
Machine Learning of fMRI Virtual Sensors of Cognitive States

Tom Mitchell Rebecca Hutchinson Marcel Just Sharlene Newman
Radu Stefan Niculescu Francisco Pereira Xuerui Wang

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
firstname.lastname@cmu.edu

December 12, 2002

Abstract

Is it feasible to train classifiers to decode the cognitive state of a human subject, based on single-episode fMRI data? If so, these trained classifiers could be used as virtual sensors to detect hidden cognitive states of a subject, providing a key tool for experimental research in cognitive science and in diagnosis of mental processes in patients with brain injuries. Whereas much work has been done on fMRI data analysis methods that average together data collected from repeated stimuli over multiple episodes, little is known about the feasibility of training classifiers to decode cognitive states from single episodes. This paper presents several case studies in which we have successfully trained such classifiers. We explore the technical issues involved in training such single-episode classifiers, and discuss areas for future research. These case studies include training a classifier to determine (1) which of twelve semantic categories of words is being read by a human subject (e.g., a word describing animals or one describing buildings), (2) whether or not a subject finds a sentence ambiguous, and (3) whether the subject is looking at a picture or at a sentence describing a picture.

1 The Problem

The study of human brain function has received a tremendous boost in recent years from the advent of functional Magnetic Resonance Imaging (fMRI), a brain imaging method that dramatically improves our ability to observe correlates of brain activity in human subjects. A typical fMRI experiment can produce one three-dimensional image of the human subject's fMRI activation every half second, at a spatial resolution of a few millimeters, resulting in a spatial-temporal data stream containing tens of millions of observations of local brain activation over the course of a single twenty minute experiment.

This fMRI technology has been used over the past several years to conduct hundreds of studies that identify which regions of the brain are activated when a human performs a particular cognitive function (e.g., reading, mental imagery). The vast majority of this work involves *averaging together* fMRI data collected during multiple time intervals, in which the subject responds to repeated stimuli of some type (e.g., reading a variety of sentences). The most common result of such studies are statements of the form:

“fMRI activity in brain region R is on average greater when performing task T than when resting.”

In contrast, this paper seeks a different kind of result. Our goal here is to study the feasibility of automatically decoding the subject's instantaneous cognitive state, given his/her fMRI activity at a single time instant or time interval. In this paper we use machine learning methods to train classifiers to decode instantaneous cognitive states such as “the subject is reading a sentence,” “the subject is examining a picture,” “the subject finds the sentence ambiguous,” and “the subject is reading a word describing some type of food.” Note our task differs in two ways from the above form of result:

- The classifier must make its decision based on fMRI data from a single time instant or interval.
- We are interested in learning the mapping from observed fMRI to the subject's state, instead of the mapping from the subject's task to activity in particular regions.

Why is this an interesting problem? Because if we could learn such classifiers they would provide a new tool for decoding the sequence of hidden cognitive states a subject passes through when performing some complex task, or for diagnosing the type of difficulty they are having performing some cognitive task. Put succinctly, such classifiers would constitute *virtual sensors* of the subject's cognitive state, which could be useful to scientists and clinicians across a range of cognitive science research and diagnostic medical applications. If such virtual sensors could be trained to classify cognitive states robustly across different human subjects, these classifiers could be published on the internet as useful tools for the research community. In this sense, the trained classifiers would themselves constitute an operational theory relating cognitive state to fMRI observations.

Is this a feasible goal? In this paper we report on several experiments in which we have trained classifiers to successfully distinguish between a fixed set of possible cognitive states (e.g., “is the subject observing a word about food, or a word about people?”). Given the noise and resolution limitations of fMRI, there will surely be cognitive states that cannot be successfully distinguished based solely on fMRI. However, our results establish that there are cognitive states that can be successfully distinguished. We now face an empirical question of which states can be distinguished, as well as a machine learning research question of what learning algorithms will work best. The main purpose of this paper is to introduce the problem of training such classifiers of cognitive state, to report on the feasibility of training virtual sensors of cognitive states in some cases, to examine the subtleties involved in training such classifiers, and to suggest relevant future research issues.

In section 2 we provide a very brief overview of fMRI, focusing only the characteristics of fMRI most relevant to our goal of training virtual sensors of cognitive states. Section

3 reviews some previous related work. Section 4 describes the general approach we have explored for training classifiers, and Section 5 describes three case studies in which we have trained classifiers of cognitive states, including discussions of the datasets, machine learning methods, and experimental results. Finally, we summarize the lessons learned from these case studies, as well as suggestions for future research, in Section 6.

2 Functional Magnetic Resonance Imaging

fMRI is a technique for obtaining three-dimensional images related to activity in the brain throughout time, as the owner of the brain performs tasks inside an MRI scanner. More precisely, fMRI measures the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood with respect to a control baseline, at many individual locations within the brain. This is often referred to as the blood oxygen level dependent (BOLD) response. The BOLD response is taken as an indicator of neural activity. It should be noted that it acts as a temporally smeared indicator of activity, having a response spanning several seconds after the stimulus and corresponding neural activity occur.

An fMRI scanner displays the values of the fMRI signal at all the points in a three dimensional grid, or *image*), covering part of the brain. In the studies described in this paper, a three dimensional image is captured every 1, 1.5, or 0.5 seconds. We refer to the cells within an image as *voxels* (three dimensional pixels). Voxels typically have a volume of a few tens of cubic millimeters, and a typical three dimensional image typically contains 10,000 to 15,000 such voxels. While the spatial resolution of fMRI is dramatically better than that provided by earlier brain imaging methods, each voxel nevertheless contains on the order of hundreds of thousands of neurons.

3 Related Work

Over the past five years there has been a growing interest within the computer science community in data processing for fMRI. This has led, for example, to two Neural Information Processing Systems (NIPS) workshops on this topic, and several other meetings.

One popular kind of processing involves using a Generalized Linear Model (GLM) approach (Friston et al., 1995; Bly 2001), in which a regression is performed on the signal value at a voxel with respect to some stimulus. This can be used to determine which voxels are related to the stimulus. Others have used *t*-statistics to determine relevant active voxels, and yet others have used more complex statistical methods to estimate parameters of the BOLD response in the presence of signal noise (Genovese 1999).

Methods for modeling time series data have been used for fMRI data. For example, Hjen-Srensen et al. (1999) used Hidden Markov Models(HMM) to learn a model of activity in the visual cortex resulting from a flashing light stimulus. Although the program was not told the stimulus, the on-off stimulus was recovered as the hidden state by the HMM.

A variety of unsupervised learning methods have also been used for exploratory analysis of fMRI data. For example, Goutte et al. (1998) discussed the use of clustering methods for fMRI data. Penny (2001) applied Expectation Maximization to estimate mixture models to cluster the data. Others have used Principle Components Analysis and Independent Components Analysis (McKeown & Makeig, 2001) to determine spatial-temporal factors that can be linearly combined to reconstruct fMRI signal.

While there has been little work on our specific problem of training classifiers to decode cognitive states, there are several papers describing work with closely related goals. For example, Haxby et al. (2001) showed that there different patterns of fMRI activity generated when a human subject views a photograph of a face versus a house, versus a shoe, versus a chair. While they did not specifically use these discovered patterns to classify subsequent single-event data, they did report that by dividing the fMRI data for each photograph category into two samples, they could automatically match the data samples related to the same category. Others (Wagner et al, 1998) reported that they have been able to predict whether

a verbal experience will be remembered later, based on the magnitude of activity within certain parts of left prefrontal and temporal cortices during that experience.

4 Approach

In the fMRI studies considered here, data were collected from neurologically normal subjects from the university community. Data were preprocessed using the FIASCO program [?] ¹ to remove artifacts due to head motion, signal drift, and other sources. The resulting preprocessed images were used as input to our classifiers.

To characterize the locations of voxels within a brain, anatomical regions of interest (ROIs) were defined individually for each subject. These ROIs were defined using the parcellation scheme of Rademacher and his colleagues [?, ?]. For each subject, the mean of the functional images (i.e., images showing the BOLD response) was co-registered to the structural volume scan (i.e., a separate scan that captures physical brain structure rather than BOLD activation), so that these ROIs could be correctly identified.

4.1 Classifiers

In this paper we explore the use of machine learning methods to approximate classification functions of the following form

$$f : \text{fMRI-sequence}(t, n) \rightarrow \text{CognitiveState}$$

where $\text{fMRI-sequence}(t, n)$ is a sequence of fMRI images collected at times t through $t + n$, and where CognitiveState is the set of cognitive states to be discriminated.

