アトラクタニューラルネットワークモデルによるカテゴリー特異性の検討

An Investigation of the Category Specifiicity by the Attractor Neural Netowrk

浅川 伸一
Shin-ichi Asakawa

東京女子大学
Tokyo Woman's Christian University
asakawa@ieee.org

Abstract

The nature of semantic memory was focused in this study. The possibility for the attractor neural network to explain the double dissociation between animate and inanimate objects in neuropsychology was discussed. In spite of the simplicity of the model, this model can describe several symptoms of the deficits. This might be the one of the major advantages of this model. Although further discussions are still required, this model could account for the confusion matrix and the delay of the reaction times. The description of each items founded on the micro–feature might be useful to describe this symptoms.

Keywords — Semantic memory, Category Specificity, Attractor Neural Network, double–dissociation between animate and inanimate objects

1. 神経心理学におけるカテゴリー特異性

従来から([9, 6, 8, 10, 7])，動物概念と非動物概念の二重乖離の症例が数多く報告されてきた。すなわち，動物を区別できない脳損傷患者で，非動物の概念は正常の範囲である患者がいる一方で，全く正反対に，動物の概念は保たれているにもかかわらず，非動物の概念，たとえば，道具であるとか，体の一部，屋外にある物品などの概念が崩壊している患者が存在する。

現在までに分かっていることを列挙すると以下のようになるだろう。

1. 動物に関する意味記憶は，機能的記憶が欠如していたとしても損傷されうる。
2. このことは，Farah と McClelland[3]，日本語の紹介については[11] のシミュレーションによって支持されている。
3. 機能的的な知識に問題がないにもかかわらず，動物に関する知識に障害を示す患者がいる[1]
4. このカテゴリー特異性から脳内での動物と非動物との処理の違いが導き出せるかもしれない。
5. 知覚的な，あるいは，機能的な，意味記憶は，脳内で別々に保持されている可能性が考えられる。それゆえ，限局した脳損傷はカテゴリー特異性を生じることが示唆される。
6. カテゴリー特異的な障害は，下側頭回の異なる部位に損傷が起こることによって生じると考えられている。

しかし，意味記憶の知覚的，機能的な側面の相違が動物—非動物の意味記憶の差異を生じる要因であったしても，脳内の情報，すなわち限局した脳損傷が本当にカテゴリー特異性を生じるのかを問うてみる価値はあると考える。意味記憶におけるカテゴリー特異的な障害は，カテゴリー内の構造と内容とが異なっていることを示しているように思われるからである。

従って，意味記憶の表象は，同一カテゴリー内で，各アイテムがどのように検索されるによって変化する。一般的に動物は意味的特徴を共有する形で脳内に保持されており，非動物は動物に比べて，より弁別特徴を持っているということができよう。特徴が共起する強度は，各特徴間の関係によって定まり，動物概念は非動物概念に比べて，高い特徴間相関を持つと考えて良いように思われる。

2. ニューラルネットワークによる表現

ニューラルネットワークを神経心理学的症状の理解のために役立てることを考える。すなわち，一旦学習が成立したニューラルネットワークに対して，部分的にコネクションやユニットを取り除くことにより，実際の脳損傷のシミュレーションを行い，患者のパフォーマンスと比較することによって，我々の脳内で起こっていることを類推しようとする試みである。

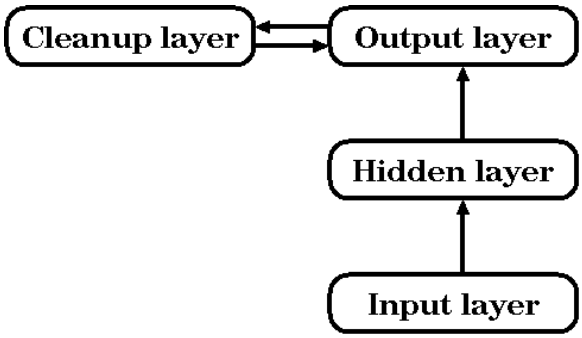

図 1 .アトラクタネットワーク[4]

