Modeling the Effect of Stimulus Perturbations on Error Correlations between Brain and Behavior

Heeyoung Choo
Beckman Institute and Department of Psychology
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
hchoo@illinois.edu

Dirk B. Walther
Department of Psychology
University of Toronto
Toronto, ON M5S 3G3, Canada
bernhardt-walther@psych.utoronto.ca

Abstract — Over the last decade, machine learning algorithms have proven to be useful tools for exploring neural representations of percepts and concepts in the brain. An important but often neglected next step is to relate neural representations to human behavior. Here, we introduce a novel approach to definitively linking neural representations to structural properties of stimuli as well as human behavior by analyzing patterns of classification errors using linear mixed-effects (LME) models. An LME model includes a priori predictive models of matching of error patterns between neural decoding and human behavior as fixed effects as well as random effects to account for subject variability. Finally, we demonstrate the viability of this approach using data from a set of fMRI and behavioral experiments testing the influence of visual properties on the neural representation of categories of real-world visual scenes.

Keywords—fMRI; neural decoding; machine learning; classification; error patterns; linear mixed-effects modeling

I. INTRODUCTION

Machine learning algorithms have rapidly gained popularity in the analysis of neuroimaging data in the last decade. Classifiers’ predictions, both accurate and erroneous, can reveal important details about the neural mechanisms underlying perception and cognition. Successful classification of a stimulus from patterns of brain activity indicates a robust neural representation of that stimulus. Behavioral experiments, on the other hand, are designed to test the causal relationship between stimuli (system input) and behavior (system output). Similarity of classification errors between neural decoders and human classification behavior allows for spatial mapping of the encoding of behaviorally relevant aspects of stimulus properties, such as the underlying categorical structure, throughout the brain [1].

In this article, we aim to elucidate neural mechanisms of perception by combining strengths of behavioral testing with brain mapping in a novel way. We perturb stimuli in specific ways and present them in behavioral and fMRI experiments. A-priori hypotheses about neural mechanisms lead to predictions about patterns of classification similarity measures across multiple experimental conditions. We here present an analytical framework for testing such hypotheses. We demonstrate the viability of the approach with a set of experiments on the categorization of real-world scenes.

II. METHODS

Let us assume that we have two competing hypotheses A and B about which stimulus properties are critical for the neural encoding of the categorical structure of the stimuli. Starting with the intact (I) stimuli, we generate two modified versions, one version (A), in which property A is preserved and property B is perturbed, and the other version (B), in which property B is preserved and property A is perturbed. All three types of stimuli are presented to human test subjects in a behavioral and in an fMRI experiment.

In the behavioral experiment, participants are asked to categorize stimuli. If the experiment is hard enough, participants will make mistakes, which are recorded in confusion matrices, separately for the three types of stimuli I, A, and B. In the fMRI experiment, participants view blocks of the three types of stimuli. Multi-voxel pattern analysis is used to decode the categories of the stimuli from patterns of fMRI activity, leading to a second set of confusion matrices.

Supported by a National Science and Engineering Research Council Discovery Grant to DBW.
matrices. Accuracy of decoding (mean of diagonal entries in the confusion matrix) for the different conditions is an indication of the strength of the neural representation of the different types of stimuli in a given region of interest. In this paper, we ignore accuracy and instead focus on the interpretation of decoding errors.

Error patterns (off-diagonal entries in the confusion matrices; $E$) are compared between experiments by computing their Pearson correlation coefficients. The coefficients for comparing each of the three behavioral with each of the three neural decoding conditions are summarized in the error match matrices:

$$M = \begin{bmatrix} C_{II} & C_{IA} & C_{IB} \\ C_{AI} & C_{AA} & C_{AB} \\ C_{BI} & C_{BA} & C_{BB} \end{bmatrix}$$

where:

$$C_{ij} = \arctanh \left( \text{corr}(E_i^{\text{neural}}, E_j^{\text{behavioral}}) \right)$$

represent error correlations under Fisher's $z$ transform to assure normality for the following analysis steps. Instead of attempting to divine meaning from such a tableau of numbers, we aim to measure quantitatively to what extent these numbers provide evidence for or against a-priori hypotheses.

