УДК 004. 7 **doi: 10.26906/SUNZ.2021.4.079**

А. В. Шостак

Національний аерокосмічний університет імені М. Є. Жуковського «ХАІ», Харків, Україна

ПРО ОСОБЛИВОСТІ ФОРМУВАННЯ ДЕСКРИПТОРІВ У СІАМСЬКІЙ НЕЙРОННІЙ МЕРЕЖІ

Анотація. Предмет дослідження − процеси розпізнавання зображень рукописних цифр із застосуванням нейронних мереж. Додаток для розпізнавання ґрунтується на архітектурі сіамської мережі з нейронними згортковими підмережами. **Мета статті** − обґрунтування вибору N-вимірних векторних уявлень вхідних зображень для опису їх властивостей, порівняння та їхнього розпізнавання. **Завдання**: експериментальне дослідження розпізнавання зображень рукописних цифр із використанням архітектури сіамської нейронної мережі. **Методи досліджень**: метод прямого пошуку для функцій з декількома змінними для визначення N-вимірних векторних представлений вхідних зображень. **Методи досліджень**: метод прямого пошуку для функцій з декількома змінними для визначення N-вимірних векторних уявлень вхідних зображень. **Результати досліджень**. Результати досліджень. Проведено визначення N-вимірних векторних уявлень вхідних зображень рукописних цифр та досліджено їх характеристики. Виконано експериментальне дослідження розпізнавання зображень із використанням векторних уявлень зображень у рамках архітектури сіамської нейронної мережі. Показано, що запропоновані методи визначення векторних N-вимірних уявлень вхідних зображень є робастними і незначно впливають на кількість помилок при тестуванні розпізнавання. Під час тестування використовувалися зображення рукописних цифр із тестового набору MNIST. Визначено, що використання наперед вибраних еталонних уявлень вхідних зображень дозволяє спростити архітектуру сіамської мережі. **Висновки**. Результати, отримані в роботі, можуть бути використані при порівняльній оцінці та визначенні N-вимірних векторних уявлень різних класів вхідних зображень з метою розпізнавання їх з використанням архітектури сіамської нейронної мережі.

Ключові слова: сіамська нейронна мережа, дескриптор, тестування нейронної мережі.

Вступ

Сіамська нейронна мережа це один із видів нейронних мереж, що складається, як правило, з трьох підмереж [1]. Вхідні дані подаються на входи перших двох підмереж, що мають однакові набори ваг та ідентичні за архітектурою. Перші дві підмережі формують векторні уявлення (дескриптори) вхідних даних. Третя підмережа на підставі результатів роботи перших двох підмереж формує підсумковий результат у порівнянні вхідних даних.

Сіамські мережі широко використовуються в системах розпізнавання обличь та інших графічних образів, для порівняння текстів, в системах перевірки підпису тощо [1, 2].

Виходом сіамської мережі є оцінка ступеню подібності або відмінності вхідних даних на двох входах мережі. На підставі цієї оцінки й виконується розпізнавання об'єктів на входах мережі [1-4].

У відомих джерелах недостатньо повно описані особливості побудови та використання дескрипторів вхідних даних.

Основна частина

На рис. 1 представлена класична архітектура сіамської нейронної мережі. i1 та i2 – входи даної мережі, на які подаються зображення рукописних цифр розміру 28*28 пікселів з набору даних MNIST [5]. Підмережі 1 і 2 однакові за архітектурою, кількістю і значенням вагових параметрів. Ці підмережі зазвичай будуються на підставі сукупності послідовних повнозв'язних або згорткових шарів.

h(i1) та h(i2) – виходи (або дескриптори вхідних зображення) обох підмереж 1 і 2. Дескриптор h(i1) представляє собою N-мірне векторне представлення зображення i1, тобто N-мірний вектор властивостей зображення i1. Евклідова відстань d(h(i1), h(i2)) для дескрипторів h(i1) і h(i2) максимальна для різних зображень i1 та i2 й мінімальна для однакових.

