A Recurrent Connectionist Model of Semantic Dementia

Timothy T. Rogers†
Matthew A. Lambon Ralph*
Karalyn Patterson*
James L. McClelland†
John R. Hodges*
†Center for the Neural Basis of Cognition and
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

*Medical Research Council
Cognition and Brain Sciences Unit
Cambridge, UK

Poster presented at the 1999 meeting of the Cognitive Neuroscience Society

Introduction

Semantic dementia is a disorder in which patients exhibit a progressive deterioration of semantic knowledge, while other cognitive faculties remain remarkably spared. The deficits witnessed in semantic dementia are typically amodal, and are not category specific. Nevertheless, the breakdown of semantic memory in the disorder is structured.

Four aspects of structured breakdown

1. General knowledge is better preserved
2. Typical properties are over-extended to inappropriate but related objects
3. Performance is better with pictures than with words
4. Different patterns of errors are observed for living and nonliving things

These phenomena are robustly observed in a variety of neuropsychological tests of semantic memory. Such tests are consistent with the view that semantic memory is a system that maps between surface representations in different modalities. For example, in picture naming, the patient must use the semantic system to derive a name (phonology) from a representation of the pictured item (vision).

Table 1: Ten common tests of semantic memory. Most require the patient to map between surface modalities; for example, drawing to name can be understood as mapping between phonological input and visual output. Some tests of semantic memory, such as sorting and word-to-picture matching, require patients to compare the similarity of internal representations. The last column shows those tests for which we have data confirming the four aspects of structured breakdown listed above.

Task	Visual	Name	Descr	Rep	Data
Drawing	Out	In			Y
Delayed copy	In/out				Y
Picture naming	In	Out			Y
Word-pic matching	In	In		Out	Y
Naming to def.		Out	In		?
Attribute listing		In	Out		?
Word definition		In	Out		?
Picture definition	In		Out		?
Word sorting		In		Out	Y
Picture sorting	In			Out	Y

The model

In learning to map properties to objects, connectionist networks can form distributed representations of instances that reflect their semantic relatedness. How might such a network's performance deteriorate as these representations break down? We explored this question in a recurrent autoencoder that learned the mappings between names, visual images, and sets of propositions.

Feature-based visual and propositional representations were constructed to match the statistics from control drawings and attribute-listing tasks. In particular, in the model there were:

1. More sensory than functional propositions
2. More functional than encyclopaedic propositions
3. More shared features for living things
4. More distinctive features for artifacts
5. More features overall for living things
6. Many more shared features for visual representations of living things

Each name was represented by a local unit. In the model diagram, units shown in cross-hatching are features shared among most members of the same domain (e.g. "can move"). By "shared," we mean that most other instances from the same domain also have the features; however, the features are not perfectly consistent. There may be some animals who do not have a "shared" feature (e.g. an animal with no legs), and some artifacts that do (e.g. a table with "legs"). Units shaded with slanted lines are features shared among members of the same superordinate category (e.g. "has four legs"), while those with no markings are idiosynctratic features (e.g. "common in Pittsburgh"). For living relative to

