ISSN 2707-1898 (print)
Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2023.02.033
Article received 20.10.2023 p.
Article accepted 26.10.2023 p.
UDK 004.93
Correspondence author
V. Y. Chornenkyi
volodymyr.y.chornenkyi@lpnu.ua

УДК 004.93

В. Я. Чорненький, І. Я. Казимира

ДОСЛІДЖЕННЯ МОДЕЛЕЙ ДЛЯ РОЗПІЗНАВАННЯ ЖЕСТІВ З ВИКОРИСТАННЯМ 3D КОНВОЛЮЦІЙНИХ НЕЙРОННИХ МЕРЕЖ ТА ВІЗУАЛЬНИХ ТРАНСФОРМЕРІВ

У роботі розглянуто актуальне завдання розпізнавання жестів з метою реформування способів навчання військових, комунікації людини та машини і вдосконалення взаємодії людина-людина та людина-машина для осіб з обмеженими можливостями. Проаналізовано методи для розпізнавання жестів руки на основі комп'ютерного зору, а також із використанням глибокого навчання.

Описано принципи роботи моделей із використанням 3D конволюційних нейронних мереж та трансформерів, наведено їх структурні схеми та проаналізовано особливості функціонування складових. У межах 3D-CNN архітектури розглянуто конволюційну нейронну мережу з двома конволюційними шарами та двома шарами групування. Кожну 3D згортку отримано у результаті згортки ядра 3D-фільтра і складання декількох суміжних кадрів разом для одержання 3D-куба. У межах ViT архітектури розглянуто візуальний трансформер з Linear Projection, Transformer Encoder, двома підшарами: шар Multi-head SelfAttention (MSA) та шаром прямого поширення, також відомим як Multi-Layer Perceptron (MLP).

На підставі досліджених архітектур здійснено навчання моделей з використанням ASL та NUS-II наборів даних та розглянуто їх ефективність після 20 навчальних епох на основі показників відтворення, точності та F1-оцінки. Визначено вплив тривалості навчання на ефективність моделі з використанням ViT архітектури після 20 та 40 навчальних епох.

Продемонстровано, в яких ситуаціях 3D конволюційні нейронні мережі та візуальні трансформери дають кращі результати точності, та обмеження, притаманні кожному підходу в умовах варіативності середовища та обчислювальних потужностей.

Розвинено інноваційні архітектури для розпізнавання жестів руки з використанням глибокого навчання для майбутніх досліджень та реалізацій у програмних продуктах.

Ключові слова: глибоке навчання; взаємодія людини та машини; ефективність нейронних мереж; набори даних для мови жестів.

Вступ / Introduction

Проблематика розпізнавання жестів займає центральне місце в академічних дослідженнях у сфері комп'ютерного зору та глибокого навчання. Передбачуваний розвиток технологій віртуальної (VR) та доповненої реальності (AR), де жести стають важливим засобом взаємодії, збільшує актуальність та необхідність подальших досліджень цього напряму. Системи розпізнавання жестів мають потенціал насамперед реформувати навчання військових, способи комунікації людини та машини, особливо – способи комунікації для осіб з обмеженими можливостями.

Систематичний огляд наукової літератури підтверджує існування численних методів та підходів, пропонованих для вирішення цієї актуальної проблеми. Від традиційних методів оброблення зображень до сучасних моделей глибокого навчання, багато пропозицій вже забезпечили вагомий внесок у підвищення точності та ефективності розпізнавання жестів. Однак ще є виклики, пов'язані з роботою в реальному часі, адаптацією до змінних умов та інтеграцією різних сенсорів.

У межах цього дослідження увагу зосереджено на інноваційних підходах до розпізнавання жестів, зокрема на аналізі 3D конволюційних нейронних мереж та трансформерів. Досліджуючи перспективи цих технологій та їх комбіноване використання із сучасними сенсорами, спробуємо зробити свій внесок у створення надійних, ефективних та адаптивних систем розпізнавання жестів.

Об'єкт дослідження – процес розпізнавання жестів руки на основі нейронних мереж.

Предмет дослідження – моделі та архітектури 3D конволюційних нейронних мереж та візуальних трансформерів для розпізнавання жестів руки.

Мета роботи – дослідження та порівняння ефективності застосування 3D конволюційних нейронних мереж та візуальних трансформерів у завданнях розпізнавання жестів руки.

Для досягнення поставленої мети визначено такі основні *завдання дослідження*:

- проаналізувати сучасні підходи до розпізнавання жестів з використанням 3D конволюційних нейронних мереж та візуальних трансформерів;
- оцінити точність розпізнавання жестів руки для 3D-CNN та ViT архітектур;
- дослідити ефективність моделей із використанням кожної з архітектур на ASL [16] та NUS-II [17] наборах даних.

Аналіз останніх досліджень та публікацій. На ранніх етапах розвитку комп'ютерного зору для розпізнавання жестів використовували класичні методи. Ці методи ґрунтувалися на ручному визначенні особливостей, таких як кольорові гістограми, контури та текстурні характеристики. Такі методи, як Canny Edge Detection, Hough Transform та Histogram of Oriented Gradients (HOG), SIFT, були популярними в сфері комп□ютерного зору для виявлення та розпізнавання об'єктів [3], [5], [7]. Ці підходи були доволі ефективними для обмеженого набору завдань, але не завжди адаптувалися до різноманітних змінних умов, таких як відхилення в освітленні, зміни в перспективі чи затіненні об'єктів. До того ж ручне виокремлення характеристик було часозатратним і не завжди забезпечувало адекватну адаптацію до нових, неочікуваних обставин.

