Recognizing Deep Grammatical Information during Reading from Event Related fMRI

Haim Shalelashvili, Tali Bitan, Alex Frid, Hananel Hazan, Stav Hertz, Yael Weiss, Larry M. Manevitz
University of Haifa,
Israel
haimshalev@gmail.com, tbitan@research.haifa.ac.il, alex.frid@gmail.com, hananel@hazan.org.il, stavhertz@gmail.com, ylweiss@gmail.com, manevitz@cs.haifa.ac.il

Abstract— This experiment was designed to see if information related to linguistic characteristics of read text can be deduced from fMRI data via machine learning techniques. Individuals were scanned while reading text the size of words in loud reading. Three experiments were performed corresponding to different degrees of grammatical complexity that is performed during loud reading: (1) words and pseudo-words were presented to subjects; (2) words with diacritical marking and words without diacritical markings were presented to subjects; (3) Hebrew words with Hebrew root and Hebrew words without Hebrew root were presented to subjects. The working hypothesis was that the more complex the needed grammatical processing needed, the more difficult it should be to perform the classification at the level of temporal and spatial resolution given by an fMRI signal. We were able to accomplish the first task completely. The second and third task did not succeed when all the data is used simultaneously. However, the third task was successful when training and testing was done within a continuous scanning run. (The experimental protocol did not allow this for the second task.) This does establish that complex linguistic information is decodable from fMRI scans. On the other hand, the need to restrict to the intra-run situation indicates that additional work is needed to compensate for distortions introduced between scanning runs.

***Index Terms*— Functional magnetic resonance imaging (fMRI), Multivoxel pattern analysis (MVPA), Machine Learning, Pattern Matching, Neural Networks, Cognitive Processing**

I. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a technique for determining which parts of the brain are activated by different types of physical sensation or cognitive activity, such as sound, the movement of a subject's arm, emotion and etc. fMRI provides indirect measurement of the neural activity through the change in blood Oxygen level, BOLD (Blood Oxygenated Level Dependent Contrast)

Use of multivoxel pattern analysis (MVPA) to predict the cognitive state of a subject during task performance has become a popular focus of fMRI studies. The input to these analyses consists of activation patterns corresponding to different tasks or stimulus types. These activation patterns are fairly straightforward to calculate for blocked trials or slow event-related designs, but for event-related designs the evoked BOLD signal for adjacent trials will overlap in time, complicating the identification of signal unique to specific trials. Rapid event-related designs are often preferred because they allow for more stimuli to be presented and subjects tend to be more focused on the task.

In this paper we are interested in seeing if this mechanism can be used for deeper grammatical cognitive tasks. The working hypothesis was that the more complex the needed grammatical processing needed, the more difficult it should be to perform the classification at the level of temporal and spatial resolution given by an fMRI signal. We use machine learning techniques on the gathered data.

Three experiments were performed corresponding to different degrees of grammatical complexity that presumably is performed during loud reading: (1) Hebrew words and pseudo-words (string of asterisks) were presented to subjects; (2) Hebrew words with diacritical marking and words without diacritical markings were presented to subjects; (3) Hebrew words with Hebrew roots and Hebrew words without Hebrew roots were presented to subjects. (In the first class words can be segmented into root and pattern, while in the second ones they cannot.) We were primarily interested in the performance on task 3.

In previous work, performed by Tali Bitan and Yael Weiss using SPM, Statistical Parametric Mapping, software and tools (SPM, http://www.fil.ion.ucl.ac.uk/spm/, [1]), a significant separation was achieved for the first two experiments. With the third, and most interesting experiment, the SPM toolset was unable to decide between the words with the Hebrew root to the words without the Hebrew root at reliability greater than chance.

II. Data

Nine subjects were presented to text the size of words in loud reading. Each of the subjects performed eight distinct runs of 360 seconds each, and exposed to pseudo words, words with Hebrew root and words without Hebrew root. Four of the eight runs presented only words with diacritical markings and the other four runs presented words only without diacritical markings. The TRs were 2 seconds and the experiment was designed in an event related design. There were baseline shifts between the runs. Each run contained twelve TRs with Hebrew root, twelve TRs without Hebrew root and twelve TRs with pseudo words. Additional unrelated conditions were presented during each run.

The scans were created in an Analyze format (Analyze, http://www.analyzedirect.com/) and preprocessed: aligned to adjust the movement between volumes (whole image), time corrected to get a signal for the whole brain from the same time point (slices alignment), normalized and smoothed.

III. Methods

The whole work was developed upon the Princeton MVPA toolbox (MVPA, www.pni.princeton.edu/mvpa). This open source toolbox facilitates exploration of multi-voxel pattern analysis techniques and further scripts development.

Before the classification stage more preprocessing was needed:

A. Conversion to AFNI format

As mentioned, the scans was saved in Analyze format. In event related fMRI experiments designs, the evoked BOLD signal for adjacent stimuli will overlap in time, complicating the identification of signal unique to specific stimuli.

To handle the convolution of signal and in order to de-convolve it, a conversion to AFNI format [2][3][4] was required. AFNI framework offers range of functionality for handling and processing fMRI scans, one of these is de-convolving.

