УДК 519.652:519.254

DOI: 10.18372/2073-4751.77.18655

Зівакін В.Д.,
orcid.org:0000-0002-0420-0558,
e-mail: zivakin1993@gmail.com,
Приставка П.О., д.т.н,
orcid.org/0000-0002-0360-2459,
e-mail: chindakor37@gmail.com

ВИКОРИСТАННЯ СПЛАЙН-МОДЕЛІ В ПРОСТОРІ ЛАТЕНТНИХ ПРЕДСТАВЛЕНЬ ПРИ ВИЛУЧЕННІ ДУБЛІКАТІВ ІЗ НАБОРУ СПОСТЕРЕЖЕНЬ

Національний авіаційний університет

Вступ

У цифрову епоху кількість згенерованих зображень зростає з неймовірною швидкістю. Соціальні медіа, веб-архіви та особисті колекції переповнені копіями зображень, які створюють не тільки питання зберігання, але й ускладнюють пошук та аналіз даних. Вилучення дублікатів із наборів зображень є критично важливим завданням для багатьох застосунків, включаючи цифрову архівацію, контент-аналіз та системи управління базами даних.

Традиційні методи, такі як порівняння хешів [1] або піксель-до-піксель аналіз [2], часто виявляються недостатньо ефективними при виявленні зображень, які були змінені або мають лише часткові співпадіння. Ці методи можуть пропускати неочевидні дублікати або помилково ідентифікувати унікальні зображення як дублікати через зовнішню схожість.

Новітні підходи, які використовують глибоке навчання і, зокрема, автоенкодери, дозволяють створити більш витончені та масштабовані системи для вирішення цієї проблеми. Автоенкодери, як описано в дослідженні [3], можуть ефективно вилучати суттєві характеристики зображень і перетворювати їх у латентні представлення, значно спрощуючи процес порівняння.

У даній статті ми пропонуємо сплайн-модель апроскимації латентних представлення, отриманих за допомогою автоенкодерів, та метод вилучення дублікатів, який грунтується на цій моделі та використанні метрики косинусної схожісті для ефективного порівняння зображень. Цей підхід не тільки покращує точність виявлення дублікатів, але й забезпечує масштабованість рішення, що є критично важливим для великих наборів даних.

Модель даних та Автоенкодер

Нехай маємо n – вимірний набір спостережень, кожне з яких належить до деякого класу, приналежність до котрого визначається міткою:

$$(x_l, y_l),$$

де $x_l -$ деяке цифрове зображення, $l = 1, N$, а $y_l = \{1 \dots k\}$, де $k \in Z$ – кількість класів, а $x_l \in \mathbb{R}^n$, де n – розмірність простору спостережень. Стверджуємо, що $x_l \in S_i$ тільки тоді, коли $y_l = i$, де $S_i = \{x_{l_i}, l_i = \overline{1, N_i}\}$ – деяка підмножина, така що $N = \sum_{i=1}^{k} N_i$ і $S_i \cap S_j = \emptyset, i \neq j$.

Автоенкодер – це тип нейронної мережі, який навчається кодувати вхідні дані в компактніші представлення і відновлювати їх назад до оригінального формату. Структура автоенкодера складається з двох основних частин: кодера/кодувальника (*encoder*) і декодера/декодувальника (*decoder*). Тобто кодувальник по суті відображає деяку величину у простір з іншою розмірністю, найчастіше – меншу, тобто стискає вхідні дані. Для цифрового зображення і автоенкодера із одним прихованим шаром процес відображення можна описати наступним чином:

$$t_l = W^1 x_l,$$

де z_l – результат відображення, а W^1 – матриця вагів прихованого шару. Для того

щоб збільшити нелінійність кінцевого перетворення достатньо додати додаткові приховані шари

Декодер, в свою чергу намагається відновити вихідні дані з цього стисненого представлення:

$$\widehat{x_l} = W^2 t_l = W^2 W^1 x_l = W x_l$$

Багатошарові згорткові автоенкодери є особливо ефективними для обробки зображень завдяки своїй архітектурі, яка дозволяє вловлювати просторові ієрархії характеристик. Їхні частини-кодувальники використовують згорткові шари (*convolutional layers*) для кодування вхідних зображень у латентні представлення. Згортки дозволяють мережі автоматично виявляти важливі характеристики на різних рівнях абстракції без втручання людини. Декодувальна частина мережі використовує транспоновані згорткові шари (*deconvolutional layers*), які розширюють латентні представлення назад до розмірів оригінального зображення.

