Attractivity in Heteroassociative Memory Networks for Image Recognition

Niijima, Koichi
Department of Control Engineering and Science Kyushu Institute of Technology

出版情報：RIFIS Technical Report. 64, 1992-10-28. Research Institute of Fundamental Information Science, Kyushu University
バージョン：
権利関係：

RIFIS-TR-CS-64

RIFIS Technical Report

Attractivity in Heteroassociative Memory Networks for Image Recognition

Koichi Niijima

October 28, 1992

Research Institute of Fundamental Information Science
Kyushu University 33
Fukuoka 812, Japan
E-mail: niijima@ces.kyutech.ac.jp

Attractivity in Heteroassociative Memory Networks for Image Recognition

Koichi Niijima
Department of Control Engineering and Science
Kyushu Institute of Technology
Iizuka 820, Japan

Abstract

A heteroassociative memory network for image recognition is constructed with the aid of the method in the paper [5]. This network is a three layered neural network which consists of an input layer, a hidden layer and an output layer. A feature of the network is to contain a sigmoid function only in the hidden units. Images to be stored in the network are real valued vectors. Weights and threshold values connecting the input layer with the hidden layer are determined such that for input reference images, a sufficiently small positive number ε or $1-\varepsilon$ is output at each unit in the hidden layer. Interconnection weights between the hidden and output layers are determined so as to reconstitute the input reference images. This approach makes possible the contraction mapping analysis for the network. As done in [5], domains of attraction in the network are sought. Regions of attraction larger than these domains are also found using the smallness of ε. Furthermore, a certain heteroassociative memory model is designed based on the shape of the fundamental domains of attraction, and successfully applied to recognition of facial images.

1 Introduction

In pattern recognition by neural networks, it is important to study recalling ability of stored patterns in the networks. The ability can be examined, for example, by finding domains of attraction of the stored patterns. There are several researches on domains of attraction in associative memory networks ([1],[2],[3],[4]). However, reference and noisy patterns treated therein are restricted to binary valued vectors. Although images with many gray codes can be represented by binary valued vectors, the dimension of pattern vectors increases extremely. This causes a great increase of input layer units. Recently in [5], we designed an autoassociative memory model recognizable real valued noisy pattern vectors, and found out its domains of attraction based on the contraction mapping analysis. The reason for this success is that the components of stored patterns consist of a sufficiently small positive number ε and $1-\varepsilon$. Along the same line as in [5], one can design autoassociative memory networks by storing reference images with a number of gray codes. However, the contraction mapping principle can not be applied to such networks, because the stored image vectors can take real elements as well as ε and $1-\varepsilon$. As seen from the analysis in [5], domains of

attraction in the network are derivable if ε or $1-\varepsilon$ is output for input reference patterns. This fact suggests that a hidden layer should be added to an autoassociative memory network in order to output ε or $1-\varepsilon$ enforcedly at each hidden unit for input reference images. The last layer is used to reconstitute the input reference images. We thus arrive at a heteroassociative memory network having an input layer, a hidden layer and an output layer with the same number of units as in the input layer. The number of hidden layer units is the same as that of reference images.

Connection weights and threshold values between the input and hidden layers are determined so that for the ν-th input reference image, ε is output at the hidden layer units except the ν-th unit from which $1-\varepsilon$ is output. Interconnection weights between the hidden and output layers are determined so as to output the ν-th input reference image. Although this condition is insufficient for determining these parameters uniquely, polyhedral domains of attraction in the network can be found by virtue of the specified binary outputs at the hidden layer. Larger domains of attraction are also sought using a feedback mechanism of the heteroassociative model. Unfortunately, the shape of such domains can not be clarified because of the nonlinearity of the sigmoid function.

A desirable heteroassociative model can be obtained in the same way as in [5]. That is, we determine unknown weights and threshold values so as to make the polyhedral domains of attraction as large as possible.

In numerical simulations, three facial images are stored in a heteroassociative memory model and its connection weights and threshold values are determined as stated above. Some noisy images are made on the base of the stored images and it is checked which domain they are contained in.