B. Z-Score

The voxels values were in the range of 0 to proximally 8000. This variance should be normalized for the classification procedure to get better results in fewer learning epochs and to avoid baseline shifts issues.

We accomplished that using Z-Scoring which implies subtracting out the mean of each voxel's time course and scaling it so that the standard deviation of the time course is normalized to one. We treated each run as a separate time course to remove any between run differences due to cross runs baseline shifts.

C. Feature Selection

It is well established in machine learning that feature selection can have a substantial effect on the quality of the results (see e.g. [5]). Note that a subset of features can often obtain better classification results due to the fact that not all of the features are essential for the classification process. (Some features can harm the classification process since they just add noise to the system.)

It is also important to note that any feature selection algorithm should take the voxel's convolution into an account.

To throw away uninformative voxels, we used the GLM (General Linear Model) functionality of AFNI. It first de-convolves the data with a hemodynamic response function (HRF) and outputs F-stat value for each voxel which shows how much the voxel's activities varies between conditions over the course of the experiment.

After getting the stat map for each voxel we choose the 2500 voxels with the highest F-stat value as the features for the classification.

We run the classification on the Z-Scored data and only on the voxels which were selected by the GLM function as the features.

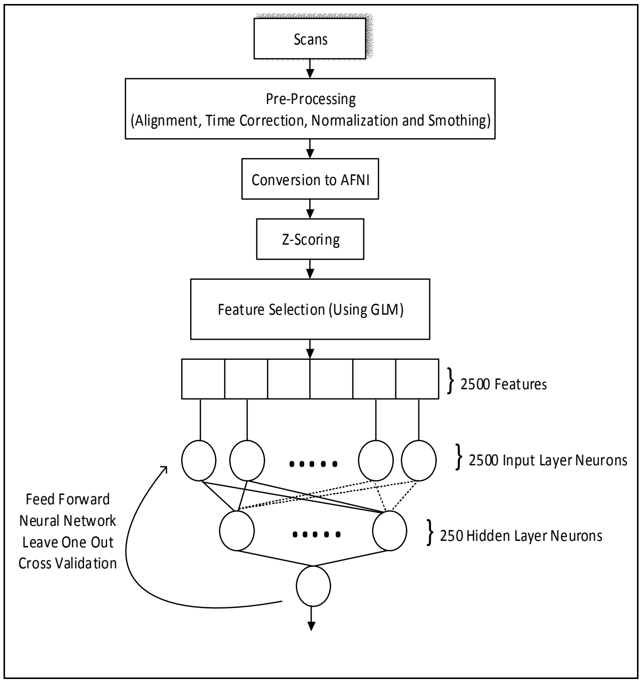
Scans
Pre-Processing
(Alignment, Time Correction, Normalization and Smothing)
Conversion to AFNI
Z-Scoring
Feature Selection (Using GLM)
} 2500 Features
} 2500 Input Layer Neurons
} 250 Hidden Layer Neurons
Feed Forward
Neural Network
Leave One Out
Cross Validation
.....
.....

Fig. 1. The processing pipeline. The feed forward neural network of 2500-250-1 architecture and was trained with the Conjugate gradient variant of back propagation

We used fully connected three-layer feed forward Neural Networks [6] as the machine learning method for the classification stage of each experiment appropriate for, in each case, a binary classification task.

Feed forward neural networks are computational models inspired by an animal's central nervous systems (in particular the brain) which is capable of machine learning as well as pattern recognition. Neural networks can be configured to do two-class classification that is by receiving labeled (training) data of two classes. During training it adjusts the weights between the neurons to minimize the error of the network's output to the actual labeling of the data. The learning algorithm which used was Conjugate gradient back propagation with Powell-Beale restarts [7]. Once optimal weights has been determined, new data are classified based on thresholding the output of the network.

The network was constructed from 2500 neurons (as the number of selected features) in the input layer which passed the features as is to the hidden layer, 250 neurons in the hidden layer and one neuron for the binary decision in the output layer.

We tested the classification of each experiment on two resolution levels: runs level and subject level.

In order to check how accurate the neural network separates the data, we mainly used cross validation tests using the "leave one out" method, taking one observation for testing and keeping all of the other observations for training the neural network.

The tested configuration for each resolution level was:

- Runs Level – for each subject and within each run.
    - N minus One - Creating twelve N minus one tests where on each test, two TRs, one from each condition, left for testing and the other 22 TRs used for training.
    - Equal Partitions – Creating two tests by randomly choosing 12 TRs, 6 from each condition, for training and the other for testing. This test increases generality by training on fewer observation and testing more observations.
- Subject Level (Cross Runs) – for each subject, training on three runs and trying to classify the observations on the fourth one, not mixing diacritical runs with non-diacritical runs thus creating N minus 1 cross validation test when N-1 runs used for training and one run left for testing to test generalization.

Trying to train over separate runs may introduce some noise as a result of subject movement, head orientation, magnet calibration etc. Thus better results are anticipated if all the work is done over the same person and on the same run.