We consider a number of classifiers for this purpose, including:

- *Gaussian Naive Bayes classifier (GNB)*. The Naive Bayes classifier assumes the features of fMRI-sequence (e.g., the individual voxel signals at different times) are conditionally independent given the CognitiveState . The Gaussian Naive Bayes classifier models the conditional probability $p(f|c)$ for each input feature f and each cognitive state c as a Gaussian distribution. Maximum likelihood estimates for the mean and variance of each such Gaussian are obtained from the training data. These estimates are then used, together with the naive Bayes independence assumption, to classify new instances according to the posterior probability of the class given the observed features.
- *Support Vector Machine (SVM)*. We used a Support Vector Machine with a linear kernel.
- *k Nearest Neighbor (KNN)*. We use k Nearest Neighbor with a Euclidean distance metric, considering values of 1, 3, and 5 for k .

To estimate the accuracy of trained classifiers, we generally employ leave-one-out cross-validation. When leaving out a particular image taken at time t , we generally also remove other images that are nearby in time. This step is designed to avoid the optimistically biased estimate that would otherwise arise, due to the similarity in activity among temporally proximate images (due to the temporal blur in fMRI BOLD response). In particular, when leaving out the fMRI image at time t for use as a test example, we also remove from the training set all images within 5 seconds of t .

4.2 Data Abstraction

We explored a variety of methods for preprocessing the raw fMRI images before providing them as input to the classifier. A typical image contains 10,000 to 15,000 voxels, and a training example can include several dozen images, so that some of our training and test examples contain hundreds of thousands of (voxel \times time) features. At the same time, our

¹FIASCO is available at <http://www.stat.cmu.edu/~fiasco>.

<i>Categories</i>	<i>Subcategories</i>
Animal	4-legged,Fish
Nature	Trees,Flowers
Food	Fruits,Vegetables
People	Family Members,Occupations
Artifact	Tools,Kitchen Items
Building	Dwellings,Building parts

Table 1: Categories of words in the Semantic Categories study. Classifiers were trained to distinguish which of the twelve subcategories of word the subject was viewing.

training example sets typically contain only a few dozen examples. Therefore, dimensionality reduction methods such as feature selection are very important for our problem. In the case studies described below, we explored a variety of methods for reducing dimensionality, involving combinations of four processes:

- *Selecting a subset of voxels.* Voxels were selected in some cases based on how well they could individually classify the training examples. In other cases they were selected based on how active they were when the subject processed any stimulus, compared to their activity when the subject was at rest.
- *Selecting a sub-interval in time.* In some experiments, the natural representation of an example is a sequence of images. This process selects a contiguous subinterval from this interval.
- *Averaging the values of multiple features.* One can reduce dimensionality by replacing a collection of voxel activations by their mean. In some cases we average voxels that occupy an anatomically defined contiguous region of the brain. In others we averaged voxels selected by alternative criteria such as the most active n voxels within a particular anatomical region.
- *Normalizing activities.* We experimented with a variety of methods for normalizing the activation values associated with different voxels (e.g., scale the activities so that each voxel achieves the same maximum and minimum value over some time interval).

5 Case Studies

This section describes our efforts to train virtual sensors of cognitive state for fMRI data collected in three distinct studies. In the semantic categories study, we train classifiers to determine which semantic category of word the subject is reading. In the syntactic ambiguity study we train classifiers to determine whether the subject is reading an ambiguous or unambiguous sentence. In the picture and sentence study, we train classifiers to determine whether the subject is viewing a picture or a sentence that describes a picture.

The following subsections describe each of these studies in turn, along with the classifiers developed for these studies and the lessons learned from each.

5.1 Semantic Categories Study

In this study, subjects were presented with words and asked to judge the semantic category to which each word belongs. Words are taken from six general semantic categories, with two subcategories of words within each, as depicted in Table 1.

Subjects were presented words one at a time, using a block design in which words from a single category are presented within each block (a 'block' is a contiguous interval in time). At the beginning of each block, the name of the semantic category is displayed for 2 seconds. After this the subject is shown a succession of 20 words, each presented for 400 msec and followed by 1200 msec of blank screen. After each word is presented the subject

must click a button in order to indicate whether the word belongs to the category that was named at the beginning of the block. This task serves to ensure the subject is performing a semantic judgment, and all subjects performed this task nearly perfectly. Nearly all words do belong to the target category (half the blocks contain no out-of-category words, and the remain blocks contain just one out-of-category word). Words were chosen from lists of high frequency words of each category, as given in [?], in order to avoid obscure or multiple-meaning words. A total of twelve blocks of words were presented, one block for each semantic subcategory. Between each block of words, a several second pause, or “fixation period,” was inserted. An fMRI image was acquired once per second.

5.1.1 Learning Problem

Our goal is to train a classifier that maps the observed fMRI activity of the subject to the semantic category that word the subject has just read². Our problem can be cast as the problem of learning a classifier function f where

$$f : \text{fMRI}(t) \rightarrow \text{WordCategory}(t)$$

Here, $\text{fMRI}(t)$ is the observed fMRI data at time t , WordCategory is the set of semantic subcategories in the study (“fish,” “vegetables” , etc.) and the value of $f(\text{fMRI}(t))$ is the word category that gave rise to the fMRI image obtained at time t . We represent $\text{fMRI}(t)$ by a vector of fMRI signal values observed at time t at selected voxels. Note that in this study, all trained classifiers use only a single image as input, and not images collected over an interval in time.

5.1.2 Methods

Classifier

In these experiments we use a Naive Bayes classifier, with normal distributions over each continuous input feature. The parameter estimates were obtained using maximum likelihood unbiased estimators for the mean and variance of a univariate normal distribution (see [?]).

This simple classifier is used so as to try to avoid the risk of overfitting, which is a considerable risk given that we have relatively few examples (a few tens per category) and a large number of features (a few thousand voxels). In addition, there is some support for the idea that a generative classifier may do better than a discriminative one for small samples ([?]). There is also empirical evidence that, in a few domains such as Text Learning ([?]), useful models can be learned despite an imbalance between the number of examples and features. Finally, we want the learned model to be simple enough to allow visual representation and interpretation by psychologists.

Evaluation

Given that our classification problem is a twelve-way classification, we measure classifier performance by the average rank of the correct label. More precisely, for each test example the classifier produces a ranking of all possible labels sorted according the posterior probabilities output by the naive Bayes classifier. The position of the correct label within this ranked list is then noted. This position is then averaged across all test examples and the ranking normalized to $[0,1]$, where 0 is first/correct and 1 is last. As an example, scores of .01,.05 and .1 in a 12 category situation correspond to average positions of 1.11,1.55 and 2.1 within the list of twelve categories. Note that guessing uniformly at random will produce an average rank of .5 on our $[0,1]$ scale, for this data set.

In order to judge how well a trained classifier can recognize cognitive states we use cross-validation estimates of the average rank of the correct label in all examples.

²In fact, the fMRI signal in response to a stimulus typically peaks several seconds after the stimulus, so strictly speaking we are learning a mapping from the fMRI signal at time t to the word category the subject observed several seconds earlier.

There are two issues to note with respect to the cross-validation method. First, when selecting examples for each test fold, we pick examples in approximately equal proportions from all the classes. The purpose of this is to maintain a balanced training set whose class priors fit the class proportions in the test set. Second, as mentioned earlier, whenever the program selects an example $\text{fMRI}(t_i)$ for inclusion in the test fold, it removes from the training set any examples that are within 5 seconds of t_i . This avoids optimistically biasing the evaluation due to the similarity of fMRI images that are nearby in time.

Feature Selection

In order to reduce the dimensionality of the input, and in an attempt to improve classifier accuracy, we considered three kinds of feature selection:

- *Select the n most active voxels (Activity p -value).* In this method, we select the voxels whose activation during reading differs the most from their activation when the subject is in a pause, or fixation period. More specifically, for each voxel, v , and each semantic word class, c , a t -test was applied to compare the mean signal level during blocks belonging to class c , versus fixation periods. The first twelve voxels were then selected by picking for each class c the voxel with the greatest t statistic. The next twelve voxels were selected by picking the second strongest voxel in each class, and so on.
- *Select most active voxels per Region of Interest (Activity p -value per ROI).* This is similar to the previous technique, but attempts to assure that voxels are selected from all ROIs within the brain. The above procedure is run on each ROI, and the union is taken of voxels selected from all ROIs. This ensures that no ROI is ignored, even if the activity level for voxels inside it is lower than that for voxels in other ROIs.
- *Select the n best individually discriminating voxels. (Feature Average Rank)* The n most discriminating voxels are selected by first training a classifier with just one voxel as input, repeating this procedure for each voxel. The n voxels with the best cross validated average rank are then selected. Note when using this feature selection method, we run feature selection separately for each cross-validation fold, in order to avoid use of data from the test fold during the feature selection process. Thus, the voxels selected may vary from fold to fold.
- *Select n voxels at random. (Random).* As a baseline, we also consider selecting features at random.

