Decoding Neural Signals of Memory Reinstatement and Affective State

Stephanie A. Gagnon [*], **James E. Sorenson** [*], **and Ian C. Ballard** [†]

[*]Department of Psychology, Stanford University, and [†]Neurosciences Program, Stanford University

Final Project for CS 229: Machine Learning, December 13, 2013

Memory is thought to depend on the reinstatement of patterns of activity across the cortex that are similar to the patterns elicited by the original learning experience. According to this model, affective states such as stress might impair certain forms of memory by disrupting or delaying the reinstatement of these patterns. Characterizing the nature of cortical reinstatement under different affective states is thus a critical initial step towards understanding the neural basis of memory. Here, we applied machine learning techniques to electroencephalography (EEG) data collected during a paired-associate retrieval experiment, when participants were under conditions of relative safety and stress. Specifically, we trained an algorithm on neural patterns representing categories of visual images, and then classified neural signals during memory retrieval to decode reinstatement of associate categories. Additionally, we classified different types of mnemonic status and affective states, and we also find limited evidence that stress may impact memory-related neural activity.

memory retrieval | cortical reinstatement | stress | pattern classification

Abbreviations: EEG, electroencephalography; SVM, support vector machine

Our ability to store and retrieve memories allows us to access knowledge about the past to inform decisions and actions in the present. The process of memory retrieval is thought to depend on *(1)* the initial formation of a cortical representation of an event during encoding, and *(2)* the cortical reinstatement of these representations later during retrieval [1]. Critically, the cortical representations formed at encoding and later reinstated at retrieval are category-specific; that is, different categories of information (e.g., faces, places, objects) are represented by separable patterns of brain activity [2].

In the present experiment, we examine whether stress, operationalized by threat of shock, might influence memory reinstatement. More specifically, our primary goal was to use machine learning techniques to assess cortical reinstatement of category-specific information during periods of relative safety and stress. We first trained a classifier to distinguish category-specific (i.e., face, place, object) patterns of neural activity from the EEG signal recorded during the category-localizer task trials. Then, we applied this classifier to the EEG signal recorded during retrieval phase trials in order to assess the relationship between cortical reinstatement of these categories and memory accuracy in stressful and stress-free conditions. Critically, we predict that under conditions of stress, participants will show reduced cortical reinstatement of category specific information; as a result, our classifier will demonstrate reduced accuracy under stress relative to safety.

Additionally, we investigated whether we could independently classify both affective state and mnemonic status during the test phase. Machine learning techniques have been successfully applied to EEG in a relatively small number of papers, and it is still an open question as to *(1)* what kinds of states can be classified and *(2)* how to best define features and run a modeling pipeline. Our secondary goal was to examine different types of feature selection, modeling decisions, and parameter choices in order to optimize classification performance across three different classification tasks.

Methods

Twenty-three paid volunteers (14 females, 18-33 years, M=23.0 years, SD=4.4) participated after giving informed written consent in accordance with procedures approved by the Institutional Review Board at Stanford University.

The session consisted of three main phases: *(a)* a learning phase (*encoding*), *(b)* a test phase (*retrieval*), and *(c)* an image-viewing phase (*category-localizer*). During encoding, images of faces were paired with either object or place associates. At retrieval, participants viewed old and new face cues while we recorded neural activity with electroencephalography (EEG; 128 channels, 500 Hz sampling rate) under conditions of stress (i.e., possibility of receiving an electric shock) and safety (i.e., no shock possible); if participants identified the face as old, they were asked to recollect the paired associate (i.e., whether that face had been originally paired with a place, or an object). After the retrieval phase, participants performed a separate category-localizer task where they were presented with images of faces, places, and objects and asked to rate their familiarity with each stimulus.

Our goal is to use the EEG signal recorded during the category-localizer task trials to train a classifier to distinguish category-specific (i.e., face, place, object) patterns of brain activity. Then, we can apply this classifier to the EEG signal recorded during retrieval phase trials in order to assess the relationship between cortical reinstatement of these categories and memory accuracy in stressful and stress-free conditions.

