A Modular Attractor Model of Semantic Access

William Power[1], Ray Frank[1], John Done[2], Neil Davey[1]

[1]Department of Computer Science, [2]Department of Psychology,
University of Hertfordshire, College Lane, Hatfield, Hertfordshire, UK, AL10 9AB.
{w.power, r.j.frank, d.j.done, n.davey} @herts.ac.uk

Abstract. This paper presents results from lesion experiments on a modular attractor neural network model of semantic access. Real picture data forms the basis of perceptual input to the model. An ultrametric attractor space is used to represent semantic memory and is implemented using a biologically plausible variant of the Hopfield model. Lesioned performance is observed to be in agreement with neuropsychological data. A local field analysis of the attractor states of semantic space forms a basis for interpreting these results.

1 Introduction

Connectionist models of neuropsychological phenomena can provide potentially useful insights into the nature of cognitive processes. In this paper we present the results of lesion experiments on a modular neural network model of semantic access. Our approach differs from previous work on three counts: the use of real pictorial data; an ultrametric representation of semantic memory that allows for an emergent prototype effect; a biologically inspired, iterative, associative memory, storage prescription. This paper reports the results from three lesion experiments, that demonstrate the model's ability to accommodate neuropsychological data. These results are subsequently interpreted in terms of distributions of local fields in the Hopfield semantic network.

2 Background

The structure and organisation of categories play a vital role in efficient access to semantic memory. Once we know that animals breathe we can infer that all exemplars of that category breathe without having to code for this feature directly in the representation of each exemplar. Early models of semantic memory explicitly distinguished between

category information and exemplar information. Rosch's work on the "basic object level" marked basic-level category knowledge as special, and distinct, from exemplar level knowledge [1]. The work of Damasio further suggests that this distinction is physical as well as functional [2]. Furthermore, recent theorising on the nature of categories has highlighted the need for dynamic forms of representation to capture the flexibility inherent in human categorisation [3]. Our model uses a integrated approach, in which category knowledge is an emergent property of the representation and network dynamics.

Neuropsychological studies of patients with Alzheimer's Disease (AD) provide striking dissociation between knowledge of categories and exemplars. These results provide converging lines of evidence that span multiple measures of semantic knowledge: attribute listing; confrontation naming and picture-name matching. The most salient result for our purpose here, is that semantic knowledge of base level categories appears disproportionately impaired relative to superordinate knowledge of categories. Particularly striking are the types of naming errors made by AD patients within the confrontation naming paradigm. In many cases an object is called by the name of its category (e.g. "animal" instead of "cow"); in others, termed semantic errors, the name of a different object from the same category is given (e.g. "horse" instead of "cow"). Detailed analysis of the patterns of errors made by AD patients across this and other methodologies suggest that semantic category knowledge is available but that specific exemplar knowledge is impaired.

Computational modelling is increasingly being used to refine neuropsychological theories [4,5]. These models have proven successful in increasing understanding of the nature of disordered brain processes. For all their success, these models can be criticised on two fronts. Firstly, the form of representation used, where instances to be categorised are generally represented with hand crafted sets of features. In so doing, a fundamental problem is side-stepped: how to determine these features. The second area of concern is the architecture of the model. Static approaches to modelling semantic memory using back-propagation in multilayered perceptrons fail to address the intrinsic dynamic nature of semantic memory. The recurrent network of Plaut & Shallice [4] and the constraint satisfaction networks of Tippet, McAuliffe & Farah [5], although addressing the dynamics of semantic memory, are hindered by a tight coupling between perceptual and semantic memory. This coupling hinders detailed analysis of perceptual and semantic processes. Our modular approach allows independent analysis of the mapping from perceptual to semantic memory, and of the structure of semantic space, in both pre- and post-morbid condition.

3 The Model

The architecture of the full model is shown in Figure 1, and builds on work by Gale et al[6, 7].

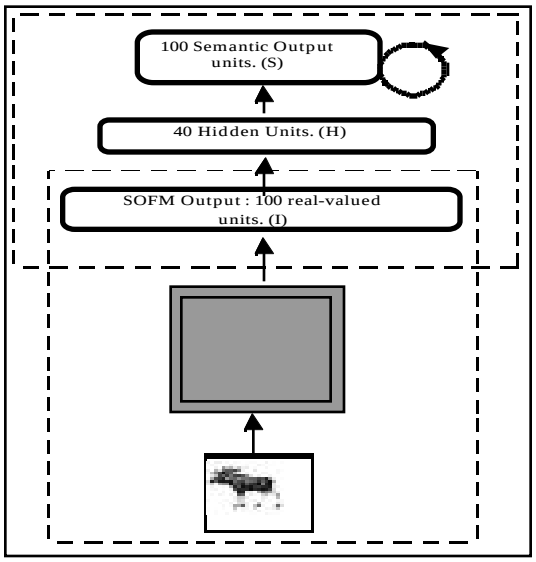

Fig. 1. Architecture of the full model showing interaction of the perceptual and semantic modules. The I-H lesions refer to modifications of the connections between the SOFM and the hidden units.

