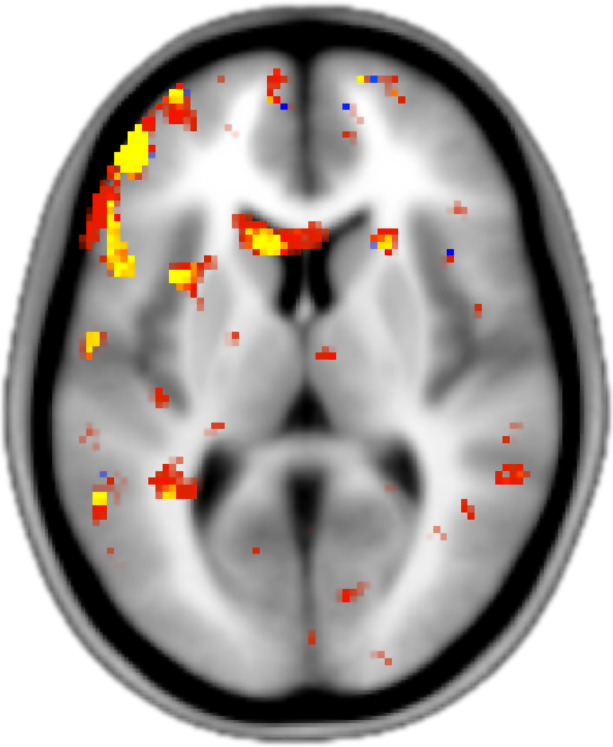
Predicting Working Memory Load from fMRI Data

Rasmus Arnling Bååth

Lund University Cognitive Science
and
Department of Computer Science
Lund University, Sweden

rasmus.arnling_baath@lucs.lu.se

June 29, 2009



Abstract

This thesis describes an attempt to answer the question: Is it possible to predict the working memory load of a subject from functional magnetic resonance imaging (fMRI) data? fMRI data was collected from experiments where subjects were performing tasks putting varying degrees of load on working memory. An artificial neural network classifier was trained to predict the working memory load given fMRI data. This was successful and further analyzes showed that activation in prefrontal cortex held the information most important to the classifier.

1 Introduction

Since the day we are born our memory is online and busy working. In a lifetime an average human learns things as diverse as how to operate many different tools, how to discriminate between the voices of hundreds of persons, and the meaning of thousands of words. It is e.g. estimated that a typical seventh grader learns the meaning of 10–15 words every day (Landauer and Dumais, 1997). Still, most of the information encountered is not remembered and some information, e.g. words in a foreign language, can be hard to remember even if one is actively trying to. What is always possible though, is to keep any small piece of information, say a foreign word or a telephone number, in memory for a short time. Here is also a clue to that memory is not a single unitary system.

One of the first models of memory that included a division of memory into subsystems were the *multi-store model* of Atkinson and Shiffrin (1968) which divides memory into *short-term* and *long-term* memory. Evidence for this division came from, e.g., studies of brain-damage patients that performed well in tasks only requiring a short memory span, while being unable to form lasting memories. Long-term memory is seen as a memory system that can store a huge amount of information and maintain this information over many years. It is not possible to process and manipulate information in long-term memory directly. This is however possible with information in short-term memory which, as opposed to long-term memory, has a very limited capacity. This limited capacity was already investigated by Miller (1956) in his famous article "The magical number seven, plus or minus two.", which title refers to the maximum capacity of short-term memory.

A further refinement of short-term memory, the *working memory* (WM) model, was proposed by Baddeley and Hitch (1974). It stresses the role of memory as a system concerned with processing of information and not only storage. It was found in some experiment setups that the performance of subjects performing two simultaneous tasks requiring the use of two separate perceptual domains, such as verbal and visual, was nearly as good as when performing the tasks individually. This finding led to the proposal that WM consists of three separate subsystems: The *phonological loop* that works with auditory information, the *visuospatial sketchpad* that holds visual information, and the *central executive* which directs the attention of WM and inhibits irrelevant information. Since the model of Baddeley and Hitch was introduced it has been criticized, built on and refined. But the concept of WM is still an active area of research[1] and will probably remain so while researchers try to answer one of the most important questions of cognitive science: How is the human memory organized and what is its neural basis?

1.1 Working Memory and the Brain

Much effort have gone into investigating the physiological mechanisms of WM. Single-cell recording studies on monkeys have been performed where electrical activity was recorded from the *prefrontal cortex* (PFC) while the monkeys where subject to a delayed response task (Fuster, 1973). An experimenter hides a piece of food in one of two cups while the monkey is watching, the cups are then concealed for a short time period and when shown again the monkey receives a reward if it selects the cup hiding the food. This task requires the monkey to keep the information regarding the location of the food in WM. It was found that some neurons in the PFC fired mostly during the time the cups were concealed which would support the hypothesis that the PFC is important to WM.

With the advent of *functional magnetic resonance imaging* (fMRI) in the early 90's it became easier to study the human brain in a non-invasive manner. Many fMRI studies have been performed investigating different aspects of memory, among others, WM. These studies have mostly focused on trying to localize what parts of the brain are important to WM. In a review of 67 fMRI studies investigating WM (Cabeza and Nyberg, 2000) it was found that increased demands on WM was almost always associated with increased activation in the PFC. This increased activation was found especially in Broadman area (BA) 6 and 44. For tasks that seem to require manipulation of WM content, activation was also increased in BA 9 and 46. It is not clear exactly which regions are associated with WM, some studies have reported activation in regions other than PFC such as occipital and cerebellar regions, what is clear though is that no single region seem to be responsible for WM.

