Giving Deep Neural Networks Semantic Dementia: Using the Picture Naming Task as a Measure of Neuromorphism

Author: Tim Ufer, s4689380
Main supervisor: Tim Kietzmann (Radboud University Nijmegen)
Direct supervisor: Émer Jones (University of Cambridge)

January 30, 2021

Abstract

Neural networks that model the way the brain works are usually evaluated based on data from exclusively healthy patients. This project provides an additional perspective for these networks by 'lesioning' them in correspondence with atrophy locations in the brains of Semantic Dementia patients. Then, their behaviour on the Picture Naming task can be compared to the robust error patterns of these patients. This allows for model comparison between two networks not based on their performance on a test-bed of healthy brain data but in an impaired state.
For this project, two vNet models with different training structures were tested. While both demonstrated a considerable degree of neuromorphism, the category-trained model performed better on the test-bed than the model trained on a semantic objective, contrary to expectations.

Introduction

Deep artificial neural networks are getting increasingly better at mimicking or even exceeding human performance in cognitive tasks such as image classification. This poses the question of how brain-like such neural networks actually are and whether their mechanistic similarity to the brain (neuromorphism) can be quantified. While performance does correlate with neuromorphism [1], comparative studies have shown that human-level performance does not imply that a network uses the same mechanisms as the brain [2]. These discoveries have led to a need for specialised evaluation systems to verify the quality of neural networks which have been engineered to emulate the brain in certain functions. An increasingly popular benchmark for these questions is *Brain-Score* [1], a scoring system that compares the responses to certain stimuli of any neural network to response data from human patients. However, *Brain-Score*'s human data exclusively consists of healthy patients and thus disregards the wealth of knowledge about the brain that is related to impaired brains.
Following this notion, the larger project under the direction of Tim C. Kietzmann at the Donders Institute aims to develop a modelling framework in the context of human ventral stream computations that promotes neural networks which are mechanistically similar to both healthy and damaged brains. To enable concrete evaluations of 'damaged' networks, a single disorder was chosen to be used as a test-bed: Semantic dementia.

Background

Semantic dementia (SD) is a form of dementia that is mainly characterised by an increasing, general loss in semantic memory, primarily in the verbal domain. Common symptoms include receptive

aphasia (difficulty understanding language), associative visual agnosia (difficulty recognising/naming objects or images) and a decreased understanding of word meaning (semantics) in general. There are two reasons why SD was chosen for this project: First, its behavioural symptoms are highly selective and robust across studies and subjects [3], and second, the atrophy that causes SD is always focal, progressive atrophy localised in the temporal lobe. To be more precise, the damage is concentrated in the ventral anterior temporal lobe and the anterior hippocampus. As the disease progresses, the atrophy extends over the temporal lobe in general and becomes more severe. This project will focus on one particular task that has been used to research SD and semantic memory as a result: Picture Naming.
Picture Naming consists of the sequential presentation of pictures of objects and animals that the patient must name. The correct answer is defined by responses collected from healthy control subjects. Patients with SD show robust error patterns wherein the ratio between the error types changes as the disorder progresses. The following list ranks these error types by their average frequency.

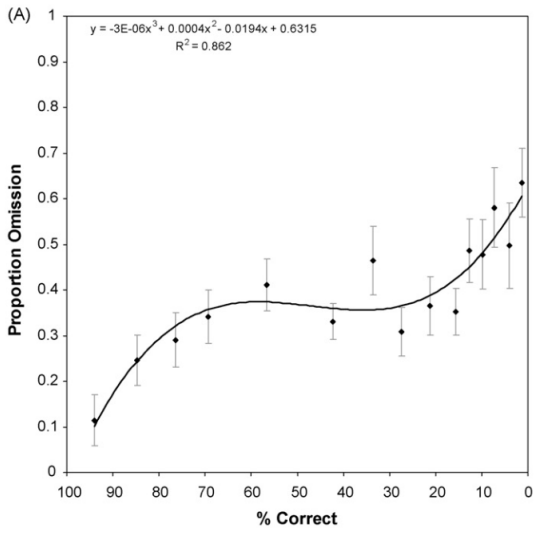
1. *Omission error*: The patient fails to provide any label.
2. *Semantic error*: The patient provides an incorrect but semantically related label.
3. *Superordinate error*: The patient provides a correct but overly general label.
4. *Cross-domain error*: The patient provides an incorrect label from a different semantic domain.

