УДК 004.932.4

Б.П. РУСИН
Фізико-механічний інститут ім. Г.В. Карпенка НАН України, Львів, Україна, e-mail: *b.rusyn.prof@gmail.com*.

О.А. ЛУЦИК
Фізико-механічний інститут ім. Г.В. Карпенка НАН України, Львів, Україна, e-mail: *olutsyk@yahoo.com.*

Р.Я. КОСАРЕВИЧ
Фізико-механічний інститут ім. Г.В. Карпенка НАН України, Львів, Україна, e-mail: *kosar2311@gmail.com*.

ОЦІНЮВАННЯ ІНФОРМАТИВНОСТІ НАВЧАЛЬНОЇ ВИБІРКИ ДЛЯ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ МЕТОДАМИ ГЛИБОКОГО НАВЧАННЯ

Анотація. Запропоновано новий підхід до оцінювання інформативності навчальної вибірки під час розпізнавання зображень, отриманих засобами дистанційного зондування. Показано, що якість навчальної вибірки можна відобразити набором характеристик, кожна з яких описує певні властивості даних. Встановлено залежність між характеристиками навчальної вибірки та точністю роботи класифікатора, тренованого на основі цієї вибірки. Розроблений підхід застосовано до різних тестових навчальних вибірок та наведено результати їхнього оцінювання. Показано, що оцінювання навчальної вибірки з використанням нового підходу здійснюється значно швидше, ніж процес навчання нейронної мережі. Це надає змогу застосовувати запропонований підхід до попереднього оцінювання навчальної вибірки в задачах розпізнавання зображень методами глибокого навчання.

Ключові слова: глибоке навчання, виділення ознак, навчальна вибірка, згорткова нейромережа.

ВСТУП

Машинне навчання набуло широкого використання в переважній більшості інтелектуальних інформаційних систем, які застосовують у різних галузях людської діяльності. Всюди, де є потреба в елементах штучного інтелекту, машинне навчання та нейромережеві методи витіснили традиційні підходи, що ґрунтуються на логічних чи евристичних алгоритмах прийняття рішення. З одного боку, це зумовлено достатньо швидким розвитком обчислювальних засобів, а з іншого — наявністю великої кількості даних для машинного навчання. Проте є обставини, які ускладнюють або унеможливлюють ефективне використання інтелектуальних систем цього типу. Однією з основних причин, яка стає на заваді їхньому ефективному використанню, є те, що отримати таку репрезентативну навчальну вибірку, для якої розпізнавання зображень здійснюватиметься з високою достовірністю, досить складно.

У задачах машинного навчання є залежність між кількістю даних, які використовуються для навчання моделі, та подальшою точністю її роботи на тестових та валідаційних даних. Найчастіше це проявляється у проблемі нестачі даних для створення якісної навчальної вибірки [1]. Репрезентативна навчальна вибірка найбільшою мірою забезпечує правильне навчання моделі під час класифікації.

На сьогодні немає універсального підходу, який би дав однозначну відповідь на таке запитання: скільки потрібно мати даних для навчання конкретної моделі з передбачуваною точністю роботи.

Зазвичай дані в навчальній вибірці формують вектор ознак заданої довжини. У цьому випадку, оперуючи статистичними підходами та елементами кластерно-

го аналізу, можна приблизно оцінити інформативність навчальної вибірки [2]. Іншим підходом до розв'язання цієї задачі є використання узагальнених динамічних моделей [3]. Є низка робіт щодо оцінювання даних у випадку використання класичних моделей для розпізнавання [4, 5]. Проте ці підходи застосовують для роботи у просторі ознак [6]. Результат оцінювання інформативності вибірки значною мірою залежить від процесу пошуку ознак, який часто має евристичний характер та не піддається строгому обґрунтуванню.

У свою чергу, є багато прикладних задач, де об'єктом аналізу виступають зображення, наприклад у разі здійснення дистанційного зондування, неруйнівного контролю стану поверхні, біометричної аутентифікації [7–10]. Інформація у вигляді зображень є більш складною для узагальнення та потребує здійснення таких проміжних етапів оброблення, як покращення зображень шляхом попереднього оброблення (усунення завад та різного роду спотворень з використанням фільтрації, сегментації, контрастування, бінаризації) та виділення системи інформативних ознак. Без виконання цих проміжних етапів оброблення зображень складно оцінити покриття класів в ознаковому просторі і, як наслідок, зробити припущення про достатність та інформативність навчальної вибірки для тестування конкретної моделі. Частковим розв'язанням цієї проблеми є використання згорткових шарів нейронної мережі, що узагальнює вибір ознак на зображенні і зводить її до процесу навчання [11, 12]. Основним інструментом встановлення репрезентативності навчальної вибірки у випадку глибокого навчання протягом тривалого часу слугували так звані криві навчання. З їхньою допомогою можна оцінити інформативність навчальної вибірки не безпосередньо, а через її вплив на конкретну досліджувану модель [13]. Проте найбільш точний підхід до оцінювання інформативності навчальної вибірки ґрунтується на методах, в яких поєднано дослідження навченої моделі з апостеріорними даними [14]. Цей підхід у багатьох роботах застосовують як еталонний, оскільки він має високу достовірність. Його недоліком вважають високу обчислювальну складність, що накладає обмеження на оперативність отримання оцінки. До прикладу, якість зображень істотно впливає на інформативність навчальної вибірки, яка формується на основі цих зображень [15].

