Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Transferring and generalizing deep-learning-based neural encoding models across subjects

Haiguang Wen [b,c], Junxing Shi [b,c], Wei Chen [d], Zhongming Liu [a,b,c,*]

[a] Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA
[b] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
[c] Purdue Institute for Integrative Neuroscience, Purdue University, West Lafayette, IN, USA
[d] Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota Medical School, Minneapolis, MN, USA

ARTICLE INFO

Keywords:
Neural encoding
Natural vision
Deep learning
Bayesian inference
Incremental learning

ABSTRACT

Recent studies have shown the value of using deep learning models for mapping and characterizing how the brain represents and organizes information for natural vision. However, modeling the relationship between deep learning models and the brain (or encoding models), requires measuring cortical responses to large and diverse sets of natural visual stimuli from single subjects. This requirement limits prior studies to few subjects, making it difficult to generalize findings across subjects or for a population. In this study, we developed new methods to transfer and generalize encoding models across subjects. To train encoding models specific to a target subject, the models trained for other subjects were used as the prior models and were refined efficiently using Bayesian inference with a limited amount of data from the target subject. To train encoding models for a population, the models were progressively trained and updated with incremental data from different subjects. For the proof of principle, we applied these methods to functional magnetic resonance imaging (fMRI) data from three subjects watching tens of hours of naturalistic videos, while a deep residual neural network driven by image recognition was used to model visual cortical processing. Results demonstrate that the methods developed herein provide an efficient and effective strategy to establish both subject-specific and population-wide predictive models of cortical representations of high-dimensional and hierarchical visual features.

Introduction

An important area in computational neuroscience is developing encoding models to explain brain responses given sensory input (Trappenberg, 2009). In vision, encoding models that account for the complex and nonlinear relationships between natural visual inputs and evoked neural responses can shed light on how the brain organizes and processes visual information through neural circuits (Paninski et al., 2007; Naselaris et al., 2011; Chen et al., 2014; Cox and Dean, 2014; Kriegeskorte, 2015). Existing models may vary in the extent to which they explain brain responses to natural visual stimuli. For example, Gabor filters or their variations explain the neural responses in the primary visual cortex but not much beyond it (Kay et al., 2008; Nishimoto et al., 2011). Visual semantics explain the responses in the ventral temporal cortex but not at lower visual areas (Naselaris et al., 2009; Huth et al., 2012). On the other hand, brain-inspired deep neural networks (DNN) (LeCun et al., 2015), mimic the feedforward computation along the visual hierarchy (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Kietzmann et al., 2017; van Gerven, 2017), match human performance in image recognition (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016), and explain cortical activity over nearly the entire visual cortex in response to natural visual stimuli (Yamins et al., 2014; Güçlü and van Gerven, 2015b; a; Wen et al., 2017, 2018; Eickenberg et al., 2017; Seeliger et al., 2017; Han et al., 2017; Shi et al., 2018).

These models also vary in their complexity. In general, a model that explains brain activity in natural vision tends to extract a large number of visual features given the diversity of the visual world and the complexity of neural circuits. For DNN, the feature space usually has a very large dimension in the order of millions (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016). Even if the model and the brain share the same representations up to linear transform (Yamins and DiCarlo, 2016), matching such millions of features onto billions of neurons or tens of thousands of neuroimaging voxels requires substantial data to sufficiently sample the feature space and

* Corresponding author. College of Engineering, Purdue University, 206 S. Martin Jischke Dr., West Lafayette, IN 47907, USA.
E-mail address: zmliu@purdue.edu (Z. Liu).

https://doi.org/10.1016/j.neuroimage.2018.04.053
Received 21 April 2018; Accepted 23 April 2018
Available online 27 April 2018

reliably train the transformation from the feature model to the brain. For this reason, current studies have focused on only few subjects while training subject-specific encoding models with neural responses observed from each subject given hundreds to thousands of natural pictures (Güçlü and van Gerven, 2015b; Eickenberg et al., 2017; Seeliger et al., 2017), or several to tens of hours of natural videos (Güçlü and van Gerven, 2015a; Wen et al., 2017, 2018; Eickenberg et al., 2017; Shi et al., 2018). However, a small subject pool incurs concerns on the generality of the conclusions drawn from such studies. Large data from single subjects are rarely available and difficult to collect especially for patients and children. It is thus desirable to transfer encoding models across subjects to mitigate the need for a large amount of training data from single subjects.

Transferring encoding models from one subject to another should be feasible if different subjects share similar cortical representations of visual information. Indeed, different subjects show similar brain responses to the same natural visual stimuli (Hasson et al., 2004; Lu et al., 2016), after their brains are aligned anatomically. The consistency across subjects may be further improved by functional alignment of fine-grained response patterns (Haxby et al., 2011; Conroy et al., 2013). Recent studies have also shown that encoding (Güçlü and van Gerven, 2015b; Wen et al., 2017) or decoding (Raz et al., 2017; Wen et al., 2017) models trained for one subject could be directly applied to another subject for reasonable encoding and decoding accuracies. Whereas these findings support the feasibility of transferring encoding and decoding models from one subject to another, it is desirable to consider and capture the individual variations in functional representations. Otherwise, the encoding and decoding performance is notably lower when the models are trained and tested for different subjects than for the same subject (Wen et al., 2017).

