Bachelor thesis

Artificial Intelligence

Radboud University

A semantic dementia-based testbed for assessing how well lesioned neural networks replicate patient behavioural data

Author:
Name : Eline Michèlle Braun
Student number : s1010232

First supervisor:
dr. Tim .C. Kietzmann
Donders Institute
t.kietzmann@donders.ru.nl

Second supervisor:
Émer. C. Jones
visiting PhD student at
the Donders Institute
e.jones@psych.ru.nl

January 29, 2021

Abstract

We introduce a way to compare neural networks on their mechanistic similarity to the brain by investigating how they perform when impaired. These networks are lesioned such that they simulate semantic dementia (SD). Two models are used for comparison: one is trained with a category objective and another with a semantics objective. Both models are used to perform the word-picture matching task, a well known task in the research and clinical assessment of SD (Rogers et al., 2004) (Rogers, Ralph, Patterson, & Jefferies, 2015). The results show that both models produce sensible results and that we have created a proof of concept for a semantic dementia based testbed. However, there is no clear answer on which of the two models performs more like SD patients: both models produce behavioural features similar to those of SD patients.

Contents

Chapter 1

Introduction

Many computational neuroscience labs are striving to create neural networks that form similar representations to the brain when performing cognitive human tasks. When the behaviour of these networks approximate brain-behaviour, the assumption can be made that the network's mechanistic structure may also be brain-like. However, this does not need to be a causal relation.

The Brain-Score platform is a benchmark that compares a model's representation to primate ventral stream neuroimaging activations and behavioural data to see how they correlate in measures of 'brain-likeness' (Schrimpf et al., 2018).

It has been shown that with the use of deep convolutional neural networks (DCNN's), functional signatures of primate visual processing can be predicted across multiple hierarchical levels at unprecedented accuracy (Kietzmann, McClure, & Kriegeskorte, 2018). Based on this, we may assume that DCNNs have a good chance to perform well on the Brain-Score platform.

A point of discussion is the fact that Brain-Score examines the submitted networks only on how similar they are to healthy brain data. However this data is not representative to all brains as not all brains are 'healthy'. Another level of similarity is to be similar to an impaired brain, showing it 'breaks down' similarly to the brain. It would be interesting to look at the neural networks ranking high on Brain-Score and see how they would perform when the networks are impaired by a brain deficit. Would the responses still be brain-like? If networks can satisfy both levels of similarity we can speak of a greater mechanistic similarity to the brain, than networks that only satisfy one level of similarity.

What we try to achieve in this paper is creating an additional testbed to be able to compare the mechanistic similarity of deep convolutional neural

networks (DCNNs) to the brain when we impair these networks. This way neural networks can be assessed on their similarity to the brain regardless of its condition.

We have chosen to use semantic dementia patient data and implement SD-like atrophy because it is uniquely well-suited for this testbed. The atrophy of SD only affects a specific area of the brain: the ventral anterior temporal lobe. This area lies in the ventral stream and as that system is what DCNNs approximate at high level, it enables us to actually implement this brain deficit. Additionally the atrophy progresses in a similar way for different patients and the patients show specific and selective behavioural impairment. Test results among patients performing on SD tasks are very robust (Rogers et al., 2004) (Rogers et al., 2015). The selective lesioning combined with robust resulting behaviour, forms a good testbed for networks to determine their resemblance to the human brain.

To verify that mechanistic similarity of a DCNN to the brain can indicate whether the performance is more brain-like, we fed two differently trained models into the test-bed. The first model is trained with a category objective while the second model is trained with a semantics objective. We will investigate whether the category-trained model or the semantics-trained model performs more similarly to SD patients when damaged with SD-like atrophy. We hypothesize that the semantics-trained model will perform more similarly to SD patients since the brain represents information by means of semantics rather than categories. It is very unlikely that categorisation is the ventral stream's objective, semantic relationships are also important.

In the following sections I will discuss relevant research which is the base of the project that we are building upon. After that, the research methods will be described. This will entail a way to reach different stages of SD atrophy in a DCNN and carrying out a commonly seen task in the research of SD. This task will be performed by the two models while impaired by the different stages of SD. We will discuss the performance of those models in the results and their actual similarity to the brain in the conclusion. At the end we will discuss whether this testbed was the best way to simulate SD lesioning and what the next step would be to further this research.

Chapter 2

Previous Work

Semantic dementia

Semantic dementia is a disease that affects a specific part in the brain that underlies semantic memory. Selective impairment of semantic memory causes severe anomia which is one of the core features of SD, as well as impaired spoken and written single-word comprehension (J. Hodges, Patterson, Oxbury, & Funnell, 1992).

Damage of the brain caused by SD is very local and similar across patients. Most atrophy is seen in the ventral anterior temporal lobe which is thought to be the end of the ventral stream (Rogers et al., 2004).

Damage starts in the left temporal lobe and as the disease progresses, the severity of atrophy increases and the right temporal lobe will 'catch up' by also experiencing atrophy (J. Hodges & Garrard, 2000). Thereafter, there is evidence the disease spreads along the inferior and middle temporal gyri posteriorly towards the occipital lobe. However, there is no evidence of atrophy in the occipital lobe, this area stays intact (J. Hodges & Garrard, 2000) (J. R. Hodges, Graham, & Patterson, 1995).

Word-picture matching

The word-picture matching task frequently used task to assess semantic memory, is impaired for people suffering from semantic dementia. The results of performing this task is robust among patients depending on the degree of their semantic dementia diagnoses. Therefor, based on their results the severity of their disease can be suggested.