Застосування усіх шарів кодувальника позначимо наступним чином:

$$t_l = Code(x_l), \qquad (1)$$

де *Code* – набір перетворень.

Так як цифрові зображення – це випадкова величина, а z – є результатом їхнього відображення в інший простір реальних чисел, тоді очевидним є існування функції розподілу такої випадкової величини:

$$F(t_1,\ldots,t_n) = P\{\omega: -\infty < \xi_1(\omega) < t_1,\ldots,-\infty < \xi_n(\omega) < t_n\}.$$

І при її неперервності існує функція щільності:

$$f(t_1,\ldots,t_n) = \frac{\partial^n F(t_1,\ldots,t_n)}{\partial t_1 \ldots \partial t_n}.$$

Для оцінки такого розподілу, про вигляд якого важко робити припущення, можна запропонувати використовувати непараметричні методи, наприклад гістограмну оцінку, яка суттєво залежить від вдалого вибору кроку розбиття. Для того, щоб забезпечити неперервність оціненої функції розподілу та її гладкість пропонується використовувати сплайн-оцінку. Слідуючи [6] маємо наступне.

Розглянемо забезпечення проведення ймовірнісної оцінки за масивом $\{t_l; l = \overline{1,N}\}$ реалізацій n-вимірної випадкової величини $\vec{\xi} = (\xi_1(\omega),\ldots,\xi_n(\omega))$ у латентному просторі авто кодувальника після перетворення (1). При умові незалежності одновимірних випадкових величин $\xi_k(\omega)$, $k = \overline{1,n}$ та зважаючи на можливість проведення ймовірнісної оцінки відповідних масивів реалізацій $\{t_{k,l}; l = \overline{1,N}\}$, шляхом уведення низки рівномірних розбиттів за вісями спостережень:

$$t_k : \Delta_{h_{t_k}},\ h_{t_k} > 0,\ k = \overline{1,n},$$

можемо розглядати й можливість ймовірнісної обробки n-вимірного варіаційного ряду, розбитого на класи, згідно рівномірного розбиття $\Delta_{h_{t_1},\ldots,h_{t_n}}$:

$$\left\{\left(t_{1,i_1},\ldots,t_{n,i_n}\right), n_{i_1\ldots i_n}, p_{i_1\ldots i_n}; i_k = \overline{0,m_k-1}, k = \overline{1,n}\right\},$$

де $(t_{1,i_1},\ldots,t_{n,i_n})$ – варіанта, яка визначає центральну (або мінімальну) точку $(i_1,\ldots,i_n)$-го елементу розбиття $\Delta_{h_{t_1},\ldots,h_{t_n}}$; m_k – кількість елементів (класів) розбиття $\Delta_{h_{t_1},\ldots,h_{t_n}}$ за напрямками t_k, $k = \overline{1,n}$; $n_{i_1,\ldots,i_n}$ – частота (кількість потраплянь точок з масиву $\Omega_{n,N}$ в $(i_1,\ldots,i_n)$-й елемент розбиття $\Delta_{h_{t_1},\ldots,h_{t_n}}$); $p_{i_1,\ldots,i_n}$ – випадковість варіанти n-вимірного варіаційного ряду:

$$p_{i_1,\ldots,i_n} = \frac{n_{i_1,\ldots,i_n}}{N},\ \sum_{i_1=0}^{m_1-1}\cdots\sum_{i_n=0}^{m_n-1} p_{i_1,\ldots,i_n} = 1,$$

причому

$$p_{i_1,\ldots,i_n} = P\{\omega: t_{1,i_1} - 0,5h_{t_1} \le \xi_1(\omega) < t_{1,i_1} + 0,5h_{t_1},\ldots,$$
$$\ldots t_{n,i_n} - 0,5h_{t_n} \le \xi_n(\omega) < t_{n,i_n} + 0,5h_{t_n}\} =$$
$$= f_{i_1,\ldots,i_n} h_{t_1} \cdot \ldots \cdot h_{t_n},$$

