

Inferring Cognition from fMRI Brain Images

Diego Sona¹, Sriharsha Veeramachaneni², Emanuele Olivetti¹, and Paolo Avesani¹

¹ FBK-irst, Trento, Italy

² Thomson R&D, MN, USA

Abstract. Over the last few years, functional Magnetic Resonance Imaging (fMRI) has emerged as a new and powerful method to map the cognitive states of a human subject to specific functional areas of the subject brain. Although fMRI has been widely used to determine average activation in different brain regions, the problem of automatically decoding the cognitive state from instantaneous brain activations has received little attention. In this paper, we study this prediction problem on a complex time-series dataset that relates fMRI data (brain images) with the corresponding cognitive states of the subjects while watching three 20 minute movies. This work describes the process we used to reduce the extremely high-dimensional feature space and a comparison of the models used for prediction. To solve the prediction task we explored a standard linear model frequently used by neuroscientists, as well as a *k-nearest neighbor* model, that now are the state-of-art in this area. Finally, we provide experimental evidence that non-linear models such as *multi-layer perceptron* and especially *recurrent neural networks* are significantly better.

1 Introduction

Thanks to the advent of functional Magnetic Resonance Imaging (fMRI), neuroscientists received impressive help in studying the functionalities of the human brain. This fMRI technology enables detailed analysis of neural activity by providing the means to collect brain activation data at high spatial and temporal resolution. Thanks to this, many studies identify regions of brain activated when humans perform specific cognitive tasks. Although traditionally fMRI scans have been used by neuroscientists to identify brain regions correlated with external conditions such as sense stimuli, there is burgeoning interest in adopting the reverse view, i.e., to use pattern recognition techniques and machine learning to predict external stimuli based on fMRI data [1].

Preliminary studies have shown that it is possible to decode visual perception cognition [2] looking at the brain state image acquired through an fMRI scan. The same approach is currently being extended to infer further high level cognitive functions [3]. The challenge of decoding mental states has been defined as a learning problem. The goal is to train a classifier that given an fMRI brain image, the mental state, predicts the associated cognitive state [4].

The machine learning community addressed the brain interpretation mainly using linear models and achieving controversial results. Previous works [5] indicate that *support vector machines (SVMs)* outperform *Gaussian naïve Bayes* and a *k-nearest neighbor classifiers (k-NN)*. On the other hand, more recently [6], it has been provided empirical evidence where a *k-NN* classifier outperformed a linear kernel model.

The above empirical analyses have been performed on datasets determined with cognitive experiments designed to address a single stimulus, i.e., the interpretation issue are just discriminative tasks between two alternative cognitive states. While this way of proceeding is effective from the point of view of neuroscientists that are looking for brain mapping, it affects the generality of empirical analysis on learning models since the evaluation is restricted to a single cognitive function.

In 2006, a team of neuroscientists organized a competition on decoding of mental states, the Pittsburgh Brain Activity Interpretation Competition (PBAIC)³. They provided fMRI data collected from three subjects while they were watching three 20 minutes movie segments from a television show. The subjects themselves later annotated the movies with respect to several ratings (e.g., language, attention, amusement etc.). The competition consisted in predicting the ratings of the third movie for all three subjects from the functional data, using the annotated ratings for the first two movies as training.

The contribution of this work is a report of our winning entry in the above competition, and an investigation of the use of non-linear learning model as feed forward neural networks and recurrent neural networks for the task of brain image interpretation. We provide empirical results on the dataset of PBAIC competition that provides the labeling of many different cognitive functions on the same brain scans.

In Section 2 we describe fMRI and feature ratings data in detail and define the prediction task. Section 3 is devoted to the description of our approach to preprocessing. In Section 4 we present the models used to predict the cognitive state of the subject from the brain images. Finally, Section 5 presents the result and in Section 6 conclusions are drawn.