To understand the relevant terminology, consider the following example: An image of a dog is presented to the patient. As for the semantic domains, assume the common distinction for this task, *man-made* and '*alive*. If the patient now labels the image as 'cat', that would constitute as a semantic error since the label was incorrect but both 'dog' and 'cat' are part of the semantic domain *alive*. If however, the patient labelled the image as 'chair', it would be a cross-domain error since 'chair' is in the *man-made* domain, contrary to 'cat'. Alternatively, labels such as 'animal' or 'pet' would be characteristic of superordinate errors while the lack of any comprehensible label corresponds to omission errors.
It is important to note that cross-domain errors are essentially never made across all stages of the disorder while omission errors account for more than 80% of errors made in later stages [4].
In terms of implementing picture naming in neural networks, the task is image classification. In the brain, this is handled by the visual ventral stream which in turn can be modelled by convolutional neural networks [5]. This is part of the reason why the picture naming task was chosen for this project, as the low inferential distance between humans performing the picture naming task and neural networks performing image classification ensures a certain baseline of validity for the test-bed.
Furthermore, the atrophy in the brain of an SD patient lies in part in the ventral stream which enables a more valid translation to the digital world as deep neural networks correspond to the ventral stream at the level of information processing. Based on that, the damage can be simulated by applying *dropout* in the last layer of a network. *Dropout* is a regularisation technique that omits units in the network during training. While it is normally used to prevent overfitting, this project will use it as digital atrophy.
With all that in mind, the purpose of this project is to contribute to a larger modelling framework for deep neural networks by creating a test-bed in the image of SD. That test-bed will provide additional insight when comparing neural networks on their neuromorphism. If successful, neural networks that maximise mechanistic similarity to the brain could be evaluated in the context of both healthy and abnormal brains by using a healthy-brain benchmark such as Brain-Score and our test-bed respectively. In any case, it should be emphasised that the project is not trying to model SD as a disorder and does not seek to optimise the performance of any given neural network used in development.
Thus my research question is: *Can the effects of Semantic Dementia be replicated in arbitrary neural networks to evaluate their similarity to the brain?*

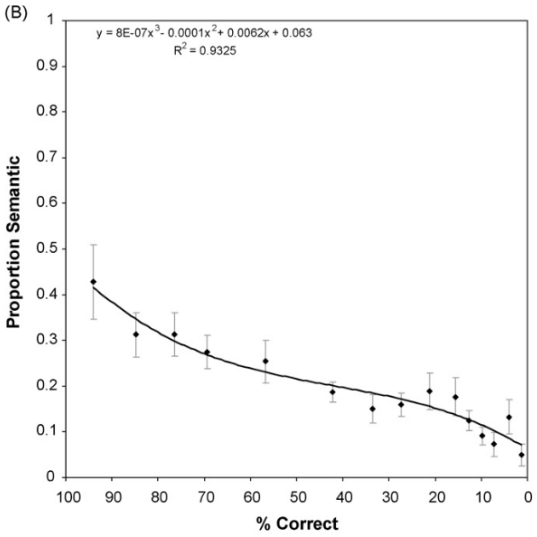
Related work

To enable accurate comparisons in our test-bed, this project requires reliable data about the behaviour of semantic dementia patients within the picture naming task. The primary source for this purpose is the 2008 paper 'Anomia: A doubly typical signature of semantic dementia' by Woollams

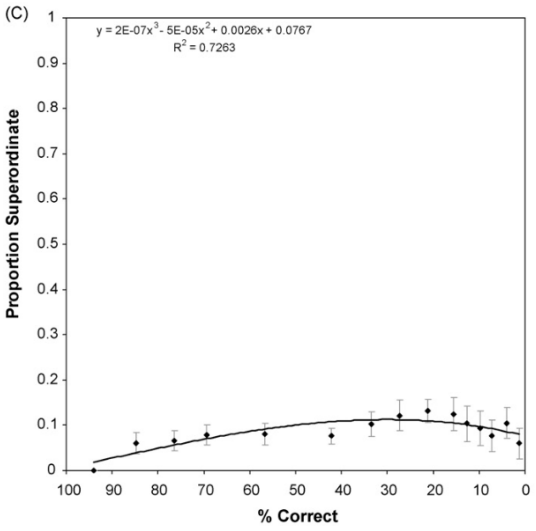
et al. [6], who conducted an extensive analysis on picture naming data over 78 patients and includes cubic curves fitted to their data as well as various graphs to describe the error patterns in patient responses during picture naming. To understand these graphs, it is important to know that all the



(a) Proportion of omission errors over correct responses



(b) Proportion of semantic errors over correct responses



(c) Proportion of superordinate errors over correct responses

Figure 1

patients (represented by the dots) are affected by SD to different extents. Additionally, it is generally assumed that the level of atrophy has a negative monotonic relationship with the correct response rate in the picture naming task. With that in mind, the three graphs nicely show how semantic errors (figure 1b) start out as the dominant error type but are surpassed by omission errors (figure 1a) at roughly 75% correctness while superordinate errors (figure 1c) are barely significant for most of the data. Note that no graph was made for cross-domain errors due to their exceeding rarity. In summary, patients often omit or mix up labels within a semantic domain, but rarely provide overly general labels and almost never confuse labels across semantic domains.
These findings were preceded by a study in 2004 which created a computational model to investigate the mechanisms behind semantic memory in general [4].