3.1 Architecture

The model consists of two modular components: an unsupervised perceptual processing module and a supervised semantic memory module. Our model uses real pictorial data. Perceptual input to the model is in the form of features extracted from pictorial data by a 100 unit unsupervised self organising feature map (SOFM). By using these

autonomously extracted features the 'homunculus problem' whereby modellers provide suitable neural inputs [8], can be avoided.

The semantic module consists of a 100 unit Hopfield associative memory trained using the iterative approach of Diederich and Opper [9]. The use of the Hopfield architecture ensures that the semantic memory module has a well defined attractor nature. A review of the biological plausibility of such an approach can be found in [10].

The perceptual module is connected to the semantic module through an intermediate layer of 40 hidden units. Traditional back-propagation is used to map SOFM output to its appropriate semantic attractor in the Hopfield network. The perceptual and semantic modules are de-coupled. Perceptual input is mapped to an initial state of the Hopfield network, which then relaxes to an attractor state under its own dynamics.

3.2 Representation

3.2.1 Perceptual representation

The total perceptual data set consists of 560 (8 bit) greyscale images derived from four superordinate categories: Animals, Musical Instruments, Clothing and Furniture. Each superordinate category comprises seven basic-level categories. Animal is subdivided into: bird, snake, spider, fish, deer, mouse and frog. Each basic-level category is further divided into five subordinate categories, for example Fish: pike, carp, salmon, herring and bass. These categorical levels reflect the tripartite hierarchy proposed by Rosch et al [1]. Each subordinate is represented by 4 versions of the same exemplar, which vary on dimensions of contrast and left-right inversion [6].

Each image is processed by a self-organising feature map (SOFM) similar to that used by Schynns [11]. The output of the SOFM forms the basis of the input to the perceptual module of the model.

Fig. 2. Examples of greyscale images that comprise the perceptual data set. Each image fits within a 50 by 50 pixel grid such that the principal dimension comes within one pixel of the grid border. These examples depict the 5 subordinate images of the basic level category 'clock' which, in turn, is one of the 7 basic level categories representing furniture.

3.2.2 Semantic representation

The semantic representation associated with each perceptual input pattern consists of a 100 unit binary (-1,1) vector. Categorical information is implicitly encoded using the ultrametric approach of Virasoro [12], where semantic vectors are arranged in a hierarchical structure. The procedure is as follows: *C* class vectors are generated by independent random choice of components. That is, each bit in the vector is chosen (independently) to be 1 or -1 with probability 0.5. For each of these class vectors, *E* exemplar vectors are generated, by correlated choice of components:

$$\begin{aligned} S_i^{\alpha\beta} &= S_i^{\alpha} \text{ with probability} (1- m) \\ S_i^{\alpha\beta} &= -S_i^{\alpha} \text{ with probability} m \end{aligned} \qquad \alpha = 1 \quad C; \quad \beta = 1 \quad E; \quad i = 1 \quad 100 \qquad (1)$$

$S^{\alpha\beta}$ specifies the exemplar vector β, within the class α. Where, α enumerates the classes and β the exemplar vectors.

For simplicity in collating results, our implementation uses exemplar patterns derived from class vectors by flipping a *fixed* number of randomly chosen bits. The number of such flips, (m × 100), determines the cluster density of the category: smaller values of m lead to more compact clusters.

These exemplar vectors, (but *not* the class vectors), are embedded in the Hopfield semantic memory. This approach allows for a two level semantic hierarchy to be implemented in the model, with category membership based on similarity among exemplar vectors, a prototype approach.

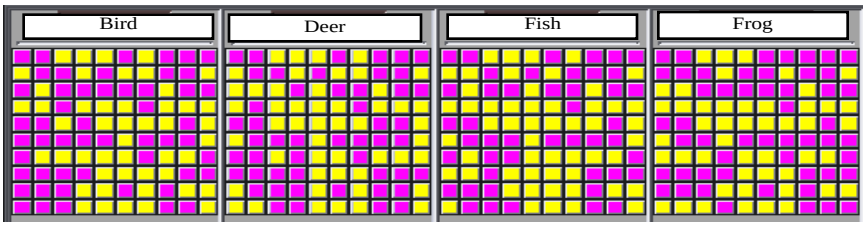

Fig. 3. Four derived exemplar semantic vectors. The cluster density m = 0.1, so that each exemplar is derived from the class vector by independent random choice of 10 bits to flip.

