УДК 621.327

МАЗІАШВІЛІ А. Р., аспірант кафедри транспортного зв'язку (Український державний університет залізничного транспорту)

Доцільність використання нейронних мереж для стиснення відеоданих

Стиснення зображень широко застосовується в різних галузях науки і техніки. Особливо великого значення набуває стиснення зображень при передачі їх по вузькосмугових каналах зв'язку. Ця сфера використання є дуже поширеною останніми роками, коли актуальними стали проблеми цифрового телебачення, передачі відеопрограм на замовлення по масових телекомунікаційних каналах зв'язку, можливість проведення відеоконференцій з використанням модемного зв'язку, питання стиснення відеоданих у випадках передачі статичних повнокольорових зображень або бінарних документів (наприклад, банківських) через модем, оскільки вартість передачі безпосередньо пов'язана з обсягом переданої інформації.

Ключові слова: *шари нейронів, стиснення даних, квантування даних, метод стиснення, стиснення відеоданих, стиснення зображень.*

Постановка проблеми і аналіз літератури

В даний час на передачу даних витрачаються великі часові ресурси. У зв'язку з цим актуальними є питання стиснення (компресії) інформації перед її передачею. Компресія дає змогу значно збільшити пропускну спроможність ліній зв'язку і є додатковим заходом забезпечення захисту конфіденційної інформації.

Проблема стиснення зображень і відеопослідовності актуальна також при створенні центрів зберігання, архівів і каталогів (баз даних) зображень і відеопослідовності в цифровому вигляді (медійні зображення, космічні зображення, отримані за допомогою датчиків дистанційного зондування, фотозображення та ін.). Вирішення цієї проблеми надасть можливість зменшити обсяг інформації, що зберігається на носіях [3]. На сьогоднішній день є ряд перспективних алгоритмів стиснення DVI, JPEG, MPEG і DWT. Однак останнім часом виникає інтерес до альтернативних способів стиснення, в тому числі за допомогою штучних нейронних мереж (ШНМ).

Мета статті: обґрунтування можливості застосування штучних нейронних мереж для стиснення відеоданих.

Основна частина

Нейромережеві технології надають сьогодні широкі можливості і для вирішення завдань прогнозування, і обробки сигналів, і розпізнавання образів. У порівнянні з традиційними методами математичної статистики, класифікації та апроксимації, ці технології забезпечують досить високу якість рішень при менших витратах [2].

Здатність нейронних мереж до виявлення взаємозв'язків між різними параметрами дає можливість надавати дані великої розмірності більш компактно.

На відміну від традиційних методів стиснення, нейронна мережа при вирішенні задачі стиснення виходить з міркувань нестачі ресурсів. Інше застосування алгоритмів стиснення зображень, на основі нейронних мереж, реалізується для створення цифрових підписів в інтересах захисту електронних об'єктів, створення радіочастотних міток підвищеної скритності, зберігання конфіденційної інформації тощо. Широкого застосування набули методи, основані на використанні властивостей надмірності відеоінформації і орієнтовані на виконання процедур стиснення [1].

Мережа Кохонена, як одна із багатьох на сьогодні варіацій нейронних мереж, часто використовується для стиснення зображень із втратою якості. Вона дозволяє виділяти схожі фрагменти даних у класи. Номер класу зазвичай займає набагато менше місця в пам'яті, ніж ядро класу. Якщо передати одержувачу всі ядра класів і номери класів, що кодують кожен фрагмент даних, то дані можуть бути відновлені.

Стиснення даних, зменшення ступеня їх надмірності, що використовує існуючі в них закономірності, може істотно полегшити подальшу роботу з даними, виділяючи дійсно незалежні ознаки. Тому самонавчені мережі найчастіше використовуються саме для попередньої обробки необроблених даних. Практично, адаптивні мережі кодують вхідну інформацію найбільш компактним кодом, при заданих обмеженнях [4].

