Universally Controversial Stimuli Reveal that Adversarial Robustness Improves DNN Prediction Accuracy across the Entire Human Auditory Cortex

David Skrill (david_skrill@urmc.rochester.edu)
Department of Biostatistics and Computational Biology
University of Rochester Medical Center
601 Elmwood Ave., Rochester, NY 14642

Jenelle Feather (jfeather@cmu.edu)
Center for Computational Neuroscience, Flatiron Institute
160 5th Ave
New York, NY 10010
Neuroscience Institute, Dept. of Psychology
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

Sam V. Norman-Haignere (samuel_norman-haignere@urmc.rochester.edu)
Depts. of Biostatistics and Computational Biology, Neuroscience
University of Rochester Medical Center
Depts. of Brain and Cognitive Sciences, Biomedical Engineering
University of Rochester
601 Elmwood Ave., Rochester, NY 14642

Abstract

Model comparison is central to all scientific progress. In sensory neuroscience, a key challenge is that distinct models often make similar neural predictions due to correlations between distinct features in the tested stimulus set. Here, we show how to distinguish models for a full neural population by designing a targeted "universally controversial" stimulus set that makes distinct, high variance predictions across an entire sensory cortical system (human auditory cortex) in every subject tested. We applied to compare the neural prediction accuracy of standard artificial neural networks (ANNs) from ANNs trained to be robust to "adversarial attacks". Standard ANNs are notoriously vulnerable to small stimulus perturbations that can substantially alter the network's decisions without meaningfully altering human perception. Yet, we find that the prediction accuracy of standard and robust ANNs in the human auditory cortex is virtually indistinguishable when measured using fMRI responses to natural sounds. In contrast, when tested with controversial stimuli, the cortical prediction accuracy of the robust model remains high throughout the auditory cortex, while the predictive power of the non-robust model drops to near zero. Universal controversiality thus opens the door to much more powerful model comparisons in sensory neuroscience and demonstrates a strikingly uncontroversial model improvement from adversarial training.

Keywords: neural encoding; model comparison; adversarial robustness; natural stimuli; fMRI; auditory cortex

Introduction

A key goal of sensory neuroscience is to build computational "encoding" models that can accurately predict the neural responses of a sensory system to a complex stimulus such as speech. Over the past decade, researchers have constructed increasingly sophisticated encoding models that are capable of predicting responses to complex natural stimuli in part by leveraging advances in deep learning and machine learning (Yamins et al. (2014), Vaidya et al. (2022), Kell et al. (2018), Tuckute et al. (2023)). A key challenge with this approach is that distinct encoding models often make highly similar predictions for natural stimuli, making it difficult to adjudicate between competing models. This problem is fundamental to model comparison in sensory neuroscience and developing effective methods to solve it would thus be a major advance.

An important example of this problem arises when comparing the neural predictive performance of different artificial neural network (ANN) models. ANNs trained on challenging tasks, such as speech and object recognition, are state-of-the-art in predicting neural responses to natural stimuli (Kell et al. (2018), Hosseini & Fedorenko (2023), Tuckute et al. (2023)). However, distinct ANN models often make very similar neural predictions for natural stimuli, even when their internal representations differ substantially (Feather et al. (2023)). For example, standard ANNs are notoriously sensitive to small perturbations that can radically change the network's behavior ("adversarial attacks") without substantially altering human perception of a stimulus (Goodfellow et al. (2015), Tramèr et al. (2018)). Incorporating adversarial examples in DNN training (adversarially-robust (AR) models) substantially improves this issue Madry et al. (2017), and helps align the model with human perceptual judgments (Gaziv et al. (2023); Feather et al. (2023)), yet standard and AR models make nearly identical neural predictions for fMRI responses to natural auditory stimuli, as we show (Fig. 1A).

Competing models can be effectively compared by finding stimuli that yield distinct behavioral responses (Wang & Simoncelli (2008); Golan et al. (2020)), but extending this approach to neuroscience has been difficult. A key challenge is that most neuroscience experiments sample hundreds or thousands of neural responses, and it is not feasible to design a new stimulus for each response. To address this problem, we develop an approach for synthesizing stimuli that are controversial across an entire sensory cortical system (universally controversial stimuli or UCSs).