2 Domains of attractivity

We describe our neural network. Let n be the dimension of image vectors and let m be the number of images to be stored. We assume that $m < n$. Our neural network is a three layered network:

$$y_i = \sum_{j=1}^{m} c_{ij} f(\sum_{k=1}^{n} w_{jk} x_k - \theta_j), \qquad i = 1, 2, ..., n. \tag{2.1}$$

Here w_{jk} denote connection weights between the input and hidden layers, and θ_j threshold values. The function f indicates the sigmoid function

$$f(t) = \frac{1}{1 + \exp(-t)}$$

which plays the role that real numbers in $(0, 1)$ are output at the hidden units. The notation c_{ij} denote weights connecting the hidden layer with the output layer. Our network has the same number of hidden units as that of the reference images. Only $\log_2 m$ hidden units suffice to classify the reference images. However, m hidden units are needed to reconstitute the m reference images at the output layer.

We put $W_j = {}^t(w_{j1}, w_{j2}, ..., w_{jn})$, $x = {}^t(x_1, x_2, ..., x_n)$ with the transpose symbol ^{t}, and rewrite (2.1) as

$$y_i = \sum_{j=1}^{m} c_{ij} f(W_j \cdot x - \theta_j), \qquad i = 1, 2, ..., n, \tag{2.2}$$

where the symbol $\cdot$ denotes the inner product. Furthermore, we put $\varphi_j(x) = f(W_j \cdot x - \theta_j), \psi_i(x) = \sum_{j=1}^m c_{ij}\varphi_j(x), \psi(x) = {}^t(\psi_1(x), \psi_2(x), ..., \psi_n(x))$ and $y = {}^t(y_1, y_2, ..., y_n)$ to write (2.2) in the operational form

$$y = \psi(x).$$

On the weight vector W_j and the threshold value θ_j, we impose the condition that for the ν-th input reference pattern, the j-th hidden unit outputs a sufficiently small number ε if $j \neq \nu$, and outputs $1 - \varepsilon$ if $j = \nu$. This condition may be written as

$$f(W_j \cdot x^\nu - \theta_j) = h_j^\nu, \qquad \nu = 1, 2, ..., m, \tag{2.3}$$

where h_j^ν indicates $1 - \varepsilon$ if $j = \nu$, and ε if $j \neq \nu$. When (2.3) is satisfied, we say that the patterns $x^\nu, \nu = 1, 2, ..., m$, are stored in the network. From (2.3), we have

$$W_j \cdot x^\nu - \theta_j = f^{-1}(h_j^\nu), \qquad \nu = 1, 2, ..., m. \tag{2.4}$$

By an easy calculation and using the sign function sgn, we get

$$f^{-1}(h_j^\nu) = \mathrm{sgn}(h_j^\nu - \frac{1}{2})\ln\frac{1-\varepsilon}{\varepsilon}.$$

Therefore, applying the change of variables

$$W_j = V_j \ln\frac{1-\varepsilon}{\varepsilon}, \qquad \theta_j = \eta_j \ln\frac{1-\varepsilon}{\varepsilon}$$

to (2.4), we have

$$V_j \cdot x^\nu - \eta_j = \mathrm{sgn}(h_j^\nu - \frac{1}{2}), \qquad \nu = 1, 2, ..., m. \tag{2.5}$$

We notice that V_j and η_j are almost independent of ε, because each component of x^ν is almost 0 or 1. The number of unknown variables of (2.5) is $n+1$, while that of the conditions of (2.5) is m. Since $m < n$, we can not determine V_j and η_j only by (2.5).

The weights c_{ij} are determined so as to satisfy

$$\sum_{j=1}^m c_{ij} h_j^\nu = x_i^\nu, \qquad \nu = 1, 2, ..., m, \tag{2.6}$$

that is, so as to reconstitute the stored images. We see from the definition of h_j^ν that $H = (h_j^\nu)$ is almost equal to the $m \times m$ unit matrix. Therefore, (2.6) can be solved explicitly to get

$$c_{ij} = x_i^j + O(\varepsilon), \qquad j = 1, 2, ..., m.$$

Combining (2.3) with (2.6) yields

$$\psi(x^\nu) = x^\nu, \qquad \nu = 1, 2, ..., m$$

which implies that x^ν are fixed points of ψ.

Under the conditions (2.5) and (2.6), we have the following theorem.

