ỨNG DỤNG MẠNG NEURAL XOẮN ĐỂ NHẬN DẠNG KÝ TỰ

Đào Minh Sơn, Nguyễn Hồng Sơn

(Bài nhận ngày 6 tháng 2 năm 2001; hoàn chỉnh sửa chữa ngày 10 tháng 5 năm 2001)

TÓM TẮT:

Mạng neural rất thường được ứng dụng để giải quyết các bài toán trong lĩnh vực nhận dạng. Trong bài viết này chúng tôi trình bày một ứng dụng cụ thể của mạng neural xoắn LeNet, hoạt động theo nguyên lý làm việc của mạng neural lan truyền ngược, trong việc nhận dạng ký tự.

Chúng tôi sẽ trình bày chi tiết kiến trúc mạng LeNet-mở rộng, dựa vào nguyên lý làm việc của mạng LeNet có kết hợp với một số chỉnh sửa nhắm vào bộ dò và các bổ sung trọng số giả newton đa ra. Mạng Lenet-mở rộng chúng tôi xây dựng gồm 7 lớp, không tính đầu ra, tất cả các lớp đều chứa thông số huấn luyện (tức các trọng số). Chúng tôi cũng giới thiệu những kết quả nhận dạng của chơng trình thực nghiệm mà chúng tôi đã phát triển.

I. Giới thiệu mạng neural xoắn (Convolutional Neural Network)

Các mạng xoắn về cơ bản là một dạng mạng neural lan truyền ngược kết hợp với 3 khái niệm kiến trúc sau:

1. Các trường tiếp thu cục bộ
2. Các bản sao trọng số
3. Các đại diện mẫu

Việc kết hợp này nhằm duy trì mức độ bất biến đối với sự dịch chuyển, tỉ lệ và biến dạng với hiệu suất cao nhất.

Thông thường *các bộ dò đặc trưng cơ bản hoạt động tốt trên một phần của ảnh thì thường cũng hoạt động tốt xuyên suốt toàn bộ ảnh.* Ta áp dụng đặc điểm này lên trên một ảnh để tạo ra các vector trọng số đồng nhất, trong đó mỗi thành phần của vector là một unit với trường tiếp thu của nó được xác định tại một vị trí được xác định trước trên ảnh. Các unit có cùng trọng số được tổ chức vào trong cùng một lớp. Tập các đầu ra của các lớp như trên được gọi là *ánh xạ đặc trưng* (feature map-FM). Do vậy, tất cả các unit trong FM đều chia sẻ cùng tập n trọng số, do vậy chúng sẽ dò tìm cùng một đặc trưng tại tất cả các vị trí có thể trên ảnh đầu vào. Mặt khác trên cùng một lớp có nhiều FM khác nhau, do vậy mỗi FM sẽ dùng các bộ trọng số khác nhau, bằng cách này ta sẽ trích ra được các kiểu đặc trưng cục bộ khác nhau. Hơn nữa ta có thể quét ảnh đầu vào với một unit đơn với trường tiếp thu cục bộ, sau đó cất giữ unit này tại các vị trí liên quan trong FM. Phép toán này tơng đơng với sự xoắn (convolution), được sinh ra bởi hàm additive bias và squashing, từ đây có tên là *mạng xoắn* (convolutional network). *Cốt lõi của sự xoắn là tập các trọng số liên kết được dùng bởi các unit trong FM.* Đặc điểm nổi bật của các lớp xoắn là nếu ảnh đầu vào bị dịch chuyển, FM đầu ra sẽ bị dịch chuyển cùng giá trị, nhng sẽ bảo toàn với các trường hợp khác. Đặc điểm này chính là điểm mạnh cơ bản của mạng xoắn-*không bị ảnh hưởng của các dịch chuyển hay biến dạng của ảnh đầu vào. Một khi đặc trưng đã được dò ra, vị*

trí chính xác của nó trở nên ít quan trọng đi. Duy nhất chỉ vị trí xấp xỉ của nó liên quan đến các đặc trưng khác là thích hợp để sử dụng.