Recently it has become popular to analyze fMRI data using *multi-voxel pattern analysis* (MVPA) (Haynes and Rees, 2006). Whereas standard univariate fMRI studies compares how the activation of individual voxels change between different conditions, MVPA studies focus on the pattern of voxel activations. In the case of WM conditions should be tasks putting load on WM. One such task is the *n-back task* (for a description see section 2.1.1)

[1] Google Scholar (http://scholar.google.com) reports that more than 13,000 articles with the phrase "Working Memory" in the title were published in 2008.

where a higher n presumably puts more pressure on WM. This is also the task studied in this thesis. MVPA studies often utilizes standard machine learning classifier algorithms such as *artificial neural networks* or *support vector machines*. These are trained to recognize patterns that code for different conditions and if this is successful some of the questions that can be probed are:

- *Is there enough information in the fMRI data to predict the cognitive state of a subject?* In an fMRI study were a subject performs various tasks, or is exposed to several categories of stimuli, it is hypothesized that this is reflected in the cognitive state of the subject. If a classifier can be trained to classify the task/stimuli a subject is exposed to, given only fMRI data, this shows that different cognitive states are actually induced and that this can be discerned from the low resolution data that fMRI is.
- *Where is information located?* That is, which voxels carry the information most important to the classifier. This question is similar to the one answered by univariate studies. The benefit of using MVPA is that it can be more sensitive when detecting cognitive states (Norman et al., 2006). For example; even if no single voxel is informative enough a MVPA approach can find patterns of activation that code for cognitive states.
- *How is information encoded?* E.g., how complex are the patterns of activation that code for cognitive states, does the classifier perform better when non-linear relations between voxels are taken into consideration or are linear relations sufficient? An interesting extension to this question is whether it is possible to predict the patterns of activation given it is known what task/stimulus the subject is exposed to. This was recently investigated by Mitchell et al. (2008) who built a model that successfully could predict the voxel activity that would result of exposing a subject to a concrete noun.

Of these questions maybe the first one is the most intriguing, as to correctly predict someones cognitive state could also be called "mind reading".

1.2 The Focus of This Thesis

This thesis describes an attempt to answer the question: Is possible to predict the WM load of a subject from fMRI data? Of the three questions outlined above, the two first will be probed, thus the second question asked is: Where is information regarding WM load located in the brain?

There are many univariate fMRI studies that investigate WM, but no MVPA study of WM has been done as far as the author knows. If found that predicting WM load is possible this would be interesting both theoretically and practically and would pave the road for more complex questions regarding the nature of WM.

2 Method

fMRI data was collected from experiments where subjects were performing tasks putting varying degrees of load on WM. This data was then analyzed using a MVPA approach, an artificial neural network classifier was implemented to predict the WM load given fMRI data from a subject. fMRI data was also collected from subjects performing tasks putting varying degrees of load on episodic memory. Even if not the main focus of this thesis, this data was analyzed in the same way as the WM data in order to compare the results.

2.1 fMRI Data

The fMRI data used in this thesis was originally collected in a study by Marklund et al. (2007) and a more detailed description of their method can be found in their article.

16 subjects participated in the fMRI study (age range: 24–37, eight male). Stimuli consisted of words presented on a screen in white font on black background. These words were drawn from a set of 138 common, arbitrarily chosen, nouns. Each subject participated in four sessions, each session consisting of four different tasks presented in blocks separated by "resting" blocks. Each task block lasted for 63 s. and exposed subjects to eight words, where words were shown for 2.5 s. followed by a break of varying length. All in all, each subject was exposed to 128 words and $16 \times 128 = 2048$ were viewed in total. Of the four tasks, two where n-back tasks and two were word recognition tasks (see section 2.1.1 for a detailed description). All tasks required a yes/no response for each word shown and this was made by the subjects by pressing yes/no on a response pad. This thesis is mainly concerned with the n-back tasks but for completeness all four tasks will be described.

2.1.1 Experimental Tasks

The n-back tasks were standard 1-back and 2-back tasks respectively (Kirchner, 1958). Each block subjects were successively shown eight words and

Figure 1: An example of 1-back and 2-back tasks, taken from session 1 of subject 14.

were supposed to indicated for each word whether it was identical to the word displayed one/two words back by pressing yes/no (see fig. 1) . The n-back task is a popular paradigm for studying WM (Owen et al., 2005), as an increase of n presumably puts a higher load on WM. In this setup 1-back is the low-load WM task and 2-back is the high-load WM task.

The word recognition tasks were also designed as one low-load and one high-load task, but putting the load on episodic memory instead of WM. Prior to the scanning sessions subjects were given two lists of 16 words, LIST 1 and LIST 2, to memorize. In the low-load *item recognition* task subjects where shown words drawn from the lists mixed with novel words and where required to discriminate between them by pressing "yes" if a word had occurred on any of the two lists and "no" otherwise. In the high-load *item source* task subjects were also shown a mix of words from the lists and novel words. Together with each word was displayed either "LIST 1" or "LIST 2" and the subjects where to indicate whether the current word was draw from the currently indicated list by pressing either "*yes*" or "no". This latter task puts a higher load on episodic memory since it not only requires subjects to remember if a word was on a list but also what list it was on.

