«Системні технології» 6 (149) 2023 «System technologies»
DOI 10.34185/1562-9945-6-149-2023-02
УДК 004.93'11

О.Р. Пеня, Є.С. Сулема

ЗАСТОСУВАННЯ ГЛИБОКИХ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ КЛАСИФІКАЦІЇ МУЛЬТИМОДАЛЬНИХ ДАНИХ

Анотація: Розвиток програмних і апаратних технологій дозволяє отримувати, зберіга ти та обробляти набори мультимодальних даних, які краще описують стан досліджува ного об'єкта, ніж дані однієї модальності. Тому дослідження та розвиток методів ефек тивної обробки таких даних є актуальною задачею. У статті запропоновано спосіб ор ганізації обробки мультимодальних даних, який адаптується до наявних обчислювальних можливостей системи, для використання у системі цифрового двійника в режимі «циф рової тіні» або відокремленої симуляції. Також описано приклад створення програмного модуля класифікації мультимодальних даних у реальному часі.
Ключові слова: мультимодальні дані, класифікація, паралельні обчислення, штучні ней ронні мережі.

Вступ. Розвиток різноманітних технічних засобів збору, зберігання та передачі даних роблять можливим різноплановий моніторинг та збір інформації про певні об'єкти, результатом яких є набори мультимодальних гетерогенних даних. Приклади систем, що генерують такі набори, наведено у табл. 1.

Таблиця 1

Приклади мультимодальних даних в різних галузях

Галузь	Дані	Область використання
Виробництво	Інформація про стан машин та продуктів виробництва із контролерів станків з ЧПУ та зовнішніх сенсорів, в тому числі мережі IIoT; введення операторів	Моделювання та прогнозування процесу виробництва, обслуговування обладнання
Медицина	Медичні дані пацієнтів (вік, вага, тощо) та результати досліджень від вимірів температури і тиску до багатошарових зображень КТ	Діагностика захворювань та виявлення ризиків для здоров'я

Продовження табл. 1

Інтернет	Неструктуровані зображення, відео та аудіо, тексти	Веб-скрейпінг, індексація, добування даних, виявлення трендів
Розумні будинки	Показники сенсорів мережі IoT	Моніторинг та автономна підтримка стану в ощадливий спосіб

Наявність більшої кількості даних про об’єкт дозволяє точніше визначити та передбачити його стан завдяки додатковій безпосередній інформації з одного боку та наявності змістовної інформації у зв’язках між модальностями, які впливають одна на одну з іншого. Окрім цього, різні джерела даних дозволяють зменшити вплив шуму та аномальних викидів на результати класифікації. Тому, методи обробки таких наборів даних є предметом численних досліджень в областях поєднання даних, класифікації та кластеризації, моделювання, прогнозування та цифрових двійників.

Постановка проблеми. Задача, що розв’язується за допомогою методів класифікації, полягає у визначені приналежності об’єкту до одного з визначених класів за певними ознаками при заданій вибірці об’єктів, для яких відома їх класова приналежність. Використання мультимодального набору даних як навчальної вибірка передбачає виділення ознак для класифікації із різних модальностей, певним чином логічно пов’язаних (як правило, синхронізованих за часом).

Ця стаття присвячена створенню програмного модуля для класифікації мультимодальних даних у режимі реального часу або близькому до нього з метою його використання як компонента системи цифрового двійника, який працює як цифрова тінь або відокремлений симулятор.

Аналіз літературних джерел. У статті [1] наведено детальний огляд принципів, задач та викликів мультимодального машинного навчання, теоретичні засади їх вирішення та їх розвиток в інших роботах.

У роботі [2] пропонується спільне використання візуальних та звукових даних для підвищення точності розпізнавання мови. Автори використовують бімодальний автокодувальник, який навчають на модифікованому наборі даних, де в частині даних відсутні значення однієї з модальностей, але мережа повинна виділити ознаки обох. Таким чином, очікується, що мережа навчанні

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

мережа виявлятиме кореляцію між ознаками, виділеними з окремих модальностей.