The word-picture matching task performed by people works as follows: a word is presented followed by 2 images. One image actually corresponds to the presented word and the other image is a distractor image. The participants should choose the image that matches the presented word. This is a two-alternative forced choice (2AFC) task, which means that chance level is 50%. It has been shown that SD patients perform worse when the

images are semantically closely related (Rogers et al., 2004) (Rogers et al., 2015). There are two papers that looked into this by testing with visual distractors that varied in there semantic distance to the target item (Rogers et al., 2004) (Rogers et al., 2015). In the range of relatedness to the target word, the papers used increasing distances for the distractors. The patients of the 2004 Rogers paper were tested with four levels of distractors: a close, dissimilar, distant and unrelated distractor. They chose to discard the dissimilar distractor for their research purposes so the results of the patients that are showed also discards this distractor level.

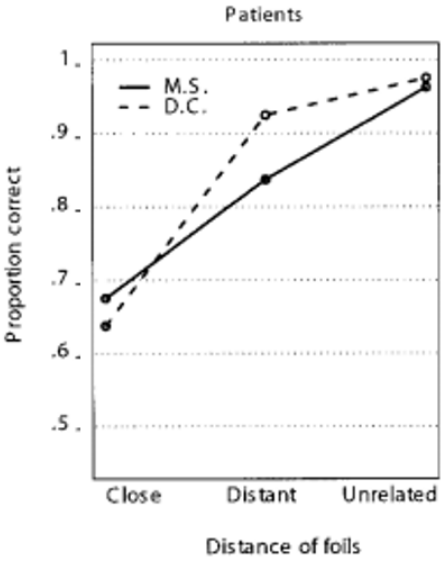
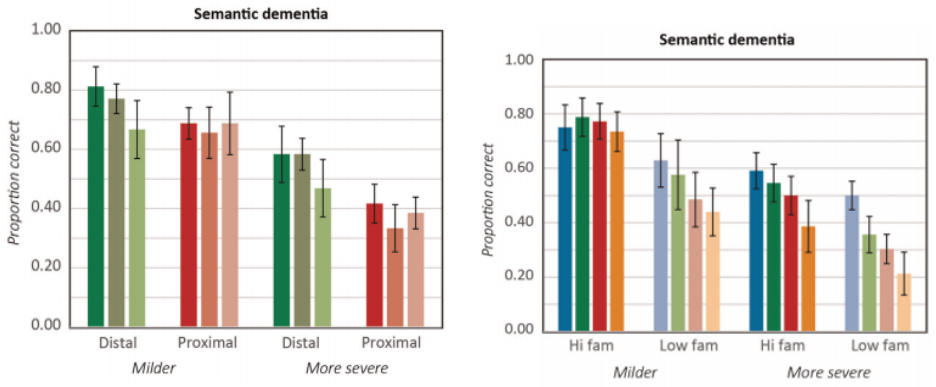


Figure 2.1: Results of the two patients, diagnosed with a severe case of semantic dementia, performing the word-picture matching task in the 2004 Rogers paper

The patients in figure 2.1 suffer from a severe case of semantic dementia so it is not surprising that the performance on the stimuli with close distractors is quite low. However, the performance is still above chance level which is 50%. This indicates that not all semantic memory is lost. The performance of the distant distractor variant is already less impaired and the performance seen for the unrelated distractor cases are nearly 100 percent which again indicates that there is a level of semantic memory still apparent. Even severe SD patients do not mix up items from different semantic domains, for example they do not mix up animate or inanimate categories.

The 2015 Rogers paper uses six distractor levels varying in typicality between the items and eight distractor levels varying in familiarity between

the items. Their study consists of fourteen patients suffering from mild to more severe cases of semantic dementia.



(a) 2015.1 (b) 2015.2

Figure 2.2: Results of the fourteen patients performing the word-picture matching task in the 2015 Rogers paper

In the left plot of figure 2.2, the distal bars represent far distractors and the proximal bars correspond to close distractors. The three bars per cluster represent more typical words to less typical words, shown by the use of darker colors for more typical words. For both severities of semantic dementia the far distractor stimuli perform better than the close distractor stimuli. For the more severe case, we observe a performance worse than in the mild case and the slope between the two distractor levels is steeper for the more severe case than in the mild case.

In the right plot, we will focus on the high familiarity clusters as the low familiarity clusters are not playing a part in our research. The blue and the green bar represent far distractors, while the red and yellow represent close distractors. The overall performance of the mild patients is higher than for the more severe patients. Here the difference between the performance of far and close distractors for more severe patients also differs more than in the mild patients, like we observed in the left plot and in the 2004 Rogers patients.

Both the 2004 and the 2015 Roger papers are relevant for our research. They show us the behaviour seen in semantic dementia patients performing the word-picture matching task with differing distractor distances. We can conclude from this data that the behaviour is robust across both papers. However, from now on we will focus on the 2004 paper, the reason for this is given in section 3.2.

Chapter 3

Research Methods

The goal of this project is to create an additional testbed to be able to compare the mechanistic similarity of DCNNs to the brain when impaired. To verify that mechanistic similarity of DCNNs to the brain indeed makes a difference in the behaviour that will arise, we are comparing the performance of a category-trained and a semantics-trained model when the models are damaged with SD-like atrophy.

The word-picture matching task is used to determine the performance of the models. Within this task we are going to follow the idea of changing the semantic distances of the distractors and look at the difference. This was only done by the two papers that we discussed before (Rogers et al., 2004) (Rogers et al., 2015).