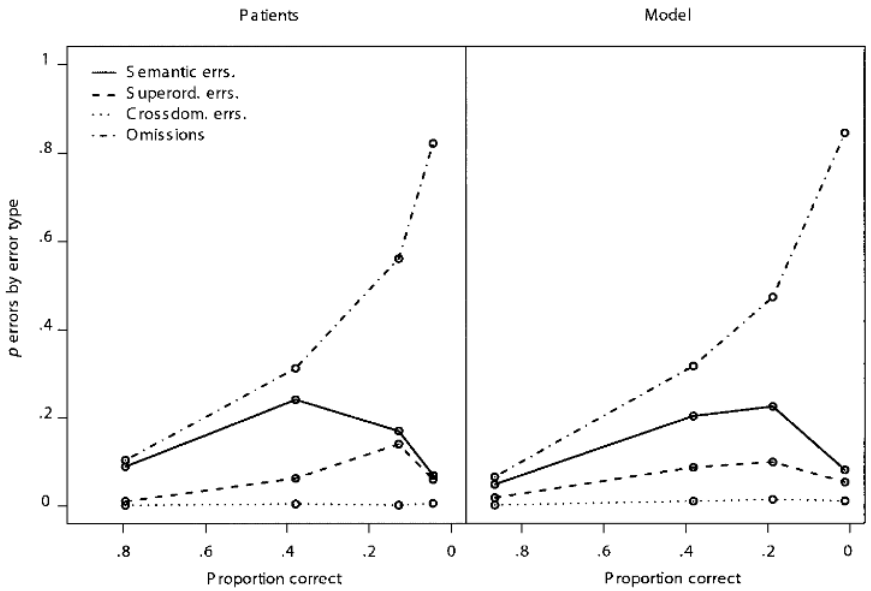


Figure 2: Proportions of the four error types over correct responses

Their research includes a comparative graph of the four error types (figure 2), that better illustrates the differences in frequency between the types. Note that their patient data starts at 80% correct, thereby skipping the brief phase where semantic errors are the dominant error type.

Research

Approach

As the project focus lies with the test-bed's ability to test arbitrary models, we use two particular pre-trained deep neural networks for image classification as a proof of concept. These networks will select labels from *Ecoset*, an image training set for neural networks that consists of 565 categories and over 1.5 million images which are designed to better represent common visual input of humans [7]. For the purpose of the picture naming task, scenes like 'desert' or 'city' were removed entirely from the set because the task always uses singular objects and animals. The remaining categories were split into two semantic domains, namely *man-made* and *natural*. To avoid any base-rate bias toward either domain, the number of categories in each domain were balanced out by removing categories from the originally larger *man-made* domain until both were of equal size (217 categories each). Categories were selected for removal based on their frequency attribute in *Ecoset*, meaning that those removed are of lesser significance in semantic memory.
For the actual input data, 4800 images were selected from 60 of the 434 categories and split evenly into a train and test set, meaning each set contains 40 unique images for each of the 60 shared categories. The aforementioned networks are then presented with these images, have their best estimate recorded and are judged on this response in the context of the picture naming task and its error types. This process is run iteratively to test out different thresholds for the omission error type where the results of each run are evaluated in terms of their correlation coefficient with the cubic curve from the 2008 paper. We chose Pearson's correlation coefficient over the mean squared error as our metric since we want to prioritise matching the shape of the error type distribution over the absolute values. This procedure does *not* encourage a common threshold among the different networks, or even between different iterations of the same network. Once the best threshold for every network in question has been found, they can be compared on their neuromorphism.

