Research Statement

Meenakshi Khosla

Cornell University, Ithaca, NY

My main research interest lies in the development of computational models that can accurately predict how the brain responds under naturalistic stimulation. In neuroscience, stimulus-response relationships can be systematically understood from two complementary standpoints. *Encoding* models map stimuli to fine-grained neural activity via complex feature transformations. Conversely, *decoding* models aim to predict stimulus attributes directly from neural recordings. In my current and intended future research, I'm particular excited about the former encoding approach as a means to understand how sensory information is represented in the activity of different brain regions. Modeling neural responses to naturalistic stimuli, in particular stimuli that reflect the complexity of real-world scenes (e.g., movies), offers significant promise to aid in understanding the human brain as it functions in everyday life and a central theme of my research is to use machine learning or predictive modelling techniques to convert neural data into understanding and fundamental knowledge about the brain. Beyond satiating the spirit of scientific curiosity, I hope that understanding the link between neural activity and complex thought will eventually improve our understanding of neuropsychiatric disorders, creating novel opportunities for rehabilitation and neural prosthetics.

The research statement is organized into sections based on the broad questions that I hope to answer as a researcher. All sections include current work that is part of my doctoral dissertation and plans for future research in the respective areas.

1 How does the brain respond to rich, naturalistic multi-modal stimuli?

Understanding the neural basis of sensory perception has been a long-standing goal of neuroscience. Brain activity recordings of healthy subjects during "free viewing" of movies present a powerful opportunity to build ecologically-sound and generalizable models of sensory systems, also known as encoding models. In my research, I have been studying auditory and visual perception via encoding models at the macroscopic level of functional Magnetic Resonance Imaging (fMRI) voxels, which reflects metabolic activity (BOLD signals) that is a consequence of aggregated activity across populations of neurons. While the blurred spatio-temporal activity measured by fMRI may not be a direct window into neural coding like single-cell recordings in other animal models, this non-invasive imaging modality nonetheless serves as a very powerful and sensitive technique to study functional organization and feature representations in the human brain.

Why encoding models?

One might question the need for encoding or 'predictive' models in neuroscience in the first place. Since the majority of my current work and envisioned future work is based on the development of predictive models, it would be apt to pause and justify their merit at this stage. In the past, sensory systems have been studied extensively using task-based paradigms where the brain activity is recorded upon stimulation with hand-crafted stimuli. This paradigm has been very successful, for example, in identifying scene-selective or face-selective regions in the brain. While successful for testing specific hypotheses, this approach is limited in the sense that no single task-based experiment can help in developing broad theories of sensory processing that generalize outside the experimental circumstance they were based on (Varoquaux and Poldrack [2019]).

Predictive models, on the other hand, are based on out-of-sample prediction and they generalize to arbitrary new stimuli and can thus offer more holistic descriptions of sensory processing. The biggest advantage is that once we have such a general model, we can use it to formulate novel hypotheses about information processing in the brain that can then be tested under more rigorous conditions. Embedded knowledge within these models of the brain could also be harnessed in other applications,

such as independent neural population control by optimally synthesizing stimuli to elicit a desired neural activation pattern.

Further, predictive models can also be useful in hypothesis testing. In this case, encoding models encapsulating competing hypotheses about neural information processing can be pitted against each other and their empirical plausibility can be directly examined by comparing their predictions on held-out data against corresponding measurements. Such an approach can shed new light on how information is represented in different parts of the brain.

Given the usefulness of predictive models in hypothesis formation, in non-invasive brain-machine interfaces, as well as in answering important research questions relating to feature representations in the brain, much of my research is aimed at developing predictive models that can capture information processing within the brain more stringently than existing approaches. In the next sections, I describe my current research with predictive models of cortical responses and several possible extensions that I hope to explore in future work.

