Using deep learning and fMRI to study auditory neural representations

Jessica Thompson and Marc Schönwiesner

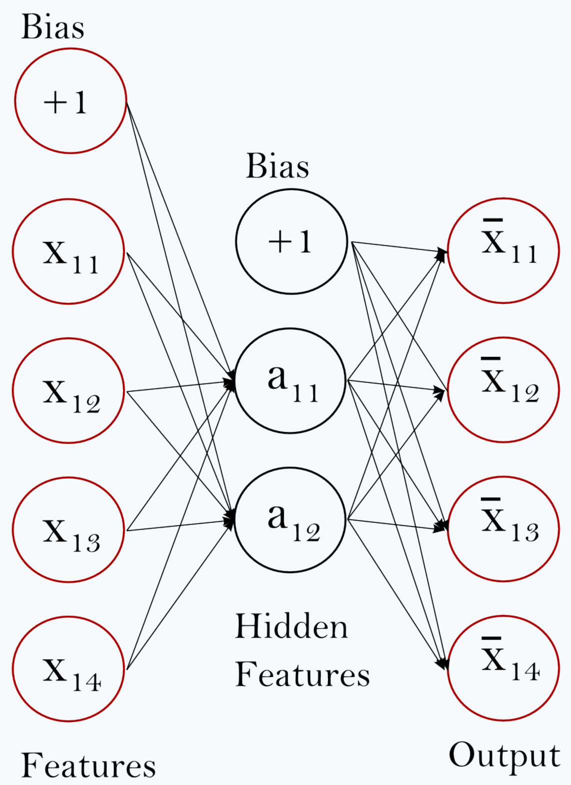
International Laboratory for Brain, Music and Sound Research (BRAMS)
Département de Psychologie, Université de Montréal
Montréal, Canada

Introduction

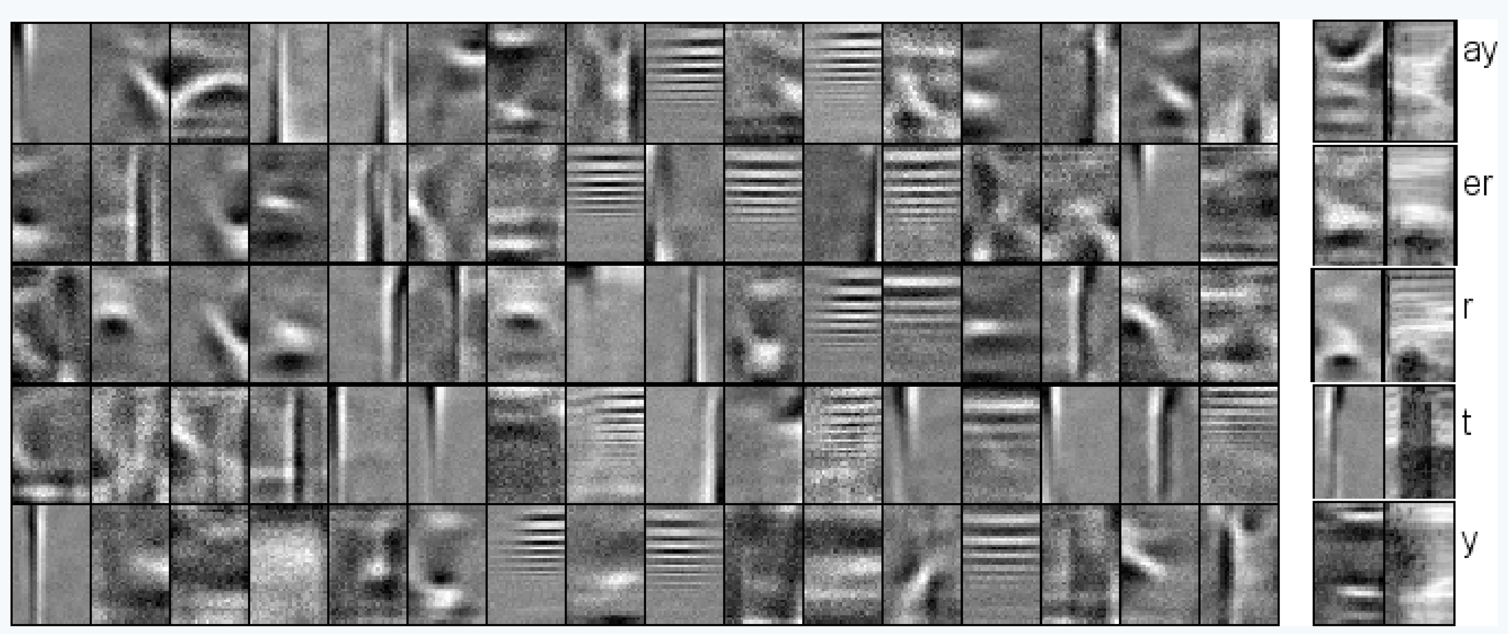
This poster introduces representation learning, an area of research of machine learning, and summarizes recent work using deep, artificial neural networks to study sensory information processing. We present ideas for future research combining representation learning, computational modeling, and neuroimaging.