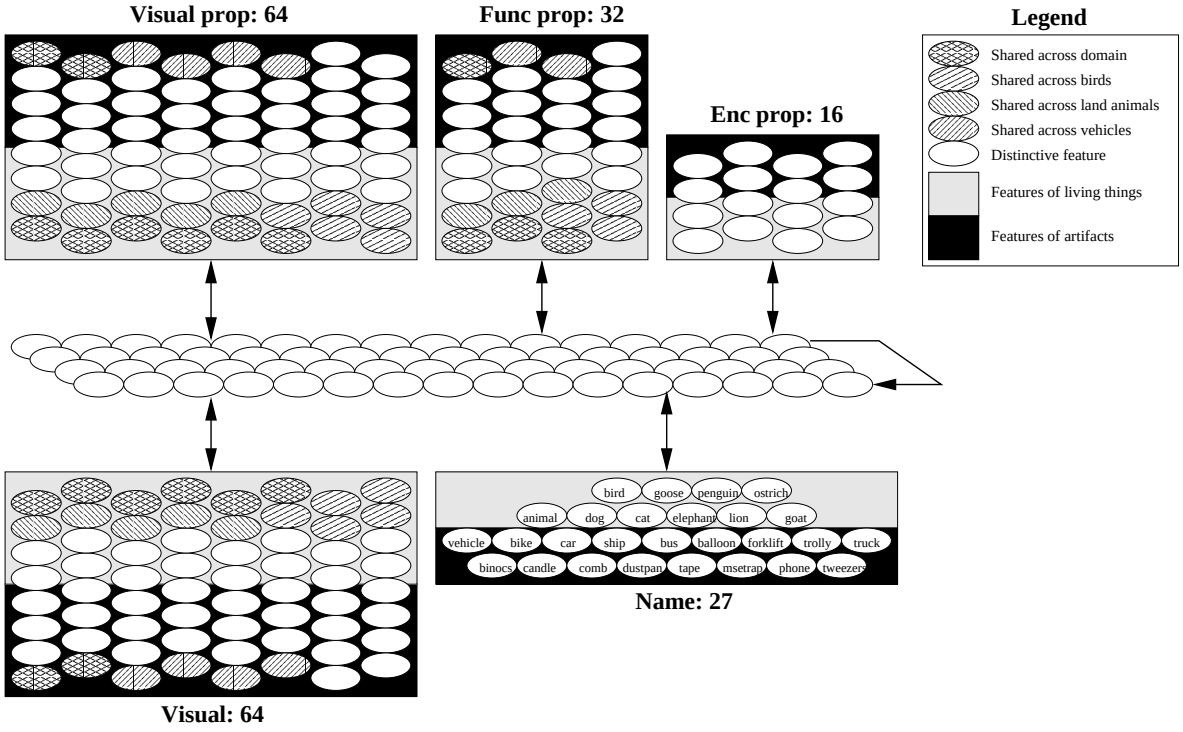

Figure 1. Model architecture showing the number of distinctive features, and the number of features shared across domain (living thing / artifact) and superordinate category. The top units are local representations of propositional features, such as "can fly" and "has a long neck." Units in the group labelled "Visual" represent high level visual features of the kind included in line drawings of objects. Each unit in the "Name" layer represents a particular name that can apply either to several unique objects (e.g. "Animal") or to a single object (e.g. "Penguin"). Note that all connections project to or from the hidden layer, which maps between names, visual representations, and descriptions of objects (in the form of propositions). The knowledge stored in these connections allows the network to make inferences when provided with partial input in any of the visible layers. Thus, these connections embody the model's "semantic" knowledge.

nonliving things, there were more features active overall, and more features shared across related instances.

Each visible layer (Proposition, Name, Visual) functioned as input on some occasions, and output on others. The network was trained to produce the correct pattern of activation across all visible layers when presented with input in one of them.

Learning to name

When presented with visual or propositional input, the network was trained to produce a single name for each instance, as shown below.

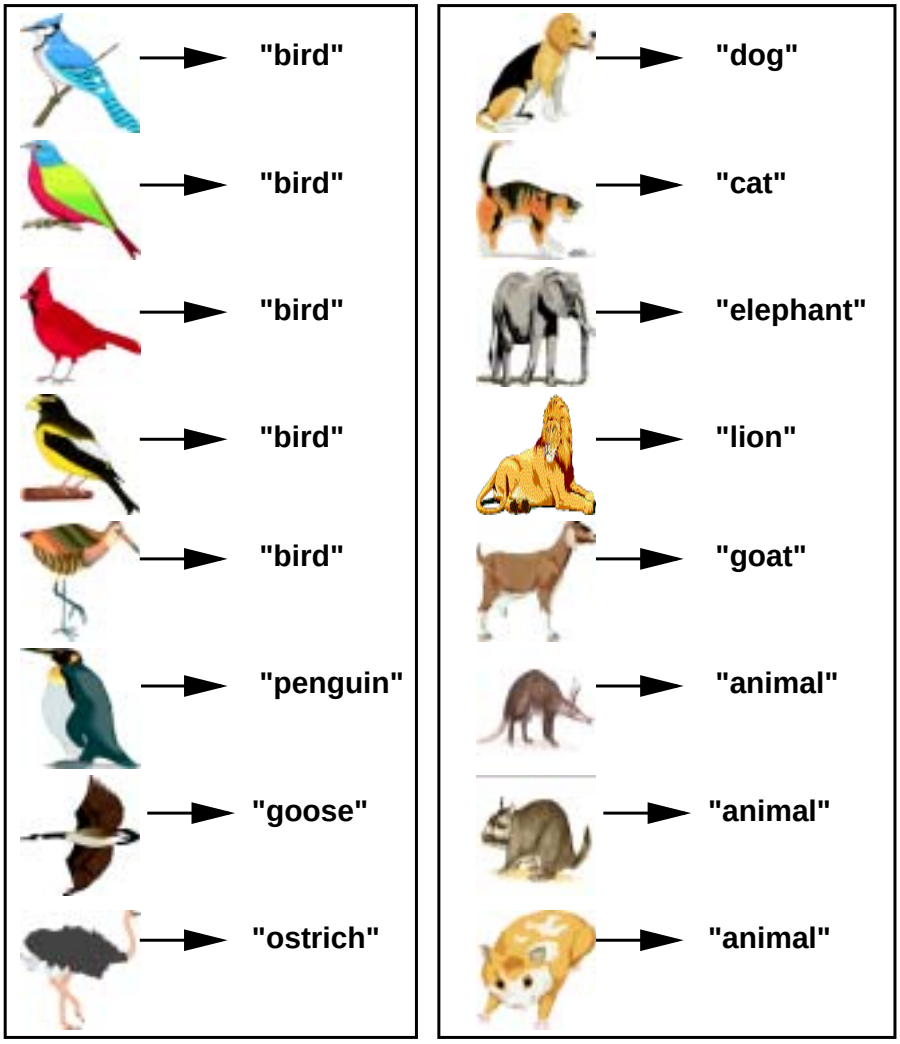

Figure 2. Training the network to name from a picture or a description, for the 16 animal items. The network was only trained to produce a single name for each object.