The target and distractor set are selected from Ecoset which is comprised of the most concrete and frequent basic-level categories in the English language (Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, in press). To impair the models with SD-like atrophy we started with determining which lesion levels would correspond to different severities of semantic dementia. The model is impaired and then submitted to the free-naming task, where the model is classifying images that are fed in. The performance of this task is observed and used to determine different lesion levels. Two models are used as test cases and are fed into the testbed. Their performance will be compared to each other and to patient data. This is done to see whether the difference in training of the models has impact on the performance, it being similar to the performance of SD patients, of the word-picture matching task.

3.1 Lesioning

From many different neuroscientific papers there is evidence that the damage done by semantic dementia is localized to the ventral anterior temporal lobe

and this extends along the inferior temporal gyrus towards the occipital lobe (J. R. Hodges et al., 1995) (J. Hodges & Garrard, 2000).

This area is part of the inferior temporal cortex (IT) which is the location of final stage of the ventral stream. The IT is especially important in visual processing and visual object recognition. There is little evidence of SD-related atrophy seen beyond the temporal lobe, which leads to the apparent conclusion that earlier stages of the ventral stream such as V4, V2 and V1 which are located in the intact occipital lobe are not impaired. For the lesioning scheme to be biologically plausible we are focusing on lesioning the last part of the DCNN. This is for the reason that the network approximates a human ventral stream and the atrophy seen in SD patients is located in the final stage of this system. We decided to start by impairing the very last layer of the networks.

Now that we know what to lesion, a lesioning scheme has to be constructed. The free-naming task is the most direct measure of behavioural impairment to extract sensible lesion levels from. This task is performed by feeding images to the network and classifying them to the corresponding category, which is simply basic classification. This is done for multiple lesion levels on the last layer to see what kind of effect each level has on the behaviour of the network. This way several lesioning levels are chosen that can resemble different degrees of the SD disease.

3.1.1 Implementation

Lesioning the last layer is done by probabilistically dropping out a certain percentage of its nodes and studying the performance of the free-naming task for every lesioning level.

The models were given 45 held-out test images per category to classify. The network classifies by selecting the image which has the highest activation for the target category. For every category the classification of 45 images was done 5 times to get a robust outcome due to the probabilistic nature of dropout .

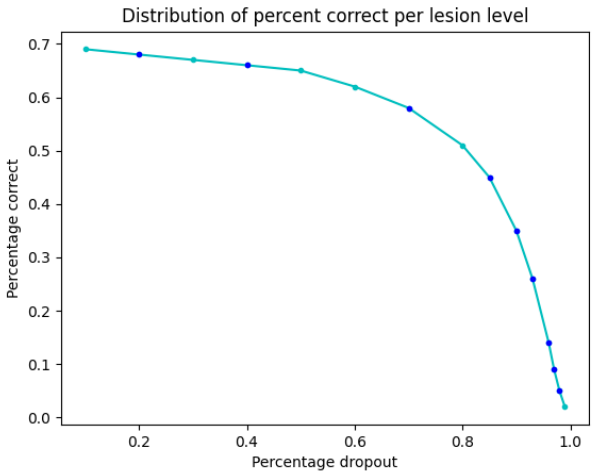


Figure 3.1: Free-naming task performed by the category-trained model at different lesion levels. Blue dots correspond to chosen lesion levels

Table 3.1: Performance of the free-naming task per lesion level. Blue rows are the chosen lesion levels.

Dropout %	% correct
0.1	69%
0.2	68%
0.3	67%
0.4	66%
0.5	65%
0.6	62%
0.7	58%
0.8	51%
0.85	45%
0.9	35%
0.93	26%
0.96	14%
0.97	9%
0.98	5%
0.99	2%

The networks we are testing were trained with 20 percent of dropout to avoid overfitting, so this level of dropout represents the 'unlesioned' performance. From 60 percent dropout onward the accuracy drops in bigger steps. To get an even distribution of steps in the distribution we included several dropout levels between 80 and 100.
From this distribution, lesion levels are chosen that sample roughly uniformly across accuracy: 0.2, 0.4, 0.7, 0.85, 0.9, 0.93, 0.96, 0.97 and 0.98. These levels of dropout are used in the word-picture matching. The level of 0.2 is included to act as a baseline as this dropout level was used when training the network.

3.2 Ecoset

The categories that will be used in the word-picture matching pipeline are carefully thought of. We agreed on 60 different categories to use from Ecoset

which is an ecologically more relevant set of image categories then ImageNet, these categories are displayed on figure 3.2 (Deng et al., 2009). It has been shown that an ecologically more relevant visual diet leads to significantly improved similarities in a DCNN to response properties in human inferior temporal cortex (IT) (Kietzmann et al., 2018).

The 60 chosen categories are partly constructed from categories used in the 2004 Rogers paper. The categories from the paper that overlap with the 565 categories of Ecoset were chosen. To fill up to 60, additional categories are chosen from Ecoset that made sure that the domains and subcategories had the same number of categories. The dataset is split into two domains: animate and inanimate categories. The categories from both domains are themselves divided into three subcategories and these consist of ten categories each.

The 2015 Rogers paper is left aside here, since they used more specific categories than were included in Ecoset. For this reason we decided to focus on the categories presented by the 2004 Rogers paper, this does not imply that the 2015 Rogers paper is not relevant. The results of both papers show that the behaviour of patients are robust.