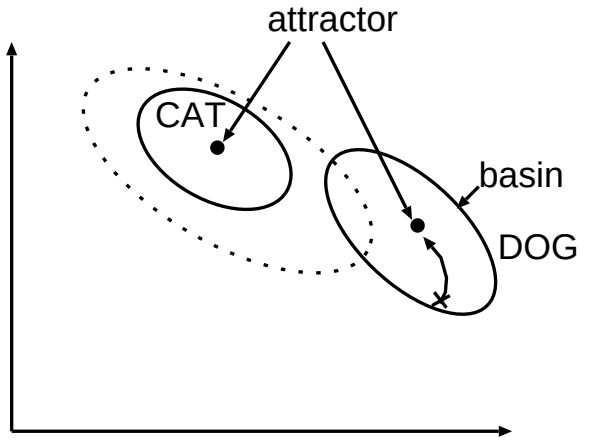

図 2 .アトラクタネットワークの流域

2.1 アトラクタネット

アトラクタネットワークはとは，図1にあるように，3 層のニューラルネットワークにに加えてクリーンナップ層と呼ばれる層が出力層と相互作用するように設計されている。情報は出力層とクリーンナップ層との間で何度か繰り返しやり取りされ，正解に達すると出力層から出力がなされて終わる。あるいは，繰り返しの上限に達して，それ以上進まなくなる。学習は適当な初期値から始めて，繰り返しの都度，勾配降下則が適用されて学習が進む。クリーンナップ層のユニット数には制限がない。出力層のユニット数は問題によって，決められていることが多いが，クリーンナップ層のユニットは，出力層のユニット数と同じでも良いし，もっと多くても良い。あるいは，少なくても，たった一つでも良い。出力層とクリーンナップ層との間の繰り返し数を正解に達するまでの反応時間と同一視することも可能であろう。その場合，繰り返しが多ければ時間がかかったことを意味する。一般には，中間層の数が適切であれば，アトラクタネットワークは，その特別な場合として 3 層パーセプトロンを含むことに注意されたい。アトラクタネットワークを意味記憶のモデルとして捉えるのなら，入力パターンを動物か非動物かを判断するような課題においては，正常な状態（健常者のそれを表すと考えられる）では，クリーンナップ層との相互作用を必要としないと考えられる。クリーンナップ層が使われるのは，もっぱら，システムのどこかに異常が起こった場合であると解釈可能であろう。

このような意味で，このネットワークはアトラクタを持つという。持っているアトラクタがポイントアトラクタだけであるか否かは，データの与え方にもよる。単なる分類課題であれば，ポイントアトラクタが形成させることとなる(図2)。アトラクタネットワークは，損傷が起こった場合，正解に至るための範囲が決まっている。この範囲（流域 basin）が，損傷よって変化し，正解に至ることができなかったり，正解に至ったとしても，繰り返し回数が増加する。これを反応時間の遅延と同一視できる(図2)などから，他のモデルにはないアドバンテージを持っているといえる。実際，アトラクタネットワークの処理能力は高い。図3に示したように，排他的論理和の拡張であるパリティ

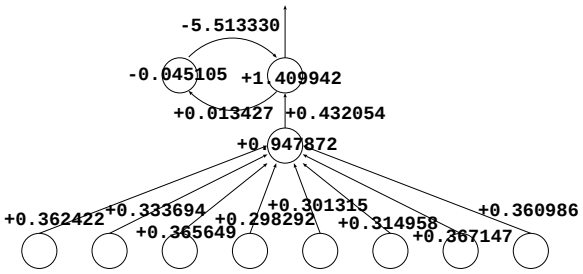

図 3 . 8 bits parity 問題を解くアトラクタネットワーク

問題（この場合 8 ビットパリティ）を，たった一つの中間層とたった一つのクリーンナップ層で解くことができる。このことは，排他的論理和を解くためには，最低 2 個の中間層素子が必要であった 3 層のパーセプトロン型のネットワークと比較して，アトラクタネットワークは本質的に異なるといえる。ここでは，アトラクタネットワークの持つ利点を生かして，神経心理学的データを説明できるのかを問うこととしたい。

3. ニューラルネットワークモデルによる検討

先述したとおり，神経心理学的データについての問題の一つは，カテゴリー効果，あるいはカテゴリー特異性である。Warrington らが1980年代に，カテゴリー特異的な症例を報告して以来，意味記

憶の構造について様々な意見やモデルが提案されてきたが，未だに解決したとは言いがたい。患者の課題成績を比較すると，意味記憶は，おおまかに，生物と非生物という2種類に分類されているらしい。先に述べたとおり，相互に排他的な二重乖離が観察されるからである。生物と非生物との概念の二重乖離によって，意味記憶の構造がどのように創発するのかが焦点である。おそらく，特定の意味記憶の知識は，脳内で分散して貯蔵されているのであろう。どのように分散していれば，上述のカテゴリー特異性が創発するのであろうか。本研究では，カテゴリー特異性障害を説明するための意味記憶表象として，生物–非生物，あるいは知覚的–機能的のような二分法に頼ることはしない。その変わり，個々のアイテムの弁別性と相関性に基づいてデータを表現した。このデータ表現方法は [2]のデータ表現を踏襲したものである。このようにして表現されたデータに対して，[5]に基づき，以下のような特徴を持たせた。すなわち，