Correlated Predictions Prevent Natural Sounds From Distinguishing Models

We first tested if it was possible to distinguish between standard and AR ANNs using natural sounds. We selected a convolutional neural network trained to predict spoken words with background sounds. This model has shown strong neural predictive performance in the human auditory cortex (Tuckute et al. (2023)), but as with most ANNs, it is highly sensitive to adversarial attacks. We compare this ANN to a variant trained to be robust to $\ell_2(\varepsilon = 1)$ adversarial attacks applied to the cochleagram (Feather et al. (2023)).

We fit encoding models using standard and AR ANNs (denoted E^S and E^{AR}, respectively) to fMRI voxel responses from 30 subjects from two prior studies (Norman-Haignere et al. (2015), Boebinger et al. (2021)). Each voxel was modeled as a weighted sum of the units from a single model layer (via ridge regression) and we selected the layer that best predicted the response separately for each voxel and model. We found that the accuracy of predictions from E^S and E^{AR} to a test set of sounds was very similar (Fig 1A), *prima facie* (55 non-overlapping sounds were used for training and testing). Similar predictive performance could imply the models are equally good neural models but could also reflect correlated predictions. To test this possibility, we correlated the predictions between the two models in the test set. We found that the correlations were very high (Fig 1B) and approached the maximum possible correlation given by training the same model on two non-overlapping training sounds. Since the model predictions (E^S and E^{AR}) are so similar, the neural data cannot adjudicate between the models and the prediction accuracy is guaranteed to be similar.

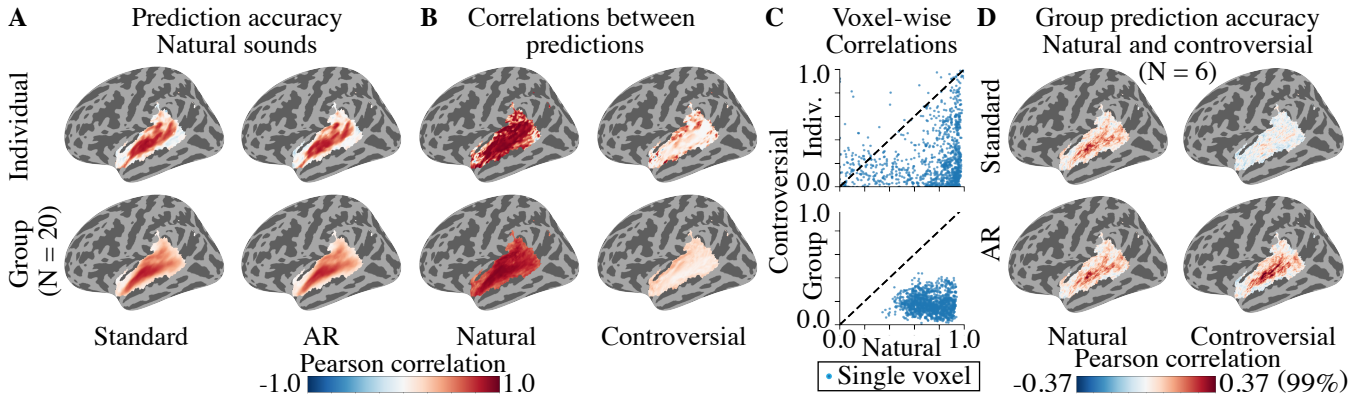

Figure 1: **A**. Individual and group-averaged prediction accuracy of standard (left) and adversarially-robust (AR, right) encoding models in predicting human fMRI responses to natural sounds. **B**. Correlations between model predictions for natural sounds (left) and synthesized controversial sounds (right). The models are the same as those shown in **A**. **C**. Scatter plots showing correlations between standard and AR model predictions for voxels from a single subject (top) and at group level (bottom) from subjects not used for sound synthesis. **D**. Prediction accuracy of standard (top) and AR (bottom) models to natural (left) and controversial sounds in a new group of subjects also not used for synthesis.