де $f_{i_1,\ldots,i_n}$ – усереднене значення щільності розподілу ймовірностей $f(t_1,\ldots,t_n)$ величини $\vec{\xi}$ на $(i_1,\ldots,i_n)$-му елементі розбиття $\Delta_{h_{t_1},\ldots,h_{t_n}}$:

$$f_{i_1,\ldots,i_n} =$$

$$= \frac{1}{h_{t_1} \cdot \ldots \cdot h_{t_n}} \int_{t_{1i_1}-0{,}5h_{t_1}}^{t_{1i_1}+0{,}5h_{t_1}} \ldots \int_{t_{1i_n}-0{,}5h_{t_1}}^{t_{1i_n}+0{,}5h_{t_1}} f(t_1,\ldots,t_n)\,dt_1 \ldots dt_n ,$$

окрім того виконується рівність

$$p_{i_1,\ldots,i_n} = p_{i_1} \cdot \ldots \cdot p_{i_n} .$$

Для непараметричної оцінки функції щільності розподілу (з точністю до константи $h_{t_1} \cdot \ldots \cdot h_{t_n}$) в просторі представлення автокодувальника пропонується така сплайн-модель на основі B-сплайнів другого порядку [6]:

$$\hat{f}(t_1,\ldots,t_n) = S_2(p,t_1,\ldots,t_n) = \qquad (2)$$

$$= \sum_{i\in Z} \ldots \sum_{j\in z} B_{r,h_{t_1}}\left(t_1 - ih_{t_1}\right) \ldots B_{r,h_{t_n}}\left(t_n - jh_{t_n}\right) p_{\underbrace{i,\ldots,j}_{n}} ,$$

де (з точністю до аргументу)

$$B_{2,h}(t) = \begin{cases} 0, & t \notin [-3h/2;3h/2], \\ (3+2t/h)^2/8, & t \in [-3h/2;-h/2], \\ 3/4-(2t/h)^2/4, & t \in [-h/2;h/2], \\ (3-2t/h)^2/8, & t \in [h/2;3h/2]. \end{cases}$$

Модель (2) є оцінкою щільності, що близька до інтерполяційної в середньому в асимптотичному сенсі [6]. Якщо в просторі представлення існує деяке афінне перетворення, наприклад за реалізації методу головних компонент, що забезпечує перехід до незалежної системи координат, то така модель може бути представлена як добуток оцінок щільності одновимірних маргінальних розподілів:

$$S_2(p,t_1,\ldots,t_n) = \hat{S}_2(t_1) \cdot \ldots \cdot \hat{S}_2(t_n),$$

де:

$$\hat{S}_2(t_k) = \sum_{i\in Z} B_{2,h_{t_k}}\left(t_k - ih_{t_k}\right) p_i .$$

Оцінки маргінальних розподілів $\hat{S}_2(t_k)$, $k = \overline{1,n}$ можутьмати використання для більш «гнучкого» дослідження особливостей набору зображень в латентному просторі автокодувальника. В [6] обґрунтовано, що, як в багатовимірному, так і в одновимірному випадку введена сплайн-молель має переваги перед параметричною оцінкою на основі нормального розподілу за рахунок того, що локальна апроксимація є більш гнучкою з позиції врахування локальних особливостей функції щільності.

Після застосування моделі (2) оцінки відкриваються широкі можливості по роботі із реалізацією випадкової величини у просторі представлення (1): проведення різноманітних перетворень, видалення аномальних значень, повторні способи кластеризації у випадку виявлення неоднорідності розподілу, тощо. Якщо ж здійснювати перехід до одновимірних маргінальних розподілів то можна проводити генерацію нових величин (зображень у нашому випадку), як описано в [7], а також вилучення дублікатів, як описано далі в даному викладені.

Косинусна схожість

Косинусна схожість – це метрика, яка вимірює косинус кута між двома векторами у багатовимірному просторі. Ця метрика використовується для визначення ступеня схожості між двома векторами на основі кута між ними, незалежно від їхньої довжини. Косинусна схожість визначається наступним рівнянням:

$$Sim(A,B) = \frac{A \bullet B}{\|A\| \bullet \|B\|},$$

де A, B – вектори, а $\|A\|, \|B\|$ – їхні норми.