З упровадженням глибокого навчання у сферу розпізнавання жестів суттєво змінилась парадигма. Останні дослідження, які використовують глибокі конволюційні нейронні мережі (CNN) з відеопослідовностями, істотно покращили точність розпізнавання динамічних жестів руки [1], [2], [7] та дій [3], [5], [6]. CNN також корисні для комбінування мультимодальних даних [2], [7], техніка, яка виявилася корисною для розпізнавання жестів у складних умовах освітлення [2], [4]. Однак системи розпізнавання динамічних ручних жестів у реальному часі створюють численні нерозв'язані виклики. Наприклад, ці системи отримують безперервні потоки необроблених візуальних даних, у яких жести з відомих класів повинні бути одночасно виявлені та класифіковані. Більшість попередніх досліджень, наприклад, [2], [7], [4], розглядають сегментацію жестів та класифікацію окремо. Два класифікатори, один, що визначає, чи зроблено жест, і інший, який характеризує тип жесту, навчали окремо, що призводить до лімітації точності системи в потоці даних.

Одним із інноваційних підходів у цій сфері стала архітектура CNN-SPP [8], яка використовувала Spatial Pyramid Pooling для захоплення ширшої просторової інформації. Це дало змогу мережі краще адаптуватися до різних розмірів і форм об'єктів на зображенні. Інший підхід, оснований на архітектурі DenseNet, модифіковано у вигляді EDenseNet [9], що забезпечувало краще узагальнення і розпізнавання об'єктів. У недавньому дослідженні [10] запропоновано метод, який використовує звичайну RGB-камеру для визначення 21 ключової точки на руці. Для цього була розроблена мережа, яка навчалася ідентифікувати ці ключові точки. В основу цієї мережі покладено архітектуру PointNet, оптимізовану для ефективної роботи на процесорах (CPU).

Трансформери, спочатку розроблені для машинного перекладу, згодом визнали революційною технологією у сфері оброблення природної мови (NLP) [11], [12]. Їх унікальна здатність враховувати великі часові контексти робить їх особливо ефективними для аналізу структурної та відносної інформації в мові.

З урахуванням цього успіху, були зроблені численні спроби адаптувати трансформери для задач комп'ютерного зору [13], [14]. Особливу увагу привернула модель Vision Transformers (ViT) [16]. На відміну від традиційних моделей комп'ютерного зору, які використовують конволюції, ViT повністю основана на архітектурі трансформерів. Це не тільки зближує підходи в NLP та комп'ютерному зорі, але й демонструє обнадійливі результати, особливо під час роботи з великими наборами даних.

Результати досліджень та їх обговорення / Research results and their discussion

Архітектура 3D-CNN. Типова концепція конволюційних нейронних мереж передбачає, що кількість карт характеристик збільшується з кожним новим шаром, що дає змогу видобувати багатогранні характеристики на основі попередніх найпростіших. 3D конволюція полягає у конволюції за допомогою 3D фільтра, об'єднуючи декілька послідовних кадрів для формування 3D-блока. Така операція дає змогу зв'язувати картки характеристик із послідовними кадрами. Якщо говорити формально, значення у позиції (x, y, z) на j-й карті характеристик у i-му шарі обчислюється за формулою (1):

$$v_{i,j}^{x,y,z} = \tanh\left(b_{i,j} + \sum_{m} \sum_{a=0}^{A_i-1} \sum_{b=0}^{B_i-1} \sum_{c=0}^{C_i-1} w_{ijm}^{abc} v_{(i-1)m}^{(x+a)(y+b)(z+c)} \right), \quad (1)$$

де C_i – розмір 3D фільтра ядра вздовж часового виміру; w_{ijm}^{abc} – значення (a, b, c) карти характеристик, пов'язані з m-м значенням ядра на попередньому шарі.

На рис. 1 подано архітектуру 3D конволюційної нейронної мережі.

Обчислення просторового розміру вихідного об'єму 3D CNN здійснюється за допомогою таких гіперпараметрів: рецептивне поле, нульове доповнення, довжина кроку та об'єм. Щоб визначити кількість нейронів у шарі згортки, використовують формулу:

$$\left(\frac{W - F + 2.P}{S} \right) + 1, \quad (2)$$

де W – ширина 3D фільтра; P – нульове доповнення; S – довжина кроку; F – глибина 3D фільтра.

За формулою (2), вхідний шар з розмірністю $\left((120-11+2.0)/1\right)+1$ відповідає вихідному об'єму $110\times110\times32$, де W, $H = 120$ вказують на висоту та ширину вхідної рамки; $F = 11\times11\times32$ представляє глибину 3D фільтра; $P = 0$ –нульове доповнення, а $S = 1$ – крок згортки.

Для досягнення успіху в налаштуванні моделі 3D CNN ми використали куб з 11 кадрами (інформація про рух) як вхідні дані [15]. На вхід моделі на основі архітектури 3D CNN (рис. 1) надходять дані розміром 120×120×11 із кількістю карт характеристик та розмірами ядер у кожному шарі. У 3D CNN два шари згортки використовують фільтри розміром 11×11×32 та 5×5×32 відповідно.

Два шари групування з ядрами розміром 2×2 додаються після конволюційних шарів, як видно зі структури CNN. Шари групування є одним із ключових компонентів конволюційних нейронних мереж. Якщо конволюційні шари відповідають за видобування характеристик зображень, то шари групування агрегують ці характеристики. Їх основна мета – поступово зменшувати просторовий розмір подання, щоб знизити кількість параметрів та обчислень у мережі, а цим самим і обчислювальні витрати, та запобігати перенавчанню.