Theorem 1. Suppose that V_j satisfies (2.5). For any fixed ρ, $0 < \rho < 1$ and each stored pattern x^ν, we define a domain $D_\rho(x^\nu)$ in R^n by

$$D_\rho(x^\nu) = \{x \in R^n \mid \max_{j=1,2,...,m} |V_j \cdot (x - x^\nu)| \leq \rho\}.$$

Then, if ε is sufficiently small for the above ρ, we have, for $x, \tilde{x} \in D_\rho(x^\nu)$,

$$\|\psi(x) - \psi(\tilde{x})\| \leq \kappa \|x - \tilde{x}\|$$

with κ which satisfies

$$\kappa = \sqrt{\sum_{i=1}^{n} (\sum_{j=1}^{m} |c_{ij}| \, \|V_j\|)^2} \; \varepsilon^{1-\rho} \ln \frac{1}{\varepsilon} < 1.$$

Proof. The proof is essentially the same as in Theorem 1 in [5]. From the definition of ψ_i, we get

$$\psi_i(x) - \psi_i(\tilde{x}) = \sum_{j=1}^{m} c_{ij}(\varphi_j(x) - \varphi_j(\tilde{x})). \tag{2.7}$$

Applying the mean value theorem to the term $\varphi_j(x) - \varphi_j(\tilde{x})$ and using the relation $f'(t) = f(t)(1 - f(t))$, we have

$$\varphi_j(x) - \varphi_j(\tilde{x}) = f(z_j)(1 - f(z_j)) W_j \cdot (x - \tilde{x}), \tag{2.8}$$

where

$$z_j = \lambda(W_j \cdot x - \theta_j) + (1 - \lambda)(W_j \cdot \tilde{x} - \theta_j), \qquad 0 < \lambda < 1.$$

In exactly the same way as in the proof of Theorem 1 in [5], we get

$$f(z_j)(1 - f(z_j)) \leq \varepsilon^{1-\rho}.$$

Hence, we have from (2.8),

$$\begin{aligned} |\varphi_j(x) - \varphi_j(\tilde{x})| &\leq \varepsilon^{1-\rho} \ln \frac{1-\varepsilon}{\varepsilon} \, |V_j \cdot (x - \tilde{x})| \\ &\leq \|V_j\| \, \varepsilon^{1-\rho} \ln \frac{1}{\varepsilon} \, \|x - \tilde{x}\|. \end{aligned}$$

Combining this with (2.7), we get the desired result.

This theorem immediately gives

Corollary 1. Let ρ and κ be defined in Theorem 1. Then we have
(i) $x = \psi(x)$ has a unique solution in each domain $D_\rho(x^\nu)$,
(ii) if $x^{(1)}$ is in $D_\rho(x^\nu)$, then a sequence $\{x^{(\ell)}\}$ generated by

$$x^{(\ell+1)} = \psi(x^{(\ell)}), \quad \ell = 1, 2, ...$$

converges to x^ν,
(iii) $D_\rho(x^\nu)$, $\nu = 1, 2, ..., m$, are mutually disjoint.

The assertion (ii) means that any noisy image in $D_\rho(x^\nu)$ can be recognized as the stored image x^ν, that is, the region $D_\rho(x^\nu)$ is a domain of attraction of x^ν.

We next consider the region $D_\rho^k(x^\nu)$ of the vector x such that its k-step recall $\varphi^k(x)$ is in $D_\rho(x^\nu)$:

$$D_\rho^k(x^\nu) = \{x \in R^n \mid \max_{i=1,2,...,n} |V_i \cdot (\varphi^k(x) - x^\nu)| \leq \rho \ \}.$$

This region can be regarded as a domain of attraction of the network by virtue of (ii) in Corollary 1.

Relations among these domains are given in the following theorem.

Theorem 2. Let ρ and κ be defined in Theorem 1. Assume that ε is so small as to satisfy

$$\sqrt{n} \max_{j=1,2,...,m} \|V_j\| \, \kappa \leq \rho. \tag{2.9}$$

Then we have

$$D_\rho^0(x^\nu) \subset D_\rho^1(x^\nu) \subset \cdots \subset D_\rho^k(x^\nu) \subset \cdots,$$

where $D_\rho^0(x^\nu)$ represents $D_\rho(x^\nu)$ defined in Theorem 1.