1. 特徴の特殊性:意味表象は，同一カテゴリー内の他の表象から，いかに検索されるかにおいて，さまざまに変動する。一般に生物は，より共有される意味特徴を持ち，人工物と比べてより少ない弁別特徴を持つ。
2. 相関関係:共起する特徴は，相互活性化によって相互に強化される。生物概念は，人工物概念より高い相関のある特徴が多い。

Tyler らの用いた刺激の相関係数行列を視覚化したものを図4に示す。図から読み取れるように，人

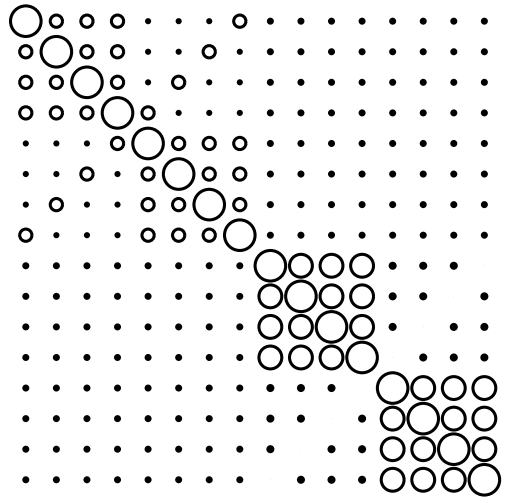

図 4 Tyler らの刺激の相関係数行列。正の相関を白丸で，負の相関を黒丸で示した。相関係数の大きさは，丸の大きさで表してある。図中の左上が人工物，右下が生物概念である。

工物（図中左上）ではカテゴリー内の相関係数は，生物概念（図中右下）よりも小さい。Tyler らは，恒等写像，すなわち入力層に提示される入力パターンを出力層で正確に再現できることをもって，概念を獲得したと定義している。

3.1 実験条件

しかし，これ以外にも正当パターンが考えられる。すなわち，

1. 生物パターンか，非生物パターンかを 1, 0 で表現した場合。この場合 16 行 2 列の行列となる。これを category 条件とする。
2. Unitary 条件。恒等写像であり，入力層のパターンと同じパターンを学習する。入力行列とターゲット行列とは，16 行 24 列の行列となる。
3. identical 条件。データのそれぞれを同定させるもの。合 16 行 16 列のデータ行列で対角要素がすべて 1 の単位行列となる。

いずれのパターンもそれなりの意味を持つ。category 条件では，個々の概念の上位概念を学習させるものであり，Unitary 条件は，それぞれのメンバーを正確に同定することを求めるものである。また，identical 条件は，各アイテムをそれぞれ同定することを意味すると考えられよう。

ここでは，中間層のユニット数を 10，クリーンアップ層のユニット数を 1，出力層とクリーンアップ層の間での繰り返しの上限を 20 とした。学習は誤差伝播学習法に従い学習係数を 0.01 とした。各結合係数の初期値は –0.1 から 0.1 までの一様乱数とした。学習の成立条件は，個々のパターンの二乗誤差がすべて 0.05 以下になったこともって学習成立とした。図5 は，アトラクタネットに Tyler et.al.の刺激を学習させた場合の学習成立までの繰り返し数の平均である。ここでは，平均自乗誤差が 0.05 以下になることをもって収束とみなすことをせず，各項目毎の自乗誤差がすべて 0.05 以下となることを学習成立の基準とした。なぜなら，各項目が，それぞれの意味概念に相当すると考えるのならば，平均して正解基準に達することの意味が曖昧だからである。それよりも，すべての概念をすべて体得したことを持って学習成立とした方が現実に近いと考える。

学習の容易さから見ると，動物と非動物とに分けるカテゴリー条件が最も簡単で，その次に unitary 条件, 最後に identical 条件となる。このことは逆に，カテゴリー内での混同が起こりやすいことをも意味しているように思われる。

