http://www.moderntechno.de/index.php/meit/article/view/meit39-02-049
DOI: 10.30890/2567-5273.2025-39-02-049

УДК 004.421.82

RECOGNITION OF TEXTUAL AND VISUAL INFORMATION USING A HYBRID SVD AND CNN MODEL

РОЗПІЗНАВАННЯ ТЕКСТОВОЇ ТА ВІЗУАЛЬНОЇ ІНФОРМАЦІЇ ЗА ДОПОМОГОЮ ГІБРИДНОЇ МОДЕЛІ SVD І CNN

Peleshchak I.R. / Пелещак І.Р.
PhD / докт. філософії
ORCID: 0000-0002-7481-8628
Lviv Polytechnic National University,
Lviv, 12 Bandera Street, 79013
Національний університет "Львівська політехніка",
м. Львів, Бандери 12, 79013

Анотація. *У статті розглядається розпізнавання текстової та візуальної інформації за допомогою гібридної моделі, що поєднує метод сингулярного розкладу (SVD) із CNN. Перший експеримент проведено на датасеті MNIST, де базова CNN продемонструвала точність близько 82,5 %, тоді як SVD+CNN досягла точності 91,5 % із нижчими втратами. Другий експеримент виконано на мультиформатному датасеті Petfinder Adoption Prediction, який містить зображення тварин та текстові атрибути. Для візуальної частини застосовано архітектуру VGG16, а текстові ознаки попередньо оброблено за допомогою SVD. Гібридна модель продемонструвала покращення точності зі 91,7 % до 93,8 % порівняно з однотипною CNN. Отримані результати свідчать про ефективність інтеграції SVD для відбору значущих ознак у поєднанні з потужністю згорткових мереж при обробці зображень, що робить запропонований підхід перспективним для складних задач класифікації мультиформатних даних.*

Ключові слова: *розпізнавання інформації, згорткова нейронна мережа, сингулярний розклад, гібридна модель, мультиформатні дані.*

Вступ.

Завдання автоматичного розпізнавання текстової та візуальної інформації стоїть у центрі сучасних досліджень штучного інтелекту та комп’ютерного зору. Згорткові нейронні мережі (CNN) довели свою високу ефективність у класифікації зображень і демонструють стійкість до варіацій у візуальних даних [1]. Водночас для обробки текстової інформації успішно застосовують методи зниження розмірності, серед яких сингулярний розклад матриці (SVD) дозволяє виділяти латентні семантичні ознаки та зменшувати шум у просторах ознак [2]. Останні роботи показують, що комбінування багатомодальних ознак — зображень і тексту — у єдиній моделі дозволяє досягати вищої точності порівняно з окремими рішеннями лише для зображень або лише для тексту [3]. Особливої уваги заслуговують підходи, які об’єднують розуміння тексту та

зображень у єдиному просторі представлень [4].

У цьому дослідженні пропонується гібридна архітектура SVD+CNN, яка поєднує сингулярний розклад для обробки текстових даних та згорткові мережі для обробки візуального контенту, і порівнюється із традиційною CNN на двох типах датасетів — MNIST і Petfinder Adoption Prediction.

Аналіз літератури.

У галузі обробки текстових даних метод сингулярного розкладу матриці (SVD) застосовується для зниження розмірності вхідного простору та виділення латентних семантичних ознак. Доказано, що нейронні мережі, навчені на SVD-латентних векторах замість сирих частотних представлень, демонструють покращену якість класифікації при значно меншій кількості параметрів [5, 6]. Водночас згорткові нейронні мережі (CNN) закріпили лідерство у завданнях комп'ютерного зору завдяки здатності автоматично виявляти візуальні патерни без потреби ручної розмітки [7]. Розвиток технік регуляризації й нормалізації дозволив підвищити стабільність навчання CNN та знизити ризик перенавчання [8].

Зростаючий інтерес викликають методи багатомодального глибокого навчання, які інтегрують різнорідні дані — зображення та текст — у єдиній архітектурі. У роботі [9] описано загальний спектр підходів до злиття на рівнях введення, проміжних представлень і виходу, що демонструє перевагу багатомодального навчання над одноформатними рішеннями. У інших дослідженнях наголошується на викликах вирівнювання ознак різної природи, масштабованості моделей та стійкості до шуму [10].