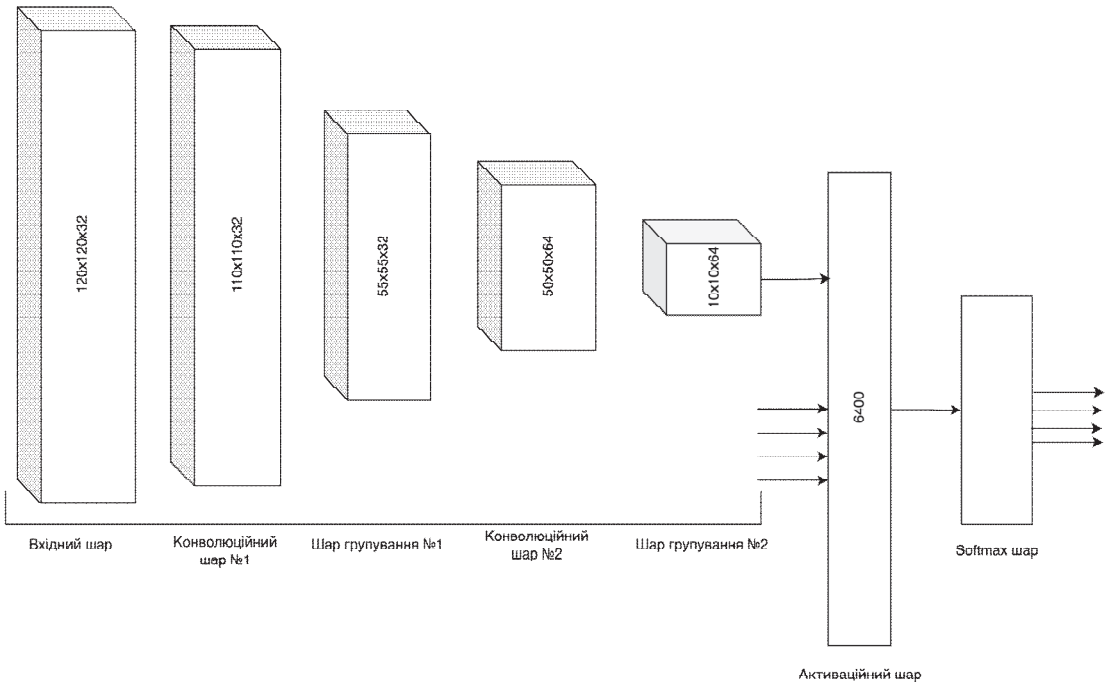

Рис. 1. Архітектура 3D CNN / 3D CNN architecture

Насамкінець, повністю під'єднаний шар містить всі активаційні величини попереднього шару, які конвертуються у векторні характеристики з 6400 компонентами. Softmax-шар містить вихідні елементи, що відображають класи дій.

Архітектура HGR-ViT. HGR-ViT є представником моделей Vision Transformer, призначеним для розпізнавання жестів. На початковому етапі вхідні зображення масштабуються до однакового розміру та нормалізуються. Потімці зображення розділяють на окремі ділянки, що розглядають як ряд піксельних значень. Щоб перетворити ці послідовності на простір з меншою розмірністю, застосовуємо шар лінійної проєкції, який можна навчати. Кожен фрагмент зображення отримує додаткове позиційне кодування для зберігання просторового контексту. Потім ці ділянки обробляються за допомогою кодувальників трансформера, що дає змогу моделі аналізувати взаємодії між різними ділянками. Завершальний етап передбачає передавання даних через лінійний проєкційний шар і функцію активації Softmax для визначення ймовірностей належності до певних класів. Усе навчання моделі відбувається на основі контрольованих даних та використання функції втрат. Рис. 2 демонструє архітектуру HGR-ViT, ураховуючи етапи оброблення ділянок, позиційного кодування, кодувальників Transformer та фінальної класифікації.

Щоб адаптувати зображення ручного жесту до розділення на сегменти розміром 32×32, його спочатку зводять до розмірів 256×256. Застосування зображень великої роздільної здатності за однакового розміру сегмента збільшує ефективну довжину послідовності, що сприяє зростанню продуктивності. Після масштабування зображення жесту розділяють на сегменти стандартного розміру. Зображення із висотою H, шириною W та C каналами перетворюється на послідовність 2D сегментів, поданих як $x_p \in R^{N*(p^2*C)}$, де $N = \frac{HW}{p^2}$ вказує на кількість сегментів і ефективну довжину послідовності для трансформера, де (P, P) визначає роздільну здатність кожного сегмента зображення.