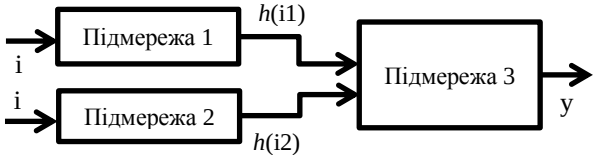


Рис. 1. Узагальнена архітектура сіамської мережі

Підмережа 3 обчислює евклідову відстань d(h(i1), h(i2)) між дескрипторами і на її підставі формує оцінку ступеня схожості або відмінності зображень на входах мережі i1 та i2.

При навчанні сіамських нейронних мереж найчастіше використовуються наступні функції втрат [2-4]: бінарна крос-ентропія, контрастна функція втрат, триплетна функція втрат.

Контрастна та триплетна функції втрат прагнуть зменшити відстань між схожими об'єктами і збільшити відстань між різними об'єктами. При використанні триплетної функції втрат застосовуються найбільш складні вимоги формування пар зображень для навчального набору даних [4].

З метою аналізу дескрипторів вхідних даних використовувалася сіамська нейронна мережа з [3]. Всі обчислення проводилися у блокноті хмарного сервісу Google Colaboratory.

Структура моделі підмережі 1 (підмережі 2) наведена на рис. 2. На вхід підмережі 1 подається одноканальне зображення у градаціях сірого кольору з рукописною цифрою i1 розміру 28*28 пікселів – Input(28*28*1). Далі йде згортковий шар (Conv2D_1(4, (5, 5), tanh)), який за допомогою ядер згортки розміру (5, 5) формує 4 карти ознак та використовує функцію активації гіперболічного тангенса

tanh. Після першого шару згортки слід субдискретизований шар AveragePooling2D(2, 2), який замінює дані в вікні розміру (2, 2) їх середнім значенням. Далі – другий шар згортки (Conv2D_2(16, (5, 5), tanh)), який за допомогою ядер згортки розміру (5, 5) формує 16 карт ознак і використовує функцію активації tanh. Наступний – шар AveragePooling2D(2, 2).

Далі шар Flatten, на виході якого з вхідних даних формується одновимірний вектор, і шар BatchNormalization, що виконує пакетну нормалізацію для прискорення процесу навчання мережі. Вихід підмережі 1 формується повнозв'язним шаром (FC(10, tanh)) з 10 вузлів з використанням функції активації tanh. Тому дескриптор вхідного зображення h(i1) є 10-мірним вектором з елементами розподіленими від -1 до +1.

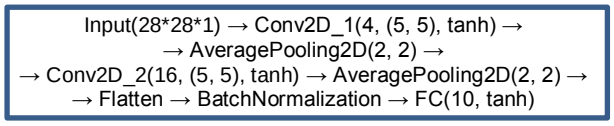


Рис. 2. Структура моделі підмережі 1

Структура моделі підмережі 3 наведена на рис. 3. На два входи (Input(10)*2) підмережі 3 подаються два 10-мірних дескриптора h(i1) та h(i2). Далі йде шар Lambda, що обчислює евклідову відстань між дескрипторами, і шар BatchNormalization. Вихід підмережі 3 формується повнозв'язним шаром (FC(1, sigmoid)) з 1 вузла з використанням функції активації sigmoid. Тому на виході всієї даної сіамської мережі формується значення y в інтервалі від 0 до 1, що показує ступінь подібності двох вхідних зображень рукописних цифр i1 та i2.

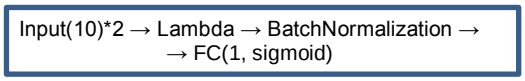


Рис. 3. Структура моделі підмережі 3

При навчанні мережі використовувалася контрастна функція втрат [2, 3], величина пакету дорівнює 16 (batch_size = 16) і кількість епох – 10 (epochs = 10). Решта значень гіперпараметрів при навчанні були значеннями за замовчуванням.