3.3 Training

There are three stages in training. Firstly, the SOFM is trained on the real picture data. The features extracted form the input to the perceptual module.
Secondly, the *exemplar* semantic vectors are embedded in the Hopfield network using the iterative approach of Diederich and Opper [9]. The procedure is as follows:

1. Begin with a zero weight matrix.
2. Repeat Until all training patterns are stable
   2.1 Set the state of the network to one of the $S^{\alpha\beta}$
   2.2 If a unit i changes state, update its incoming weights according to

$$\Delta w_{ij} = \frac{1}{N-1} S_i^{\alpha\beta} S_j^{\alpha\beta} \qquad (2)$$

This approach overcomes the poor storage performance associated with standard 'one-shot' training. A maximum useful capacity of N - 1 linearly independent patterns can be stored in this way. Furthermore, this approach uses only locally available information, thereby increasing the biological plausibility of the model.

Finally, the network is trained to associate perceptual input with semantic output. Two methods of training were investigated: using back-propagation to train a perceptual input to fall just within the boundary of its associated semantic attractor; training perceptual input directly to the associated semantic attractors. Results reported here are based on the latter.

4 Experiments

4.1 Experiment 1

This experiment investigates the effect of lesions on the perceptual to semantic, (direct) route. Training sets of size 28 (4 categories of 7 exemplars) are generated as mentioned above and the model trained to 100% performance. Independent lesions of varying severity, (5%-70%) are then performed on the I-H and H-S connections, see Figure 1. Results are summarised in tables 1 and 2.

4.1.1 Results and Discussion

In order to measure the success of the network in categorising the input, the overlap (normalised dot product) of the networks semantic response with the training set patterns is taken. The model is considered to give a correct exemplar response, if its output is closer to the desired exemplar than to any other exemplar. Additionally, this overlap must be sufficiently close, that is, the overlap must be greater than the overlap with the class vector of the category. Furthermore, the gap to the next nearest exemplar must exceed 0.05, see [4] for motivation and justification. Correct category response, is achieved if the network's response is closer to the desired prototype than to any other prototype. To ensure that this response is sufficient a cut-off of 0.7 was enforced.

Damage to the direct pathway can be visualised as a change in the mapping from perceptual input to semantic output, or as perturbation in the initial point projected to in semantic space. The attractor nature of semantic memory will ensure that small perturbations will be corrected, provided they stay within the appropriate basin of attraction (see 10% H-S lesion in Table 1). But, as lesion severity increases the projection point falls outside its correct basin of attraction. Table 2 shows a typical response profile. At 30% lesion, 6 of the 28 patterns get mapped into an emergent category prototype attractor. This attractor is *not* directly stored, but gets embedded as a result of storing the correlated exemplar patterns, (a mixture state). Additionally, there are 4 semantic errors, where the input is mapped into the attractors of semantically related exemplar, ('Dog' to 'Cat'). This behaviour mirrors that observed in patients with semantic deficit [6].

In summary, lesions to the direct pathway can result in an overall loss of exemplar level accuracy over the category level. Closer inspection of the data reveals that this is the result of the model making emergent prototype and semantic error responses.

Lesion Severity		*10%*	*20%*	*30%*	*40%*	*60%*
Type I-H	Category	95%	84%	70%	79%	54%
	Exemplar	75%	70%	41%	50%	13%
Type H-S	Category	100%	100%	98%	97%	86%
	Exemplar	100%	84%	79%	66%	34%

Table 1. Percentage of correct category/exemplar responses for lesions at different sites on the direct pathway. Later stages of processing (H - S) are more resistant to damage. There is superior preservation of category information over exemplar information. All results are averaged over 10 training sets with each lesion performed 10 times.

Lesion severity	*20%*	*30%*	*50%*	*60%*
Correct exemplars	23	18	12	9
Semantic errors	0	4	4	6
Correct prototypes	5	6	9	11

Table 2. Typical response profile for H-S lesions. Based on 4 categories of 7 exemplars with cluster density m = 0.1. Correct here indicates going to an attractor state. For example, at 50% lesion severity: 12 patterns go to the appropriate attractor; 4 to a semantically related attractor; 9 to the emergent prototype attractor; 3 fail the response criteria.

4.2 Experiment 2

This experiment investigates the effect of diffuse damage to the semantic memory module.

4.2.1 Results and Discussion

Unlike damage to the direct pathway, damage to the Hopfield network results in change to the semantic attractor space: the locations and basins of attraction get modified. The performance of the network is summarised in table 3.

Lesion Severity	*10%*	*25%*	*30%*	*40%*	*50%*	*60%*
Category	100%	100%	93%	86%	64%	34%
Exemplar	100%	96%	75%	61%	11%	0%

Table 3. Percentage of correct category/exemplar responses after diffuse damage to the Hopfield semantic memory.