Автори статті [3] використовують декілька незалежних мереж для обробки модальностей із подальшим поєднанням результатів класифікації завдяки зваженому середньому, ваги якого підбираються на значеннях валідаційного набору. В контексті задачі дослідження цей підхід має декілька переваг: мережі не пов'язані між собою, тому обчислення їх результатів може відбуватись паралельно, при цьому вони функціонують як "чорна скринька" і можуть бути швидко замінені іншими, більш точними мережами або взагалі іншими засобами класифікації із однаковим форматом результату.

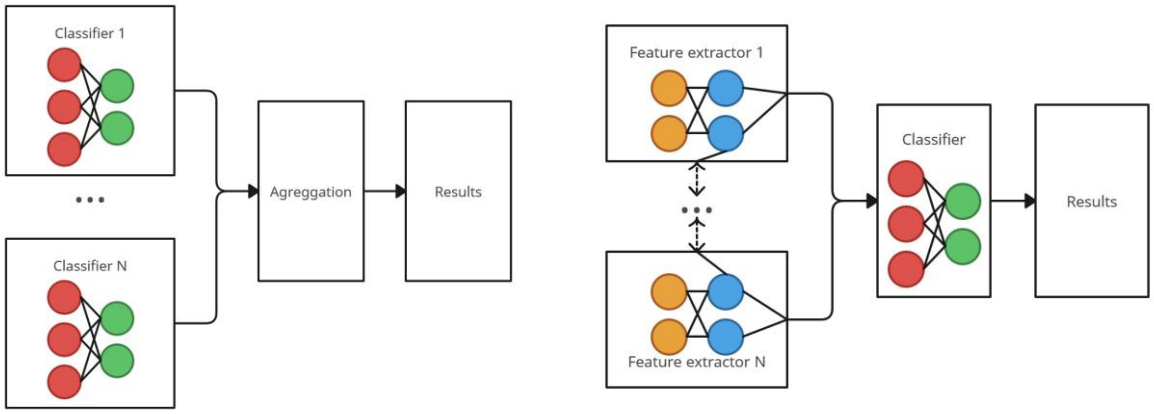
У статті [4] пропонується використання генеративної рекурентної мережі, яка використовує ознаки зображень, виділені згортковою мережею. Сигнали рекурентної мережі та виділені ознаки поєднуються у мультимодальному шарі, забезпечуючи спільну імовірнісну генерацію речень для опису зображень.

Ці та інші нещодавні публікації фокусуються на конкретних задачах, які або не потребують подальшого розвитку, наприклад у статті [3] розв'язується задача діагностики раку, але не приділено увагу інтеграції отриманих рішень у програмні системи. Отже, дослідження цього напрямку обробки мультимодальних даних є актуальними.

Побудова системи класифікації мультимодальних даних в реальному часі. Поставлена задача в загальному вигляді передбачає обробку даних декількох модальностей. Залежно від того, наскільки вони пов'язані між собою, результати класифікації можуть або доповнювати стан об'єкта, у випадку, коли модальності слабко пов'язані (рис. 1а), або уточнювати його завдяки злиттю результатів (рис. 2б).

При використанні штучних нейронних мереж як методу класифікації, реалізація такого поєднання передбачає обчислення результату мереж для кожної модальності з подальшою обробкою результатів допоміжним блоком. Це може бути як просте зважене середнє результатів, так і ще одна мережа, навчена класифікувати об'єктами за результатами обробки мереж, що відповідають за окремі модальності. Дослідження показують, що другий спосіб побудови класифікатора є більш точним. Він використовує мережі для виділення ознак із точок мультимодальних даних, в той час як класифікатор приймає рішення, спираючись на поєднаний вектор ознак. При цьому мережі, як правило, інтегровані і окремі мережі попередньо навчаються на відповідних окремих модальностях, після чого вони і класифікатор донавчаються на повному наборі.