Endowing neural encoding models with both audition and vision

Deep neural networks trained on image or sound recognition tasks have emerged as powerful models of computations underlying sensory processing, surpassing traditional models of image or sound representation based on Gabor filters and spectrotemporal filters, respectively, in mid-level and higher-order visual and auditory regions (Yamins et al. [2014], Kell et al. [2018]). While this success is promising, existing encoding models based on deep neural networks have been limited in their focus on limited portions of the sensory space under naturalistic stimulation, ignoring the complex and dynamic interactions of modalities (audio and vision) in this inherently context-rich paradigm. This reductionism leads to sub-optimality in predictive models of cortical responses as neural patterns evoked by movies are not simply a conjunction of activations in modality-specific cortices by their respective uni-sensory inputs; rather, there are known cross-modal influences as well as regions that receive afferents from multiple senses. Longer narratives or movies further have an inherent temporal structure; much of the meaning we infer is from stimulation sequences rather than from instantaneous visual or auditory stimuli alone. To address this limitation, we recently proposed a Deep Neural Network (DNN)-based encoding model that captures three critical inductive biases about information processing in the brain: namely, *hierarchical processing, assimilation over longer timescales* and *multi-sensory auditory-visual interactions*. By developing and evaluating this model on a large-scale movie-watching dataset, we demonstrated how incorporating this joint information leads to remarkable prediction performance across large areas of the cortex, well beyond the visual and auditory cortices into multi-sensory sites and frontal cortex. Further, we demonstrated how these neural encoding models trained solely on naturalistic data can allow us to interrogate the temporal and sensory sensitivity of different brain regions. A recent pre-print describes the main results of this project:

- Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski and Mert R. Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. bioRxiv 2020.09.11.293878.

In this study, we also demonstrated that encoding models trained with naturalistic data are not limited to modeling the responses of their constrained stimuli set. Instead, by learning high-level concepts of sensory processing, these models can also generalize to out-of-domain data and replicate results of alternate task-bound paradigms. While our models were trained on complex and cluttered movie scenes, we tested their ability to predict response to relatively simple stimuli from specialized fMRI task batteries, such as faces and scenes. The remarkable similarity between the predicted and measured contrasts in all cases suggests that 'synthetic' brain voxels, predicted by the trained DNNs, correspond well with the target voxels they were trained to model. We thus provide evidence that these encoding models are capsulizing stimulus-to-brain relationships extending beyond the experimental world they were trained in. In the coming years, I plan to broaden the scope of this line of work and investigate the use of encoding models trained on naturalistic stimuli to fully map the functional modules that are typically characterized by hand-crafted stimuli in task-based experiments.

Attentional modulation in neural encoding

Naturalistic visual scenes typically contain a myriad of objects that occur all at once. Amidst this complexity, clutter, and often, occlusion, humans have a remarkable ability to interpret such rich scenes rapidly because of their ability to direct attention quickly to the most salient regions. It is well known that multiple objects in natural scenes compete for neural resources and attentional guidance helps to resolve the ensuing competition by biasing neural activity in favor of the attended location (Kanwisher and Wojciulik [2000]). Two questions come to mind in this context: (1) Can we leverage information about overt attention, such as gaze location, to improve response predictions? If so, how? (2) If gaze is useful, which consequently confirms the strong link between where humans attend and the evoked neural response, can we learn where humans attend in natural scenes solely with supervision from neural data without any eye-tracking? Our experimental approach using concurrent eye-tracking and fMRI recordings from a large cohort of human subjects watching movies evinced an affirmative answer to both these questions. Interestingly, we demonstrated that it is possible to learn visual attention policies that agree with human fixation data by end-to-end learning merely on fMRI response data, and without relying on any eye-tracking. This study, referenced below, was recently selected for an oral presentation at the NeurIPS conference (2020).

- Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski and Mert R. Sabuncu. Neural encoding with visual attention. Advances in Neural Information Processing Systems 2020.

