Precision Brain Encoding Under Naturalistic Conditions

Dora Gozukara (dora.gozukara@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour,
Radboud University,
Thomas van Aquinostraat 4, 6525 GD
Nijmegen, Netherlands
Djamari Oetringer (djamari.oetringer@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour,
Radboud University,
Thomas van Aquinostraat 4, 6525 GD
Nijmegen, Netherlands
Linda Geerligs (linda.geerligs@donders.ru.nl)*
Donders Institute for Brain, Cognition and Behaviour,
Radboud University,
Thomas van Aquinostraat 4, 6525 GD
Nijmegen, Netherlands
Umut Güçlü (umut.gueclue@donders.ru.nl)*
Donders Institute for Brain, Cognition and Behaviour,
Radboud University,
Thomas van Aquinostraat 4, 6525 GD
Nijmegen, Netherlands

Abstract:

Convolutional Neural Networks (CNNs) are often used as a model of the visual system. Using CNN features to train brain encoding models requires a lot of data and conventional modelling practices also require these data to be collected under controlled conditions. By enhancing our models with additional measures, such as eye-tracking and receptive field maps, we can use data from more ecologically valid tasks such as free movie viewing, while decreasing the amount of data needed. Here, we showcase this by training precision brain encoding models on the Study Forrest dataset. Combining the population receptive field estimate of a voxel with eye-tracking data at each frame, we create subject and voxel specific feature timeseries by sampling only the relevant parts of the CNN feature map for only the relevant timepoints for a given voxel. We show that our precision encoders overperform conventional models and enable encoding under naturalistic viewing conditions.

Keywords: brain encoding; artificial neural networks; convolutional neural networks; precision encoding; naturalistic neuroscience

Introduction

Convolutional Neural Networks (CNNs) have had great success in modeling the visual processing in the brain thanks to their architectural similarities. Just like in the brain, neurons in convolutional neural networks (CNNs) have receptive fields, which means that they only respond to features in a particular part of the visual field. However, this selectivity is typically ignored in studies that use brain encoding models to predict brain activity from CNN features. Features from an entire CNN layer are often grouped together to predict the activity in a specific brain voxel, regardless of their spatial selectivity or size (Guclu & van Gerven, 2015; Horikawa & Kamitani, 2017; Wen et al., 2018). This means that brain-encoding models have large parameter spaces because they need to learn which spatial location is relevant for the voxel to which they are being fitted. These large parameter spaces in turn require large amounts of data to be able to achieve a good model fit. In addition, in such a large parameter space, it is likely that there are multiple combinations of features and locations that could explain the signal of a voxel, therefore the model cannot be guaranteed to learn the correct set of features. Here we propose to leverage this spatial selectivity and include voxel population receptive field (pRF) estimates in our encoding models.

Another way we might be able to build better and more ecologically valid brain encoding models is by using data in which participants are able to freely move their eyes. Nearly all studies that build brain-encoding models, and in fact most studies in cognitive neuroscience, use gaze fixation to control where the subject is directing their attention to and minimize extraneous eye movements. This allows for precise measurement of neural activity in response to visual stimuli presented in specific locations, but it also has important downsides. Because perception is an active process that naturally includes eye movements, fixating the eyes on a specific location is a very unnatural way of viewing. Even when eye-tracking data is included in studies, it is mostly used to check if participants indeed fixate. Here we propose to use eye-tracking data to eliminate fixation as a constraint altogether and include information about the gaze location in the feature-selection stage of model building.

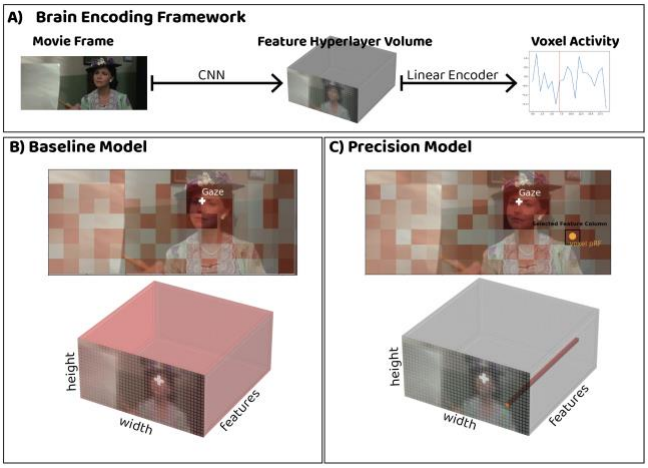

Figure 1: A) The summary of the encoder framework. CNN layer features are combined in a hyperlayer volume and used to predict voxel activity. B) The baseline model uses the entire hyperlayer feature volume, as indicated by red. C) The precision model selects a unique hyperlayer column using pRF estimates of the voxel and gaze location for each frame and subject, as indicated by red.

