Generative Network Explains Category Formation in Alzheimer Patients

Péter Aszalós[1], Szabolcs Kéri[2], Gyula Kovács[2], György Benedek[2], Zoltán Janka[2], and András Lörincz[1]
[1]Eötvös Loránd University, Pázmány Péter sétány 1/D
Budapest, Hungary H-1117, emails: aszalos@alpha0.iki.kfki.hu, lorincz@valerie.inf.elte.hu
[2]Albert Szent-Györgyi Medical School, Dóm tér 10,
Szeged, Hungary H-6720, emails: {szkeri,gkovacs,benedek,janka}@phys.szote.u-szeged.hu

Abstract

This paper presents a generative data reconstruction neural network model equipped with plastic lateral connections. The model is capable of capturing basic phenomena related to category formation. It explains category formation as an effect of cumulative memory traces at the level of lateral connectivity. The formed memory traces change network activity that is the basis of categorization according to the model. This change however depends on the structure of the lateral connectivity and on the stimuli used in demonstrations. We argue that the model resolves the seemingly contradictory demonstrational results carried out with Alzheimer disease (AD) patients on category formation. We consider different stimulus sets and degraded lateral connectivity and show that the categorization probability can change from monotone to non-monotone functions depending on the sets.

I. Introduction

Implicit learning is reflected when subjects exhibit knowledge of stimuli that they fail to recollect consciously. A typical example of implicit learning is implicit category formation. Implicit category formation occurs when subjects encountering a series of stimuli learn about what all the stimuli have in common, with the result that information is acquired about the category defined by the objects [1]. In a typical paradigm, subjects are presented with a series of stimuli, together with a task that does not involve conscious categorization of the stimuli (e.g. to point to the right upper corner of the picture) [2]. The stimuli are different exemplars of the same category, e.g. distorted versions of a predefined stimulus, a dot pattern, called the prototype. The measure of distortion is the distance of the dots of the stimulus from the dots of the prototype. After this training procedure, the subjects are informed that they have seen members of the same category. Next, a test procedure is performed, where the subjects are presented with different exemplars and random patterns. The subjects have to decide whether a given stimulus belongs to the category or not. Category endorsement is measured for the previously seen or unseen exemplars of the category and for random patterns. Interestingly, the prototype is judged with the highest rate to be a member of the category, (this is called the prototype effect), even though it was not presented during training. Category membership of a stimulus (the rate with which subjects judge it to be a member of the category) depends statistically on the 'distance' (e.g. the measure of distortion) between the stimulus and the prototype. This process is called prototype-based (implicit) category formation.

Sinha et al. [3] found similar categorization and prototype formation phenomena in AD patients as in normal subjects. Kéri et al. [4] reported distorted prototype effect in AD patients, while categorization performance seemed to be intact. Both groups followed the classical paradigm introduced by Posner and Keele [2] and a contradiction seemed to evolve. The only difference in demonstrational procedure was that the first group used stimuli with smaller distortions than the second group.

This article presents an artificial neural network model which is capable of capturing basic phenomena related to category formation. The model network is a member of a generative or data reconstruction network paradigm [5-7]. The paradigm is oriented to sensory data abstraction and involves the optimization of information transfer.

Information optimization principles have been introduced to form long-term memories from external data in a self-organizing manner [8-14]. Information optimization leads to non-orthonormal memories and requires development of the direct computation of the pseudoinverse matrix of the long-term memories, or a dynamic reconstruction architecture that computes the pseudoinverse indirectly

[15-17]. It has been shown that the latter is more robust against tuning imprecisions [18]. Our model is a minimal extension of the basic generative Data Compression and Reconstruction (DCR) network [15].

II. Description of the model

A novel generative neural network equipped with lateral connections was utilized shown by the following figure:

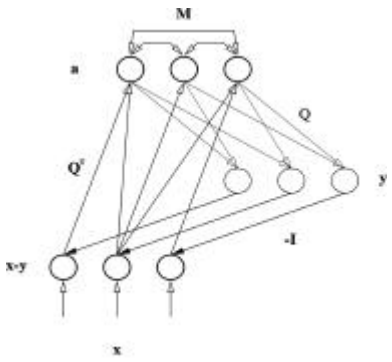

Figure 1. Associative Compression and Reconstrauction (ACR) network. x: input, y: reconstructed input, a: internal representation, Q: memory matrix, I: identity matrix, M: associative matrix at the level of the internal representation.