Втім, для більш ефективної з точки зору часу виконання реалізації, окремі мережі доцільно тримати ізольованими з ряду причин. По-перше, це робить програмну систему, яка реалізує класифікатор, модульною, що дозволяє розподілити обчислення паралельно, а також дозволяє частково оновлювати компоненти без необхідності повного повторного навчання та розгортання системи. Навіть підходи з використанням інтегрованих мультимодальних мереж можуть бути розділені на окремі мережі, за умови відсутності великої кількості наскрізних зв'язків (shortcuts). В такому випадку, деякі внутрішні шари стають частиною вхідного.



(а) (б)

Рисунок 1 - а – незалежні класифікатори слабко пов'язаних модальностей. б – спільні класифікатори пов'язаних модальностей

Розглянемо механізм організації обчислень, який керує роботою мереж. Для ефективного паралельного виконання необхідний спосіб регулювання навантаження. Для цього використовується паралельна черга за шаблоном producer-consumer. Процес отримання даних з певною частотою отримує нову точку даних, виконує їх попередню обробку для приведення у формат, придатний для подальшої обробки та поміщає їх у чергу. За наявності вільного процесу-обробника, він зчитує дані з черги та розподіляє їх по процесам, які виконують обчислення, визначені відповідними нейронними мережами. Отримані результати синхронізується та агрегуються. Як зазначалось раніше, агрегація може передбачати алгоритм класифікації за отриманими результатами або просте їх поєднання. Результат обробки передається далі як виведення, наприклад, для його відображення користувачу. При цьому, сигнал про завершення обробки передається на процес отримання даних (рис. 2).

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

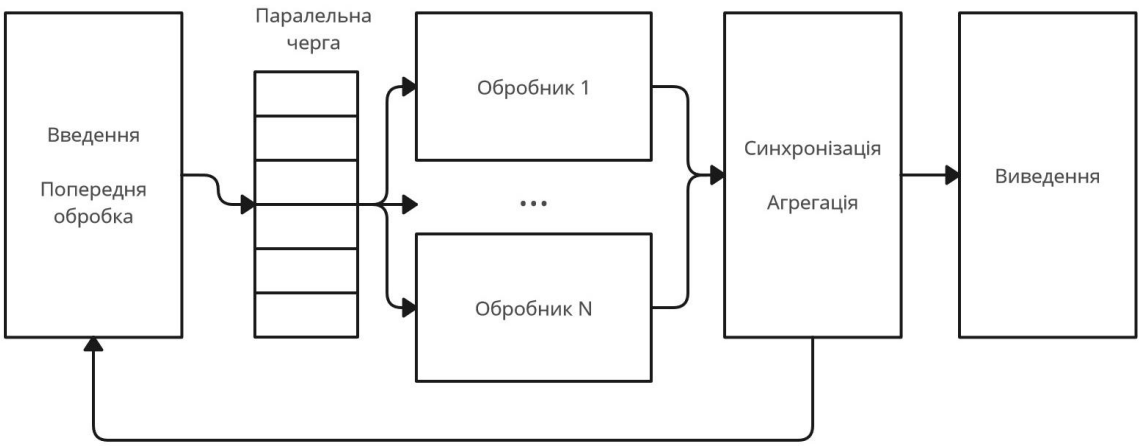


Рисунок 2 - Схема системи паралельної обробки мультимодальних даних

Частота отримання точки даних оновлюється таким чином, щоб максимізувати пропускну здатність системи. При цьому, якщо процеси-обробники зайняті довше, то черга виступає буфером, який забезпечує, хоч і з затримкою, обробку отриманих даних, але частота їх зчитування в подальшому зменшиться, допоки черга не буде опрацьована. Якщо ж обробники очікують на надходження даних довше, ніж частота зчитування, то вона збільшиться, дозволяючи отримати більшу швидкодію системи.

Створення експериментальної програмної системи. Для підтвердження такої організації системи було проведено експеримент з класифікації відео з використанням кадрів для визначення об’єкта та звуку для уточнення його стану. Для цього було використано відкриті набори даних VGG Sound [5] та Dogs vs. Cats [6].