Ecological validity

Creating a digital version of the picture naming task for neural networks naturally led to some key differences with the original. First of all, we attempt to simulate Semantic Dementia, or rather its characteristic atrophy, in the given networks by applying dropout 'lesions' to an unusual extent. This is only an experimental approximation of the actual physical changes that are caused by SD and may present an inherent inaccuracy for this project. Additionally, SD patients might well have different omission thresholds for different objects, but as this internal threshold can not be verified by current research, this project assumes a single, uniform threshold for all objects shown in one iteration of the task. As mentioned before, the extent of atrophy in SD patients and their correct response rate are thought to have a negative monotonic relationship, so that was used as a quality criterion during the development of our digital SD.
The image data represents another issue as its large volume effectively made it impossible to verify the images in terms of topic clarity and quality. This is mainly because such data sets are sourced automatically and not validated in their entirety by a human. The original picture naming task was done in person with hand-picked drawings which allowed the researchers to control the attention of patients towards the target object or a purposefully placed distractor. As this project utilises merely 4800 images instead of 1.5 million, a few bad images per category can be sorted out but it still sets a higher bar for object recognition than the original. It should be noted that although the picture naming task is usually done with line drawings, patient behaviour has been shown to be highly robust when using either drawings or images. The correlation coefficient between the two versions was found to be 0.99 [6].
Lastly, *superordinate* errors are not included as a proper implementation would require a cascading design which would be equivalent to additional prompting that actual patients never receive in this task. Such a design would make the network attempt to find a label on the regular image-level first, and move on to higher, more general levels upon failure. As the omission errors depend on a definitive threshold in the decision process, this concept would cause a contradiction between omission and superordinate errors which would be a highly inaccurate depiction of the error ratios in patients. For these reasons, we do not expect any network to perfectly mimic the patient data but rather to show similar tendencies over many data-points, which is why the correlation coefficient was chosen as the evaluation metric.

Networks & Models

Two specific neural networks are used as test models; the first was trained on *Ecoset* categories and used to optimise the task implementation. The second model was trained on 300-dimensional fasttext word embedding vectors and its readout layer was re-trained on *Ecoset*'s categories, which was a necessary adaption to homogenise the output format of the two networks for comparative purposes. The latter approach represents a semantic objective where the semantic distance between nouns is simulated. They will be further referred to as the category-trained and semantics-trained model respectively. Both models are vNets with 10 layers and kernel sizes which roughly correspond to the biological foveal receptive field sizes in the visual system. After the task has been successfully executed, the performance of the two lesioned networks is compared and their differences evaluated in the context of neuromorphism.
The severity of these 'lesions' is an increasing parameter to reflect the progressive nature of the disease. The steps in which that severity is increased and which levels correspond to which stage of SD is to be decided uniformly for the entire group project and therefore not set here. An important criterion in that regard is whether the ratio between the different error types over the different severity levels reflects the changes that occur in errors made by patients as their disorder progresses. If it does, the threshold for *omission* errors is the most significant parameter in the test-bed as these errors account for an ever-increasing share in the error ratio in patients.
While we generally did not make any structural changes to the code of the networks, their exact implementation required more adaptions than initially expected as it was discovered that several commonly used *TensorFlow* functions were causing certain problems for our implementation. First, it turned out that the standard *Dropout()* function has built-in re-scaling of node values to improve a model's robustness. While generally desirable, this runs counter to the purpose of using dropout as a form of deliberate impairment. Therefore, the node values are manually re-scaled in the opposite

direction after calling *Dropout()* in the last layer of a network. Secondly, the softmax function was inherently affected by the value distortion caused by the lesioning of the models with dropout, which necessitated the use of a custom normalisation function instead (see Implementation: Normalisation).
It is important to keep in mind that these direct changes to the networks are to be seen as giving any model a fair chance rather than optimising them specifically for better scores.

Implementation

This project's implementation builds on top of previous work done by the Kietzmann Lab at the Donders Institute and is therefore based in *Python* and *TensorFlow*. Regarding the four error types, only *omission*, *semantic*, and *cross-domain* are implemented in the following way:

- *Omission* errors are coded as a threshold on the confidence value of nodes in the readout layer of a given neural network. A node is only counted when it reaches this threshold and omitted of not. This means that a threshold value is introduced and optimised for each lesion level.
- *Semantic* errors are coded as an incorrect output label from the same semantic domain (either 'man-made' or 'natural') and are not possible if the *omission* threshold value is not reached.
- *Cross-domain* errors are coded as an incorrect output label from the other semantic domain. As they rarely occur in patients, large quantities of such errors produced by a network would indicate that that network is lacking in similarity to the brain but it does not necessarily mean that the network is bad at the task. *Cross-domain* errors are not possible if the *omission* threshold value is not reached.

As mentioned before, the initial goal is to determine the best omission error threshold for each given network and then the picture naming task is run again with only that threshold and the second set of input images to generate final results. The procedure is outlined below and the actual code is attached to my submission.