Animate		
Land	*Water*	*Insect*
Squirrel	Whale	Ant
Horse	Dolphin	Bee
Monkey	Fish	Beetle
Rabbit	Jellyfish	Butterfly
Mouse	Octopus	Caterpillar
Elephant	Shrimp	Cockroach
Deer	Lobster	Grasshopper
Tiger	Starfish	Mantis
Lion	Tadpole	Mosquito
Rhino	Crawfish	Moth
Inanimate		
Household objects	*Vehicles*	*Musical instruments*
Refrigerator	Airplane	Clarinet
Kettle	Bicycle	Cymbals
Lamp	Boat	Drum
Toaster	Bus	Guitar
Vase	Car	Kazoo
Phone	Helicopter	Mandolin
Table	Motorcycle	Piano
Chair	Ship	Ukelele
Stove	Train	Violin
Sink	Truck	Bugle

Figure 3.2: A selection of Ecoset that is used for the word-picture matching task

For each of those categories a close, middle and far distractor are chosen from within these 60 categories to create a dataset with distractors from varying semantic distance to the target. The close distractor is chosen from the same subcategory as the target. The middle distractor is chosen from another subcategory while still staying within the same domain. The far distractor is chosen from a subcategory of the other domain. As an example from our dataset, I highlighted a target (red) and its close (orange), middle (yellow) and far distractor (blue).

All distractors are chosen in a way that there is little to no ambiguity to whether it is a close, middle or far distractor. We verified this by having 5 people rank the distractors per target word. The distractors were presented next to the target word in random order such that the people had to determine which of the distractors were in their opinion the closest, medium and furthest away related to the target without getting hints from the document. The average distance for every distractor per target was chosen in the way we had created it. The distractors were placed of increasing semantic distance to the target. The results of our small inquiry, which includes the averaged number of placing for every distractor and their standard deviation, are included in Appendix A.2.

The full distractor dataset is included in Appendix A.1.

3.3 Word-picture matching task

The task is implemented as follows: two images are fed into a specific neural network and the activation of node corresponding to the the target category for each image is compared. The image with the highest activation for the target category is selected as the chosen match of the word target the picture. The implementation will be discussed in more detail in section 3.3.1.

The 2004 paper uses line drawings and the 2015 paper uses images, is this going to affect our results when comparing behaviour? The behaviour of the patients in the 2004 and the 2015 Rogers papers are robust despite other methods of input. In addition, we look upon a 2008 paper that observed the performance of the picture naming task done by semantic dementia patients. The input stimuli they used in their research consisted of both line drawings and images and they showed robustness across stimulus types. The correlation between the scores of these two input types was near-perfect, $\rho = 0.99$ (Woollams, Cooper-Pye, Hodges, & Patterson, 2008).

3.3.1 Implementation

For each target category, the target and distractor image are inputted to the model. The image that elicits the highest value in the readout node corresponds to the target image is taken to be the model's selection of the image. This is done for the 180 pairs of target and distractors in our distractor dataset. We distinguish the results based on the semantic distance of the distractor that was used. There are 60 pairs per distractor level and the performance of all these pairs are averaged. The whole distractor set is run 5 times, for robustness due to probabilistic dropout, to attain the overall average outcome.
This task is then performed by the network for every lesioning level that

was chosen. The results for close, middle and far distractors are calculated for every lesioning level.

To show the variability of the performance for every distractor level, we calculated the standard error or uncertainty per level. These bars show how far from our averaged out value, the true value might lie. The standard error is calculated over the different distractor pairs in one distractor level, after averaging the multiple runs and images used for the pairs. The standard error indicates the variability over different close, middle or far pairs.

3.4 Two models

Both models consist of a vNet architecture (Mehrer et al., in press). This architecture consists of 10 layers where the sizes of convolutional kernels in each layer approximates the biological foveal receptive field sizes from early to late regions of the visual system.

The first network is a model with a category objective trained on Ecoset categories (Mehrer et al., in press).

The second network is a model with a semantic objective trained on 300-dimensional fast-text word embedding vectors corresponding to Ecoset categories. Words with more relation to each other than with other words have more similar word embedding vectors. Semantically similar words are closer in this 'semantic space' than words with less similar meaning.

The last layer of the network is retrained with nodes corresponding to the Ecoset categories such such that it can be inputted to the Ecoset-based testbed pipeline.

I hypothesize that between the two differently trained models, the model trained with a semantics objective will perform more similarly to SD patients when damaged with SD-like atrophy because its representations are constrained to group semantically-related objects. Therefore, the semantics model's objective enforces that similar objects (whose word embedding vectors are similar) should be positioned closeby in the model's representation space. This means that *horse* and *bear* will be close together in the model's 'semantic space' because of similar vectors, whereas *chair* will be grouped far away from both *horse* and *bear*. By lesioning this system, closely related categories will likely be mixed-up whereas mix-ups between remotely related categories such as *horse* and *chair* are less likely. In the category-trained model however, this can still be the case. This model does not have a form of semantic representation for its categories. The categories will be randomly placed, in this way *horse* can be represented next to *chair*. When lesioned,

the model will likely confuse categories within and between domains.

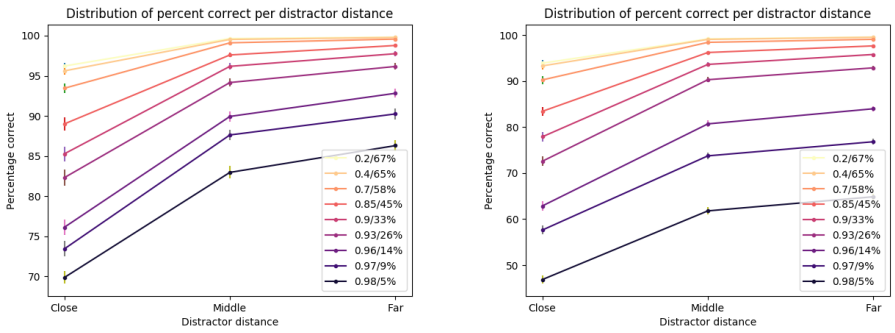
This suggests that when lesioned, the semantic model will likely confuse objects from the same domain, while rarely or never confusing objects from different semantic domains, similarly to SD patients. Whereas the category model will likely confuse objects from both same and different semantic domains, not similarly to SD patients

Chapter 4

Results

Two models, one category-trained and one semantics-trained, performed the word-picture matching task while being lesioned on different levels. The task uses three different distractors for every target with varied semantic distance. The performances of the models are compared to behavioural data from semantic dementia patients at different stages of disease progression presented in the two papers that were also using different distractor distances (Rogers et al., 2004) (Rogers et al., 2015).