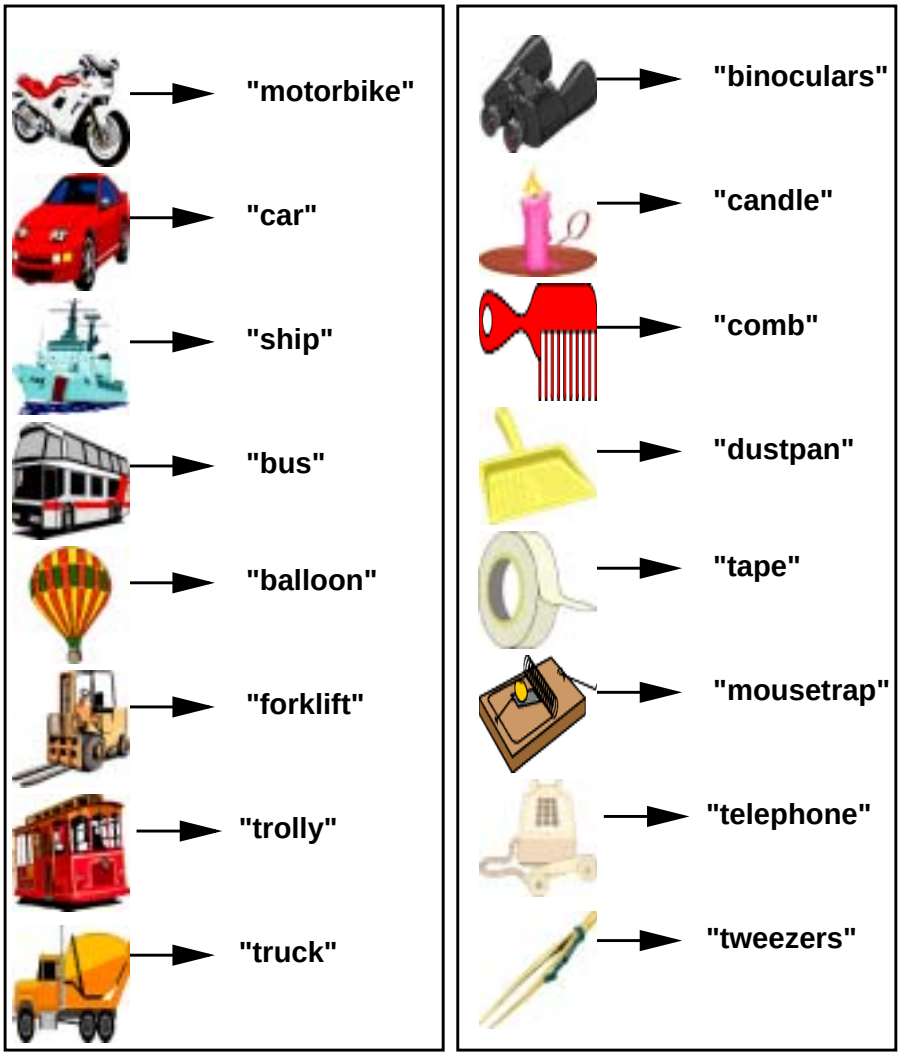

Figure 3. Training the network to name artifacts. Note there was no superordinate name for the "household objects" category.

However, when shown a name that could map to any of several visual or propositional representations (e.g. “Animal”), the network was trained with a particular representation chosen each trial at random from among correct alternatives.

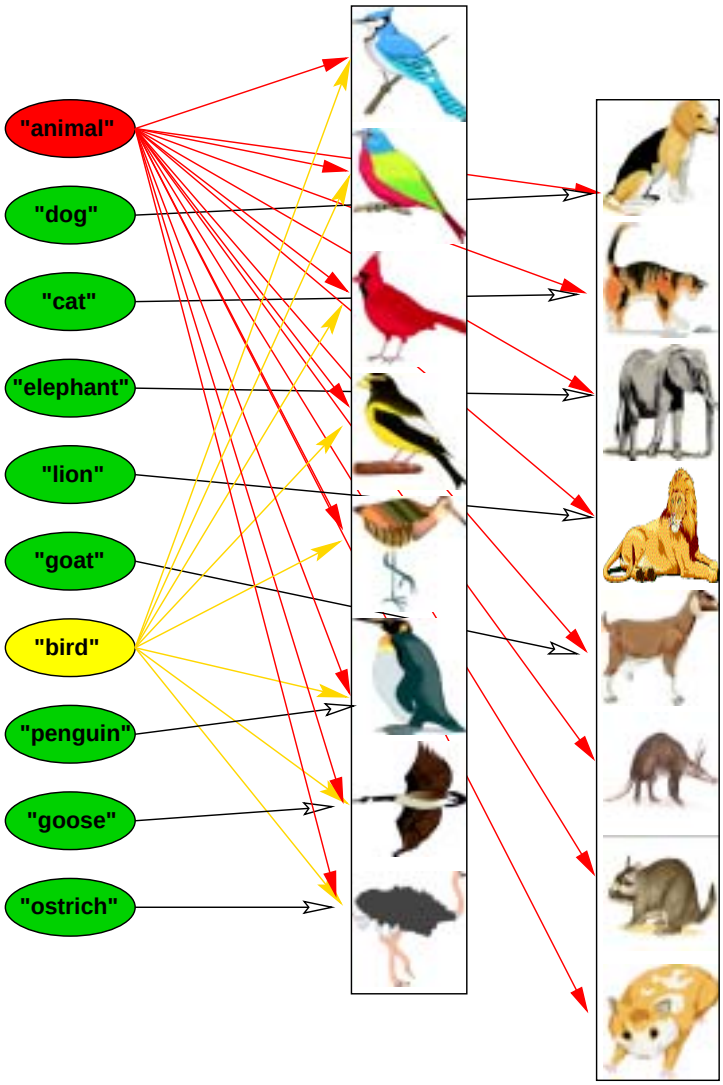

Figure 4. Training the network to produce a description of picture from name input. For more superordinate names, on each trial a particular pattern was chosen from among the correct alternatives. The net result was that, when the model was presented with a word such as “Animal” for input, the features shared by all animals would be active in the “Visual” and “Proposition” layers, whereas idiosyncratic features would not.