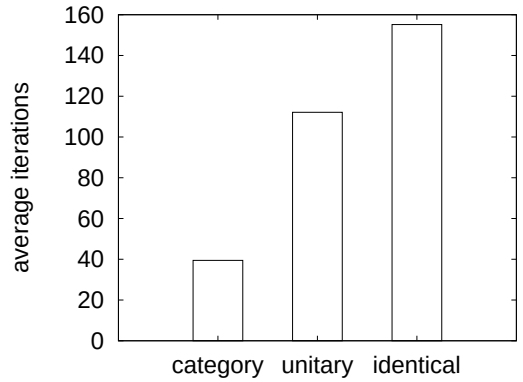

図 5 学習が成立するまでの平均繰り返し数。各学習項目の 2 乗誤差が0.05 以下になることをもって，学習が成立したとみなした。

3.2 破壊実験

次に任意の中間層ユニットを取り除いて成績を観察した。中間層を1,2,3 個取り除いても，学習能力に優れるこのネットワークは，すぐに再学習してしまう。脳損傷とは，ニューロンが損傷を受けて動作しなくなるだけではなく，再学習不能な事態に陥ってしまったことを指すのであろう。このことを表現するために，中間層のユニットを取り除くだけでなく，中間層から出力層への結合強度を固定して，再学習を行わせた。その結果，どの条件においても，損傷によって，一旦低下した成績は回復することなく学習回数の上限値に達する場合が多かかった。このときの中間層とクリーンナップ層のユニット間での項目間の相関係数を図6に示す。図4と比較すると，図から明らかなとおり，項目間の相関係数が高くなっている。このことは，混同が起きやすくなっていると解釈でき，それ故，損傷を受けたネットワークは各項目を混同しやすくなっているということができよう。

次に，クリーンナップ層のユニット数を2にして，学習させた後，1つの除去した場合のパフォーマンスを示す。それぞれ初期値を変えて実行した結果を示す。

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 1 2 3 4 5 6 7 (5) 9 (5) 11 (5) (5) (5) (5)
0 1 2 3 4 5 6 7 (7) (7) (7) (7) 12 (7) (7) 15
0 1 2 3 4 5 6 7 (6) (6) (6) (6) (6) (6) (6) (6)
0 1 2 3 4 5 6 7 (1) (1) (1) (1) (1) (1) 14 (1)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 1 2 3 4 5 6 7 (1) (1) 10 (1) (1) 13 (1) (1)
0 1 2 3 4 5 6 7 8 (1) 10 11 (1) (1) (1) (1)
0 1 2 3 4 5 6 7 (4) (4) (4) 11 12 (4) 14 15
0 1 2 3 4 5 6 7 (7) 9 10 11 (7) (7) (7) (7)

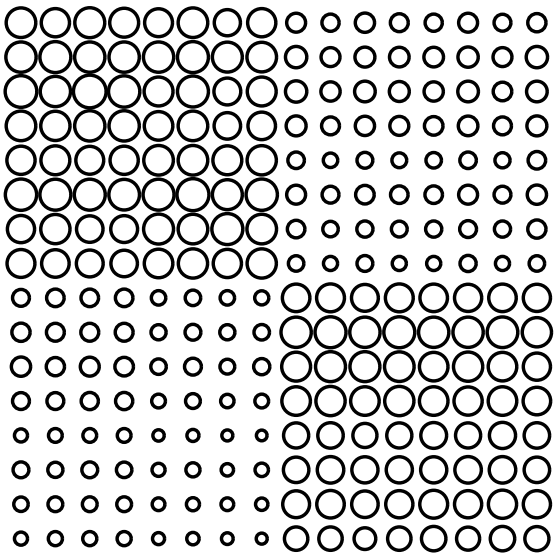

図 6 損傷を受けたネットワークの再学習後の中間層とクリーンナップ層のユニットによる学習項目間の相関係数行列を視覚化したもの。カテゴリー条件で，10個の中間層ユニット，1個のクリーンナップ層ユニットで学習させた後，中間層のユニットを3つ除去して再学習させた後の各項目間の相関行列。図4との比較されたい

0 から 7 までが非動物，8 から 15 までが動物概念である。カッコつきの数字は出力層とクリーンナップ層との繰り返しの上限 20 回でも，収束基準である自乗誤差 0.05 に達しなかったことを示している。収束基準には達してないものの，もっとも当てはまる項目を表示してある。これを見ると損傷によりアトラクタの流域が変化し，それによって他の項目と混同が起こっているのではないかという可能性が指摘できる。動物概念に比べて，個々の相関係数が小さい非動物概念は，全く誤らなかった。