4.1 The two models



(a) Performance of category-trained model (b) Performance of semantics-trained model

Figure 4.1: Results of the two networks performing the word-picture matching task. Left is the category-trained model and right is the semantics-trained model. Chance level is at 50% as the task is choosing between two images

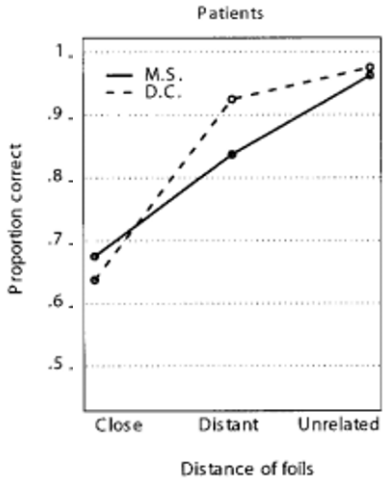


Figure 4.2: Behaviour of the two patients in the 2004 Rogers paper performing the word-picture matching task

Figure 4.2 displays the behavioural data of the patients M.S and D.C on the word-picture matching task from the 2004 Rogers paper (Rogers et al., 2004). These patients are diagnosed with a severe case of semantic dementia. The plot uses distant and unrelated distractors where we call them middle and far distractors.

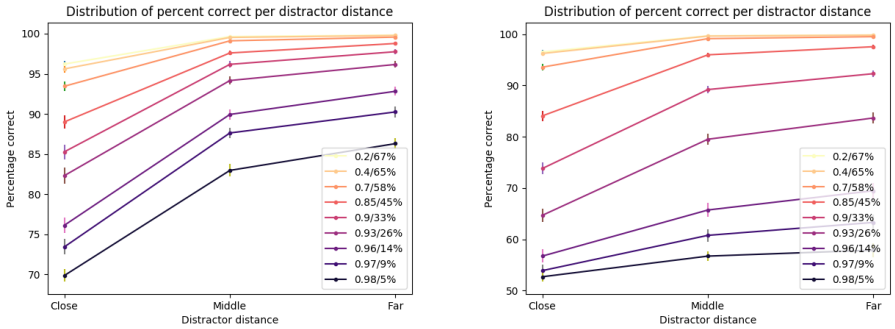
Figure 4.1 displays the performance of the two different models on our testbed. The color of the lines grow colder as the lesioning level increases. The legend shows the correspondence between the line color and lesioning level. For every lesion level the legend states the corresponding performance of the free-naming task impaired by the lesion level. This gives an idea of what every lesion level actually means or what kind of effect is has.

Figure 4.1(a) displays the performance of the category-trained model in our testbed. There is a nice spread of performance between every lesion-level. The drop in performance seems to be gradual in reference to the increase of lesioning rate. We observe a steeper slope in general between close and middle distractors, whereas the slope between middle and far distractors is much flatter. The error-bars show that the variance of the different category pairs within 1 lesioning level is small. The biggest variance is observed within close distractors.
The 0.97 and the 0.98 lesion levels have a similar performance as the patients M.S and D.C for the close and middle distractor stimuli. However, only low lesion levels show a similar behaviour for the far distractor stimuli. There is no direct correspondance of the performance from the category-trained model to patients M.S. and D.C.

Figure 4.1(b) displays the performance of the semantics-trained model in our testbed. The overall performance is seen along a greater performance spread than seen in the category-trained plot. The highest lesion level, 0.98, is performing around chance level for the close distractor stimuli. This means that the network lost the ability to distinguish closely related images when its last layer lost 98% of its nodes. Although the slope seems flatter between close and middle distractors, the difference in performance between the two distractor level is bridging the same amount as the one observed in the category-trained model plot. What is striking is the gaps between the 93/96 and the 97/98 % lesion-levels. This implies the possibility for more differentiated behaviour between those lesion-levels. The error-bars show that the variance within each distractor stimuli is very small. The averaged performance is thus very robust.
The performance portrayed by the 93% lesion-level is quite similar to the patients M.S. and D.C.. The close and middle distractor performances are close to the one observed in the patients. The far distractor performance is getting close to the behavioural data, closer than we observed for similar lines in the category-trained model.

4.2 First-layer lesioning control

We include some controls for our testbed in order to determine whether our models are performing sensibly.
The first control is switching the lesioning from the last to the first layer. This way we can observe if lesioning the last layer indeed simulates semantic dementia-like impairment. The behaviour produced by lesioning the first layer should be distinct from the behaviour seen in the models that are lesioned in the last layer. The first layer in our DCNN can correspond to early visual stages like v1 that plays a role in detecting simple visual features. Impairing this area can correspond to pattern of stroke in the early visual system. A stroke in the inferior temporal cortex (IT) would not result in the same behaviour for example.