Data Preprocessing. The raw EEG data was detrended with a high pass filter (.1 Hz Hamming windowed sinc FIR filter, order = 1650), and then low pass filtered (35 Hz, order = 500). Data from the four encoding runs were then concatenated into a single file. Next, the data was segmented into epochs -500 to 3000 ms relative to stimulus onset, and baseline corrected (-200 to 0 ms pre-stimulus baseline). Channels with spectral power between 10 and 20 Hz that were more than 4 standard deviations greater than the other channels were replaced with an average of its neighbors by spherical spline interpolation; other noisy channels were identified by visually inspecting the data, and interpolated.

Electrocardiogram (EKG) artifact was extracted using the extended-independent component analysis (ICA) algorithm of Lee, Girolami & Sejnowski, annealing rate of 0.98, with principal components analysis (PCA) dimension reduction to select the top 50 components; we identified the component weight(s) that accounted for the variance related to the EKG artifacts (0 to 3 components per participant, see Figure 1 for an example). Then, the same method was used to remove component weight(s) that accounted for the variance related to the blink artifacts (e.g., component spectral plot highlighted frontal channels near the eyes, with component weight(s) corresponding to blinks in the raw EEG signal); this allows us to retain more training/testing trials by excluding the components of the data contributing to signal, without excluding the entire trial. Data was then re-referenced to a common average.

Feature Space. After preprocessing, each trial consists of a 3500 ms voltage trace (measured in μV) corresponding to a 500 ms pre-stimulus period, and the 1500/3000 ms following stimulus onset. To decompose this voltage trace into features for classification analyses, we first decomposed the continuous voltage trace for each channel into five frequency bands of interest: delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and low-gamma limited by our initial low-pass filter (30-35 Hz). The amplitude component of the signal was extracted by applying the Hilbert transform on the band-passed signal; we then computed power by squaring the signal and transforming it to log-scale (dB). Then, we down-sampled the power timeseries for each frequency band into 100-ms time bins using an eighth-order low-pass Chebyshev Type I filter.

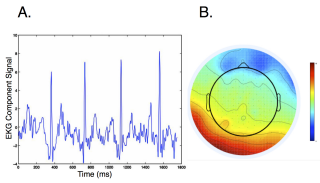

Fig. 1: Example EKG ICA component plotted for a single trial (A) and spectral topographical plot (B)

Results

Classification of Visual Category Representations. Our first goal was to classify the neural representations of the three visual categories (i.e., faces, places, and objects) using the independent category localizer datasets. During a localizer trial, participants viewed novel images of either a famous face (e.g., *Jennifer Aniston*), a famous place (e.g., *Pyramids of Giza*), or an object (e.g., *beach ball*), and rated their general familiarity with that image; we labeled each trial as a face, place, or object. Then, for each trial for each participant, we constructed a feature vector containing the first 10 time samples for each trial (averaged over 100 ms bins, spanning 1000 ms in total), for each scalp channel (104 total, full-scalp coverage), for the five frequency bands; each feature vector corresponds to a "pattern" of neural activation.

We implemented pattern classification analyses to discriminate between face, place, and object trials in Python using the scikit-learn package. Classification was assessed separately on each participant's data using a 25-fold stratified cross-validation procedure; trials from each of the 3 categories were randomly divided into $k = 25$ balanced subsets, preserving the percentage of samples for each category. The trials from $k-1$ of these subsets were then used for classifier training, and the held-out trials were used as a test set for assessing generalization performance; this was repeated iteratively k times.

We tested a variety of machine learning algorithms, including L_2 regularized logistic regression and support vector machines (SVM) with a linear kernel ($\langle x, x' \rangle$; n.b., we also tried 2nd order polynomial kernels and rbf kernels, but a linear kernel provided the best cross-validated accuracy); both of these algorithms yielded qualitatively similar results. Here, we will focus on the results using a multiclass linear SVM (implementation based on libsvm) with a one-vs-one scheme, and a penalty parameter $C = 10$. We selected the penalty parameter $C = 10$ to maximize generalization performance, as shown in Figure 2(a).