The order of the task blocks in each session was permuted so that the order was different for all sessions, this to avoid side effects that might arise if the tasks were presented in similar order each session. An example of an experiment setup is given

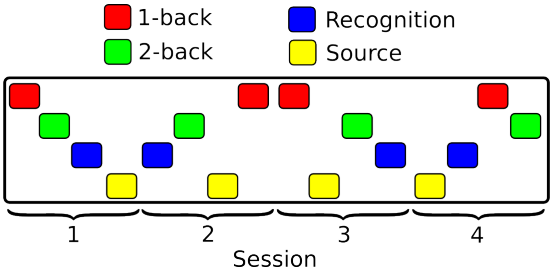


Figure 2: The experiment setup for subject 1.

in fig. 2. In total a subject performed each task four times, once in every session.

2.1.2 fMRI Data Acquisition Details

Data was collected on a 1.5-T Philips Intera scanner (Philips Medical Systems, Netherlands). Functional T2*-weighted images were obtained with a single-shot gradient echo EPI sequence used o for BOLD imaging. The sequence had the following parameters: echo time, 50 ms; repetition time, 3000 ms (33 slices acquired); flip angle, 90°; field of view, 22 × 22 cm, 64 × 64 matrix and 4.4 mm slice thickness. Functional imaging data was preprocessed in a number of steps prior to statistical analyzes using the SPM2 software package[2] on Matlab[3]. All image volumes were first corrected for variability in slice timing acquisition. Image realignment to the first volume was then performed using a truncated sinc interpolation. The image volumes were then normalized to an approximate Talairach space (Talairach and Tournoux, 1988) as defined by the SPM2 T1-weighted MNI template and smoothed with an isotropic 8-mm full-width at half-maximum Gaussian kernel. Finally the image volumes of each session underwent voxel-wise linear detrending and z-scoring. For each subject 4×119 fMRI images were collected, each having the size $79 \times 95 \times 79$. After masking away voxels outside the brain each fMRI image contained approximately $240,000$ voxel activation values.

2.2 The Predictive Model

The predictive model was built using Matlab and the Princeton Multi-Voxel Pattern Analysis toolbox[4]. Training examples were created for each subject by labeling voxel response patterns for each word onset as belonging to a category (e.g. 1-back, 2-back, etc.). An artificial neural network classifier was trained on the examples on a per-subject basis

[2] http://www.fil.ion.ucl.ac.uk/spm/
[3] http://www.mathworks.com/products/matlab/
[4] http://www.csbmb.princeton.edu/mvpa/

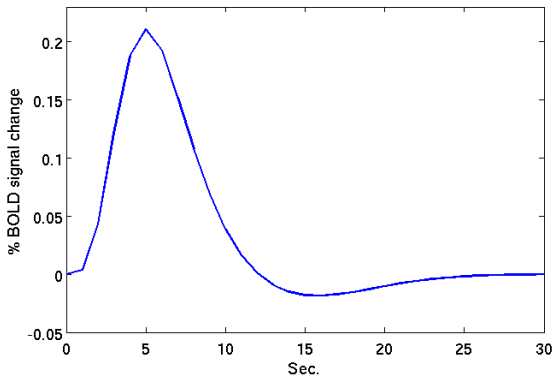


Figure 3: The canonical haemodynamic response function as specified in SPM2. This function is an approximation of the blood-oxygen-level dependent (BOLD) signal in the brain as the response to some stimuli.

and evaluated by calculating the percentage of correct answers. Training and testing was done using cross-validation; training on examples from three of the sessions and testing on the fourth, repeating this for each session. To ascertain the location of the areas influencing classifier, importance maps where created for each trained classifier, highlighting the most influential voxels.

2.2.1 Creation of Training Examples

For each of the 2048 word onsets w_i a weighted mean onset image M_i was created by using a canonical *haemodynamic response function* specified in SPM2 (see fig. 3). A $|w| \times |I|$ matrix H was constructed, where I is a vector of all fMRI images and $H_{i,j}$ is the *haemodynamic response* (HR) resulting from w_i at the time of I_j as predicted by the canonical HR function. An example of an H matrix is given in fig. 4. The onset image M_i was then created for each w_i by

$$M_i = \sum_{j=1}^{|I|} \left(I_j t_{i,j} \right),$$

$$t_{i,j} = \frac{t'_{i,j} t''_{i,j}}{\sum_{h=1}^{|I|} \left(t'_{i,j} t''_{i,j} \right)},$$

$$t'_{i,j} = \frac{H_{i,j}}{\sum_{k=1}^{|I|} \left(H_{k,j} \right)},$$

$$t''_{i,j} = \frac{|H_{i,j}|}{\sum_{l=1}^{|w|} \left(|H_{i,l}| \right)}.$$

The intuition behind this equation is that the contribution of the image I_j to M_i, that is $t_{i,j}$, should depend on the HR of w_i at the time of I_j relative to all HR values of w_i, that is $t'_{i,j}$, and the HR of w_i relative to all word onsets w at the time of I_j, that is $t''_{i,j}$.

Training examples were made by partitioning the training data by labeling each onset image as belonging to a category. Three different partitionings of the training data were made:

- *1-back vs. 2-back.* This partitioning was made to test the main question of this thesis; is it possible to train a classifier to discriminate between a low-load and a high-load WM task? Onset images were labeled depending on whether they were from a 1-back task block or a 2-back task block. The two first onset images in every task block were discarded as it is plausible that the WM load of the 2-back task does not kick in until two words have been viewed. Only onset images from the n-back task blocks were labeled, onsets images from the word recognition tasks were not used. This partitioning yielded 12 training examples per session, in total $4 \times 12 = 48$ training examples per subject.

