позиціях, що відносяться до нот даного акорду і 0 на інших позиціях. Ці вектори часто називаються бінарними шаблонами акордів. Хорда (N) не відповідає нульовому вектору, і тому можна виявити, коли елемент результуючого вектора з максимальним абсолютним значенням не більш Δ – параметр, який коригується дослідним шляхом.

Бінарні шаблони для всіх акордів даного типу (наприклад, мажорні акорди) можна отримати з шаблону акордів до мажору шляхом циклічних перестановок його компоненти. Таким чином, для кожного стовпця спектрограми можна згенерувати 12 вхідних векторів, і 12 відповідних вихідних векторів з відповідною міткою хорди, шляхом застосування циклічної перестановки в одну сторону 12 разів. Таким чином, на рівну кількість всіх акордів може бути отриманий один тип навчальних даних. Але циклічна перестановка спектра компоненти неприродна, бо найвищі спектральні компоненти переміщаються до найнижчих одиниць. Так що краще замінити його ковзаючим вікном і дозволити вихідному спектру охопити на 1 октаву більше, забезпечуючи 12 додаткових значень.

Циклічна перестановка входів або компонентів вектора кольоровості –це механізм коли параметри навчальної моделі в алгоритмах розпізнавання акордів. Пропонується його використовувати як при навчанні автокодувальника з шумозаглушенням, так і при його тестуванні на невідомих входах. Вектори кольоровості для нової спектрограми можуть бути розраховані 12 разів з 12 різними параметрами, потім повернені до одного і того ж кореня і усереднені. Це потенційно може привести до кращої якості розпізнавання.

Щоб компенсувати різницю в загальній кількості мажорних і мінорних акордів в межах навчального набору, необхідно обмежити цю різницю значенням 100 (що перетворюється в 1200 через циклічну перестановку) в створеному наборі прикладів для підготовки нейронної мережі.

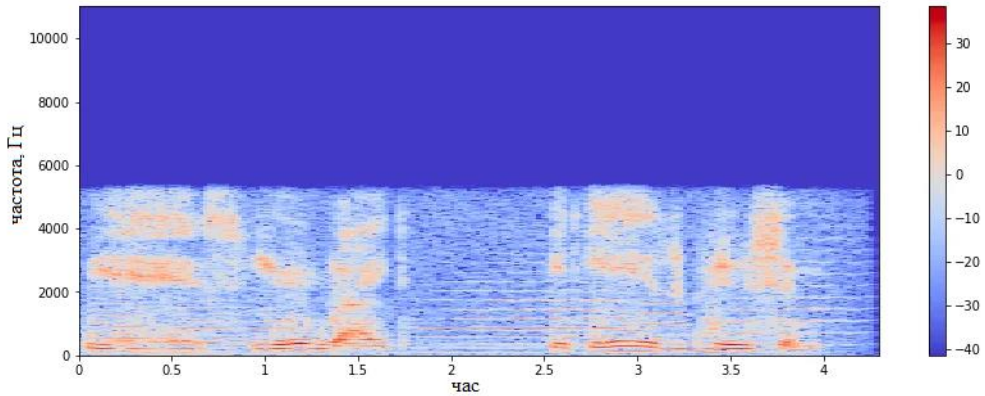

Рис. 2. Початкова спектрограма

Початкова спектрограма (рис. 2) охоплює 6 октав: від C2 (65,41 Гц) до B7 (4751 Гц) і має 1 рядок. Отже, кожен стовпець спектрограми складається з 72 значень, які представляють інтенсивність відповідної ноти на даному звуковому фрагменті. Коли менше компонентів необхідних для розрахунків, завжди використовуються найнижчі з них. Логарифмічні перетворення застосовується до кожного значення, щоб імітувати людське сприйняття інтенсивності звуку: кожне значення v замінюється на $log10\,(1000v\,+\,1)$, як запропоновано в [16]. Кожен стовпець спектрограми потрібно нормалізувати, щоб його значення знаходились у інтервалі [0, 1] перед передачею в нейронну мережу.