5.1.3 Experiments

Our experiments address the following questions:

1. Can a classifier learn the mapping between an fMRI image and the cognitive state, when trained using voxels from the entire cortex?
2. Can the results be improved by reducing the number of voxels considered by the classifier?
3. What are the features/voxels that most help a classifier be successful?
4. Are particular semantic categories more identifiable than others?

We use 32 images from each stimulus block as examples of the fMRI activity for the category word being shown during the block.

We performed experiments on data collected from 10 subjects, considering only those voxels found in the union of several anatomically demarcated regions of interest (ROIs). Prior to any feature selection, the section of cortex in question contained between 8470 and 11136 voxels, depending on the subject.

The tables presented contain the average rank of the correct category across all examples, for each of the 10 subjects.

<i>Experiment</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>
6 classes	0.45	0.52	0.4	0.5	0.42	0.38	0.52	0.35	0.50	0.29
12 classes	0.13	0.17	0.04	0.12	0.06	0.069	0.2	0.04	0.14	0.05

Table 2: Average rank for Semantic Category classifiers using all voxels as inputs. Each column indicates the average rank for a different human subject. The top row describes the classifier trained to distinguish the six abstract semantic categories. The second row describes the classifier trained to distinguish the twelve subcategories. Note random classification yields an expected average rank of 0.5, perfect classification a rank of 0.0, and worst case classification a rank of 1.0.

<i>Feature Selection Method</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>
Feature Average Rank	0.11	0.18	0.04	0.12	0.065	0.085	0.20	0.047	0.13	0.054
Activity p-value	0.049	0.042	0.024	0.07	0.029	0.032	0.049	0.02	0.078	0.013
Activity p-value per ROI	0.13	0.15	0.051	0.12	0.058	0.068	0.13	0.039	0.15	0.042
Random	0.14	0.18	0.065	0.13	0.072	0.082	0.22	0.052	0.15	0.059
Use all voxels	0.13	0.17	0.044	0.12	0.058	0.069	0.2	0.043	0.14	0.05

Table 3: Average ranks for classifiers using different feature selection strategies. Columns indicates different subjects, rows indicate different feature selection strategies. The final row represents using all voxels with no feature selection, and is equivalent to the 12 class row in Table 2.

Full Cortex Experiments

The Naive Bayes classifier was first trained using each voxel in the full cortex as an input, and using the six categories (not the twelve subcategories) as the target classes. The two subcategory blocks for each category we used as a pool of examples for that class. For several subjects the resulting classifier accuracies were somewhat better than those expected of a classifier predicting labels at random. However, when we trained a second classifier using the twelve subcategories as target classes, we obtained much greater accuracy. In fact, for five of the human subjects, the average rank of the correct word class in the twelve-class experiment was lower than 0.09, which means that on average the correct word category was ranked first or second out of the twelve categories. The results of both classifiers are compared in Table 2. Given the much higher accuracy for the twelve-category classifier, we use the twelve-category classification problem in all subsequent experiments reported here. Accuracies are measured using 32-fold cross-validation, which corresponds to using a single example of each class per test fold.

One might reasonably ask at this point how we can be sure that what the classifier is learning is in fact a discriminator for the semantic category of a word, and not an indicator of some other time-varying feature of the subject. We believe the preponderance of evidence supports our belief that the classifier is indeed learning to distinguish semantic categories, and we discuss this topic in greater detail at the end of this section.

Feature Selection Experiments

The four feature selection approaches described in section 5.1.2 were applied to the data, and the selected voxels were used to train new classifiers for the twelve class problem. Recall these feature selection methods are Feature Average Rank, Activity p-value and Activity p-value per ROI, and Random.

Each feature selection method (except Random) was used to select different numbers of voxels, and all were found to produce the best results when selecting between 1200 and 1600 voxels. We therefore report the average rank scores for classifiers selecting 1200 voxels according to each of the three methods, plus the Random method (see Table 3). Reported average ranks were obtained using 32-fold cross-validation.

Several trends are apparent from the results of Table 3. First, the most effective feature selection method is Activity p-value, and this method produces at minimum a factor of two reduction in the average rank cost function when compared against using all voxels, for each of the 10 subjects. Second, it is apparent that randomly selecting features is detrimental, and performs worse than simply using all available voxels with no feature selection.

5.1.4 Discussion

The above experiments indicate that one can train a classifier to predict the semantic category of word observed by a subject, based on the instantaneous fMRI activity produced while reading blocks of such words. Despite the very high dimension of the data, and the relatively small number of training examples, a naive Bayes classifier appears to perform reasonably well.

Automatic feature selection to select a subset of the voxels can further improve accuracy, reducing the average rank loss by a factor of two. Interestingly, selecting voxels by how good an individual discriminator they are (i.e., the Feature Average Rank method) barely produces an improvement over the baseline of using all voxels, whereas selecting them by how active they are (i.e., the Activity p-value method) works much better. We believe the reason for the weaker performance of the Feature Average Rank method may be related to the fact that we have a twelve-class classification problem.

What prompted us to consider individual voxel discriminability in the first place was the fact that highly discriminating voxels are often physically contiguous within the brain, even though the feature selection method is unaware of physical proximity. This clustering of highly discriminating voxels within the brain is visible with the naked eye (see Figure 1).

In most subjects, highly discriminating voxels concentrate in the inferior temporal (parts of Brodmann areas 20, 37), inferior extrastriate (parts of BA 20, 37, 37 and parcellation units OF, T3a, T3p, Oli and LG), temporal (parts of BA 21, 22 and 37) and parts of dorsolateral prefrontal cortex (parts of BA 6, 8, 9 and 46, anterior to and excluding frontal eye fields and posterior precentral sulcus) rois (areas as defined in [?]). For examples across three subjects, see panel (a) of Figure 1.

These locations make sense in the light of the results in [?], [?], [?] and [?]. These suggest that semantic object information is represented in distributed networks spanning ventral temporal cortex and lateral temporal cortex. In addition, they also remark that different semantic categories produce different spatial patterns of signal (that are consistent across exemplars in a given category). We were surprised by the presence of discriminability in inferior dorsolateral prefrontal cortex, and are in the process of seeking other reports of it in the literature.

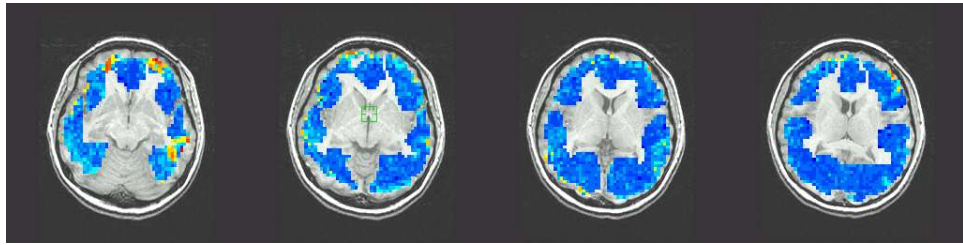
Interestingly, these were not necessarily the places where more activation was present. When doing feature selection, the subsets of active voxels and discriminating voxels overlapped by about 40% across subjects. The overlap is shown in panel (b) of Figure 1).

Given all of the above, what could justify the superior performance of the Activity p-value method for feature selection, compared to the Feature Average Rank method? Our current explanation can be illustrated by comparing voxels appearing in only one of these groups, and voxels common to both groups. This is depicted in Figure 2.

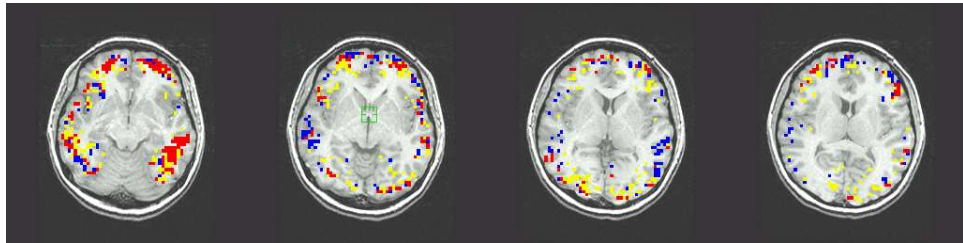
Voxels like (a) in Figure 2 mostly do not help or hinder classification, as all the category densities are roughly the same and hence contribute equally to the final $p(\text{category}|\text{voxels})$ that is used in predicting category.