The network is initially resistant to small amounts of damage. The intrinsic error correcting abilities of the Hopfield network ensure 100% correct performance up to 20% lesion severity. As the lesion severity reaches 25% exemplar level loss begins. This is the result of 'drift' in the location of the attractors in semantic space. The direction of this 'drift' is predominately towards the category prototype vector; that is, loss of exemplar bits. This corresponds to the response profile observed in patients with AD, where there is predominant loss of exemplar level features over category level features. The next experiment investigates the computational reason for this.

4.3 EXPERIMENT 3

It is instructive to relate the effect of lesions to the semantic network's interconnections to a measure of noise in the system. The local field (net-input), h_i, to a particular semantic unit determines its activation state. This in turn, is determined by the weighted sum over the activations of the remaining semantic units. The effect of damage to network connectivity can be understood as the addition of a noise term, δ_i, to the 'pre-damage' local fields:

$$h_i = \sum_j w_{ij} S_j^{\alpha\beta} \qquad h_i + \delta_i$$

A bit flip will occur only if the absolute value of the local 'pre-damage' field is less than that of the noise term. From this, it follows that larger fields will lead to more robust bit patterns. The results of experiment 2 would indicate that the field strengths for category bits should on average be stronger than those of exemplar bits. Such a result would be important as it would make the results of experiment 2 independent of precise implementation of diffuse damage. See [12] for a detailed theoretical treatment.

In this experiment the distribution of local field strengths for exemplar and category bits are analysed. Exemplar bits, where $S_i^{\alpha\beta} = -S_i^{\alpha}$, are those that identify the exemplar within a category. Category bits, on the other hand, are those where $S_i^{\alpha\beta} = S_i^{\alpha}$, The distributions of fields for various network loadings, ρ, the number of stored patterns/ number of units and various cluster densities, m, are summarised in Table 4.

	Cluster density, m	*0.1*	*0.15*	*0.2*
Load density, ρ				
0.28	Exemplar field	1.41	1.44	1.48
	Category field	2.11	2.08	1.98
	Difference	0.7	0.68	0.5
0.56	Exemplar field	1.42	1.56	1.59
	Category field	2.11	2.08	2.01
	Difference	0.69	0.52	0.42

Table 4. Analysis of exemplar and category field strengths for various load and cluster densities. Category bit fields are substantially stronger and therefore more resistant to noise.

4.3.1 Results and Discussion

There is substantial difference between the strengths of category and exemplar fields. The cumulative distribution of field strengths is particularly informative. For example, with ρ = 0.56, m = 0.1, we found approximately 80% of category bits have field strengths above

the 1.5 level compared to 30% for exemplar bits; increased loss of exemplar information follows.

5 Conclusions

In this paper we report the results of lesion experiments on a modular connectionist model of semantic access. With appropriate use of representation and architecture we were able to overcome some of the problems associated with previous approaches. Furthermore, prototypes are an emergent property of the model. Finally, damage to the semantic memory was shown to preferentially effect exemplar level information; analysis traced this to asymmetry between exemplar and category field distributions.

References

1. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic Objects in Natural Categories. Cognitive Psychology. 8 (1976) 382-439
2. Damasio, A.R.: Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. Cognition. 33 (1989) 25-62
3. Thelen, E., Smith, L.B.: A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge Mass. (1994)
4. Plaut, D.C., Shallice, T.: Deep Dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology. 10 (1993) 377-500
5. Tippett, L.J., McAuliffe, S., Farah, M.J.: Preservation of categorical knowledge in Alzheimer's disease: A computational account. Memory. 3 (1995) 519-533
6. Gale, T.M.: Perceptual and semantic information in object in object recognition: A neuropsychological and connectionist study. Ph.D. Thesis, University of Hertfordshire.(1997).
7. Done, D.J., Gale, T.M.: Attribute verification in dementia of Alzheimer's Type: Evidence for the preservation of distributed concept knowledge. Cognitive Neuropsychology. 14 (1997)
8. Reeke, G.N., Sporns, O.: Behaviourally based modelling and computational approaches to neuroscience. Annual Review of Neuroscience. 16 (1993) 597-623
9. Diederich, S., Opper, M.: Learning of Correlated Patterns in Spin-Glass Networks by Local Learning Rules. Physics Review Letters. 58 (1987) 949-952
10. Amit, D.J.: The Hebbian paradigm reintegrated: local reverbations as internal representations. Behavioral and Brain Science. 18 (1995) 617-657
11. Schynns, P.G.: A modular neural network model of concept acquisition. Cognitive Science. 15 (1991) 461-508
12. Virasoro, M.A.: The Effect of Synapse Destruction on Categorization by Neural Networks. Europhysics Letters. 7 (4) (1989) 293-298