Довжина опису даних пропорційна, по-перше, розрядності b, яка визначає можливу різноманітність прийнятих ними значень і, по-друге, розмірності

даних, тобто числу компонент вхідних векторів. Відповідно, можна розрізнити два граничних типи кодування, що використовують протилежні способи стиснення інформації:

1 Зниження розмірності даних з мінімальною втратою інформації. (Мережі, наприклад, здатні здійснювати аналіз головних компонент даних, виділяти набори незалежних ознак.)

2 Зменшення різноманітності даних за рахунок виділення кінцевого набору прототипів і віднесення даних до одного з таких типів. (Кластеризація даних, квантування безперервної вхідної інформації.)

Розглянемо, які можливості щодо адаптивної обробки даних має одиничний нейрон і як можна сформулювати правила його навчання. Через локальність нейромережевих алгоритмів, це базове правило можна буде потім легко поширити і на мережі з багатьох нейронів.

Побудуємо одношарову мережу. Сконцентруємось на тому, що для даної мережі нелінійність функції активації не принципова. У простій постановці нейрон з одним виходом і d входами навчається на наборі d-мірних даних. Тому можна спростити розгляд, обмежившись лінійною функцією активації. Вихід такого нейрона є лінійною комбінацією його входів. На рис. 1, а) зображено одношарову мережу, на рис. 1, б) – вектор ваг нейрона.

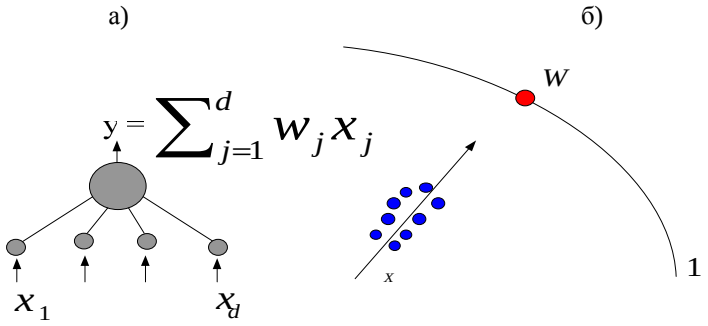

Рис. 1. Одношарова мережа (а); вектор ваг нейрона (б)

Амплітуда цього виходу після відповідного навчання (тобто вибору ваг по набору прикладів) може служити індикатором того, наскільки даний вхід відповідає навчальній вибірці. Іншими словами, нейрон може стати індикатором належності вхідної інформації до вибраної групи прикладів.

Правило навчання окремого нейрона-індикатора за потребою локальне, тобто базується тільки на інформації, безпосередньо доступній самому нейрону – значеннях його входів і виходу.

Якщо сформулювати навчання як задачу оптимізації, ми побачимо, що нейрон, який навчається за правилом Хебба, прагне *збільшити* амплітуду свого виходу

$$\langle \Delta W \rangle = -\eta \frac{\partial E}{\partial W}, \; E\left(W \cdot X^{a}\right) = \frac{1}{2}\left\langle (W \cdot X)^2 \right\rangle = -\frac{1}{2}\left(y^2\right). \tag{1}$$

Усереднення, згідно з формулою (1), проводиться за навчальною вибіркою $\{X^a\}$. За відсутності еталона, мінімізувати значення немає сенсу: мінімізація амплітуди виходу призвела б лише до зменшення чутливості виходів до значень входів.

Зазначена відмінність з метою навчання має принциповий характер, тому що мінімум помилки $E(W)$ в даному випадку відсутній. Тому навчання за Хеббом у тому вигляді, в якому воно описано вище, на практиці не застосовується, тому що призводить до необмеженого зростання амплітуди ваг.

Від цього недоліку, а саме від зростання амплітуди ваг, можна досить просто позбутися, додавши компоненту, що перешкоджає зростанню ваг. Так, за навчанням Ойа, записуємо вираз який має вигляд:

$$\Delta w_j^{\tau} = \eta y^{\tau}\left(x_j^{\tau} - y^{\tau} w_j\right), \tag{2}$$

