

神経回路網におけるスパース化と記憶容量について On the Capacity and Sparse Encoding Scheme in Neural Networks

赤穂 昭太郎
Shotaro AKAHO

電子技術総合研究所
Electrotechnical Laboratory

Abstract

神経回路網の能力を記憶容量という観点から眺め、未解決問題である3層回路網の記憶容量の下限の計算を中心に考察する。また記憶パターンのスパース化が記憶容量に対してどのように働くかについても述べる。

1. ま え が き

神経回路モデルの記憶容量を考えることはそのモデルの能力をはかる上で興味深い問題であり、従来からいろいろな手法を用いて研究されてきた。(Appendix C 参照)

一方、記憶に伴う学習について考えてみると、脳は外界からの知識を例として与えられ、知識を蓄えるとともに、さらにその知識を含むより一般的な法則(関数)を学習する。このとき、どれだけの例を与えれば十分な一般的な法則を学び取ることができるかを明らかにする必要がある。これは“帰納的学習における汎化能力”という問題で、これについては神経回路網を含むいろいろな枠組から最近活発に議論されはじめた[14, 13, 5]。与えるべき例の数はモデルの記憶容量、つまりモデルがどれだけのパターンを分類する能力があるかに関係していて、記憶容量が増えれば増えるほど与えるべき例の数を増大させなければならない。

さて記憶容量に関しては、これまでに単一のニューロン(しきい素子)の容量が $2n$ (ただし n は入力次元)であることはCover[6]によって示された(Appendix B)。これはそのまま2層回路の容量となる。

一方、中間層をもつ3層回路網の容量 $C(n, h, k)$ (ここで n, h, k はそれぞれ入力層、中間層、出力層の素子の数とする)については、最近 Mitchison ら[9]によって次式の上限が与えられただけである。

$$C(n, h, k) < O\left(\frac{nh}{k} \log \frac{h}{k}\right) \quad (1)$$

しかし計算の上からもシミュレーションの結果によっても、実際の容量はこの上限より少ないことがわかっている。

本報告では記憶容量 m が、

$$m = O\left(\frac{nh}{\log(nh^2)}\right) \quad (2)$$

であたえられるモデル (ただし $k < \log_2 h$) を構成することによって $C(n, h, k)$ の下限を与える。

このモデルで中心的役割を果たすのはパターンベクトルのスパース化である。スパースなベクトルとは、その成分のほとんどが 0 であって 1 である成分の数が非常に少ないものをいう。連想記憶モデルにおいて、記憶させるパターンをスパースにすると記憶容量が著しく増加することが知られている [15, 12, 3]。またスパースなパターンの存在は海馬 (記憶の制御機能を果たしていると考えられる脳の部位) における生理学的知見からも得られている [10]。

本報告で示すモデルでは入力パターンをグループ化することによって中間層のパターンを減らすとともに、中間層のパターンのスパース化によって全体の容量を増加させることができる。まず 2 節、3 節、4 節でモデルについて説明し、5 節でモデルの容量を求める。