Methods

To demonstrate precision encoding models, we used the publicly available Study Forrest dataset (Hanke et al., 2016). This fMRI dataset includes a total of 2 hours of fixation-free movie watching runs, eye-tracking data for the duration of the movie, and a retinotopic mapping run for each subject that allows us to estimate the receptive field of each voxel. We used a convolutional neural network (vgg-19) that was pre-trained to classify natural images and extracted its features in response to movie frames from all convolutional layers. As a result, we obtained a timeseries for each feature in all layers at each spatial location in response to the movie. Combining the population receptive field estimate of a voxel with eye-tracking locations at each frame, we created subject and voxel specific feature timeseries by sampling only the relevant parts of the visual scene for only the relevant timepoints for a given voxel. This

sampled timeseries is about 30.000 times smaller than the entire layer map for the biggest layer, and about 112 times smaller for the smallest layer. We then trained a linear encoder that tries to learn the best combination of features in this timeseries that predicts the activity of a voxel. Our proposed pipeline not only reduces the size of the training data needed to be able to achieve meaningful performance, but also captures the inter- and intra-individual variability introduced by the freely moving eyes.

Results

Our precision models overperform the baseline model for the majority voxels within the visual hierarchy. Additionally, there is very little correlation between the performances of both models, showing that the voxels they fit well to are different. This can be seen in the scatterplot in Figure 2. Complimentarily, Figure 3 shows that even though we see performance gains across the whole visual stream, the most consistent gain comes from areas that are in the middle of visual hierarchy, such as the lingual and fusiform gyri.

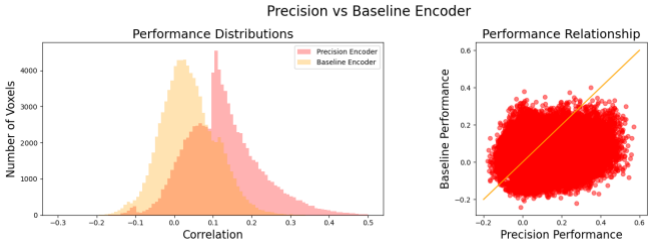

Figure 2: On the left, a histogram of Pearson's correlation coefficients over all significant voxel models (same voxels are shown for all models). On the right, scatterplot of Pearson's correlation coefficients for both models. Each dot is one voxel.

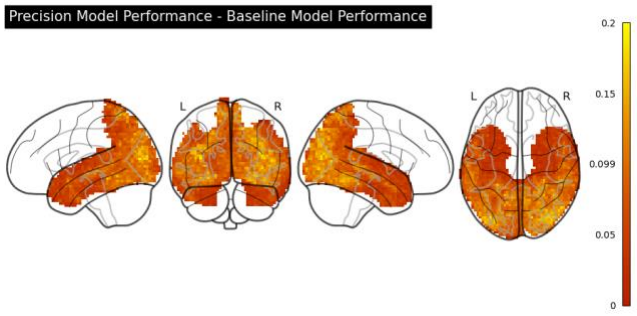

Figure 3: Brain map showing the precision model performance gain over baseline model (precision score – baseline score).

Discussion

We demonstrate that by building precision brain encoding models, we can use less data while simultaneously improving encoding performance. By allowing for free-viewing, this also opens the way for building brain-encoding models with data collected in more naturalistic settings. A general shift in more precise and personalized encoding models that take inter- and intra-individual variances into account will prove to be valuable a practice in brain encoding and decoding, especially using artificial neural networks.

References

Guclu, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., & Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.92

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 15037. https://doi.org/10.1038/ncomms15037

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, *28*(12), 4136–4160. https://doi.org/10.1093/cercor/bhx268