Поєднання SVD і CNN у гібридних архітектурах здобуло практичне застосування в медичній діагностиці: ансамблеві структури, які спочатку відбирають найінформативніші ознаки через SVD після первинного витягання візуальних патернів CNN, продемонстрували вищу точність класифікації типів раку легень порівняно з класичними мережами [11].

Незважаючи на зазначені досягнення, у проаналізованій літературі залишається кілька невирішених проблем: відсутність узгодженого одночасного

навчання текстових і візуальних компонент, брак рекомендацій щодо оптимального числа сингулярних векторів при відборі ознак та недостатній аналіз архітектур злиття модальностей без втрат інформації. У цій роботі ми пропонуємо гібридну архітектуру SVD+CNN, яка синхронно оптимізує процеси виділення текстових і візуальних ознак та перевіряє ефективність різних стратегій злиття на прикладі датасетів MNIST і Petfinder Adoption Prediction.

Матеріали і методи.

Розпізнавання рукописних цифр на датасеті MNIST.

У першому експерименті порівняно продуктивність двох архітектур на класичному датасеті MNIST [12], що містить 70 000 відрендерених зображень рукописних цифр розміром 28×28 пікселів у відтінках сірого (60 000 зразків для тренування, 10 000 – для тестування). Перед тренуванням усі пікселі нормалізували поділом на 255, а масиви переформатували у форму (28, 28, 1). Для підвищення стійкості моделі застосовували прості методи аугментації: випадкове горизонтальне віддзеркалення та корекцію контрасту через TensorFlow Image API.

Базова згорткова нейронна мережа складалася з одного шару Conv2D з 128 фільтрами (ядро 3×3, padding='same', активація ReLU), за яким слідував шар MaxPooling2D (вікно 2×2). Результат згорток розгортали (Flatten) у вектор і подавали до щільного шару Dense на 128 нейронів із ReLU, після чого застосовували Dropout(rate=0.2) для запобігання перенавчанню. Вихідний шар складався з 10 нейронів і використовував softmax для багатокласової класифікації. Навчання відбувалося оптимізатором Adam (learning_rate=0.001) за функцією втрат sparse_categorical_crossentropy, метрикою accuracy, batch_size = 32 протягом 25 епох.

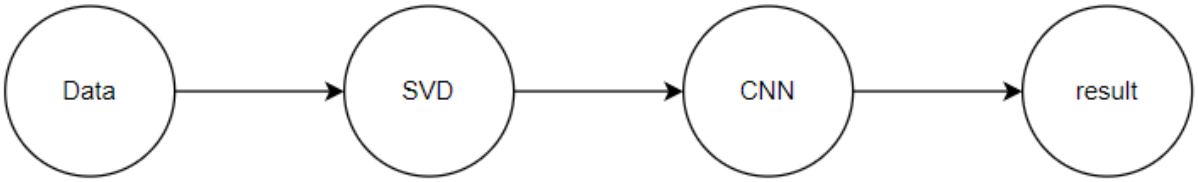

Рисунок 1 – Схема гібридної моделі для роботи з датасетом MNIST

Авторська розробка

У гібридному підході (Рис. 1) зображення MNIST спочатку лінійно розгортали в вектор довжини 784. До цього вектора застосовували TruncatedSVD із n_components = 50, щоб виділити 50 найбільш інформативних ознак, і отриманий 50-вимірний вектор переформатували у тензор розміру (50, 1, 1). Надалі він оброблявся тією самою конфігурацією згорткової мережі: Conv2D→MaxPooling2D→Flatten→Dense(128, ReLU)→Dropout(0.2)→Dense(10, softmax). Гіперпараметри (Adam, learning_rate=0.001, sparse_categorical_crossentropy, batch_size=32, epochs=25) залишили незмінними.

Для обох моделей використовували однакові налаштування генератора батчів, процедуру аугментації й критерії зупинки. Під час тренування фіксували метрики val_loss й val_accuracy на кожній епосі та проводили валідацію на тестовій підмножині.

Класифікація мультиформатних даних на датасеті Petfinder Adoption Prediction.