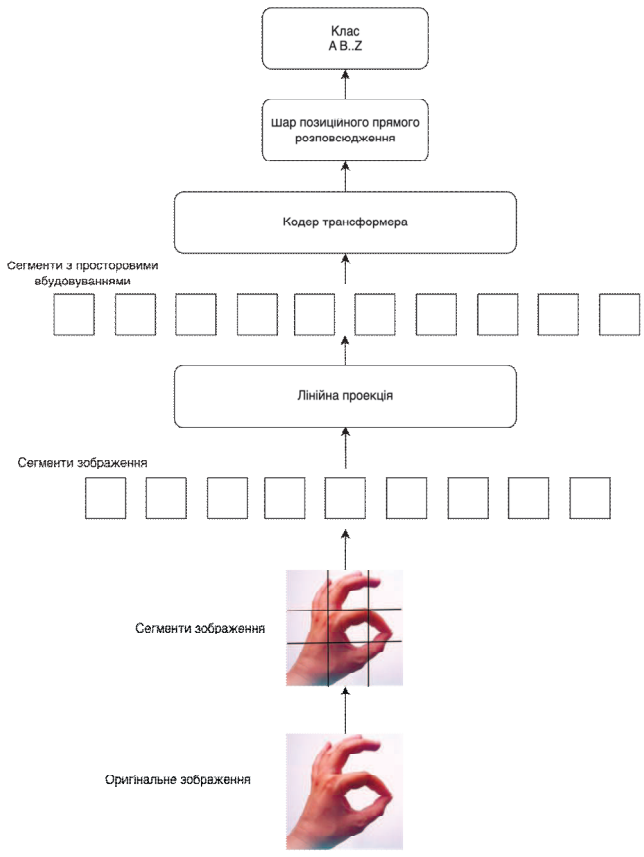

Рис. 2. Архітектура HGR-VIT / HGR-ViT architecture

Перед передаванням сегментів до блоків кодера трансформера здійснюється процес їх лінійної проєкції. Спочатку кожний сегмент перетворюється на вектор x_{n_p} довжиною $P^2 \times C$, де $n = \overline{1, N}$. Потім навчена матриця вбудовувань E відображає згладжені сегменти в D вимірах, створюючи послідовність вбудованих сегментів зображення. Щоб ввести просторову інформацію і полегшити навчання, вбудовування сегментів доповнюються одновимірними просторовими вбудовуваннями E_{pos}, які навчаються під час процесу тренування. Щоб подати вихідний результат класифікації y, навчене вбудовування класу x_{class}, аналогічно до маркера класу в Bidirectional Encoder Representations from Transformers (BERT), додається на початок послідовності вбудованих сегментів зображення. Вихід процесу лінійної проєкції z_0 подамо таким рівнянням:

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; ...; x_p^n E \right] + E_{pos}, \quad (3)$$

де $E \in R^{N \times (p^2 \times C) \times D}$ – навчена матриця вбудовувань; $E_{pos} \in R^{(N+1) \times D}$ – одновимірне просторове вбудовування.

Вбудовування сегментів слугує вхідними даними для кодера трансформера, даючи змогу ViT ефективно виявляти глобальні закономірності та залежності на зображенні, зберігаючи деяку просторову інформацію за допомогою сегментів.

Кодер трансформера є важливою частиною моделі трансформера. Він складається із набору L шарів, кожен з яких містить два підшари: шар багатоголової самоуваги (Multi-Head Self-Attention або MSA) та шар позиційного прямого поширення, відомий також як багатошаровий перцептрон (MLP). Ці підшари розташовані послідовно, а вихід кожного шару є входом для наступного шару, як показано на рис. 3.

На кожному шарі l вхідна послідовність із попереднього шару z_{l-1} нормалізується за допомогою нормалізації шару (LN), яка незалежно нормалізує входи за розмірністю для кожного прикладу. Це підвищує стабільність представлення моделі та загальну продуктивність. Вихід LN передається через шар MSA, а отримана послідовність знову нормалізується за допомогою LN. Наостанок, вихід другого LN проходить через шар MLP, який виробляє набір оновлених вбудовувань сегментів.

До шару MLP додаються залишкові з'єднання, щоб вирішити проблему зниклих градієнтів, даючи моделі можливість вчити залишкові функції. Потік процесів у блоці кодера трансформера можна подати за допомогою рівнянь:

$$zl' = MSA(LN(zl-1)) + z_{l-1}, \quad (4)$$

$$zl = MSA(LN(zl')) + z'_l, \quad (5)$$

де $l = \overline{1, L}$ – індекс шару; z_{l-1} – вхідна послідовність із попереднього шару; MSA – шар багатоголової самоуваги; LN – нормалізаційний шар.

Підсумовуючи, кодер трансформера використовує механізми самоуваги (self-attention) для виявлення глобальних залежностей між вхідними значеннями та багатошаровими перцептронами для оброблення отриманих представлень. Залишкові з'єднання та нормалізація шару забезпечують ефективне навчання та кращу продуктивність.

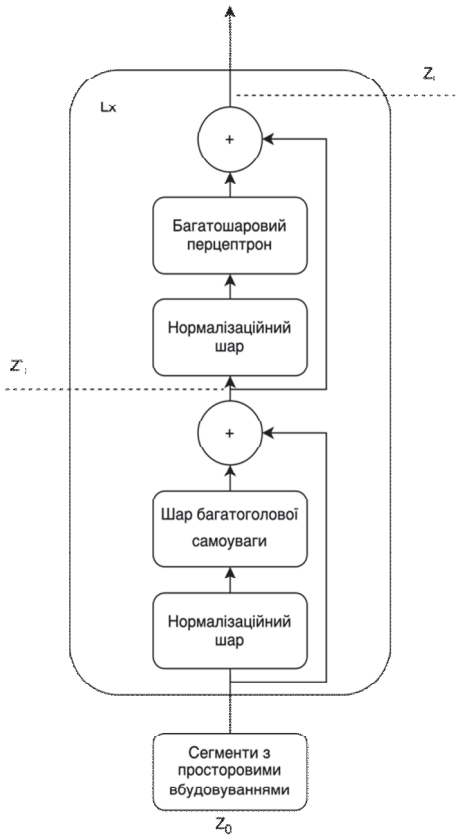

Рис. 3. Архітектура трансформерного кодера / Architecture of the Transformer Encoder

У моделі Vision Transformer (ViT) ключовим елементом є шар Multi-Head Self-Attention (MSA). Він допомагає моделі виявляти контекстні відносини між частинами зображення, розглядаючи взаємодії між вбудовуваннями сегментів. Основний інструмент цього шару – механізм self-attention, який оцінює важливість окремих сегментів на основі їх взаємодій з іншими сегментами на зображенні.