- *Recognition vs. source.* This partitioning was made to test whether it would be possible to train a classifier to discriminate between a low-load and a high-load episodic memory task. Onset images were labeled depending on whether they were from a item recognition task block or a source recognition task block. Onsets images from the n-back task blocks were not used. This partitioning yielded 16 training examples per session, in total $4 \times 16 = 64$ training examples per subject.

- *n-back vs. word recognition.* This partitioning was made to test how well a classifier can discriminate between the two types of tasks; n-back and word recognition. As these tasks are not easily comparable the outcome of this test is difficult to interpret, but since the tasks are so different the performance of the classifier is expected to be well above chance. This partitioning yielded 32 training examples per session, in total $4 \times 32 = 128$ training examples per subject.

For each of these partitionings a randomized partitioning was also made, e.g. the randomized partitioning of 1-back vs. 2-back included the same training data but assigned onset images labels by coin flipping. The performance of the classifier when classifying these randomized partitionings is

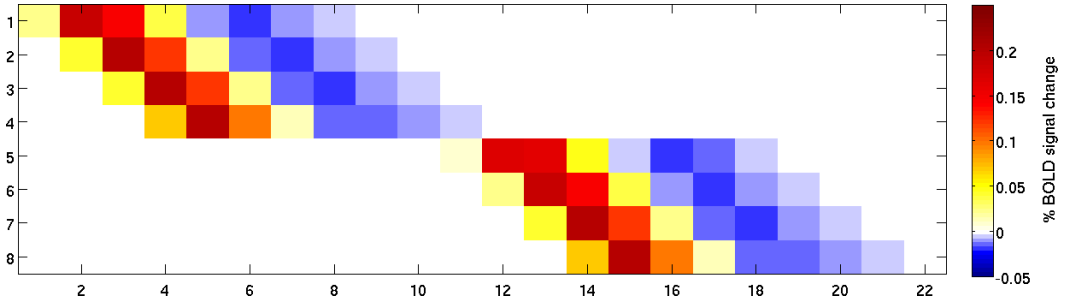


Figure 4: The H matrix from subject 1, session 2, task block 2-back. Each row is the HR of a word onset, each column is the HR of all onsets at the time of an fMRI image. Compare with fig. 3.

expected to be no better than chance. These partitionings were included in the analysis to ascertain that the classification of the regular partitionings was not artificially boosted.

2.2.2 Training Setup

The classifier was trained and tested on a per-subject basis. To avoid cheating by training and testing on the same data a *K-fold cross-validation* approach was used. The training examples of each subject went trough four training–test cycles. Each cycle the classifier trained on examples from three of the four session, the classifier was then tested on the examples from the remaining session. After four cycles all examples from all session had been tested on. Note that for the three non-randomized partitionings every session contain the same number of training examples from both categories. This is important since if there are more training examples from one category the classifier will seemingly achieve above chance performance by classifying all examples as belonging to that category.

Training the classifier on all voxels is possible but better performance is often reached if uninformative voxels are filtered away by a feature selection procedure (Pereira et al., 2009). In order to find the voxels that significantly deviate between different labels an ANOVA was run on a voxel-by-voxel basis. This generated a p-value for each voxel and the 10000 voxels with the lowest p-values were singled out to be used in classification. This feature selection was only made on the training examples of each training-test cycle to avoid "peeking" at the test examples.

As the classifier is not deterministic the performance of a single subject training–test run could be misleading. Therefore all classification results presented in this thesis is the average of 50 training–test runs.

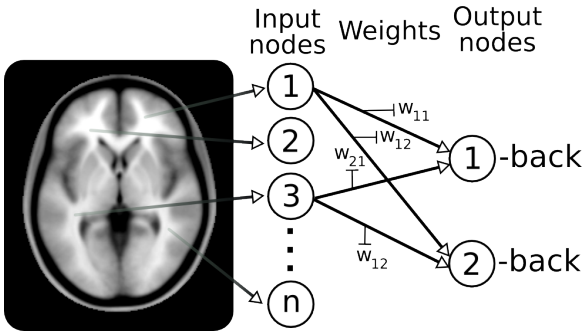


Figure 5: Schematic of the artificial neural network classifying fMRI images from the n-back tasks.

2.2.3 Artificial Neural Network Classifier Details

The classifier used was an *artificial neural network* (ANN), a standard machine learning algorithm that is widely used to solve many different pattern classification problems (for a tutorial see e.g. Jain et al. (1996)). The ANN was implemented using MathWorks Neural Network Toolbox[5], and was a feed-forward multi-layer perceptron network (see fig. 5 for a simplified schematic). No hidden layer was used. This makes the ANN a linear classifier, thus it is only able to find linear relations between voxels. There are two reasons for the choice of a single-layer network: It is not obvious that nonlinear classifiers perform better than linear ones when training on fMRI data (Pereira, 2007) and it is easier to interpret the importance of the weights in a single-layer network.

The ANN had one input node for each voxel of the 10000 selected voxels and one output node for each category. The output nodes used a log-sigmoid transfer function and the category predicted, given an example as input, was the category of the output node with the highest output. Before training the weights of the network were initialized

[5]http://www.mathworks.com/products/neuralnet/

to random values. During training the performance of the ANN was measured as the mean squared error between the response of the output nodes and the correct response. As a training function the ANN used conjugate gradient descent backpropagation. Training stopped when performance was less than 0.00001 or when 200 training epochs had been completed.