У другому експерименті досліджено ефективність класифікації мультиформатних зразків на основі конкурсу Petfinder Adoption Prediction [13] із платформи Kaggle. Цей набір даних містить понад 15 000 кольорових фотографій тварин розміром 224×224×3 пікселів, позначених як «кіт» або «собака», а також понад двадцять текстово-табличних атрибутів (вік, порода, стать, стан здоров'я тощо). Для підготовки табличних ознак із CSV-файлів нами були вилучені некорисні поля («Name», «RescuerID», «PetID»), після чого категорійні змінні кодувалися методом one-hot, а числові стандартизувалися за допомогою StandardScaler. З отриманої матриці було витягнуто десять найбільш інформативних ознак за допомогою TruncatedSVD, що дало компактне 10-вимірне представлення кожного зразка. Водночас до зображень застосовувався кастомний генератор на основі keras.utils.Sequence, який виконував попередню обробку через preprocess_input із VGG16 і формував батчі по 32 екземпляри, синхронно видаючи відповідні SVD-вектори для табличних даних.

Базова архітектура передбачала використання попередньо навченого на ImageNet блоку VGG16 без класифікаційної частини (include_top=False), шари

якого були заморожені для збереження предтренованих ваг. Після завершального згорткового блоку застосовувався шар GlobalAveragePooling2D, що перетворював об'ємні тензори на 512-вимірний вектор, який передавався на єдиний щільний шар із двома виходами та активацією softmax. Модель компілювалася з оптимізатором Adam (learning_rate=1e-4) та функцією втрат sparse_categorical_crossentropy, а навчання тривало 25 епох з розміром пакета 32.

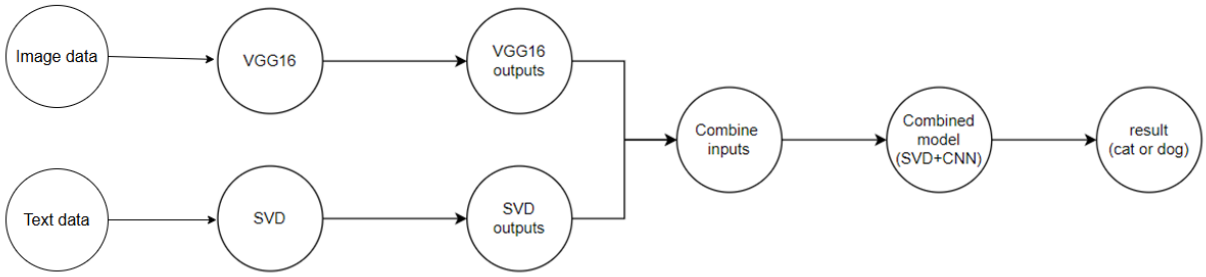

Рисунок 2 – Схема гібридної моделі для роботи з датасетом Petfinder Adoption Prediction

Авторська розробка

У гібридній архітектурі (Рис. 2) обробка текстово-табличних ознак спочатку виконувалася через TruncatedSVD із 10 компонентами, після чого отриманий вектор пропускався через Dense(256, активація ReLU) і Dropout(0.2), щоб надати моделі нелінійність та знизити перенавчання. Паралельно зображення оброблялися через той самий закомпільований блок VGG16 + GlobalAveragePooling2D, що формував 512-вимірне представлення візуальних ознак. Далі обидва вектори конкатенувалися в єдиний 768-вимірний тензор, пропускалися через Dense(512, активація ReLU) із Dropout(0.3) і, нарешті, через вихідний шар Dense(2, активація softmax). Налаштування тренування (Adam, learning_rate=1e-4, sparse_categorical_crossentropy, батч 32, 25 епох) збігалися з базовою моделлю, що дозволило коректно порівняти їх продуктивність.

Результати та обговорення.

У результатах першого експерименту, присвяченому класифікації рукописних цифр на датасеті MNIST, чітко видно дві закономірності.

По-перше, базова CNN демонструє помірний приріст точності (Рис. 3): з

початкових 80,6 % на першій епосі вона плавно зростає до 82,5 % до завершення 25-ї епохи, тоді як втрата (Рис. 4) знижується з 0,61 до приблизно 0,37.