2. 3 層神経回路網

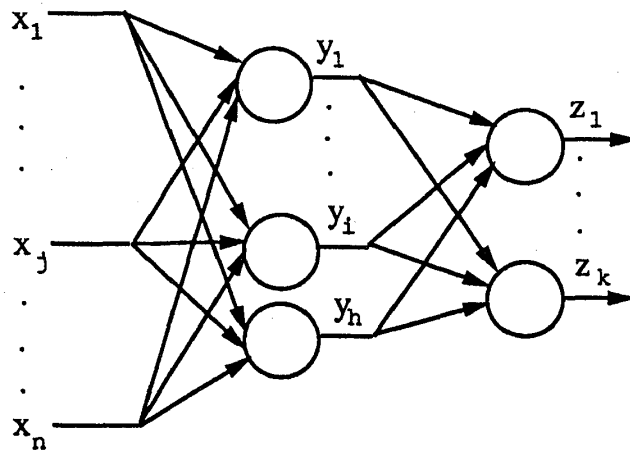


図 1. 3 層回路網

図 1. に、3 層回路網を示す。入力 n 次元、中間層及び出力層のユニット (しきい素子) の数をそれぞれ h, k とする。入力ベクトル

$$\mathbf{x} = (x_1, \dots, x_n)$$

は変換 $\mathbf{y} = T_w \mathbf{x}$ によって中間層のベクトル

$$\mathbf{y} = (y_1, \dots, y_h)$$

に変換される。成分で書けば

$$y_i = f\left(\sum_j w_{ij} x_j - \theta_i\right) \quad (3)$$

となる。ここで $f(u)$ は

$$f(u) = \begin{cases} 1 & u \geq 0 \\ 0 & u < 0 \end{cases} \quad (4)$$

なる階段関数とする。中間層のベクトル $\mathbf{y}$ はさらに変換 $\mathbf{z} = T_v \mathbf{y}$ によって出力層のベクトル

$$\mathbf{z} = (z_1, \dots, z_k)$$

に変換される。成分で書けば

$$z_l = f\left(\sum_i v_{li} y_i - \psi_l\right) \quad (5)$$

となる。

この回路に m 個の入力パターン $\mathbf{x}^\mu (\mu = 1, \dots, m)$ と、対応する出力パターン $\mathbf{z}^\mu (\mu = 1, \dots, m)$ を記憶させる。つまり w_{ij} や v_{li} を適当にとって

$$\mathbf{z}^\mu = T_v(T_w(\mathbf{x}^\mu)) \quad (\mu = 1, \dots, m) \quad (6)$$

となるようにする。 $\mathbf{x}_i^\mu, \mathbf{z}_i^\mu$ がランダムに決められるとき、(6) 式が成立する m の最大値を記憶容量と呼ぶ。(Appendix A 参照)

3. 中間層のスパース化

中間層にあらわれるパターンは、 $\mathbf{x}^\mu$ によっていろいろな値をとりうるが、ここではスパースな s 個のベクトル $\mathbf{y}^\alpha (\alpha = 1, \dots, s)$ をとる。スパースなベクトルとは $c \ll n$ なる整数 c をとって、ベクトルの成分のうち c 個だけが 1 で残りは 0 であるものをいう。スパースなパターンは海馬などでも見つかっており、また連想記憶モデルにおいて、出力パターンをスパースに決めると記憶容量や情報容量が増大することが知られている。 $\mathbf{y}^\alpha$ のとりかたはスパースであればランダムでよいが、ここでは

$$s = h \quad (7)$$

として次のように決めることにする。

$$y_i^\alpha = \begin{cases} 1 & i \equiv \alpha, \dots, \alpha + c - 1 \pmod{h} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

すると次の補題が成り立つ。

補題 1. 行列 $Y = (y_i^\alpha)$ は h と c が 1 以外の共通因数をもたないときに限って正則となる。

ここで出力パターンの場合の数がただか h であるという制限をつける。つまり、出力素子の数 k が、

$$2^k \leq h \quad (9)$$

であるとする。このとき、 Y が正則であれば、中間層と出力層の間のシナプス荷重が存在する。つまり出力パターン $\mathbf{z}_i^\alpha$ が 0, 1 の 2 値をとるとき、

$$z_i^\alpha = v_i \cdot \mathbf{y}^\alpha \quad (10)$$

なる h 次元ベクトル $\mathbf{v}_i$ が存在し、それを用いれば

$$z_i^\alpha = f\left(\mathbf{v}_i \cdot \mathbf{y}^\alpha - \frac{1}{2}\right) \quad (11)$$

と書ける。以下この補題の成り立つ範囲で考える。 $(\mathbf{y}_i^\alpha)$ を上のように決めなくても漸近的には確率 1 で中間層から出力層への結合は存在する。従ってこの仮定は本質的なものではなく通常はランダムに決めて良い。2 層回路の容量を考えると s としては $2h$ までとれる。

4. 入力パターンのグループ化

入力パターンの各成分 x_i^μ はそれぞれ独立に、標準正規分布 $N(0,1)$ に従うとする (これは x_i^μ をそれぞれ独立に、1, -1 を確率 1/2 でとるとしたときと結果的に全く同じになる). ここで $X = \{x^\mu\}$ を、同じ出力パターン z^α をもつもの同士 $M = m/h$ 個ずつのグループ $X^1, \dots, X^h$ に分割する. つまり,

$$X = \bigcup_{\alpha=1}^h X^\alpha, \quad (12)$$

ただし,

$$X^\alpha = \{x^{\alpha,1}, \dots, x^{\alpha,M}\} \quad (13)$$

とする. ここでグループ X^α に入っているパターンを $x^{\alpha,\nu}$ ($\nu = 1, \dots, M$) とおいた.