Розглянемо основні етапи створення експериментальної програмної системи.

Отримання та підготовка даних. Основний набір даних для валідації і тренування - VGG Sound. Формат записів включає ідентифікатор відео на Youtube, час початку фрагменту, мітку класу, яка містить назву об’єкта та відповідний звук та мітку розподілу: навчальні або тестові дані (табл. 2).

Для отримання безпосередніх даних, було створено утиліту для завантаження відео, виділення фрагменту околом 2 секунди навколо зазначеного в наборі часу та збереження елементів цього фрагменту:

- кадрів відео з інтервалом в 1 секунду у форматі JPG;
- звукової доріжки у форматі WAV.

Таблиця 2

Формат даних набору VGG Sound

YouTubeID	StartSeconds	Label	Split
---g-f_I2yQ	1	people marching	train
--0PQM4-hqg	30	waterfall burbling	train
--5OkAjCI7g	40	people belly laughing	train

Збереження проміжних кадрів із відео призводить до низької якості більшості зображень, тому набір даних для візуального розпізнавання було доповнено зображеннями з набору Dogs vs. Cats до загальної кількості у 650 зображень. Звук із відео було кодовано у моно-каналі з частотою 44,1 кГц. Для навчання було відібрано якісні, менш зашумлені фрагменти, які належать до трьох класів, по 250 файлів на клас.

При навчанні нейронних мереж отримані набори даних було випадково розділено на навчальні (85%) та тестові (15%) дані.

Навчання мережі розпізнавання об'єктів. У задачі розпізнавання образів себе добре зарекомендували згорткові нейронні мережі. Дослідження фокусуються на двох типах архітектури таких мереж: R-CNN та YOLO. R-CNN та різноманітні модифікації в цьому класі вважаються більш точними, але є більш обчислювально витратними і, відповідно, потребують більше часу для навчання та класифікації. Натомість архітектури типу YOLO мають меншу точність, особливо для невеликих об'єктів на зображенні, але працюють швидше і є придатними для розпізнавання у режимі реального часу, тому для даної задачі було обрано мережу YOLOv3 із розмірністю вхідного шару 416x416x3 [7].

Нейронні мережі створені за допомогою відкритої бібліотеки машинного навчання Tensorflow v2. Під час навчання використовуються поширені засоби покращення якості навчання глибоких нейронних мереж: використання пакетного навчання, ініціалізація початкових ваг мережі невеликим випадковими числами, використання шарів пакетної нормалізації, проходження декількох розігрівальних (warm-up) епох, після яких використовується поступово спадаючий до заданого ліміту темп навчання (рис. 3), проведення декількох спроб навчання з різними початковими вагами та поділом даних на навчальні та тестові для уникнення перенавчання [8, 9].

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

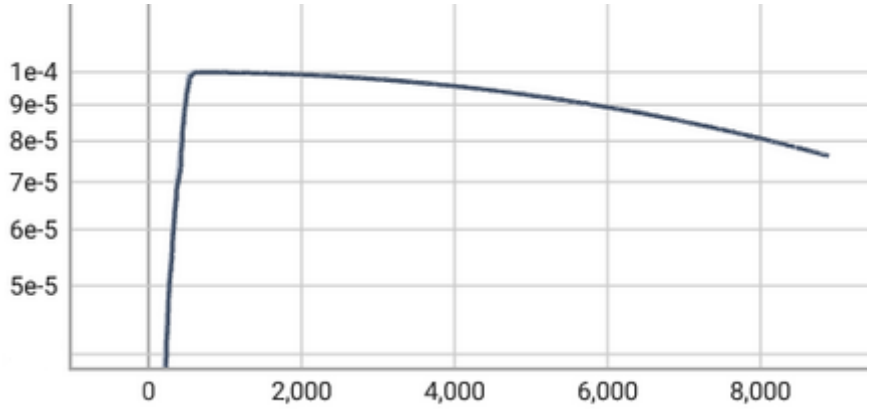


Рисунок 3 - Темп навчання під час тренування мережі в залежності від кроку

Для оцінки якості навчання використовується моніторинг функції втрат, яка складається з втрат координат (використано метрику GIoU), втрат впевненості та втрат класифікації. Значення втрат найкращих мереж з трьох спроб навчання наведено у табл. 3.