IV. Results

After preprocessing the data, z-scoring, de-convolving and finding the best features for each specific test we discovered that:

A. Words versus pseudo-words

These could be told apart during loud reading at the Subject and Run levels at the level of 72.9% accuracy for all the subjects.

Each class of words: rooted words with diacritical signs, rooted words without diacritical signs, un-rooted words with diacritical signs and un-rooted words without diacritical signs was tested against the pseudo-words, strings of asterisks.

These results are consistent with more standard methods run by Tali Bitan and Yael Weiss using SPM software and used as sanity checks for the experiment.

TABLE I. Word classes vs pseudo words.

Tested word class	**Resolution Level**					
	Subject Level		***Run Level N Minus One***		***Run Level Equal Partitions***	
	Mean Accuracy	***Mean STD***	***Mean Accuracy***	***Mean STD***	***Mean Accuracy***	***Mean STD***
Words with Root and with Diacritical Signs	69.3%	10%	74.2%	9%	71%	10%
Words without Root and with Diacritical signs	70.7%	7%	75%	10%	73.6%	11%
Words with Root and without Diacritical signs	70.6%	9%	77.1%	9%	74%	12%
Words without Root and without Diacritical signs	71%	9%	75.3%	8%	74%	11%

B. Diacritical versus non-diacritical

The diacritical/non-diacritical experiment was unable to separate with a reliability greater than chance within a single subject. Run level classification could not be executed directly because of the experimental design. (There weren't any runs which mixed words with diacritical signs and words without diacritical signs.) We interpret these results as indicating that the experiment introduced too much noise in the fMRI between runs and additional data treatment is needed to obtain significant separation. (We note that in some earlier work by Bitan and Weiss, where some “session effect” corrections on the same data using standard SPM tools were done, did obtain significant distinction for voxel areas.)

C. Words with Hebrew root versus Words without Hebrew root

The words with Hebrew root versus without Hebrew root was unable to be separated with a reliability greater than chance on the subject level.

However, reducing the resolution of the tests to the run level, shows much improvement and it seems we can then reliably separate between the conditions.

TABLE II. GENERALIZATION RESULTS UNDER ALL TEST CASES BY THE RESOLUTION LEVEL.

Resolution Level	**Words with Hebrew Root Vs Words without Hebrew Root**			
	With Diacritical Signs		***Without Diacritical Signs***	
	Mean Accuracy	**Mean STD**	**Mean Accuracy**	**Mean STD**
Subject Level	48.61%	8.3%	49.34%	8%
Run Level N Minus One	74.03%	8.8%	72.40%	9%
Run Level Equal Partitions	73.14%	9.8%	70.53%	9%

In this case the mean average raised to around 72.5% with a standard deviation of 8-9%. The results for the runs with Diacritical signs was a bit higher from the results on the tests without the Diacritical signs and the results for the N minus one tests was a bit higher probably because of the amount of data used for training.

V. CONCLUSIONS

At this time, we are able to classify our most shallow grammatical task, words versus pseudo-words, on the Subject and Run level. We were unable to classify the other "deeper" tasks at more than chance levels at the Subjects level. When reducing the resolution of the tests to the Run level we can reliably separate rooted words versus un-rooted words.

Over all the results indicate that the experimental protocol introduces too much noise in fMRI between individual runs. Further work on eliminating this noise should be pursued.

ACKNOWLEDGMENT

This research is part of Haim Shalelashvili M.Sc. thesis under the supervision of Prof. Larry M. Manevitz of the Neuro-Computation laboratory, located at the Caesarea Rothschild Institute (CRI), University of Haifa, Haifa, Israel

REFERENCES

[1] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, Eds., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, 1 edition. Amsterdam ; Boston: Academic Press, 2006.

[2] R. W. Cox, "AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages," *Comput. Biomed. Res.*, vol. 29, no. 3, pp. 162–173, Jun. 1996.

[3] R. W. Cox and J. S. Hyde, "Software tools for analysis and visualization of FMRI Data," *NMR Biomed.*, vol. 10, pp. 171–178, 1997.

[4] S. Gold, B. Christian, S. Arndt, G. Zeien, T. Cizadlo, D. L. Johnson, M. Flaum, and N. C. Andreasen, "Functional MRI statistical software packages: a comparative analysis," *Hum. Brain Mapp.*, vol. 6, no. 2, pp. 73–84, 1998.

[5] O. Boehm, D. R. Hardoon, and L. M. Manevitz, "Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms," *Int. J. Mach. Learn. Cybern.*, vol. 2, no. 3, pp. 125–134, Sep. 2011.

[6] I.-V. Onut and A. Ghorbani, "Classifying cognitive states from fMRI data using neural networks," in *2004 IEEE International Joint Conference on Neural Networks, 2004. Proceedings*, 2004, vol. 4, pp. 2871–2875 vol.4.

[7] E. M. Johansson, F. U. Dowla, and D. M. Goodman, "Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method," *Int. J. Neural Syst.*, vol. 02, no. 04, pp. 291–301, Jan. 1991.