Voxels like (b) exhibit different means for some of the semantic categories, but for several classes they exhibit strongly overlapping distributions. This could indicate that several classes get boosted simultaneously, one of which might be the correct one, thus the reasonable average rank scores in the group.

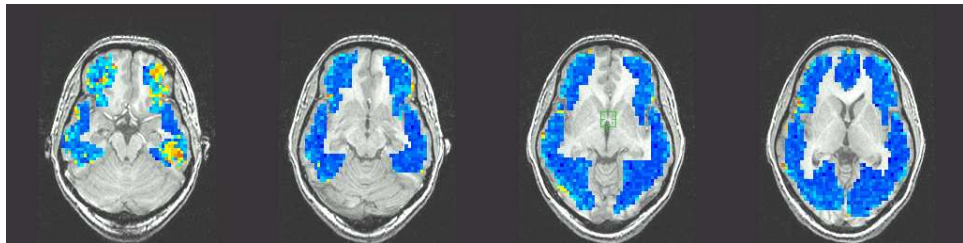
Voxels like (c), however, have a relatively clear separation between just a few of the categories. In this case, one of the categories has a particularly high mean signal level, another



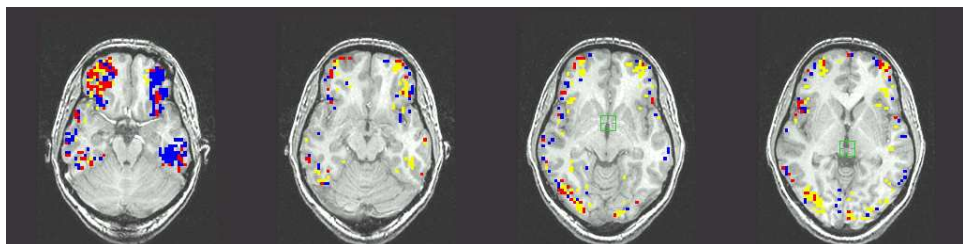
(a) Distribution of discriminating voxels for subject I (the more red the better)



(b) Overlap (red) of the active (yellow) and discriminating (blue) 1200 voxel subsets for sub. I



(c) Distribution of discriminating voxels for subject J (the more red the better)



(d) Overlap (red) of the active (yellow) and discriminating (blue) 1200 voxel subsets for sub. J

Figure 1: Panels (a) and (c) depict the overall distribution of Feature Activity Rank scores for each voxel, for two different subjects. Panels (b) and (d) compare two subsets of 1200 voxels, one containing voxels selected by activity level (using Activity p-value), the other containing discriminating voxels (selected by Feature Average Rank). There are 4 slices for each of 2 subjects, inferior to superior from left to right. The left hemisphere is on the right side of each image.

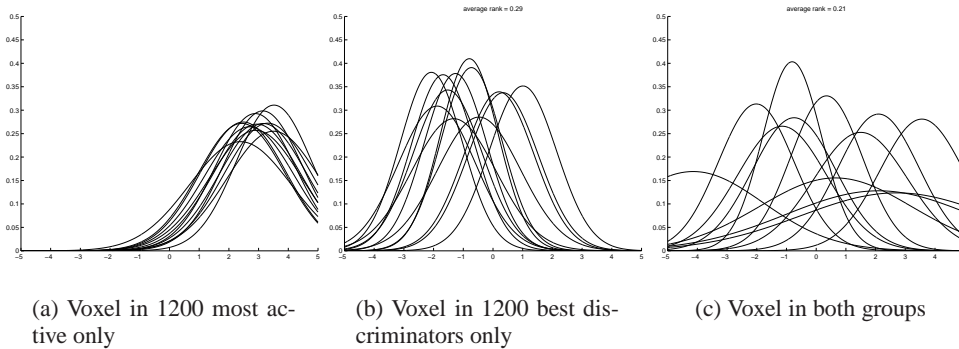


Figure 2: When learning signal distribution model for a particular category, we fit a Gaussian probability density function to the sample of signal values obtained at a given voxel during the block of that category. Each plot contains the 12 class-conditional densities for one voxel, the x axis is the percent signal above control. The leftmost voxel is selection by the Activity p-value method, the middle voxel is selected by Feature Average Rank, and the rightmost voxel is selected by both methods.

has a particularly low one, and others are in the middle. Being picked as an active voxel means that at least one of the classes has a high mean signal level when compared with the mean during fixation. This increases the chance of a separation.

Given this, we think that the good performance of the 1200 active voxel subset has two causes. First, the fact that the voxels are active for at least one condition (and often more) ensures that a few class densities are spread apart, which should increase the average rank. This should account for most of the overlap group voxels, which have good scores. The second cause is that the remaining active voxels, those not in the overlap group, do not greatly affect results, for better or for worse. Given that the class densities mostly overlap, these voxels just contribute equally to $p(\text{category}|\text{voxels})$ for all categories. In practice, it is as if they were not there. The 1200 discriminating voxel subset contains the same good discriminating voxels, but also many others whose decisions conflict with them (but are still reasonable discriminators).

Finally, we would like to address the question of what exactly is being learned by the classifier. We began with the goal of training a classifier to decode the semantic category of word the subject is considering. However, one limitation of the data we are analyzing here is that each of the twelve semantic categories also corresponds to a particular contiguous interval in time. Therefore, one must consider the possibility that what the classifier is actually learning is some other property of the subject that varies over time (i.e., perhaps the subject is becoming increasingly bored over time, and the classifier is learning this property; or perhaps there is a long-term signal drift in the fMRI data).

We consider it highly unlikely that the classifier is learning some time-varying property of the subject other than semantic word category, for several reasons: First, the consistency of accuracies across multiple human subjects suggests the classifier is learning a property common to all of these subjects. Second, the same brain regions appear to be involved in making the classification decisions across subjects, and these brain regions are consistent with earlier reported results in the neuroscience literature on semantic categorization. Third, in examining the behavior of individual voxels found by the classifier to be relevant, one can observe a response that is clearly a strong function of the semantic category block. One such voxel time series is depicted on the left side of Figure 3. However, a second voxel that does exhibit a longer term time drift is also included on the right side of the figure. Given the importance of this question, we are currently collecting data from additional subjects using a variation of the stimulus that will allow a more definite test of our conjecture.

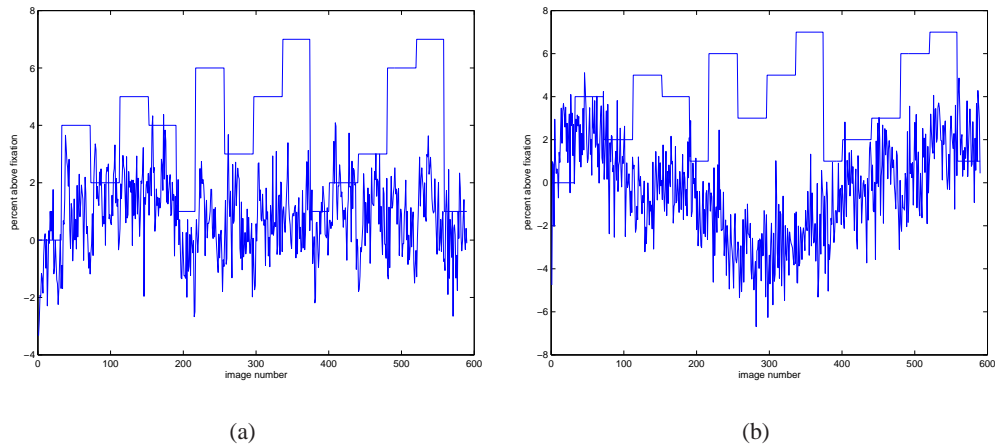


Figure 3: fMRI signal over time, for two voxels. The horizontal axis is time, and covers the entire twelve-block session. The straight lines superimposed on the activity plot indicate which of the six high-level semantic word categories (indicated as conditions 2 through 7) the subject was viewing at that time. The periods of rest, or fixation, are indicated by conditions 0 and 1.

5.2 Syntactic Ambiguity Study

In this study, subjects were presented with two types of ambiguous sentences and two types of unambiguous sentences, and were then asked to respond to a yes-no question about the content of the sentence. The questions were designed to ensure that the subject was in fact reading the sentences. An example of the four types of sentences is given below. This experiment is based on work by (Mason et al.,2002).

1. The experienced soldiers warned about the dangers conducted the midnight raid.
2. The experienced soldiers warned about the dangers before the midnight raid.
3. The experienced soldiers who were told about the dangers conducted the midnight raid.
4. The experienced soldiers spoke about the dangers before the midnight raid.