Кожний self-attention блок у MSA містить три повністю з'єднані шари для створення запитів, ключів та значень. Ці компоненти визначають важливість кожного сегмента, залежно від його взаємодій з іншими сегментами. Така структура дає змогу шару MSA ефективно обробляти зображення, зберігаючи при цьому просторову інформацію.

MSA шар обробляє послідовність вбудованих сегментів зображення, перетворюючи їх за допомогою лінійних проєкцій на матриці запитів, ключів та значень.

За MSA йде блок MLP, що дає змогу перетворити вхідні дані на простір вищого рівня. Цей процес передбачає два послідовні лінійні перетворення з активаційною функцією між ними. Наприкінці кодера вибирається перше вхідне значення послідовності й генерується подання зображення.

Після завершення роботи блока кодера він вибирає перше значення з послідовності та створює представлення зображення y за допомогою нормалізації шару, поданого формулою $y = LN(z_0^L)$. Після цього вихід y передається у шар згортки для отримання єдиного вектора характеристик зображення, який потім "випрямляється" та обробляється у шарах нормалізації та SoftMax для класифікації.

Під час навчання використовуються оптимізатор Rectified Adam та функція втрат категоріальної перехресної ентропії (categorical cross-entropy loss function), подана рівнянням:

$$L_{CE} = -\sum_{i=1}^{N} T_i \log(S_i), \quad (6)$$

де S – представляє ймовірності SoftMax, а T – мітки. Для запобігання перенавчанню та для підвищення продуктивності моделі під час навчання також використовують методи ранньої зупинки та адаптивного навчального коефіцієнта.

Використані набори даних. У цій статті використано два тестові набори даних для дослідження пропонованих моделей: набір даних з жестами руки American Sign Language (ASL) з цифрами та набір даних з жестами руки National University of Singapore (NUS).

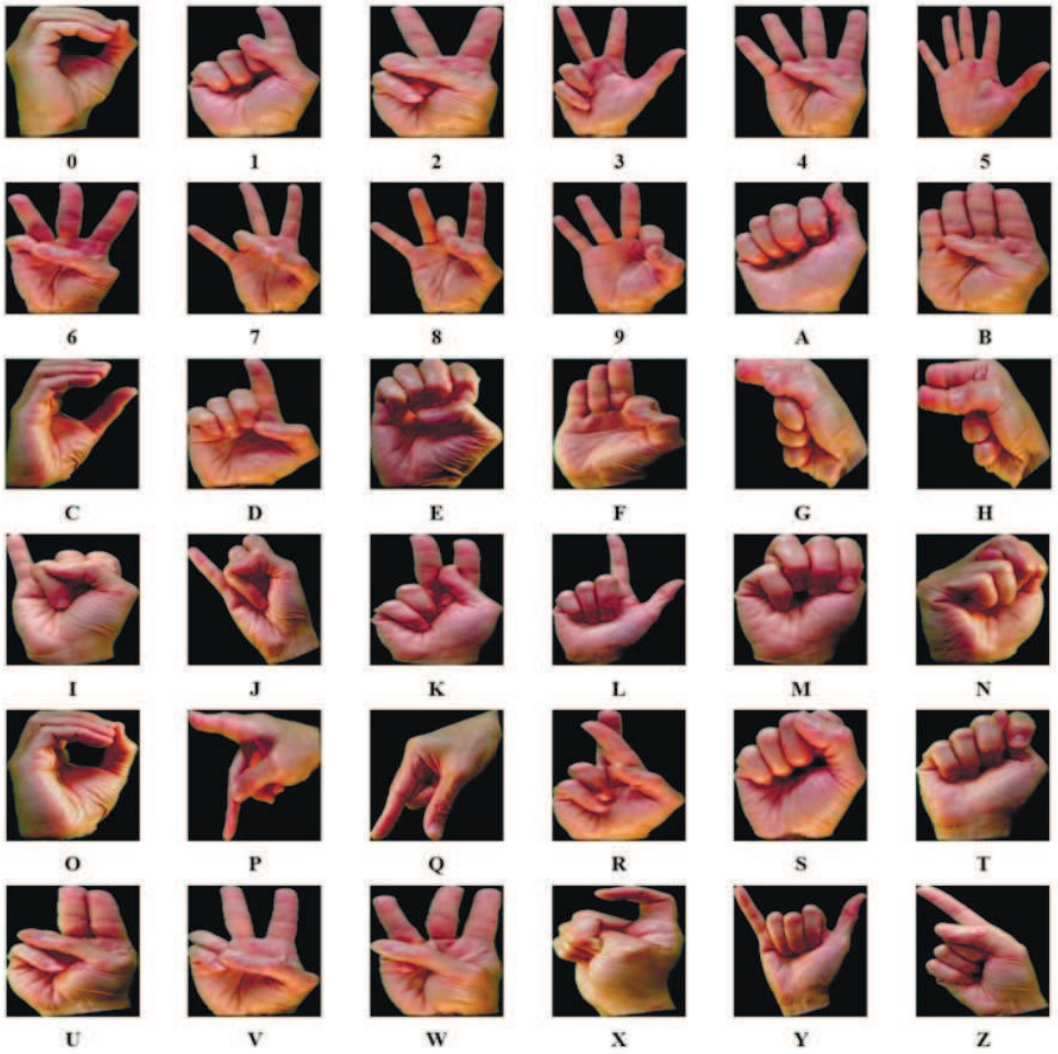

Рис. 4. Зразки зображень для кожного класу з набору даних ASL із цифрами / Image samples for each class from the ASL dataset with numbers

