Classifying Instantaneous Cognitive States from fMRI Data

Tom M. Mitchell PhD, Rebecca Hutchinson, Marcel Adam Just PhD, Radu S. Niculescu, Francisco Pereira, Xuerui Wang

Carnegie Mellon University, Computer Science, 5000 Forbes Avenue, Pittsburgh, PA 15213

{<firstname>.<lastname>}@ cmu.edu

ABSTRACT

We consider the problem of detecting the instantaneous cognitive state of a human subject based on their observed functional Magnetic Resonance Imaging (fMRI) data. Whereas fMRI has been widely used to determine average activation in different brain regions, our problem of automatically decoding instantaneous cognitive states has received little attention. This problem is relevant to diagnosing cognitive processes in neurologically normal and abnormal subjects. We describe a machine learning approach to this problem, and report on its successful use for discriminating cognitive states such as "observing a picture" versus "reading a sentence," and "reading a word about people" versus "reading a word about buildings."

1. INTRODUCTION

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. This fMRI technology has now been used 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, remembering). The vast majority of published research summarizes average fMRI responses when the subject responds to repeated stimuli of some type (e.g., reading sentences). The most common results of such studies are statements of the form "fMRI activity in brain region R is on average greater when performing task T than when in control condition C." Other results describe the effects of varying stimuli on activity, or correlations among activity in different brain regions. In all these cases the approach is to report descriptive statistics of effects averaged over multiple trials, and often over multiple voxels and/or multiple subjects.

In contrast, we propose here a qualitatively different way to utilize fMRI data, based on machine learning methods. Our goal is to automatically classify the instantaneous cognitive state of a human subject, given his/her observed fMRI activity at a single time instant or time interval. We describe here initial results, in which we have successfully trained classifiers to distinguish between instantaneous cognitive states such as "the subject is reading an ambiguous sentence" versus "the subject is reading an unambiguous sentence." Note that our objective differs in two ways from the objective of earlier statistical analyses of average effects: First, we are interested in learning the mapping from observed fMRI data to the subject's instantaneous mental state, instead of the mapping from a task to brain locations typically activated by this task. Second, we seek classifiers that must make decisions based on fMRI data from a single time instant or interval, rather than statements about average activity over many trials.

Why this is an interesting problem? Because if we could develop such classifiers they would provide a new tool for decoding and tracking the sequence of hidden cognitive states a subject passes through when performing some complex task, or for diagnosing their hidden difficulties in performing that task. Put succinctly, such classifiers would constitute *virtual sensors* of the subject's cognitive states, which could be used across a broad range of cognitive science research and diagnostic medical applications.

2. fMRI

An fMRI experiment produces time-series data that represent brain activity in a collection of 2D slices of the brain. Multiple 2D slices can be captured, forming a 3D image (the aggregate of all the slice planes) that may contain on the order of 15,000 voxels, each of which can measure the response of a 3x3x5-mm^3 region of the brain. Images of 15,000 voxels can be acquired at the rate of one or two per second with high field (3 Tesla) echoplanar imaging. The resulting fMRI time series thus provides a high-resolution 3D movie of the activation across a large fraction of the brain.

The actual "activation" we consider at each voxel is called the Blood Oxygen Level Dependent (BOLD) response, and reflects the ratio of oxygenated to de-oxygenated hemoglobin in the blood at the corresponding location in the brain. Neural activity in the brain leads indirectly to fluctuations in the blood oxygen level, which are measured as the BOLD response by the fMRI device. Despite the limitations in its spatial and temporal resolution, fMRI provides arguably the best view into activity across the human brain that is available today.

3. PRIOR WORK

The approach most commonly used to analyze fMRI data is to test hypotheses regarding the location of activation, based on regressing the signal on stimuli and behavioral information. One widely used package for doing so is SPM [2].