We accomplish this with a linear mixed-effects (LME) model [2]. An LME model describes the linear relationship between observed data and design matrices just like other linear models. In our case, these fixed effects are determined by a-priori hypotheses that are formulated during the experiment design process. In addition to fixed-effects terms, an LME model includes terms that describe random effects, which are determined by the random draw of a small number of participants from the general population. The standard form of an LME model is:

$$y = X\beta + Zb + \varepsilon.$$

Here, $y$ are the observed error correlations summarized in $M$, $X$ is the fixed-effects design matrix, $\beta$ are the fixed-effects coefficients, $Z$ is the random-effects design matrix, $b$ are the random-effects coefficients, and $\varepsilon$ are the error terms (residuals).

Specifically, we are positing three a-priori hypotheses for the effects that influence $M$. The first is the default hypothesis $H_D$, which assumes that error patterns between neural decoding and behavior match only if the same stimulus type was presented in both experiments. Specific hypotheses $H_A$ and $H_B$ assume that error patterns match between the intact stimulus I and the degraded stimuli A and B, respectively. These hypotheses for $M$ can be formalized as:

$$H_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, H_A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, H_B = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

The fixed-effects design matrix $X$ is composed of vectorized and normalized versions of these hypotheses, replicated to account for the number of participants as well as a column of ones to model the intercept. The random-effects design matrix $Z$ associates each observation with the correct participant.

Coefficients for random effects and residuals are assumed to be normally distributed:

$$b \sim N(0, D(\theta)) \quad \varepsilon \sim N(0, \sigma^2 I),$$

where $D$ is the random effects covariance matrix with variance components vector $\theta$ of length $q$, and $\sigma^2$ the error variance.

Solving the model can be formulated as a likelihood maximization problem. The likelihood function is:

$$L(y|\beta, \theta, \sigma^2) = \int P(y|b, \theta, \sigma^2) \cdot P(b|\theta, \sigma^2) db$$

where

$$P(b|\theta, \sigma^2) = \frac{1}{(2\pi \sigma^2)^{q/2} |D(\theta)|^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} b^T D^{-1} b \right\}$$

and

$$P(y|b, \beta, \theta, \sigma^2) = \frac{1}{(2\pi \sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta - Zb)^T (y - X\beta - Zb) \right\}$$

$L$ is first maximized with respect to $\beta$ and $\sigma^2$ for fixed $\theta$. The solutions are substituted back into the expressions above, which is then optimized with respect to $\theta$.

The estimates of the fixed-effects coefficients $\hat{\beta}$ are subsequently interpreted as the strength of evidence in favor of the three hypotheses $H_D$, $H_A$, and $H_B$. At the same time, variance due to individual differences between participants is accounted for by
III. EXPERIMENTS

We applied LME modeling of patterns of error correlations using the data from [3]. The experiments were designed to measure the importance of contour orientation versus contour junctions for the categorization of real-world scenes. To this end, 15 participants passively viewed line drawings (LD) of real-world scenes while their brain activity was recorded using fMRI. The drawings depicted scenes from six categories: 
beaches, city streets, forests, highways, mountains, and offices.
Intact LD (I) were manipulated in two different ways: (1) the drawings were rotated by a random angle to disrupt contour orientation (R), and (2) individual contours of intact LD were randomly shifted (S) to disrupt contour junction statistics (Fig. 1).

In the fMRI experiment, participants passively viewed eight runs consisting of 18 LD blocks. A linear support vector machine predicted the scene category of each block in a leave-one-run-out cross-validation procedure, using the block fMRI activity in a set of pre-defined regions of interest (ROI): early visual areas V1-4, the parahippocampal place area (PPA), the occipital place area (OPA), the restrosplenial cortex (RSC), and the lateral occipital complex (LOC). In addition, an exploratory searchlight procedure was performed, which used a Gaussian Naive Bayesian classifier in a 1.95 cm³ (5x5x5 voxels) cubic searchlight.