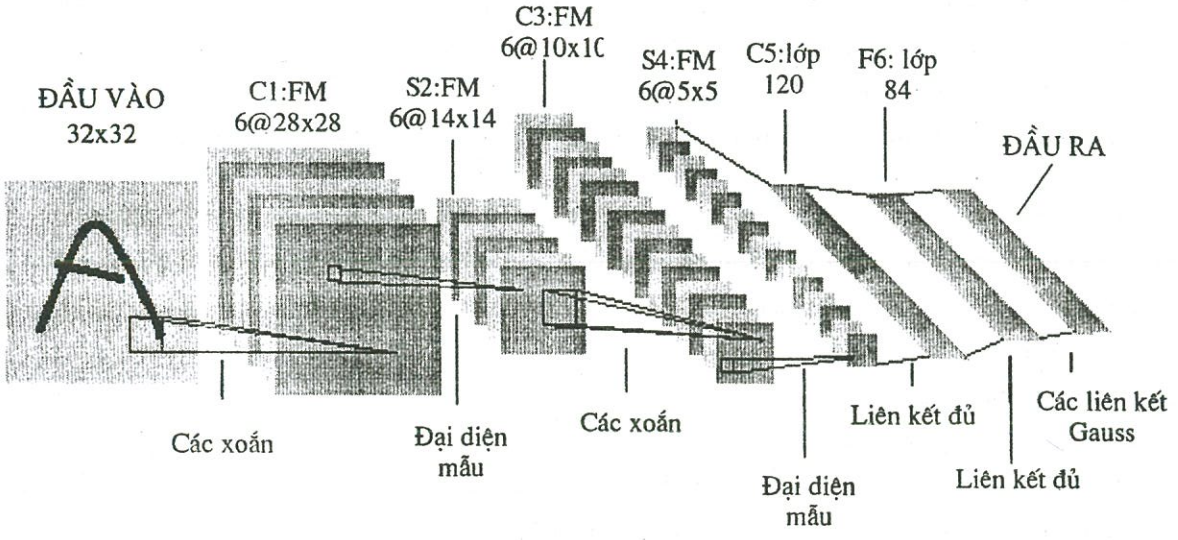
Các biến dạng hoặc dịch chuyển của đầu vào có thể là nguyên nhân làm cho vị trí của các đặc trưng nổi bật bị thay đổi. Không những vị trí chính xác trong các đặc trưng đó là không thích hợp cho việc nhận ra mẫu (pattern), mà nó còn tiềm tàng lỗi do bởi các vị trí có khuynh hướng thay đổi đối với các thể hiện khác nhau của cùng một ký tự. Một cách đơn giản để giảm thiểu độ chính xác trong vị trí là mã hóa vị trí của các đặc trưng phân biệt trong FM, điều này làm giảm thiểu không gian lời giải của FM. Điều này có thể đạt được với *lớp đại diện mẫu* (sub-sampling layer SSL), lớp này có vai trò làm giảm thiểu lời giải của FM, và làm giảm độ nhạy cảm của đầu ra đối với dịch chuyển và biến dạng.

Bởi vì tất cả các trọng số được học với phương pháp lan truyền ngược, các mạng xoắn có thể được xem như việc tổng hợp đầu dò đặc trưng của nó. Kỹ thuật chia sẻ trọng số có một tác động phụ rất hay là giảm được số lượng các thông số tự do, do đó làm giảm *dung lượng* của máy và làm giảm không gian sai số giữa lỗi kiểm tra và lỗi huấn luyện [5].