Таблиця 3

Значення функцій втрат декількох спроб навчання

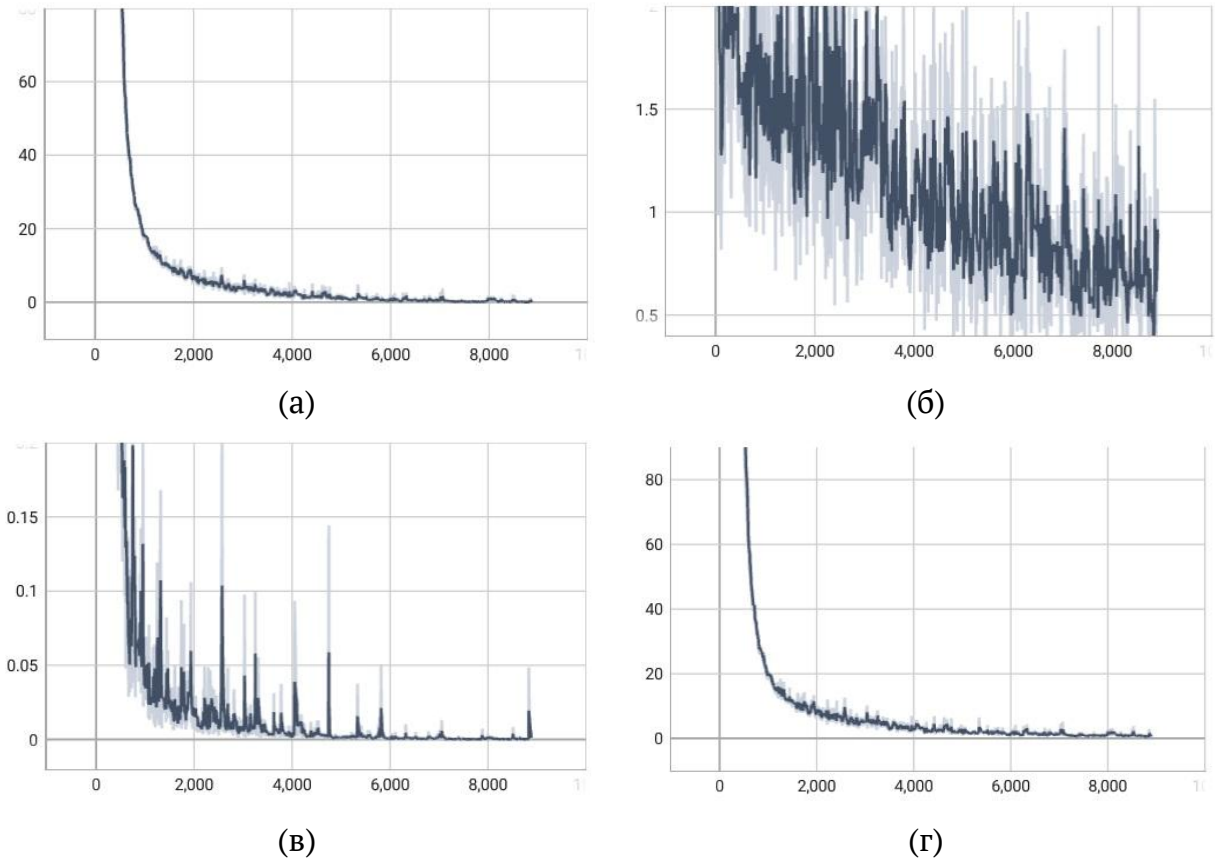
№	Втрати координат	Втрати впевненості	Втрати класифікації	Загальні втрати
1	0.0464	0.5032	0.0186	0.5682
2	**0.0413**	**0.3748**	**0.0117**	**0.4278**
3	0.0670	0.7037	0.0102	0.7809

На рис. 4 приведено динаміку втрат під час навчання використаної мережі.