Задачі дистанційного зондування пов'язані з отриманням та аналізом великих обсягів даних. Здебільшого ці дані представлені у вигляді матриць зображень у градації сірого або мультиспектральних зображень. Для розв'язання прикладних задач дистанційного зондування потрібно розробити автоматизовані методи класифікації, розпізнавання, пошуку, детектування об'єктів інтересу. Це зумовило появу великих баз даних, які можна застосувати для глибокого навчання [16, 17]. Навчання нейронних мереж є обчислювально складним процесом, який не завжди гарантує бажаний результат. Згідно з дослідженнями процес коректного навчання нейронної мережі істотно залежить від властивостей навчальної вибірки, а саме від того, наскільки вдало вона представляє сукупність дискримінаційних ознак зображень.

Експерименти з навчальними вибірками свідчать про те, що немає сенсу у великій кількості даних, якщо їхня якість є поганою. Під якістю даних будемо розуміти комплексну характеристику, що описує властивості даних, які виконують поставлену задачу, а саме: компактність представлення, розбалансування, узгодженість класів, відхилення класу в межах вибірки. Поняття якості вибірки є абстрактним, тому будемо використовувати інформативність навчальної вибірки.

З огляду на це, виникає потреба у здійсненні попереднього оцінювання навчальної вибірки, тому потрібно розробити підходи до оцінювання інформативності навчальної вибірки [18]. Відповідно, запропоновано підхід, який передбачає попереднє оцінювання інформативності навчальної вибірки на основі сукуп-

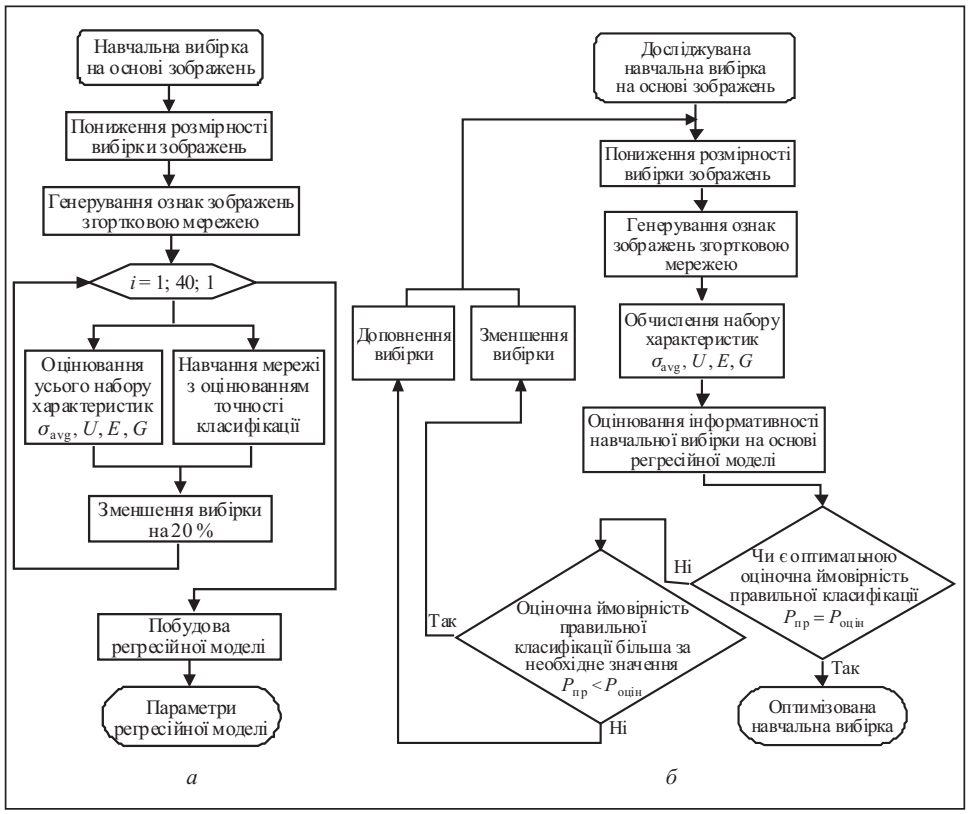

Рис. 1. Блок-схема запропонованого підходу: побудова регресійної моделі (*а*); попереднє оцінювання навчальної вибірки (*б*)

ності зазначених складових характеристик. Є два випадки, в яких попереднє оцінювання інформативності вибірки може дати ефект:

1. Якщо навчальна вибірка є мало інформативною, то тренування моделі не дасть очікуваного результату. Тому таку вибірку необхідно доповнити у такий спосіб, щоб вона відповідала комплексній характеристиці інформативності.

2. Якщо навчальна вибірка є надлишковою, тоді процес навчання моделі є обчислювально складним. Зменшення її надлишковості суттєво не вплине на точність навчання, але може сприяти значному скороченню часу навчання.

Цей підхід надає змогу здійснити попереднє оцінювання властивостей навчальної вибірки ще до навчання нейронної мережі на цій вибірці. З одного боку, при цьому можна оптимізувати загальний час, витрачений на створення навчальної вибірки, її формування та оцінювання, а з іншого — можна автоматизувати процес створення інформативної навчальної вибірки для обчислювально складних моделей класифікації.

Відповідно до запропонованого підходу оцінювання навчальної вибірки здійснюють на основі всіх даних вибірки. Після цього навчальну вибірку поступово проріджують і знову оцінюють. У найпростішому випадку прорідження вибірки реалізують у випадковий спосіб. За наявності додаткової інформації можна вилучати надлишкові дані та ті, що додають шум у процес навчання. Зменшення навчальної вибірки є виправданим тільки до тієї межі, коли починає знижуватись її репрезентативність. Цю межу встановлює запропонований підхід. Його блок-схему наведено на рис. 1.