To decode neural patterns of visual categories, we first normalized each feature. Then, within each iteration of k-fold cross validation, we implemented feature selection (fit to the training data only) to eliminate uninformative features. Here, we tried several types of feature selection, including truncated singular value decomposition (SVD) and univariate feature selection, which both performed similarly. We also tried recursive feature elimination, but due to the size of our feature space, this technique was impractical. Here, we will focus on the analyses using univariate feature selection; this method runs a one-way analysis of variance (ANOVA) test for each feature, and extracts the p-value. We then extract some percentage of features with the highest $-Log$(p-value), i.e., the features best able to discriminate between the categories. Here, the percent of features selected did not have a large impact on generalization performance, with all percentages yielding cross-validated accuracy between 40-44% on average; see Figure 2(b).

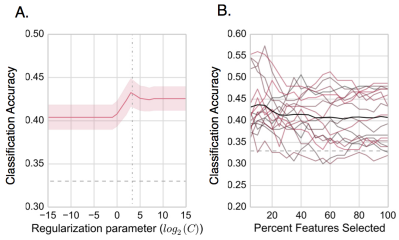

Fig. 2: Group accuracy for classification of visual category representations with a linear SVM as a function of penalty parameter C (A) and percent features selected (B)

With the linear SVM (penalty parameter $C = 10$), one-vs-one classification, and 25-fold cross-validation, we successfully classified visual category with 44% accuracy on average, significantly greater than chance, $t(16) = 6.83, p < 0.001$. See Figure 3(A) for the distribution of classification accuracy across participants. Looking at the group confusion matrix (Figure 3(B)), it appears that classification is best when participants were looking at images of faces (47.5% accuracy); classification was worse for the objects and places, which tended to get mis-classified as places and objects.

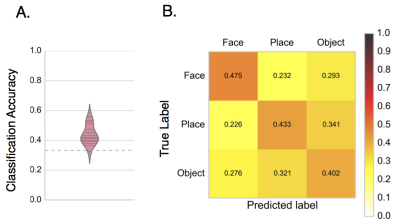

Fig. 3: Group accuracy for classification of visual category representations with a linear SVM (A) and confusion matrix

Next, we investigated classification accuracy as a function of time-bin width. More specifically, we trained and tested linear SVMs on subsets of the channel/frequency band features from specific time-bins (e.g., 0-200 ms post-stimulus onset, including time features from 0-100 ms and 101-200 ms); this provides us with some intuition about information contained in the neural signal at different points in the trial (e.g., the first 400 ms post-stimulus onset), as well as over specific temporal

windows (e.g., span of 200 vs. 400 ms). In general, classification accuracy increased with the width of the time-bin for training/testing with SVMs (25-fold cross-validated); see Figure 4. That is, when more temporal features were included in classification, generalization performance increased from 35% (when training/testing on 100 ms of data at a time) to 44% (when training/testing on features from the entire 1000 ms). From these results, it seems that information across the entire trial, and not some discrete subset, is valuable for discriminating visual stimulus category.

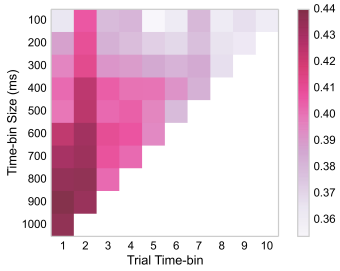

Fig. 4: Localizer classification accuracy across time-bins as a function of time-bin size

Decoding Test Phase Neural Activity. Using the independent set of category-localizer data for each participant, we next sought to decode *cortical reinstatement*. During the test phase, participants were presented with image cues of faces that were either new (never seen before) or old (learned during the study phase, paired with an object or a place). To examine cortical reinstatement, we focused specifically on the trials where participants correctly remembered that a face cue was old, and were able to retrieve the specific category of associate (e.g., the face cue had been paired with the Eiffel tower, a place); these trials will be referred to as *source hits*. Additionally, these source hit test phase trials were either under conditions of stress (i.e., threat of shock) or safety; we analyzed source hits from these conditions separately. Here, we were interested in the temporal dynamics of reinstatement, specifically *(a)* when during a test trial might we be able to decode the reinstated memory, and *b* what time window of visual category information was reinstated.