以下ではグループ X^α に属するパターン $x^{\alpha,\nu}$ から y^α を想起するように重みを決める. そうすれば補題 1 より, $x^{\alpha,\nu}$ から z^α が想起できることになる. 図 2 にパターンの写像関係を示す.

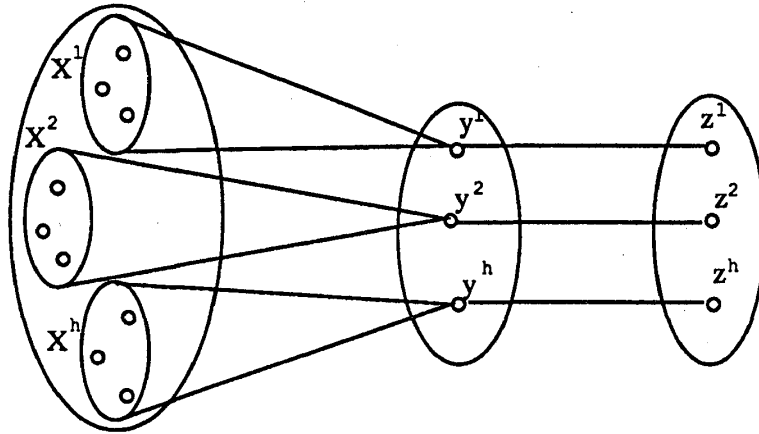


図 2. パターンの写像

X^α から y^α への想起は、単純かつ自然なモデルである相関型の連想記憶モデル [11, 8] で行ない、重みは

$$w_{ij} = \frac{1}{n} \sum_{\alpha=1}^h \sum_{\nu=1}^M y_i^\alpha x_j^{\alpha,\nu} \quad (14)$$

ととる.

5. モデルの容量

本節では、2節から 4節で述べたモデルの容量を計算する.

グループ X^1 の第一要素 $x^{1,1}$ からパターン y^1 が想起できるのは、 $u_i = \sum_j w_{ij} x_j^{1,1}$ として、 $y_i^1 = f(u_i - \theta)$ となる場合である.

$$\begin{aligned} u_i &= \sum_j w_{ij} x_j^{1,1} \\ &= \frac{1}{n} y_i^1 x^{1,1} \cdot x^{1,1} + \frac{1}{n} y_i^1 \sum_{\nu \neq 1} x^{1,\nu} \cdot x^{1,1} \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{n} \sum_{\alpha \neq 1} y_i^\alpha \sum_{\nu \neq 1} x^{\alpha, \nu} \cdot x^{1, 1} \\
& = (1 + L_i) y_i^1 + N_i.
\end{aligned} \tag{15}$$

ここで,

$$L_i = \frac{1}{n} \sum_{\nu \neq 1} x^{1, \nu} \cdot x^{1, 1}, \tag{16}$$

$$N_i = \frac{1}{n} \sum_{\alpha \neq 1} y_i^\alpha \sum_{\nu \neq 1} x^{\alpha, \nu} \cdot x^{1, 1}. \tag{17}$$

中心極限定理により L_i, N_i は漸近的に正規分布とみなせる. 平均はともに 0 であり, 分散は,

$$V[L_i] = \frac{M}{n}, \tag{18}$$

$$V[N_i] = \frac{Mc}{n}. \tag{19}$$

となる. 簡単のためしきい値を $1/2$ とすれば, y_i^1 の条件付誤り確率は,

$$Pr(y_i^1 \neq f(u_i - \frac{1}{2}) | y_i^1 = 1) = F(\frac{1}{2\sigma}) \tag{20}$$

$$\begin{aligned}
Pr(y_i^1 \neq f(u_i - \frac{1}{2}) | y_i^1 = 0) & = F(\frac{1}{2V[N_i]}) \\
& < F(\frac{1}{2\sigma})
\end{aligned} \tag{21}$$