The architecture can be realized by three layers of units. Circles represent units, layer of circles represent layer of units. Lines and arrows represent connections and connection directions. Lines with arrows on both ends represent symmetrical connections between units. White arrows represent excitation, while black ones represent inhibition. For the sake of simplicity, not all connections are shown, although layers are fully connected. Input (x) excites the network and, in turn, drives the internal representation layer (a) through memory matrix Q^T. From the current internal representation, the input is reconstructed by computing y(=Qa). The input and the reconstructed input are subtracted (x-*y*), and this difference corrects the internal representation via memory matrix Q. This closed loop (the Data Construction and Reconstruction [DCR] network) minimizes the energy:

$$J(a) = \tfrac{1}{2} (x-Qa)^2 \qquad (1)$$

and solves the overdetermined set of equations Qa=x for a, which value appears after relaxation. We note that the input-to-internal representation matrix and the internal representation-to-reconstruction vector matrix can be different. Identical matrices were kept for the sake of notational simplicity.

We have extended this scheme by a full lateral connectivity matrix (M) at the level of the internal representation. The resulting network is called Associative Compression and Reconstruction (ACR) network. Network dynamics is described by the following equations:

$$r = Q^T(x-Qa) \qquad (2)$$

$$\partial_t a = r + \alpha r \bullet Ma \qquad (3)$$

$$\partial_t M = \eta(-M+aa^T) \qquad (4)$$

Here ∂_t denotes temporal derivation, Q is the long-term memory matrix, α is a parameter, η is the learning rate of M, lateral connection matrix, and $\bullet$ denotes component-wise multiplication. M obeys a slow learning rule and builds up slowly decaying memory traces for every input (x) at the level of the internal representation (a). The equations describe the dynamics of the network and the learning of the lateral connectivity matrix, respectively. Equation for a stops if r=0, which guarantees, that the internal representations formed are unique and do not depend on M. The componentwise multiplication of r and Ma expresses that M transfers activities to units which have not achieved their unique internal representation value. Parameter α controls the influence of the lateral connections on the formation of the internal representation, while η is the learning rate. The equation for M expresses that units influencing the dynamics of each other are statistically dependent. In other words, we assume that, even if memory vectors are trained to minimize mutual information (i.e. to form independent components), higher order correlation between them may still exist and this correlation is represented by the associative structure. We intend to group together those units which (at a lower level of analysis) display statistically dependent behavior. We assume that the learning of memory matrix Q based on the statistical properties of the inputs (i.e. the minimization of mutual information) is even slower and can be neglected on the time scale considered here.

III. Method

In the test phase subjects have to categorize random and different distorted stimuli. Class membership for a given stimulus is measured as the rate of correct endorsement of stimuli. Normal subjects demonstrate the classical psychophysical class membership function that decreases with the distortion.

To test the modeling capacities of the network, we carried out computer demonstrations. In the demonstrations, matrix Q was held constant with overlapping memory vectors. Memory vectors represent local filters to a two-dimensional 10 by 10 grid world. Local filters were sensitive within a 3 by 3 area; units of the network had nonzero weights within different 3 by 3 areas. 100 local filters were utilized and the local filters were also

positioned on a 10 by 10 grid, i.e. the internal representation had 100 elements (N=n=100, x,a $\in R^{100}$, Q,M$\in R^{100x100}$ and the filters of the elements were distributed evenly. Parameters α and η were set to 0.1 and 15.0, respectively. Matrix M was initialized to zero and self-excitation was not allowed, i.e. lateral connections that connect a unit with itself were kept zero. The presentation time of the inputs was set to be constant, modeling the transient response of real neocortical neurons under sustained inputs, and was defined to allow approximate convergence in all cases. Learning was 'ON' in every instant.

Gaussian inputs were presented to the network. Figure 2 shows an input and the corresponding internal representation after relaxation (panel A and B, respectively). Note that with these memory vectors the appearance of the internal representation is somewhat narrower but it is very similar to the input.