Навчання мережі класифікації звуку. Для обробки звукового сигналу, виділені фрагменти були доповнені до однакової довжини в 4 секунди та перетворені в частотний спектр за допомогою віконного перетворення Фур'є із частковим перекриттям вікон. Для класифікації спектрограм використовується нейронна мережа з двома згортковими шарами для виділення ознак та повнозв'язним шаром для класифікації. Для кращої чисельної стабільності навчання, значення частот на вході нормалізуються. Також, для уникнення перенавчання використовується випадкове випадання ваг після виділення ознак та класифікації.

Метрики навчання мереж: значення функції втрат (розріджена категоріальна перехресна ентропія) та точність класифікації. Навчання проходить у 20 епох, але передчасно припиняється, якщо функція втрат не зменшується про-

тягом трьох епох поспіль. Значення метрик для трьох спроб навчання наведено в табл. 4.



(а) (б)

(в) (г)

Рисунок 4 - Складові функції втрат під час тренування мережі в залежності від кроку. а – впевненість. б – координати. в – класифікація. г – загальні втрати

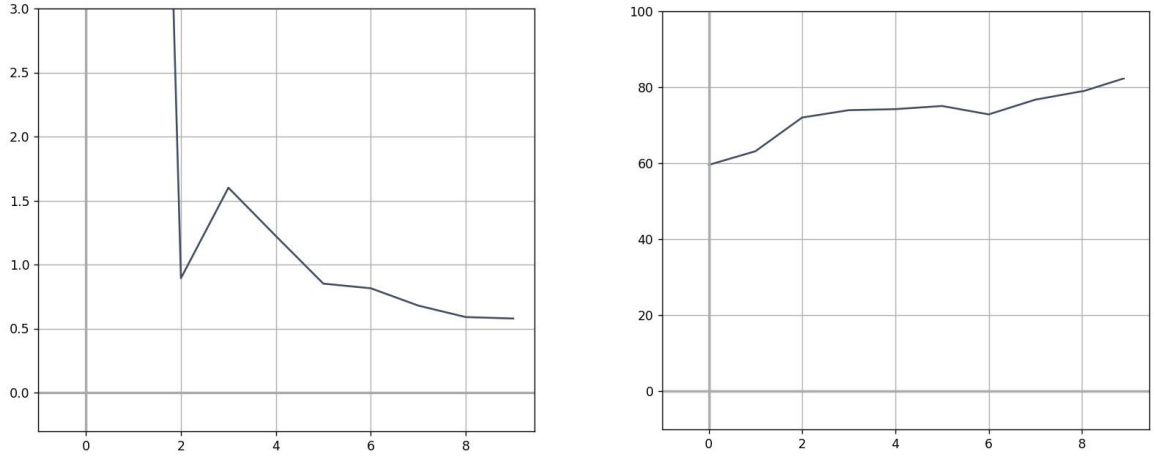
Таблиця 4

Значення метрик декількох спроб навчання

№	Фактичні епохи	Втрати	Точність (%)
1	7	0.75	72
2	15	0.68	78
3	**11**	**0.62**	**83**

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

На рис. 5 наведено поведінку метрик під час навчання.



(а) (б)

Рисунок 5 - Значення метрик під час тренування мережі в залежності від епохи. а – функція втрат. б – точність

Створення програмного модуля мультимодальної класифікації. Створений програмний модуль, який використовує підготовані мережі для класифікації відео в реальному часі, майже повністю відповідає схемі, наведеній на рис. 2. Блок введення отримує дані із відео, виконує необхідну попередню обробку для приведення їх у формат, придатний до обробки з певною періодичністю:

– для зображень: отримання кадру з відео, перетворення кольорової схеми, пропорційне масштабування карду, щоб найбільша розмірність складала 416 пікселів та заповнення іншої частини зображення чорним;

– для звуку: виділення чотирьохсекундного фрагменту, виконання відповідного кодування та отримання частотного спектру.