First, we trained 3-way linear SVM classifiers on the face vs. place vs. object category localizer data for each participant seperately; we trained classifiers separately for each 200 ms subset of the localizer data with a sliding window (e.g., 0-200 ms, 100-200 ms). Then, we tested each of these classifiers on each 200 ms subset of the test phase data, to assess cortical reinstatement of the associate image (9×14 classifications, see Appendix for full spectrum of classifiers). This analysis indicated that the first 200-300 ms of the localizer data contained information that was reinstated during memory retrieval. That is, when training a linear SVM on category-localizer data from 0-200 ms or 100-300 ms post-stimulus onset, this classifier was able to acheive above chance accuracy decoding reinstatement of a "object" or a "place".

Specifically, during safe block trials, classification of the reinstated image category was above chance starting at approximately 200-400 ms post-cue onset, and was sustained until approximately 600-800 ms post-cue onset; there was a second peak around 800-1000 ms that lasted through 1500 ms; see Figure 5. In contrast, during stress blocks, the rise to above-chance accuracy for reinstatement was delayed by approximately 200-400ms; here, classification of the retrieved associate category was above chance, but only when a longer amount of time had elapsed since cue-onset. We ran a linear mixed effects model to examine the effects of stress condition, the linear and quadratic effects of test time, and the interaction between stress and time on accuracy decoding reinstatement. This analysis revealed a significant interaction between the quadratic effect of time and stress condition, $\beta = 0.21, t = 2.06, p < 0.05$, such that during stress, test time had a strong quadratic relationship with classification accuracy, $\beta = -0.70, t = -4.96, p < 0.001$, rising above chance for a shorter time than during safe blocks; in contrast, during safe blocks the quadratic effect of test time on accuracy was wider, $\beta = -0.29, t = -2.04, p < 0.05$, such that accuracy was above chance for a longer period of time. Along these lines, when training on the first 200 ms of the category-localizer blocks, there was an overall main effect of stress condition on accuracy, such that safe blocks had marginally higher accuracy decoding the reinstated associates, relative to stress blocks, $\beta = 0.01, t = 1.9, p = 0.057$.

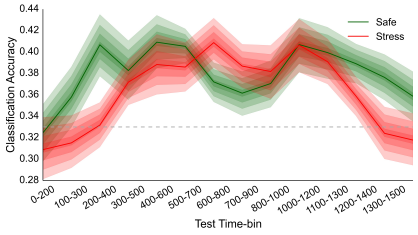

Fig. 5: Decoding reinstatement across the test trial: trained on category-localizer from 0-200 ms

Classification of Mnemonic Status. We next asked the question of whether we could classify mnemonic state using only data from the test phase. Since participants had fairly good memory performance on the task, we sought to distinguish between neural patterns during *correct rejections* (i.e., correctly indicating that a face was new), *general hits* (correctly indicating that a face was old without recalling the paired associate), and *source hits* (correctly indicating that a face was old as well as the paired object or place). We implemented both regularized logistic regression and linear SVMs across a variety of parameter values. Our results were roughly consistent across modeling choices, and we present data using linear SVMs with $C = 10$, keeping the top 15% of the features as determined by univariate feature selection (i.e., linear ANOVA). Other modeling procedures were the same as above unless stated otherwise. We used $k = 8$ fold cross validation because this was the largest value we could use in order to be consistent across all the subjects, who had variable numbers of trials per label. We successfully classified memory status (37.8% accuracy), which was significantly higher than chance, $t(16) = 2.47, p < .05$. See Figure 6(A) for the distribution of classification accuracy across subjects.

Examining the confusion matrix for this classification (Figure 6(B)), we observed that the classifier was best at distinguishing correct rejections from source hits, and tended to classify test items as correct rejections too often. We were concerned that the fewer number of general hits (mean number of correct rejections, general hits, source hits per subject = 58.1, 37.4 50.7), might be biasing results. In addition, general hits and source hits may be more similar to one another, since the subject is remembering an old stimulus, than correct rejections where a subject is classifying a stimulus as novel.