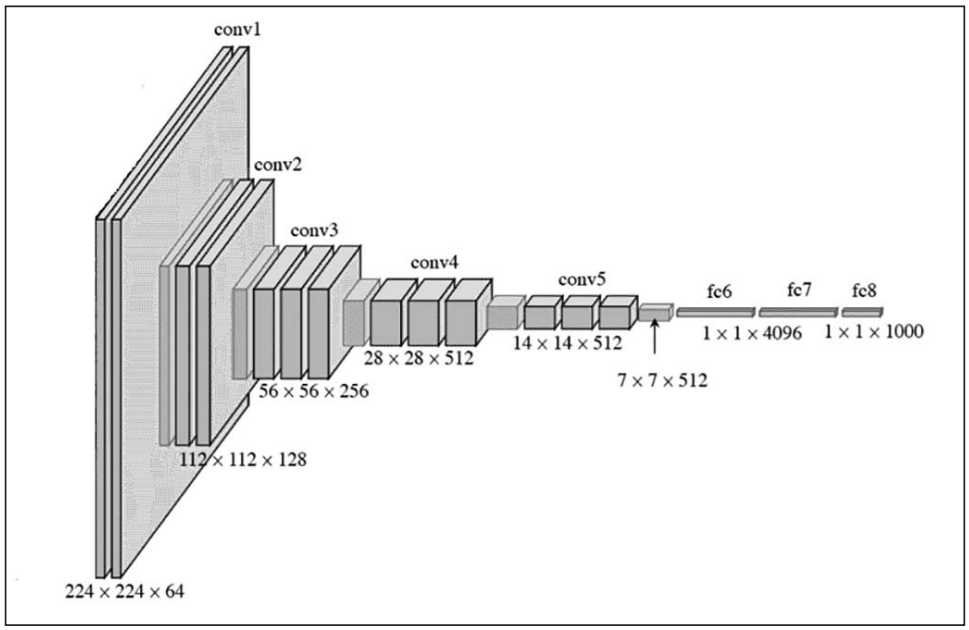

Рис. 2. Архітектура згорткової мережі VGG

ПОНИЖЕННЯ РОЗМІРНОСТІ ДАНИХ

Упродовж тривалого часу для генерування ознак зображень застосовували евристичні підходи [19]. Пізніше хороший результат отримали з використанням методу головних компонент та вейвлетів [20]. В описаному в цій статті підході до оцінювання інформативності бази даних зображень дистанційного зондування як генератор ознак запропоновано використати багатошарову згорткову мережу. Шляхом тренування згорткову мережу навчають генерувати ознаки зображень. Архітектура та кількість шарів мережі залежить від типу та кількості зображень. У цьому випадку застосовано архітектуру VGG, наведену на рис. 2.

Ця архітектура добре підходить до розв'язання задач класифікації зображень дистанційного зондування, оскільки вона має просту рівномірну структуру, що складається з послідовно упорядкованих згорткових та об'єднувальних шарів, за яких глибина моделі становить 16 шарів. Робота мережі в режимі генератора ознак є збалансованою, а фільтри ефективно захоплюють корисні ознаки [21, 22]. Ще однією перевагою застосування цієї структури є те, що можна провести навчання лише один раз для достатньо великої вибірки з метою тренування згорткових фільтрів та застосувати обчислені коефіцієнти без потреби у перенавчанні.

Згорткові шари виводять дані активації у вигляді тривимірного масиву значень, де зріз за третьою координатою відповідає фільтру, який слугує входом наступного шару. Інформація, що виводиться повноз'вязними шарами на виході мережі, є комбінацією характеристик, які шляхом тренування згенерував попередній шар мережі. Тоді згортку можна записати у такому вигляді:

$$\text{conv}\,(a^{l-1}, Z^n)_{x,y} = \psi^l\left(\sum_{i=1}^{n_H^{l-1}}\sum_{j=1}^{n_W^{l-1}}\sum_{k=1}^{n_C^{l-1}} Z^n_{i,j,k}\, a^{l-1}_{x+i-1,\,y+j-1,k} + b^l_n\right),$$

де n_H, n_W — висота і ширина зображення; n_C — кількість каналів зображення; Z — фільтр, який у цьому випадку має квадратну форму; a — розмір конкретного шару мережі; b — початковий поріг; l — розмірність.

Представимо отриманий шар мережі у вигляді тензора

$$\dim\,(\mathrm{conv}\,(a^{l-1}, Z^n)) = (n_H^l, n_W^l).$$

Попереднє виокремлення ознак зображень зумовлено потребою в оптимізації великої кількості даних, що містяться в базі зображень. Тоді за допомогою об'єднання інформативних характеристик зображення можна підвищити дискримінаційні властивості ознак в ознаковому просторі.

Розглянемо процедуру пониження розмірності даних на прикладі опрацювання бази даних, що містить 92000 зображень, отриманих засобами дистанційного зондування. Початкову базу даних позначимо $X = \{x_1, x_2, \ldots, x_n\}$, де $x_1, x_2, \ldots, x_n$ — вектор ознак окремих зображень довжиною 4096 відліків. Тоді базу даних пониженої розмірності можна записати у вигляді $M = \{m_1, m_2, \ldots, m_k\}$. Для подальшого оцінювання даних з меншими обчислювальними витратами потрібно понизити розмірність бази даних. Метою зменшення розмірності є збереження структури даних високої розмірності у просторі з пониженою розмірністю. Одним з класичних підходів пониження розмірності є аналіз головних компонент (PCA). Він є лінійним методом і забезпечує представлення низької розмірності для даних, які відрізняються між собою. Під час проведеного дослідження було виявлено, що класичні підходи не можуть повною мірою відобразити локальну та глобальну структуру даних. Тому запропоновано застосовувати ймовірнісні підходи до пониження даних, що ґрунтуються на відстані Кульбака–Лейблера (КЛ) [23].