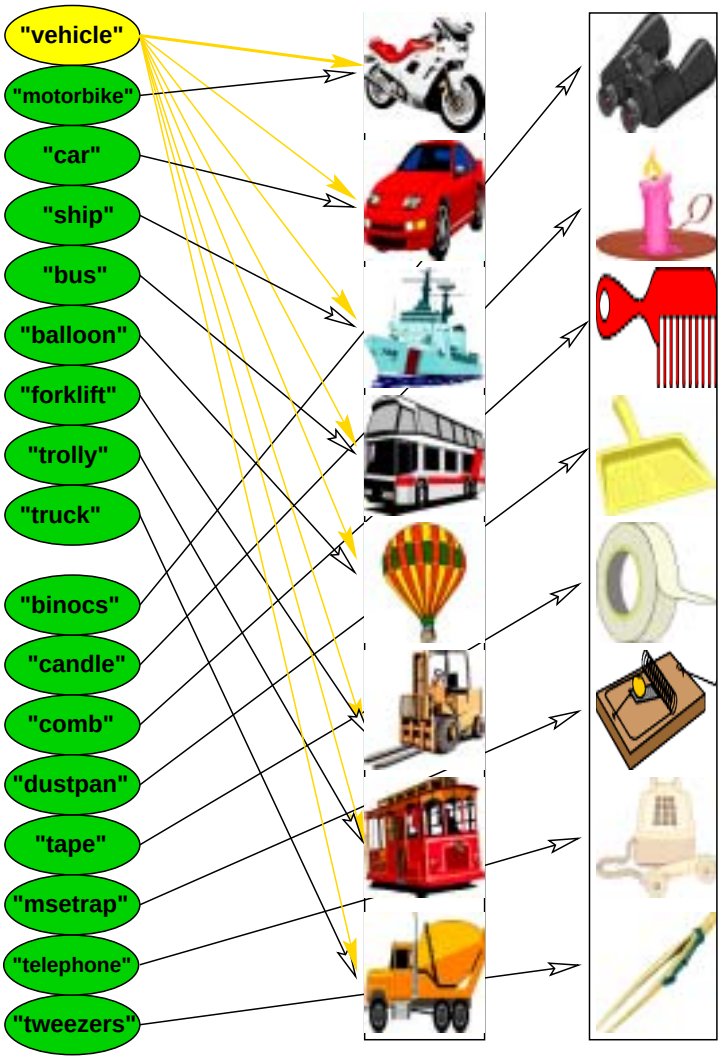

Figure 5. Training the network to produce descriptions or pictures for artifact name input.

The network learned about 32 objects in total: 8 birds, 8 land animals, 8 vehicles, and 8 household objects.

Simulating dementia

When it finished learning, the model could be probed with input in one modality (word, picture, or description), and would settle to a state in which all the correct features for the instance in each modality were active. Also, the distance between hidden unit states for various instances reflected the semantic relations between objects, such that similar objects had similar internal representations. The model provides an analog version of each of the semantic tasks shown in Table 1.

To simulate dementia, we removed an increasing proportion of all the connections in the network, and examined how its representations and performance changed.

Picture naming

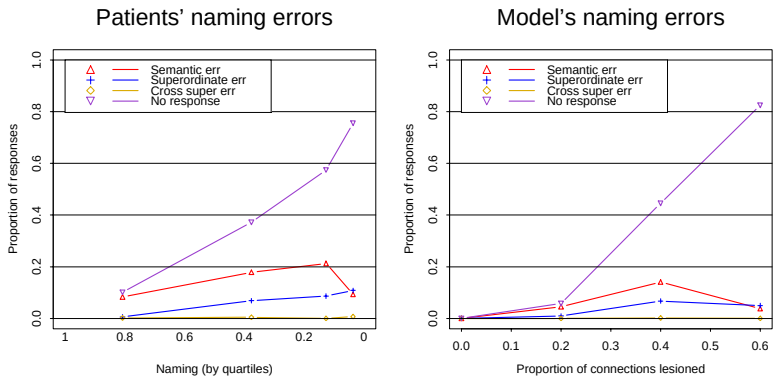

Figure 6. Picture naming errors from averaged patient data and from the model. Patient data are plotted against overall naming performance. We show the proportion of total responses for each kind of error. Note: i) An increasing number of "no responses," ii) A rise and decline in semantic errors, iii) A small but increasing proportion of superordinate errors, iv) No cross-domain errors.

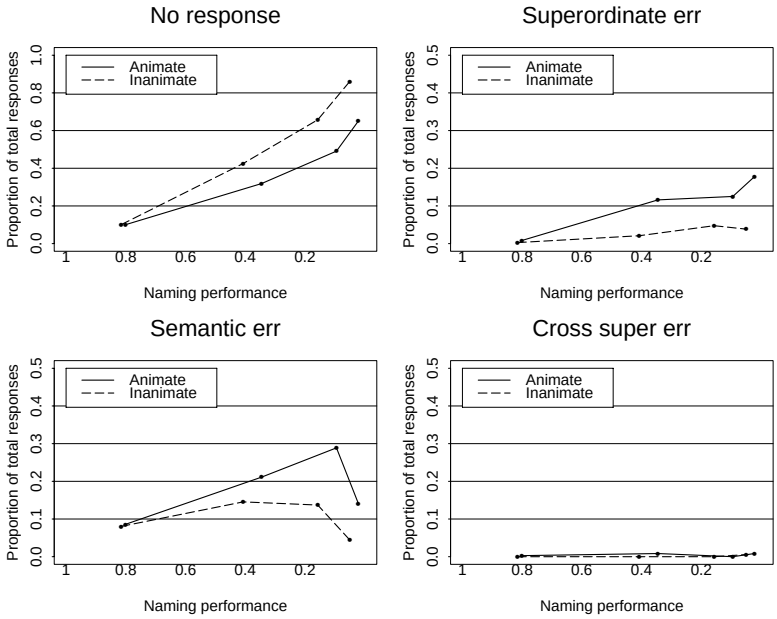

Figure 7. Picture naming errors split by domain for averaged patient data, and different error types. Patients typically give more "No responses" in the inanimate domain, but make more semantic and superordinate errors in the animate domain.