2.2.4 Creation of Importance Maps

The ANN classifier importance maps shown in the Results section were created by the method used by Polyn et al. (2005). The importance $v_{i,j}$ of a voxel, where i and j are numbers uniquely defining each voxel and category respectively, is given by:

$$v_{i,j} = 100 \cdot w_{i,j} a_{i,j},$$

where $w_{i,j}$ is the weight between the input unit of voxel i and output unit of category j and $a_{i,j}$ is the average activity of the input unit of voxel i for the training examples drawn from category j. The value 100 is an arbitrary upscaling factor. $v_{i,j}$ can be both positive and negative, a positive/negative value indicates that voxel i has a positive/negative influence on the output unit of category j.

3 Result

The performance of the classifier when classifying the different partitionings is summarized in table 1 and fig. 6. The following results, unless noted otherwise, are the combined results of 50 training–test runs with a classifier, from all subjects. For all partitionings the performance of a classifier relying on chance alone is expected to be 0.5.

In the case of 1-back vs. 2-back performance was above chance ($mean = 0.62$) and significantly so (t-test, $p < 0.003$). Performance in the case of n-back vs. word recognition was even better ($mean = 0.79$ and t-test, $p < 10^{-8}$) but this was expected as it is plausible that the differences between the two categories of tasks are greater than the differences within these categories. In stark contrast to these results was the performance in the case of recognition vs. source, where it seems like the classifier performed no better than chance ($mean = 0.50$ and t-test, $p = 0.94$). When classifying the three randomized partitionings the performance was close to chance ($mean = 0.51, 0.51, 0.49$) and did not significantly deviate from it (t-test, $p > 0.63, 0.47, 0.43$), this was to be expected and supports the claim that the performance when classifying the non-randomized partitionings was not artificially boosted.

	1-back	2-back
accuracy, mean	0.96	0.95
accuracy, SD	0.04	0.05
Response time, mean	920 ms	1001 ms
Response time, SD	99	106
Difficulty rating, mean	1.2	2.7

Table 2: Summary of the behavioral analysis of the n-back tasks.

3.1 Behavioral Analysis

A behavioral analysis was performed by Marklund et al. (2007) investigating the difficulty of the n-back and word recognition tasks. Response time and accuracy were recorded for all tasks and subjective measures of task difficulty were acquired by asking subjects to rank the effort associated with each task on a scale from 1 (low effort) to 5 (high effort). The results from this analysis are shown in table 2. It was found that there was no significant difference (t-test, $p > 0.7$) between the accuracy of the n-back tasks. There was a difference however in response time (t-test, $p < 0.02$) and difficulty rating (t-test, $p < 0.001$) where the 2-back task had longer response time and was rated as more difficult than the 1-back task. This shows that, even though performance was similar, there is a genuine difference in difficulty, both subjective and objective, between the 1-back and 2-back tasks.

3.2 Evaluation of Performance

To establish the significance of the performance of a classifier an independent one-sample t-test is used. The number of degrees of freedom is 15 (as the number of subjects are 16) and the null hypothesis is that the mean is 0.5. As an alternative to the t-test it would be possible to use the binomial test and view every choice of the classifier as a trial. This test is used when establishing the significance of the performance on a single subject but it is questionable if it is sound to use this test over all subjects. This is because the performance of the classifier probably is dependent on the subject, thus the trials would not be independent. Never the less, a binomial test of the performance of the classifier over all subjects in the 1-back vs. 2-back case shows that, as a t-test does, performance is significantly above chance ($trials = 768$, $successes = 473$, $mean = 0.62$, $p < 10^{-9}$).

3.2.1 1-back vs. 2-back

As reported, the performance of the classifier in the 1-back vs. 2-back case was significantly above chance. The best performance was reached when

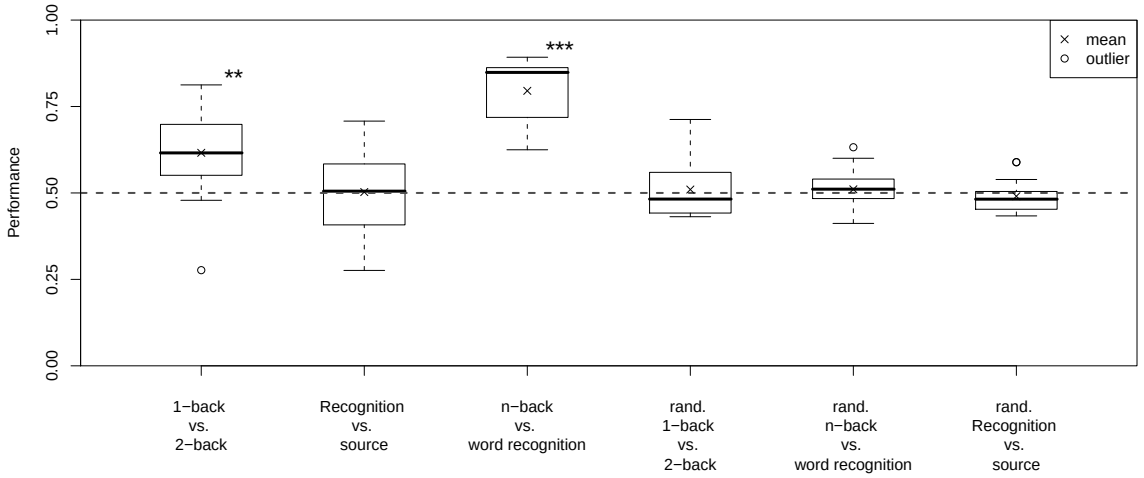


Figure 6: Box plot over the performance of the classifier on a per-subject basis when classifying the six partitionings described in section 4. A value is considered an outlier if it is 1.5 × the interquartile range outside the first and third quartile.