Як міру подібності між векторами ознак, що відповідають за окремі зображення x_i та x_j, приймемо умовну ймовірність $p_{i|j}$. Для x_i та x_j, що лежать близько один від одного в гіперпросторі, умовна ймовірність $p_{i|j}$ набуватиме великих значень і матиме вигляд

$$p_{i|j} = \frac{\exp\left(\dfrac{-||x_j - x_i||^2}{2\sigma_j^2}\right)}{\sum\limits_n \exp\left(\dfrac{-||x_j - x_n||^2}{2\sigma_j^2}\right)},$$

де σ_j^2 — дисперсія, локалізована в околі x_j.

Такий самий вираз умовної ймовірності можна записати для даних m_i та m_j у пониженій розмірності: $q_{i|j} = \dfrac{\exp\,(-||m_j - m_i||^2)}{\sum\limits_k \exp\,(-||m_j - m_k||^2)}$.

Якщо база даних пониженої розмірності M правильно відображає початкову базу даних X, то умовні ймовірності $p_{i|j}$ і $q_{i|j}$ будуть сумірними. У цьому випадку задача зводиться до знаходження представлення низької розмірності для даних за умови мінімізації відмінностей між $p_{i|j}$ і $q_{i|j}$. Мірою, яка вказує на те, наскільки один розподіл відрізняється від іншого, є відстань КЛ. У нашому випадку розподіли ймовірностей P і Q є дискретними і визначеними в одному й тому самому ймовірнісному просторі. Тоді цільова функція набуде такого вигляду:

$$\sum_j D_{KL}(P_i||Q_i) = \sum_j \sum_i p_{i|j} \log \frac{p_{i|j}}{q_{i|j}}.$$

Цей вираз можна записати так:

$$\sum_j D_{KL}(P_i||Q_i) = \sum_j \sum_i p_{i|j} \log p_{i|j} - p_{i|j} \log q_{i|j}.$$

Його мінімізацію здійснюємо за m_i з використанням чисельних градієнтних методів, а саме методу градієнтного спуску. Його потрібно ініціалізувати початковими значеннями. У цьому випадку вибирають випадкові значення з околу початку координат.

За допомогою описаного підходу можна понизити розмірності даних на основі встановленої умовної ймовірності попарних метричних відстаней. Це надає змогу перейти до формування ознак інформативності бази даних із значно меншими обчислювальними витратами та компактнішим представленням ознакового простору.

ФОРМУВАННЯ ІНФОРМАТИВНИХ ХАРАКТЕРИСТИК ВИБІРКИ

Інформативною навчальною вибіркою для глибокого навчання вважають набір даних, у разі використання якого мережа у результаті навчання забезпечує бажаний рівень розпізнавання чи класифікації. Ці характеристики навчальної вибірки закладені у структуру даних. До особливостей зазначеної структури можна віднести кластеризаційні властивості, перекриття та розбалансування класів [21, 23], розрідженість даних [24], компактність представлення класів, репрезентативність. Інформативним параметром вибірки є її повнота. Характерною ознакою для повноти є усереднене відхилення класу в межах вибірки, яке має вигляд

$$\sigma_{\text{avg}} = \exp\left(\sum_{c}^{N} \left(\frac{1}{N} - \frac{K_c}{K} \right)^2 \right),$$

де N — загальна кількість класів, K_c — кількість компонент, що належать окремому класу, c — порядковий номер конкретного класу, K — загальна кількість компонент. Бажано насамперед перевірити найважливіші дані, оскільки залежність повноти від маловажних даних є неістотною.

Нерівномірність або розбалансування навчальної вибірки за класами визначимо показником:

$$U = \frac{1}{(N - K_c)^2} \sum_{c}^{N} \left(K_c - \frac{K}{N} \right).$$

Якщо показник нерівномірності U демонструє велике розбалансування, це означає, що якийсь окремий клас представлений значно меншою кількістю векторів ознак. Тому в процесі навчання нейронної мережі ним можна знехтувати, або трактувати його як шуми.

Внутрішньою ознакою навчальної вибірки є рівномірність покриття векторів ознак в ознаковому гіперпросторі, яка має такий вигляд:

$$E = \frac{1}{K^2} \sum_{i}^{K} \sum_{d=1}^{L} \sum_{c=1}^{N} \exp\left(\left(x_d - \frac{1}{2N} (2r-1)(\max(x_{i,c}) - \min(x_{i,c}) \right)^2 \right),$$

де r — коефіцієнт покриття. На практиці встановлено, що рівномірність покриття векторів ознак в ознаковому гіперпросторі має обернено пропорційний зв'язок із кластеризаційними властивостями навчальної вибірки в цілому. Розглянемо наступну інформативну характеристику вибірки, а саме компактність представлення даних у просторі ознак G. Ця ознака є прямо пропорційною до кластеризаційних властивостей навчальної вибірки і слугує досить точним маркером, що вказує на простоту побудови класифікатора у процесі навчання нейронної мережі:

$$G = 1 - \frac{\sum_{c=1}^{N} \sum_{d=1}^{N} \sum_{i=1}^{K} (x_{i,c} - x_{i,d})}{(K_c^2 - K_c) \sum_{i=1}^{N} (\max_{c}(x_{i,c}) - \min_{c}(x_{i,c}))^2}.$$