Synthesized Sounds Decorrelate Model Predictions Throughout Auditory Cortex

We tested whether it was possible to synthesize a single sound set $\mathcal{S}$ that would be universally controversial in the auditory cortex (Fig 1B), defined as yielding low correlations between the predicted responses $E^S(\mathcal{S})$ and $E^{AR}(\mathcal{S})$. Additionally, we wanted the neural response to the new sound set to maintain high variance to ensure high signal power relative to MRI measurement noise. To accomplish these two goals, we minimized the following loss function:

$$L(\mathcal{S}) = \sum_{i \in \mathcal{V}} \Big(\left[\mathrm{cov}(E_i^S(\mathcal{S}), E_i^{AR}(\mathcal{S})) - C_i^* \right]^2 \quad (1)$$

$$+ \quad \frac{1}{2} \sum_{m \in \{S,AR\}} \left[\mathrm{var}(E_i^m(\mathcal{S})) - V_i^* \right]^2 \Big) \quad (2)$$

where $\mathcal{V}$ is the set of voxel indices, and C_i^* and V_i^* are prespecified target values for the covariance and variance of predictions for voxel i, respectively. (1) measures the covariance between the prediction of two models for all sounds, and (2) measures the difference between predicted variance and a desired variance target. In this work, we target 0 covariance for all voxels ($C_i^* = 0 \ \forall \ i$). It is not important that we hit our variance target, only that we push the model to generate high variance stimuli, and we therefore set the variance target to be high (5 times that for natural sounds), which did not impede our ability to minimize the covariance term.

The synthesized sounds differed from natural sounds, as expected, but nonetheless exhibited interesting and complex acoustic structure. There were speech-like sounds with tone-complexes and formant-like structure, frequency-modulated inharmonic tones, amplitude-modulated noise, and dense tone and noise clouds with varied frequencies and spectrotemporal modulation patterns. To evaluate whether our sound set was consistently controversial, we measured the correlation of our predictions for an entirely new set of 20 subjects whose voxels were not used to synthesize the sounds.

We found that this approach was strikingly successful. In nearly every voxel at the individual subject level and in literally every voxel at the group level, we found that the controversial sounds had lower correlation than the natural sounds (Fig 1C; $p < .001$, Wilcoxon). The predicted variance was also somewhat higher than that for natural sounds as intended ($p < .001$, Wilcoxon). Because the voxels and subjects used to measure these statistics were not used for sound synthesis, these results demonstrate that we have synthesized a set of sounds that are universally controversial for any region of the auditory cortex in any subject tested.

Universally controversial sounds reveal near universal improvements from adversarial training

We conducted an fMRI experiment where we measured responses to both natural and controversial sounds from a new set of 6 subjects. We again fit encoding models using the natural sounds and we correlated the measured fMRI response with the model-predicted response to both the controversial sounds and independent set of natural sounds not used to fit the model. For natural sounds, we found that the prediction accuracy was nearly indistinguishable for standard and AR models, replicating our prior results. In contrast, for the controversial sounds, we observed a dramatic difference. Specifically, the prediction accuracy for the robust model was in fact slightly higher for controversial sounds than for natural sounds, likely due to variance maximization. In contrast, the prediction accuracy for the standard ANNs dropped to near 0. Thus, in virtually every voxel in the auditory cortex, controversial stimuli provide a stronger approach for model comparison, compared with natural sounds, revealing a near universal improvement in model prediction accuracy from adversarial training that is masked when using natural stimuli alone.

References

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021, Jun 1). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*, *125*(6), 2237–2263. (Epub 2021 Feb 17) doi: 10.1152/jn.00588.2020

Feather, J., Leclerc, G., Madry, A., et al. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, *26*, 2017–2034. doi: 10.1038/s41593-023-01442-0

Gaziv, G., Lee, M., & DiCarlo, J. J. (2023). Strong and precise modulation of human percepts via robustified anns. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 65936–65947). Curran Associates, Inc.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330-29337. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1912334117 doi: 10.1073/pnas.1912334117

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples.* Retrieved from https://arxiv.org/abs/1412.6572

Hosseini, E. A., & Fedorenko, E. (2023). Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. In *Thirty-seventh conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=h3lTrt4Ftb

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018, May). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.e16. (PMID: 29681533) doi: 10.1016/j.neuron.2018.03.044

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015, Dec 16). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*(6), 1281–1296. doi: 10.1016/j.neuron.2015.11.035

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *International conference on learning representations (iclr).*

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2023, 12). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, *21*(12), 1-70. Retrieved from https://doi.org/10.1371/journal.pbio.3002366 doi: 10.1371/journal.pbio.3002366

Vaidya, A. R., Jain, S., & Huth, A. (2022, 17–23 Jul). Self-supervised models of audio effectively explain human cortical responses to speech. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 21927–21944). PMLR. Retrieved from https://proceedings.mlr.press/v162/vaidya22a.html

Wang, Z., & Simoncelli, E. P. (2008). Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, *8*(12), 8–8.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1403112111 doi: 10.1073/pnas.1403112111