2 Description of the Task

The task consists on the analysis of fMRI brain data of human subjects watching movie segments of a tv-series. The data was produced by the neuroscience group at the University of Pittsburgh for a competition held in 2006. Each movie segment is rated by the subjects themselves with multi-valued categories, such as the presence of faces, tools, sadness, arousal, individual actors, language, music, etc.⁴. The challenge is to interpret the brain activity of the human subject allowing predicting what he/she is experiencing.

³ <http://www.ebc.pitt.edu/2006/competition.html>

⁴ The subject cognitive experience is made of 13 feature ratings, 3 actor presence ratings and 3 location ratings.

In more detail, the same segments of movie were shown to three subjects (2 male and 1 female, with mean age of 26 years) that afterwards rated the movies with their personal impressions⁵. Both fMRI data and features ratings were sampled at a frame-rate of one frame every 1.75 seconds.

Each frame of the fMRI data is a 3-dimensional image made of 64x64x34 voxels. The intensity of the voxel represents the amount of blood arriving at the particular area of the brain measured using *blood-oxygen-level dependant fMRI contrast*. This is an indirect measure of the brain activity in the corresponding area. The image sequences were preprocessed for motion correction, slice time correction, linear trend removal, and spatial normalization. Each brain image has an associated vector of ratings provided by the human subjects (the target of our task). These ratings were temporally convolved with a hemodynamic filter that makes these features real-valued⁶. The features scored in the competition are of two types:

- Content:** body parts, environmental sounds, faces, food, language, laughter, motion, music, and tools;
- Reaction:** amusement, attention, arousal, and sadness;

but there were also other optional scores which were, actors and locations. We refer the reader to [7] for a more detailed description of the task and the data.

For the remaining part of the paper, we evaluated the proposed models using a subset of the available features: Amusement, Body parts, Faces, Language, and Motion. The reason is that for these ratings the evaluation of the quality of the used models is more robust. All the discarded ratings are characterized by the absence of a significant number of positive samples, hence the model evaluation are not as robust as for those ratings well sampled that we selected.

3 Image Processing and Dimensionality Reduction

The fMRI data and the corresponding ratings that we used for the experiments in the present paper are made of two movies collected in a total time span of 40 minutes at the frame-rate of 1.75 seconds. Hence, the sum of the temporal sequences is made of about 10^3 brain images and rating samples. From these sequences, we have to select training and testing sets. In addition, each of these 3-D brain images are extremely noisy and high dimensional (10^5 voxels). This suggests that these images need to be reduced in their dimensionality both from a computational standpoint as well as to alleviate the loss of prediction accuracy due to the *curse of dimensionality*. In the remainder of this section, we describe

⁵ The ratings given by the subjects resulted to have high discrepancies, i.e., the correlations between the ratings of different subjects were sometimes very low.

⁶ The reason for this convolution is the need for temporal realignment of the fMRI sequences with the events in the movies. The blood flow increases in the interested brain area few seconds after the area activity. This delay is well studied and the hemodynamic filter compensates this delay.

the process of image feature attribute generation. The main idea is to select the most informative voxels in the brain and then to cluster the voxels with similar behavior in time. Each cluster then is used to extract one image feature attribute.

3.1 Noise Removal

We observed that the variation of intensity of a voxel over time has much higher frequency components than the feature ratings. Therefore, the first step in the preprocessing phase consists on the filtering of high frequency noise for all voxels in the brain images by a low pass filter, i.e., a moving average window. The dimension of the moving window was 5 time steps. The reason for the choice of this window size is that it preserves the important part of the signal without losing those frequencies that also appear on the ratings (see Fig. 1 for an example of signal smoothing).