Haxby and colleagues [3] showed that different patterns of fMRI activity are 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 [9] 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

Our approach to classifying instantaneous cognitive states is based on a machine learning approach (see [6]), in which we train classifiers to predict the subject's cognitive state given their observed fMRI data. The trained classifier represents a function of the form:

$$f\text{: } fMRI(t,t+n) \rightarrow Y$$

where $fMRI(t,t+n)$ is the observed fMRI data during the interval from time t to $t+n$, Y is a finite set of cognitive states to be discriminated, and the value of $f(fMRI(t,t+n))$ is the classifier prediction regarding which cognitive state gave rise to the observed fMRI data $fMRI(t,t+n)$. The classifier is trained by providing examples of the above function (i.e., fMRI observations along with the known cognitive state of the subject). The input $fMRI(t,t+n)$ is represented as a feature vector, where each feature may correspond to the observed fMRI data at a specific voxel at a specific time. In some cases, we use features computed by averaging the fMRI activations over several voxels, or by selecting a subset of the available voxels and times.

Our learning method in these experiments was a Gaussian naïve Bayes (GNB) classifier (see, e.g., [8]). The GNB classifier uses the training data to estimate the probability distribution over fMRI observations given the subject is in cognitive state Y_i, $P(fMRI(t,t+n) \mid Y_i)$ for each cognitive state Y_i under consideration. To estimate this distribution, it assumes the features x_j describing $fMRI(t,t+n)$ are conditionally independent given Y_i, and thus models $P(fMRI(t,t+n) \mid Y_i)$ as the product over all features x_j of $P(x_j \mid Y_i)$. Each $P(x_j \mid Y_i)$ is modeled as a Gaussian distribution, whose mean and variance are estimated from the training data using maximum likelihood estimates. The GNB also estimates the class priors $P(Y_i)$ from the training data. New examples are classified using the learned $P(fMRI(t,t+n) \mid Y_i)$ along with Bayes rule to calculate the posterior probability $P(Y_i \mid fMRI(t,t+n))$ of cognitive state Y_i given the new observation $fMRI(t,t+n)$. The most probable Y_i is output as the classifier prediction.

5. EXPERIMENTAL RESULTS

Successfully training classifiers to decode cognitive states rests on two key assumptions: (a) the fMRI data contains sufficient information to distinguish interesting sets of cognitive states, and (b) machine learning algorithms can successfully learn the spatial-temporal fMRI patterns that distinguish these cognitive states. In this section we present experimental results indicating that both assumptions are warranted. In particular, we describe the use of machine learning methods to train classifiers for a variety of cognitive states, using data from four distinct fMRI studies. These results and data analysis methods are described in greater detail in [7].

5.1 Word Semantic Categories Study

In this fMRI study, ten human subjects were presented words one at a time, using a block design in which words from a single semantic category were presented within each block (a 'block' is a contiguous interval in time). The twelve categories of words presented were 'fish' 'four-legged animals,' 'trees,' 'flowers,' 'fruits,' 'vegetables,' 'family members,' 'occupations,' 'tools,' 'kitchen items,' 'dwellings,' and 'building parts.' At the beginning of each of the twelve blocks, the name of a category was displayed for 2 seconds, and the subject was then shown a sequence of 20 words, each presented for 400 msec followed by 1200 msec of blank screen. The subject was instructed to push a button after each word, to indicate whether the word belonged to the category named at the beginning of the block. At least 19 of the 20 words in each block were in the category -- the task was designed merely to ensure the subject attended to the meaning of each word. Between each block of words, a several second pause, or 'fixation period', was inserted. fMRI images were acquired once per second. We restricted our analysis to 30 anatomically defined regions of interest within the brain, yielding a total of 8,470 to 11,136 voxels, depending on the subject.

In this case we considered the task of learning to decode the semantic category of the word the subject was viewing, based on their instantaneous fMRI activity, i.e., learning the following classifier function:

$$f_s: fMRI_s(t) \rightarrow Y$$

where $fMRI_s(t)$ is the observed fMRI data for subject s at time t, Y is the set of 12 possible word categories, and the value of $f_s(fMRI_s(t))$ is the word category that gave rise to the fMRI image in subject s at time t. We trained a GNB classifier, representing the classifier input $fMRI_s(t)$ by the vector of fMRI signal values observed at time t at selected voxels.

We selected voxels based on their activity during word blocks compared to fixation (the periods between blocks of words). In particular, for each word category y_i and for each voxel v a t-test was conducted, comparing the activity of v during the stimulus y_i versus fixation. Voxels were then selected by choosing for each y_i the voxel with the largest t-statistic, removing these voxels from the pool, and then repeating this process until a total of 1200 voxels were selected.