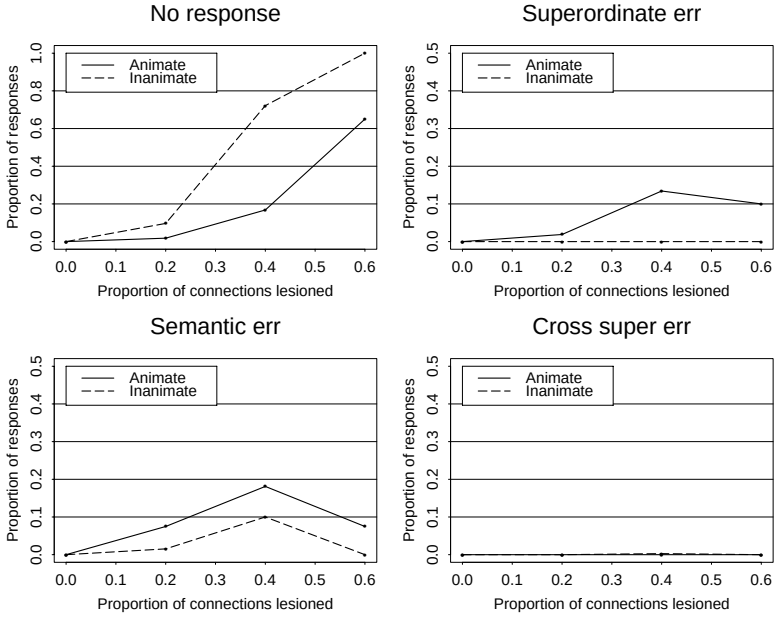

Figure 8. Picture naming errors split by domain for the model, which shows a qualitatively similar patterns of errors.

Word and picture sorting

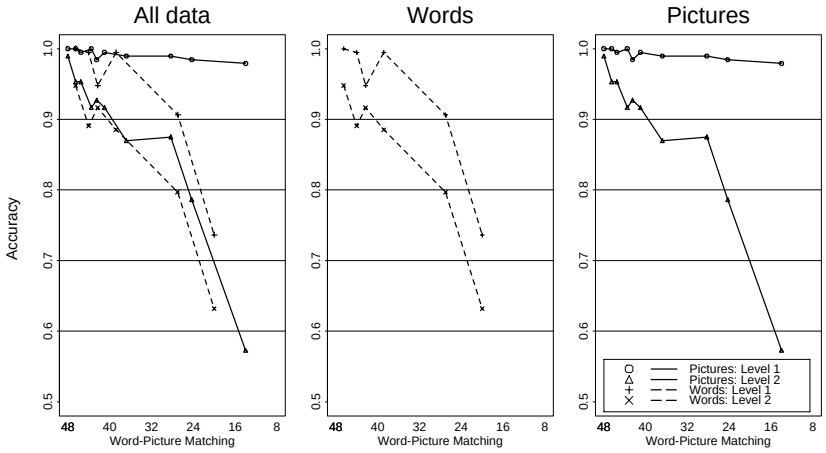

Figure 9. Average patient data for sorting words or pictures, at two levels of granularity: living vs. nonliving, or water vs. land vs. air creature. Patients are better at the more general sort, and better for words than for pictures.

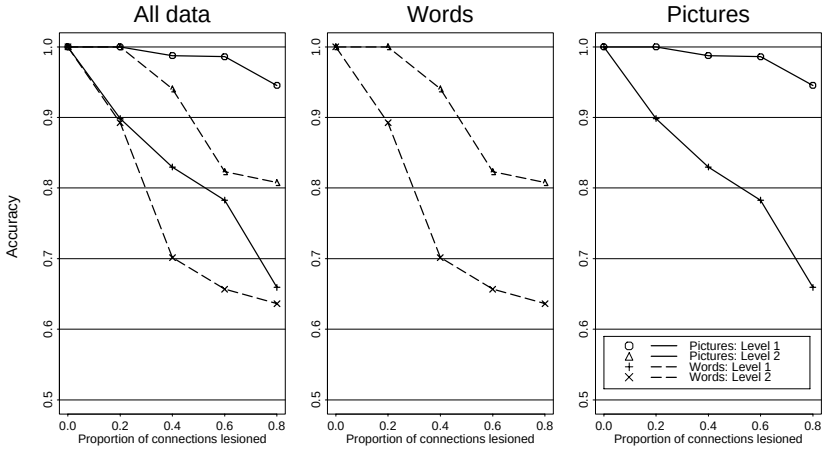

Figure 10. Analagous data for the model. The network divided instances into groups on the basis of the similarity among their learned internal representations. The accuracy measure indicates the proportion of times the network included, for example, animals in a group consisting primarily of artifacts.