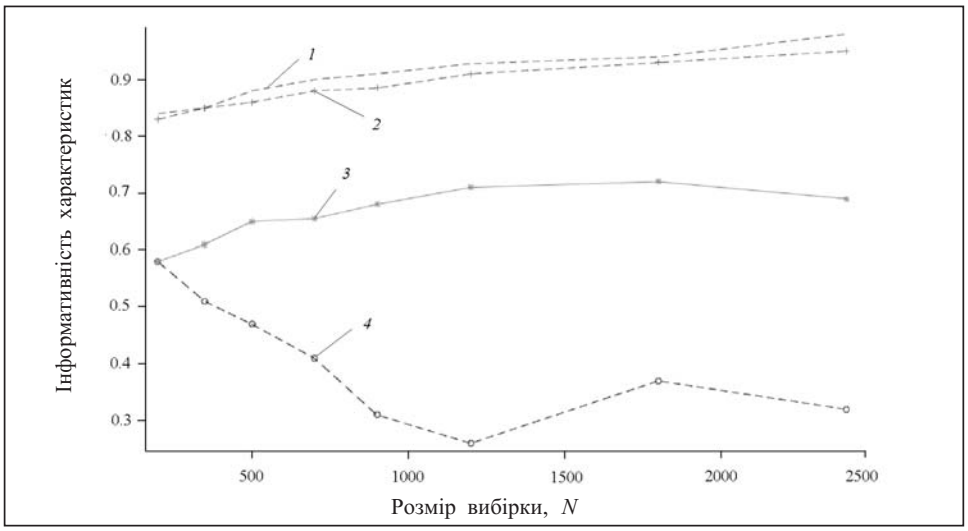

Рис. 3. Залежність запропонованих характеристик σ_{avg} (крива *1*); *U* (крива *2*); *G* (крива *3*); *E* (крива *4*); від розміру навчальної вибірки

Встановлено, що наявність дубльованих даних N_d чинить негативний вплив на процес навчання і згодом призводить до розбалансування. Прикладом цього є наявність у навчальній вибірці декількох записів одного і того самого вектора ознак, а також випадки, коли вектори ознаки лежать дуже близько один від одного. Тому, якщо це виявиться під час попереднього оцінювання навчальної вибірки, доцільно локалізувати та вилучити відповідні дані.

Формування ознак, інваріантних до конкретної навчальної вибірки, надасть змогу скористатися набором цих ознак для попереднього оцінювання інформативності навчальної вибірки, не виконуючи обчислювально складного процесу навчання нейронної мережі. До того ж, завдяки цьому підходу можна скоригувати наявну навчальну вибірку, суттєво скоротивши її. У свою чергу це дасть можливість зменшити час навчання мережі. Найпростішим варіантом пониження розмірності навчальної вибірки є вилучення даних у випадковий спосіб після кожного кроку перевірки її на репрезентативність. Складнішим, але точнішим варіантом пониження розмірності навчальної вибірки є видалення надлишкових даних, які вносять шум у процес навчання.

На рис. 3 наведено залежність характеристик інформативності від розміру навчальної вибірки. Навчальну вибірку створено на основі синоптичної бази дослідження хмарності зображень. Як видно з рис. 3, три з чотирьох кривих зростають у разі збільшення кількості елементів навчальної вибірки. Цю властивість використовують під час побудови навчальної вибірки. Вона надає змогу встановити поріг, за якого ефективність навчання нейронної мережі залишається на бажаному рівні.

АНАЛІЗ ТА УЗАГАЛЬНЕННЯ ХАРАКТЕРИСТИК НАВЧАЛЬНОЇ ВИБІРКИ

Сформовані характеристики для оцінки навчальної вибірки забезпечують достовірне представлення про структуру даних у ній. Проте метою є визначення інтегральної характеристики. Іншими словами, потрібно з'ясувати, якою є інформативність навчальної вибірки. У практичний спосіб було встановлено залежність між описаним у попередньому розділі набором характеристик, отриманих із навчальної вибірки, та ймовірністю розпізнавання моделі глибокого навчання, тренованої на тій самій навчальній вибірці. На основі цього

зроблено припущення про відповідність набору характеристик навчальної вибірки можливій імовірності розпізнавання, якої може досягти модель глибокого навчання в результаті навчання на цій самій вибірці.

Один із варіантів встановлення залежності між характеристиками навчальної вибірки та ймовірністю навчання полягає у використанні лінійної регресії. У цьому випадку скористаємося множинною лінійною регресією в багатовимірному просторі, оскільки маємо набір декількох характеристик. Це дасть можливість отримати єдину величину, яка з деякою ймовірністю буде описувати інформативність конкретної вибірки.

Для цього з усієї бази, що містить 92000 зображень, формуємо вторинну сукупність вибірок шляхом випадкового вилучення зображень, причому кожна наступна навчальна підвибірка скорочується за розміром відносно попередньої на 20 %. Після отримання 40 підвибірок для кожної з них виконують обчислення всього набору характеристик та здійснюють навчання мережі з оцінюванням точності класифікації. При цьому точність класифікації кожної із 40 нейромережевих моделей лежить у межах 54–93 % і залежить від конкретних характеристик даних, сформованих у результаті скорочення навчальної вибірки. Це роблять для того, щоб отримати регресійну модель, яка встановить відповідність між інформативністю навчальної вибірки і точністю класифікації моделі на її основі.

У цьому випадку рівняння множинної лінійної регресії має вигляд

$$Y_r = d_0 + d_1 J_1 + d_2 J_2 + \ldots + d_n J_n,$$

де J_i — i-та характеристика вибірки; d_i — i-й ваговий регресійний коефіцієнт, обчислюваний методом найменших квадратів.

Регресійна модель лінійного типу досить добре описує характеристики навчальної вибірки. Тому можна використовувати цей підхід для попереднього оцінювання інформативності довільної навчальної вибірки.

На рис. 4. наведено результат правильної класифікації залежно від розміру навчальної вибірки для трьох моделей глибокого навчання, а саме: тришарової мережі, мережі типу ResNet та ансамблю з CNN мереж. Навчальну вибірку створено на основі синоптичної бази зображень дослідження хмарності.