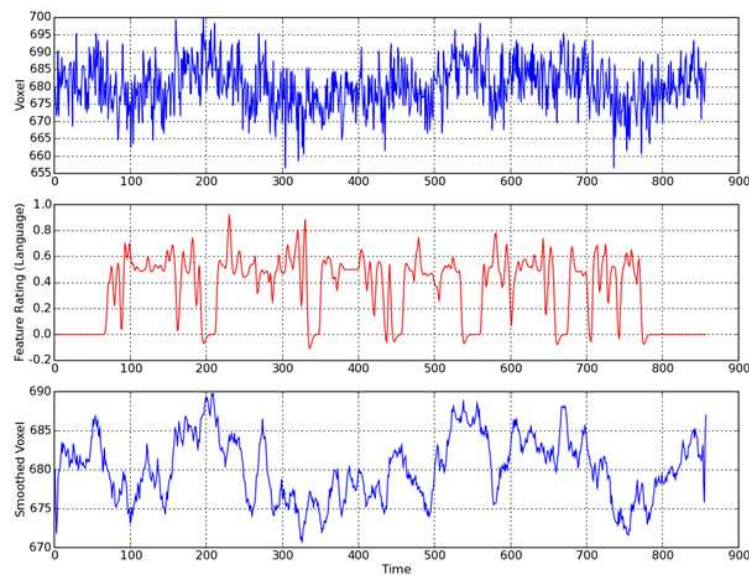


Fig. 1. The top graph describe the temporal behavior of a voxel randomly selected in the brain. The middle graph describes the temporal trend of the Language feature rating. The bottom graph is the result of low-pass filtering applied to the top graph.

3.2 Feature Selection

The second preprocessing step is the selection of the most informative voxels in order to discard all the voxels not informative for the task. We adopted the mutual information measure to evaluate the informativeness of a voxel with respect

to the desired target. The feature extraction was conducted as follows. For each feature rating, we found its mutual information to the value at every voxel in the image⁷, separately for every subject. For each subject and feature rating, we ranked the voxels according to their mutual information and we selected the best 10% voxels obtaining a subject-feature mask of approximately 10^4 voxels. Figure 2 shows an example of the values of mutual information computed for the voxels in some slices of a brain image.

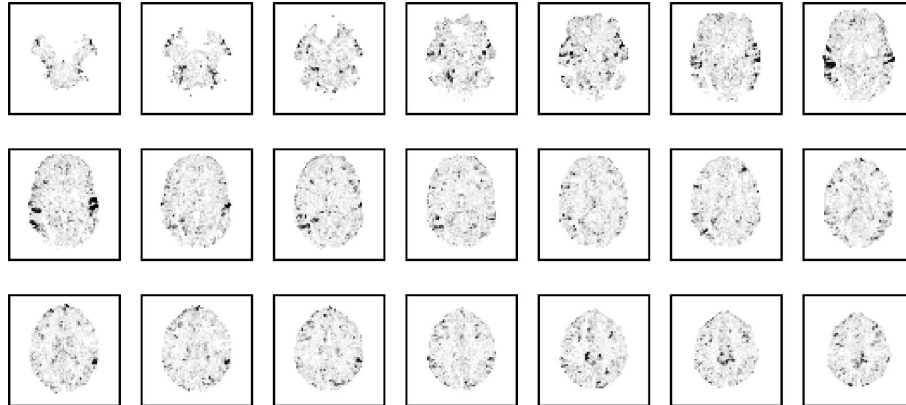


Fig. 2. Mutual information of some slices of brain image of subject 1 against Language feature ratings. The scale goes from 0.0 (bright voxels) to 0.4 (dark voxels). Notice the evident highlight of the hearing center in the temporal lobe of both left and right hemispheres

3.3 Features Extraction with Voxel Clustering

The third step in the preprocessing phase consisted in a further reduction of the dimension of the brain representation from 10^4 to 200 attributes. This reduction is obtained by grouping the voxels into a smaller set of representatives. To obtain these representatives we clustered the voxels with a simple *k-means* algorithm. However, in the neuroscience community there is still a debate about whether the cognitive processes are centered in a few specific and well-organized areas of the brain, or are distributed in many smaller and sparse agglomerates of neurons. Hence, we decided to adopt a measure of similarity taking into account these

⁷ To compute mutual information between the time-course of each voxel and a given feature rating we quantized both signals (50 steps for voxels and 16 steps for feature rating) and estimated probabilities with the help of Laplace smoothing. The choice of the quantization grids has been motivated by two opposite factors: representing signals without losing relevant information and reducing estimation problems.