Proof. It suffices to replace φ by ψ, n by m in the proof of Theorem 2 in [5].

By Theorem 2, the result (iii) of Corollary 1 is extended as follows:

Corollary 2. Suppose that the same conditions as in Theorem 2 are satisfied. Then, for any integer $p, q \geq 0$ and $\nu \neq \mu$, we have

$$D_\rho^p(x^\nu) \cap D_\rho^q(x^\mu) = \phi.$$

3 A heteroassociative memory model

As stated in Section 2, V_j and η_j can not be uniquely determined from (2.5) since $m < n$. According to Theorem 1 and Corollary 1, it is desirable to make large the domains of attraction $D_\rho(x^\nu)$. One way of doing so is to maximize the distances from x^ν to the hyperplanes forming the boundaries of $D_\rho(x^\nu)$. One can easily calculate the distance from x^ν to the hyperplane $|V_j \cdot (x - x^\nu)| = \rho$ as $\rho / \|V_j\|$. This yields the following minimization problem:

$$\|V_j\|^2 \longrightarrow \min \tag{3.1}$$

subject to

$$V_j \cdot x^\nu - \eta_j = \operatorname{sgn}(h_j^\nu - \frac{1}{2}), \quad \nu = 1, 2, ..., m. \tag{3.2}$$

The answer of this problem is given by

Theorem 3. Assume that the stored images x^ν, $\nu = 1, 2, ..., m$, are linearly independent. Let G be the $m \times m$ Gram matrix whose (ν, μ)-element is the inner product (x^ν, x^μ). Denote the m-dimensional vector ${}^t(2, 2, ..., 2)$ by $\boldsymbol{a}$. We put $s_j = {}^t(s_{j1}, s_{j2}, ..., s_{jm})$, where $s_{j\nu} = -2 \operatorname{sgn}(h_j^\nu - \frac{1}{2})$, and $\xi_j = {}^t(\xi_{j1}, \xi_{j2}, ..., \xi_{jm})$. Then the equation

$$\begin{pmatrix} G & \boldsymbol{a} \\ {}^t\boldsymbol{a} & 0 \end{pmatrix} \begin{pmatrix} \xi_j \\ \eta_j \end{pmatrix} = \begin{pmatrix} s_j \\ 0 \end{pmatrix}$$

has a unique solution ξ_j, η_j, and the solution V_j of the problem (3.1) and (3.2) is given by

$$V_j = {}^t(v_{j1}, v_{j2}, ..., v_{jn}),$$

where

$$v_{jk} = -\frac{1}{2}\sum_{\mu=1}^{m} \xi_{j\mu} x_k^{\mu}.$$

Since the proof is almost the same as that of Theorem 3 in [5], we omit it.

4 Numerical simulations

For numerical simulations, the following three facial images are stored in the model designed in the previous section.

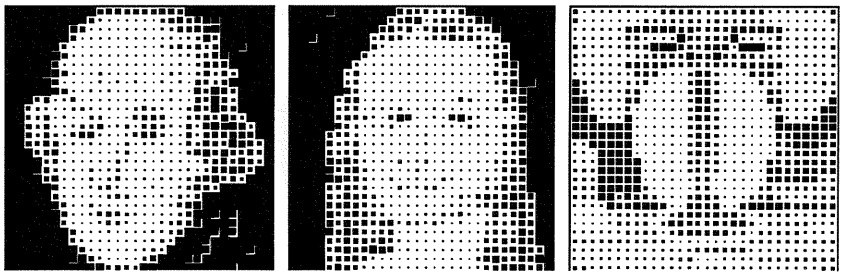

Figure 1: Shakespeare, Mona Liza and Mandrill.

Each stored image consists of 30×30 units, and so it is a 900-dimensional vector. The gray codes of the images were represented by 9 real numbers 0.1,0.2,...,0.8 and 0.9. The area of each square is proportional to these values.

Using these images, we computed the weights w_{jk} and the threshold values θ_j following Theorem 3, where we have chosen $\varepsilon = \exp(-1000)$. Since $n = 900$ and $m = 3$, the number of w_{jk} is 2700 and that of θ_j is 3. The number of c_{ij} is 2700. If $\rho = 0.92$ is chosen, then all the conditions of Theorems 1 and 2 are satisfied.