Вектори ознак кольоровості отримують з нейронної мережі, навченої, як описано вище. Мережа у всіх експериментах має 12 виходів, але кількість входів, кількість прихованих шарів і їх розмір змінювалися в залежності від експерименту. Були розглянуті наступні варіанти:

– описана мережа з 48 або 60 входами, 1, 2 або 3 прихованими шарами і без повторюваних з'єднань;

– описана мережа з 48 або 60 входами, 1, 2 або 3 прихованими шарами і повторюваними з'єднаннями внутрішнього прихованого шару з самим собою.

Спочатку кожен вектор ознак замінюється лінійною комбінацією 10% найбільш схожих векторів, де вага кожного вектора дорівнює його подібності з вихідним вектором ознак. Подібність розраховується як евклідова відстань між двома векторами. Це дає можливість врахувати повторення музичних фраз.

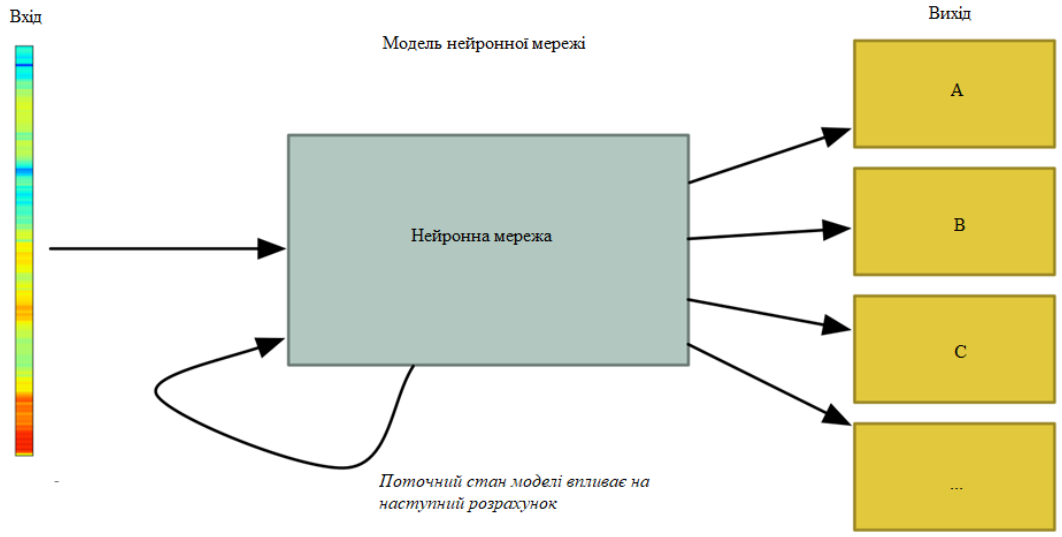

Рис. 3. Модель глибокого навчання нейронної мережі

Крім того, були реалізовані дві евристики для виправлення помилок визначення акорду. Перша евристика полягає в пошуку всіх інтервалів, де кілька акордів одного кореня, але іншого типу розташовуються поруч один з одним, а потім замінюючи кожен інтервал з однією міткою акорду. Цей акорд обраний як найближча k сума всіх векторів ознак в межах інтервалу. Була введена ще одна евристика для фіксації послідовності акордів, наприклад (A, B, C), де 3 послідовні частки позначені 3 різними акордами. Кожна подібна послідовність замінюється однією з (A, A, C), (A, C, C), (B, B, C), (A, B, B), для яких сума відстаней від слідуючих один за одним векторів ознак до відповідних хордових міток мінімальна.