Ten sentences of each of the 4 types were presented to each of 6 subjects in the same quasi-random order (using a Latin square design). Each sentence was presented to the subject in the fMRI scanner for 10 seconds. Next a question was presented, which the subject was given 4 seconds to answer. After the subject answered the question, or 4 seconds elapsed, an "X" appeared on the screen for a 12 second rest period. The fMRI scanner collected 7 image planes from each subject every 1.5 seconds.

Sentences of 1 and 2 are considered ambiguous because there are two possible interpretations of the phrase "the experienced soldiers warned" when it is first encountered while reading the sentence. The sentences are ordered from most ambiguous to least ambiguous, and our experiments attempt to distinguish whether the subject is reading a question of type 1 or type 4. Note in this study there are only ten sentences of each type.

5.2.1 Methods

In this case study, we consider the following three experimental questions:

- Is it possible to distinguish whether a subject was looking at the most ambiguous or least ambiguous types of sentences based on the measured fMRI signal?
- What kind of feature selection achieves the highest accuracies for this task?

- How do the Naive Bayes classifier variance estimates affect the accuracy of the classifier?

We use a Naive Bayes classifier to address these three questions. We measure the performance of the classifiers based on the accuracy, or fraction of examples classified correctly, using leave-one-out cross-validation. As discussed later, we also investigate some variations on the Naive Bayes classifier.

Of the six subjects whose fMRI data we have for this study, one was eliminated because it had different trial lengths than the other five subjects, all of whose trial lengths matched. Across subjects, the LB (left Broca) ROI showed the highest activity relative to the rest of all the ROIs, and classifiers that were restricted to using only LB (or abstractions of it) performed better than on any other ROIs. Thus, we restricted all further data analysis experiments to use only data from this region of the brain, and investigated further feature selection questions by abstracting from this baseline. Classifiers achieved the highest and most consistent accuracies on Subject 1, and we focus mainly on this subject.

5.2.2 Experiment 1: Making the Distinction

The most fundamental experimental question we can ask about this dataset is whether or not sufficient evidence exists in the fMRI signal for training a classifier to distinguish between ambiguous and unambiguous sentences. We found that there is sufficient evidence for this task. Classification accuracies that were better than random were obtained for all five subjects, when appropriate time intervals and voxels were selected. We discuss the details of this feature selection in the following section. The best obtained accuracies ranged from a low of 65% for one subject to the high of 85% for Subject 1. In comparison, a default classifier that simply predicts the most common class would achieve an accuracy of only 47% in our leave-one-out experiments³.

5.2.3 Experiment 2: Feature Selection

In this case study, we experimented with varying three different parameters that define the representation of the input data for the classifier: 1) the time interval on which each example was defined, 2) the subset of voxels within the LB cortical region that were considered, and 3) whether the data from voxels were presented separately as classifier inputs, or whether the voxel activities were averaged into a single classifier input.

The time interval on which classifiers in this case study were trained was clearly an important aspect of feature selection. Each trial, which consists of the sentence presentation, question presentation, response, and rest period, lasted 24 seconds, or for 16 time points. Our experiments show that restricting a classifier's input data to time intervals including the images taken just after the midpoint of the trial (e.g. images 9 and 10) produces higher accuracies than intervals at the beginning or end of the trials. Figure 4(a) illustrates this trend for Subject 1. This trend is consistent across subjects, as shown by the corresponding trend for four subjects in Figure 4(b). In both figures, each horizontal line indicates a specific time interval, and its vertical coordinate indicates the accuracy achieved by a classifier trained on that interval.

Another type of feature selection that we looked at in this case study is the number of voxels in the LB region to use as inputs. Figure 5 shows plots of the accuracy of classifiers that use the n most active voxels either averaged or using their individual values. The level of activity of a voxel is calculated using a statistical t -test between fixation and non-fixation trials. Whether the voxel values are averaged or used separately, it appears that picking a relatively small number of voxels yields the best accuracies. This is not surprising given that we have a total of only 20 training examples.

The question of whether the voxels that are picked are averaged into one input for the classifier, or all taken separately is also important to feature selection. In the plots of

³Each leave-one-out fold will contain 9 examples from the same class as the test example, and 10 examples of the opposite class (9/19=.47).

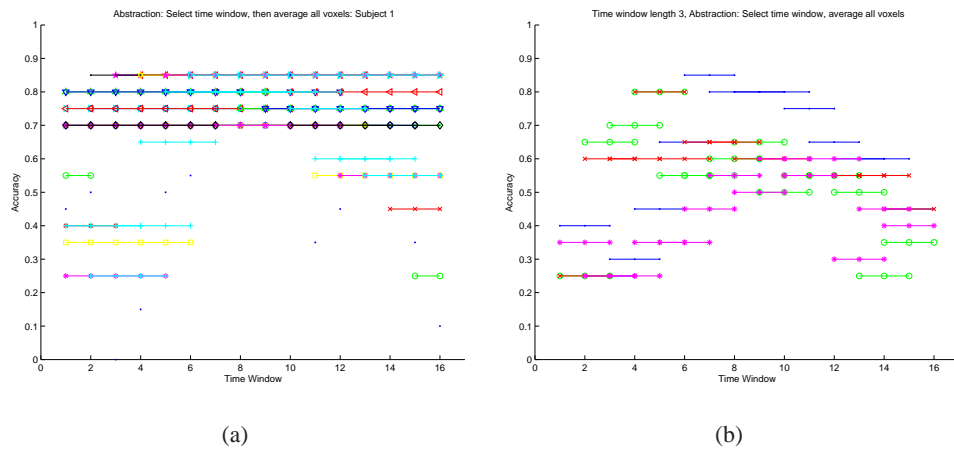


Figure 4: Dependence of classifier accuracy on the time interval over which the classifier input is defined. Figure (a) shows the dependencies for Subject 1 and Figure (b) for four subjects. Each colored horizontal line indicates a subinterval within the 16 image interval containing the full data for a single sentence-question-response trial. Note the higher accuracies for subintervals containing images 9 and 10, in both (a) and (b).

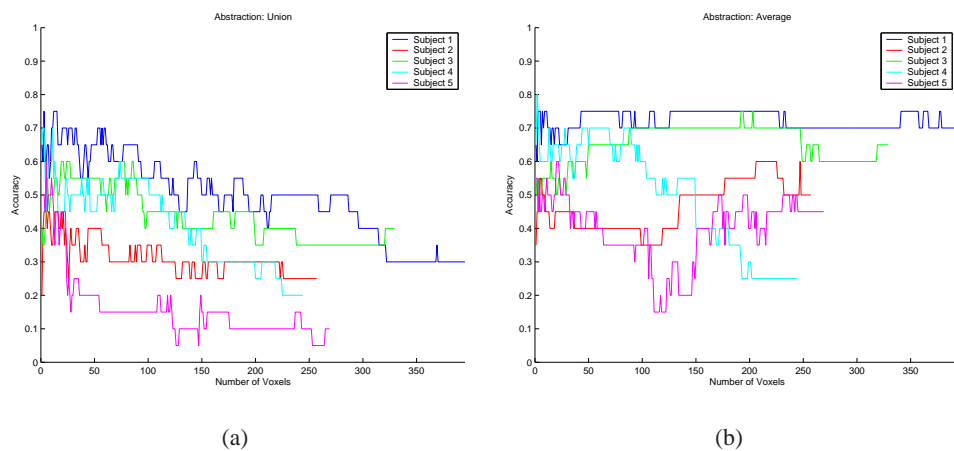


Figure 5: Dependence of classifier accuracy on the number of voxels used to define the input. Figure (a) shows the accuracy versus number of voxels for each of the five subjects, when each selected voxel is used as an input. Figure (b) shows accuracy for the same voxel sets when their activities are averaged to form a single input to the classifier.

<i>Subject</i>	1	2	3	4	5
<i>Accuracy1</i>	30%	25%	40%	20%	10%
<i>Accuracy2</i>	70%	15%	60%	65%	30%
<i>Variance over all voxels</i>	0.20788	0.22783	0.22767	0.30133	0.30275

Table 4: Impact of setting variances in GNB to fixed values. Accuracy1 is obtained by allowing the GNB to estimate the variance, and Accuracy2 is obtained by instead setting the variance to the variance over all voxels.