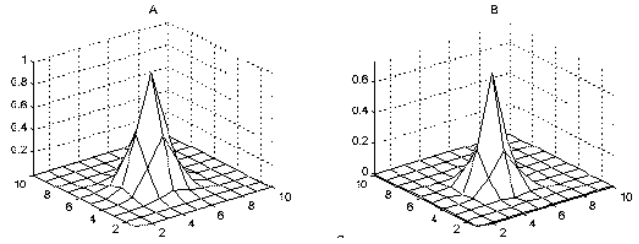

Figure 2. Input (panel A) and relaxed internal representation (panel B)

An activity measure, expressed by $F=\int|M(t)a(t)|dt$ was introduced to capture category formation phenomena. The value of *F* represents the integral of the activity flowing into the units of the internal representation via M. Gaussian input centered to the grid was chosen as the 'prototype input'. The network was trained with a sequence of 'distorted' (i.e., displaced) versions of the prototype. The measure of distortion, *d*, was held constant in every computer demonstration. Class membership was measured by increased lateral activity flow (the difference between *F* values for trained and untrained M matrices). Class membership function was computed for different *d* values.

IV. Results

Two sets of computer demonstrations were carried out to differentiate between the two cases corresponding to the two demonstrations with AD patients. The distortion measure, *d*, was set to 1.0 and to 2.0 in the two cases, respectively, modeling the different distortion values used in the two experiments.

The first case re-presented the classical psychophysical class membership function, i.e. the class membership was a monotonously decreasing function of the distance from the prototype. In the second case, however, the class membership function had a maximum at around 2.0 (the value of the distortion), and displayed a minimum at the prototype. This is due to the structure of the cumulative memory trace formed. The following figure shows the results:

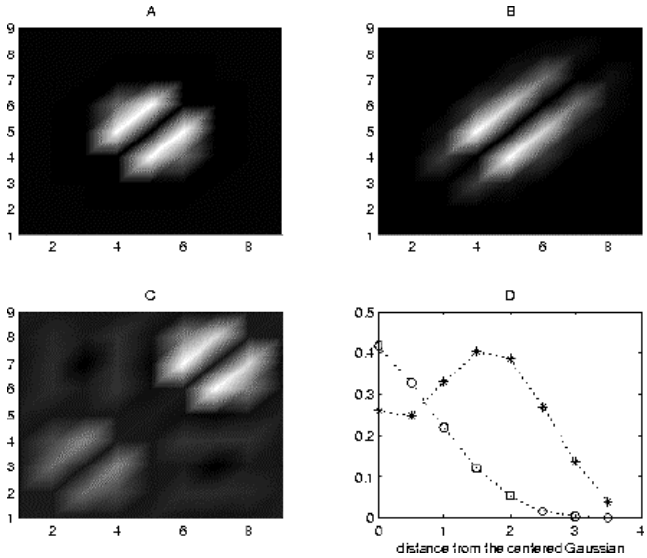

Figure 3. *Panel A:* the gray-scale interpolated plot of the relevant part (a 6 by 6 square of the 100 by 100 matrix) of the lateral memory matrix M showing a typical part of the memory trace left by the prototype input, a Gaussian, centered to the input grid. Connection strengths are represented in gray level and the discrete points are blurred for easier visualization. The larger the connection strength the lighter the gray level is. The particular shapes in the figure are emerging from the arrangements: elements of a 10 by 10 matrix are listed on both axis. *Panel B:* the relevant part of the memory trace left by a training sequence that contained inputs distorted around the prototype with the distortion *d*=1.0. Note the similarity to the memory trace shown in Panel A. *Panel C:* the relevant part of the memory trace left by the training sequence with a distortion *d*=2.0. Note that the features of the figure became longer and a 'gap' opened up in the middle of those. This 'gap' causes the deviation from the standard class membership function. *Panel D:* Class membership function as measured after training in the two cases described above.

The graphs show the activity measure. The first function ('o') corresponds to the classical class membership function, measured in psychophysical demonstrations, i.e. class membership depends monotonously on the distance from the prototype. The second one ('*'), however shows a minimum at the prototype. This is due to the fact that the training sequence contained only highly distorted stimuli.

For small *d* values the network produced the classical class membership function. For larger *d* values the class membership function had a maximum at around *d*, and

showed a minimum at the prototype. This is due to the structure of the formed cumulative memory trace.