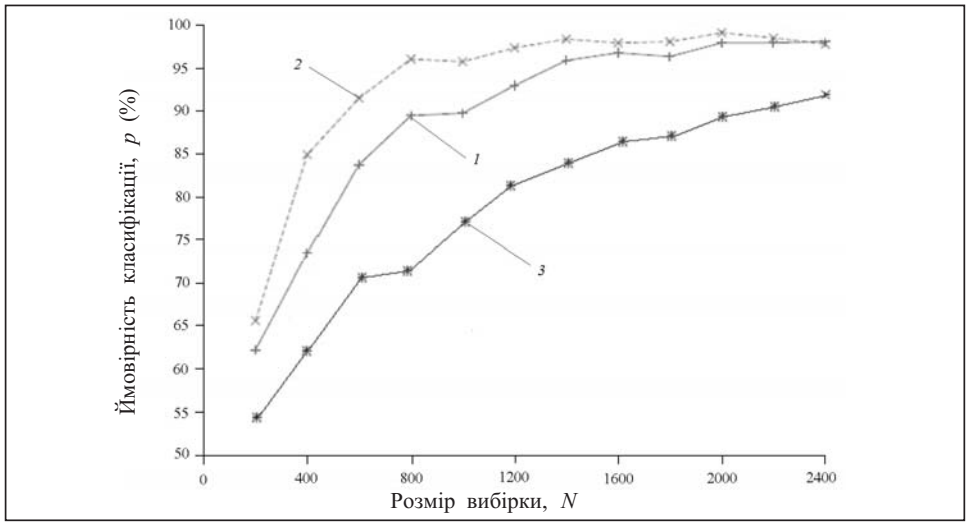

Рис. 4. Залежність ймовірності правильної класифікації від розміру навчальної вибірки: мережа типу RESNET (крива *1*); ансамбль з CNN мереж (крива *2*); тришарова мережа (крива *3*)

З рис. 4. можна дійти висновку, що різні моделі глибокого навчання демонструють різні властивості класифікації навіть для однієї навчальної вибірки.

Для перевірки запропонованого підходу використано базу зображень для дослідження хмарності, повний розмір якої становить 2400 зображень. Перевірка полягає в одночасному застосуванні регресійної моделі до оцінювання прогнозованого результату точності класифікації та повного циклу з навчанням моделі глибокого навчання й оцінюванням реальної точності. Результати перевірки наведено в табл. 1.

Таблиця 1

Розмір навчальної вибірки	Точність правильної класифікації на навчальній вибірці розміром 2400 зображень (%)	
	Запропонований підхід	Наявний підхід
2400	94	92
1800	93	86
1200	85	81
900	78	74
700	71	72
500	67	63
350	61	56
200	55	54

Як видно з табл. 1, результати оцінювання за допомогою запропонованого підходу є сумірними з результатами роботи реальної нейронної мережі, навченої на тестовій навчальній вибірці. Деяка розбіжність результатів пов'язана з тим, що регресійна лінійна модель, використана для оцінювання, була створена на основі результатів імовірностей класифікації за іншою моделлю глибокого навчання з відмінною структурою та кількістю внутрішніх параметрів.

Окремі моделі глибокого навчання дають різні результати ймовірностей правильної класифікації за умови навчання на тій самій навчальні вибірці. У свою чергу, запропонований підхід ґрунтується на результатах моделі, взятої як еталон. У табл. 2 наведено результати порівняння точності правильної класифікації таких моделей глибокого навчання, як VGG16, Alex Net, GoogLe Net та запропонованого підходу. Навчання всіх моделей здійснено на базі зображень Cifar-10 з різним розміром навчальних вибірок. Результати підтверджують ефективність запропонованого підходу, який дає можливість виконати попереднє оцінювання навчальної вибірки та виявити момент, коли її подальше збільшення не спричиняє підвищення результатів правильної класифікації.

У табл. 3 наведено результати запропонованого підходу для порівняння з реальними результатами розпізнавання для навчальної вибірки MNIST. База даних MNIST є стандартом, запропонованим Національним інститутом стандартів і тех-

Таблиця 2

Розмір навчальної вибірки	Точність правильної класифікації для різних моделей, тренованих на основі навчальної вибірки Cifar-10 (%)			
	Запропонований підхід	VGG16	Alex Net	GoogLe Net
40000	93	92	84	98
35000	91	92	83	98
30000	88	92	83	98
25000	84	91	81	97
20000	79	90	78	97
15000	75	84	76	95
10000	70	81	63	93
5000	64	72	61	85

Таблиця 3

Розмір навчальної вибірки	Точність правильної класифікації моделі на навчальній вибірці MNIST (%)	
	Запропонований підхід	Наявний підхід
60000	97	94
30000	95	93
15000	91	88
10000	88	86
5000	76	77
3000	66	71
2000	58	62
1000	51	56

нологій США з метою калібрування та зіставлення методів розпізнавання зображень за допомогою машинного навчання. Вона представлена 10 класами та містить 60000 зображень для навчання і 10000 зображень для тестування вже натренованої мережі. Як модель класифікатора для навчальної вибірки MNIST використано ансамбль з CNN мереж. Як бачимо, запропонований метод попереднього оцінювання інформативності вибірки демонструє ймовірності, дещо нижчі за ймовірність правильної класифікації з використанням реальної моделі на малих розмірах вибірки. Це пояснюється тим, що ансамбль CNN мереж є однією з найкращих моделей і за своїми властивостями має перевагу над класифікатором, на основі даних якого була створена регресійна модель оцінювання якості навчальної вибірки.