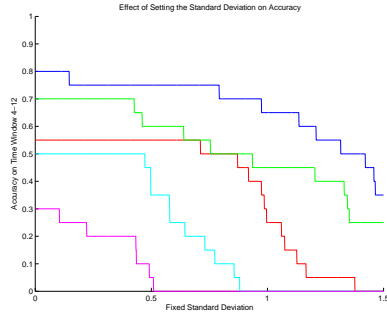


Figure 6: Dependence of Naive Bayes classifier accuracy on the value of the fixed variance. In all these experiments the variances in the Naive Bayes classifier were set to a constant. Each line shows for a single subject how the classifier accuracy (vertical axis) depends on the value of this constant (horizontal axis).

Figure 5, the maximum accuracies achieved for each subject using the averaging abstraction are greater than or equal to the maximum accuracies achieved taking all voxel values as separate inputs to the classifier. Our experiments show that averaging is also helpful when used in conjunction with restricting the time window to a subinterval as discussed above.

5.2.4 Experiment 3: Variance Estimates and Accuracy

At the outset of our experiments with this study, we measured the accuracy of classifiers trained on the entire LB ROI with no feature selection for each subject to obtain a baseline comparison for feature selection abstractions. The results were surprising: almost all the classifiers reported accuracies significantly worse than random. After some analysis, we conjectured that perhaps a Naive Bayes classifier using 10 or fewer examples, and using leave-one-out cross-validation estimates its variance parameter so badly that the variance estimates negatively impact the accuracies.

To test this conjecture, we tried fixing the variance parameters of all the Naive Bayes classifiers learned during leave-one-out to the variance of the data set with no feature selection, including all conditions (fixation trials and all four types of sentences). This change improved the accuracies of three of the five no-feature-selection classifiers to significantly better than random, as shown in Table 5.2.4.

Fixing the variance parameters of the classifiers also improved the leave-one-out accuracies with some types of feature selection. For example, when using the union abstraction, setting the variance to a fixed value produces an improvement in the maximum achievable accuracy in four of the five subjects.

To further investigate the relationship between the variance estimates and accuracies of the Naive Bayes classifiers, we ran an experiment varying the value of the fixed variance between 0.001 and 1.5, without regard to the actual variance of the data. The results of this experiment for averaging all voxels and cropping to the time window [4,12] are shown in Figure 6. The evidence in this plot suggests that when setting the variance to a fixed constant, lower values of the constant result in higher accuracy.

We experimented with fixing versus calculating the variance in several additional experi-

mental settings as well. The general trend we observed was that fixing the variance typically results in substantial improvement in accuracy. While the improvement varied from one experiment to another, the improvements ranged from none to 40%, and in no cases resulted in reduced accuracy.

5.2.5 Discussion

Our preliminary work on the syntactic ambiguity case study leads us to several conclusions and directions for future work. We conclude from the results presented here that the problem of classifying sentences as ambiguous or unambiguous based on fMRI data is learnable with proper feature selection techniques. Three important aspects of feature selection are choosing the time window within the trial, the subset of voxels to be used to define the classifier input, and whether or not to average the values of these voxels. We believe that the time window selected for this classification problem is especially important because of the cross-subject consistency that we have observed, and that it may be possible to make this feature selection decision based on multi-subject data even when the classifiers are trained for individual subjects. Finally, we are intrigued by the experimental results that we get by fixing the variance parameters of the classifiers to a fixed constant. We do not understand this phenomenon yet, and plan to investigate it in more depth.

5.3 Picture and Sentence Study

In this study subjects were presented trials in which they were shown a sentence and a simple picture, and asked whether the sentence correctly described the picture. We use this data to explore the feasibility of training a classifier to distinguish whether the subject is examining a sentence or a picture. We also explore the feasibility of training classifiers that can be used successfully across multiple human subjects.

In each trial, a sentence is presented for 4 seconds (e.g., “The *plus* sign is above the *star* sign.”). A blank screen is then presented for 4 seconds, and finally a picture is presented for 4 seconds. Pictures are geometric arrangements of the symbols +, * and/or \$, such as

$$\frac{+}{*}$$

When the picture is presented, the subject must press a button for “yes” or “no”, depending on whether the sentence matches the picture seen or not. The subject is instructed to rehearse the sentence until the picture is presented rather than try to visualize the sentence immediately. Finally, a rest or fixation period of 15 seconds is inserted before the next trial begins. Thus, each trial lasts approximately 27 seconds. fMRI images are collected every 500 msec, yielding a 54 time-point time course of images per trial. Each subject is presented a total of 20 trials, interspersed with five “fixation” trials. In a fixation trial the subject simply stares at a fixed point on the screen, providing a control condition. Of the 20 trials containing sentences, 10 have negated sentences (i.e., the sentence contains the phrase “does not contain”) and 10 have affirmative sentences.

In a second variant of this experiment, which we also use in this paper, the pictures are presented first within the trial and the sentences second. In this case, the subject is instructed to keep the picture in mind until the sentence is presented. We denote the data set in which sentences are presented before pictures as the *SP dataset* and the set in which pictures are presented before sentences as *PS dataset*.

Clean data were obtained from 13 neurologically normal subjects, and we use the data from all 13 subjects in the experiments reported below.

5.3.1 Learning Problem

We are interested in learning a classifier that takes as input an 8 second (16 image) interval of fMRI activity, and determines whether the subject is viewing a picture or sentence. In

other words, we wish to learn a classifier of the form

$$f : \text{fMRI-sequence}(t, 16) \rightarrow \text{Stimulus}$$

where Stimulus = {Picture, Sentence}.

5.3.2 Methods

Here we experiment with three classifiers: Naive Bayes classifier (GNB), k Nearest Neighbor (KNN), and a Support Vector Machine with a linear kernel (SVM).

Feature Selection, Abstraction, and Normalization

In this case study, we consider the following methods for selecting and abstracting the raw fMRI data. These methods operate by selecting and/or averaging voxels taken from individual ROIs. They include:

- *ROIselected.* This method selects 20 voxels from each ROI, according to the following procedure. First, all trials are averaged into a single trial, T, which represents the average time series for each voxel within the ROI. Second, each voxel is assigned a score which is the average of its activation over time within T. Third, the 20 voxels with the highest scores are selected to be used as inputs to the classifier.
- *ROIallAveraged.* This method simply averages all voxels within each ROI, to create a single “supervoxel” per ROI.
- *ROIselAveraged.* This method both selects voxels and averages their activities, producing a single “supervoxel” for each ROI. For each ROI, voxels are first chosen for each non-fixation condition, by selecting the five voxels with largest t -statistics when comparing this condition against fixation. Second, the union of these chosen voxels is averaged into the supervoxel.

We initially assumed that seven ROIs (CALC, LDLPFC, LIPL, LIPS, LOPER, LT, and LTRIA) were most probably involved in performing the task in this study. Our experiments revealed that some of these ROIs were more predictive and others less predictive given our classification task. We therefore also made some selections of ROIs by hand, based on these observations.

In contrast to the other case studies where we trained only subject-specific classifiers, in this study we are interested in training classifiers that work across multiple subjects. For this reason, abstracting the data is especially important. In addition, it is useful to normalize activity values across subjects. We adopted a normalization method in which the algorithm linearly rescales the data in each trial of each subject into $[0, 1]$, that is,

$$Y_t = \frac{X_t - \min_t X_t}{\max_t X_t - \min_t X_t} \quad t = 1, \dots, 32$$

where X_t 's and Y_t 's are the data before and after normalization, respectively.

5.3.3 Picture vs. Sentence Experiment for Single Subjects

We first consider training a classifier using either the *SP* data set or the *PS* dataset. In either case, we take the first 8 seconds of each trial as one example (i.e., a sentence in the case of the *SP* dataset), and the second 8 seconds as a second example (i.e., a picture in the case of the *SP* dataset). Thus, we have two examples per trial, times 20 trials, or 40 examples total. The task for the classifier is to classify an arbitrary 8 second interval taken from this set. Note the expected accuracy of a random classifier in this case is 50%.

In these experiments we trained kNN, SVM and GNB classifiers. We also tried all three of the feature selection and abstraction methods described above (note these methods select and/or average voxels by ROI, and we used the abstracted data from all seven voxels). In general we found significantly better than random classification accuracy for all 13 subjects.

Detailed results of this comparison of classifiers and abstraction methods are shown in Table 5. The data in this table is for one subject – the subject with the highest classification accuracy. In the next experiment we report the mean accuracies for all subjects. Notice that for this subject, the classifier achieves near perfect classification accuracy discriminating sentences from pictures in the *SP* dataset, and strong but less dramatic accuracy for the *PS* dataset. In general, SVM and GNB seem to outperform KNN (a trend we have also seen in other experiments not reported here). Among abstraction methods, ROIallAveraged and ROIselAveraged appear somewhat more effective than ROIselected.