V. Discussion

The main statements and findings relating to our work are as follows:

1) The ACR network is a model of category formation. The ACR network explains category formation as a consequence of the fast synaptic changes at the level of lateral connections. The cumulative memory traces left by previous stimuli form the category and categorization processes use the measure of activity flow at the lateral connectivity level.

2) The present model is a good candidate for explaining the findings found in AD patient's category formation capacities.

We consider this model to be adequate to characterize some important aspects of the activity of the cortical architecture and dynamics. The ACR network has a full lateral connectivity at the level of internal representation. It can be seen as a model of a small part of the neocortex. To model larger parts of the cortex one should restrict the length of the lateral connections. The network preserves topography, i.e. similar inputs elicit similar internal representations. This kind of mapping is known to be characteristic in several areas of the neocortex. The model is formulated in a general fashion that involves abstract inputs and abstract internal representations. It is not intended to capture the details characteristic to certain visual and/or cortical areas.

The dynamic elements of the model (input and internal representation units) are intended to model neurons and capture the dynamics of real neurons at the level of spike frequency rate, i.e. no spiking activity is modeled here. The learning rule of the associative matrix in the ACR network is Hebbian, considered to be close to reality. Learning of matrix M is fast as compared to long-term memory changes, and assumes that fast synaptic changes are responsible for implicit category formation. Such fast synaptic changes exist in the neocortex [19], but their relation to category formation is unknown. There is some evidence that neurons along the visual pathway accomplish information maximization [20]. In contrast, there is almost no direct evidence concerning the dynamics of the reconstruction network or the learning mechanism of the lateral connections. However, we subscribe to the view that vision (and possibly neocortical processing in general) has a reconstructive (generative) component. (see, e.g. [5,21,22]). A similarly general and well-known idea is that the grouping of information at several levels of representation may play a fundamental role in vision [23]. There are suggestions in the literature that lateral connections subserve the grouping of information (see, e.g., [24] and references therein).

The model network has the same input and output dimension, i.e. neither compression nor expansion (development of a sparse representation [25]) is modeled. This feature does not restrict the dynamics, but the representation capabilities. The model neglects several important characteristics of the neocortex, e.g. whether firing is synchronous or not, the interlaced diverse networks of excitatory and inhibitory cells, the connection structure between different areas, etc. We consider the model as a minimal model that can capture complex dynamic properties of the full architecture.

The model has important psychophysical connections. The network exhibits the most important phenomena related to implicit category formation. This is achieved in a self-organizing manner, i.e. none of the processes need external teacher or reinforcement, which is basic for such implicit processes. We measured categorization ‘strength’ as the difference between local integrals of activity flows along the lateral connections. Our measure is defined locally in the model neural tissue, in the region, where category formation has happened. In other words, the model suggests that local neural activity may provide sufficient information concerning class membership strength.

The main suggestion of the model is that the relative strength of the prototype effect may decrease if the distortion of the stimuli in the training sequence is raised. This may explain the different results of experiments carried out in AD patients because the two experiments used different distortion levels. The difference between the categorization capacity around the prototype of AD patients and normal subjects in [4] may be explained as a result of the degraded lateral connectivity of the cortex in AD patients.

VI. Conclusions

The model is a candidate for modeling implicit category formation. The model suggests that the shape of the class membership function changes with the distortion of the exemplars. It further suggests that implicit categorization is a local process in the neocortex and is a result of fast synaptic changes of lateral connections. Numerical studies show that seemingly contradictory results of experiments carried out with AD patients may be given an unified view.

We have presented a neural network model which is capable of reproducing the basic aspects of prototype-based category formation. The model network is a member of a novel class of artificial neural networks featuring maximization of information transfer achieved by dynamic

data reconstruction. The model is a minimal extension of the DCR network; here we have empowered the DCR network with lateral connections. The lateral connectivity exhibits fast learning, which leads to memory trace formation at the level of the internal representation. The model suggests that lateral connections of the local circuitry of the neocortex may play an important role in implicit processes such as category formation. It also suggests that the classical psychophysical class membership function changes systematically on variation of the training stimuli. The novel neural network model can explains implicit category formation for AD patients.

