Unsupervised deep neural network for fMRI feedback modelling

Michele Svanera[1] and **Andrew T. Morgan** and **Lucy S. Petro** and **Lars Muckli**
Centre for Cognitive Neuroimaging
Institute of Neuroscience and Psychology
University of Glasgow
UK

Abstract

The brain is constantly dealing with two streams of information: the feedforward stream which carries sensory inputs, and the feedback stream that contains predictions derived from internal models that the brain has about the world. During a brain imaging experiment, an effective method to study the feedback stream is to occlude a portion of the image presented, thereby isolating it from the feedforward signal. The predictive coding framework suggests that the brain is trying to reconstruct the image under the occlusion; this operation is conceptually similar to the image processing task of inpainting, in which an artificial model predicts the missing image part. Using an encoder/decoder network architecture, trained to fill occlusions and reconstruct an unseen image, we investigated similarities and contradictions between brain visual pathway and artificial neural networks. We will perform comparisons between brain data collected at 3T and 7T during the vision of static images and different layers of the encoder/decoder network. These analyses will be conducted using Representational Similarity Analysis (RSA) and by creating encoding models to investigate visual pathways and V1 layers. Understanding how information is integrated together in the early visual cortex will provide insight to fundamental neuroscientific questions about human vision, cognition, and perception.

Keywords: encoder/decoder, 3T/7T brain data, RSA, encoding.

Introduction

The brain is a prediction machine. In addition to receiving sensory information, it actively generates sensory predictions. It does so by creating internal models about the world which are used to predict upcoming sensory inputs. Therefore, the brain is constantly dealing with two streams of information: the feedforward stream (carrying the sensory input) and the feedback stream (carrying the predictions). This perspective has led to a paradigm shift in cognitive neuroscience over the last two decades. One of the relevant theories on this topic is "predictive coding", which interprets the brain as an inference engine, optimising representations built by sensory input (Clark, 2015).

A number of studies have investigated the composition and interaction of the two information streams, using different brain data across different species. An effective method to investigate the problem is to apply an occlusion paradigm, where images and videos are partially occluded (Smith & Muckli, 2010; Muckli et al., 2015). This blocks the feedforward stream allowing the isolation of cortical feedback signals and lateral connections. In the brain data analysis, this is important for the retinotopic nature of V1, since there is a direct correspondence between image and cortical spaces.

There are two approaches to analyse brain data in this context: a more indirect analysis with representational similarity analyses (RSA, Kriegeskorte, Mur, and Bandettini (2008); Nili et al. (2014)) and a direct one, with encoding models. RSA tests the similarity between brain responses and other type of representations (as for example computational models or behavioural responses) obtained on the same stimuli. Instead, *encoding* models (Naselaris, Kay, Nishimoto, & Gallant, 2011) consider the association between multi-dimensional features of a stimulus and the value of each brain voxel. Using this generative encoding model, the brain response associated with a new stimulus reconstructed by pooling separate voxel responses.

Importantly, the success of both encoding models and RSA relies on the validity of the feature space used to represent the stimuli. In this respect, Deep Learning (DL) methods (in particular Convolutional Neural Networks - CNN) have become the leading methods for automatic feature learning because they provide image and movie feature representations at different degrees of abstraction. Recently, multiple advances in deep learning have made important contributions in unsupervised learning. Examples include reconstruction of unseen or damaged parts of images (termed inpainting; Pathak, Krahenbuhl, Donahue, Darrell, and Efros (2016)), image segmentation (Ronneberger, Fischer, & Brox, 2015), and representation learning (automatically discovering pattern in data). These advances have allowed the development of effective generative models such as deep encoder/decoder or generative adversarial networks (Mirza & Osindero, 2014).

The attempt to relate DL models with brain imaging data started only few years ago. Some studies which revealed interesting similarities between CNN architectures and the hierarchy of biological vision (Yamins & DiCarlo, 2016). For example, one study showed how a CNN resembles representational similarity of Inferior Temporal (IT) intra- and inter-categories (Khaligh-Razavi & Kriegeskorte, 2014). Another study (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016) described how a CNN captured the stages of human visual processing in time and space from early visual areas towards the dorsal and ven-

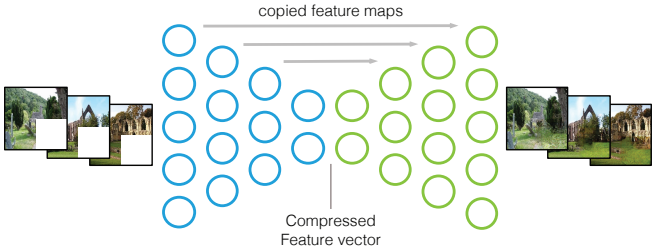
[1] Corresponding author: Michele.Svanera@glasgow.ac.uk

tral streams. In a noteworthy finding, fMRI encoding modelling work by (Güçlü & van Gerven, 2015) indicated that a stimulus decomposition based on selected layers from a pre-trained CNN outperformed the Gabor-based approach proposed in (Kay, Naselaris, Prenger, & Gallant, 2008).