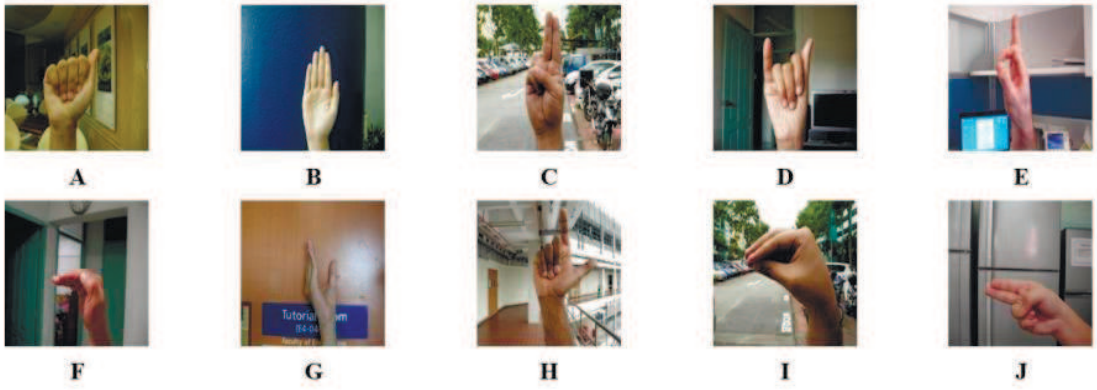

Рис. 5. Зразки зображень для кожного класу з набору даних жестів рук NUT-II / Image samples for each class from the NUT-II dataset

Набір даних ASL з цифрами, поданий у [16], складається із 36 класів жестів руки, ураховуючи літери від A до Z та числа від 0 до 9. Він містить 2515 зразків з варіаціями, які підписали п'ять різних виконавців. Перший і другий виконавці показали по 25 разів кожний символ, за винятком літери T, яка має 20 зразків, тоді як третій і четвертий виконавці показували символи по п'ять разів, і останній виконавець продемонстрував кожен символ десять разів. На рис. 4 показано зразки для кожного класу.

Набір даних із жестами руки NUS-II, запропонований у [17], містить 10 класів жестів руки, ураховуючи літери від A до J, із загальною кількістю 2000 зображень. У наборі даних представлено 40 виконавців, кожен з яких показував жести по п'ять разів для кожного класу, це дало 200 зразків на клас для збільшення різноманітності жестів руки. На рис. 5 наведено зразки зображень для кожного класу в наборі даних.

Результати експериментів із використанням 3D-CNN та ViT. Аналіз виконано на Python 3.11 з використанням OpenCV 3.3 та TensorFlow на macOS Ventura на ноутбуці із процесором Apple M1 Pro та RAM 16 GB. Моделі 3D CNN та ViT оцінено на наборах даних ASL [16] та NUS-II [17].

Внаслідок здійснення 20 навчальних епох відбувалося тренування моделей 3D-CNN та ViT. Важливо зазначити, що 20 ітерацій недостатньо для демонстрації ідеальних результатів, які можна спостерігати в інших дослідженнях, але вони дають контекст, важливий для оцінювання точності та ефективності під час подальшого використання.

Одними із основних метрик вимірювання якості навченої моделі є відтворення (recall), точність (precision) та F1-оцінка. У випадку з розпізнаванням жестів recall – це частка людей, які махають руками в тестовому наборі даних, які правильно ідентифікувала модель і які визначають за формулою (7):

$$recall = \frac{tp}{tp + fn}, \tag{7}$$

де tp – результат тесту, який правильно вказує на наявність стану або характеристики; fn – результат тесту, який помилково вказує на те, що певна умова або ознака відсутня.

Precision – це частка людей, яких модель ідентифікує як тих, що махають руками, і які справді махають руками, що визначається за формулою (8):

$$precision = \frac{tp}{tp + fp}, \tag{8}$$

де tp – результат тесту, який правильно вказує на наявність стану або характеристики; fp – результат тесту, який помилково вказує на наявність певної умови або ознаки.

Оцінку F1 розраховують як середнє гармонійне значення оцінок точності та повноти за формулою (10). Що вищий показник F1, то краще працює модель. Ідеальний показник F1, що дорівнює 1, означає, що модель має ідеальну точність і відтворення.

$$F1score = 2 \times \frac{precision \times recall}{precision + recall}, \tag{9}$$

де *precision* та *recall* – значення, розраховані за формулами (7) та (8).

Детальні результати з порівнянням частини жестів за допомогою 3D-CNN наведено у табл. 1.

3D-CNN показала кращі результати для ASL набору даних, а саме F1-оцінка дорівнює 0,881536, як середній показник для усіх символів. Це можна пояснити ізольованістю символів, на яких була тренована модель, де навколишнє середовище не так впливає на результат. Середню F1-оцінку 0,872366 досягнуто на наборі даних NUS-II.