We re-ran the analysis in 2 separate ways: *(1)* excluding general hits, *(2)* combining general and source hits; both resulted in above chance classification accuracy. We conclude that the neural representations of the three categories are in fact distinct, and classification is best when separating them.

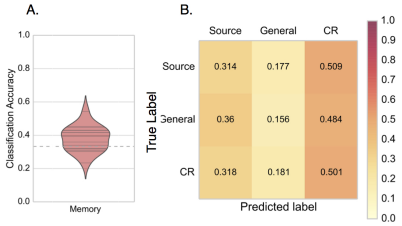

Fig. 6: Classification accuracy of mnemonic status across subjects (A), confusion matrix (B)

Since classification accuracy was close to chance, we examined whether our model suffered from high bias or high variance. To this end, we analyzed training and test accuracy on subsets of the training data, with between 0-90% of the data held out (Figure 7). We observed that although training accuracy declines modestly with more training data, test accuracy does improves only slightly. The large gap between training error and testing error indicate high variance rather than high bias. Indeed, when reducing the number of features used from 15% to .1%, accuracy increased to 42.0%, $t(16) = 6.68, p < 0.001$).

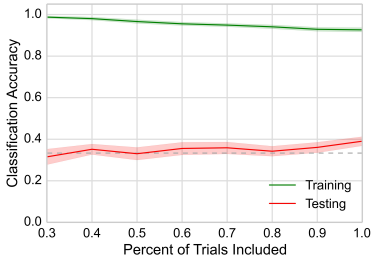

Fig. 7: Training vs. testing accuracy by percent of data for training

Finally, we investigated whether stress impacted the ability to classify memory retrieval by constructing separate classifiers for threat and safety conditions. When keeping the top .1% of features, we were able to correctly classify memory significantly above chance in both conditions (accuracy threat: 40.8%, accuracy safety: 42.9%); classification accuracy was not different between conditions, $t(16) = .96, p = .34$.

Classification of Affective State. Another goal for the project was to classify *affective state*; that is, to decode whether a given trial was from a "stress" block (i.e., during threat of shock), or if it was from a "safety" block (i.e., no threat of shock). Data used for decoding affective state came exclusively from the test phase of the experiment. EEG signal from the first 1500 ms of each source hit, general hit or correct rejection trial was preprocessed in the manner described above. To specifically examine stress related to *anticipatory threat* of shock (and to prevent excess motion artifact), we excluded trials in which participants received a shock, or had received a shock on the previous trial.

Two-way (threat vs. safety) affective state classification was implemented using L_2-regularized logistic regression, with univariate feature selection (keeping the most informative 10% of features) within 10-fold cross-validation. Overall, we had limited success classifying affective state. A classifier trained on all eligible trials did not classify trials above chance [Accuracy: 52.9%, $t(16) = 1.68, p > .05$]; see Figure 8. As we did with mnemonic status, we trained and tested the classifier with data sets of differing size; our results were similar to Figure 7, indicating that our modeling suffered from high variance but increasing data may offer minimal gains in performance.

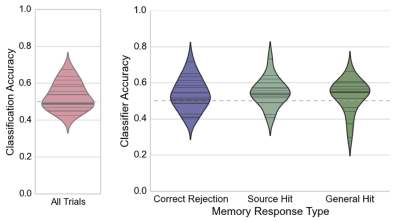

Fig. 8: Classification accuracy decoding affective state

Notably, we had more success classifying stress and safety when splitting our data into trials from correct rejections, source hits and general hits, while using a more aggressive feature selection (keeping the top 1% of trials). Here we achieved classification significantly above chance for source hits [Accuracy: 54.8% $t(16) = 2.55, p < .05$], but not for correct rejections [Accuracy: 53.7% $t(16) = 1.78, p = .09$], or general hits [Accuracy: 52.7% $t(16) = 0.37, p > .05$]; see Figure 8. Interestingly, the stress condition impaired source memory retrieval behaviorally; we hope to examine whether stress classification accuracy for source trials is correlated with memory performance across safe and stress blocks.