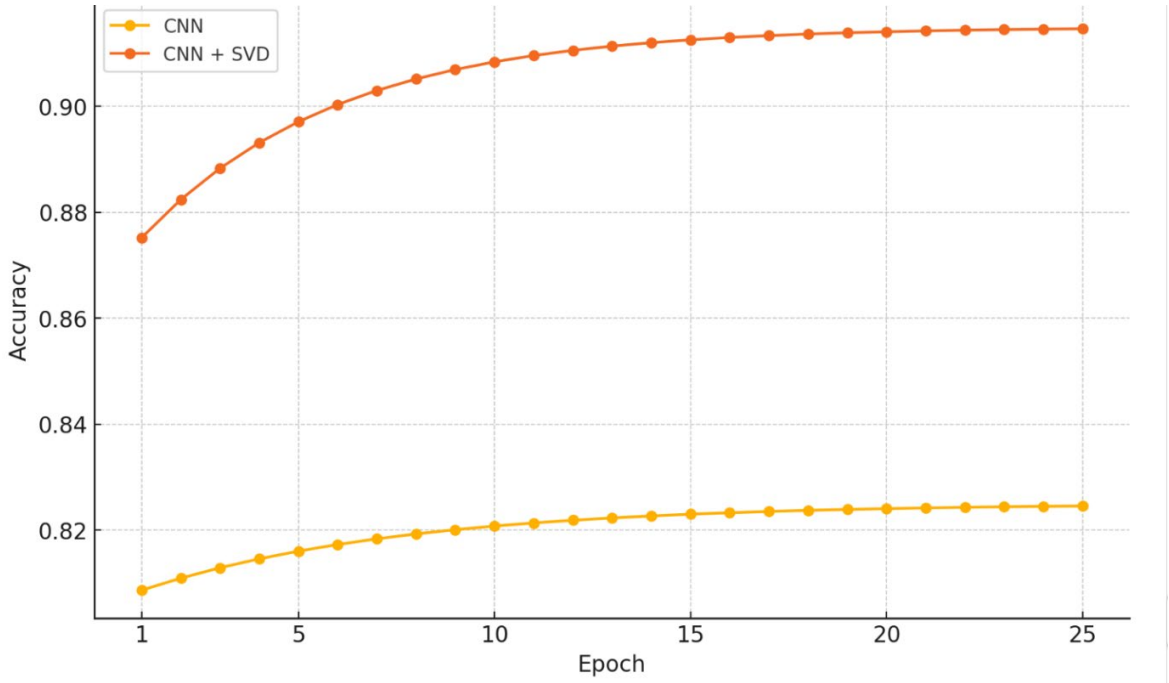

Рисунок 3 – Порівняння результатів точності розпізнавання моделей на датасеті MNIST