Setup

First of all, the range of omission thresholds and lesions must be set. As we had no expectations regarding which thresholds might be better than others, we chose to generate a linearly spaced vector with 20 values:

$$\{0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\} \quad (1)$$

As for the lesions, early experimentation with the category-trained model suggested using an uneven spread of 14 values to maximise information gain per value without overly increasing the required runtime:

$$\{0.2, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.8, 0.9, 0.93, 0.96, 0.99\} \quad (2)$$

These values represent the dropout parameter indicating the portion of nodes to be 'dropped'. For those unfamiliar with dropout, it should be noted that a drop rate of 0.2 (20%) was used in both of our test networks to make them more robust and less prone to overfitting. Because of that, a lesion value of 0.2 is not considered as significant damage in the frame of our digital atrophy and thus serves as a baseline to the higher lesion values. It should be noted that the last three values $(0.93, 0.96, 0.99)$ were only included to better illustrate the complete decline of general performance but are less relevant for the error patterns and ratios. They were therefore not included in some runs to reduce the required runtime.
Once those ranges are set, a completely unlesioned version of the model (drop rate $= 0.0$) is run through a shorter version of the picture naming task to collect values necessary for later normalisation (see corresponding subsection). Next, the model is instantiated for each lesion level and runs the following steps in combination with each possible omission threshold. To simulate multiple patients and because dropout is a probabilistic function, every such combination is run 5 times and aggregated into a single set of values by computing the mean.

Preprocessing

The image train set is retrieved from storage, where an equal number of images is taken from each of the 60 selected categories. During that process, the correct label and semantic domain for each image are recorded for later. Black & white or corrupt images are filtered out. The preprocessed images are then split into appropriately sized batches to avoid resource issues during computation.

Picture Naming

All image batches are run sequentially through the model of interest to obtain its classifications. The classification procedure itself is inherent to the model and not directly modified. These values are normalised with a custom normalisation function (see following subsection) before being processed with respect to the category labels. Using the previously saved list of correct labels and corresponding semantic domain, the model's responses are evaluated in terms of chosen SD error types.

Normalisation

The confidence values that a vNet initially generates are not bound to a fixed range like $[0, 1]$. Normalising the values to such a range is immensely useful as it provides a reference frame for sensible thresholds and makes them easier to interpret as ratios. The models initially used TensorFlow's softmax function for normalisation, but the use of dropout for lesioning distorted the values in the denominator of the function and thus its normalised output as well. To understand this effect, one should recall the softmax equation:

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}} \tag{3}$$

Every input value i is turned into an exponential (e^{x_i}) and divided by the sum of all the exponentials generated from the input set ($\sum_{j=1}^{n} e^{x_i}$), thus creating a proportional relation that converts the input values to a probability distribution where all resulting values add up 1. However, this relation is disturbed by dropout since it effectively removes values from the input set by 'dropping' nodes in the network without reducing the size of the set. That means that the numerator and denominator of equation (3) are no longer conceptually connected and the resulting probability distribution will not add up to 1 anymore.
Attempts to avoid this issue with a custom softmax function which utilised the denominator from an unlesioned model's softmax caused the resulting values to vastly increase beyond the supposed range of $[0, 1]$, as the exponential nature of softmax translates the linear changes in dropout rate to exponential changes in its results. As this defied the original purpose of using softmax in the first place, this issue could only be solved by switching to a simpler min-max normalisation as a replacement:

$$x' = \frac{x - min(x)}{max(x)} \tag{4}$$

As mentioned above, an unlesioned version of the model is run that stops after collecting the raw responses of the model and before these values are evaluated with respect to the error types. The normalisation function is called to only extract the denominator term of equation (4), $max(x)$, and save it in the overarching environment. All later iterations of the picture naming task that are instantiated to test the model over the various lesion levels and omission thresholds, import this previously saved term into their normalisation process to avoid the value distortion caused by their respective lesions.

Omission Threshold Fitting

Upon closer inspection of figure 1a, one might notice that the equation in the top left corner of the graph does not accurately reproduce the line drawn. As their equation was built on top of the depicted data-points, we assume that it was depicted with insufficient precision and therefore used a slightly different equation to better represent the paper's findings in our test-bed.

$$\begin{aligned} &-0.00000311x^3 + 0.000438x^2 - 0.0194x + 0.632 \\ &\text{where } x = \text{correct response rate} \end{aligned} \tag{5}$$

To determine this more accurate equation, a web tool[8] for extracting data points from charts was used to generate a corresponding data set which in turn was fed into a function approximation calculator. As the constant value in the original equation was already sufficiently precise, it was fixed during fitting to avoid unnecessarily diverging from said original. The resulting equation (5) required a minimum of three significant figures to produce a curve indistinguishable from the original graph, which is what we used.
Using this new equation, a predicted omission error rate is generated based on a given correct response rate. That prediction is then compared with the omission error rate of the model to compute the correlation coefficient between the two. Based on the highest correlation score, the best threshold for the given model is identified. All above steps are then repeated with the image test set and the top threshold alone to verify the results.
The procedure is repeated for every given network and their respective results can be compared at the end to determine which of them is more brain-like.

