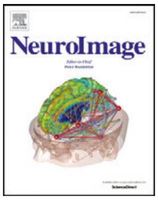
NeuroImage 145 (2017) 329–336

Contents lists available at ScienceDirect

NeuroImage

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



Increasingly complex representations of natural movies across the dorsal stream are shared between subjects

Umut Güçlü *, Marcel A.J. van Gerven

Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Accepted 22 December 2015
Available online 24 December 2015

Keywords:
Encoding
Decoding
Deep neural network
Dorsal stream
Hyperalignment

ABSTRACT

Recently, deep neural networks (DNNs) have been shown to provide accurate predictions of neural responses across the ventral visual pathway. We here explore whether they also provide accurate predictions of neural responses across the dorsal visual pathway, which is thought to be devoted to motion processing and action recognition. This is achieved by training deep neural networks to recognize actions in videos and subsequently using them to predict neural responses while subjects are watching natural movies. Moreover, we explore whether dorsal stream representations are shared between subjects. In order to address this question, we examine if individual subject predictions can be made in a common representational space estimated via hyperalignment. Results show that a DNN trained for action recognition can be used to accurately predict how dorsal stream responds to natural movies, revealing a correspondence in representations of DNN layers and dorsal stream areas. It is also demonstrated that models operating in a common representational space can generalize to responses of multiple or even unseen individual subjects to novel spatio-temporal stimuli in both encoding and decoding settings, suggesting that a common representational space underlies dorsal stream responses across multiple subjects.

Introduction

The human visual system is devoted to the analysis of increasingly complex properties of our environment as one moves from upstream to downstream visual areas. Traditionally, the ventral visual pathway is hypothesized to be devoted to object recognition and the dorsal visual pathway is thought to be devoted to motion processing and action recognition (Mishkin et al., 1983; Haxby et al., 1991; Goodale and Milner, 1992).

An important question is what stimulus properties are processed as one traverses these pathways toward more downstream areas. Recently, we have shown that deep neural networks (DNNs) (Schmidhuber, 2015; LeCun et al., 2015) can be used to predict with high accuracy how voxels in different areas of the ventral stream respond to naturalistic stimuli (Güçlü and van Gerven, 2015). Moreover, this analysis revealed that artificial neurons in deeper hidden layers of the neural network gave better predictions for more downstream areas.

It remains unclear, however, whether DNNs can also be used to accurately predict neural responses across the dorsal stream up to and including area MT. Furthermore, if this property holds, an interesting secondary question is whether representations in particular visual areas are highly individualized or rather shared between subjects. If the latter is the case, then it may be possible to predict neural responses in a particular subject using computational models that are estimated using data from other subjects (Yamada et al., 2015). Furthermore, if such a common representational space exists, decoding of stimuli from observed neural responses can be improved by combining data from multiple subjects.

The current paper addresses these questions using a sophisticated computational model, commonly referred to as an encoding model (Naselaris et al., 2011). The encoding model, depicted in Fig. 1, consists of a deep convolutional neural network (Fukushima, 1980) that nonlinearly maps stimuli to their constituent features, as well as a response model that linearly maps features to observed blood–oxygen-level-dependent (BOLD) responses.

The deep neural network was trained using tens of thousands of action videos, yielding spatio-temporal filters that are important for action recognition, ostensibly yielding a representation suitable for probing dorsal stream responses. The linear mapping was estimated using data by (Nishimoto et al., 2011) in which subjects were watching natural movies. Estimation proceeded by first mapping data from different number of subjects to a common representational space and then averaging responses across subjects (Haxby et al., 2011). Next, deep neural network features were regressed onto averaged responses.

Using this framework, we were able to show (1) the existence of a correspondence between DNN layers and dorsal stream areas of individual subjects such that deeper layers better predict downstream areas and (2) the existence of a common representational space that can facilitate the estimation of common models for individual subject prediction such that responses of individual subjects to novel spatio-

* Corresponding author.
E-mail address: u.guclu@donders.ru.nl (U. Güçlü).

http://dx.doi.org/10.1016/j.neuroimage.2015.12.036
1053-8119/© 2015 Elsevier Inc. All rights reserved.