Representation learning

Recent advances in the field of representation learning (including *deep learning*) have enabled significant progress on various machine learning tasks including object and speech recognition. These methods automatically learn intermediate representations of a signal of interest (e.g. image, speech, music) that maximize performance on a given task. This distinguishes representation learning from other machine learning approaches which rely on "hand-coded" features which have been carefully designed using domain knowledge. Much of this line of research is motivated by knowledge about biological neural networks and the manner in which humans learn and encode information. Common architectures include regularized auto-encoders, restricted Boltzmann machines, convolutional neural networks (CNN), deep belief networks and other artificial neural networks (ANNs).



A simple autoencoder

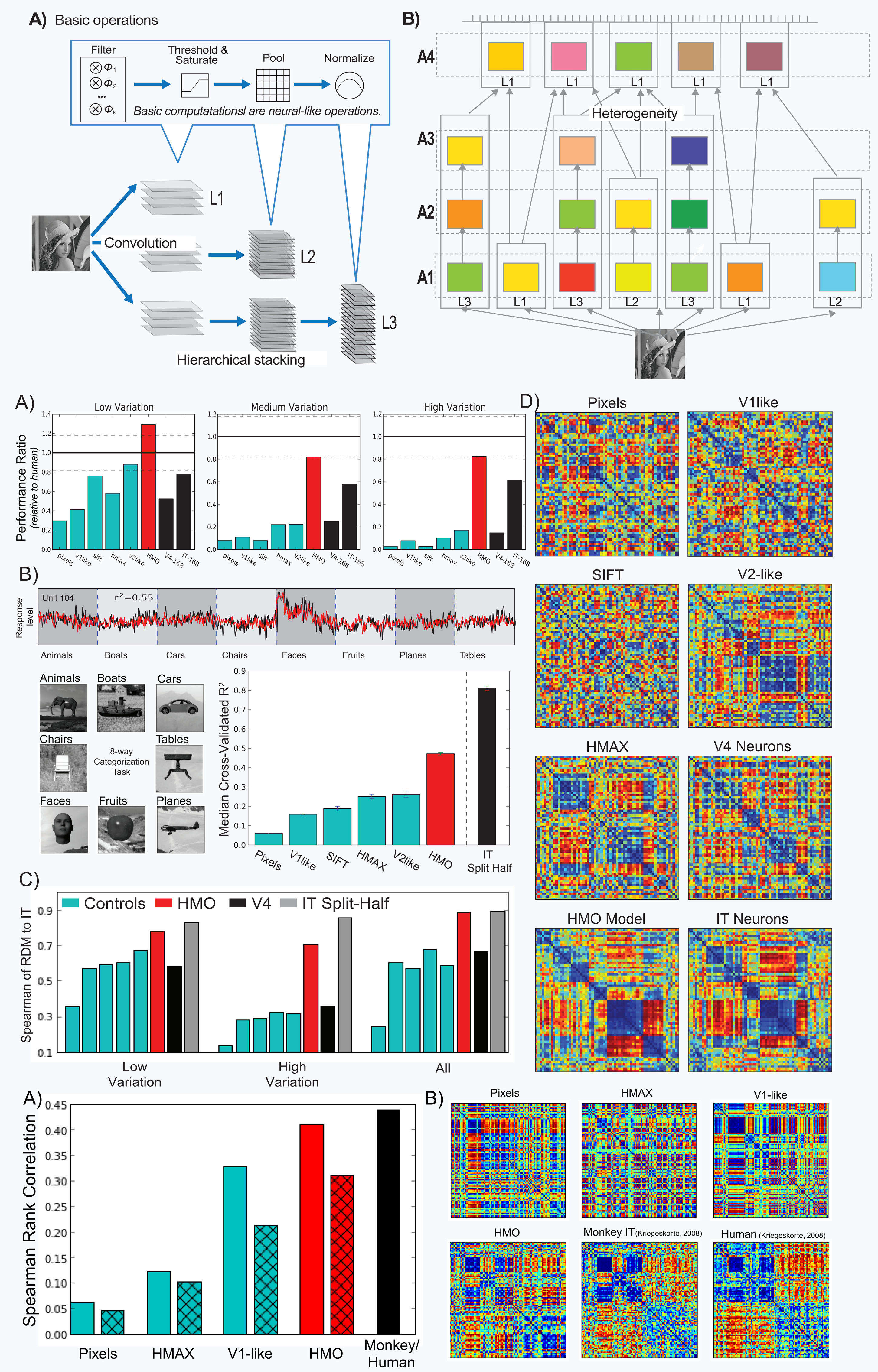