Оцінка якості моделі на тестових даних показала для метрики accuracy значення 0,9824.

Еталонні дескриптори для цифр будувалися на підставі дескрипторів, отриманих на виході повнозв'язного шару FC(10, tanh) (рисунок 2) з тренувального набору зображень рукописних цифр. Для побудови еталонних дескрипторів цифр були проаналізовані наступні способи:

1. h1 – дескриптор, який відповідає конкретному зображенню цифри і мінімізує суму евклідових відстаней від нього до всіх інших дескрипторів зображень з відповідною цифрою,

2. h2 – дескриптор, визначений за допомогою одного із методів прямого пошуку для функції з кількома змінними [6], і який мінімізує суму евклідових відстаней від нього до всіх інших дескрипторів зображень з відповідною цифрою,

3. h3 – дескриптор, визначений за допомогою одного з методів прямого пошуку для функції з кількома змінними [6], і який мінімізує суму евклідових відстаней від нього до k випадково вибраних дескрипторів зображень з відповідною цифрою.

Еталонні дескриптори для десяти цифр обчислювалися на підставі набору з 30000 10-мірних дескрипторів, отриманих з зображень цифр тренувального набору даних.

Дескриптор h10 мінімізує сумарну всередині-кластерну евклідову відстань (SR) для кластера з 2961 дескриптора зображень цифри 0 тренувального набору даних. Наприклад, дескриптор h10 для цифри 0 має такий вигляд:

h10 = [0.9993759, - 0.99132943, - 0.99756294,
- 0.94506454, - 0.9919112, 0.9991755, - 0.963361,
- 0.69670916, 0.99568534, 0.9948068].

У табл. 1 наведено кількість зображень різних цифр в тренувальному наборі, сумарні всередині-кластерні відстані та середні всередині-кластерні відстані (SSR) (з точністю до чотирьох знаків після коми), що забезпечуються дескрипторами h1 для відповідної цифри. Найкомпактніший кластер утворюють дескриптори цифри 3 – середня всередині-кластерна відстань для дескрипторів цифри 3 складає 0.2802, кількість цифр в цьому кластері 3073. Найнекомпактніший кластер утворюють дескриптори цифри 0 – SSR = 0.595, кількість цифр в цьому кластері 2961.

Таблиця 1 – **Сумарна всередині-кластерна відстань і середня всередині-кластерна відстань для h1**

Цифра	Число цифр	SR	SSR
0	2961	1761.74	0.5950
1	3423	1311.64	0.3832
2	2948	1654.30	0.5612
3	3073	861.07	0.2802
4	2926	1523.16	0.5206
5	2709	1352.85	0.4994
6	2975	1181.38	0.3971
7	3107	1308.17	0.4210
8	2875	1281.36	0.4457
9	3003	1235.98	0.4116

Для пошуку оптимальних дескрипторів h2 або h3 використовувався метод Нелдера-Міда [6] з пакету optimize бібліотеки SciPy в Python. Дескриптори h2 розраховувалися на підставі дескрипторів всіх цифр відповідного кластера тренувального набору, дескриптори h3 – на підставі k=10 випадково вибраних дескрипторів відповідного кластера.

У табл. 2 для зображень з кожною цифрою наведені кількість цифр в тренувальному наборі, сумарні всередині-кластерні відстані і середні всередині-кластерні відстані, забезпечене дескрипторами h2 та h3 для відповідної цифри. Наприклад, дескриптори h20 і h30 для цифри 0 мають такий вигляд:

h20=[0.98109656, - 0.98173788, - 0.98839225,
- 0.96168772, - 0.99090116, 0.98956289, - 0.97216287,
- 0.69049328, 0.98800528, 0.99558528],

h30=[0.9428392, - 0.93947548, - 0.98871858,
- 0.98101832, - 0.99888479, 0.96929498, - 0.99031109,
- 0.77560263, 0.98796668, 0.99986114].