Discussion

Overall, we had success with our primary goal: to classify the reinstatement of visual category information from neural signal during memory retrieval. This consisted of two steps: *(1)* decoding neural representations of stimulus category based on visual input, and *(2)* detecting the reinstatement of these visual categories during the memory retrieval test phase. First we were able to classify image category (face, object or place) from neural signal related to visual input when participants were viewing images during the category localizer; we achieved above chance accuracy for all subjects using both linear SVMs and L2-regularized logistic regression. Here, temporal features proved important; as more temporal features were added, classification accuracy increased.

Secondly, we were also able to train a classifier on this category localizer data, and use it to successfully classify the category (object vs. scene) of reinstated associate images during the memory retrieval task. Notably, during retrieval, any information participants had about the objects and scenes was internally generated from memory. The success of our classifier therefore indicates that during retrieval, participants reinstated representations of associate visual category (objects and scenes), that were similar to those representations present during the viewing of these images. Here, classification of re-

instated memory category was above chance when training on the first 300 ms of localizer data. This suggests that neural activity related to stimulus visual category during recall is most consistently similar to neural activity during the initial stages of visual stimulus processing. Because previous EEG work has shown neural activity in the first 300 ms to be related to basic attention and perceptual processes, our findings would seem to indicate that the category information reinstated during recall may share some overlap with neural processes related to visual perception.

We had somewhat less success classifying mnemonic status and affective state. First, when classifying mnemonic status, a linear SVM classifier was best at distinguishing correct rejections from source hits, and tended to classify test items as correct rejections too often. Overall, however, we were able to successfully classify mnemonic status. Secondly, we were only able to classify neural representations of stress vs. safety on trials where participants correctly recalled source information (i.e., source hits). This result may be related to a behavioral effect in which stress selectively impacted participants' memory for source information. Our results also indicate that neural activity related to affective state may be more heterogenous than activity related memory and visual perception.

Analyses varying the percent of data used for training on accuracy indicated that our modeling of mnemonic status and affective state suffered from high variance. Testing on human subjects limits the amount of data that can be acquired, so constraining feature selection is likely the most critical step for optimizing classifier performance. Indeed, our modeling choices of classifier algorithm (SVM, logistic regression), penalty parameters, and feature selection algorithm (linear, SVD), did not have a large impact on performance. Varying the percentage of features included was important for mnemonic and affective state processing, although not for visual classification. Since EEG data is correlated across space and time, devising ways of decomposing the feature space is critical. A promising first step was to examine the classification accuracy across different time bins in the trial, determine where in time the most informative features were, and retrain the classifier on just that time window. This analysis allowed us to successfully classify reinstated associations based on a classifier trained on category localizer data. Other methods for constraining the feature space, such as examining accuracy as a function of frequency band, are promising avenues for future work.

The main contribution of this project is demonstrating successful classification for three different problems: memory reinstatement, memory status, and affective state. Additionally, we identified that our models suffered from high variance and explored a variety of modeling choices and feature selection methods to reduce this problem. We determined that strategies for reducing the feature space will be particularly important, especially when it is not possible to collect large datasets. In particular, we developed a technique for limiting feature space to informative time bins which may be of broad use to EEG classification problems. Taken together, our results indicate the potential for machine learning techniques to decode neural representations from EEG data. It will be exciting to examine the nature of the representations decoded, their frequency, spatial location, and temporal latency, in order to characterize the underlying neural mechanisms.

ACKNOWLEDGMENTS. This experiment was conducted in the Stanford Memory Laboratory, advised by Anthony D. Wagner, and supported by a grant from the John D. and Catherine T. MacArthur Foundations Law and Neuroscience Project to ADW. SG is supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship and a Stanford Graduate Fellowship (SGF), IB is supported by the National Science Foundation Graduate Fellowship (NSFGF). We thank Alex Gonzalez for assistance with EEG time-frequency decomposition.

1. **Gordon, A. M., Rissman, J., Kiani, R., & Wagner, A.D. (in press). Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. Cerebral Cortex.**
2. **Polyn, S.M., Natu, V.S., Cohen, J.D., & Norman, K.A. (2005). Category-specific cortical activity precedes recall during memory search. Science, 310, 1963-1966.**

Appendix: Test phase classification accuracy for reinstatement (object/place)