two possibilities. Our clustering algorithm therefore is designed to group voxels which are both near in space and similar in the temporal trend. This is obtained adopting a distance measure able to combine these two kinds of information:

$$d = d_{spatial}^{\alpha}(1 - r_{temporal})^{1-\alpha}, \quad (1)$$

which is a geometric mean of a standard Euclidean distance ($d_{spatial}$) that uses the 3D coordinates in the brain, and the Pearson’s correlation over time ($r_{temporal}$). The weighting factor α can be used to give priority to space or correlation. With $\alpha = 1$ only the Euclidean distance is used, hence a spatial clustering is performed. On the contrary, with $\alpha = 0$ only correlation is used, performing in this way a sort of temporal clustering. In our experiments we gave the same importance to spatial distance and correlation (i.e., $\alpha = 0.5$).

In both movies there are some scenes lasting for a few seconds each, where nothing was projected on the screen⁸. Since the location of these blank sections was known, we decided to remove these parts from the brain image sequences while computing the correlation. This was done to eliminate noise due to the possibly random states of the brain during these blank sections.

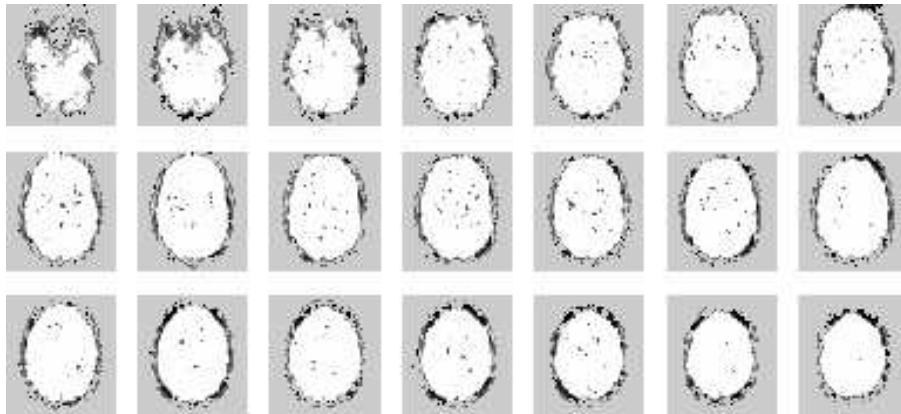


Fig. 3. Example of an important cluster (with approximately 300 voxels). The voxels are organized into sparse small agglomerates.

Each cluster is therefore a spatio-temporally proximal set of “informative” voxels. Figure 3 shows an example of a cluster of voxels determined with the above processing. The example shows that the voxels in the cluster are sparse agglomerates in the brain. At the end, the representation of the brain at a given time instance, i.e., an fMRI volume, was reduced to 200 features. These features

⁸ These “blank” parts were used during data collection to refine the calibration of the MRI machine.

were computed as the average intensity of all the voxels in the image associated to the corresponding cluster (Figure 4 shows an example of how the voxels in a cluster participate to the creation of a unique average feature). We constructed the feature attribute datasets for each subject and feature rating.

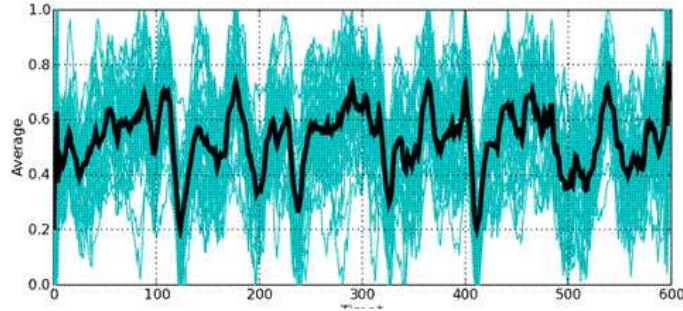


Fig. 4. The dark black line is the average behavior of all the voxels belonging to a cluster.