となる. ただしここで,

$$\sigma^2 = V[L_i] + V[N_i] = \frac{M(1+c)}{n} \tag{22}$$

であり, $F(u)$ は正規分布の分布関数

$$F(u) = \int_u^\infty \frac{\exp(-u^2/2)}{\sqrt{2\pi}} \tag{23}$$

である. 従って全てのパターンについて, 誤って想起される成分の全個数の期待値は,

$$E = mhF(\frac{1}{2\sigma}) \tag{24}$$

で上から抑えられる. ここで, u が十分大きいとき,

$$F(u) \simeq \frac{\exp(-u^2/2)}{\sqrt{2\pi}u} \tag{25}$$

となることをもちいると,

$$E \simeq \exp(-\frac{1}{8\sigma^2} + \log(mh) + O(\log \sigma)) \tag{26}$$

となるから $E \rightarrow 0 (m, n \rightarrow \infty)$ となるのは (22) 式より

$$m < \frac{nh}{8(c+1)\log(nh^2)} \tag{27}$$

のときであることがわかり, (27) 式の右辺がモデルの容量を表し, $c=1$ とおけば 3 層回路の容量の下限を与える.

定理 1. 3 層回路の記憶容量 $C(n, h, k)$ は, 次の下限を持つ.

$$C(n, h, k) > \frac{nh}{16 \log(nh^2)} \quad (28)$$

容量の導出の際, しきい値を計算が容易になるように決め, 成分の誤り個数の評価も上から抑えたのでモデルの容量は (28) 式で得られた値よりも多くなる. しかしこれらのことを考慮に入れても定数倍増加するだけでオーダーは変わらない.

6. モデルの考察

本報告で述べたモデルは非常に単純であり, 実際の 3 層回路の容量はこれより多くなるだろうと思われる. 問題点をいくつか挙げると, まず第 1 に, 中間層のパターンはスパースであるという仮定を満たせばランダムまたは適当に固定してとって良かった. 第 2 には, 出力素子の数が非常に少ない場合 (例えば $k = 1$ のとき), 同じ出力を持つ入力パターンの集合も, 異なるいくつかのグループに分ける必要があるが, ここで示したモデルでは全くランダムに分けて良かった. 第 3 に, 入力層から中間層の間で相関型の連想記憶を行なっている. 容量をさらに増加させるような他の結合が存在するかも知れない.

第 2 点に関して, k が大きい場合 ($\log_2 h$ 程度) を考えると, このモデルの記憶容量は良い下限を与えていると思われる. 実際 $k = \log_2 h$ のとき, Mitchison らの与えた上限の値は $O(nh)$ となる. 実際の容量はこれより少ないはずだから, 定理 2 で与えた容量は実際の容量に近いと思われる. このことは, k が大きくなるとグループわけのしかたに自由度がなくなることが一つの原因かも知れない. 従って $k = 1$ の場合にはグループ分けをうまくやって容量を増やすことができる可能性もある.

次に, このモデルが情報処理システムとして持つ意味について考えてみる. 画像や言葉といった個々の“事物”(または“事象”)がシステムに入力され, それに対する“行動”(または“反応”)を出力するシステムを考える. これを 2 層の回路で実現しようとする, $2n$ 個 (n は入力ベクトルの次元) の事物およびそれにたいする行動だけしか記憶できないことになる. 一方, 3 層回路にして, 中間層の各素子に事物の“種類”(または“(抽象)概念”)をわりあて, 事物と種類との関係を入力層と中間層の間で記憶し, 種類と行動との関係は中間層と出力層の間で独立して学習を行なうことにする. すると種類の数としては中間層の素子の数だけ記憶でき, 1 つの種類に対して $O(n/\log(nh^2))$ 個の事物を対応させられ, 全体として $O(nh/\log(nh^2))$ 個の事物を記憶することができることになる. またこの場合, 同じ種類に属する事象同士は全く相関を持っていなくてもよい. このモデルは従来から言われている“おばあさん細胞”, つまり個々の事物に対して 1 つの細胞がわりあてられるとするモデルとは似て非なるものであることがわかる.