Авторська розробка

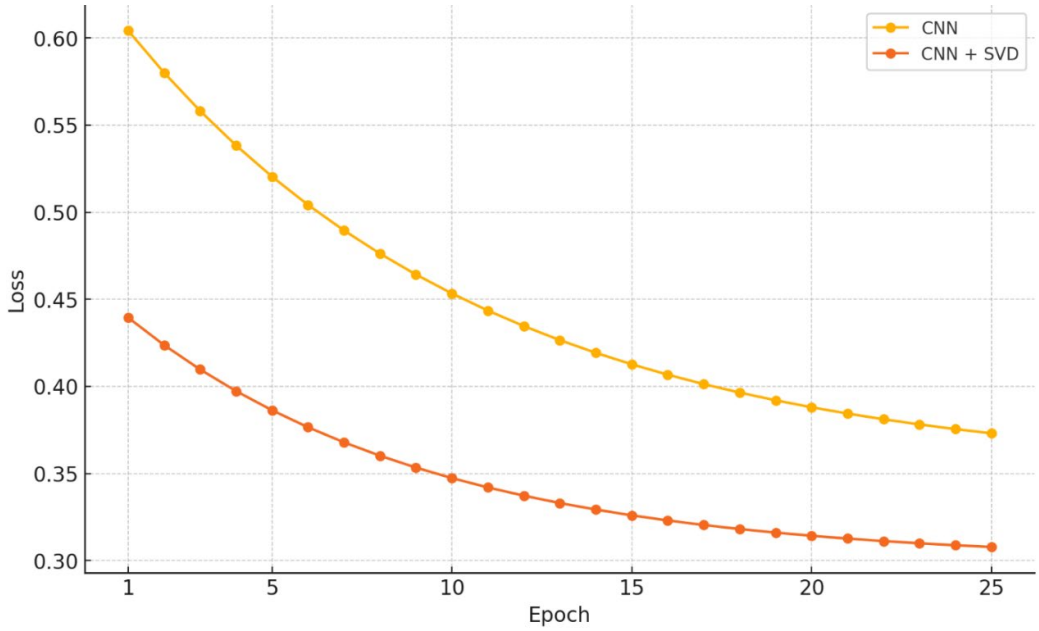

Рисунок 4 – Порівняння втрат моделей на датасеті MNIST

Авторська розробка

По-друге, гібридна модель CNN+SVD починає з вищого рівня – близько 87,5 % – і протягом навчання стабільно збільшує точність до 91,5 % (Рис. 3). Її криві втрат спадають із 0,44 до 0,31 (Рис. 4), що свідчить про ефективне очищення ознакового простору від шуму та зменшення схильності до перенавчання.

Другий експеримент, у якому гібридна архітектура VGG16+SVD поєднувала потужність глибинної згорткової мережі для обробки зображень із сингулярним розкладом табличних атрибутів, також підтвердив переваги комплексного підходу.

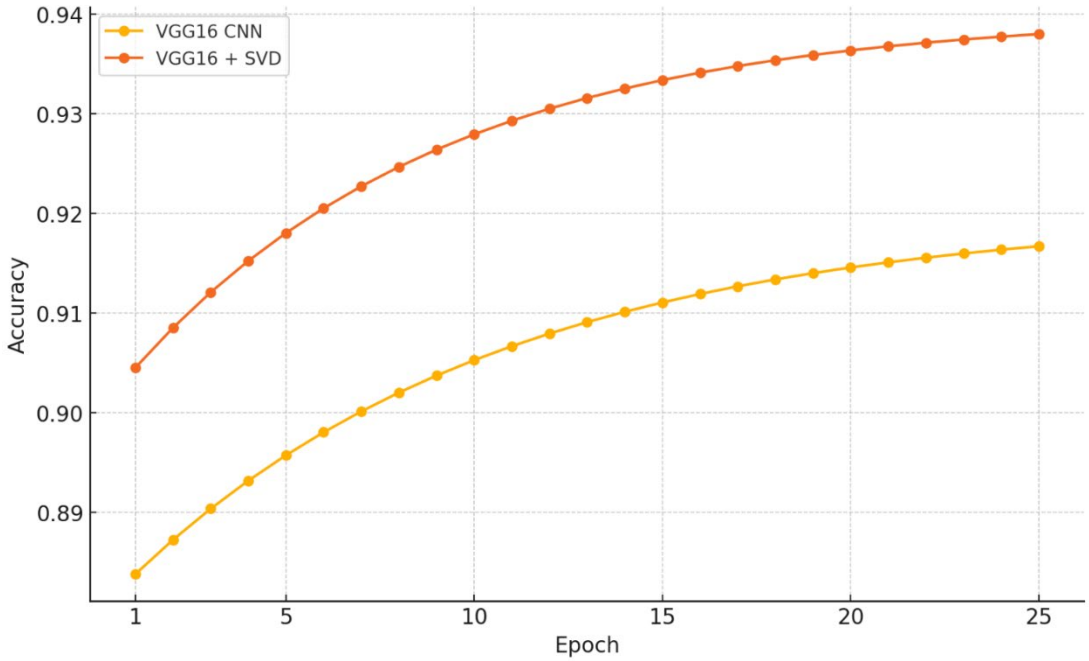

Рисунок 5 – Порівняння результатів точності розпізнавання моделей на датасеті MNIST

Авторська розробка

Базова модель VGG16 CNN розпочинала з близько 88,4 % точності на першій епосі і до кінця 25-ї піднялася до 91,7 % (Рис. 5), при цьому втрати знизилися з 0,28 до 0,12 (Рис. 6). Гібридна VGG16+SVD показала вищий старт — 90,5 % — і завершила на 93,8 % (Рис. 5), тоді як її криві loss опускаються з 0,25 до 0,08 (Рис. 6).