Косинусна схожість має декілька ключових переваг, зокрема:

- Ігнорування розмірності векторів: Ця метрика вимірює лише напрямок векторів, ігноруючи їх масштаб. Це робить її ідеальною для порівняння векторів, де абсолютні значення компонентів не мають значення, як у випадку з текстовими даними або зображеннями.
- Ефективність в обчисленнях: Косинусна схожість зазвичай потребує менше обчислювальних ресурсів, особливо коли працює зі стислими або розрідженими даними.
- Робота з нормалізованими даними: Ця метрика особливо корисна, коли дані були нормалізовані або коли важливо

зменшити вплив різних масштабів характеристик.

Постановка задачі

Із урахуванням наведеної інформації, поставимо наступну задачу. Нехай маємо набір зображень, які належать до певних класів, розмірністю 64 на 64 пікселя. Класи можуть представляти різні категорії об'єктів (наприклад, автомобілі, рослини, цільові класи, тощо). Необхідно дослідити чи покращить новий метод вилучення дублікатів набір даних для подальшого навчання.

Для вирішення задачі скористаємося описаною вище моделлю (2) сплайн-апроксимції функції щільності випадкової величини у латентному просторі.

Інструменти для проведення експерименту

Набором даних для дослідження виступала частина датасету «Аерозйомка» [4], а для безпосереднього проведення досліджень, з використанням бібліотеки pytorch [5] був організований та навчений власний автокодувальник, архітектура якого представлена в табл. 1.

Таблиця 1. Структура автокодувальника

Шар	Вихідна форма даних	Кількість параметрів
Conv2d-1	32, 32, 32	896
ReLU-2	32, 32, 32	0
Conv2d-3	64, 16, 16	18,496
ReLU-4	64, 16, 16	0
Conv2d-5	128, 8, 8	73,856
ReLU-6	128, 8, 8	0
Conv2d-7	256, 4, 4	295,168
ReLU-8	256, 4, 4	0
Conv2d-9	256, 2, 2	590,08
ReLU-10	256, 2, 2	0
ConvTranspose2d-11	256, 4, 4	590,08
ReLU-12	256, 4, 4	0
ConvTranspose2d-13	128, 8, 8	295,04
ReLU-14	128, 8, 8	0
ConvTranspose2d-15	64, 16, 16	73,792
ReLU-16	64, 16, 16	0
ConvTranspose2d-17	32, 32, 32	18,464
ReLU-18	32, 32, 32	0
ConvTranspose2d-19	3, 64, 64	867
Tanh-20	3, 64, 64	0

Беззаперечним плюсом такої нейромережі, як аквтокодувальник є можливість використання вже натренованої моделі для отримання латентних представлень будь-яких даних. Головне, щоб навчальний набір був якісним. Після тренування, будь-яке зображення може бути представлене у формі латентного вектора, який отримується при роботі частини-кодувальника. Такий вектор може бути використаний для визначення ступеню схожості між зображеннями. Це робиться через вимірювання косинусної схожості між векторами, що дає можливість визначити, чи є два зображення дублікатами одне одного.

Для полегшення обчислень, а також розширення даних експерименту дублікати ідентифікуватимуться у кожному класі окремо. Для побудови матриць косинусної схожості з латентних представлень для кожного класу потрібно виконати наступні кроки:

- Обрахунок векторів: Спочатку потрібно отримати латентні вектори для кожного об'єкта у датасеті, використовуючи модель, таку як автоенкодер. При цьому зберегти мікти класів y_l.
- Для кожного класу виділити набір латентних векторів кількості N_c, такої що $N = \sum_{c=1}^{k} N_c$.
- Розрахунок косинусної схожості: Для кожної пари векторів в класі обраховується косинусна схожість згідно вищезазначеної формули.
- В результаті будується матриця косинусної схожості для кожного класу, де кожен елемент матриці відображає схожість між парою зображень. Великі значення в матриці позначають високу схожість, тоді як низькі значення вказують на низьку схожість (0 .. 1).
- Задається граничний рівень T, в порівнянні з яким приймається рішення про видалення: якщо $Sim(x_i, x_j) > T$, то x_i, x_j – дублікати.

Проведення експерименту та оцінка результатів

Для проведення дослідження було взято 33 класи набору даних «Аерозйомка» і визначено коефіцієнти ексцесів перших трьох головних компонент кожного класу до і після вилучення дублікатів.