Таблиця 2 – **Сумарна всерединікластерна відстань і середня всерединікластерна відстань для h2 та h3**

Цифра	Число цифр	Для h2		Для h3	
		SR	SSR	SR	SSR
0	2961	1755.98	0.5930	1799.13	0.6076
1	3423	1304.64	0.3811	1344.94	0.3929
2	2948	1650.18	0.5598	1708.59	0.5796
3	3073	859.71	0.2798	861.86	0.2805
4	2926	1512.94	0.5171	1518.41	0.5189
5	2709	1349.86	0.4983	1363.79	0.5034
6	2975	1175.42	0.3951	1182.45	0.3975
7	3107	1303.09	0.4194	1308.62	0.4212
8	2875	1278.35	0.4446	1309.50	0.4555
9	3003	1229.58	0.4095	1240.28	0.4130

На рис. 2 представлена архітектура сіамської нейронної мережі для тестування застосування дескрипторів h1, h2 та h3. На перший вхід мережі подається еталонний дескриптор j-й цифри hj, на другий вхід мережі подається зображення рукописної цифри і розміру 28*28 пікселів з набору даних MNIST.

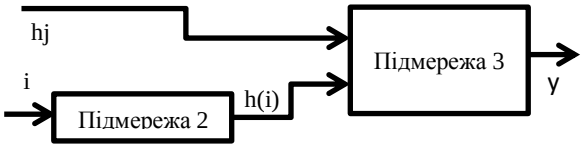


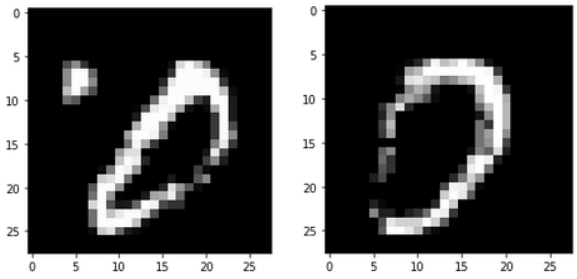
Рис. 2. Архітектура сіамської мережі при тестуванні

Підмережі 2 і 3 повністю аналогічні відповідним підмережам архітектури мережі на рис. 1. Дескриптор h(i) є N-мірним векторним представленням зображення і. Очевидно, що після визначення дескрипторів для цифр h використання сіамської нейронної мережі може виконуватися відповідно до архітектури на рис. 2.

У табл. 3 представлені результати тестування 980 зображень з цифрою 0 з тестового набору з використанням дескрипторів h1, h2 і h3 для цифр від 0 до 9. У стовпці з цифрою 0 показаний результат тестування, коли на вхід h мережі (рис. 2) подається значення дескриптора h10 для цифри 0 (рядки 2, 3) (дескриптору h20 відповідають рядки 4, 5, а дескриптору h30 – рядки 6, 7), а на вхід і – одне з 980 зображень з цифрою 0 з тестового набору. Цифра 7 в другому рядку стовпця з цифрою 0 означає, що для дескриптора h10 величина виходу нейронної мережі для 7 зображень цифри 0 з 980 була менше 0.5, причому мінімальне значення виходу мережі було 0.0078 (з точністю до чотирьох знаків після коми). На рис. 3, а показано це зображення рукописної цифри з тестового набору з міткою 0, яке відповідає значенню виходу мережі 0.007797569. У стовпці для цифри 0 в рядках для виходу мережі у вказується мінімальне значення виходу мережі. У стовпцях для цифр від 1 до 9 в рядках для виходу мережі у вказується максимальне значення виходу мережі.