Word to picture matching varying semantic distance

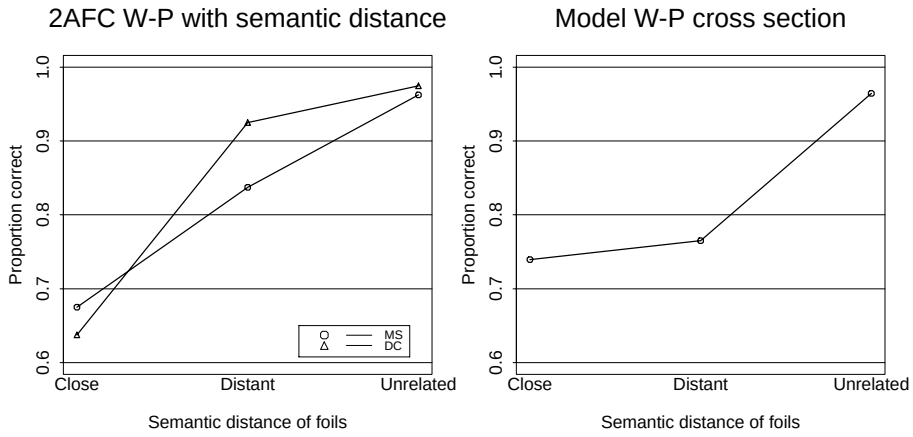

Figure 11. Word-to-picture matching data in a two-alternative forced choice paradigm, in which the distractor varied in its semantic relatedness to the target. The left graph shows the performance of two semantic dementia patients. Both performed near ceiling when the distractor was unrelated to the target, and near chance when it was closely related. The right graph shows the model's performance in an analagous task. The network's "choice" was made by comparing the similarity of its internal representations in response to the word input, the target picture, and a distractor picture. Like the patients, the model performs best when the distractor comes from an unrelated category, and worst when the distractor is similar to the target.

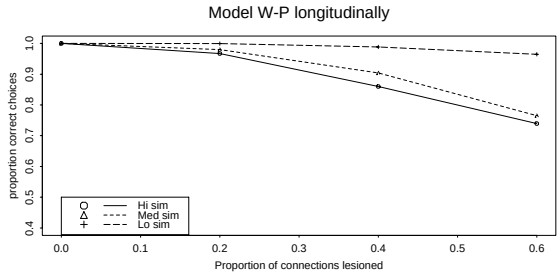

Figure 12. The patient and model data above is cross-sectional; however, the model allows us to predict how performance on this task should decline longitudinally. This plots shows the model's performance under increasing amounts of damage. Throughout the decline, the model is best when the distractor is unrelated to the target.

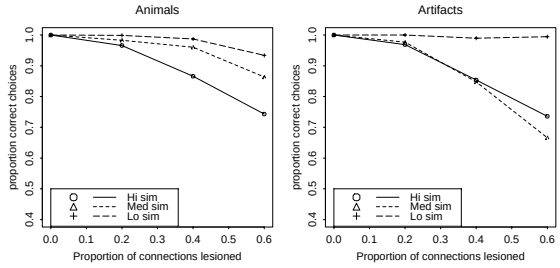

Figure 13. In the artifact domain, the model is equally bad when the distractor highly or moderately related to the target. In this simulation, there were few features reliably discriminating between vehicles and household objects. Thus the network under damage had a difficult time maintaining this distinction.

Drawing and delayed copy

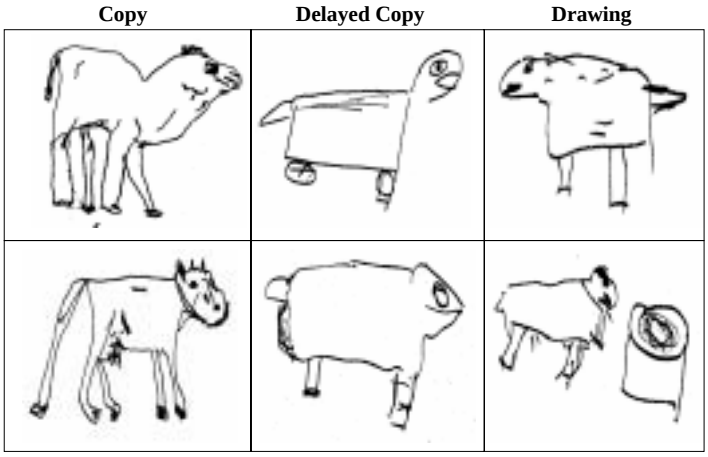

Figure 14. Immediate copy, delayed copy, and drawing to name for patient I.F., from two Snodgrass and Vanderwort pictures. These drawings show: i) fairly good immediate copying, ii) an overall loss of distinguishing features, iii) a striking overregularization in the delayed copy and drawing tasks, iv) the delayed copying seems to be less impaired than the drawing.

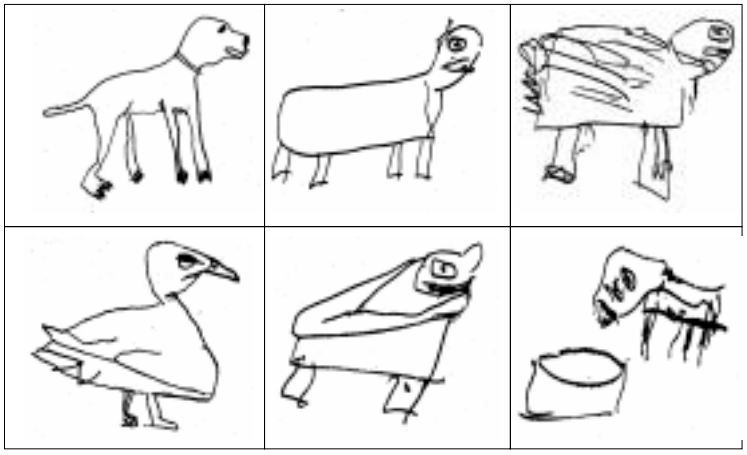

Figure 15. More of I.F.'s animal drawings.