або у векторному вигляді:

$$\Delta W^{\tau} = \eta y^{\tau}\left(X^{\tau} - y^{\tau} W\right). \tag{3}$$

Саме це навчання максимізує чутливість виходу нейрона при обмеженій амплітуді ваг. У цьому легко переконатися, прирівнявши середню зміну ваг до нуля. Помноживши потім праву частину на W, бачимо, що в рівновазі

$$\left(y^2\right)\left(1 - |w|^2\right) = 0 \,. \tag{4}$$

Записавши правило змагального навчання в градієнтному вигляді, отримуємо

$$\langle \Delta W \rangle = -\eta \frac{\partial E}{\partial W}. \quad (5)$$

Виходячи з цієї формули, легко можна переконатися, що змагальне навчання мінімізує квадратичне відхилення вхідних векторів від прототипів-ваг нейронів переможців:

$$E = \frac{1}{2}\sum_a \left| X^a - W^a \right|. \quad (6)$$

Іншими словами, мережа здійснює кластеризацію даних: знаходить такі усереднені прототипи, які мінімізують помилку огрубіння даних. Недолік такого варіанта кластеризації очевидний – "нав'язування" кількості кластерів, рівного числу нейронів.

Порівняльний аналіз шарів нейронів широко використовується для квантування даних (vector quantization), що відрізняється від кластеризації лише великим числом прототипів. Це дуже поширений на практиці метод стиснення даних. При досить великій кількості прототипів, щільність розподілу ваг змагального шару добре апроксимує реальну щільність розподілу багатовимірних вхідних векторів. Стиснення даних у цьому випадку досягається за рахунок того, що кожен прототип можна закодувати меншим числом біт, ніж відповідні йому вектори даних.

Проведення оцінки обчислювальної складності

Сигмоїда застосовується в нейронних мережах як функції активації, оскільки дозволяє як посилювати слабкі сигнали, так і не насичуватися від потужних сигналів. Похідна сигмоїда може бути виражена через саму функцію, що дає можливість істотно скоротити обчислювальну складність методу зворотного поширення помилки, зробивши його придатним на практиці. Згідно з формулами, алгоритм навчання мереж, що знижують розмірність, зводиться до звичайного навчання з вчителем. Таке навчання потребує $\sim PW^2$ операцій, де W – це число синаптичних ваг мережі, P – число навчальних прикладів. Для одношарової мережі з d-входами та m-вихідними нейронами число ваг дорівнює $W \approx dm$. Складність навчання можна оцінити як:

$$C_1 \sim P^2 d^2 m^2 = \frac{Pd^4}{K^2}, \quad (7)$$

де $K = \frac{d}{m}$ – коефіцієнт стиснення інформації.

Щодо квантування та кластеризації, то вони вимагають налаштування набагато більшої кількості ваг – через неефективний спосіб кодування. В той же час, таке надлишкове кодування спрощує алгоритм навчання. Число ваг, як і раніше, так само дорівнює $W \approx dm$, але ступінь стиснення інформації в даному випадку визначається по-іншому:

$$K = \frac{db}{\log_2 m}. \quad (8)$$

Складність навчання як функція ступеня стиснення запишеться у вигляді:

$$C_2 \sim P d m \sim P d 2^{\frac{db}{K}}. \quad (9)$$

При однаковій мірі стиснення відношення складності квантування до складності даних зниження розмірності запишеться у вигляді:

$$\frac{C_2}{C_1} = \frac{K^2 2^{\frac{db}{K}}}{d^3}. \quad (10)$$

Іноді, навіть проста заміна лінійної функції активації нейронів на сигмоїдну в знайденому вище правилі навчання призводить до нової якості. Такий алгоритм, зокрема, з успіхом застосовувався для поділу змішаних сигналів (так званий blind signal separation). Це завдання кожен з нас змушений вирішувати, коли хоче виділити мову однієї людини в шумі загальної розмови.

Тому замінимо лінійну функцію активації нейронів на сигмоїдну і отримаємо новий вираз:

$$\Delta W_i^\tau = \eta f\left(y_i^\tau\right)\left(X^\tau - \sum_k f\left(y_k^\tau\right) W_k\right). \quad (11)$$

При обробці сигналів і зображень часто використовуються різного роду лінійні і нелінійні перетворення даних, що забезпечують їх стиснення.