In a separate behavioral experiment, a separate group of 39 participants was asked to classify briefly presented LDs. Responses were recorded in confusion matrices, separately for intact, rotated, and contour-shifted line drawings. Group-averaged confusion matrices were used as a reference to compute error matching matrices for each of the fMRI participants. The behavioral and fMRI experiments were performed separately to avoid conflation of brain activity due to perceptual similarity with decision making and motor planning processes (e.g., for pressing a particular button).

We can directly apply the mechanisms developed in the previous section. Stimulus conditions I, R, and S give rise to the hypotheses about error correlations between the neuroimaging and the behavioral experiment:

\[ H_0: \text{Errors are correlated when the same stimulus type is presented in both experiments (II, RR, SS).} \]

\[ H_4: \text{Errors are correlated when contour junctions are preserved, which is the case when the line drawings are rotated (IR, RI).} \]

\[ H_6: \text{Errors are correlated when orientations are preserved, which is the case when the contours are randomly shifted (IS, SI).} \]

IV. RESULTS

We computed error correlations for all 15 participants and fitted the LME model as described in the previous section. Fig. 2 shows the estimated coefficients and their standard deviations corresponding to the three hypotheses for all ROIs. In V1-3, none of the three hypotheses significantly explained observed patterns of error correlations, even though decoding accuracy for stimulus types I, R, and S was still significantly above chance for V1-3 (all \( q < 0.01 \); corrected using false discovery rate, data not shown here). The finding that error patterns from decoding scene categories from neural activity in V1-3 had little resemblance with behavioral error patterns is consistent with previous research [4].

![Fig. 2. Estimated coefficients and estimated standard errors of means of the fixed effects, \( H_0 \) in pale magenta, \( H_1 \) in red, and \( H_2 \) in blue in all eight regions of interest. Significance was corrected using false discovery rate, \(* q < 0.05, ** q < 0.01, *** q < 0.001.**\)](image)
As the stimulus processing progresses along the visual hierarchy, we found that both $H_D$ and $H_A$ significantly explained error correlations in area V4 as well as high-level visual regions (PPA, OPA, RSC, and LOC), demonstrating that preserving contour junctions was critical for error correlations between neural decoding and human behavior in these brain regions. In contrast, $H_B$ could not explain patterns of error correlations in any of the ROIs.

The searchlight results bolster the ROI-based results (Fig. 3). Both $H_D$ and $H_A$ (orange) could explain patterns of error correlations in the majority of visually active cortex, overlapping with V4, the PPA, OPA, and LOC. On the other hand, error correlations in searchlight locations overlapping with V1-3 are frequently explained by $H_B$, but only with negative coefficients (dark blue). Notably, in a large region that overlaps with the PPA, patterns of error correlations are explained by all three fixed-effects; positively by $H_D$ and $H_A$ and negatively by $H_B$ (yellow).

A linear model with only fixed effects yielded almost identical results. However, including random effects to account for subject variability showed superior model fits, especially in the PPA and LOC.

The results presented here allow for the strong conclusion that preservation of contour junctions but not orientation is critical for a neural representation of real-world scenes that matches human behavior. The importance of contour junction becomes explicit as early as V4, and continuous to dominate in the later scene- and object-specific brain regions.

The same conclusion is also supported by the analysis of decoding accuracy from decoding within and across stimulus types [3]. The argument in favor of the importance of contour junctions is built on the corroborating evidence from both analyses.

V. DISCUSSION

We have introduced a novel LME modeling approach for testing specific hypotheses about the relation between neural representations and human behavior under targeted stimulus manipulations. The method can equally be applied to neural and behavioral data that are collected concurrently or separately. The formulation allows for hypothesis-driven testing of specific regions of interest as well as exploratory whole-brain analyses. The simplicity of the linear modeling approach allows for a straightforward interpretation of the regression coefficients. We have demonstrated the application of the method to data from [3] on the critical properties for the representation of natural scenes.

The analysis of error patterns, while similar in spirit to representational similarity analysis (RSA) [5], uses a different processing paradigm. Instead of correlations of activity vectors, we here use error patterns that arise in the course of the decoding analysis, which makes the comparison to behavior straightforward. Furthermore, the explicit modeling of the effect of stimulus manipulations described here goes beyond the scope of RSA.

REFERENCES