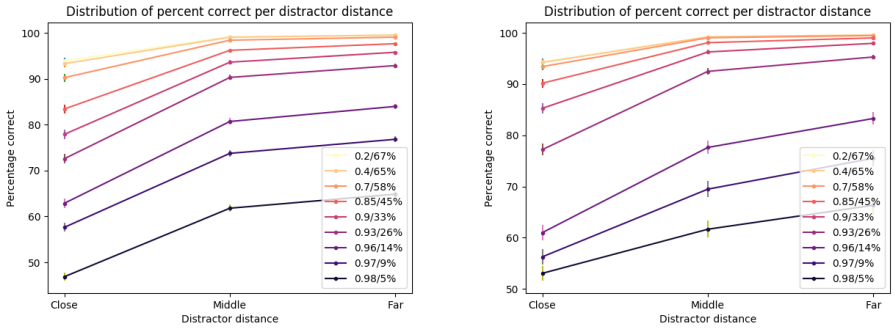


(a) Last layer lesioned

(b) First layer lesioned

Figure 4.3: Category-trained model lesioning in different layers

Figure 4.3 displays the performance of the category-trained model when its last layer is lesioned (a) and when its first layer is lesioned (b). The lines of plot b are on average much flatter than in a. The average performance of the first layer lesioned plot is also much lower than in the last layer lesioned plot. There are however, some lines that display a change in slope between the close and the middle distractor. This can be explained by the fact that abstract features, created in later layers, are still learned. Close distractors can have the same or similar abstract features which can hinder the network in choosing the right image.



(a) Last layer lesioned

(b) First layer lesioned

Figure 4.4: Semantics-trained model lesioning in different layers

Figure 4.4 displays the last (a) vs first (b) layer lesioned for the semantics-trained model. Here the two plots are more similar to each other than were the two plots for the category-trained model. The lines for the lower lesion-levels are somewhat flatter and clustered together. The higher lesion-levels do have a steeper slope between the different distractor types.

4.3 Random network control

This control is used to determine whether the learned representations or weights of our base model is causing the behaviour we see. We also try to determine that the retrained readout is not causing the behaviour. For this control we used a model with the same architecture as was used for our two other models, but this one has random weights. The last layer of this network was retrained with an Ecoset readout, like for the other models.

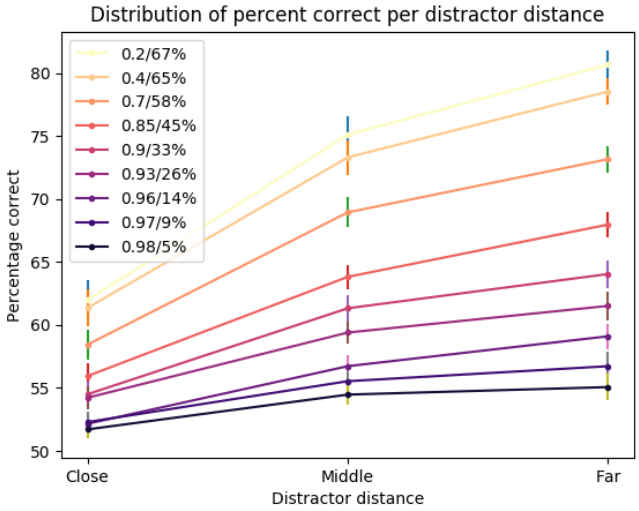


Figure 4.5: Random network retrained with an Ecoset readout with the last layer lesioned.

Figure 4.5 displays the random network performance on the word-picture naming task. The model is performing on average just above chance level. The performance of the higher lesion-levels does not change much between the distractor types. However, lower lesion-levels show some increase in performance especially for middle and far distractor stimuli.
The error-bars show that there is some variance happening for every distractor type results.

Chapter 5

Discussion

Two models, one category-trained and one semantics-trained, performed the word-picture matching task while being lesioned on different levels. The task uses three different distractors for every target, varying the semantic distance. The performances of the models are being compared to behavioural data from semantic dementia patients at different stages of disease progression.

The performance of the models are also compared to the performance of the models when impaired in the first layer rather than the last layer. This way we can observe if lesioning the last layer indeed simulates semantic dementia-like impairment. At last a random network with a retrained last layer on a Ecoset readout is compared to the performance of the models to determine whether our base-model is acting sensible.

5.1 Category vs semantics model

To go back to the question we are researching, does the category-trained model or the semantics-trained model perform more similarly to semantic dementia patients when damaged with SD-like atrophy?

There is no clear-cut answer for this question. Neither model has produced an output that correlates precisely to the behavioural data seen from the patients M.S. and D.C. (Rogers et al., 2004). The lesion levels of 0.96 and 0.97 for the category-trained model and the 0.93 lesion level for the semantics-trained model resemble the observed behaviour the most.
The high lesion levels for the category-trained model resemble the close and middle distractor performance well. The slope is steep between these two points. However, the slope between the close and far distractor stimuli is much flatter: a near to perfect performance is expected for the far distractor stimuli.

The 0.93 lesion level for the semantics-trained model is also close to the behaviour of the patients. There is a steep slope between the close and middle distractors stimuli and the performance of the far distractor is also getting higher towards the 100%, still not getting there completely. Also the performance of higher lesion levels in the semantics model is lower than we observed in the patients especially the performance on close distractor stimuli is around chance level for the highest lesion levels. This is not something that is observed by the patients M.S. and D.C. who suffer from a severe case of semantic dementia.

We expected the category-trained model to perform worse than it actually did. We hypothesized that the semantics model would perform better because it consists of semantic representations, where we did not anticipate the category model to have any. However, the category-trained model appears to have some form of semantic representation because otherwise we would have observed flatter lines across the distractor types as the category would not care about semantic relatedness. The semantic features are possibly a result of similar visual features. For instance, sea creatures are mostly depicted in an aquatic surrounding. The fact that these creatures have similar visual features may result in the clustering of these creatures within the model trained with a category objective.