	1-back vs. 2-back	Recognition vs. source	n-back vs. word recognition
performance , mean	0.62	0.50	0.79
t-test, p	< 0.003	0.94	$< 10^{-8}$

Table 1: Summary of the performance of the classifier.

classifying subject 5 (binomial test, $trials = 48$, $successes = 39$, $mean = 0.81$, $p < 10^{-4}$). Classification of 14 out of 16 subjects yielded performance above chance, the other two subjects were subject 14 ($mean = 0.48$) and 16 ($mean = 0.28$). One explanation of the variation in performance across subjects is given by Mitchell et al. (2008) who have shown that there is a strong negative correlation between performance and the estimated head motion of subjects. In the current study this information is not available however. As the performance when classifying subject 16 is much lower than than the mean (more than 2.5 SDs) it could be considered an outlier.

3.2.2 Recognition vs. Source

In this case performance was no better than chance. Classification of half of the subjects yielded performance above chance. Both mean and median of the performance was 0.50.

3.2.3 n-back vs. Word Recognition

Of the three partitionings, the performance of the classifier was best when classifying the n-back vs. word recognition case. Classification of all subjects reached performance above chance with the highest performance when classifying subject 5 (binomial test, $trials = 128$, $successes = 114$, $mean = 0.89$, $p < 10^{-15}$). Even classification of the subject that yielded the lowest performance still was significantly above chance (binomial test, $trials = 128$, $successes = 80$, $mean = 0.62$, $p < 0.01$).

3.3 Importance Maps

Importance maps were created for all subjects. As the main focus of the thesis is the WM tasks, only importance maps from 1-back vs. 2-back are shown here. Fig. 7 shows importance maps from subject 5 and mean importance maps created from all subjects. Subject 5 was chosen since classifying that subject resulted in the best performance. The mean importance maps were created by taking importance maps from all subjects and calculating the mean value of each voxel.

As can be seen in fig. 7 the color scale of the importance maps from subject 5 differ from the color scale of the mean importance maps. This is because the importance maps of a subject only contains values for those 10000 voxels that were selected by the ANOVA feature selection procedure, all other voxels receive an importance value of 0. The 10000 voxels are not necessary the same for every subject and when all importance maps are averaged together the result is a map with lower peaks than for any single subject.