In other variations of this experiment we found that high accuracies could also be achieved by using subsets of images from the 8 second interval that defines each example. In particular, the highest accuracy achieved using a single image was achieved using the image that follows stimulus onset by 5.5 seconds (which corresponds to the general rule of thumb that fMRI response typically peaks approximately 4 to 5 seconds following stimulus onset).

We also found in other experiments that normalizing the data, using the procedure defined in Section 5.3.2, could significantly improve accuracy. In several cases, classification error was reduced by one third due to normalization. Given this, the remaining experiments in this study were done using normalization.

<i>Classifier</i>	<i>Abstraction</i>	<i>SP</i>	<i>PS</i>
GNB	ROIallAveraged	100%	85%
GNB	ROIselected	97.5%	80%
GNB	ROIselAveraged	100%	85%
KNN	ROIselected	97.5%	62.5%
KNN	ROIselAveraged	97.5%	82.5%
SVM	ROIselected	100%	82.5%
SVM	ROIselAveraged	97.5%	95%

Table 5: Accuracies for Picture vs. Sentence for Subject 04847 using different classifiers and abstractions of the data. The expected accuracy of a default random classifier is 50%.

5.3.4 Picture vs. Sentence Experiment for Multiple Subjects

In this experiment we considered several questions, using the GNB classifier, the ROIallAverage data abstraction, and normalizing the data as described in Section 5.3.2.

First, we considered the question what is the average accuracy of classifiers trained for each of the 13 subjects, and how does this accuracy vary if we manually select subsets of the ROIs? The results are shown in the first three columns of Table 6. As can be seen there, the average accuracy of the 13 classifiers trained on each of the subjects is 84-92% for the *SP* dataset, and 75-82% on the *PS* dataset. It can also be seen that accuracies are slightly higher when using only 4 of the 7 ROIs.

The second question we considered is whether one can train a single classifier to be used on multiple subjects. First we tried training a single classifier based on the data from all 13 subjects. The leave-one-example-out accuracy for this 13-subject classifier is shown in the fourth column of Table 6. Note the accuracies of this single cross-subject classifier are comparable to the average accuracy of the 13 individual-subject classifiers in column 3.

Finally, we tried training a multi-subject classifier using only 12 of the 13 subjects, then testing the classifier on the 13th subject. The result of this leave-one-subject-out experiment is shown in the rightmost column of Table 6. The surprisingly positive result here is that the accuracy of this multi-subject classifier, when tested on new subjects outside the training set, is comparable to the average accuracy achieved when training on data from the test subject itself (as indicated in the third column).

<i>Dataset</i>	<i>Selected ROIs</i>	<i>Avg SingleSubj Accuracy</i>	<i>Leave-1-example-out</i>	<i>Leave-1-subject-out</i>
SP	7 ROIs	90%	88%	86%
SP	4 ROIs	92%	88%	88%
SP	1 ROI	84%	84%	83%
PS	7 ROIs	80%	80%	80%
PS	4 ROIs	82%	81%	81%
PS	1 ROI	75%	78%	77%

Table 6: Accuracies for multiple subject classifier. The third column shows the average accuracy of classifiers trained for single subjects. The fifth column shows the accuracy of a multi-subject classifier when applied to a subject withheld from the training set. Results are obtained using the ROIallAverage abstraction, normalization, and the GNB classifier. Here, 7 ROIs denote CALC, LDLPFC, LIPL, LIPS, LOPER, LT, and LTRIA. 4 ROIs denote CALC, LIPL, LIPS and LOPER. 1 ROI denotes CALC. The average accuracy denotes the mean value of all single subject accuracies.

5.3.5 Picture vs. Sentence Experiment across Different Contexts

In our final experiment, we explored the impact of merging data from the *SP* and the *PS* datasets. In this case we have a total of 40 trials and 80 examples. However, the examples now arise in differing temporal contexts. By “differing temporal context” we mean that half of the “picture” examples now occur immediately following a fixation period (the examples from the *PS* dataset), and half of them occur immediately following a sentence (the examples from the *SP* data set). These differing contexts are likely to make the classification problem more difficult for several reasons. First, the temporal smear of the fMRI response (it may take 8 or more seconds for the fMRI stimulus response to decay back to baseline levels) makes it likely that activity from the preceding event may still be apparent during the subsequent event. Second, the cognitive state of the subject, and the way they respond to a new sentence, may well be influenced when they have just seen a picture which they expect to compare to the upcoming sentence.

We trained a multi-subject classifier on the merged *SP* and *PS* data to explore the feasibility of training classifiers to work across such differing contexts. We used GNB, normalization, and the ROIallAverage abstraction in this experiment, as in the multi-subject experiments of the previous section. The results, shown in Table 7, show that the classifier in this case is less accurate than when trained on single-context *SP* or *PS* data sets. Nevertheless, the leave-one-subject-out accuracy for this multi-subject and multi-context classifier (75%) is still very significantly higher than the 50% one would expect from a random classifier.

<i>Selected ROIs</i>	<i>Leave-1-example-out</i>	<i>Leave-1-subject-out</i>
7 ROIs	70% (71%)	69% (70%)
4 ROIs	74% (75%)	73% (75%)
1 ROI	75% (75%)	75% (75%)

Table 7: Accuracies for multiple subjects across contexts (Using 13 subjects). The accuracies are based upon the average abstraction and got by GNB. The numbers in parenthesis are the accuracies after normalization mentioned in Section 5.3.2. The accuracy of a default classifier is 50%. Here, 7 ROIs denote CALC, LDLPFC, LIPL, LIPS, LOPER, LT, and LTRIA. 4 ROIs denote CALC, LIPL, LIPS and LOPER. 1 ROI denotes CALC.

5.3.6 Discussion

In this case study we have shown again that it is possible to train single-event classifiers that successfully distinguish between predefined cognitive states. We also found that normalizing the data was very useful in many of our experiments, and that SVM’s and GNB outperformed the KNN classifier.

Perhaps most significantly, we found that we could train multi-subject classifiers based on an abstraction of the data that is driven by summarizing activity within predefined ROIs. Note some abstraction is necessary, because the differences in physical size and shape of brains among subjects makes it extremely difficult to align voxels in one brain to the corresponding location in another. In this case study, the key seems to be using the anatomically defined ROIs as a basis for spatially abstracting the data.

We also conducted a variety of additional experiments, not reported here, attempting to train classifiers for other cognitive states such as “is the subject reading a negated or an affirmative sentence?” and “does the picture contain a + symbol?” While we obtained experimental results that were better than random in these two cases, the test sample sizes used to estimate these accuracies were small and it is therefore difficult to judge their statistical significance at this time. However, we do find the ROIs that support the most accurate predictions in these cases are LTRIA (widely thought to be involved in language processing), and LIPS (widely thought to be involved in spatial visualization). Thus, background knowledge from neuroscience argues in favor of the significance of these results. These experiments and others are described in (Niculescu, Wang, & Mitchell, 2002).

6 Summary and Conclusions

6.1 Training Classifiers to Detect Instantaneous Cognitive States

We began this paper with the question, “Is it feasible to train classifiers to decode the cognitive state of a human subject, based on single-episode fMRI data?” While much research has been done on analysis of *average* fMRI-measured behavior of various brain regions under different task conditions, we are unaware of other studies that have examined this question of classifying the *instantaneous, single-episode* cognitive states of subjects based on fMRI.

We have presented experimental results from three different fMRI studies, demonstrating the feasibility of training such classifiers to discriminate cognitive states such as, “Is this person examining a word describing fish or flowers?”, “Is this person reading an ambiguous or unambiguous sentence?”, and “Is this person examining a picture, or a sentence describing a picture?”. In all three case studies, we found it was possible to train classifiers to levels of accuracy significantly above that expected of random classifiers, given the size of test sets we used to evaluate the classifiers.

6.2 Training Classifiers Across Different Subjects and Contexts

In all three case studies, we found that it was feasible to train classifiers over individual subjects, and to recognize cognitive states occurring within a fairly restricted context. By “restricted context” we mean that the states to be distinguished (e.g., “examining an ambiguous versus unambiguous sentence”) occur in the context of a fixed temporal sequence (e.g., the sentence is preceded in both cases by a period of rest or fixation). Of course if we desire to use such classifiers in more general settings, it is natural to ask whether they can also be trained across more diverse contexts, and across subjects.

In the Picture and Sentence case study, we demonstrated that it is possible to train a classifier to operate over multiple contexts (i.e., to discriminate viewing of pictures versus sentences that are presented in sequence, regardless of which is presented first). We also showed that it is possible to train a classifier on one set of human subjects, then use this classifier successfully for single-episode classification of “picture” versus “sentence” in new subjects. As described above, an accuracy of 75% is achieved when classifying “picture” versus “sentence” in new subjects outside the training set, when the pictures and sentences appear in multiple contexts.