Спочатку запис був розділений на сегменти з використанням усіх кордонів акордів як від еталонної, так і від розрахункової послідовності. Потім на кожному сегменті здійснили оцінку тріад, її значення s дорівнює 1 на відрізку при оцінці та опорний акорд на цьому відрізку зведені до відповідних їм тріад (перші 3 ноти) вони рівні або в обох відсутні акорди. В іншому випадку $s = 0$. Тоді коефіцієнт накладання розраховується для цього запису як:

$$K_n = \sum_{i=1}^{N_{\text{сегм}}} s_i l_i \Big/ \sum_{i=1}^{N_{\text{сегм}}} l_i$$

де s_i – оцінка на і-му сегменті, l_i – довжина і-го відрізка, $N_{\text{сегм}}$ – загальна сума кількості сегментів у аудіо.

Середньозважений коефіцієнт перекриття визначений для весь набір даних як:

$$K_p = \sum_{i=1}^{N_{\text{трек}}} OR_i L_i \Big/ \sum_{i=1}^{N_{\text{трек}}} L_i$$

деOR_i– коефіцієнт перекриття для і-го запису, L_i– загальна тривалість і-гоаудіо записуі $N_{трек}$– це розмір набору даних (318 записів).

Отримані значення зведемо до таблиці 1.

Таблиця 1

Значення

№	Конфігурація	Частковий збіг	Сегментація
Мережа без повторюваних з'єднань			
1	48, 200	0,7640	0,8135
2	48, 200, 200	0,7620	0,8019
3	60, 100	0,7620	0,7999
4	60, 200	0,7642	0,8011
5	60, 300	0,7648	0,8021
6	60, 200, 100	0,7666	0,8021
7	60, 300, 300	0,7655	0,8022
8	60, 100, 100, 100	0,7679	0,7998
9	60, 300, 300, 300	0,7677	0,8010
Мережа з повторюваними з'єднаннями внутрішнього прихованого шару з самим собою			
1	48, 200	0,7620	0,7955
2	48, 200, 200	0,7663	0,7979
3	60, 300	0,7615	0,7949
4	60, 300, 300	0,7680	0,7980
5	60, 300, 300, 300	0,7686	0,7999

Кількість входів у мережі відносно невелика. Тому експерименти проводилися лише з 1, 2 та 3 прихованими шарами. Швидкість навчання для попередньої підготовки автокодерів та мережевого навчання була встановлена на 0,03 та 0,01 відповідно; як попередня підготовка, так і глибоке навчання проводились протягом 15 епох. Розмір партії для градієнтного спуску партії був встановлений на 5. Стандартне відхилення для ізотропного гаусового шуму було встановлено на 0,2, і шум застосовувався до кожної спектральної складової з імовірністю p = 0,7. Ймовірність встановлення векторного компонента на 0 була встановлена на 0,2 для маскування шуму.

Конфігурація мережі вказує на кількість одиниць у вхідному шарі та приховані шари нейронних мереж або кількість спектральних компонентів, що використовуються для обчислення характеристик аудіо-сигналів. В експериментах повторювали лише останній попередньо підготовлений шар.

Перший експеримент спрямований на виявлення найкращого макета мережі. Використовувались однакові макети з повторюваними підключеннями та без них. Спектр обчислювали з 6 або 5 октав, використовуючи ковзне 5 або 4 октави (60 та 48 входів відповідно). Застосовувались як логарифмічне перетворення, так і пост-обробка.

Незважаючи на те, що конфігурації з більшою кількістю одиниць досягли трохи кращого результату, істотних відмінностей між результатами не було. Ні наявність повторюваних зв'язків, ні збільшення розміру вхідного вектора значно не покращують розпізнавання аудіо-сигналів. Лише для деяких пар різниця була статистично значущою. Загальний час глибокого навчання та тестування сильно зростає із додаванням більшої кількості шарів та періодичних зв'язків.

Висновки та перспективи подальших досліджень. У роботі досліджено, як глибоке навчання нейронної мережі, яка обробляє спектрограму стовпець за стовпцем (на відміну від згорткових нейронних мереж), може створити ефективні кольорові функції, які можна використовувати для розпізнавання звукових акордів у аудіо-додатку.