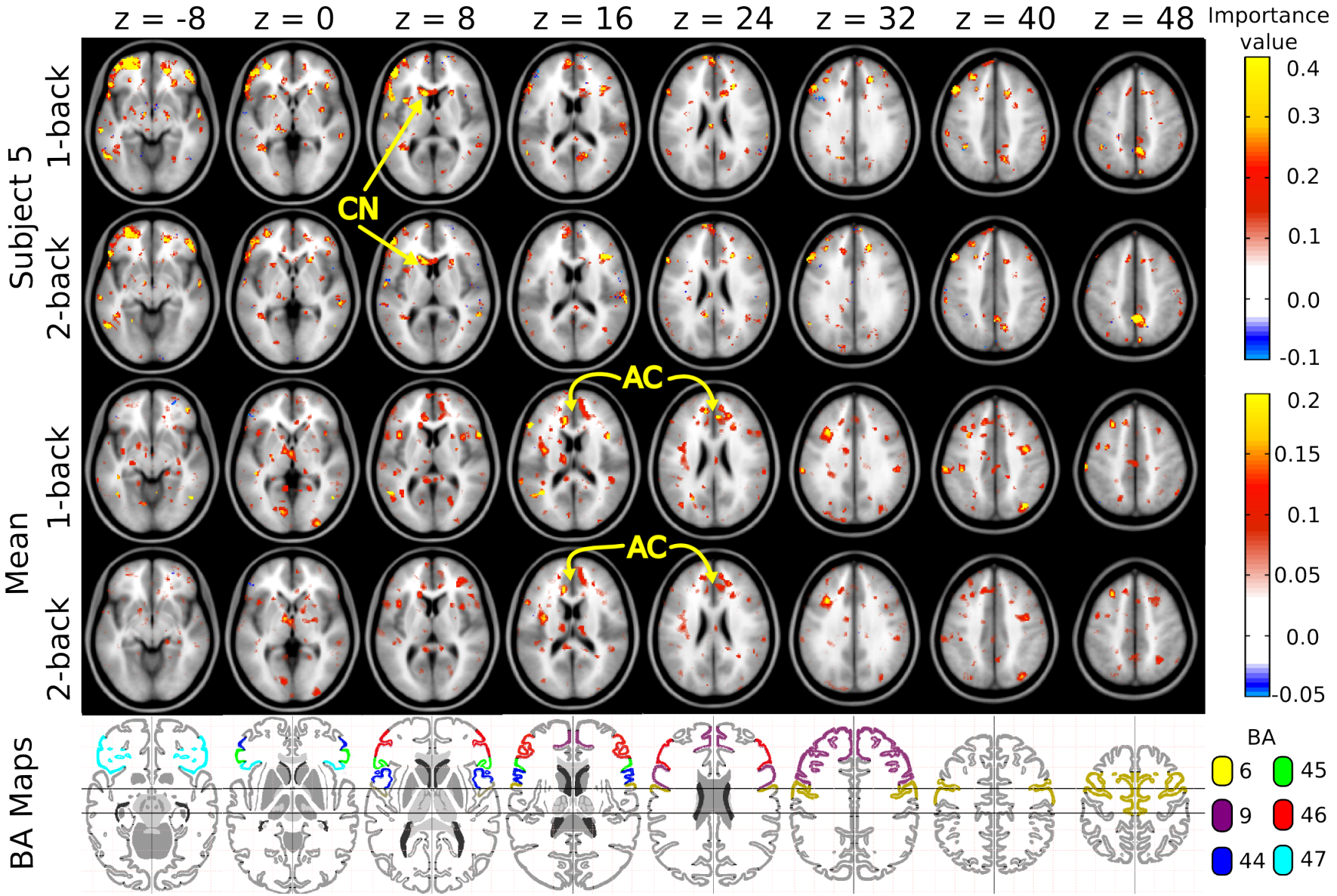


Figure 7: Importance maps from classification of the 1-back vs. 2-back case. The two top rows show maps from subject 5, the subject the classifier performed best on (binomial test, $trials = 48$, $successes = 39$, $mean = 0.81$, $p < 10^{-4}$). The following two rows are maps that show the mean of all 16 subjects, note though that classification was done on a per-subject basis. Red/blue indicate voxels that positively/negatively influences the classifier to choose the category of the row of the map. When the importance of a voxel is small it fades into transparency, shown as white on the color bars. The colored Broadman areas (BA) in the bottom row are areas shown to be related to WM in other fMRI studies (Cabeza and Nyberg, 2000; Owen et al., 2005). These were created with the help of the Talairach Daemon (Lancaster et al., 2000). The CN and AC labels point to the left lateralized caudate nucleus and anterior cingulate cortex respectively. The value given Z indicates in mm the position in the sagittal plane of the corresponding column of slices.

4 Discussion

The purpose of this thesis was to attempt to answer the question: Is possible to predict the WM load of a subject from fMRI data? Given the results presented it is concluded that the answer to this question is "yes". One might also answer "yes, to a degree" when considering that the mean performance of the classifier was 0.62, but this result is not so bad when considering the similarity of the 1-back and 2-back tasks.

It might also be that the WM requirements of 1-back and 2-back are not as different as one might think. The assumption is that an n-back task requires a subject to keep n items in working memory. But this might not always be the case as there is nothing stopping a subject performing a 1-back task to keep more than only the last seen item in memory. A tendency to keep more items in memory might be especially strong if a subject recently performed a 2-back task, which have been the case in the experiments described in this paper.

While this might be a concern, there are also reasons why the n-back tasks described in this thesis really have produced differences in WM load. As the behavioral analysis show (section 3.1) there was a real difference in difficulty between the n-back tasks. The reaction time was longer for the 2-back task and it was rated as more difficult by the subjects. The fact that subjects performed nearly as well on both n-back tasks does not imply that there was not a difference in difficulty. Instead this is desirable since a discrepancy in performance could result in e.g. irritation or lessened ability to concentrate, which might effect the fMRI data.

It is interesting to note that while classifying in the 1-back vs. 2-back (WM) case worked well, the recognition vs. source (episodic memory) case did not. This, even if there also was a difference in performance, reaction time and experienced difficulty between the two word recognition tasks, the item source task being the more difficult in all respects (Marklund et al., 2007). It is hard to infer anything from this result, but a conclusion one might draw is that the activation evident in the fMRI data does not contain as much information about the distinction between the word recognition tasks as it does about the n-back tasks.

That the performance of the classifier in the case of n-back vs. word recognition was better than in any of those single categories is not surprising. What might be surprising is that performance was so high ($mean = 0.79$, $max = 0.89$), since the two categories of task were not that different. The stimuli shown (words in white on a black surface), the intervals between stimuli, and the reaction required (the press of a button) were identical for both categories of tasks. A speculation is that one reason why classification performance was good is because the difference in WM requirement between the two categories of tasks helps the classifier to separate them. Presumably n-back in general requires more of WM than word recognition.

4.1 Importance Maps

While studying the importance maps in fig. 7 it is important to note that these maps do not show what is shown in standard fMRI contrast image: Areas that have higher activation compared to another condition or a baseline. What these importance maps show is what voxels positively or negatively influenced the classifier to choose a certain category. One issue that makes these maps difficult to interpret is that they will include noise as the performance of the classifier was not perfect. Another issue is that importance maps of subject 5 differs from the mean importance maps, it's hard to tell which to trust the most. On one hand the mean importance maps should be more reliable as noise from the single subject importance maps is averaged away. On the other hand some of the single subject maps are derived from classifiers with performance below chance, thus the importance maps of subject 5 should be more reliable as classification of that subject led to the best performance ($mean = 0.81$). However, both kinds of importance maps show areas of importance that are in accordance with findings from standard univariate fMRI studies of WM and both will be considered in this discussion.

Before going into details, two things to note are that there are very few voxels with negative influence (blue voxels) and that the areas of importance for 1-back is mostly overlapping 2-back. This is the case for both the mean maps and the maps of subject 5. The reason for this is probably the similarity of the n-back tasks, as the voxels that codes for both tasks are overlapping.

Many fMRI studies have been investigating WM. Cabeza and Nyberg (2000) reviews 67 fMRI studies using different experimental paradigms to study WM and Owen et al. (2005) reviews 24 fMRI studies using the n-back experimental paradigm. When below, a brain area is claimed to be of importance to WM, these two studies are referred to.