First, we try recognition of the following noisy image x:

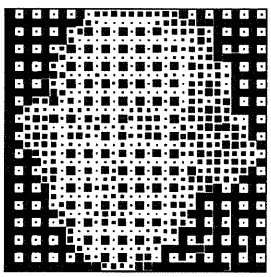

Figure 2: The first noisy image.

We shall check which domain of attraction this x is contained in. For this purpose, we compute $\max_{j=1,2,3} |V_j \cdot (x - x^{\nu})|$ for $\nu = 1, 2, 3$:

ν	1	2	3
	0.643	1.911	1.446

Table 1: The values of $\max_{j=1,2,3} |V_j \cdot (x - x^{\nu})|$.

Since the value for $\nu = 1$ is less than $\rho = 0.92$, this x is in $D_{0.92}(x^1)$. That is, the first noisy image x was recognized as Shakespeare image x^1.

The next noisy image x is

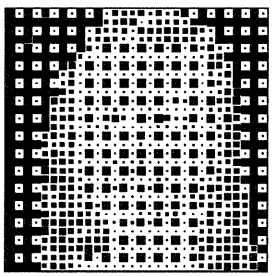

Figure 3: The second noisy image.

For this x, we compute $\max_{j=1,2,3} |V_j \cdot (x - x^\nu)|$ for $\nu = 1, 2, 3$:

ν	1	2	3
	1.677	0.916	1.406

Table 2. The values of $\max_{j=1,2,3} |V_j \cdot (x - x^\nu)|$.

The value for $\nu = 2$ is less than $\rho = 0.92$. Hence, this x belongs to $D_{0.92}(x^2)$ and was recognized as Mona Liza image.

The third noisy image is given by

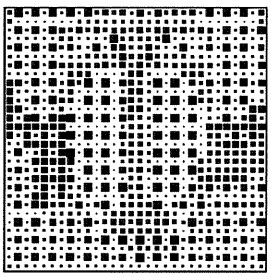

Figure 4: The third noisy image.

The values of $\max_{j=1,2,3} |V_j \cdot (x - x^\nu)|$ for this x are

ν	1	2	3
	1.753	1.854	0.392

Table 3. The values of $\max_{j=1,2,3} |V_j \cdot (x - x^\nu)|$.

This table shows that x is in $D_{0.92}(x^3)$. That is, this x was recognized as Mandrill image.

5 Conclusion

We designed a heteroassociative memory model for image recognition on the base of the contraction mapping principle. The strength of our method is that domains of attraction in the model can be calculated explicitly. Therefore, if a certain noisy image is contained in the domains, then it can be recognized as one of the stored images. In numerical simulations, considerably noisy patterns succeeded in recognition because of only 3 stored images.

As easily seen from Theorems 1 and 2, the shape of the domains of attraction does not depend on the stored patterns. It is natural that the shape of domains of attraction varies together with stored patterns. The reason seems to lie in the enforced output of the same number ε at almost all the hidden units. The enforced output at the hidden units except one unit suffices to be any positive number smaller than ε. An approach under such a loose condition has a possibility of discovering domains of attraction depending on stored patterns.

References

[1] Amari,S., Characteristics of Sparsely Encoded Associative Memory, *Neural Networks*,Vol.2, 1989, pp.451-457.

[2] Amari,S. and Maginu,K., Statistical Neurodynamics of Associative Memory, *Neural Networks*,Vol.1, 1988, pp.63-73.

[3] Cottrell,M., Stability and Attractivity in Associative Memory Networks, *Biological Cybernetics*,Vol.58, 1988, pp.129-139.

[4] McEliece,R.J., Posner,E.C., Rodemich,E.R., and Venkatesh,S.S., The Capacity of the Hopfield Associative Memory, *IEEE Transactions on Information Theory*,Vol.33, 1987, pp.461-482.

[5] Niijima,K., Domains of Attraction in Autoassociative Memory Networks for Character Pattern Recognition, *Proceedings of the Third Workshop on Algorithmic Learning Theory, ALT'92*, 1992, pp.87-98.