The use of attention modules in neural encoding may not be restricted to the visual domain. For example, the selective role of auditory attention is well known in speech perception, as popularly demonstrated by the “cocktail party" effect which highlights our ability to segregate sounds of interest from complex, multi-speaker environments. In general, I think the field of neural encoding can benefit from the addition of modules that mimic attention by *amplifying the relevant (salient) features* of the environment while *suppressing the irrelevant* features. With the availability of large-scale brain activity data at an individual-level through open source projects such as BOLD5000 or Natural Scenes Dataset (NSD), I hope to extend this line of work to understand inter-individual variability in attention policies and its relation to inter-individual variability in brain activity. In the future, I also plan to continue investigating neural encoding models augmented with different types of attention for my dissertation and subsequent research.

On biological plausibility and connections with specialized cognitive processes

It is important to acknowledge the crucial ways in which the developed models diverge from biological networks. Feedforward DNNs, such as those used in our current experiments, fail to capture known properties of biological networks, such as local recurrence, however, they have been found to be useful for modelling neural activity across different sensory systems. Although the functional significance of intra-regional recurrent circuits in core object recognition is still a matter of vigorous debate, mounting evidence suggests that they may be subserving recognition under challenging conditions (Kar et al. [2019]). As future work, I would like to delve into the investigation of more neurobiologically plausible models of the cortex that innately model intra-regional recurrent computations, especially in relation to their role in visual recognition and neural response prediction.

Over the course of my current research and review on this topic, I have realized the importance of incorporating modules that mimic cognitive and perceptual processes, such as working memory and attention, from a neural encoding perspective. For instance, our research with human fMRI, as well as previous research with animal models (Sinz et al. [2018]), shows that models that can efficiently store and access information over longer spans, such as Recurrent Neural Networks (RNNs) with sophisticated gating mechanisms, are much more suitable for modeling neural computations that unfold over time (as in stimulation with natural videos) in comparison to non-recurrent approaches. Since activations of units within RNNs depend not only on the incoming stimulus, but also on the “current” state of the network as influenced by past stimuli, they are capable of holding short-term events into memory. Adding the RNN module can thus be viewed as augmenting the encoding models with working memory. Further, in our study on attentional modulation, we modelled the phenomenon of *attention* as a novel module that selects certain portions of visual stimuli, the so-called attention “spotlights", for subsequent processing at the expense of others. The use of multiplicative scaling factors to modulate feature maps has some biological grounding insofar as it is loosely inspired from

the notion of gain modulation in existing studies on biological attention. However, a biologically plausible computational model of attention would capture both bottom-up and top-down influences of working memory and context as they may ultimately constrain which spatial locations are selected for further processing. In the future, I hope to draw upon the visual attention and saliency literature to develop models that mimic biological attention more closely.

2 How to effectively share information across subjects to improve individualized neural encoding models?

Over the last year, I have also started to think about practical problems in the development of subject-specific neural encoding models with fMRI. Building accurate individual-level models of brain function often requires large amounts of data per subject for good generalization. The problem is further exacerbated by the variability in anatomy and functional topographies across individuals, making inter-subject knowledge transfer difficult. In light of these challenges, an important question arises: how can one leverage information from one dataset's (or subject's) stimulus-response relationship to better inform or regularize the encoding model of another novel subject? In a recent study, we proposed a shared convolutional neural encoding approach to address this challenge by drawing inspiration from recent studies which suggest that coarse-grained response topographies are highly similar across subjects and individual differences manifest in fine-grained patterns. Our proposed approach has several merits. First, it allows us to combine data from multiple subjects watching same or different movies to build a global model of the brain. At the same time, it can capture meaningful individual-level deviations from the global model which can potentially be related to individual-specific traits or functional organization. Second, it is amenable to incremental learning with diverse, varying stimuli across same or novel subjects, and with less constraints on data collection from single subjects. Our study, which was published in the proceedings of MICCAI last year, is referenced below:

- Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski and Mert R. Sabuncu. A shared neural encoding model for the prediction of subject-specific fMRI response. MICCAI 2020.