Results

Category-trained Model

Applying the previously described procedure to this model resulted in the selection of the 0.45 omission threshold based on a correlation coefficient of 0.717.

Figure 3: Comparison of the omission error rate created by the category-trained model and the omission error rate predicted by the 2008 paper for the omission threshold found to be optimal for the category-trained model

On first glance, the trend of the model's omissions (figure 3) may appear too extreme, but this is simply because all the model's data points fall into the range of $[0.7, 0.2]$. Recalling the graph from the 2008 paper (figure 1a), it becomes clear that the model successfully emulated the trend of the patient data but failed to capture the cubic nature of the curve. With that in mind, the different error types compared to each other for this model are impressive as well:

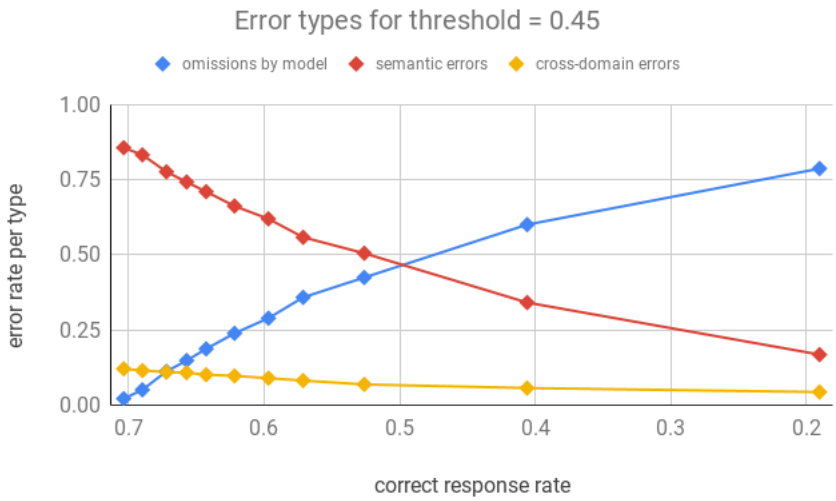


Figure 4: Comparison of the omission, semantic and cross-domain error rates created by the category-trained model for the omission threshold found to be optimal for the category-trained model

Looking back at the patient data for semantic errors in figure 1b, the category-trained model does take longer to produce more omission than semantic errors but is still able to correctly show the change in trends despite missing the cubic feature for the omission errors' curve (figure 4). The cross-domain error rate starts out as being more significant than it should be but decreases to less relevant levels over time.

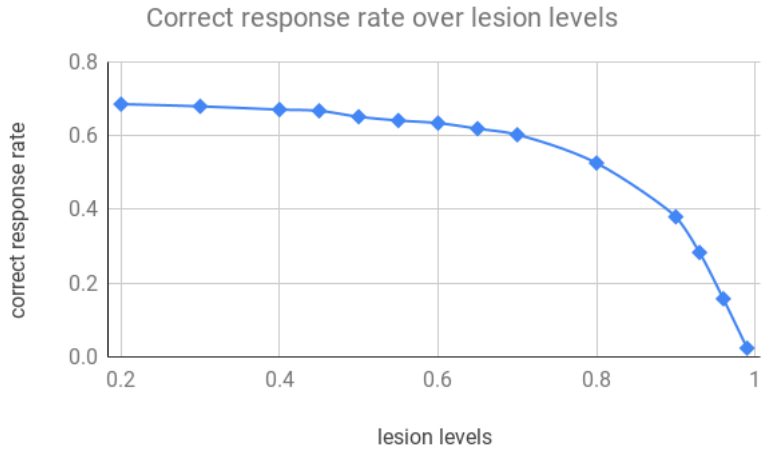


Figure 5: The correct response rate of the category-trained model over the increasing lesion levels without an omission threshold

Lastly, the model can be judged on its trend of the correct response rate over increasing lesion levels. As explained before, a negative monotonic relationship is assumed in patients but there has not been a precise quantification of the brain's damage levels for SD. Therefore, the development seen in figure 5 should be judged positively as the continuous decline of performance is correctly emulated.

Semantics-trained Model

Testing the semantics-trained model in the same manner yielded an omission threshold of 0.15 with a correlation coefficient of 0.558.