Важливою характеристикою моделей глибокого навчання є час тренування на навчальній вибірці. Залежності часу навчання моделей глибокого навчання від розміру навчальної вибірки наведено на рис. 5.

Для оцінювання інформативності навчальної вибірки велике значення мають обчислювальна складність та час оцінювання. З рис. 4. видно, що час оцінювання навчальної вибірки запропонованим підходом є більше ніж на порядок меншим за час прямого оцінювання після навчання, оскільки процес навчання моделі глибокого навчання є обчислювально складним, а запропонований підхід потребує лише виділення ознак попередньо навченою згортковою мережею типу VGG та застосування регресійної моделі для встановлення прогнозованих точностей класифікації. Експеримент здійснено з використанням процесора core i5, 16 GB RAM та графічного прискорювача Nvidia GTX 1060.

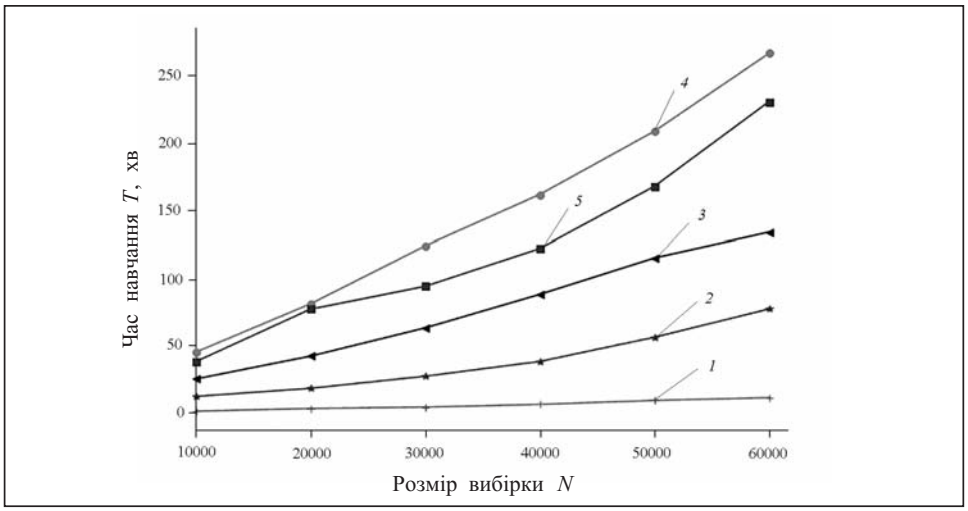

Рис. 5. Залежності часу навчання моделей глибокого навчання від розміру навчальної вибірки: запропонований підхід (крива *1*); VGG Net (крива *2*); Alex Net (крива *3*), ансамбль CNN (крива *4*); Google net (крива *5*)

ВИСНОВКИ

Для глибокого навчання передбачається наявність великої навчальної вибірки. Дуже часто навчальну вибірку формують за таким принципом: чим більшою є вибірка, тим краще. Проте не завжди легко отримати велику кількість даних для навчання. Тому потрібно знайти спосіб, у який можна попередньо оцінити інформативність наявних даних ще до початку навчання моделі глибокого навчання, зберігши час та обчислювальні ресурси.

Запропоновано новий підхід до оцінювання інформативності навчальної вибірки, створеної на основі зображень, отриманих засобами дистанційного зондування. Підхід ґрунтується на припущенні, що інформативність навчальної вибірки можна відобразити набором деякої кількості характеристик, кожна з яких описує певні властивості даних. Встановлення залежності між характеристиками навчальної вибірки та точністю роботи класифікатора, тренованого на основі цієї вибірки, реалізовано за допомогою лінійної регресійної моделі.

На обчислювану складність обрахунку характеристик навчальної вибірки впливає кількість даних і особливо їхня розмірність. Для зменшення обчислювальної складності було запропоновано використовувати метод пониження розмірності даних без втрати структури даних, що ґрунтується на мінімізації відстані Кульбака–Лейблера. Це надає змогу перейти до формування характеристик навчальної вибірки зі значно меншими обчислювальними затратами та компактнішим представленням ознакового простору.

Запропонований підхід було апробовано на різних тестових навчальних вибірках і показано, що він дає результати сумірні з тими, які отримані в результаті навчання нейронної мережі. Водночас є можливість за рахунок розробленого підходу оцінювання інформативності навчальної вибірки на порядок пришвидшити отримання результату класифікації. Це надає змогу використовувати запропонований підхід для попереднього оцінювання навчальної вибірки, завдяки чому є можливість скорегувати її розмір ще до початку навчання мережі.

СПИСОК ЛІТЕРАТУРИ

1. Khan A., Sohail A., Zahoora U., Qureshi A. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*. 2020. Vol. 53, Iss. 8. P. 5455–5516. https://doi.org/10.1007/s10462-020-09825-6.
2. Rusyn B.P., Lutsyk O.A., Tayanov V.A. Upper-bound estimates for classifiers based on a dissimilarity function. *Cybernetics and Systems Analysis*. 2012. Vol. 48, N 4. P. 592–600. https://doi.org/10.1007/s10559-012-9439-2.
3. Boyun V.P. The principles of organizing the search for an object in an image, tracking an object and the selection of informative features based on the visual perception of a person. *Proc. International Conference on Data Stream Mining and Processing (DSMP 2020)* (21–25 April 2020, Lviv, Ukraine) Lviv, 2020. Communications in Computer and Information Science. 2020. Vol. 1158. P. 22–44. https://doi.org/10.1007/978-3-030-61656-4 sub 2.
4. Vapnik V. The Nature of statistical learning theory. New York: Springer-Verlag, 2000. 314 p.
5. Bishop C.M. Pattern recognition and machine learning (Information science and statistics). London: Springer, 2006. 738 p.
6. Chen Y., Lin Z., Zhao X., Wang G., Gu Y. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2014. Vol. 7, Iss. 6. P. 2094–2107. https://doi.org/10.1109/JSTARS.2014.2329330.
7. Ma L., Liu Y., Zhang X., Ye Y., Yin G., Johnsonf B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing.* 2019. Vol. 152. P. 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.
8. Li Y., Zhang H., Xue X., Jiang Y., Shen Q. Deep learning for remote sensing image classification: A survey. WIREs Data Mining and Knowledge Discovery. 2018. 17 p. https://doi.org/10.1002/widm.1264.
9. Cheng G., Xie X., Han J., Guo L., Xia G. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensin*g. 2020. Vol. 13. P. 3735–3756. https://doi.org/10.1109/JSTARS.2020.3005403.