За відсутності апріорної інформації побудова перетворень для стиснення сигналів та зображень зручно здійснювати в рамках нейромережевого підходу. В його основі лежить застосування штучних нейронних мереж (НМ), яких навчають за сукупністю реалізацій випадкових векторів, що відображають взаємозв'язки тимчасових і просторових фрагментів аналізованих процесів і полів.

Далі розглянемо задачу, яка пов’язана з можливістю застосування нейронних мереж для побудови перетворюючих стиснень.

Нехай $z \in R^N$, $N = N1 + N2$ – випадковий вектор, що представляє деяку область випадкового поля $\Omega \subset \Psi$ і отриманий шляхом розгортки $w(x, y), (x, y) \in \Omega$ у довільному порядку. Для визначеності будемо вважати, що математичне очікування $M[z] = 0$, при цьому матриця коваріації вектора z $R_z = M\left[zz^T\right]$.

Потрібно зазначити, що випадкові вектори z_1, z_2 пов'язані співвідношенням:

$$z_2 = z_{2/1} + V = H z_1 + V, \quad H = R_{z21} R_{z11}^{-1}, \quad M[V] = 0, \quad M\left[VV^T\right] = R_{z22} - R_{z21} R_{z11}^{-1} R_{z12} \tag{12}$$

де z_{21} – має значення оптимальної (в лінійному класі) оцінки z_2;

V – стохастична складова, яка не корельована з $z_{2/1}$;

$$R_{z11} = M\left[z_1 z_1^T\right], \quad R_{z22} = M\left[z_2 z_2^T\right], \quad R_{z21} = M\left[z_2 z_1^T\right].$$

Як перетворювачі можуть використовуватися нейронні мережі прямого поширення, вагові коефіцієнти яких можуть налаштовуватися шляхом безпосереднього їх обчислення або на основі ітеративного навчання за методом зворотного поширення помилки [3]. Типова архітектура НМ, яка може бути використана для стиснення даних, подана на рис. 2, а. Для подальшого аналізу потрібно також розглянути перетворювач (рис. 2, б), який може бути реалізований у вигляді одношарової НМ.

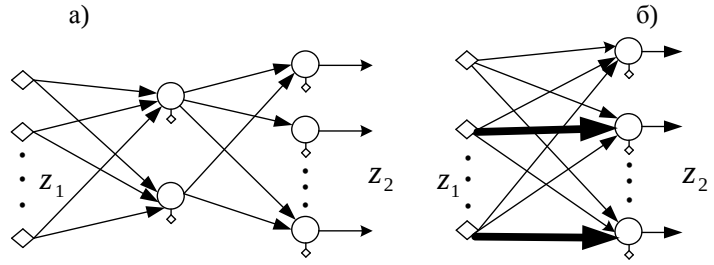

Рис. 2. Архітектура нейронних мереж, які використовуються для стиснення даних (а); перетворювач, реалізований у вигляді одношарової НМ (б)

Особливістю архітектури мережі є використання скороченого числа нейронів $M < N_1, M < N_2$ у прихованому шарі по відношенню до розмірності вхідного та вихідного векторів. Загальна матриця вагових коефіцієнтів $W^{(1,2)}$ цієї НМ має вигляд $W^{(1,2)} = W^{(2)} W^{(1)}$, де $W^{(2)}, W^{(1)}$ – матриці ваг першого та другого шарів. При виконанні перетворення стисненням вхідного вектора z_1 на виході НМ отримаємо вектор:

$$\tilde{z}_2 = W^{(2)} W^{(1)} z_1 = W^2 \nu_{12}, \tag{13}$$

де ν_{12} – перехідний сигнал на виході обчислювальних елементів (нейронів) першого шару, який надходить на вхід другого шару нейрона. Для подальшого аналізу потрібно також розглянути перетворювач (рис. 2, б), який може бути реалізований у вигляді одношарової НМ.

