Revisiting the Role of Relearning in Semantic Dementia

Devon Jarvis (devon.jarvis@wits.ac.za)
University of the Witwatersrand, 1 Jan Smuts Ave,
Braamfontein, Johannesburg, 2000, South Africa
and
Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, University College London, 25 Howland Street,
London W1T 4JG, United Kingdom

Verena Klar (verena.klar@psy.ox.ac.uk)
Department of Experimental Psychology, Anna Watts Building, Woodstock Rd,
Oxford OX2 6GG United Kingdom

Richard Klein (richard.klein@wits.ac.za)
University of the Witwatersrand, 1 Jan Smuts Ave,
Braamfontein, Johannesburg, 2000, South Africa

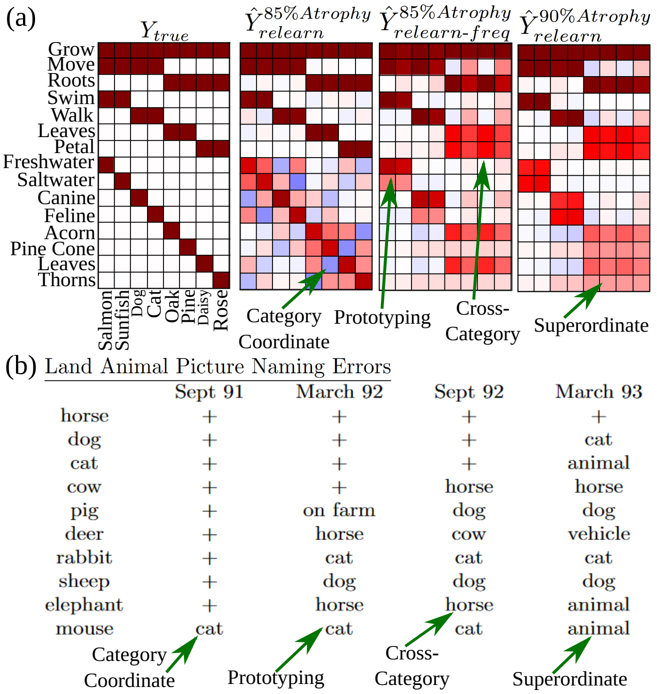
Benjamin Rosman (Benjamin.Rosman1@wits.ac.za)
University of the Witwatersrand, 1 Jan Smuts Ave,
Braamfontein, Johannesburg, 2000, South Africa

Andrew Saxe (a.saxe@ucl.ac.uk)
Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, University College London, 25 Howland Street,
London W1T 4JG, United Kingdom

Abstract

Patients with semantic dementia (SD) present with remarkably consistent atrophy of neurons in the anterior temporal lobe and behavioural impairments, such as graded loss of category knowledge. While relearning of lost knowledge has been shown in acute brain injuries such as stroke, it has not been widely supported in chronic cognitive diseases such as SD. Previous research has shown that deep linear artificial neural networks exhibit stages of semantic learning akin to humans. Here, we use a deep linear network to test the hypothesis that relearning during disease progression rather than particular atrophy cause the specific behavioural patterns associated with SD. After training the network to generate the common semantic features of various hierarchically organised objects, neurons are successively deleted to mimic atrophy while retraining the model. The model with relearning and deleted neurons reproduced errors specific to SD, including prototyping errors and cross-category confusions. This suggests that relearning is necessary for artificial neural networks to reproduce the behavioural patterns associated with SD in the absence of *output* non-linearities. Our results support a theory of SD progression that results from continuous relearning of lost information. Future research should revisit the role of relearning as a contributing factor to cognitive diseases.

Keywords: Semantic Dementia, Linear Neural Networks, Relearning, Anterior Temporal Lobes, Representation Learning



Land Animal Picture Naming Errors

	Sept 91	March 92	Sept 92	March 93
horse	+	+	+	+
dog	+	+	+	cat
cat	+	+	+	animal
cow	+	+	horse	horse
pig	+	on farm	dog	dog
deer	+	horse	cow	vehicle
rabbit	+	cat	cat	cat
sheep	+	dog	dog	dog
elephant	+	horse	horse	animal
mouse	cat	cat	cat	animal

Figure 1: As with a patient with SD ((b) taken from Hodges et al. (1995)), the model (a) makes category coordinate errors first and progresses to cross-category and superordinate errors. Additionally, if odd data points are shown twice as often during relearning ($\hat{Y}_{relearn\text{-}freq}$) then the more frequently seen features will dominate the representations. As a result, the less frequent objects are mistaken for the prototypical ones when making category coordinate and cross-category errors. Thus, a linear network with relearning reproduces the pattern of errors and prototyping effect associated with SD in humans.