7. むすび

記憶容量が $O(nh/\log(nh^2))$ で与えられるモデルを構成し, 3 層回路の容量の下限を与えた. このモデルで重要な役割を果たすのは, 中間層のパターンのスパース化であることをしめしたが, 容量が最大となるのは非常にスパースな場合, つまり中間層の素子のうち 1 つだけが発火するようなパターンのときであることは興味深い. 残された問題としては, 出力層の素子が少ない場合

に入力パターンのグループ化をうまくやることで(オーダーの意味で)容量の増大があるかどうかを解析することである。

本報告で提案したモデルは非常に単純なモデルであるから3層回路の容量は実際にはこれより多くなると思われる。しかし一般に3層回路にパターンを記憶させる学習法は(バックプロパゲーション法を含め)、そのポテンシャルが数多くのローカルミニマムをもっている。本報告では構成的にモデルを作って記憶容量を増加させることを考えたが、3層回路の実際の容量と同じ個数のパターンを記憶させるにはローカルミニマムに落ちにくい学習法を工夫する必要がある。

また、帰納的学習において、十分な一般化を実現するためには、記憶容量が増えるに従って与える例の数も増やさなければならない。従って学習する関数の複雑さに応じて、適当な回路網のサイズがあるはずである。そのような問題も今後の課題として残されている。

(本報告は参考文献 [2] より一部転載した)

参考文献

- [1] S. Akaho and S. Amari: On the capacity of three-layer networks. In *Proc. of IJCNN '90*, volume III, pp. 1-6, 1990.
- [2] 赤穂、甘利: スパース符号化を用いた3層ニューラルネットの容量の下限の計算. 信学技法 NC89-73, 1989.
- [3] S. Amari: Characteristics of sparsely encoded associative memory. *Neural Networks*, Vol. 2, No. 6, pp. 451-457, 1989.
- [4] 甘利 俊一: 神経回路網の数理. 産業図書, 1978.
- [5] E.B. Baum and D. Haussler: What size net gives valid generalization? *Neural Computation*, Vol. 1, pp. 151-160, 1989.
- [6] T.M. Cover: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.*, Vol. 14, pp. 326-334, 1965.
- [7] E. Gardner: The space of interactions in neural network models. *J. Phys. A: Math. Gen.*, Vol. 21, pp. 257-270, 1988.
- [8] T. Kohonen: Correlation matrix memories. *IEEE Trans. Comput.*, Vol. 21, pp. 343-359, 1972.
- [9] G.J. Mitchison and R.M. Durbin: Bound on the learning capacity of some multi-layer networks. *Biol. Cybern.*, Vol. 36, pp. 646-658, 1989.
- [10] Y. Miyashita and H.s. Chang: Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, Vol. 331, pp. 68-70, 1988.

- [11] K. Nakano: Associatron — a model of associative memory. *IEEE Trans. Sys. Man Cybern.*, Vol. 2, pp. 381–388, 1972.
- [12] G. Palm: On the storage capacity of an associative memory with randomly distributed storage elements. *Biol. Cybern.*, Vol. 36, pp. 646–658, 1981.
- [13] L.G. Valiant: A theory of learnable. *Comm. ACM*, Vol. 27, No. 11, pp. 1134–1142, 1984.
- [14] V.N. Vapnik and A.Ya. Chervonenkis: On the uniform convergence of relative frequencies of events to their probabilities. *Theory Prob. Appl.*, Vol. 16, pp. 264–280, 1971.
- [15] D.J. Willshaw and H.C. Longuet-Higgins: Associative memory models. In B. Meltzer and O. Michie (eds.), *Machine Intelligence*, Vol. 5. Edinburgh University Press, 1970.

Appendix A. 記憶容量の定義について

記憶容量の直感的な定義を一言で言えば“与えられたシステムで識別可能な入力ベクトルの最大個数”となる。ここではこれをもう少し厳密に論じる。

厳密に言えば記憶容量の定義には次の 2 つの立場がある。

一つはどのような入力ベクトルが与えられてもその個数だけは絶対に識別できるとする立場である（この意味での容量を絶対容量と呼ぶことにする）。式で書けば

$$\begin{aligned} C &= \max m \\ \text{s.t. } &\text{任意の } \mathbf{x}_i^\mu, \mathbf{z}_i^\mu \text{ に対し } W \text{ が存在し,} \\ &\mathbf{z}_i^\mu = T_W \mathbf{x}_i^\mu \quad (\mu = 1, \dots, m). \end{aligned} \tag{29}$$