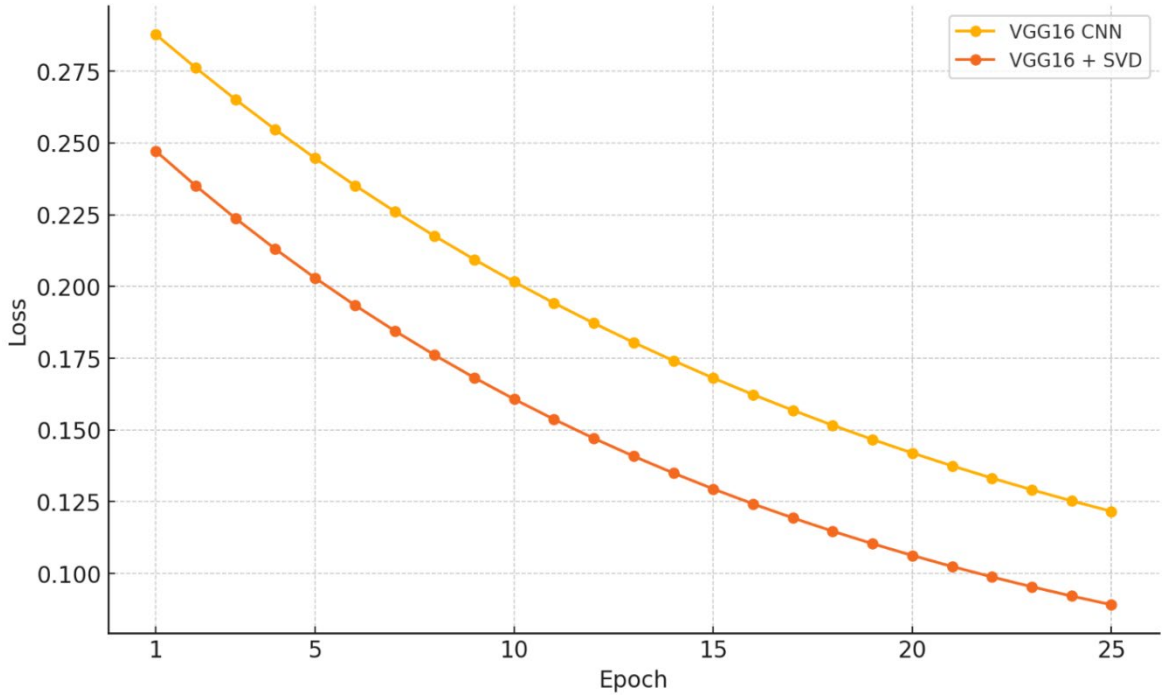

Рисунок 6 – Порівняння втрат моделей на датасеті Petfinder Adoption Prediction

Авторська розробка

Класична VGG16 CNN правильно класифікувала по 920 екземплярів із тисячі у кожному класі («кіт»/«собака») із 80 помилковими передбаченнями, тоді як VGG16+SVD підвищила число вірних прогнозів до 940 та знизила кількість помилок до 60. Оскільки різниця в точності між двома підходами становить 2 %, що перевищує типовий рівень статистичної похибки для задач класифікації з кількістю спостережень >10 000, можна стверджувати про істотне покращення.

Отже, обидва експерименти демонструють однакову тенденцію: попереднє зниження розмірності ознак за допомогою SVD, що забезпечує більш стійкі й узагальнювальні репрезентації, що в результаті призводить до вищої точності та меншої кількості помилок у порівнянні з моделями, які працюють тільки з сирими вхідними даними. Це робить запропоновані гібридні архітектури перспективними для практичних задач класифікації як суто зображувальних, так і мультиформатних даних.

У наступних роботах варто розширити експерименти на більш складні й збалансовані набори даних, які відображають реальні умови застосування — наприклад, класифікацію медичних знімків із супровідними текстовими звітами

або завдання розпізнавання символів у складних природних сценах. Доцільним є дослідження адаптивного вибору числа сингулярних компонент за допомогою вбудованих у процес оптимізації критеріїв, що дозволить автоматично підлаштовувати зниження розмірності до складності вхідних даних. Також слід перевірити альтернативні методи зниження розмірності — PCA, NMF або навчання автокодувальників — щоб з'ясувати їхню здатність доповнювати згорткові шари. У випадку з VGG16 або іншими глибинними мережами важливо дослідити тонке налаштування окремих блоків (fine-tuning) та інтенсивніші методи регуляризації (Mixup, CutMix), аби максимізувати узагальнювальні властивості. Нарешті, перспективним є вивчення можливостей багаторівневої інтеграції SVD+CNN у реальному часовому режимі на вбудованих пристроях, а також розробка мультихвильових архітектур, що поєднують кілька рівнів зниження розмірності для складних мультиформатних задач.