After preprocessing, the data was normalized to improve the quality of the models. We decided to perform a linear scaling of the feature ratings to the interval $[0, 1]$. The brain data was, instead, normalized according to mean and variance⁹. For each of the 200 features, the mean was forced to 0 and the variance to 1.

4 The Prediction Models

As previously mentioned our intention was to let the “data do the talking”, i.e., to let the data determine whether it was possible to find useful dependencies between brain activation and cognitive states. Therefore, excluding the spatio-temporal assumption made for “data compression” during clustering, we made no any other particular assumption on the distribution of data or on the dependencies between brain activation and stimuli. Starting from this hypothesis, we did not know whether there exists any linear or non-linear dependence between the brain activations and the desired output. For this reason, we preferred to select a non-linear model. Clearly, the drawback was that a bad choice of the model complexity (i.e. the number of free parameters) can cause a severe overfitting of the model on the given data if the data is not as complex as the model. Our choice went to Neural Networks, even considering that these models cannot help to identify the regions of interest in the brain (regions devoted to specific cognition tasks).

⁹ All 200 features were normalized with $x' = \frac{x - \bar{x}}{\sigma}$.

Moreover, in the data there are at least two kinds of temporal dependencies inherent in the above brain activation sequences. The first kind of dependence is what we like to call “latency” (or “inertia” of the scenes). Since the subjects are watching a video, there are few drastic changes in the scene from one frame to the next. We can assume that the features appearing in a scene of the movie (music, a face, a location, some food, etc.) have some persistence, i.e., they will last for a certain time in the movie. Hence, the “instantaneous” forecasting given the current input can be reinforced by the hypothesis made in the past scenes.

The second kind of dependence encoded within the sequence can be referred to as “adaptivity” of the brain. When a stimulus or a combination of stimuli arrives to a subject, his/her brain is activated in certain locations according to the cognitive process. These locations and the strength of activity, however, change in time according to changes in the cognitive process. For example, the first time that a subject sees an actor in a movie his/her brain may be involved in solving several cognitive problems (identifying who is the actor, comparing the appearance with respect to his/her memory of past movies, memorizing the face if unknown, and many other specific unconscious activities). As far the actor continuously appears in the movie, the cognitive efforts change in time, maybe reducing to just recognizing the character. The idea is that the brain activation can slowly change in time for the same stimulus.

Both the above-described varieties of dependence motivate a temporal analysis of the data. Hence, we believe *recurrent neural networks* (*RNN*) to be appropriate models both because of the ease with which time dependence can be modeled as well as their ease of learning or estimation.

As comparison versus *RNN* we also tested other linear and non-linear models. The first obvious choice was to compare the *RNN* versus the *multi-layer perceptron* (*MLP*). This was necessary to see whether the hypotheses on temporal dependencies were sustained by experiments. Then we decided to evaluate also a very simple *general linear model* (*GLM*) described by linear equations:

$$Y = XA \tag{2}$$

where the parameters A are determined with an ordinary least squares estimation. The reason for this evaluation is that this is a reference model in the neuroscience community; hence, it was quite natural to consider it as a baseline.

The other model we tested is a *k-NN* that, together with *SVMs*, is considered the state-of-art for the current application. Many authors showed that for the current task *k-NN* frequently gives better results than *SVMs*.

4.1 Experimental Setting

We used the data of the two movies both for training and for testing adopting a cross-validation approach. Since each movie is composed of segments separated by intervals of blank video, we used these blank intervals to split the two movies into 12 consistent segments (6 for each movie). Each of these segments is made

of approximately 100 temporally ordered samples. Then we performed the leave-one-segment-out training on all possible combinations of 11 segments, iterating the test set over all the 12 segments. Each feature rating was predicted separately with preprocessing of data and training of the model performed separately for that rating.