Запишемо цільову функцію, яку треба мінімізувати при навчанні НМ відносно сукупності реалізацій z_1, z_2, у вигляді:

$$E = \frac{1}{2} \sum_{p=1}^{P} (z_2^p - \tilde{z}_2^{(p)})^T (z_2^{(p)} - \tilde{z}_2^{(p)}) = \frac{1}{2} \sum_{p=1}^{P} \sum_{j=1}^{N2} (z_{2,j}^{(p)} - \sum_{k=1}^{M} w_{jk}^{(2)} \sum_{i=1}^{N1} w_{ki}^{(1)} z^{(p)}{}_{1,i})^2, \tag{14}$$

де $\tilde{z}^{(p)}{}_2 = W^{(2)}W^{(1)}z^{(p)}{}_1$ – реакція на вхідний вплив $z^{(p)}{}_1$.

Подамо величину середньоквадратичної помилки у вигляді:

$$
\begin{aligned}
E &= \frac{1}{2}tr\left(\frac{1}{2}\sum_{p=1}^{P}\left(z_2^p - \tilde{z}_2^{(p)}\right)\left(z_2^{(p)} - \tilde{z}_2^{(p)}\right)^T\right) = \\
&= \frac{1}{2}tr\left(\sum_{p=1}^{P}\left(z_2^{(p)} - \tilde{z}_{2/1}^{(p)}\right)\left(z_2^{(p)} - \tilde{z}^{(p)}{}_{2/1}\right)^T\right) + \\
&+ \frac{1}{2}tr\left(\sum_{p=1}^{P}\left(\tilde{z}_{2/1}^{(p)} - \tilde{z}_2^{(p)}\right)\left(\tilde{z}_{2/1}^{(p)} - \tilde{z}_2^{(p)}\right)^T\right) + \\
&+ \frac{1}{2}tr\left(\sum_{p=1}^{P}\left(z_2^{(p)} - \tilde{z}_{2/1}^{(p)}\right)\left(\tilde{z}_{2/1}^{(p)} - \tilde{z}_2^{(p)}\right)^T\right) + \\
&+ \frac{1}{2}tr\left(\sum_{p=1}^{P}\left(\tilde{z}_{2/1}^{(p)} - \tilde{z}_2^{(p)}\right)\left(z_2^{(p)} - \tilde{z}^{(p)}{}_{2/1}\right)^T\right)
\end{aligned}
\tag{15}
$$

де $\tilde{z}_{2/1}^{(p)} = \tilde{R}_{z21}R_{z11}^{-1}z_1^{(p)}$ – лінійна оцінка $z_2^{(p)}$ відповідно до спостереження $z_1^{(p)}$, яка є реакцією на виході перетворювача (рис. 2, б). Оскільки:

$$
\begin{aligned}
&\sum_{p=1}^{P} z^{(p)}{}_2\tilde{z}_{2/1}^{(p),T} = (P-1)\tilde{R}_{z21}\tilde{R}_{z11}^{-1}\tilde{R}_{z12}^{(p)}, \\
&\sum_{p=1}^{P}\tilde{z}_2^{(p),T} = (P-1)\tilde{R}_{z21}W^{(1,)T}W^{(2),T}, \\
&\sum_{p=1}^{P}\tilde{z}^{(p)}{}_{2/1}\tilde{z}_{2/1}^{(p),T} = (P-1)\tilde{R}_{z21}\tilde{R}_{z11}^{-1}\tilde{R}_{z11}^{(p)}\tilde{R}_{z11}^{-1}\tilde{R}_{z12} = \\
&= (P-1)\tilde{R}_{z21}\tilde{R}_{z11}^{-1}\tilde{R}_{z12}, \\
&\sum_{p=1}^{P}\tilde{z}^{(p)}{}_{2/1}\tilde{z}_2^{(p),T} = (P-1)\tilde{R}_{z21}\tilde{R}_{z11}^{-1}\tilde{R}_{z11}W^{(1,)T}W^{(2),T} = \\
&= (P-1)\tilde{R}_{z21}W^{(1,)T}W^{(2),T}
\end{aligned}
\tag{16}
$$

Тоді два останніх доданки у виразі для E дорівнюють нулю. Це означає, що помилка, яка мінімізується відносно коефіцієнтів матриць $W^{(2)}W^{(1)}$, складається з двох доданків.