Таблиця 3 – **Результати тестування на тестовому наборі зображень з цифрою 0 з використанням дескрипторів h1, h2 та h3 для цифр від 0 до 9**

Цифра	0	1	2	3	4	5	6	7	8	9
Помилок для h1	7	0	3	0	0	0	2	1	0	0
y	0.0078	0.0500	0.8800	0.1115	0.3770	0.4442	0.8717	0.9961	0.3857	0.1309
Помилок для h2	7	0	3	0	0	0	2	1	0	0
y	0.0077	0.0522	0.8857	0.1119	0.3840	0.4534	0.8744	0.9961	0.3939	0.1360
Помилок для h3	7	0	3	0	0	0	2	1	0	0
y	0.0063	0.0580	0.8793	0.1121	0.3951	0.4642	0.8664	0.9960	0.3944	0.1338



а б

Рис. 3. Зображення цифри з міткою 0
(а – для h10 y = 0.007797569; б – для h17 y = 0.996109)

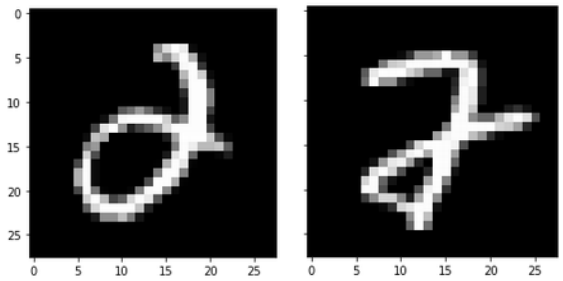
На рис. 3, б показано єдине зображення рукописної цифри з тестового набору з міткою 0, для якого для h17 значення виходу мережі більше 0.5 (стовпець таблиці 3 з цифрою 7, рядок 2 – значення виходу мережі y = 0.9961).

У табл. 4 представлені результати тестування 1032 зображень з цифрою 2 з тестового набору з використанням дескрипторів h1, h2 та h3 для цифр від 0 до 9. У стовпці з цифрою 2 показаний результат тестування, коли на вхід h мережі (рис. 2) подається значення дескриптора h12 для цифри 2 (рядки 2, 3) (дескриптору h22 відповідають рядки 4, 5, а дескриптору h32 – рядки 6, 7), а на вхід і – одне з 1032 зображень з цифрою 2 з тестового набору. Цифра 5 у другому рядку стовпця з цифрою 2 означає, що для дескриптора h12 величина виходу нейронної мережі для 5 зображень цифри 2 з 1032 була менше 0.5, причому мінімальне значення виходу мережі було 0.0625 (з точністю до чотирьох знаків після коми). У стовпці для цифри 2 в рядках для виходу мережі у вказується мінімальне значення виходу мережі. У стовпцях для інших цифр в рядках для виходу мережі у вказується максимальне значення виходу мережі.

На рис. 4, а показано зображення рукописної цифри з тестового набору з міткою 2, відповідне для дескриптора h12 із значенням виходу мережі 0.12663102. На рис. 4, б показано зображення рукописної цифри з тестового набору з міткою 2, відповідне для дескриптора h38 із значенням виходу мережі 0.63038. При тестуванні 980 зображень з цифрою 0 з тестового набору кількість помилок дорівнює 13 і незмінна для дескрипторів h1, h2 та h3 (таблиця 3). При тестуванні 1032 зображень з цифрою 2 кількість помилок дорівнює 21 для дескрипторів h2 та 20 – для дескрипторів h1 і h3 (табл. 4).

Таблиця 4 – **Результати тестування на тестовому наборі зображень з цифрою 2 з використанням дескрипторів h1, h2 та h3 для цифр від 0 до 9**

Цифра	0	1	2	3	4	5	6	7	8	9
Помилок для h1	1	0	5	4	1	0	0	3	6	0
y	0.7878	0.4548	0.0625	0.9702	0.8674	0.0093	0.3104	0.9848	0.9970	0.0721
Помилок для h2	1	0	5	4	1	0	0	4	6	0
y	0.7843	0.4621	0.0647	0.9703	0.8749	0.0098	0.3196	0.9852	0.9971	0.0745
Помилок для h3	1	0	5	4	1	0	0	3	6	0
y	0.7715	0.4560	0.0618	0.9702	0.8806	0.0100	0.3066	0.9846	0.9972	0.0711



а б

Рис. 4. Зображення цифри з міткою 2
(а – для h12 y = 0.12663102; б – для h38 y = 0.63038)

Наведені в табл. 3 і 4 результати тестування показують, що запропоновані способи вибору еталонного дескриптора h незначно впливають на кількість помилок при тестуванні.