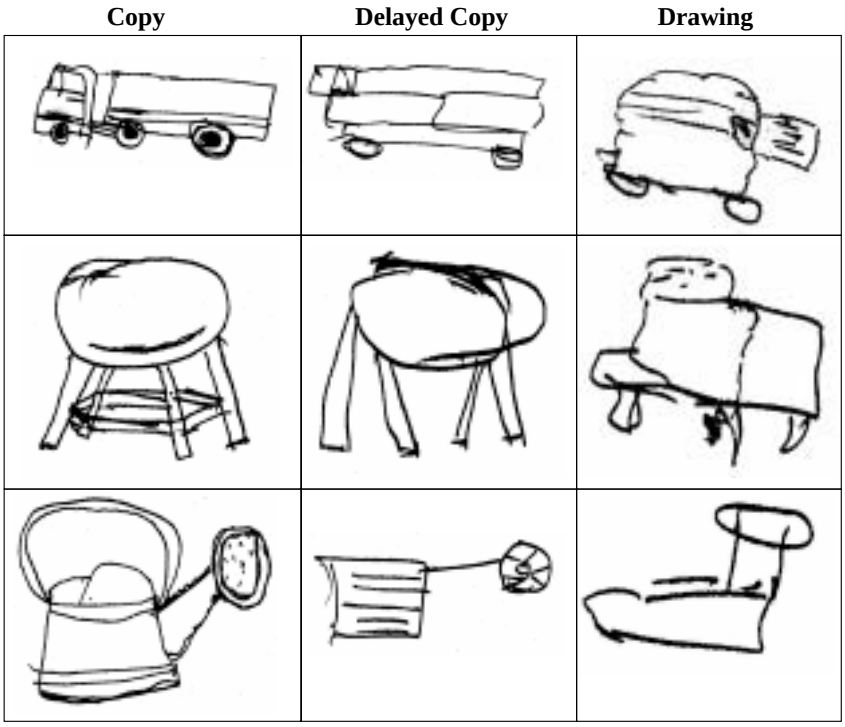

Figure 16. I.F.'s artifact drawings. Like the animal drawings, distinguishing features have dropped out in the delayed copy and drawing conditions; however, there is little overregularization. Again the drawings are more impoverished than are the delayed copies.

In the model, we examine the proportion of visual features missed (inactive when they should be active), and the proportion of false positives (active when they should be inactive).

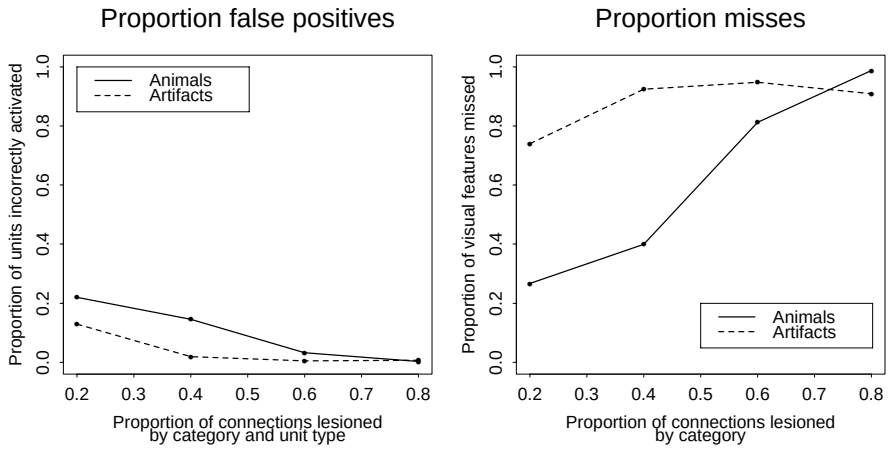

Figure 17. This figure shows the model's errors in producing visual output when probed with a name as input, split by domain. Overall, the network shows a higher proportion of false positives for living things, and a higher proportion of features missed for nonliving things.

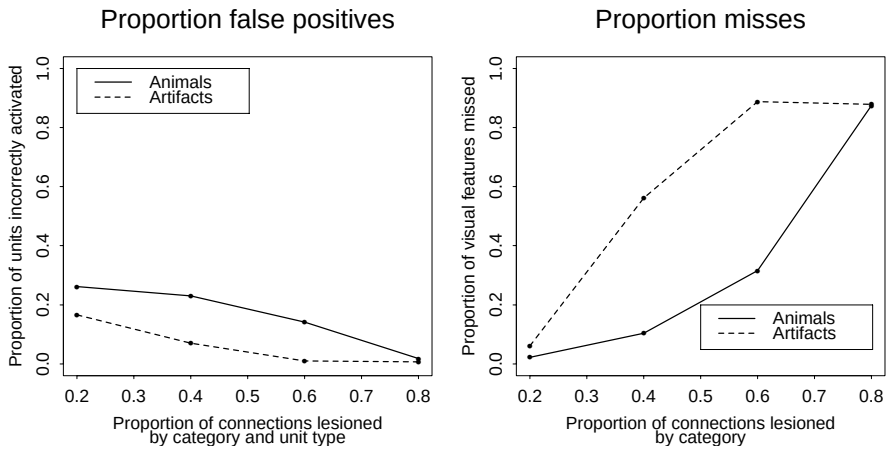

Figure 18. Here we show the same data for the model in a "delayed copy" task, in which the network is probed with visual input, and allowed to cycle for several time steps. The healthy network can maintain the correct visual representations; however, the damaged network cannot. The pattern of errors across living and nonliving things is similar to the model's "drawing" behaviour (above), but there are fewer errors overall.