II. Mạng LeNet-x

Một trong các mạng xoắn đạt hiệu suất nhận dạng cao nhất là mạng LeNet-x (hình 1) [1]. Về cơ bản, mạng LeNet-x vẫn tuân thủ theo cấu trúc của mạng neural cổ điển: lớp đầu vào, lớp ẩn và lớp đầu ra, đồng thời nó có các đặc điểm riêng như sau:

- Lớp đầu vào nhận các ảnh đã được chuẩn hóa về kích thước và trung tâm hóa (tức là trọng tâm các ký tự chính là trọng tâm của ảnh).
- Các unit tại các lớp ẩn và lớp đầu ra tuân thủ theo nguyên tắc sau: mỗi unit trong lớp này



Hình 1. Cấu trúc mạng neural xoắn LeNet-x.

nhận các đầu vào từ tập các unit được định vị trong vùng lân cận nhỏ (cục bộ) ở lớp trước nó.

Trong đó:

- ***Lớp C1*** là lớp xoắn với 6 FM. Mỗi unit trong mỗi FM được liên kết đến 5x5 láng giềng trong đầu vào. Cỡ của các FM là 28x28 để ngăn ngừa liên kết từ đầu vào xuất phát từ các biên của ảnh. C1 chứa 156 thông số huấn luyện, và 122,304 liên kết.
- ***Lớp S2*** là lớp đại diện mẫu với 6 FM có cỡ 14x14. Mỗi unit trong mỗi FM được liên kết với 2x2 láng giềng trong FM liên quan tại lớp C1. Bốn đầu vào đối với unit trong S2 được thêm vào, sau đó được nhân với hệ số huấn luyện và cộng với bias huấn luyện. Kết quả được gởi về hàm sigmoid. Các trường tiếp thu 2x2 là không chồng lấp, do đó các FM trong S2 có một nửa số dòng và một nửa số cột là FM trong C1. Lớp S2 có 12 thông số huấn luyện và 5,880 liên kết.
- ***Lớp C3*** là lớp xoắn với 16 FM. Mỗi unit trong mỗi FM được liên kết với một vài 5x5 láng giềng tại các vị trí đồng nhất/y hệt nhau trong tập con của FM của S2 Các FM khác nhau bị ép để trích ra các đặc trưng khác nhau (hy vọng là bổ sung cho nhau) bởi vì chúng nhận được các tập đầu vào khác nhau. Các FM được liên kết nh sau: (1) 6 FM đầu tiên của C3 lấy các đầu vào từ mỗi tập con kề nhau của 3 FM trong S2 (2) 6 FM kế tiếp của C3 lấy các đầu vào từ mỗi tập con kề nhau của 4 FM trong S2 (3) 3 FM kế tiếp của C3 lấy các đầu vào từ một số tập con không kề nhau của 4 FM trong S2 (4) FM còn lại của C3 lấy các đầu vào từ tất cả các FM của S2. Nh vậy lớp C3 có tất cả 1,516 thông số huấn luyện và 156,000 liên kết.
- ***Lớp S4*** là lớp đại diện mẫu với 16 FM cỡ 5x5. Mỗi unit trong mỗi FM được liên kết với 2x2 láng giềng trong FM liên quan trong C3 tơng tự nh C1 và S4. Lớp S4 có 32 thông số huấn luyện và 2,000 liên kết.
- ***Lớp C5*** là lớp xoắn với 120 FM. Mỗi unit được liên kết với 5x5 láng giềng trên tất cả 16 FM của S4. Bởi vì cỡ của S4 cũng là 5x5, nên cỡ của FM của S4 là 1x1: đây là số lượng liên kết đầy đủ giữa S4 và C5. C5 được đánh nhãn nh lớp xoắn, thay vì là lớp liên kết đủ, bởi vì nếu đầu vào của LeNet-mở rộng được tạo lớn hơn với mọi thứ khác giữ nguyên kích cỡ thì chiều của FM có thể là lớn hơn 1x1. Lớp C5 có 48,120 liên kết huấn luyện.
- ***Lớp F6***, chứa 84 unit và được liên kết đủ với lớp C5. Nó có 10,164 thông số huấn luyện.