Підсумок

У роботі проаналізовано 3 способи побудови еталонних дескрипторів для порівняння і розпізнавання рукописних цифр для сіамської нейронної мережі. Архітектура сіамської нейронної мережі, що використана для тестування, складається із двох підмереж. На вхід першої підмережі подається зображення, яке аналізується, а її виходом є дескриптор цього зображення. Друга підмережа має два входи: на перший вхід подається еталонний дескриптор, а на другий вхід - дескриптор аналізованого зображення з виходу першої підмережі.

Тестування проводилося на зображеннях рукописних цифр з тестового набору MNIST для вихідної сіамської мережі [3]. Результати тестування показали, що запропоновані способи вибору еталонного дескриптора h незначно впливають на кількість помилок при тестуванні. При тестуванні 980 зображень з цифрою 0 кількість помилок склала 13 для всіх трьох способів побудови дескрипторів. При тестуванні 1032 зображень з цифрою 2 кількість помилок склала від 21 для дескрипторів h2 до 20 - для дескрипторів h1 та h3.

Подальші дослідження слід спрямувати на аналіз впливу розмірності дескриптора на точність порівняння зображень у сіамській нейронній мережі, а також на способи побудови дескрипторів при малому розмірі тренувального набору даних.

СПИСОК ЛІТЕРАТУРИ

1. Chicco D. Siamese Neural Networks: An Overview. Artificial Neural Networks. MIMB, vol. 2190, 2020, pp. 73-94.
2. Contrastive loss for Siamese networks with Keras and TensorFlow [Электронный ресурс]. – Режим доступа: https://www.pyimagesearch.com/2021/01/18/contrastive-loss-for-siamese-networks-with-keras-and-tensorflow/.
3. Image similarity estimation using a Siamese Network with a contrastive loss [Электронный ресурс]. – Режим доступа: https://keras.io/examples/vision/siamese_contrastive/.
4. Schroff F., Kalenichenko D., Philbin J. FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE CSC on CVPR, 2015, pp. 815-823.
5. The Mnist database of handwritten digits [Электронный ресурс]. – Режим доступа: http://yann.lecun.com/exdb/mnist/.
6. Банди Б. Методы оптимизации. Вводный курс. – М.: Радио и связь, 1988. – 128 с.

Received (Надійшла) 22.09.2021
Accepted for publication (Прийнята до друку) 27.10.2021

On the features of the formation of descriptors in the Siamese neural network

A. Shostak

Abstract. The subject of research − the processes of image recognition of handwritten numbers using neural networks. The recognition application is based on the architecture of the Siamese network with neural convolutional subnets. **The purpose of the article** is to substantiate the choice of N-dimensional vector representations of input images to describe their properties, compare and recognize them. **Objective**: experimental study of handwritten number image recognition using the architecture of the Siamese neural network. **Research methods**: direct search method for functions with several variables to determine N-dimensional vector representations of input images. **Research results**. The definition of N-dimensional vector representations of input images of handwritten numbers is carried out and their characteristics are investigated. An experimental study of image recognition using vector representations of images within the architecture of the Siamese neural network. It is shown that the proposed methods for determining vector N-dimensional representations of input images are robust and have little effect on the number of errors in recognition testing. Images of handwritten numbers from the MNIST test set were used during testing. It is determined that the use of preselected reference representations of the input images can simplify the architecture of the Siamese network. **Conclusions**. The results obtained in this work can be used in the comparative evaluation and determination of N-dimensional vector representations of different classes of input images in order to recognize them using the architecture of the Siamese neural network.

Keywords: siamese neural network, descriptor, neural network testing.