Отримані результати

На основі наведеного математичного апарату було показано, що застосування апарату штучних нейронних мереж (ШНМ) для стиснення даних надає певні переваги, з точки зору зниження часу виконання процедури стиснення при наявності навченого перетворювача. Таким чином, нейроалгоритми є важливим інструментом нелінійного аналізу, що дає змогу відносно легко знаходити способи глибокого стиснення інформації та виділення нетривіальних ознак.

Висновки

Таким чином, перспективним напрямком розвитку є підхід, в основі якого закладені ШНМ. При цьому ШНМ можуть використовуватися як при стисненні без втрат (наприклад, у статистичних методах кодування для оцінки ймовірностей появи символів), так і при реалізації стиснення з втратами, наприклад, у стандарті JPEG 2000, основаному на вейвлет-перетворенні. В останньому випадку доцільним є застосування ШНМ, що дублює векторне квантування або кластеризацію [5].

У зв’язку з цим, штучні нейронні мережі можуть досить успішно справлятися із завданнями стиснення, як зображення, так і відео.

Література

1. Комашинский, В. И. Нейронные сети и их применение в системах управления и связи [Текст] / В. И. Комашинский, Д. А. Смирнов, – М.: Телеком 2003. – 94 с.
2. Саймон Хайкин. Нейронные сети: полный курс [Текст]: пер. с англ. / Хайкин Саймон. – 2-е изд. – М.: Издательский дом «Вильямс», 2006. – 1104 с.
3. Руденко, О. Г. Сжатие изображений на основе нейронной сети ART [Текст] / О. Г. Руденко, М. С. Сныткин // Кибернетика и системный анализ. – 2008. – № 6. – С. 7.
4. Кохонен, Т. Самоорганизующиеся карты [Текст] / Т. Кохонен, – М.: БИНОМ. Лаборатория знаний, 2008. – 655 с.

5. Siva Nagi Reddy K. Image Compression and Reconstruction Using a New Approach By Artificial Neural Network [Text] / K. Siva Nagi Reddy, Dr. B.R. Vikram, L. Koteswara Rao, B. Sudheer Reddy // International Journal of Image Processing (IJIP) – Taiwan, 2012. – Volume 6.

Мазиашвили А. Р. Целесообразность использования нейронных сетей для сжатия видеоданных. Сжатие изображений находит широкое применение в различных областях науки и техники. Особенно большое значение приобретает сжатие изображений при передаче их по узкополосным каналам связи. Эта область использования становится широко распространенной в последние годы, когда актуальной стала проблема цифрового телевидения, передача заказных видеопрограмм по массовым телекоммуникационным каналам связи, возможность проведения видеоконференций с использованием модемной связи, вопросы сжатия видеоданных в случаях передачи статических полноцветных изображений или бинарных документов (например, банковских) через модем, поскольку стоимость передачи непосредственно связана с объемом передаваемой информации.

Ключевые слова: слои нейронов, сжатие данных, квантование данных, метод сжатия, сжатие видеоданных, сжатие изображений.

Maziashvili A. R. The expediency of using neural networks for video compression. Image compression is widely used in various fields of science and technology. Of particular importance is the image compression when transferring their narrowband channels. This area of use is becoming widespread in recent years, as has become urgent in digital television transmission customized video on mass telecommunication channels, video conferencing is possible with the use of modem, issues video compression in cases of transfer of static full-color images or binary documents (such as bank) via a modem, since the cost of transmission is directly related to the volume of information transmitted.

Unlike traditional compression methods, neural network with the task of compressing out for reasons of lack of resources. Another application of algorithms for image compression based on neural networks implemented digital watermarks for the protection of electronic objects, creating radio frequency tags increased stealth, keeping confidential information and more. Widespread methods based on the use of video and redundancy features focused on the procedures of compression.

Key words: layers of neurons, data compression, quantization data, compression method, video compression, image compression.

Надійшла 25.11.2016 р.

Maziashvili A. R., *postgraduate student, Ukrainian State University of Railway Transport, Kharkiv, Ukraine.*

Мазіашвілі Артур Рамазійович, *аспірант кафедри транспортного зв'язку, Український державний університет залізничного транспорту, Харків, Україна.*