Результати попередньої обробки поміщаються в чергу, звідки вони зчитуються і паралельно обробляються мережами для класифікації, результати якої синхронізуються та подаються як виведення модуля. Виведення використовується для відображення результатів, в даній реалізації окрема допоміжна програма накопичує результати класифікації та відповідні часові мітки і монтує копію початкового відео з відображенням результатів (рис. 6).

Періодичність обробки стабілізується в середньому на 0.2 секунди, дозволяючи опрацювання 5 кадрів в секунду, що близько до режиму реального часу. Час оброблення можна суттєво зменшити із застосуванням обчислень на графічному процесорі.

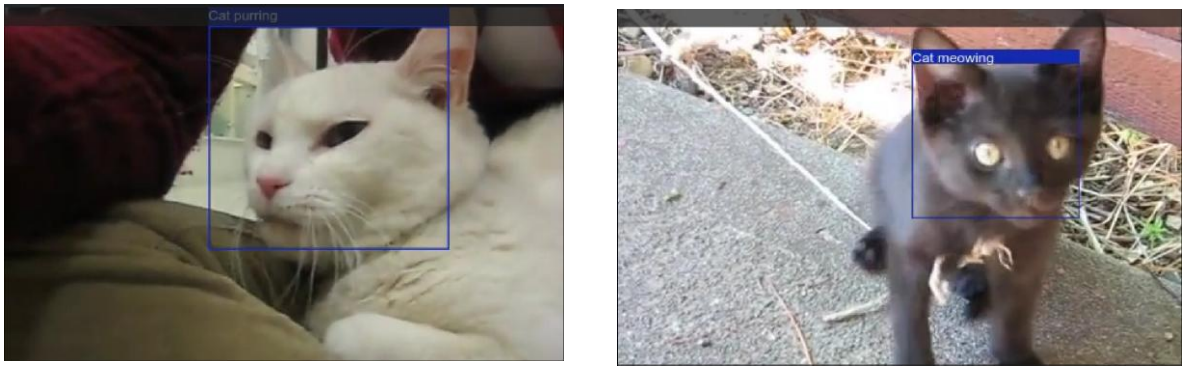


Рисунок 6 - Приклади результатів класифікації

Висновки і подальша робота. У статті запропоновано метод створення програмного модуля для виконання класифікації мультимодальних даних у режимі реального часу або близькому до нього.

Також було розглянуто створення прикладу програмного модуля із використанням штучних нейронних мереж для класифікації мультимодальних даних. Створений модуль здатен виконувати класифікацію 5 кадрів з відео в секунду з прийнятною точністю. Його ефективність може бути значно підвищена за рахунок використання спеціалізованого апаратного забезпечення (напр. графічного процесора) або використання більшого ступеня паралельності, за якого декілька процесів можуть виконувати задачі класифікації.

Запропонований підхід є модульним і дозволяє інкрементне покращення окремих складових з мінімальних впливом на всю систему. Також запропонований підхід до розподілення навантаження дозволяє зробити систему масштабованою, додаючи обчислювальної потужності. Подальша робота буде направлена на дослідження цієї властивості, в тому числі і з використанням хмарних сервісів, які надають можливість використовувати майже довільну кількість віртуалізованого апаратного забезпечення і підтримують паралельні системи потоків даних.

ЛІТЕРАТУРА

1. Liang P. P., Zadeh A., Morency L. P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions //arXiv preprint arXiv:2209.03430. – 2022.

2. Ngiam J. et al. Multimodal deep learning //Proceedings of the 28th international conference on machine learning (ICML-11). – 2011. – С. 689-696.