We evaluated the performance of trained classifiers using standard leave-one-out cross validation, in which each fMRI image in the training set was held out in turn as a test example while training on remaining data. When holding out the image at time t as a test example, we also removed from the training data any image within 5 seconds of t, to avoid training on images likely to be correlated with the test image (due to the prolonged BOLD response).

For each test input, the trained classifier output a rank ordered list of the twelve word categories, from most to least probable according to its learned model. Classifier accuracy was measured by the percentile rank of the correct class in the output sorted list (perfect=1.0, random guessing=0.5, worst = 0). This accuracy measure for each of the ten trained classifiers (one per human subject) is as follows:

Subject	A	B	C	D	E	F	G	H	I	J
Accuracy	.89	.88	.96	.90	.94	.93	.89	.96	.87	.96

Note if the classifiers were guessing at random the expected accuracy would be 0.5. Given that these results were obtained independently for each subject, it is highly improbable that such high accuracies could have arisen by chance. One way to understand the trained classifiers is to examine which portions of the brain were used by the learned classifier to decode the word category. To achieve this, one can create a false-color brain image in which each voxel is colored by the accuracy of a classifier that is allowed to use only this single voxel. These false-color images are shown in Figure 1 for three of the ten subjects. Higher accuracy voxels are shown in darker color. Note the higher accuracy voxels cluster spatially, and that these clusters are found in similar regions across all three brains. The lower, larger cluster is in a location consistent with an earlier study of semantic categories reported by Haxby et al., [3] which found distinctive activation in a similar region when subjects were shown photographs (in contrast to our words) from different semantic categories such as faces, buildings, chairs, and shoes. The two upper clusters (located in inferior dorsolateral prefrontal cortex) are in a region not reported by Haxby's study.

The significant accuracy of these classifiers, the fact that highly discriminating voxels occur in similar brain regions across subjects, and the correspondence to Haxby's results on semantic categories all support the conclusion that our classifiers successfully learned to decode the semantic category of the word based on a single fMRI image.

5.2 Picture-Sentence Study

In this study (see [1] for a description of a similar task) 13 subjects were each presented 20 trials. In each trial they were shown a sentence (for 4 seconds, followed by a blank screen for 4 seconds), and a simple picture (for 4 seconds, followed by a blank screen for 4 seconds), and then answered (by pressing a mouse button) whether the sentence correctly described the picture. In 10 of these trials the sentence was presented first, in the other 10 the picture was presented first. Pictures were geometric arrangements of two of the symbols '*', '+', and '$', and sentences were descriptions such as "It is not true that the plus is below the star." Images were collected every 500 msec. We restricted our analysis to seven anatomically defined brain regions of interest containing a total of 1,397 to 2,864 voxels, depending on the subject.

We used this data set to explore the feasibility of training classifiers to distinguish whether the subject is visually examining a sentence or a picture, based on their fMRI activation. More precisely, for each subject s we trained a classifier of the form:

$$f_s: fMRI_s(t,t+8) \rightarrow \{P,S\}$$

where $fMRI_s(t,t+8)$ is the sequence of 16 observed fMRI images for subject s throughout the 8-second time interval $[t, t+8)$, and the target value for $f_s(fMRI_s(t,t+8))$ is P if the subject was viewing a picture during this interval, or S if the subject was viewing a sentence. In this study, we fix t to 1 or 9 so that the only