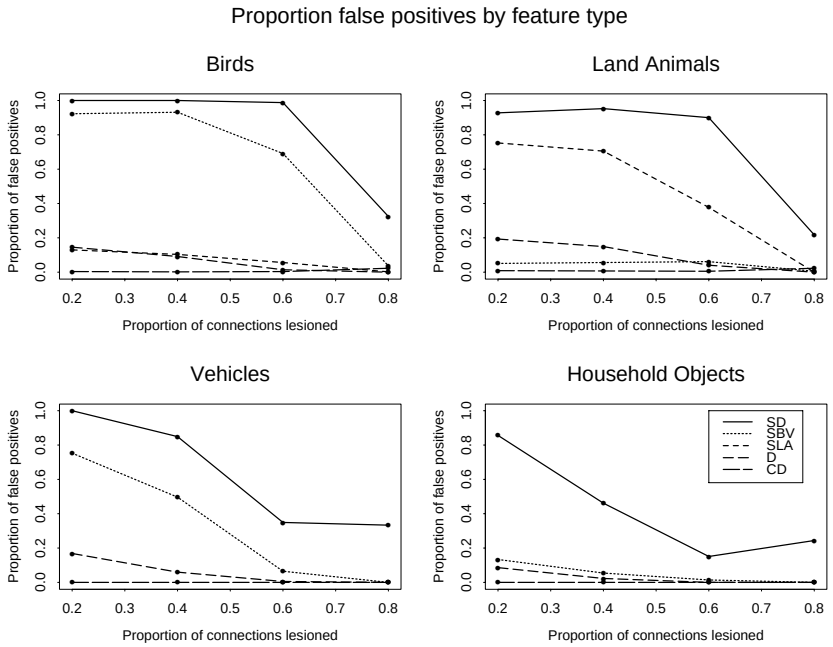

Figure 19. This figure shows the model's tendency to make false positive errors split by feature type and input for the delayed copy task. Each graph shows performance for the subset of objects indicated; for example, the graph labelled "Bird" shows average performance for all the bird objects. The lines indicate the proportion of false positives generated by the model for different kinds of features, as follows: SD are features generally shared across all members of the domain, SBV are features shared across birds (for the top two graphs) or vehicles (for the bottom two graphs), SLA are features that tend to be shared across land animals, D are distinctive features, and CD are features from the wrong domain. Distinctive features and features from the opposite domain are almost never inappropriately activated. Features that tend to be shared across domain or superordinate category are much more likely to be inappropriately activated under damage, but only when the object is from the appropriate category. For example, a feature such as "can fly" gets inappropriately activated by non-flying birds (e.g. penguin), but not by land animals or artifacts.

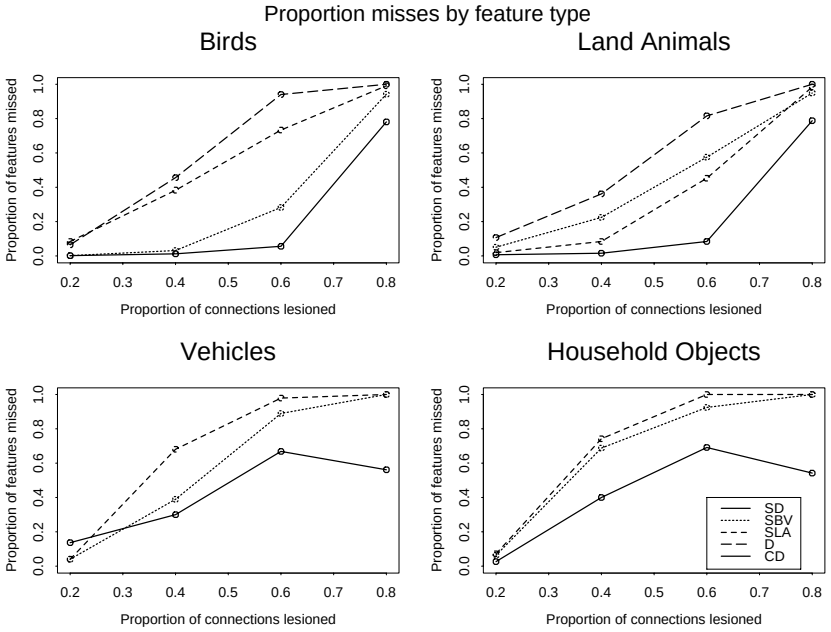

Figure 20. By contrast, features shared across the domain or superordinate category are much less likely to drop out under damage. Here we show the proportion of misses by object and feature type. Distinctive features are always the first to drop out. Features shared across a particular superordinate category are less likely to be inapporpriately inactive, but only when the instance belongs to the category. For example, a feature such as "has wings" is not likely to be missed for a bird, but is more likely to be missed for a land animal with wings (such as a bat). See the caption for the previous figure for an explanation of the legend.

Conclusions

In each of the six tasks for which we have data, both the model and the patients consistently show the four aspects of structured deterioration described in the introduction. How does the model help us understand these phenomena?

Because animate objects are generally richer and share more properties than inanimate objects, the network learns to build deep, highly-structured attractors for these instances. As the network's knowledge degrades, idiosyncratic object properties are lost. In highly structured domains, this causes the network to drift out of the correct state and into a more general attractor. In less well-structured domains, the loss of idiosyncratic information does not cause the network to drift, because there are no proximal competing attractors. As a consequence, the network tends to make errors of omission in the nonliving domain, and errors of both omission and commission in the living domain.