Acknowledgments. This work was partially supported by OTKA (T14566, T17100), the US-Hungarian Joint Fund (No. 519) and the Flemish-Hungarian Bilateral Fund (BIL 97-31).

References

1. Squire LR, Knowlton BJ (1995) Learning about categories in the absence of memory, Proc Natl Acad Sci, USA, 92: 12470-12474
2. Posner MI, Keele SW (1970) Retention of abstract ideas, Journal of Experimental. Psychology: Learning, Memory and Cognition, 15:282-304.
3. Sinha R, Heindel W, Ott B (1997) Classification learning and recognition memory for prototype dot patterns in Alzheimer's disease Society of Neuroscience, Abstract No. 32, p. 1580.
4. Kéri Sz, Kálmán J, Antal A, Rapcsák I, Benedek Gy, Janka Z (1999) Classification learning in Alzheimer's disease. Brain (in press).
5. Hinton GE, Ghahramani Z (1997) Generative models for discovering sparse distributed representations. Proc Trans of the Royal Society (London) B 352: 1177-1190.
6. Rao RPN, Ballard DH (1997) Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex. Neural Computation 9:721-763.
7. Roweis A, Ghahramani Z (1999) A unifying review of linear Gaussian models. Neural Computation (In press).
8. Jutten C, Herault J (1991) Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24:1-10.
9. Comon P (1994) Independent component analysis, a new concept? Signal Processing 36:287-314.
10. Bell AJ, Sejnowski TJ (1995) An information maximization approach to blind separation and blind deconvolution. Neural Computation 7:1129-1159.
11. Wang L, Karhunen J, Oja E (1995) A bigradient optimization approach for robust PCA, MCA, and source separation. Proceedings of the IEEE ICNN, Perth, Australia, USA: IEEE Publishing, pp. 1684-1689.
12. Amari SL, Cichocki A, Yang HH (1996) A new learning algorithm for blind signal separation. In: Advances. In: Neural information processing systems 8, (Touretzky D, Mozer M, and Hasselmo M, eds.) pp. 757-763. Cambridge MA: MIT Press.
13. Cardoso JF, Laheld B (1996) Equivalent adaptive source separation. IEEE Trans on Signal Processing 44:3017-3030.
14. Karhunen J, Oja E, Wang L, Vigario R, Joutsensalo J (1997) A class of neural networks for independent component analysis. IEEE Trans on Neural Networks 8:486-504.
15. Fomin T, Körmendy-Rácz J, Lörincz A (1997) Towards a unified model of cortical computation I: Data compression and data reconstruction architecture using dynamic feedback, Neural Network World 7:121-136.
16. Lörincz A (1998) Forming independent components via temporal locking of reconstruction architectures: a functional model of the hippocampus. Biological Cybernetics 79: 263-275.
17. Lörincz A, Buzsáki Gy (1999) Computational model of the entorhinal-hippocampal region derived from a single principle, IJCNN'99 (in this volume).
18. Körmendy-Rácz J, Szabó Sz, Lörincz J, Antal Gy, Kovács Gy, Lörincz A (1999) Winner-take-all network utilizing pseudoinverse computing subnets demonstrates robustness on the hand written charater recognition problem. Neural Computation and Applications (in press).
19. Varela JA, Sen K, Gibson J, Fost J, Abbott LF, Nelson SB (1997) A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. Journal of Neuroscience 17:7926-7940.
20. Baddeley R, Abbott LF, Booth MCA, Sengpiel F, Freeman T, wakeman EA, Rolls ET (1997) Responses of neurons in primary and inferior temporal cortices to natural scenes. Proc of the Royal Society (London) B 264:1775-1783.
21. Horn BKP (1977) Understanding image intensities. Artificial Intelligence 8:201-231.
22. Kosslyn SM (1994) Image and brain, MIT Press, Cambridge.
23. Marr D (1982) Vision, W.H. Freeman and Company, San Francisco.
24. Kovács I (1996) Gestalten of today: Early processing of visual contours and surfaces. Behavioral Brain Research 82:1-11.
25. Olshausen BA, Field DJ (1996) Emergence of Simple-Cell Receptive field properties by learning a sparse code for natural images. Nature 381: 607-609.