6.3 Feature Selection and Feature Abstraction

To achieve these results, we found it important to use methods for automatic feature selection and for abstracting the data. The most successful combinations of voxel selection methods, time interval selection methods, averaging values over multiple voxels, normalization of voxel activities, and using information about anatomically defined regions of interest, varied by case study. However, it seems clear that these and other methods for feature selection and abstraction are important given the very high dimensional data, and the relatively sparse training data available. When learning multi-subject classifiers rather than single-subject, we found that spatial abstraction according to anatomically defined regions was a successful approach.

6.4 Future Research

We foresee many opportunities for further research toward machine learning methods for decoding cognitive state from observed fMRI and other brain imaging data. Some arise from the unanswered questions raised by the experimental results we present here. For example, why do we get dramatically better performance in the semantic categories study when selecting voxels based on t -tests of activity, rather than based on discrimination capability? Or, why does accuracy improve dramatically in the syntactic ambiguity study when we assign a fixed variance rather than allowing the classifier to use a maximum likelihood estimate for variance? In addition, we see several broad directions that appear promising for research at the intersection of machine learning and fMRI-based studies of cognitive processes, including:

- *Learning temporal models.* One direction to expand this work is to move beyond training classifiers that operate on single time slices or single intervals, and more toward learning models of the *sequence* of cognitive states occurring over time. For example, we might conjecture that sentence processing involves a sequence of subprocesses including determining the main verb in the sentence, the subject, and the object. Approaches such as Hidden Markov Models and temporal Bayes nets seem relevant here, because they may offer a way to learn the mapping between observable fMRI data and hidden cognitive states, while at the same time learning the rules that govern transitions among these cognitive states, and using these rules to more accurately segment subsequent fMRI time series data.
- *Discovering useful intermediate abstractions of the data.* While we have explored a variety of ways of abstracting the fMRI data above the level of raw voxel activities, the work reported here is just scratches the surface of what is possible. Approaches such as Independent Components Analysis (ICA) and Principle Components Analysis (PCA) provide a very different way to re-express the high dimensional data in a lower dimensional representation. Beyond this, it might be useful to go beyond ICA and PCA, which seek representations that minimize reconstruction error, to develop methods that will seek abstract representations particularly useful for minimizing classification error (similar to the way in which the hidden layer of a neural network forms an intermediate data abstraction), or that seek abstract representations that best represent the signal from two different subjects.
- *Using prior knowledge in the form of proposed cognitive models.* The work we report here is based on using very little prior knowledge to guide the machine learning system. One direction for future research is to use models of human cognitive processing (e.g., the 4CAPS system of Just and colleagues, or the ACT-R model of Anderson and colleagues) as a basis for guiding the interpretation of the data, and as prior knowledge to guide the learning of classifiers. For example, the ACT-R model posits that problem solving involves short term imaginal memory, and that this involves activity in particular brain regions. When learning classifiers for cognitive states related to such problem solving, it should be helpful to use these kinds of assertions to generate priors to guide learning. More generally,

we may benefit by moving from tabula rasa learning of classifiers to instead fitting the data to ACT-R and 4CAPS models, then using this fit to refine these cognitive models.

- *Merging data from multiple sources.* While fMRI provides a useful window into human cognitive processes, it is nevertheless a limited window. One direction for future work is to combine fMRI data with data from other sources, including behavioral data (e.g., response times and error rates) and ERP. ERP is a technology for recording electrical signals from skin electrodes on the scalp, which provides much higher temporal resolution than fMRI, but more coarse spatial resolution. In principle, if one can develop methods that analyze data from fMRI together with ERP, one might gain the high spatial resolution of fMRI together with the high temporal resolution of ERP, while simultaneously obtaining measurements with uncorrelated noise.

Acknowledgments

We are grateful to Luis J. Barrios for helpful discussions and detailed comments on various drafts of this paper. Thanks to Vladimir Cherkassky and Joel Welling for the useful observations/suggestions during the course of this work. We thank Paul Bennett for many helpful discussions and for writing part of the code used for the Word Categories study.

Radu Stefan Niculescu was supported by a Graduate Fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University established by the Merck Company Foundation. Francisco Pereira was funded by a PRAXIS XXI scholarship from Fundação para a Ciência e Tecnologia, Portugal, and a PhD scholarship from Fundação Calouste Gulbenkian, Portugal. Rebecca Hutchinson was supported by an NSF Graduate Fellowship.

7 References

Aguirre, G.K., Zarahn, E. and D'Esposito, M. (1998). "An Area within Human Ventral Cortex Sensitive to Building Stimuli: Evidence and Implications". *Neuron*, 21, 373-383.

Battig, and Montague. "Category Norms for Verbal Items in 56 Categories". *Journal of Experimental Psychology*.

Bly, B.M. (2001). "When you have a General Linear Hammer, every fMRI time-series looks like independent identically distributed nails". Concepts and Methods in NeuroImaging workshop at NIPS*01.

Caviness, V.S.Jr., Kennedy, D.N., Bates, J., and Makris, N. (1996). "MRI-based parcellation of human neocortex: an anatomically specified method with estimate of reliability". *Journal of Cognitive Neuroscience*, 8, 566-588.

Chao, L., Haxby, J.V., and Martin, A. "Attribute-based Neural Substrates in Temporal Cortex for Perceiving and Knowing about Objects". *Nature Neuroscience*, 2, 913-919.

Chao, L., Weisberg, J., and Martin, A. (2002). "Experience-dependent Modulation of Category-related Cortical Activity". *Cerebral Cortex*, 12, 545-551.

Eddy, W., et al., The Challenge of Functional Magnetic Resonance Imaging, *Journal of Computational and Graphical Statistics*, Volume 8, Number 3, Page 545-558

Friston, K.J., Holmes, A. et al. (1995). "Statistical Parametric Maps in Functional Imaging: A General Linear Approach". *Human Brain Mapping*, 2, 189-210.

Friston, K.J., Holmes, A.P., Poline, J.-B., Grasby, P.J., Williams, S.C.R., Frackowiak, R. S.J., and Turner, R. (1995). "Analysis of fMRI time-series revisited". *NeuroImage*, 2, 45-53.

Genovese, C. (1999). "Statistical Inference in Functional Magnetic Resonance Imaging".

CMU Statistics Tech Report 674.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.A. and Hansen, L.K. (1998). "On clustering fMRI time series". Technical Report IMM-REP-1998-11.

Haxby, J., et al. (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex". *Science*, 293, 2425-2430.

Hjen-Srensen, P., Hansen, L.K., and Rasmussen, C.E. (1999). "Bayesian Modeling of fMRI Time Series". NIPS*99. Denver, November 29 - December 4, 1999.

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., and Haxby, J.V. (1999). "Distributed Representation of Objects in the Human Ventral Visual Pathway". *Proc. Nat. Acad. Sci. USA*, 96, 9379-9384.

Lazar, N., et al., Statistical Issues in fMRI for Brain Imaging, *International Statistical Review*(2001), Volume 69, Number 1, Page 105-127

Mason, R., Just, M., Keller, T., and Carpenter, P. (2002). Ambiguity in the Brain: What brain imaging reveals about the processing of syntactically ambiguous sentences, under review.

McKeown, M.J., Makeig S., et al. (2001). "Analysis of fMRI data by blind separation into independent spatial components". *Human Brain Mapping*, Vol. 6, No. 3, pp. 160-188.

Niculescu, S., Wang, X., and Mitchell, T.M., "Detecting Cognitive States Using Machine Learning" Internal report, 2002.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). "Text Classification from Labeled and Unlabeled Documents using EM". *Machine Learning*, No. 39, pp. 103-134.

Ng, A.Y., and Jordan, M. (2002). "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes". *Neural Information Processing Systems*, Vol. 14.

Penny, W. (2001). "Mixture Models with Adaptive Spatial Priors". *Concepts and Methods in NeuroImaging workshop at NIPS*01*. Vancouver, British Columbia, Canada, December 3 - 8, 2001.

Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., Rosen, B.R., and Buckner, R.L. (1998). "Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity". *Science*, 281, 1188-1191.

Rademacher, J., Galaburda, A.M., Kennedy, D.N., Filipek, P.A., and Caviness, V.S.Jr. (1992). "Human cerebral cortex: Localization, parcellation, and morphometry with magnetic resonance imaging". *Journal of Cognitive Neuroscience*, 4, 352-374.