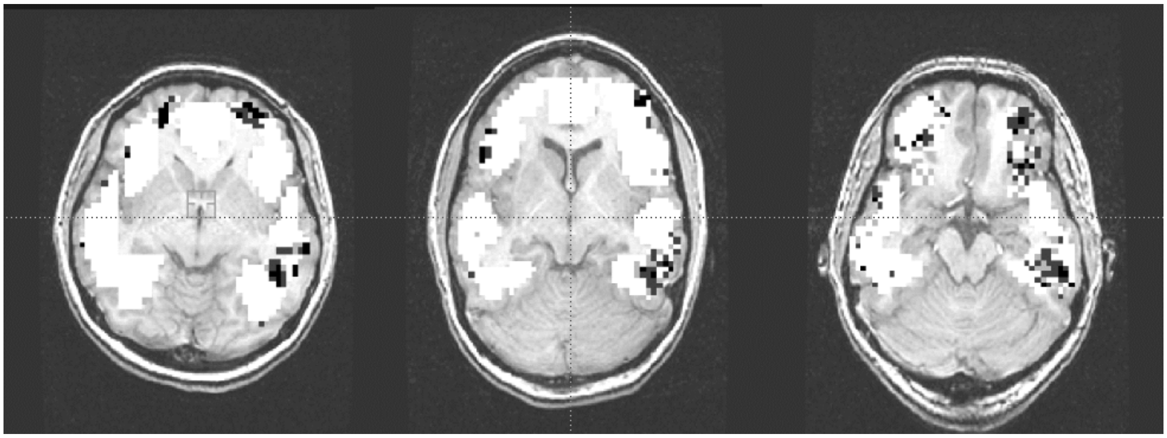

Figure 1. Plots show, for each of three human subjects, locations of voxels that best predict the semantic category of word read by the subject. Darkened voxels are those that produce the highest prediction accuracy; white voxels were also considered but were found to be less informative. Note the similar locations of predictive voxels across these three subjects.

time intervals considered are those that align with the 8-second intervals during which the subject was seeing either a picture or a sentence.

To train a GNB classifier for each subject, we first rescaled the data for each voxel to give it the same maximum and minimum activation across all trials, to compensate for variation in activity over time. We then selected a subset of voxels using the same voxel selection algorithm described in section 5.1. The values of each selected voxel for each of the 16 images were concatenated to form a feature vector describing the activity of these selected voxels over time. Hence, *n* selected voxels produce a feature vector of length *16 n.* This feature vector was used as the GNB classifier input.

We trained distinct classifiers for each of 13 subjects. The average accuracy of these 13 classifiers was .80 ±.034 when considering only trials in which the picture was presented before the sentence, and .90 ± 0.26 when considering only trials in which the sentence was presented before the picture. Here the intervals on accuracy reflect 95% confidence intervals calculated using a standard Binomial analysis (see, e.g., [6]). We also calculated the probability P_0 of achieving these accuracies by chance, under the null hypothesis that the classifier was guessing at random. The P_0 values in these two cases are 10^{-46} and 10^{-86} respectively, indicating that the classifiers are indeed learning predictive patterns over the fMRI data. The different accuracies in these two cases are presumably due to differing brain activity arising from the two contexts.

5.2.1 Multi-Subject Classifiers

We next considered training a single classifier to fit data from 12 of the subjects, then testing it on a 13th subject. A key technical difficulty when combining data from multiple subjects is that different brains have physically different shapes, making it unclear how to align voxels across subjects (e.g., see the brains illustrated in Figure 1). To address this issue, we spatially abstracted the voxel activities for each subject. In particular, we first manually marked up each brain by identifying seven anatomically-defined regions of interest (ROIs), such as the left temporal lobe. We then treated each of these seven ROI's as a very large "supervoxel," defining its activity to be the mean activity of the voxels it contained. The data for each subject was represented by the activity of these seven supervoxels over time, providing a common representation across subjects, at a considerable cost in spatial resolution.

We trained a multi-subject GNB for Picture vs. Sentence classification, using a leave-one-subject-out testing regime. Here each of the 13 subjects was used in turn as the test subject, while training on the remaining 12 subjects. The average accuracy of predictions for the left out subject was .81 ± .034 ($P_0 = 3.38 \times 10^{-49}$) for the picture-then-sentence data, and .88 ± .028 ($P_0 = 5.97 \times 10^{-76}$) for the sentence-then-picture data. Using the union of both data sets, the average accuracy was .75 ± .026 ($P_0 = 1.17 \times 10^{-71}$) over the left out subject. Again these results are highly significant compared to the null hypothesis that classification was by chance (expected .50 accuracy). Based on these results, it is clear that these classifiers discovered cross-subject regularities enabling them to distinguish cognitive states in new subjects, with accuracies rivaling those obtained when training single-subject classifiers. This is encouraging, given our goal of training virtual sensors of cognitive states that work across many subjects. Follow-on research is described in [10].

