Master’s Thesis (DATS 5970) Research Proposal: Transfer learning to improve memory state classification from EEG

Abstract: Predicting future memory success from neural signals has proven to be a useful tool for guiding algorithmic interventions that improve memory. Accurate out-of-sample memory prediction, however, is difficult to achieve due to substantial variation in both noise sources and memory-related signals across recording sessions and across people. We propose testing both unsupervised and supervised methods of normalizing model input features in order to stabilize classifier training and improve generalization to out-of-sample data. Improvements in classifier accuracy will improve the efficacy of therapeutic interventions for people with memory deficits.

Memory is a unique phenomenon in human psychology: an integration of external stimuli into the mind for future reference. It is a complex process, operating on multiple timescales and in a context-dependent fashion (Howard & Kahana, 1999). Whereas failures of perception are generally consistent and predictable thanks to biological and psychological understanding of human sensory systems, the reasons for failure to encode and retrieve memories continue to elude researchers due to uncertainty about the mechanistic underpinnings of memory. Consequently, cognitive neuroscientists have spent much of the last two decades trying to identify neural correlates of human memory success and failure (Kahana et al., 2001; Buzsáki & Moser, 2013; Herweg et al., 2020).

This line of research, though incomplete, has helped memory scientists identify the features of neural timeseries that correlate with success in laboratory memory tasks. These features can be used as inputs to machine learning classifiers that binarize complex brain states as “good” or “bad” for encoding new memories. Predictive power opened the door for therapeutic interventions; a landmark study showed that electrical brain stimulation - selectively applied based on classifier predictions - was able to rescue “bad” memory states and improve subsequent recall (Ezzyat et al., 2018).

Of course, the effectiveness of this type of therapeutic treatment for individuals with memory deficits depends directly on the performance of the classifiers deployed. This converts a question of cognitive neuroscience into a machine learning challenge. Much effort has gone into exploring model and feature selection as well as data preprocessing (Arora et al., 2018; Meisler et al., 2019; Phan et al., 2019; Owen et al., 2020), but scientists working to improve brain-state classification have hit a performance plateau that is difficult to overcome for several reasons.

First, and perhaps most importantly, electrophysiological recordings are complex signals rife with noise. This includes noise that is native to the brain and represents stochastic processes uncorrelated with the behavior of interest, as well as more systematic noise that arises from technical challenges during recording. Noise is mitigated by pre-processing techniques like band-pass filtering to remove electrical line noise or more complex feature extraction methods that isolate signal components of interest; ultimately, signal-to-noise ratio noise enforces an upper bound on classification accuracy.

Brain signals, in addition to being complex and noisy, are often highly non-stationary. This makes it difficult to normalize neural features over the course of an experiment that lasts multiple hours. Many classifiers depend on normalized training and test features in order to make accurate predictions. Applying methods for normalizing neural features that are not vulnerable to the non-stationarity of the signal would likely improve classification accuracy.

Another key problem for training models on neural data from psychological experiments is heterogeneity across subjects – not only is every person’s brain different, but the recording sources that are used as input features do not match. This problem is especially troublesome for intracranial EEG, where subjects have vastly different recording montages that are based on their individual clinical needs. For scalp EEG the problem is not as severe since subjects use identical electrode caps, but precise contact placement can still vary. The consequence of this challenge is that classifiers are “personalized” for every subject, which requires collecting a huge amount of training data for each. This burden is highly expensive, and having independent classifiers for each subject is likely to lead to model overfitting. Moreover, inability to pool training data across subjects drastically reduces the size of the training set and makes many data-intensive deep learning techniques infeasible. It would be hugely advantageous to share information across sessions and subjects in order to stabilize and improve classifier predictions.

For my master’s thesis, I propose to work under the joint guidance of Konrad Körding (Bioengineering and Neuroscience) and Michael Kahana (Psychology) to improve the prediction of subsequent memory success from neural timeseries by applying newer methods in domain adaptation and transfer learning. Our plan is to study the classification of scalp EEG data from a free recall memory task, which has over 150 subjects recruited by the Kahana lab as part of the [Penn Electrophysiology of Encoding and Retrieval Study (PEERS)](https://memory.psych.upenn.edu/PEERS). We hope to first apply some well-developed domain adaptation techniques like Correlation Alignment (Sun et al., 2015) and DANNs (Ganin et al., 2016) to see if classifier generalization improves after directly controlling for differences in the feature space across days and across subjects. Next, we look to design a customized neural network architecture that will allow us to minimize error on memory classification while also learning a non-linear transformation to pre-condition the classifier’s input features across all recording channels, sessions, and subjects. By doing this, we will leverage the full dataset to teach a neural network to mitigate noise in our features that impedes correct classification; the network will learn what “good” or informative signal features look like because those features will lead to correct classification more often. Future work will hopefully include extending our approach to intracranial EEG, which poses a greater challenge than scalp EEG; as mentioned above, every subject with implanted electrodes has a different set of recording channels that are chosen for clinical purposes rather than for research. This complicates training a model across subjects, since each has features of different dimensions.

Improving memory classification could lead to the further development of both invasive and non-invasive therapeutic technologies to help people with memory and learning impairments. We are hopeful that this project will further a meaningful line of work over the course of the next two semesters.

References:

Arora, A., Lin, J.-J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M., & Lega, B. (2018). Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. *Journal of Neural Engineering*, *15*(6), 066028. https://doi.org/10.1088/1741-2552/aae131

Buzsáki, G., & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, *16*(2), 130–138. https://doi.org/10.1038/nn.3304

Ezzyat, Y., Wanda, P. A., Levy, D. F., Kadel, A., Aka, A., Pedisich, I., Sperling, M. R., Sharan, A. D., Lega, B. C., Burks, A., Gross, R. E., Inman, C. S., Jobst, B. C., Gorenstein, M. A., Davis, K. A., Worrell, G. A., Kucewicz, M. T., Stein, J. M., Gorniak, R., … Kahana, M. J. (2018). Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. *Nature Communications*, *9*(1), 365. https://doi.org/10.1038/s41467-017-02753-0

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., & Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, *17*(59), 1–35.

Herweg, N. A., Solomon, E. A., & Kahana, M. J. (2020). Theta Oscillations in Human Memory. *Trends in Cognitive Sciences*, *24*(3), 208–227. https://doi.org/10.1016/j.tics.2019.12.006

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–941. https://doi.org/10.1037/0278-7393.25.4.923

Kahana, M. J., Seelig, D., & Madsen, J. R. (2001). Theta returns. *Current Opinion in Neurobiology*, *11*(6), 739–744. https://doi.org/10.1016/S0959-4388(01)00278-1

Meisler, S. L., Kahana, M. J., & Ezzyat, Y. (2019). Does data cleaning improve brain state classification? *Journal of Neuroscience Methods*, *328*, 108421. https://doi.org/10.1016/j.jneumeth.2019.108421

Owen, L. L. W., Muntianu, T. A., Heusser, A. C., Daly, P. M., Scangos, K. W., & Manning, J. R. (2020). A Gaussian Process Model of Human Electrocorticographic Data. *Cerebral Cortex*, *30*(10), 5333–5345. https://doi.org/10.1093/cercor/bhaa115

Phan, T. D., Wachter, J. A., Solomon, E. A., & Kahana, M. J. (2019). Multivariate stochastic volatility modeling of neural data. *ELife*, *8*, e42950. https://doi.org/10.7554/eLife.42950

Sun, B., Feng, J., & Saenko, K. (2015). *Return of Frustratingly Easy Domain Adaptation* (arXiv:1511.05547). arXiv. https://doi.org/10.48550/arXiv.1511.05547