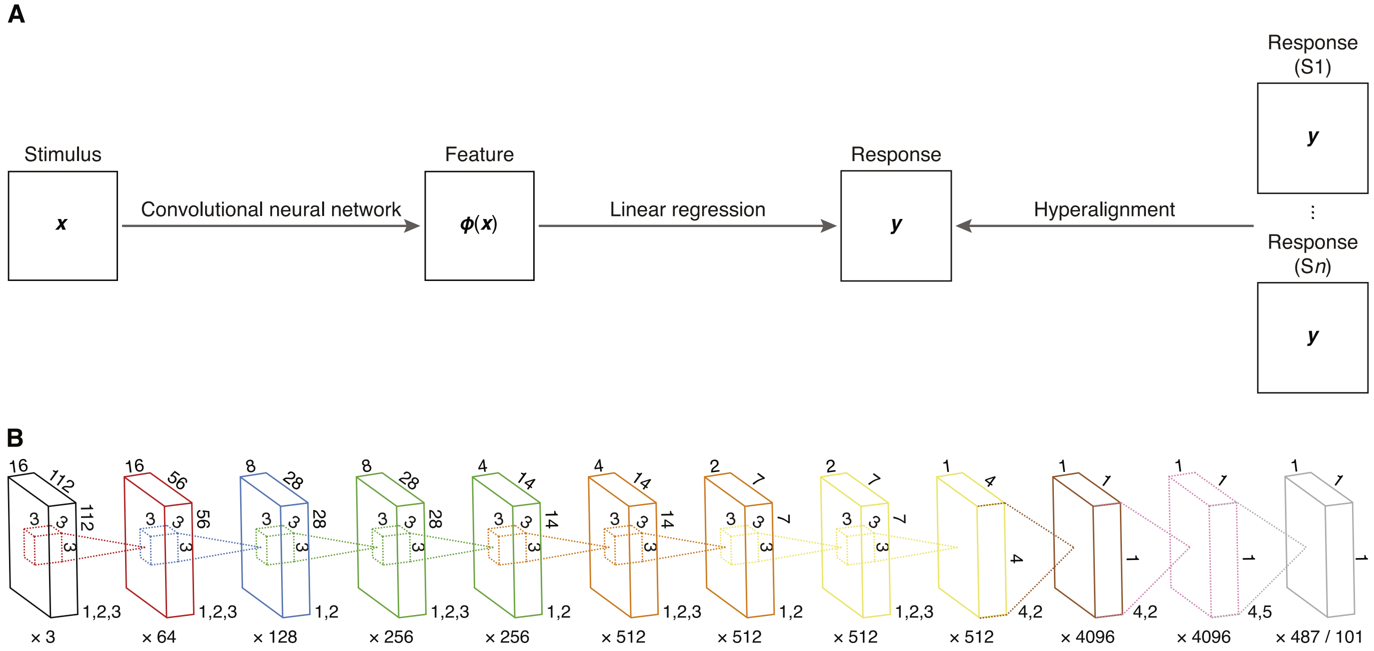


Fig. 1. Framework that combines feature, response and representational space models. (A) Encoding model. (B) Convolutional neural network. Large boxes show a stimulus and feature maps, and numbers around them show their dimensionality. Similarly, small boxes and their projections show neurons, and numbers around them show their dimensionality. Number of feature maps and neurons in each layer is indicated below the boxes. Note that the dimensionality of a neuron in a fully-connected layer is the same as that of the feature maps in the previous layer. Each neuron filters the feature maps in the previous layer and returns the corresponding feature map in the current layer. Transformations in each layer are indicated bottom-right of the boxes: 1. Convolution. 2. Rectifier. 3. Max pooling. 4. Dot product. 5. Softmax function.

temporal stimuli can be predicted with models estimated from responses of other subjects in both encoding and decoding settings.

Material and methods

Data set

We used the vim-2 data set (Nishimoto et al., 2014), which was originally published in Nishimoto et al. (2011). The experimental procedures are identical to those in Nishimoto et al. (2011). Briefly, the data set has twelve 600-s blocks of stimulus–response pairs in a training set and nine 60-s blocks of stimulus–response pairs in a test set. Stimuli are videos (128 px × 128 px or 20° × 20°, 15 FPS) that were drawn from various sources. Responses are BOLD responses (voxel size $= 2 \times 2 \times 2.5$ mm^3, TR $= 1$ s) that were acquired from occipital cortices of three subjects (S1, S2 and S3). Stimuli in the test set were repeated ten times. Responses in the test set were averaged across repetitions.

Stimuli in the data set were spatially downsampled to 112 px × 112 px and temporally upsampled to 16 FPS. Responses in the data set have already been preprocessed as described in (Nishimoto et al., 2011). Briefly, they have been realigned to compensate for motion, detrended to compensate for drift and z-scored. Additionally, the first six seconds of the blocks were discarded. No further preprocessing was performed.

Regions of interests were localized using the multifocal retinotopic mapping technique on retinotopic mapping data that were acquired in separate sessions (Hansen et al., 2004). We restricted our analyses to dorsal stream visual areas (V1, V2, V3, V3A, V3B and MT).

Hyperalignment

In addition to analyzing the data in the individual representational spaces, we analyzed them in a common representational space (Haxby et al., 2011). A representational space model that uses Procrustes transformation for hyperaligning the data of the individual subjects to the common representational space was estimated from the training set per cerebral hemisphere and visual area as follows: the common representational space was first set to the data of the individual subject that has the most number of voxels (Table 1). The common representational space was then iteratively updated. At each iteration, the data of the individual subjects were first projected to the common representational space. The common representational space was then set to the mean of the projections of the data of the individual subjects. After the final iteration, the data of the individual subjects were projected to the common representational space. *PyMVPA* (http://www.pymvpa.org) was used for representational space model estimation (Hanke et al., 2009).

Encoding

Feature model

We used a deep convolutional neural network for nonlinearly transforming stimuli to multiple layers of hierarchical feature representations. The architecture of the DNN is identical to the C3D architecture in (Tran et al., 2014). The architecture was developed for learning generic features for video analysis, building on previous insights in DNNs for image recognition. Here, we provide an overview

Table 1
Number of voxels per cerebral hemisphere and visual area. Bold numbers show the dimensionality of the corresponding cerebral hemisphere and visual area in the common representational space.

	Left hemisphere						Right hemisphere					
	V1	V2	V3	V3A	V3B	MT	V1	V2	V3	V3A	V3B	MT
S1	494	726	598	92	**104**	**197**	514	781	562	160	**152**	**152**
S2	470	733	**734**	135	83	83	573	**928**	**670**	**202**	140	116
S3	**653**	**746**	504	**164**	88	166	**713**	650	637	118	138	64