Табл. 1. Результати експериментів для вибірки символів із наборів даних, архітектура 3D-CNN / Experiment results for set of characters from the used datasets with application of the 3D-CNN architecture

	ASL набір даних з 3D-CNN			NUS набір даних з 3D-CNN		
Символ	Precision	Recall	F1 Score	Precision	Recall	F1 Score
A	0,907859	0,902831	0,905338	0,907859	0,863429	0,885087
B	0,918905	0,914286	0,916589	0,918905	0,874239	0,896015
E	0,875763	0,676577	0,763391	0,875763	0,852618	0,864036
G	0,864380	0,617762	0,720553	0,864380	0,848549	0,856391
I	0,905250	0,925267	0,915149	0,905250	0,853278	0,878496
P	0,915479	0,916582	0,916030	0,915479	0,852178	0,882695
S	0,884388	0,827916	0,855221	0,884388	0,837844	0,860487
Z	0,873488	0,764462	0,815346	0,873488	0,838659	0,855719

Табл. 2. Результати експериментів для вибірки символів з наборів даних, використовуючи архітектуру ViT / Experiment results for set of characters from the used datasets with application of the ViT architecture

Символ		A	B	E	G	I	P	S	Z
ASL набір даних з HGR-ViT	Precision	0,9065	0,8838	0,9005	0,8994	0,8849	0,9059	0,8978	0,8838
	Recall	0,7989	0,8075	0,8073	0,8123	0,7978	0,8084	0,8043	0,8131
	F1 Score	0,8493	0,8439	0,8514	0,8536	0,8391	0,8544	0,8485	0,8470
NUS набір даних з HGR-ViT	Precision	0,8765	0,9184	0,8852	0,9124	0,8905	0,9155	0,8773	0,9054
	Recall	0,7954	0,7835	0,7854	0,8048	0,7884	0,7979	0,7975	0,8080
	F1 Score	0,8340	0,8456	0,8323	0,8552	0,8364	0,8527	0,8355	0,8539
HGR-ViT після 40 епох	Precision	0,9079	0,9189	0,8758	0,8644	0,9052	0,9155	0,8844	0,8735
	Recall	0,9028	0,9143	0,8166	0,7978	0,9253	0,9166	0,8279	0,8545
	F1 Score	0,9053	0,9166	0,8451	0,8297	0,9151	0,9160	0,8552	0,8639

Умови тестування ViT були такими самими і дали гірші результати. Для ASL набору даних показник F1 сягнув 0,848405 та 0,843213 для NUS-II. Цікаво зазначити, що ViT показав кращі результати в умовах варіативного навколишнього середовища. Це зумовлено механізмом самоуваги, що притаманний трасформерам.

Для того, щоб ViT модель показала кращі результати, здійснено додаткові 20 навчальних епох. Внаслідок цього середня F1-оцінка по всіх символах моделі зросла до 0,880884.

Детальні результати з порівнянням частини жестів, із використанням ViT, наведено у табл. 2.

Обговорення результатів дослідження. Отже, здійснено тестові випробовування розроблених архітектур для 3D-CNN та ViT в контексті розпізнавання жестів. Розглянуто їх результативність на основі ASL та NUS-II наборів даних. Виділено переваги та недоліки використання згаданих підходів для завдань розпізнавання жестів. Продемонстровано вищу ефективність 3D-CNN порівняно з ViT в умовах обмежених ресурсів, а саме показник F1 становить 0,881536 проти 0,848405, що демонструє перевагу 3D-CNN на 3,8 %. Ці показники порівнянні із результатами, наведеними у [13, 15].

За результатами виконаної роботи можна сформулювати наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів: отримано оцінку ефективності 3D конволюційних нейронних мереж (3D-CNN) та візуальних трансформерів (ViT) для глибокого навчання у процесі розпізнавання жестів в умовах обмежених ресурсів.

Практична значущість одержаних результатів: виділено переваги та недоліки, досліджено ефективність і продуктивність моделей глибокого навчання 3D-CNN та ViT на основі дослідження ASL та NUS-II наборів даних з метою подальшого впровадження в IT рішення для автоматизованого розпізнавання ручних жестів.

Висновок / Conclusions

Під час порівняння моделей конволюційних нейронних мереж та візуальних трансформерів виявлено відмінність у вимогах до навчання, точності та продуктивності моделей на основі показників відтворення (recall), точності (precision) та F1-оцінки. Моделі CNN традиційно відомі компактністю та ефективним використанням пам'яті, що робить їх придатними для ресурсообмежених середовищ. Доведено їх високу ефективність у завданнях оброблення зображень та продемонстровано точність у різних областях комп'ютерного зору [8], [10], [15]. З іншого боку, Vision Transformer пропонує потужний підхід для визначення глобальних залежностей та контекстуального розуміння на зображеннях, що сприяє підвищенню продуктивності у певних завданнях [13], [14], [17]. Однак у Vision Transformer, як правило, більші розміри моделі та вищі вимоги до пам'яті порівняно з CNN. Хоча вони можуть досягати кращої точності, особливо коли працюють із великими наборами даних, обчислювальні вимоги можуть обмежувати їх практичність у ситуаціях із обмеженими ресурсами.

У роботі продемонстровано вищу ефективність 3D-CNN порівняно із ViT в умовах обмежених ресурсів, а саме показник F1 становив 0,881536 проти 0,848405, що демонструє перевагу 3D-CNN на 3,8 %.

Вибір між моделями CNN та Vision Transformer залежить від конкретних вимог завдання, враховуючи такі фактори, як наявні ресурси, розмір набору даних та компроміс між складністю моделі, точністю та продуктивністю. Оскільки завдання розпізнавання жестів залишається актуальним, то подальше дослідження і вдосконалення обох архітектур дасть змогу дослідникам та практикам приймати обґрунтованіші рішення, із урахуванням конкретних потреб та обмежень.