Nh trong một mạng neural kinh điển, các unit trong các lớp cho đến F6 tính toán bằng cách nhân giữa vector đầu vào và vector trọng số, sau đó cộng thêm bias. Tổng trọng số a_i của unit i, sau đó được gởi vào hàm sigmoid squash để tính ra trạng thái x_i của unit i:

$$x_i=f(a_i)$$

Hàm squash được tính nh sau:

$$f(a)=A \tanh(Sa)$$

với A là biên độ của hàm và S chỉ ra độ dốc của nó tại gốc tọa độ. Hàm f là hàm lẻ với đường tiệm cận ngang tại +A và -A. Hằng A được chọn bằng 1.7159 (đạt được sau quá trình thực nghiệm).

Cuối cùng, lớp đầu ra là các unit RBF (Euclidean Radial Basis Function), một cho mỗi lớp, với 84 đầu vào cho mỗi lớp. Các đầu ra của mỗi unit RF y_i được tính như sau:

$$y_i = \sum_j (x_j - w_{ij})^2$$

áp dụng và cải tiến mạng LeNet-x để nhận dạng tốt hơn

Thực ra mạng LeNet-x nêu trên được thiết kế chi tiết để nhận dạng chữ số viết tay, do đó đầu ra chỉ có 10 lớp. Tuy nhiên, ta có thể vận dụng, mở rộng thiết kế của mạng này để nhận dạng chữ viết tay.

Ban đầu, chúng tôi thử thiết kế mạng LeNet mở rộng với 26 lớp đầu ra để nhận dạng 26 ký tự từ A đến Z. Song khi đó, số lượng các lớp, các thông số liên kết, các thông số huấn luyện tăng lên một cách nhanh chóng. Trở ngại lớn hơn là tốc độ học của mạng trở nên vô cùng chậm chạp, các trường hợp nhận dạng lỗi và không phân biệt được lớp đầu ra tăng lên đáng kể.

Do vậy, sau một thời gian thử nghiệm, chúng tôi không tiếp tục phát triển theo hướng này, mà quyết định thiết kế hệ thống bao gồm 2 tầng phân lớp: (1) tầng phân lớp nhóm và (2) tầng phân lớp chi tiết.

III.1 Tầng phân lớp nhóm

Chúng tôi nhận thấy các chữ cái có các đặc điểm gần giống nhau thì dễ bị nhận dạng lầm sang nhau. Do vậy, dựa sự gần giống nhau của các chữ cái, sau một thời gian thực nghiệm, chúng tôi phân 26 chữ cái và 10 chữ số vào các nhóm sau:

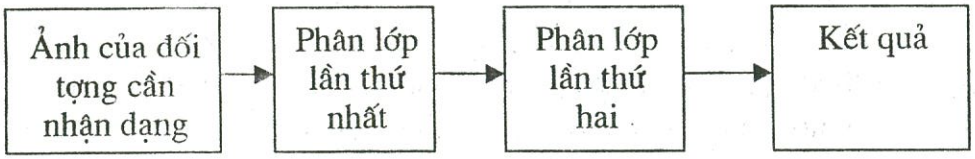
a. *Nhóm CS* : bao gồm các chữ số <0, 1, 2, 3, 4, 5, 6, 7, 8, 9>
b. *Nhóm CC1*: bao gồm các chữ cái sau <A, H, C, G, K, L, M, N, R, W>
c. *Nhóm CC2*: bao gồm các chữ cái sau <D, O, S, Z, T, U, V, Y, X, H>
d. *Nhóm CC3*: bao gồm các chữ cái sau <I, J, T, Q, D, O, P, B, E, F >

Nh vậy, ở tầng này chúng tôi chỉ phải thiết kế mạng LeNet với 4 lớp đầu ra, do vậy các lớp FM, các thông số liên kết và trọng số sẽ giảm đi, thời gian học cho lớp này nhanh hơn.