10. Hoque M., Burks R., Kwan C., Li J. Deep learning for remote sensing image super-resolution. *Proc. IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (10–12 Oct. 2019, New York, NY, USA). New York, 2019. P. 286–292. https://doi.org/10.1109/UEMCON47517.2019.8993047.
11. Van Niel T.G., McVicar T.R., Datt B. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote Sensing of Environment*. 2005. Vol. 98, Iss. 4. P. 468–480. https://doi.org/10.1016/j.rse.2005.08.011.
12. Zou Q., Ni L., Zhang T., Wang Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*. 2015. Vol. 12, Iss. 11. P. 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299.
13. Hinterstoisser S., Lepetit V., Wohlhart P., Konolige K. On pre-trained image features and synthetic images for deep learning. In: *Computer Vision — ECCV 2018 Workshops. Proc. 15th European Conference on Computer Vision (ECCV2018)* Workshops (8–14 September 2018, Munich, Germany). Munich, 2018. P. 178–186. https://doi.org/10.1007/978-3-030-11009-3_42.
14. Genc B., Tunc H. Optimal training and test sets design for machine learning. *Turkish Journal of Electrical Engineering and Computer Sciences.* 2019. Vol. 27(2). P. 1534–1545. https://doi.org/10.3906/elk-1807-212.
15. Dodge S., Karam L. Understanding how image quality affects deep neural networks. *Proc. 2016 Eighth International Conference on Quality of Multimedia Experience* (6–8 June 2016, Lisbon, Portugal). Lisbon, 2016. https://doi.org/10.1109/QoMEX.2016.7498955.
16. Cheng G., Yang C., Yao X., Guo L., Han J. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing.* 2018. Vol. 56, Iss. 5. P. 2811–2821. https://doi.org/10.1109/TGRS.2017.2783902.
17. Ma X., Geng J., Wang H. Hyperspectral image classification via contextual deep learning. *Journal on Image and Video Processing*. 2015. Article number: 20 (2015). https://doi.org/10.1186/s13640-015-0071-8.
18. Subbotin S.A. The training set quality measures for neural network learning. *Optical Memory and Neural Networks.* 2010. Vol. 19, Iss. 2. P. 126–139. https://doi.org/10.3103/S1060992X10020037.
19. Forsati R., Moayedikia A., Safarkhani B. Heuristic approach to solve feature selection problem. *Proc. International Conference on Digital Information and Communication Technology and Its Applications (DICTAP 2011)* (21-23 June 2011, Dijon, France). Dijon, 2011. Communications in Computer and Information Science. Vol. 167. P. 707–717. https://doi.org/10.1007/978-3-642 22027-2_59.
20. Huang K., Aviyente S. Wavelet feature selection for image classification. *IEEE Transactions on Image Processing*. 2008. Vol. 17, Iss. 9. P. 1709–1720. https://doi.org/10.1109/TIP.2008.2001050.
21. Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric. *Journal of Classification.* 2020. Vol. 37, Iss. 3. P. 696–708. https://doi.org/10.1007/s00357-019-09345-1.
22. Belov D., Armstrong R. Distributions of the Kullback–Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology*. 2011. Vol. 64, Iss. 2. P. 291–309. https://doi.org/10.1348/000711010X522227.
23. Prati R.C., Batista G.E., Monard M.C. Class imbalances versus class overlapping: an analysis of a learning system behavior. *Proc. Third Mexican International Conference on Artificial Intelligence (MICAI 2004)* (26–30 April 2004, Mexico City, Mexico). Mexico City, 2004. Lecture Notes in Computer Science. Vol. 2972. P. 312–321. https://doi.org/10.1007/978-3-540-24694-7_32.
24. Shepperd M., Cartwright M.. Predicting with sparse data. *Proc. 7th IEEE International Software Metrics Symposium* (4–6 April 2001, London, UK). London, 2001. P. 28–39. https://doi.org/10.1109/METRIC.2001.915513.

B.P. Rusyn, O.A. Lutsyk, R.Y. Kosarevych

EVALUATING THE INFORMATIVITY OF TRAINING SAMPLE FOR CLASSIFICATION OF IMAGES BY DEEP LEARNING METHODS

Abstract. A new approach to evaluate the informativeness of the training sample when recognizing images obtained by means of remote sensing is proposed. It is shown that the informativeness of the training sample can be represented by a set of characteristics, each of which describes certain properties of the data. A relationship between the characteristics of the training sample and the accuracy of the classifier trained on the basis of this sample is established. The proposed approach is applied to various test training samples and the results of their evaluation are presented. When evaluating the training set by the proposed approach, the process is shown to be much faster than training a neural network. This makes it possible to use the proposed approach for preliminary estimation of the training sample in the problems of image recognition using deep learning methods.

Keywords: deep learning, feature selection, training sample, convolution network.

Надійшла до редакції 15.01.2021