The figure above shows a random subset of features learned in an unsupervised way using a sparse auto-encoder neural network with rectified linear units and natural speech as input. The vertical axis is over frequencies and horizontal axis is over time (26 frames, 10ms long each). The right panel shows examples from the validation set.

Images copied from Zeiler, M., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., ... Hinton, G. E. (2013). On rectified linear units for speech processing. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.

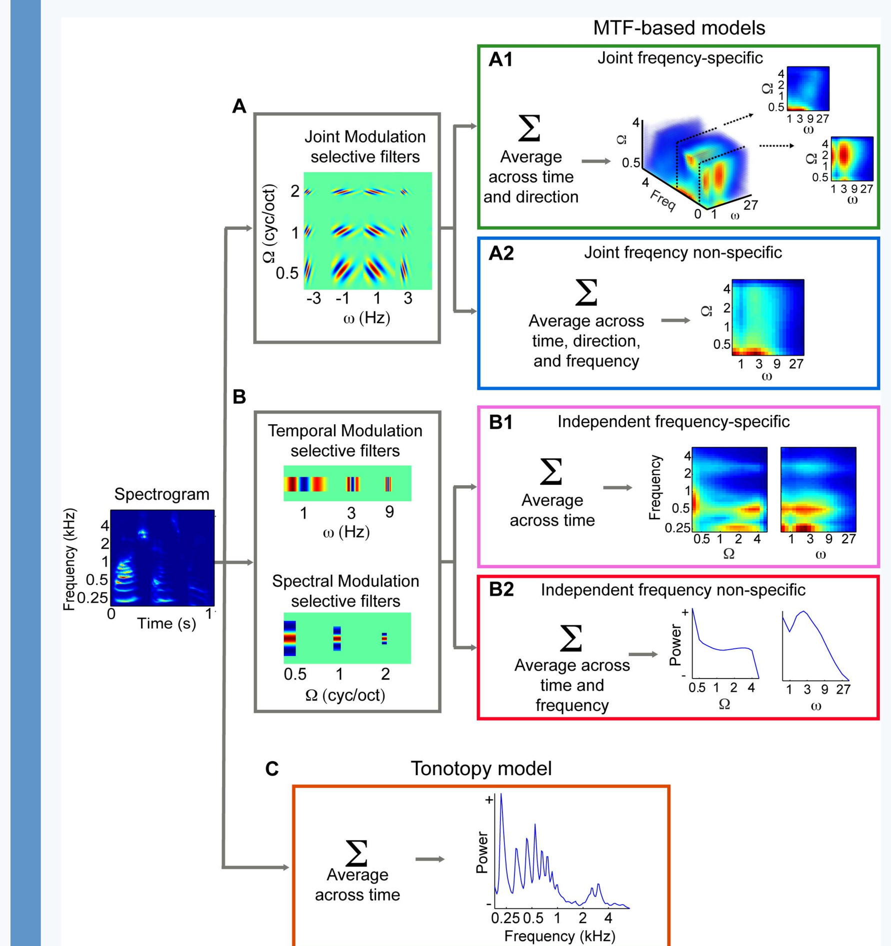
Convolutional networks achieves representations similar to macaque IT and human ventral stream

Features learned using a deep, CNN trained for object recognition have recently been compared to recordings from macaque IT and human ventral stream during visual stimulation. A representation similarity analysis (RSA) revealed that the similarity structure of the features learned by the CNN was significantly more similar to that of the brain data than any other model, suggesting that the brain may perform similar information processing.