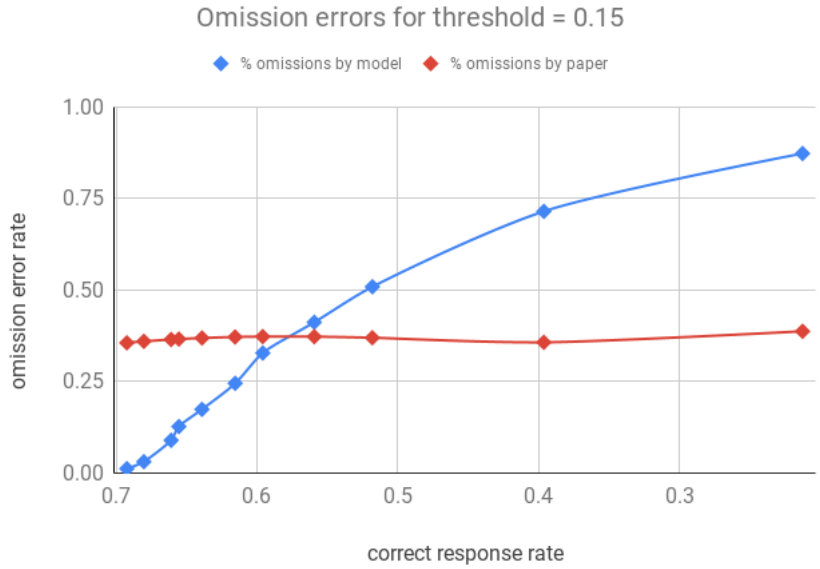


Figure 6: Comparison of the omission error rate created by the semantics-trained model and the omission error rate predicted by the 2008 paper for the omission threshold found to be optimal for the semantics-trained model

The semantics-trained model shows similar behaviour as the category-trained model but does so at a much lower omission threshold and correlation coefficient (figure 6).

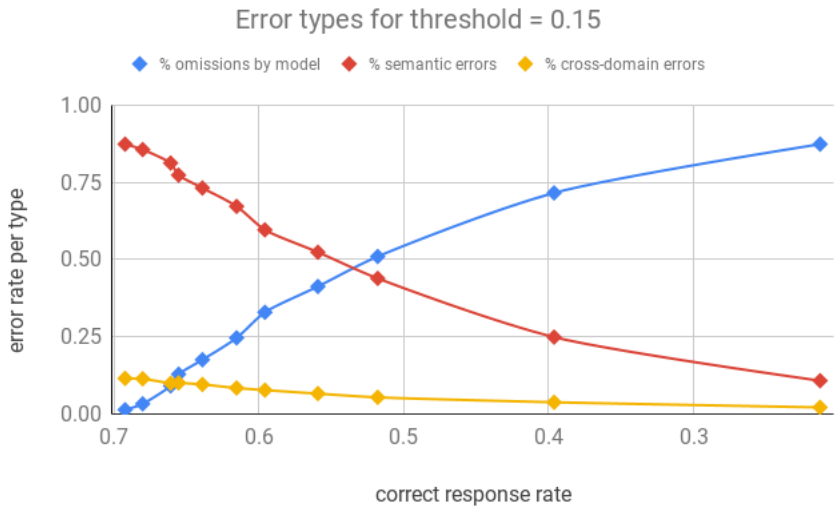


Figure 7: Comparison of the omission, semantic and cross-domain error rates created by the semantics-trained model for the omission threshold found to be optimal for the semantics-trained model

When comparing the different error types for the semantics-trained model (figure 7), it becomes evident that the omission error rate increased even steeper than for the category-trained model, which explains the lower correlation coefficient. The trend changes between the error types are quite similar to the other model, albeit with less linear curves. The semantic errors rate also decreases steeper than the other model, but the cross-domain error rate behaves very similarly.

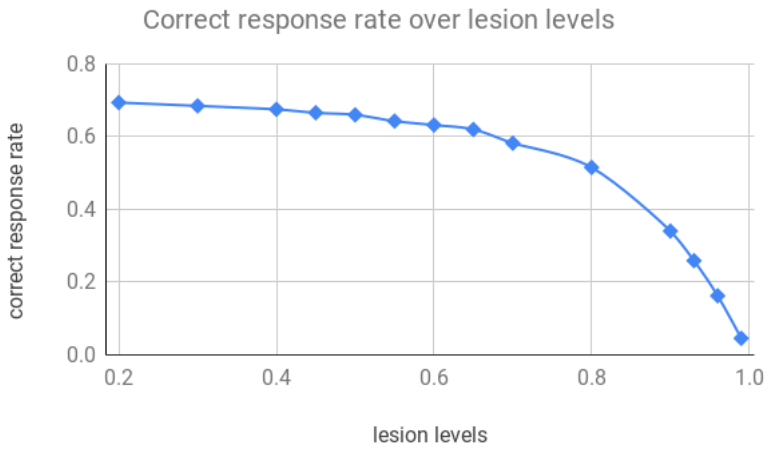


Figure 8: The correct response rate of the semantics-trained model over the increasing lesion levels without an omission threshold

The semantics-trained model also achieves a continuously decreasing performance curve (figure 8).