The model trained with a semantics objective performed worse than we expected. The lesion levels used for the category-trained model seems to affect the performance of the semantics-trained model more. For the 0.98 lesion level the model performed around chance level for close distractor stimuli and as we have seen from the behaviour of very severe SD patients, this is not what we see in behavioural data.
This can be the result of us choosing the different lesion levels based on the performance of the free-naming task run by the category-trained model. The chosen lesion levels might not correspond exactly to a uniformly spread across the performance when we had run the free-naming task in the semantics-trained model. We observe in the results of the semantics-trained model that there is quite a gap between the performance on the 0.93 and the 0.96 lesion levels, which might indeed indicate that there is room for more different behaviour. The semantics-trained model might be more fragile than the category-trained model.
Another explanation for the semantics model to perform worse than expected would be the training time it had. Both category- and semantics-trained model were trained for the same number of epochs. The performance of the semantic model was nearing to its plateau. However, the model is not trained beyond this number of epochs. We cannot be certain of whether it actually reached its top point in behaviour. This also raises the question of when to cut-off the training part. Assuming that the semantics-trained

model needed more time to train, when is a fair cut-off point? Both models must be able to learn a same amount to compare them in a fair way.

5.2 Controls

First layer vs last layer lesioning

The first layer lesioning showed less similar behaviour to SD-like behaviour than the last layer lesioning as expected. The lines across the different distractor types is much flatter than observed in the last layer lesioning performances. The overall behaviour also dropped in the first layer lesioning, this is sensible since recognizing basic features is impaired. However, we do see similar features of the performance on the first layer lesioning compared to the last layer lesioning. Their behaviour is not completely dissimilar.
Overall we can conclude that lesioning the last layer produces more similar behaviour to semantic dementia patients than the first layer lesioning. Therefor lesioning the last layer indeed simulates SD-like atrophy best.

Random network

The performance of the randomly initialized network, of which last layer is retrained on an Ecoset readout, is very dissimilar to the performance of both category- and semantics-trained model. The overall average of the lines is quite flat across the distractor types. The performance on the close distractor stimuli are around chance level. Some lower lesion levels show increase in performance especially for midlel and far distractor stimuli. There is a possibility that the retrained readout has managed to find some combination of the random weights that is useful for animate vs inanimate. However, the performance of the random network is not well-enough to match the SD patient behaviour.
Hereby we can conclude that the readout is not behind all behaviour, but rather the trained weights and the learned representations.

5.3 Future steps

As both models did not produce the exact similar behaviour to the patients M.S. and D.C., further research may include different lesioning schemes and room for development of more brain-like models. We matched the whole process happening in the inferior temporal cortex, also known as the last part of the ventral stream, to the very last layer of the model. To simulate SD-like atrophy, further research may look into lesioning more layers simultaneously and for different percentages of dropout in the different layers.

Chapter 6

Conclusion

Our results show that both models produce sensible results and that we have created a proof of concept for a semantic dementia based testbed. Neither model is performing better or similar to SD-impaired patients.

The performance of the models is roughly similar to that of SD patients. We observe correlation of high lesion levels to the behaviour of the severe patients. The performance of close distractor stimuli are always worse than of the middle distractor stimuli and there is an increase in performance between the middle distractor stimuli to the far distractor stimuli. However, we had expected the performance on the far distractors to be near to perfection. As a consequence this means that neither model fully displayed SD-like behaviour, because even severe SD patients almost never mix-up in target and distractors from different domains.

Hereby we reject our hypothesis of the semantics-trained model performing better because of its representations being constrained to group semantically-related objects. Both models seem to perform similarly which leads us to believe that the category-trained model also shows to have some form of semantic knowledge. There is no clear answer to conclude which of the two models is performing more like semantic dementia patients when impaired by SD-like atrophy.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. , 248–255.

Hodges, J., & Garrard, P. (2000). Semantic dementia: clinical, radiological and pathological perspectives. *J Neurol 247*, 409–422.

Hodges, J., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: progressive fluent aphasia with temporal lobe atrophy. *Brain 115*, 1783–1806.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, *3*, 463–496.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (in press). An ecologically motivated image dataset for deep learning yields better models of human vision.

Rogers, T. T., Garrard, P., McClelland, J. L., Ralph, M. A. L., Bozeat, S., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychology Review*, *111*, 205–235.

Rogers, T. T., Ralph, M. A. L., Patterson, K., & Jefferies, E. (2015). Disorders of representation and control in semantic cognition: Effects on familiarity, typicality, and specificity. *Neuropsychologia*, *76*, 220–239.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

Woollams, A. M., Cooper-Pye, E., Hodges, J. R., & Patterson, K. (2008). Anomia: A doubly typical signature of semantic dementia. *Neuropsychologia*, *46*, 2503–2514.

Appendix A

Appendix

A.1 Dataset distractors

Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Squirrel	Mouse	Jellyfish	Motorcycle
Horse	Deer	Dolphin	cymbals
Monkey	Squirrel	Octopus	sink
Rabbit	Mouse	Grasshopper	helicopter
Mouse	Rabbit	Mantis	violin
Elephant	Rhino	Ant	chair
Deer	Horse	Lobster	Vase
Tiger	Lion	Bee	Ship
Lion	Tiger	mosquito	Kettle
Rhino	Elephant	Moth	Mandolin
Target word	**Close distractor**	**Middle distractor**	**Far distractor**
whale	fish	Mouse	Lamp
dolphin	fish	Squirrel	ukelele
fish	whale	grasshopper	Airplane
jellyfish	Octopus	Lion	Truck
octopus	Jellyfish	Monkey	phone
shrimp	tadpole	horse	Table
lobster	tadpole	deer	bicycle
starfish	jellyfish	Butterfly	Refrigerator
tadpole	shrimp	Rhino	Stove
Crawfish	fish	mantis	chair

Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Ant	Cockroach	Whale	Guitar
Bee	Mosquito	Rabbit	Car
Beetle	Grasshopper	Fish	Piano
Butterfly	Moth	Starfish	Clarinet
Caterpillar	Grasshopper	tiger	drum
Cockroach	Ant	Elephant	toaster
Grashopper	Mantis	Shrimp	Boat
Mantis	Mosquito	Deer	Bus
Mosquito	Mantis	tadpole	Train
moth	Mosquito	Crawfish	Kazoo
Target word	**Close distractor**	**Middle distractor**	**Far distractor**
refrigerator	stove	Piano	Ant
kettle	vase	Car	mantis
lamp	Vase	boat	Lion
toaster	Stove	Bicycle	Dolphin
vase	kettle	kazoo	deer
phone	Toaster	guitar	jellyfish
table	Stove	ship	Mosquito
chair	table	cymbals	butterfly
stove	refrigerator	violin	squirrel
sink	Toaster	clarinet	caterpillar
Target word	**Close distractor**	**Middle distractor**	**Far distractor**
airplane	helicopter	Chair	monkey
bicycle	motorcycle	refrigerator	lobster
boat	ship	toaster	bee
bus	truck	phone	starfish
car	bus	mandolin	octopus
helicopter	airplane	vase	mouse
motorcycle	bicycle	ukelele	shrimp
ship	boat	drum	horse
train	truck	kettle	crawfish
truck	car	lamp	moth

Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Clarinet	Kazoo	table	jellyfish
Cymbals	Ukelele	Airplane	whale
drum	Piano	train	fish
guitar	Piano	sink	rhino
kazoo	clarinet	motorcycle	tiger
mandolin	ukelele	stove	grasshopper
piano	clarinet	helicopter	butterfly
ukelele	guitar	truck	tadpole
violin	mandolin	bus	beetle
bugle	clarinet	Chair	Rabbit

A.2 Survey results

Target word	Close distractor	Middle distractor	Far distractor
Squirrel	1 ±0	2 ±0	3 ±0
Horse	1 ±0	2 ±0	3 ±0
Monkey	1 ±0	2 ±0	3 ±0
Rabbit	1 ±0	2 ±0	3 ±0
Mouse	1 ±0	2 ±0	3 ±0
Elephant	1 ±0	2 ±0	3 ±0
Deer	1 ±0	2 ±0	3 ±0
Tiger	1 ±0	2 ±0	3 ±0
Lion	1 ±0	2 ±0	3 ±0
Rhino	1 ±0	2.2 ±0.4	2.8 ±0.4
Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Whale	1 ±0	2 ±0	3 ±0
Dolphin	1 ±0	2.2 ±0.4	2.8 ±0.4
Fish	1 ±0	2 ±0	3 ±0
Jellyfish	1 ±0	2 ±0	3 ±0
Octopus	1 ±0	2 ±0	3 ±0
Shrimp	1 ±0	2 ±0	3 ±0
Lobster	1 ±0	2 ±0	3 ±0
Starfish	1 ±0	2 ±0	3 ±0
Tadpole	1 ±0	2 ±0	3 ±0
Crawfish	1 ±0	2 ±0	3 ±0

Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Ant	1 ±0	2 ±0	3 ±0
Bee	1 ±0	2 ±0	3 ±0
Beetle	1 ±0	2 ±0	3 ±0
Butterfly	1 ±0	2 ±0	3 ±0
Caterpillar	1 ±0	2 ±0	3 ±0
Cockroach	1 ±0	2 ±0	3 ±0
Grashopper	1 ±0	2 ±0	3 ±0
Mantis	1 ±0	2 ±0	3 ±0
Mosquito	1 ±0	2 ±0	3 ±0
Moth	1 ±0	2 ±0	3 ±0
Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Refrigerator	1 ±0	2 ±0	3 ±0
Kettle	1 ±0	2 ±0	3 ±0
Lamp	1 ±0	2 ±0	3 ±0
Toaster	1 ±0	2 ±0	3 ±0
Vase	1 ±0	2 ±0	3 ±0
Phone	1 ±0	2 ±0	3 ±0
Table	1 ±0	2 ±0	3 ±0
Chair	1 ±0	2 ±0	3 ±0
Stove	1 ±0	2 ±0	3 ±0
Sink	1 ±0	2 ±0	3 ±0

Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Airplane	1 ±0	2 ±0	3 ±0
Bicycle	1 ±0	2 ±0	3 ±0
Boat	1 ±0	2 ±0	3 ±0
Bus	1 ±0	2 ±0	3 ±0
Car	1 ±0	2 ±0	3 ±0
Helicopter	1 ±0	2 ±0	3 ±0
Motorcycle	1 ±0	2 ±0	3 ±0
Ship	1 ±0	2 ±0	3 ±0
Train	1 ±0	2 ±0	3 ±0
Truck	1 ±0	2 ±0	3 ±0
Target word	**Close distractor**	**Middle distractor**	**Far distractor**
Clarinet	1 ±0	2 ±0	3 ±0
Cymbals	1 ±0	2 ±0	3 ±0
Drums	1 ±0	2 ±0	3 ±0
Guitar	1 ±0	2 ±0	3 ±0
Kazoo	1 ±0	2 ±0	3 ±0
Mandolin	1 ±0	2 ±0	3 ±0
Piano	1 ±0	2 ±0	3 ±0
Ukelele	1 ±0	2 ±0	3 ±0
Violin	1 ±0	2 ±0	3 ±0
Bugle	1 ±0	2 ±0	3 ±0