3. Sun D., Wang M., Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data //IEEE/ACM

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

transactions on computational biology and bioinformatics. – 2018. – Т. 16. – №. 3. – С. 841-850.
4. Mao J. et al. Explain images with multimodal recurrent neural networks //arXiv preprint arXiv:1410.1090. – 2014.
5. Chen H. et al. Vggsound: A large-scale audio-visual dataset //ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2020. – С. 721-725.
6. Dogs vs. Cats Dataset. URL: https://www.kaggle.com/c/dogs-vs-cats
7. Redmon J., Farhadi A. Yolov3: An incremental improvement //arXiv preprint arXiv:1804.02767. – 2018.
8. Bjorck N. et al. Understanding batch normalization //Advances in neural information processing systems. – 2018. – Т. 31.
9. You Y., Gitman I., Ginsburg B. Large batch training of convolutional networks //arXiv preprint arXiv:1708.03888. – 2017.

REFERENCES

1. Liang P. P., Zadeh A., Morency L. P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions //arXiv preprint arXiv:2209.03430. – 2022.
2. Ngiam J. et al. Multimodal deep learning //Proceedings of the 28th international conference on machine learning (ICML-11). – 2011. – P. 689-696.
3. Sun D., Wang M., Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data //IEEE/ACM transactions on computational biology and bioinformatics. – 2018. – Т. 16. – №. 3. – P. 841-850.
4. Mao J. et al. Explain images with multimodal recurrent neural networks //arXiv preprint arXiv:1410.1090. – 2014.
5. Chen H. et al. Vggsound: A large-scale audio-visual dataset //ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2020. – P. 721-725.
6. Dogs vs. Cats Dataset. URL: https://www.kaggle.com/c/dogs-vs-cats
7. Redmon J., Farhadi A. Yolov3: An incremental improvement //arXiv preprint arXiv:1804.02767. – 2018.
8. Bjorck N. et al. Understanding batch normalization //Advances in neural information processing systems. – 2018. – Т. 31.
You Y., Gitman I., Ginsburg B. Large batch training of convolutional networks //arXiv preprint arXiv:1708.03888. – 2017.

Received 04.12.2023.
Accepted 08.12.2023.

Using deep artificial neural networks for multimodal data classification

Multimodal data analysis is gaining attention in recent research. Pu Liang et al. (2023) provide a comprehensive overview on multimodal machine learning, highlighting its foundations, challenges and achievements in recent years. More problem-oriented works propose new methods and applications for multimodal ML, such a Ngiam et al. (2011) propose to use joint audio and video data to improve speech recognition accuracy; Sun, Wand and Li (2018) describe application of multimodal classification for breast cancer prognosis prediction; Mao et al. (2014) propose an architecture of multimodal recurrent network to generate text description of images and so on. However, such works usually focus on the task itself and methods therein, and not on integrating multimodal data processing into other software systems.

The goal of this research is to propose a way to conduct multimodal data processing, specifically as a part of a digital twin systems, thus efficiency and near-real-time operation are required.

The paper presents an approach to conduct parallel multimodal data classification, adapting to available computing power. The method is modular and scalable and intended for in digital twin application as a part of analysis and modeling tools.

Later, the detailed example of such a software module is discussed. It uses multimodal data from open datasets to detect and classify the behavior of pets using deep learning models. Videos are processed using two artificial neural networks: YOLOv3 object detection network to process individual frames of the video and a relatively simple convolutional network to classify sounds based on their frequency spectra.

Constructed module uses a producer-consumer parallel processing pattern and allows processing 5 frames per second of a video on available hardware, which can be sufficiently improved by using GPU acceleration or more paralleled processing threads.

Пеня Олександр Романович – аспірант кафедри програмного забезпечення комп'ютерних систем Київського політехнічного інституту ім. Ігоря Сікорського.

Сулема Євгенія Станіславівна – д.т.н., доцент, завідувач кафедри програмного забезпечення комп'ютерних систем Київського політехнічного інституту ім. Ігоря Сікорського.

Oleksandr Penia – Post-Graduate Student of Computer Systems Software Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

Yevgeniya Sulema – DSc, Associate Professor, Head of the Computer Systems Software Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)