Conclusions

Based on the results collected for both models, we believe the test-bed has succeeded in evaluating a neural network in its neuromorphism. Both networks managed to replicate the behavioural patterns from SD patients with high accuracy, suggesting that the challenge posed by the test-bed is not unreasonably difficult. As for the strong similarity between the two models' results, the collected data is unable to provide a definite explanation. Nonetheless, the models achieved these results under very different omission thresholds, so it is unlikely that the test-bed is causing tested networks to converge on a single solution to its task.

Discussion

Model Comparison

Overall, the category-trained model surpassed our expectations in its neuromorphism both on its own and especially in comparison to the semantics-trained model. We had expected the latter to do better than the first since its training was focused on simulating a semantic objective instead of disconnected categories of data, but this was apparently not a deciding factor for our test-bed. However, whether that indeed means that such semantic objectives are a worse representation of the brain's mechanisms remains unclear since the semantics-trained model might have been disadvantaged by having its readout layer retrained to match *Ecoset*'s categories for evaluation purposes. Alternatively, our understanding of the internal representations of the models might be inaccurate or their preparatory training might have been inadequate in its extent and/or equality. Another possibility would be that visual and semantic features of an image overlap more than currently thought, meaning that visually based models like the category-trained model have an implicit understanding of semantic relations. Nevertheless, that should not be enough to surpass the semantic understanding of the semantics-trained model, which further suggests that the latter did not maximise its potential for such a task like picture naming.

Future Research

The questions raised previously could possibly be answered by constructing another test-bed that relies on a semantic objective readout layer for evaluation and thus disadvantaging the category-

trained model. Alternatively, the lesioning could be expanded in-depth instead of only 'damaging' the last layer, as it was done for this project. Such a study might gain further insight as to whether our choice of simulating SD was sufficiently realistic for our test-bed. Similarly, the omission threshold procedure could be improved by using the correlation coefficient as the loss function of an actual fitting process contrary to our testing process of a pre-determined range of values.

Healthy versus Impaired Patient Data

Since the results of this project showed a remarkable degree of neuromorphism for comparatively simple neural networks, it would be interesting to see how the models which rank highly on *Brain-Score* would fare on this test-bed. Another intriguing angle would be the relation between models which do well on the one data source or the other. Would high scores on healthy patient data correlate with high scores on impaired patient data or vice versa? Depending on those results, the importance of impaired patient data might require new consideration in the field of neuromorphic neural networks.

References

[1] Martin Schrimpf et al. *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* en. preprint. Neuroscience, Sept. 2018. DOI: 10.1101/407007. URL: http://biorxiv.org/lookup/doi/10.1101/407007 (visited on 09/26/2020).

[2] Samuel Ritter et al. "Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study". In: *arXiv:1706.08606 [cs, stat]* (June 2017). arXiv: 1706.08606. URL: http://arxiv.org/abs/1706.08606 (visited on 12/18/2020).

[3] Timothy T. Rogers et al. "Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity". en. In: *Neuropsychologia* 76 (Sept. 2015), pp. 220–239. ISSN: 00283932. DOI: 10.1016/j.neuropsychologia.2015.04.015. URL: https://linkinghub.elsevier.com/retrieve/pii/S002839321500161X (visited on 09/19/2020).

[4] Timothy T. Rogers et al. "Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation." en. In: *Psychological Review* 111.1 (2004), pp. 205–235. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.111.1.205. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.111.1.205 (visited on 09/19/2020).

[5] Grace W. Lindsay. "Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future". en. In: *Journal of Cognitive Neuroscience* (Feb. 2020), pp. 1–15. ISSN: 0898-929X, 1530-8898. DOI: 10.1162/jocn_a_01544. URL: https://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_01544 (visited on 09/19/2020).

[6] Anna M. Woollams et al. "Anomia: A doubly typical signature of semantic dementia". en. In: *Neuropsychologia* 46.10 (Aug. 2008), pp. 2503–2514. ISSN: 00283932. DOI: 10.1016/j.neuropsychologia.2008.04.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0028393208001280 (visited on 09/19/2020).

[7] Mehrer, J., Kietzmann, T. C., & Kriegeskorte, N. "Deep neural networks trained on ecologically relevant categories better explain human IT". In: *Conference on Cognitive Computational Neuroscience. New York, NY, USA* (2017).

[8] Ankit Rohatgi. *Webplotdigitizer: Version 4.4*. 2020. URL: https://automeris.io/WebPlotDigitizer/ (visited on 01/25/2021).