Interestingly, in this study, we also showcased a potential application of the shared model in personalized brain mapping by demonstrating its ability to characterize meaningful individual differences in response to traditional task-based facial and scenes stimuli. I believe this is an important research direction as developing accurate individualized neural encoding models that can rapidly characterize high-resolution functional maps in novel subjects with limited data can pave the way for novel neurotechnologies and diagnostics that are based on aberrations in neural representations of information across disorders.

3 'Brain-like' representations and 'brain-like' capabilities: Is there a causal link?

Thus far, in my current research, I have explored the use of artificial neural networks in modelling neural responses. One can also turn the question around and ask how neural data or neuroscientific insights can help inform and inspire the next generation of models in Artificial Intelligence (AI). Recently, there has been a growing interest in this direction and some recent studies have already laid the seed for cross-fertilization between machine learning and neuroscience. For instance, an integrative benchmark was recently proposed in Schrimpf et al. [2020] wherein they developed a comprehensive 'Brain-score' to measure the similarity of DNNs to brain's core object recognition using neural and behavioral measures. These benchmarks aim to guide model development in AI by adopting the reverse-engineering approach. Other studies have shown that explicitly encouraging neural networks to build representations that are similar to neural representations or principles can improve their robustness under stimulus noise and adversarial attacks (Li et al. [2019], Dapello et al. [2020]). Attempts at aligning the activity of neural networks with brain-like representations are based on the hypothesis that networks favoring brain-like representations will demonstrate brain-like capabilities, for example, *stronger generalization* with limited supervision, *robustness* to stimulus noise, *lower sample complexity* in few-shot learning settings etc. I am interested in delving into this

exciting topic with human fMRI data in my research to study if increasing the similarity of model representations to human brain activity patterns can indeed lend them brain-like capabilities.

4 Individualized diagnosis and functional mapping

While uncovering the workings of the human brain using neuroimaging is a thrilling and satisfying scientific endeavor in its own right, harnessing this window into the brain to understand and alleviate the afflictions of neurological and psychiatric conditions is an important goal that I hope to keep revisiting during my research.

Recently, there has been a surge of studies harnessing resting-state functional connectivity for a wide range of supervised prediction tasks. For instance, the application of machine learning methods to resting-state fMRI (rs-fMRI) data has shown great promise in investigations of the developing connectome, as well as in predicting individual differences in cognition and behavior. Over the last decade, substantial effort has been devoted to using rs-fMRI for classification of a wide range of neuropsychiatric conditions, such as alzheimer's disease, schizophrenia, autism spectrum disorder etc. Predictive approaches can also be used to address research questions of interest in neuroscience. For example, to what extent is resting-state functional connectivity heritable or how does it vary across different vigilance states? Last year, we presented a high-level overview of this promising intersection of machine learning and resting-state fMRI analysis in a review article for a special issue on 'Artificial Intelligence in MRI':

- Meenakshi Khosla, Keith Jamison, Gia H. Ngo, Amy Kuceyeski and Mert R. Sabuncu. Machine learning in resting-state fMRI analysis. Magnetic Resonance Imaging, 2019.

Linking functional connectivity and mental disorders with machine learning

Besides encoding models, I have been interested in developing machine learning approaches for individual level predictions from neuroimaging data, in particular, resting-state fMRI scans. To this aim, we previously proposed a novel convolutional neural network based approach that takes advantage of the full-resolution 3D spatial structure of rs-fMRI data and fits non-linear predictive models. The key insight of our approach was to use the 'connectivity fingerprint', or functional coupling of each voxel to distinct target regions of interest, as input features for a traditional volumetric convolutional neural network, represented as a multi-channel image volume. The use of spatial convolutions allowed us to capture local structural or topographic patterns in the data, such as connectivity gradients. We further expanded upon this approach to build an atlas-agnostic ensemble using stochastic brain parcellations, leading to more robust and generalizable prediction models. This approach led us to outperform existing models in autism classification on a large-scale publicly available dataset, while shedding light on the salient brain regions useful for this discrimination task.