III.2 Tầng phân lớp chi tiết

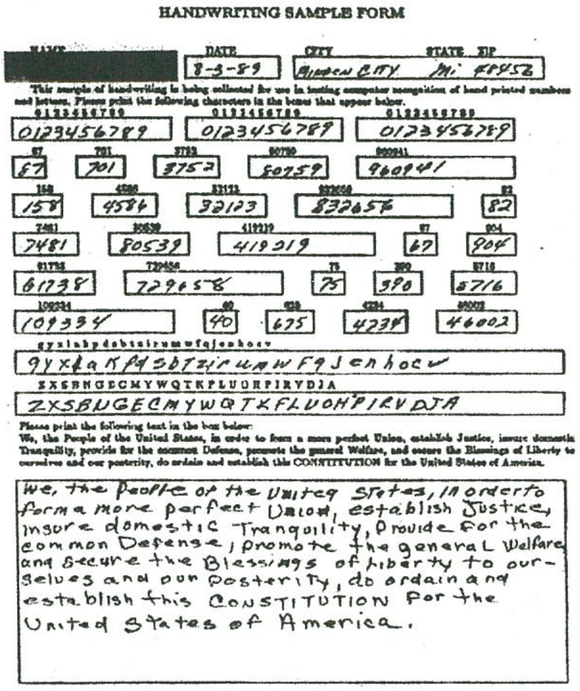
Tầng này chúng tôi sử dụng mạng LeNet nguyên thuỷ có 10 lớp đầu ra. Mỗi mạng LeNet con sẽ dùng tập trọng số tơng ứng.

Hệ thống thực hiện quá trình nhận dạng nh sau:



- **Bộ dữ liệu mẫu**

Chúng tôi sử dụng bộ dữ liệu mẫu của National Institute of Standards and Technology : "NIST Special Database 19 — Handprinted Forms and Characters Database", với kiểu chữ viết của 1500 ngời khác nhau được viết trên form có dạng ·nh sau:


HANDWRITING SAMPLE FORM

NAME [redacted] DATE CITY STATE ZIP

8-3-89 | Bimsen CITY | Mi | 48452

This sample of handwriting is being collected for use in testing computer recognition of hand printed numbers and letters. Please print the following characters in the boxes that appear below.

0123456789 | 0123456789 | 0123456789

ZXSBNGECMYWQTKFLUOHPIRVDJA

Please print the following text in the box below:

We, the People of the United States, in order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common Defense, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our posterity, do ordain and establish this CONSTITUTION for the United States of America.

- **Các thông số dữ liệu được sử dụng như sau:**

Mạng	Bộ dữ liệu học	Bộ dữ liệu kiểm tra	Bộ dữ liệu đánh giá	Tổng cộng
LeNet cho tầng phân lớp thứ nhất	40000 mẫu cho mỗi nhóm	40000 mẫu cho mỗi nhóm	20000 mẫu cho mỗi nhóm	60000 mẫu
LeNet cho nhóm CC1	10000 mẫu cho một chữ cái	10000 mẫu cho một chữ cái	5000 mẫu cho một chữ cái	25000 mẫu
LeNet cho nhóm CC2	10000 mẫu cho một chữ cái	10000 mẫu cho một chữ cái	5000 mẫu cho một chữ cái	25000 mẫu
LeNet cho nhóm CC3	10000 mẫu cho một chữ cái	10000 mẫu cho một chữ cái	5000 mẫu cho một chữ cái	25000 mẫu
LeNet cho nhóm CS	10000 mẫu cho một chữ số	10000 mẫu cho một chữ số	5000 mẫu cho một chữ số	25000 mẫu

Kết quả thực nghiệm

Ứng với bộ test được nêu như bảng tham chiếu trên, kết quả thu được như sau:

Chữ số	Tỷ lệ lỗi	Tỷ lệ nhầm lẫn	Đạt độ chính xác
0	0,83659	0,16084	99,00257%
1	0,39428	0,56071	99,04501%
2	0,63268	1,61026	97,75706%
3	0,45257	0,94863	98,59880%
4	0,82900	0,07263	99,09837%
5	0,61592	0,59699	98,78709%
6	0,99770	1,98864	97,01242%
7	0,22182	0,37790	99,40028%
8	0,99894	1,96191	97,03915%
9	0,99126	1,81926	97,19848%

Chữ cái	Tỷ lệ lỗi	Tỷ lệ nhầm lẫn	Đạt độ chính xác
A	0,98670	1,93269	97,08061%
B	0,92246	0,91531	98,16623%
C	0,83159	0,49757	98,67084%
D	0,73588	0,93898	98,32514%
E	0,99996	1,95301	97,04703%
F	0,99870	1,97284	97,02846%
G	0,99916	1,97309	97,02775%
H	0,71549	0,90798	98,37653%
I	0,35276	0,32104	99,32620%
J	0,99195	1,99001	97,01804%
K	0,99159	1,96413	97,04428%
L	0,99977	1,99670	97,00353%
M	0,29199	1,21946	98,48855%
N	0,96712	1,94745	97,08543%
O	0,60465	0,12079	99,27456%
P	0,74110	0,19731	99,06159%
Q	0,98482	0,39639	98,61879%
R	0,99261	1,98183	97,02556%
S	0,54895	1,53345	97,91760%
T	0,74892	0,22264	99,02844%
U	0,79744	1,77728	97,42528%
V	0,09347	1,89394	98,01259%
W	0,93019	1,66133	97,40848%
X	0,09474	0,66613	99,23913%
Y	0,54895	1,53345	97,91760%
Z	0,95784	0,03564	99,00652%

TÀI LIỆU THAM KHẢO

[1] LeCun, Y., Kanter, I., and Solla, S. (1990). Eigenvalues of covariance matrices: aplication to neural-networks. *Neural Computation*, Vol.1.

[2] Hubel, D.H. and Wiesel, T.N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, 160:106-154.

[3] Fukushima, K. (1975). Cognitron· A self-organizing multilayered neural network. Biological Cybernetics, 20:121-136.

[4] Martin, G.L. (1993). Centered-object integrated segmentation and recognition of overlapping hand-printed characters. *Neural Computation*, 5:419-429.

[5] LeCun, Y., Generalization and network design strategies. In Pfeifer, R., Schreter, Z., Fogelman, F., and Steels, L., editors, *Connectionism in Perspective*, Zurich, Switzerland. Elsevier. An extended version was published as a technical report of the University of Toronto.

[6] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1990). Handwritten digit recognition with a back-propagation network. In Touretzky, D., editor, *Advances in Neural Information Processing System 2 (NIPS*89)*, Denver, CO. Morgan Kaufmann.

[7] Wang,J. and Jean, J. (1993). Multi-resolution neural networks for omnifont character recognition. In *Proceedings of International Conference on Neural Networks*. Vol.3, pp. 1588-1593.

[8] Bengio, Y., LeCun, Y., Nohl, C., and Burges, C. (1995). LeRec: A NN/HMM hybrid for on-line handwriting recognition. *Neural Computation*, 7(5).

[9] Lawrence, S., iles, C.L., Tsoi, A.C., and Back, A.D. (1997). Face recognition: A convolutional neural network approach. *IEEE Transactions on Neural Networks*, 8(1):98-113.

[10] Lang, K.J. and Hinton, G.E. (1998). A time delay neural network architecture for speech recognition. Technical Report CMU-CS-88-152, Carnegie-Mellon University, Pittsburgh PA.

[11] LeCun, Y., Kanter, I., and Solla, S. (1991). Eigenvalues of covariance matrices: aplication to neural-networks. *Neural Computation*, 3(3).