Introduction

Patients with semantic dementia (SD) show a particular progression of behavioural impairments which result from the disease (Hodges et al., 1995) (examples shown in Figure 1(b)). Specifically, in SD errors of hierarchically organised semantic knowledge occurs first for fine grained distinctions at the bottom of a hierarchy (termed category coordinate errors) and progresses upwards until errors are made for more semantically distinct objects (cross-category errors) and fine grained distinctions are forgotten (superordinate errors) (Jefferies & Lambon Ralph, 2006). There is also a distinct typicality effect where patients are better able to identify frequently seen objects of a category with very typical features. This results in prototyping, where all objects of a category are named as the most stereotypical object in the category. SD has been linked to bilateral atrophy of the anterior temporal lobe (ATL). This finding in conjunction with converging evidence from brain imaging (Devlin et al., 2000; Visser et al., 2010) has led to the region being regarded as an amodal semantic hub (Jackson et al., 2021).

In this work, we explore the ability of simple linear neural network models to reproduce the pattern of behavioural impairment from SD. Such a model has been used successfully in the past to reproduce regularities in the development of semantic cognition during early childhood (Saxe et al., 2019). The networks hidden layer can be thought to represent the ATL. We test the hypothesis that disruption (atrophy) to the hidden layer itself is not sufficient as a model of SD, but that progression of impairment is better explained by a combination of atrophy and relearning.

Methods

Model We trained a linear neural network to describe the features of an object in a semantically hierarchical dataset, where features are shared based on how close objects are in the hierarchy. In contrast to prior approaches (Rogers et al., 2004), we do not rely on **output** non-linearities (specifically thresholding) to reproduce SD with our model. Instead, our network outputs real values representing the network's confidence in the presence of each feature. This allows for more direct comparison with human behavioural tasks with fine-grained actions or responses, such as picture selection tasks (Jefferies & Lambon Ralph, 2006) or object drawing tasks (Bozeat et al., 2003).

Training regime We first trained the network to convergence using (full-batch) gradient descent from small initial weights. This results in the network learning weights which identify the hierarchical nature of the input and map each distinction in the hierarchy to a corresponding set of output features. After convergence we deleted neurons in the hidden layer to model

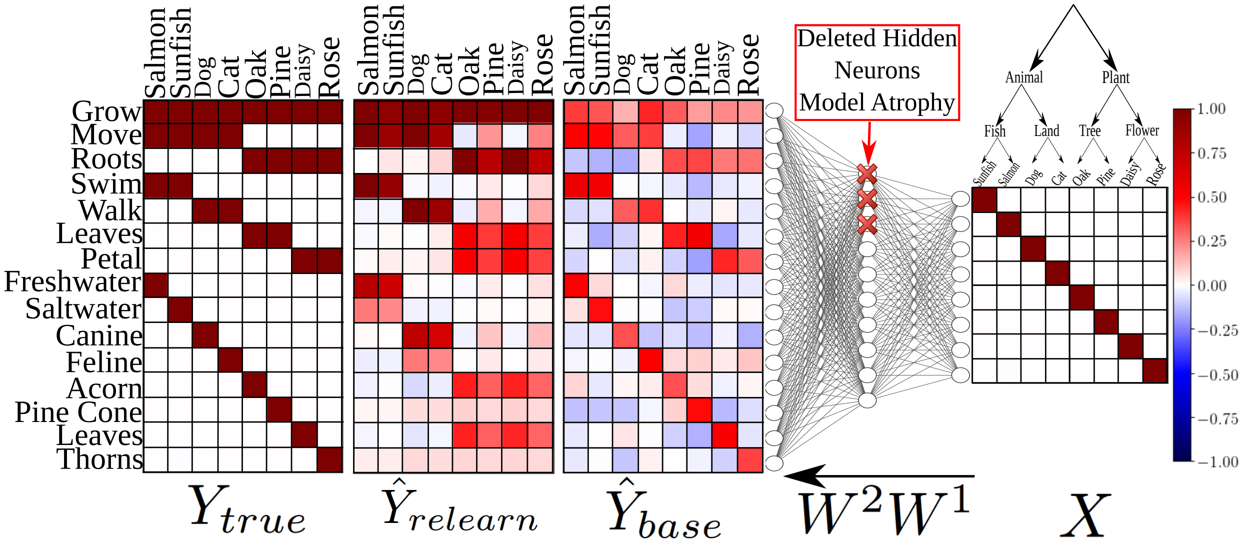


Figure 2: Setup and Primary Results: The linear network (W^2W^1) learns to map hierarchically structured data (X) to the corresponding features (Y_{true}). We model SD by deleting hidden neurons after training is complete. We compare two models: one with no relearning ($\hat{Y}_{base}$) and one with relearning ($\hat{Y}_{relearn}$). A model with no relearning loses information across all levels of the hierarchy at once, contradicting the patterns associated to SD. A model with relearning loses specific feature information before general features, consistent with what is expected of SD.