Beyond the level of single subjects, what is also lacking is a method to train encoding models for a group by using data from different subjects in the group. This need rises in the context of "big data", as data sharing is increasingly expected and executed (Teeters et al., 2008; Van Essen et al., 2013; Paltoo et al., 2014; Poldrack and Gorgolewski, 2014). For a group of subjects, combining data across subjects can yield much more training data than are attainable from a single subject. A population-wise encoding model also sets the baseline for identifying any individualized difference within a population. However, training such models with a very large and growing dataset as a whole is computationally inefficient or even intractable with the computing facilities available to most researchers (Fan et al., 2014).

Here, we developed methods to train DNN-based encoding models for single subjects or multiple subjects as a group. Our aims were to 1) mitigate the need for a large training dataset for each subject, and 2) efficiently train models with big and growing data combined across subjects. To achieve the first aim, we used pre-trained encoding models as the prior models in a new subject, reducing the demand for collecting extensive data from the subject in order to train the subject-specific models. To achieve the second aim, we used incremental learning algorithm (Fontenla-Romero et al., 2013) to adjust an existing encoding model with new data to avoid retraining the model from scratch with the whole dataset. To further leverage both strategies, we employed functional hyper-alignment (Guntupalli et al., 2016) between subjects before transferring encoding models across subjects. Using experimental data for testing, we showed the merits of these methods in training the DNN-based encoding models to predict functional magnetic resonance imaging (fMRI) responses to natural movie stimuli in both individual and group levels.

Methods and materials

Experimental data

In this study, we used the video-fMRI data from our previous studies (Wen et al., 2017, 2018). The fMRI data were acquired from three human subjects (Subject JY, XL, and XF, all female, age: 22–25, normal vision) when watching natural videos. The videos covered diverse visual content representative of real-life visual experience.

For each subject, the video-fMRI data was split into three independent datasets for 1) functional alignment between subjects, 2) training the encoding models, and 3) testing the trained models. The corresponding videos used for each of the above purposes were combined and referred to as the "alignment" movie, the "training" movie, and the "testing" movie, respectively. For Subjects XL and XF, the alignment movie was 16 min; the training movie was 2.13 h; the testing movie was 40 min. To each subject, the alignment and training movies were presented twice, and the testing movie was presented ten times. For Subject JY, all the movies for Subjects XL and XF were used; in addition, the training movie also included 10.4 h of new videos not seen by Subjects XL and XF, which were presented only once.

Despite their different purposes, these movies were all split into 8-min segments, each of which was used as continuous visual stimuli during one session of fMRI acquisition. The stimuli $(20.3^\circ \times 20.3^\circ)$ were delivered via a binocular goggle in a 3-T MRI system. The fMRI data were acquired with 3.5 mm isotropic resolution and 2 s repetition time, while subjects were watching the movie with eyes fixating at a central cross. Structural MRI data with T_1 and T_2 weighted contrast were also acquired with 1 mm isotropic resolution for every subject. The fMRI data were preprocessed and co-registered onto a standard cortical surface template (Glasser et al., 2013). More details about the stimuli, data acquisition and preprocessing are described in our previous papers (Wen et al., 2017, 2018).

Nonlinear feature model based on deep neural network

The encoding models took visual stimuli as the input, and output the stimulus-evoked cortical responses. As shown in Fig. 1, it included two steps. The first step was a nonlinear feature model, converting the visual input to its feature representations; the second step was a voxel-wise linear response model, projecting the feature representations onto the response at each fMRI voxel (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Huth et al., 2012; Güçlü and van Gerven, 2015b; a; Wen et al., 2017, 2018; Eickenberg et al., 2017; Seeliger et al., 2017; Han et al., 2017; Shi et al., 2018). The feature model is described in this sub-section, and the response model is described in the next sub-section.

In line with previous studies (Güçlü and van Gerven, 2015b; a; Wen et al., 2017, 2018; Eickenberg et al., 2017; Seeliger et al., 2017), a deep neural network (DNN) was used as the feature model to extract hierarchical features from visual input. Our recent study (Wen et al., 2018) has demonstrated that deep residual network (ResNet) (He et al., 2016), a specific version of the DNN, was able to predict the fMRI responses to videos with overall high and statistically significant accuracies throughout the visual cortex. Therefore, we used ResNet as an example of the feature model in the present study for transferring and generalizing encoding models across subjects. Briefly, ResNet was pre-trained for image recognition by using the ImageNet dataset (Deng et al., 2009) with over 1.2 million natural images sampling from 1000 categories, yielding 75.3% top-1 test accuracy. The ResNet consisted of 50 hidden layers of nonlinear computational units that encoded increasingly abstract and complex visual features. The first layer encoded location and orientation-selective visual features, whereas the last layer encoded semantic features that supported categorization. The layers in between encoded increasingly complex features through 16 residual blocks. Passing an image into ResNet yielded an activation value at each unit. Passing every frame of a movie into ResNet yielded an activation time series at each unit, indicating the time-varying representation of a specific feature in the movie. In this way, the feature representations of the training and testing movies could be extracted, as in previous studies (Wen et al., 2017, 2018). Here, we extracted the features from the first layer, the last layer, and the output layer for each of the 16 residual blocks in ResNet (Wen et al., 2018).