つまり入力ベクトル $\mathbf{x}^\mu$ と出力ベクトル $\mathbf{z}^\mu$ とをどんな意地悪く選んでもパラメータ W が存在して $\mathbf{x}^\mu$ から $\mathbf{z}^\mu$ を出力することができるという定義である。これは組合せ論的立場からの定義であって確率論的手法には馴染みにくい。

そこでもう一つの立場として、入力ベクトルや出力ベクトルはある確率分布に従ってランダムに決まっていると見て（たとえば各成分が独立に $1/2$ ずつの確率で 0 または 1 をとるとして）、確率 1 で識別できるとする定義法があり、本報告ではこちらの立場をとる。式で書けば、

$$\begin{aligned} C &= \max m \\ \text{s.t. } &\text{Prob}[W \text{ が存在する; } \mathbf{z}_i^\mu = T_W \mathbf{x}_i^\mu \quad (\mu = 1, \dots, m)] = 1 \end{aligned} \tag{30}$$

となる。

Appendix B. 単一のニューロンの記憶容量 [6]

ここでは単一ニューロンの記憶容量が $2n$ であることの証明の概略を述べる。

今 m 個の入力ベクトル (n 次元) が与えられたとし、それぞれ 0 を出力すべきか 1 を出力すべきかが決まっているとす。そこで、0 を出力すべき入力ベクトルの集合を S_0 , 1 を出力すべき入

力ベクトルの集合を S_1 とおく. 単一ニューロンは n 次元の入力空間を超平面で切って, 片方の半空間に属する入力に対して 0 を 他方の入力に対しては 1 を出力する. つまり S_0 と S_1 とは線形分離可能でなければならない. S_0 と S_1 の与えられ方の総数は 2^n であるが, そのうち線形分離可能である場合の数 $L(n, m)$ はどれだけあるかを考える. 入力ベクトル集合が一般の位置にあれば, 実は入力ベクトルの配置によらず

$$L(n, m) = 2 \sum_{i=0}^{n-1} \binom{m-1}{i} \quad (31)$$

となることがわかっている. 従って $m \leq n+1$ ならば常に線形分離可能となり, 絶対容量は $n+1$ である. また S_0 と S_1 の与えられ方が全くランダム (2^n 個の場合が等確率で起きる) とすれば, 線形分離可能である確率は n が十分大きい時, $m/n < 2$ で 1, $m/n > 2$ で 0 となるため, 確率論的に決まる容量は $2n$ となる.

Appendix C. 神経回路網における解析手法

本報告の計算においては確率論的な手法を用いたが, そのほかに Cover が用いた計算幾何学的手法はより厳密な結果を得る方法として考えられる. しかし計算幾何学的手法はその厳密性ゆえに, 一般に解析が難しい.

一方, 統計物理学的なアプローチは強力な解析手法として従来から用いられてきた. 種々の近似手法によって, 特に神経回路網のダイナミクスに関しては多くの興味ある解析結果がある.

記憶容量に関しても, 最近, Gardner[7] らによって統計物理学的な解析がなされた. そこで用いられている近似手法には鞍点法のほか, スピングラスとのアナロジーからレプリカ法や, 平均場近似を用いている. そうして得られた解析結果 (特にレプリカ対称解) は, 定量的に疑問が残る点もある. レプリカ法や平均場近似の数学的基礎を明らかにする必要がある. また神経回路網の特長の一つである学習による自己組織化に関しても統計物理学的な解析が期待される.

しかしながら, 統計物理学としての神経回路網と情報処理システムとしてのそれとは, 興味の対象が微妙にずれやすいという問題もある. 統計物理学は物理的視点からより現象をうまく説明できるモデルを構築し, その数学的な性質を調べることに重きをおいてきた. 一方情報処理システムとしての神経回路網研究は究極的には脳内の現象を説明するという目標があるにしろ, 一旦実際の脳とは距離をおいて, 高い能力をもつ情報処理システムとは何かということを追求め, そのためのモデル構築およびその数学的解析に重きをおいてきた. そのような視点においても統計物理学的手法は強力な解析手段となり得るであろう.