このときの出力層とクリーンナップ層との間の繰り返し数を見てみると，

0 0 0 0 0 0 0 0 2 2 2 2 1 2 2 2
2 0 0 0 0 0 0 0 (20) 2 (20) 2 (20) (20) (20) (20)
0 0 0 0 0 0 0 0 (20) (20) (20) (20) 2 (20) (20) 2
0 0 0 0 0 0 0 0 (20) (20) (20) (20) (20) (20) (20) (20)
0 0 0 0 0 0 0 0 (20) (20) (20) (20) (20) (20) 3 (20)
0 0 0 0 0 0 0 0 2 2 2 2 3 2 2 2
0 0 0 0 0 0 0 0 (20) (20) 2 (20) (20) 2 (20) (20)
0 0 0 0 0 0 0 0 2 (20) 2 3 (20) (20) (20) (20)
0 0 0 0 0 0 0 0 (20) (20) (20) 2 2 (20) 2 3
0 0 0 0 0 0 0 0 (20) 2 2 2 (20) (20) (20) (20)

となって，やはり動物概念の方が繰り返し数が

増えていることがわかる。繰り返し数と反応時間とを同一視できると考えるのであれば，アトラクタネットワークはカテゴリー特異性を説明するモデルと考えることができると言えよう。

誤り方の性質の分析としては，Tyler のデータが各項目間であまりにも近すぎ，一箇所を破壊されてしまうと，まったく同じ項目に誤るようになると考えられる。実際，Tyler のデータを多次元尺度構成法によって2次元までの解を求めると，以下のようになった。

-0.000000 -0.968246
-0.000000 -0.968246
-0.000000 -0.968246
-0.000000 -0.968246
-0.000000 -0.968246
-0.000000 -0.968246
-0.000000 -0.968246
-0.000000 -0.968246
-1.414214 0.968246
-1.414214 0.968246
-1.414214 0.968246
-1.414214 0.968246
1.414214 0.968246
1.414214 0.968246
1.414214 0.968246
1.414214 0.968246

上 8 行が非動物概念の座標値，下 8 行が動物概念の座標値である。すなわち，非動物に関しては 2 次元までの解では判別不能なのである。従って，ある項目に近い項目があった場合，その項目は，他の多数の項目の最近解である可能性が高い。このように考えれば，カテゴリー内エラーばかりが起こる訳ではなく，カテゴリーを飛び越えてしまう誤りを説明できるように思われる。これは，データの与え方をより現実に近いように変更すれば，実現できるかもしれない。

4. 考察

モデルの単純さにもかかわらず，アトラクタニューラルネットワークモデルは，神経心理学におけるカテゴリー特異性の症状を記述できる。このことは，このモデルの大きな特徴であろう。動物–非動物概念の二重乖離をより詳細に記述できる可能性さえ残されていると考える。しかしながら，この二重乖離を説明するためには，さらなる検討が必要である。ここでは，記憶表象の二分法的な分類を採用せず，各項目を記述する特徴間の相関行列を用いて説明を試みた。すなわち，このように記述することで，予め生物–非生物の概念を分類した脳内器官を持って生まれてくるというような前提を設けるようなことをせずに済む。むしろ，経験により動物概念と非動物概念とが分離されると考えるべき根拠となっていると考えられる。そして，このようにして創発した意味記憶が損傷を受けるとカテゴリー特異的な障害を生じると考えることができるように思われる。

参考文献

[1] A. Caramazza and J.R. Shelton. Domain specific knowledge system in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1):1–34, 1998.

[2] J.T. Devlin, L.M. Gonnerman, E.S. Andersen, and M.S. Seidenberg. Category-specific semantic deficits in focal and widespred brain damage: A computational account. *journal of cognitive neuroscience*, 10(1):77–94, 1998.

[3] Martha J. Farah and James L. McClelland. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4):339–357, 1991.

[4] Gregory E. Hinton and Tim Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

[5] L.K. Tyler, H.E. Moss, M.R. Durrant-Peatfield, and J.P. Levy. Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, 75:195–231, 2000.

[6] E.K. Warrington and R. McCarthy. Category specific access dysphasia. *Brain*, 106:859–878, 1983.

[7] E.K. Warrington and R. McCarthy. Multiple meaning systems in the brain: A case for visual semantics. *Neuropsychologica*, 32:1465–1473, 1994.

[8] E.K. Warrington and T. Shallice. Category-specific semantic impairment. *Brain*, 107:829–854, 1984.

[9] Elizabeth K. Warrington. Neuropsychological studies of verbal semantic systems. *Phil. Trans. R. Soc. Lond. B*, 295:411–423, 1981.

[10] Elizabeth K. Warrington and R.A. McCarthy. Categories of knowledge further fracitonations and an attempted integration. *Brain*, 110:1273–1296, 1987.

[11] 浅川伸一. 脳損傷とニューラルネットワークモデル — 神経心理学への適用例 —. In 守一雄, 都築誉史, and 楠見孝, editors, **コネクショニストモデルと心理学**, chapter 5, pages 51–66. 北大路書房, 2001.