- Meenakshi Khosla, Keith Jamison, Amy Kuceyeski and Mert R. Sabuncu. Ensemble learning with 3D convolutional neural networks for functional connectome based prediction. NeuroImage, 2019.

From 'connectivity' to 'activity'

Aside from serving as a diagnostic or prognostic indicator for different mental disorders, resting-state fMRI has also demonstrated potential in clinical settings as a useful technique for pre-surgical mapping of brain function. Evidence suggests that resting-state connectivity already contains the wide repertoire of cognitive states that can enable efficient mapping of functional networks at the individual level. Recently, the above insight of using resting-state 'connectivity fingerprints' as input to predictive models was applied in an MRI modality translation problem to predict individual task contrasts from their resting-state fingerprints.

- Gia H. Ngo, Meenakshi Khosla, Keith Jamison, Amy Kuceyeski and Mert R. Sabuncu. From connectomic to Task-evoked Fingerprints: Individualized Prediction of Task Contrasts from Resting-state Functional Connectivity. arxiv:2008.02961 (2020). [MICCAI 2020]

5 Looking ahead

My research thus far has been focused on the broad themes of utilizing brain imaging for understanding cognition and making individual-level predictions of clinical phenotypes. Broadly, to summarize my vision for future work, I hope my research can answer questions of the kind that David Marr elegantly and inspiringly formulated in his influential book (Marr [2010]): *What kind of an information processing device is the brain? How is information from our environment robustly transformed into a coherent percept of the world? What are the fundamental principles underlying neural computations?* While we have made significant strides over the last couple of decades; to date, we have few satisfying answers for the questions above and there is much to be learned about the function of different brain areas, the means by which the function is achieved, how it comes into being given the constraints faced by the system (over evolution or development) and how disparate brain networks/regions collectively support complex human behavior. At the end of the day, I hope my research can make significant contributions towards understanding the broad general principles that can explain biological intelligence and consequently inform artificial intelligence. Towards the fulfilment of this dream, I envision the new data revolution in neuroscience, with large-scale compilation of neural data and dissemination through open-source initiatives, to play a crucial role in narrowing down the numerous (often incompatible) theories and hypotheses about how the mind works.

References

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *bioRxiv*, 2020. doi: 10.1101/2020.06.16.154542. URL `https://www.biorxiv.org/content/early/2020/06/17/2020.06.16.154542`.

Nancy Kanwisher and Ewa Wojciulik. Visual attention: insights from brain imaging. *Nature reviews neuroscience*, 1(2):91–100, 2000.

Kohitij Kar, J. Kubilius, Kailyn Schmidt, Elias B Issa, and J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22:974–983, 2019.

Alexander J. M. Kell, Daniel Yamins, Erica Shook, Sam V Norman-Haignere, and J. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98:630–644.e16, 2018.

Zhe Li, W. Brendel, E. Y. Walker, E. Cobos, Taliah Muhammad, J. Reimer, M. Bethge, Fabian H Sinz, Xaq Pitkow, and Andreas S. Tolias. Learning from brains how to regularize machines. In *NeurIPS*, 2019.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2020.07.040. URL `http://www.sciencedirect.com/science/article/pii/S089662732030605X`.

Fabian H Sinz, Alexander S. Ecker, P. G. Fahey, E. Y. Walker, E. Cobos, Emmanouil Froudarakis, D. Yatsenko, Xaq Pitkow, J. Reimer, and A. Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *bioRxiv*, 2018.

G. Varoquaux and R. Poldrack. Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, 55:1–6, 2019.

Daniel Yamins, H. Hong, C. Cadieu, E. Solomon, Darren Seibert, and J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111:8619 – 8624, 2014.