In general WM is associated with PFC and Broadman areas (BA) 6, 44, 9, and 46. Of these areas, activation in 9 and 46 seem to occur mostly for tasks requiring manipulation of WM content, such as n-back. Consequently BA 6 and 44 are not especially highlighted in the importance maps,

while 9 and 46 are, though some important voxels can be found in BA 44 in the mean maps (Z=8). Important voxels in BA 9 and 46 are found in both the maps of subject 5 and the mean maps (BA 9 in Z=16, 24, 32 and BA 46 in Z=8, left lateralized). In the mean maps (Z=16, 24) there are important voxels in the anterior cingulate cortex (AC). Increased activity in this area is often connected with increased effort or attention. For subject 5 there is a cluster of left lateralized important voxels in BA 44, 45, and 47 (Z=-8, 0, 8). This is notable since activation in *Broca's area* (BA 44, 45 left lateralized) is associated with verbal WM tasks and BA 47 is especially associated with verbal identity monitoring. More generally, the mid-ventrolateral frontal cortex (BA 45, 47) have shown activation in processes connected to the n-back task such as holding non-spatial information online. Finally important voxels are found in the left caudate nucleum of subject 5 (Z=8), which is associated with manipulation of information in WM (Lewis et al., 2004).

4.2 Future Research

The result from the study of the importance maps is interesting but mostly confirm much that has already been shown in standard univariate fMRI studies. The most interesting result presented is that WM load is possible to predict given only fMRI data. Except for being interesting theoretically, this find has practical implications too.

The work presented in this thesis is a first step towards being able to predict WM load, not just for the n-back task, but more generally for any task. If this was made possible it would be a very useful tool, e.g. when studying human-computer interaction. When building human-computer interfaces it is desirable that an interface does not demand to much of the human WM. If it was possible to measure WM load, different interface implementations could be compared in this regard. Another phenomenon that could be investigated is *chunking*, first introduced by Miller (1956). Chunking refers to our capability to group together familiar "chunks" of information into one unit, for example when trying to remember a telephone number. Good chunking capabilities are often found in experts, a classical example being the ability of chess experts to remember the positions of a chess board given just a brief look (Chase and Simon, 1973). It is presumed that good chunking capabilities lighten the burden of WM. If general prediction of WM load was possible this hypothesis could be tested. This could be done using the n-back experimental paradigm but using WM demanding stimuli such as chess positions.

To make general prediction of WM load possible more experimental paradigms than n-back has to be studied, as n-back does not activate all processes ascribed as WM processes. E.g. activations when performing verbal a n-back task are already different from those when performing a spacial or pictorial n-back task (Owen et al., 2005). It would also be desirable to repeat the analysis of this thesis using data from n-back tasks where $n > 2$. One difficulty with doing this is the sheer difficulty of n-back for high n, already 3-back is very demanding. If subjects perform poorly when doing n-back tasks for high n this could lead to subjects feeling irritation or anxiety and one might end up classifying that instead of WM load.

4.3 Conclusion

This thesis has described an attempt to answer whether it is possible to predict working memory load from fMRI data. To test this, data from a previous study was used (Marklund et al., 2007) where subjects were exposed to one low-load and one high-load task requiring working memory. Using a multi-voxel pattern analysis approach an artificial neural network classifier was trained to, given an fMRI image, predict the corresponding task. This was successful as the classifier performed significantly better than chance. It was concluded that this find is interesting both theoretically and practically, but that more research is needed before general prediction of working memory load, that is prediction not bound to a specific task, is possible.

5 Acknowledgments

I would like to thank my supervisor Sverker Sikström. I would also like to thank Christian Balkenius for all his support. I would like to thank Petter Marklund and Petter Kallionen for supplying the fMRI data and for helping me out with the analyzes. I thank LUNARC[6] for supplying enough computing power to make the analyzes of this thesis possible. At last I would like to thank my father, Erland Bååth, for coaching me in questions regarding statistics.

~

[6] www.lunarc.lu.se

References

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2:89–195.

Baddeley, A. D. and Hitch, G. J. (1974). Working memory. *Recent advances in learning and motivation*, 8:47–90.

Cabeza, R. and Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of cognitive neuroscience*, 12(1):1–47.

Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, 4(1):55–81.

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology*, 36(1):61–78.

Haynes, J. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.*, 7(7):523–534.

Jain, A. K., Mao, J., and Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, 29(3):31–44.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4):352–358.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., and Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*, 10:120–131.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Lewis, S. J. G., Dove, A., Robbins, T. W., Barker, R. A., and Owen, A. M. (2004). Striatal contributions to working memory: A functional magnetic resonance imaging study in humans. *European Journal of Neuroscience*, 19(3):755–760.

Marklund, P., Fransson, P., Cabeza, R., Larsson, A., Ingvar, M., and Nyberg, L. (2007). Unity and diversity of tonic and phasic executive control components in episodic and working memory. *Neuroimage*, 36(4):1361–1373.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci. (Regul. Ed.)*, 10(9):424–430.

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). n-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59.

Pereira, F. (2007). *Beyond Brain Blobs: Machine Learning Classifiers as Instruments for Analyzing Functional Magnetic Resonace Imaging Data*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, 45(1 Suppl.):199–209.

Polyn, S. M., Natu, V. S., Cohen, J. D., and Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963–1966.

Talairach, J. and Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* Thieme, Stuttgart.