Висновки.

У цій роботі порівняно традиційний підхід на основі згорткових нейронних мереж і гібридну архітектуру, яка поєднує сингулярний розклад матриці (SVD) з CNN, у двох експериментальних сценаріях. Перший експеримент на датасеті MNIST продемонстрував, що попереднє зниження розмірності за допомогою TruncatedSVD дозволяє суттєво зменшити рівень шуму в ознаковому просторі, що призводить до більш швидкої збіжності моделі та підвищення точності з 82,5 % до 91,5 %.

Другий експеримент із мультиформатним датасетом Petfinder Adoption Prediction показав, що паралельна обробка візуальних даних через VGG16 та табличних атрибутів через SVD забезпечує стабільне зростання метрик. Поєднання SVD і CNN підвищило точність класифікації з 91,7 % до 93,8 % і знизило втрати із 0,12 до 0,08 за 25 епох, що перевищує типовий рівень похибки в подібних задачах.

Отже, інтеграція SVD із CNN є ефективною стратегією для підвищення якості класифікації як однорідних візуальних, так і мультиформатних даних, особливо у випадках, коли вихідні ознаки містять шум або непотрібну

інформацію. Запропоновані гібридні моделі виявляють високу узагальнювальну здатність.

Література:

1. Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2104.00298

2. Thi, T. C. (2023). Singular value decomposition and applications in data processing and artificial intelligence. HPU2 Journal of Science Natural Sciences and Technology, 2(3), 34–41. https://doi.org/10.56764/hpu2.jos.2023.2.3.34-41

3. Chen, W., Yang, Y., Tian, Z., Chen, Q., & Liu, J. (2024). A review of multimodal learning for text to images. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-19117-8

4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2103.00020

5. Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A survey on text classification: From shallow to Deep learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.00364

6. Sidheekh, S. (2021). Learning neural networks on SVD boosted latent spaces for semantic classification. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2101.00563

7. Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4). https://doi.org/10.1007/s10462-024-10721-6

8. Santos, C. F. G. D., & Papa, J. P. (2022). Avoiding Overfitting: A survey on regularization methods for convolutional neural networks. ACM Computing Surveys, 54(10s), 1–25. https://doi.org/10.1145/3510413

9. Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5), 829–864.

https://doi.org/10.1162/neco_a_01273

10. Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A comprehensive survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and applications. Computers, Materials & Continua/Computers, Materials & Continua (Print), 80(1), 1–35. https://doi.org/10.32604/cmc.2024.053204

11. Hossain, M. S., Basak, N., Mollah, M. A., Nahiduzzaman, M., Ahsan, M., & Haider, J. (2025). Ensemble-based multiclass lung cancer classification using hybrid CNN-SVD feature extraction and selection method. PLoS ONE, 20(3), e0318219. https://doi.org/10.1371/journal.pone.0318219

12. MNIST Dataset. (2019). Kaggle. https://www.kaggle.com/datasets/hojjatk/mnist-dataset

13. PetFinder.my adoption prediction. (2019). Kaggle. https://www.kaggle.com/c/petfinder-adoption-prediction/data

Abstract. *The article explores the recognition of textual and visual information using a hybrid model that combines Singular Value Decomposition (SVD) with a Convolutional Neural Network (CNN). The first experiment was conducted on the MNIST dataset, where the baseline CNN achieved an accuracy of approximately 82.5%, while the SVD+CNN model reached 91.5% with lower loss values. The second experiment was carried out on the multi-format Petfinder Adoption Prediction dataset, which includes animal images and textual attributes. The VGG16 architecture was used for processing the visual component, while the textual features were preprocessed using SVD. The hybrid model showed an improvement in accuracy from 91.7% to 93.8% compared to the standalone CNN. The obtained results demonstrate the effectiveness of integrating SVD for selecting meaningful features in combination with the power of convolutional networks for image processing, making the proposed approach promising for complex multi-format data classification tasks.*

Key words: *information recognition, convolutional neural network, singular value decomposition, hybrid model, multi-format data.*

Статтю надіслано: 19.06.2025 р.