Можна зробити висновок, що характеристики аудіо-додатків, розраховані за допомогою нейронної мережі з глибоким навчанням, дозволяють досягти практично тієї ж якості розпізнавання аудіо-сигналів, що нейронні мережі без навчання. Але конфігурація мережі та тип шуму, доданий під час пошарової попередньої підготовки, мають менше значення, ніж метод підготовки даних. Дані кращого спектру можуть покращити як звичайні функції, так і такі, що базуються на нейромережах. Іншим важливим висновком є те, що додавання шарів та одиниць покращуєрезультат за рахунок помноженого часу навчання. Те саме стосується і періодичних з'єднань. Крім того, вони роблять процес тестування набагато повільнішим, оскільки результат отриманого вектору на кожному кроці

залежить від усіх попередніх кроків, а отже, окремі вхідні вектори не можуть бути оброблені паралельно.

Перспективи подальших досліджень базуються на об'єднанні цих функцій із звичайними, що може призвести до кращої якості розпізнавання аудіо-сигналів.

Список бібліографічного опису.

1. Бродкевич, В. М., & Ремесло, В. Я. (2018). Алгоритми машинного навчання (МН) та глибокого навчання (ГН) і їх використання в прикладних додатках. *Міжнародний науковий журнал Інтернаука*, (11 (1)), 56-60.
2. Кривохата, А. Г., Кудін, О. В., & Лісняк, А. О. (2018). Огляд методів машинного навчання для класифікації акустичних даних. *Вестник Херсонского национального технического университета*, (3-1 (66), 327-331.
3. Кривохата А. Г. (2020). Нейромережеві математичні моделі звукових сигналів у задачах розпізнавання : дисертація канд. фізико-матем. наук: 01.05.02; Запорізький національний університет Міністерства освіти і науки України, Запоріжжя. 160 с.
4. Глибовець, М. М. Жиркова, А. П. (2019). Використання машинного навчання у задачах класифікації звуків. *Наукові записки НаУКМА. Комп'ютерні науки*, 2, 22-31.

References.

1. Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, *6*(5), 143, 1–44.
2. Xu, Y., Huang, Q., Wang, W., Foster, P., Sigtia, S., Jackson, P. J., & Plumbley, M. D. (2017). Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(6), 1230-1241.
3. Camastra, F., & Vinciarelli, A. (2015). *Machine learning for audio, image and video analysis: theory and applications*. Springer.
4. Sturm, B. L. (2012, October). A survey of evaluation in music genre recognition. In *International Workshop on Adaptive Multimeia Retrieval* (pp. 29-66). Springer, Cham.
5. Sturm, B. L. (2012, October). A survey of evaluation in music genre recognition. In *International Workshop on Adaptive Multimedia Retrieval* (pp. 29-66). Springer, Cham.
6. Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011, June). Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*.
7. Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011, June). Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*.
8. Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 776-780). IEEE.
9. Xu, Y., Kong, Q., Huang, Q., Wang, W., & Plumbley, M. D. (2017, May). Convolutional gated recurrent neural network incorporating spatial features for audio tagging. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 3461-3466). IEEE.
10. Stastny, J., Skorpil, V., & Fejfar, J. (2013, July). Audio data classification by means of new algorithms. In *2013 36th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 507-511). IEEE.
11. Wichern, G., Yamada, M., Thornburg, H., Sugiyama, M., & Spanias, A. (2010, March). Automatic audio tagging using covariate shift adaptation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 253-256). IEEE.
12. Zaccone, G., Karim, M. R., & Menshawy, A. (2017). *Deep learning with TensorFlow*. Packt Publishing Ltd.
13. Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206-219. https://doi.org/10.1109/JSTSP.2019.2908700
14. Music Genre Classification With Python. (2021). https://towardsdatascience.com/music-genreclassification-with-python-c714d032f0d8. – Title from the screen