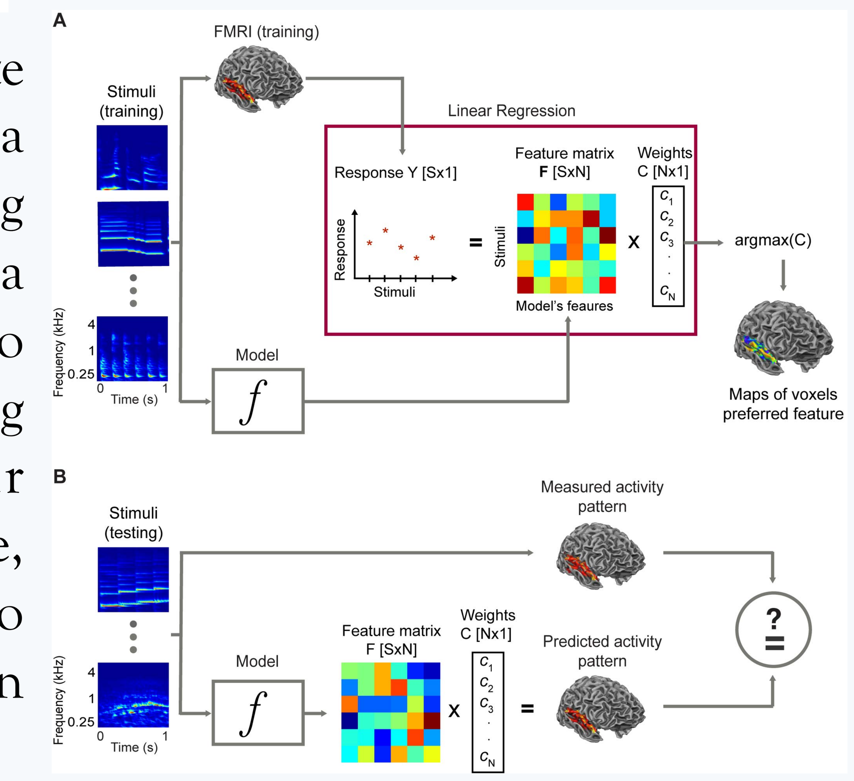


Images copied from Yamins, D., Cadieu, C., & Dicarlo, J. J. (2013). Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. Neural Information Processing Systems (NIPS).

Modeling the encoding of natural sounds



Recent work by Formisano et al. (2014) suggested the combination of computational modeling, machine learning and neuroimaging to pursue a more mechanistic understanding of human sensory processing. In the auditory domain, the modeling would typically consist of explicit transformations of spectral and temporal information. Each model represents a different hypothesis about neural information processing.



The authors proposed to evaluate their encoding models using a standard machine learning paradigm. fMRI responses to a variety of natural sounds are used to estimate the encoding models using regularized multivariate linear regression. To evaluate performance, each encoding model is used to predict brain responses to an independent set of stimuli.

Images copied from Santoroa, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. PLOS Computational Biology, 10(1), e1003412.

Future Directions

In the above description, hypotheses about auditory information processing are embedded in the specific audio features passed to the linear regression. We propose to replace the explicit audio features with representation learning models where hypotheses about neural information processing are embedded in the architectural design of the representation learning model (e.g. convolutional, feature pooling, non-linear response characteristics). Such representation learning models could be trained to perform tasks such as speech reconstruction, recognition, or synthesis and then used to calculate features or to generate stimuli. The underlying assumption behind this approach is that the brain uses intermediate representations that are optimized for complex auditory behaviours. We argue that representation learning offers valuable tools to the study of neural information processing.