References

[1] Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. http://dx.doi.org/10.1109/CVPRW.2015.7301342

[2] Molchanov, P., Gupta, S., Kim, K., & Pulli, K. (2015). Multi-sensor system for driver's hand-gesture recognition. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, *1*, 1–8. https://doi.org/10.1109/FG.2015.7163132

[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition, 223*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223

[4] Ohn-Bar, E., & Trivedi, M. M. (2014). Hand Gesture Recognition in Real Time for Automotive Interfaces: A Multimodal Vision-Based Approach and Evaluations. *IEEE Transactions on Intelligent Transportation Systems*, *15*, 2368–2377. https://doi.org/10.1109/TITS.2014.2337331

[5] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition. https://doi.org/10.48550/arXiv.1406.2199

[6] Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *2015 International Conference on Computer Vision*, 9, 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[7] Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2014). Multiscale deep learning for gesture detection and localization, 474–490. http://dx.doi.org/10.1007/978-3-319-16178-5_33

[8] Yong, T., Kian, L., Connie, T., Chin-Poo, L., & Cheng-Yaw, L. (2021). Convolutional neural network with spatial pyramid pooling for hand gesture recognition. *Neural Computing and Applications,* 33, 1–13. https://doi.org/10.1007/s00521-020-05337-0

[9] Yong, T., Kian, L., & Chin-Poo, L. (2021). Hand Gesture Recognition via Enhanced Densely Connected Convolutional Neural Network. *Expert Systems with Applications*, 175. https://10.1016/j.eswa.2021.114797

[10] Osimani, C.; Ojeda-Castelo, J. J.; & Piedra-Fernandez, J. A. (2023). Point Cloud Deep Learning Solution for Hand Gesture Recognition. *International Journal of Interactive Multimedia and Artificial Intelligence*. https://doi.org/10.9781/ijimai.2023.01.001

[11] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics.* https://doi.org/10.18653/v1 %2FN19-1423

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[13] Hengshuang, Z., Jiaya, J., & Vladlen, K. (2020). Exploring Self-Attention for Image Recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10073–10082. https://doi.org/10.1109/CVPR42600.2020.01009

[14] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. https://10.1007/978-3-030-58452-8_13

[15] Ji, S. Xu, W., Yang, M., & Yu, K. (2010) 3 d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35 (1), 495–502. http://dx.doi.org/10.1109/TPAMI.2012.59

[16] Barczak, A. L. C., Reyes, N. H., Abastillas, M., Piccio, A., & Susnjak, T. A. (2011). New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures.

[17] Pisharady, P. K., Vadakkepat, P., & Loh, A. P. (2013). Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision*, 101, 403–419. https://doi.org/10.1007/s11263-012-0560-5

V. Ya. Chornenkyi, I. Ya. Kazymyra

Lviv Polytechnic National University, Lviv, Ukraine

RESEARCH OF THE MODELS FOR SIGN GESTURE RECOGNITION USING 3D CONVOLUTIONAL NEURAL NETWORKS AND VISUAL TRANSFORMERS

The work primarily focuses on addressing the contemporary challenge of hand gesture recognition, driven by the overarching objectives of revolutionizing military training methodologies, enhancing human-machine interactions, and facilitating improved communication between individuals with disabilities and machines. In-depth scrutiny of the methods for hand gesture recognition involves a comprehensive analysis, encompassing both established historical computer vision approaches and the latest deep learning trends available in the present day.

This investigation delves into the fundamental principles that underpin the design of models utilizing 3D convolutional neural networks and visual transformers. Within the 3D-CNN architecture that was analyzed, a convolutional neural network with two convolutional layers and two pooling layers is considered. Each 3D convolution is obtained by convolving a 3D filter kernel and summing multiple adjacent frames to create a 3D cube. The visual transformer architecture that is consisting of a visual transformer with Linear Projection, a Transformer Encoder, and two sub-layers: the Multi-head Self-Attention (MSA) layer and the feedforward layer, also known as the Multi-Layer Perceptron (MLP), is considered.

This research endeavors to push the boundaries of hand gesture recognition by deploying models trained on the ASL and NUS-II datasets, which encompass a diverse array of sign language images. The performance of these models is assessed after 20 training epochs, drawing insights from various performance metrics, including recall, precision, and the F1 score. Additionally, the study investigates the impact on model performance when adopting the ViT architecture after both 20 and 40 training epochs were performed.

This analysis unveils the scenarios in which 3D convolutional neural networks and visual transformers achieve superior accuracy results. Simultaneously, it sheds light on the inherent constraints that accompany each approach within the ever-evolving landscape of environmental variables and computational resources.

The research identifies cutting-edge architectural paradigms for hand gesture recognition, rooted in deep learning, which hold immense promise for further exploration and eventual implementation and integration into software products.

Keywords: deep learning; human-machine interactions; neural networks performance; sign language datasets.

Інформація про авторів:

Чорненький Володимир Ярославович, аспірант, кафедра автоматизованих систем управління.
Email: volodymyr.y.chornenkyi@lpnu.ua; https://orcid.org/0009-0000-0569-6623

Казимира Ірина Ярославівна, канд. техн. наук, доцент, кафедра автоматизованих систем управління.
Email: iryna.y.kazymyra@lpnu.ua; https://orcid.org/0009-0000-0569-6623

Цитування за ДСТУ: Чорненький В. Я., Казимирal. Я. Дослідження моделей для розпізнавання жестів з використанням 3D конволюційних нейронних мереж та візуальних трансформерів. *Український журнал інформаційних технологій.* 2023. Т. 5, № 2. С. 33–40.

Citation APA: Chornenkyi, V. Ya. , & Kazymyra, I. Ya. (2023). Research of the models for sign gesture recognition using 3D convolutional neural networks and visual transformers. *Ukrainian Journal of Information Technology*, 5(2), 33–40. https://doi.org/10.23939/ujit2023.02.033