Навіщо розглядати ексцес основних компонентів?

В разі зменшення ексцесу можна стверджувати про збільшення ентропії, що може позитивно сказатися при навчанні нейронних мереж. Адже більша ентропія зменшить імовірність перенавчання моделі.

Для більш детального дослідження було проведено 3 вилучення дублікатів із відповідними граничними рівнями $T = [0.93, 0.95, 0.97]$ і для кожного новоотриманого набору була проведена оцінка коефіцієнтів ексцесу по 3-х перших головних компонентах.

Для відображення залежності зміни ексцесу після видалення, для кожної головної компоненти була відтворена лінійна регресія методом Тейла (через стійкість до викидів), де на осі абсцис розташовано значення ексцесу по класах до видалення, а на осі ординат – після. Для прикладу наводяться графіки регресій для перших головних компонент на рисунках 1-3.

В табл. 2 та табл. 3 відповідно представлена кількість зображень, що залишається після процедури видалення із відповідним граничним рівнем та рівняння лінійної регресії кожної компоненти при різних граничних рівнях.

Таблиця 2. Зміна кількості зображень при видаленні дублікатів

***T* (граничний рівень)**	**Кількість**
Без видалення	128401
0.97	81018
0.95	64953
0.93	48607

Таблиця 3. Рівняння регресій для відповідних головних компонент.

***T* (граничний рівень)**	**Номер компоненти**	**Рівняння**
0.97	1	$y = 0.82x + 0.05$
0.97	2	$y = 0.78x + 0.08$
0.97	3	$y = 0.74x + 0.22$
0.95	1	$y = 0.51x + 0.10$
0.95	2	$y = 0.52x + 0.18$
0.95	3	$y = 0.61x + 0.43$
0.93	1	$y = 0.30x + 0.14$
0.93	2	$y = 0.30x + 0.26$
0.93	3	$y = 0.42x + 0.80$