5.3 Additional Studies

We also trained classifiers using data from two additional studies (see [7] for details). In one study [5], subjects were asked to read sentences. Some of the sentences were ambiguous (e.g., "The experienced soldiers warned about the dangers conducted the midnight raid.") and others were unambiguous but of the same length. We trained GNB classifiers for single subjects, to decode whether the subject was reading an ambiguous or unambiguous sentence. Accuracies ranged from .65 to .80 for individual subjects, compared to .50 by random chance.

Another study involved showing individual words to subjects, and asking them to indicate whether the word was a noun or verb. We trained single-subject GNB classifiers to decode whether the word was perceived as a noun or verb, achieving accuracies from .75 to .81 for individual subjects.

6. SUMMARY AND CONCLUSIONS

The experimental results presented here demonstrate the feasibility of training classifiers to distinguish a variety of instantaneous cognitive states of human subjects based on their observed fMRI data. Subject-specific classifiers were successfully trained using data from four fMRI studies, and in one of these studies a classifier was trained to apply across multiple subjects, including subjects outside the training data. Interestingly, it was possible to train accurate classifiers despite the limited number of training examples available (for each subject, approximately 30 examples per class in the semantic categories study, and 40 per class in the sentence-picture study). While these classifiers were limited to distinguishing among a predefined set of cognitive states, and were applied only to a predefined time window of fMRI data, they provide encouraging evidence that additional research in this direction is warranted.

One direction for future research involves developing improved feature selection and classifier learning algorithms for single-subject and cross-subject classifiers, and testing these classifiers in additional settings. A second direction is to explore the applicability of this approach to diagnostic classification problems, such as early detection of Alzheimer's symptoms from fMRI data (perhaps in combination with structural MRI and other clinical data). In both of these applications, we see the opportunity for considerable improvement over the baseline algorithms reported here. For example, we would like to develop methods for simultaneously decoding networks of interrelated cognitive states such as the sequence of states involved in performing complex cognitive tasks, and for discovering new abstractions of fMRI data that can be used to combine data from multiple subjects.

Acknowledgements

We thank Luis Barrios for very helpful inputs throughout this research. Rebecca Hutchinson was supported by an NSF Graduate Fellowship. 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 and by National Science Foundation (NSF) grant no. CCR-0122581. Francisco Pereira was supported by a PRAXIS XXI fellowship from Fundação. para a Ciência e Tecnologia, (III Quadro Comunitário de Apoio, Fundo Social Europeu), a PhD fellowship from Fundação Calouste Gulbenkian, and by the Center for the Neural Basis of Cognition at Carnegie Mellon University and the University of Pittsburgh.

REFERENCES

[1] Carpenter, P., Just, M., Keller, T., Eddy, W., and Thulborn, K. "Time Course of fMRI-Activation in Language and Spatial Networks during Sentence Comprehension". *Neuroimage* 1999;10:216-224.

[2] Friston, K. et al. "Statistical Parametric Maps in Functional Imaging: A General Linear Approach". *Human Brain Mapping* 1995;2:189-210.

[3] Haxby, J., et al. "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex". *Science* 2001;293:2425-2430.

[4] Lazar, N., Eddy. W., Genovese. C., and Welling, J. "Statistical Issues in fMRI for Brian Imaging". *International Statistical Review* 1999;69:105-127.

[5] Mason, R.., Just, M.., Keller, T.., and Carpenter, P. "Ambiguity in the Brain: What Brain Imaging Reveals about the Processing of Syntactically Ambiguous Sentences". *Journal of Experimental Psychology: Learning, Memory, and Cognition* (in press).

[6] Mitchell, T. *Machine Learning*, McGraw Hill 1997.

[7] Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M. and Newman, S. "Learning to Decode Cognitive States from Brain Images", submitted to *Machine Learning*, April, 2003.

[8] Ng, A., and Jordan, M. "A Comparison of Logistic Regression and Naïve Bayes". *Neural Information Processing Systems* 2002;14.

[9] Wagner, A., Schacter, D., Rotte, M., Koutstaal, W., Maril, A., Dale, A., Rosen, B. and Buckner, R..."Building Memories: Remembering and Forgetting of Verbal Experiences as Predicted by Brain Activity", *Science* 1998;281:1188-1191.

[10] Wang, X., Hutchinson, R., and Mitchell, T., "Training fMRI Classifiers to Detect Cognitive States across Multiple Human Subjects". Submitted to *Neural Information Processing Systems* 2003.