Model	Level 1 Error (Percentage)	Level 4 Error (Percentage)
Linear; $n = 0$	58.25	79.88
Linear; $n = 200$	0.0	62.50
ReLU; $n = 0$	51.50	84.50
ReLU; $n = 200$	2.75	62.50

Table 1: Relearning is necessary for models with linear and ReLU activations on the hidden layer to lose fine grained features before higher-level features due to atrophy in the hidden layer. Note that the output layer is still linear for all models (error is as a percentage compared to a naive model to aid comparison between hierarchy layers).

atrophy and retrained the model for n epochs after each deletion. This was repeated until no hidden neurons remained. A summary of the hierarchical dataset, network and findings are shown in Figure 2.

Results

The behavioural impacts of atrophy depends on how information is distributed in the hidden neurons and are represented by the errors made by the network. Our results show that a linear network is unable to reproduce the behaviour patterns of SD without relearning ($n = 0$) as the loss of semantic knowledge occurs for all levels of the hierarchy at once (see $\hat{Y}_{base}$ in Figure 1(a) and the first row of Table 1). However, if relearning is used ($n > 0$) the network reproduces these patterns (see $\hat{Y}_{relearn}$ in Figure 1(a) and the second row of Table 1). Table 1 also shows that the same effect can be seen when non-linearity is used on the hidden layer of the network.

Specifically, the network makes errors on the fine grained semantic features after fewer neurons are deleted (less atrophy) than for higher-level semantic features. Thus, category coordinate errors occur first. As the atrophy increases the network makes errors on semantic features for higher levels of the hierarchy (cross-category errors) and produces near 0 output for fine grained features which are forgotten (superordinate errors). An example of the pattern of errors from the model is shown in Figure 1(a) and can be compared to the example progression of the human patient from Hodges et al. (1995). Finally, Figure 1(a) also demonstrates that our model is sensitive to the frequency that it encounters features with. The model more rapidly forgets the semantic features which are encountered less often and makes errors by providing the features for the most stereotypical object at a given level of the hierarchy. Thus, our model also reproduces the prototyping effect observed in human SD patients (Bozeat et al., 2003).

Discussion

We have demonstrated that the typical progression of behaviour impairment from SD: category coordinate errors leading to cross category and superordinate errors; only emerges with relearning in an artificial neural network with linear output producing semantic features for hierarchically structured objects. Our results suggest that it is relearning after atrophy – the brain's attempt to adapt to the disease – rather than atrophy itself which causes patterns of memory degradation in SD. A similar role of relearning has been established in herpes simplex virus encephalitis (Ralph et al., 2007), but remains unexamined in other chronic cognitive diseases. Our results call for consideration of the role of relearning as a primary cause of the known behavioural patterns in these diseases.

Acknowledgements

This work was supported by the Commonwealth Scholarship and Google PhD Fellowship to DJ, the ESRC ES/P000649/1 and New College 1379 Old Members Scholarship to VK, a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (216386/Z/19/Z) to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755). A.S. and B.R. are CIFAR Azrieli Global Scholars in the Learning in Machines & Brains program.

References

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., ... Hodges, J. R. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive neuropsychology*, *20*(1), 27–47.

Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Wilson, J., Moss, H. E., ... Tyler, L. K. (2000). Susceptibility-induced loss of signal: comparing pet and fmri on a semantic task. *Neuroimage*, *11*(6), 589–600.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, *3*(3-4), 463–495.

Jackson, R. L., Rogers, T. T., & Lambon Ralph, M. A. (2021). Reverse-engineering the cortical architecture for controlled semantic cognition. *Nature human behaviour*, *5*(6), 774–786.

Jefferies, E., & Lambon Ralph, M. A. (2006). Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain*, *129*(8), 2132–2147.

Ralph, M. L., Lowe, C., Rogers, T. T., et al. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, hsve and a neural network model. *Brain*, *130*(4), 1127–1137.

Rogers, T. T., McClelland, J. L., et al. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Visser, M., Jefferies, E., & Lambon Ralph, M. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of cognitive neuroscience*, *22*(6), 1083–1094.