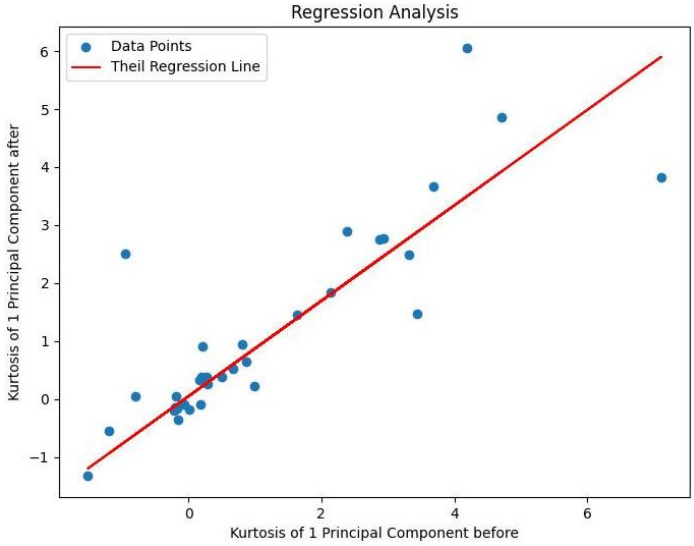

Рис. 1. Регресія ексцесу 1 ГК після видалення із $T = 0.97$

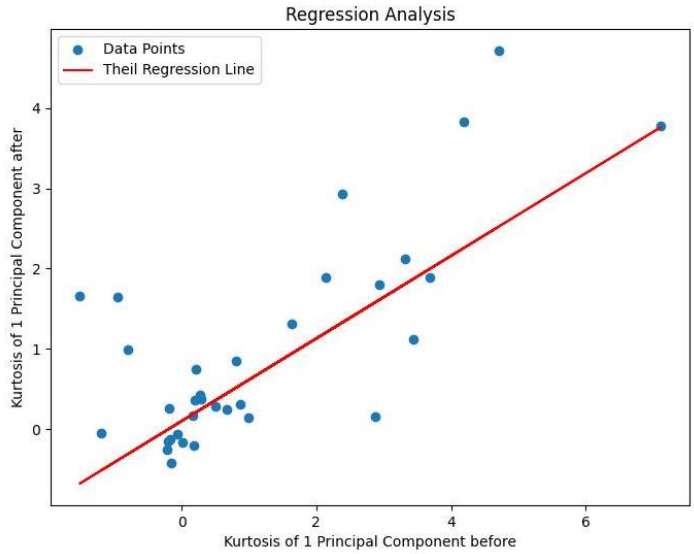

Рис. 2. Регресія ексцесу 1 ГК після видалення із $T = 0.95$

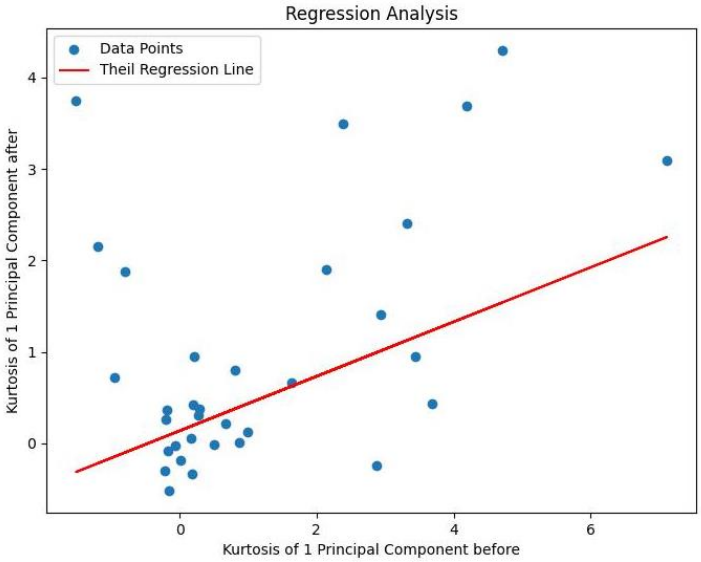

Рис. 3. Регресія ексцесу 1 ГК після видалення із $T = 0.93$

Висновки

В роботі введено до розгляду модель сплайн-оцінки розподілу випадкової величіни у просторі представлень (2) і на її основі наведено метод видалення дублікатів із навчального набору зображень.

З наведених даних видно, що після виконання процедури видалення не тільки зменшується загальна кількість зображень у датасеті (що, очевидно і має відбуватися), а й при кожному новому граничному рівні спостерігається зменшення ексцесу на рівні всього датасету, що підтверджується відтвореними регресіями при головних компонентах. Так, при T=0.97, коефіцієнт нахилу (*scope*) компонент знаходиться в межах 0.7-0.8. При T=0.95 вже у межах 0.5-06, тобто ексцес після видалення практично вдвічі менший ніж у початкових даних. І при T=0.93 нахил змінюється в межах 0.3-0.4. можна навіть сказати, що існує залежність між пороговим коефіцієнтом T та зміною ексцесів, а саме чим менший T, ти сильніше спадає ексцес, а тому збільшується ентропія. В подальшому передбачається вдосконалення методу шляхом динамічної зміни порогового рівня.

Результати досліджень засвідчують, що запропонований метод вилучення дублікатів зображень призводить до зменшення коефіцієнту ексцесу маргінальних розполідів у моделі (2). Це, у свою чергу, забезпечує збільшення ентропії таких розподілів. Тому можна стверджувати, що запропонований підхід забезпечує меншу схильність глибоких нейронних моделей, що навчаються на даних без дублікатів, до перенавчання.

Література

1. Liang X. et al. Robust Hashing with Local Tangent Space Alignment for Image Copy Detection. *IEEE Transactions on Dependable and Secure Computing*. 2023. P. 1–13. DOI: 10.1109/TDSC.2023.3307403.

2. Zhu C., Zhou Y., Xie Z. A Pixel-to-Pixel Convolutional Neural Network for Single Image Dehazing. *Lecture Notes in Computer Science. Vol. 10636. Neural Information Processing. 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017. Proceedings, Part III.* / ed. by D. Liu et al. Cham, 2017. P. 270–279. DOI: 10.1007/978-3-319-70090-8_28.

3. Masci J. et al. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. *Lecture Notes in Computer Science. Vol. 6791. Artificial Neural Networks and Machine Learning - ICANN 2011. 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011. Proceedings, Part I* / ed. by T. Honkela et al. Berlin, 2011. P. 52–59. DOI: 10.1007/978-3-642-21735-7_7.