Independently of the model one of the major problems of the neural networks is the choice of the network topology (i.e., the number of free parameters). During our experiments we observed that very few hidden units were usually enough to create overfitting problems both for *RNN* and *MLP*. Hence, to avoid overfitting we decide to adopt cross-validation as stopping criterion. From the 11 training segments, we were holding-out 2 randomly selected segments used to stop the back-prop training algorithm on the remaining 9 segments. In particular *RNNs* were trained with a standard *back-propagation through time*. Both networks were using the hyperbolic tangent output function for hidden units and the logistic output function for the output unit. The recurrences in *RNNs* were only on the hidden layer. Weights were randomly initialized in the interval $[-0.05, 0.05]$ and updated with a dynamic learning parameter and with moment.

The evaluation of the results was done using the Pearson’s correlation between estimated and real feature ratings as suggested by the competition guidelines. The reason for this measure of correctness is that we have sequences of real values to compare. In this case, there is not any kind of correctness but just similarity; hence, it was not possible to determine a standard precision/error evaluation. The *NNs* models, due to their indeterministic behavior, were experimented with 5 trials and the results were then averaged.

5 Results and Discussion

Both *MLP* and *RNN* were tested with 4 hidden units, still having severe problems of overfitting. Regarding *k-NN*, since this is a regression problem, we have seen that the best solution was to average the ratings of the most similar *k* brain activations. Almost all studies of *k-NN* applied to this task are designed on controlled experiments, where the subjects are repeatedly presented with a controlled set of stimuli (usually positive and negative). In this task, however, there is not control on the sequence and the combination of stimuli, because they are consequence of a real experience (watching a movie). Hence, a lot of noise appears in the brain signals. For this reason, the best solution was to smooth the noise averaging the best *k* ranked elements. The average was weighted with values inversely proportional to the corresponding Euclidean distances. In particular, we discovered that for all features the best results were with $115 \leq K \leq 120$. We chosen $K = 118$.

In Tab. 1 are shown the average results of the different models for the 5 different feature ratings. Apparently, non-linear models are always better than linear models. Moreover, the exploitation of temporal autocorrelation, as expected, gives a little improvement in the quality of results.

Table 1. The results of the 4 models over all feature rating and their average.

	Amusement	Body parts	Faces	Language	Motion	Average
<i>GLM</i>	0.209	0.327	0.311	0.426	0.381	0.331
<i>k-NN</i>	0.087	0.334	0.394	0.439	0.446	0.340
<i>MLP</i>	0.285	0.432	0.468	0.605	0.506	0.459
<i>RNN</i>	0.306	0.446	0.480	0.621	0.543	0.479

6 Conclusions

The PBAIC team provided an extremely rich data set that includes complex spatial and temporal dependencies among brain voxels and cognitive states. For the first time it was possible to evaluate the performance of a learning model across many and cognitive functions. Our preliminary study shows that non-linear models as feed-forward and recurrent neural networks are effective in dealing with the complex task of decoding brain states as acquired by fMRI scan. The behavior has been shown quite stable with respect to different classes of cognitive tasks.

The slight enhancement introduced by the recurrent neural networks suggests that the extraction of relational knowledge, both at temporal and spatial level, remains an open challenge. Anyway, it is still not clear whether the temporal response of the brain is different with respect to concurrent stimuli rather than a single stimulus.

References

1. Editorial, B.: What's on your mind. *Nature Neuroscience* **9**(8) (2006)
2. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* **8**(5) (2005) 679–685
3. Haynes, J.D., Rees, G.: Decoding mental states from brain activity in humans. *Nature Neuroscience* **7**(7) (2006)
4. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to decode cognitive states from brain images. *Machine Learning* **9**(8) (2004)
5. Wang, X., Hutchinson, R., Mitchell, T.M.: Training fmri classifiers to detect cognitive states across multiple human subjects. In: *International Conference on Neural Information Processing Systems Foundation*. (2003)
6. Vishwajeet Singh, K.P. Miyapuram, R.S.B.: Detection of cognitive states from fmri data using machine learning techniques. In: *Proceedings of Twentieth International Conference on Artificial Intelligence*. (2007) 587–592
7. Schneider, W., Siegle, G.: Pittsburgh brain activity interpretation competition guidebook. <http://www.ebc.pitt.edu/2006/competition.html> (2006)