4. Zivakin V. et al. Training set AERIAL SURVEY for data recognition systems from aerial surveillance cameras. *IX International Scientific Conference "Information Technology and Implementation" (IT&I-2022)* : proceedings, Kyiv, Ukraine, November 30 – December 02, 2022 / Taras Shevchenko National University of Kyiv. 2023. P. 246–255. URL: https://ceur-ws.org/Vol-3347/Paper_21.pdf.

5. PyTorch. PyTorch: An open source machine learning framework that accelerates the path from research prototyping to production deployment. URL: https://pytorch.org.

6. Приставка П. О. Поліноміальні сплайни при обробці даних : монографія. Д. : Вид-во Дніпропетр. ун-ту, 2004. С. 155–164.

7. Зівакін В. Дослідження імітації одновимірних вибірок із використанням поліноміальних сплайнів; *Таврійський науковий вісник. Серія: Технічні науки*. 2021. Вип. 6. С. 23–30.

Зівакін В.Д., Приставка П.О.

ВИКОРИСТАННЯ СПЛАЙН-МОДЕЛІ В ПРОСТОРІ ЛАТЕНТНИХ ПРЕДСТАВЛЕНЬ ПРИ ВИЛУЧЕННІ ДУБЛІКАТІВ ІЗ НАБОРУ СПОСТЕРЕЖЕНЬ

В роботі вперше запропоновано використання сплайн-моделі на основі локальних B-сплайнів другого порядку для оцінки щільності розподілу навчального набору цифрових зображень в латентному просторі багатошарового нелінійного авто кодувальника. На основі моделі запропоновано метод видалення дублікатів зображень у латентному просторі представлення мережі-автокодувальника. Проведено дослідження та статистично доведено збільшення ентропії розподілів даних, що сприяє меншому перенавчанню нейронних моделей. Дослідження зосереджено на вивченні цифрових зображень за допомогою багатошарового нелінійного автокодувальника, інструменту глибинного навчання, що дозволяє здійснювати зниження розмірності та витягування корисної інформації з вхідних даних. Розроблена сплайн-модель надає нові можливості для оцінювання і візуалізації розподілів, що може бути корисним для подальших аналітичних досліджень у сфері обробки зображень.

Основний фокус роботи сконцентрований на методі видалення дублікатів зображень у латентному просторі, який використовує дані про щільності розподілів, отримані зі сплайн-моделі. Це дозволяє не тільки очистити набір даних від повторюваних зразків, але й оптимізувати процес навчання нейронних мереж, зменшуючи перенавчання та підвищуючи загальну ефективність моделей.

***Ключові слова**: сплайн-модель; локальні B-сплайни; латентний простір; авто кодувальник; видалення дублікатів зображень; зменшення перенавчання; ентропія розподілів; глибинне навчання; статистичне дослідження; щільність розподілу; нейронні мережі.*

Zivakin V.D., Prystavka P.O.

USING A SPLINE MODEL IN THE SPACE OF LATENT REPRESENTATIONS WHEN REMOVING DUPLICATES FROM A SET OF OBSERVATIONS

In the paper, for the first time, the use of a spline model based on local B-splines of the second order is proposed to estimate the density of the distribution of a training set of digital images in the latent space of a multilayer nonlinear auto encoder. Based on the model, a method for removing duplicate images in the latent space of the autoencoder network representation is proposed. Research has been conducted and statistically proven to increase the entropy of data distributions, which contributes to less retraining of neural models. The research focuses on learning digital images using a multi-layer nonlinear autoencoder, a deep learning tool that allows for dimensionality reduction and extraction of useful information from input data. The developed spline model provides new opportunities for estimating and visualizing distributions, which may be useful for further analytical research in the field of image processing.

The main focus of the work is concentrated on the method of removing duplicate images in the latent space, which uses data on the density of distributions obtained from the spline model. This allows not only to clean the data set from repeated samples, but also to optimize the learning process of neural networks, reducing overtraining and increasing the overall efficiency of the models.

***Keywords**: spline model; local B-splines; latent space; auto encoder; removal of duplicate images; reduction of retraining; entropy of distributions; deep learning; statistical research; distribution density; neural networks.*