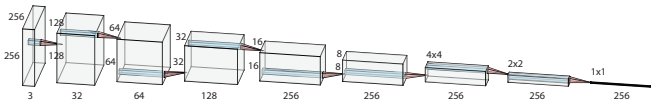
The work we propose here is tries to benefit from these generative models to gain a better understanding of how cortical prediction works. Exploiting the occlusion paradigm (where portion of the scene is occluded suppressing feed-forward signals), we aim to interpret the detected cortical feedback signals in brain data by comparing them to encoder/decoder layers trained on an inpainting task to reconstruct occluded images. This comparison, previously done between brain signals and convolutional neural network layers activations (Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2016), is now carried out with an encoding scheme using encoder/decoder layers in order to predict fMRI feedback signals.

Model training

The model we trained to fill occlusion for solving inpaiting task is a fully-convolutional neural network, with the encoder/decoder architecture, and with skip connections (known as U-Net and described in Ronneberger et al. (2015)); the model is shown in Figure 1.



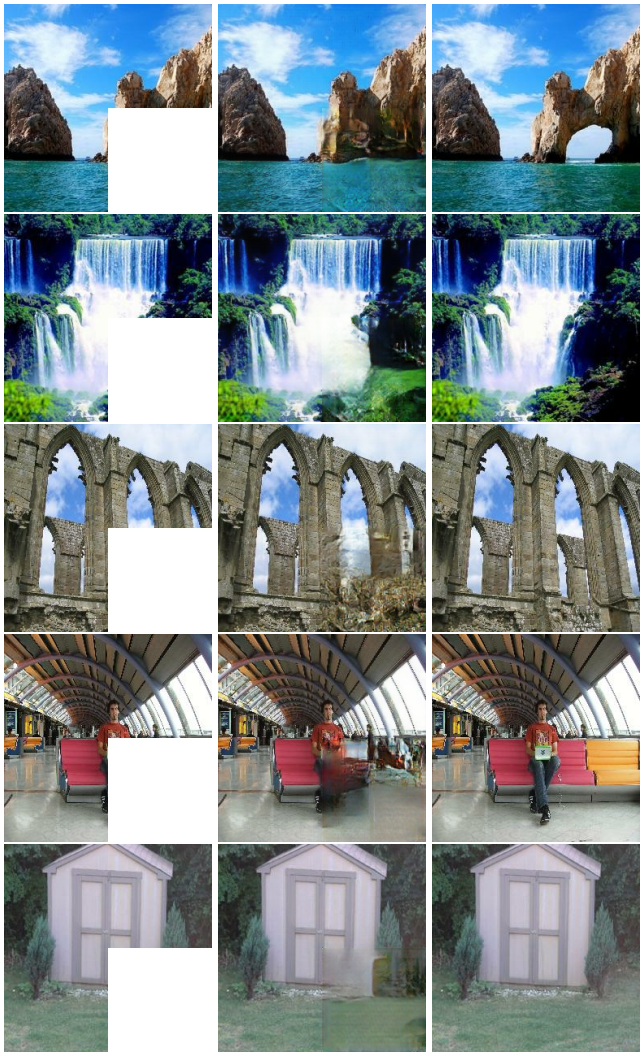
(a) Encoder/decoder



(b) Encoder details

Figure 1: Model architecture with encoder details (decoder is mirrored).

The way we trained the network is similar to what has been proposed in (Isola, Zhu, Zhou, & Efros, 2017); they proposed to learn not only the mapping from input image to output image, but also the best loss function to train this mapping. This is crucial in unsupervised learning, where labels or categories are not available; this is accomplished in their work using conditional adversarial neural networks (Goodfellow et al., 2014). The training has been carried out using images from SUN database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), with occluded images as input of the network and original images - i.e., to reconstruct - as output; some images from the database where discarded because they were too small, which led to a total of ~ 70000 images for the training. A graphical result to appreciate the quality of the output of the network is shown in Figure 2. For the experimental part, we



(a) Input (b) Output (c) Target

Figure 2: Model results with (a) the input to the network, (b) the output of the network, and (c) the target used to trained the network (i.e., the original image).

will train the network to reconstruct grayscale images with occlusion and masked with a circular aperture; images used in the experiment were not part of the train set.

Brain data

We used 3T and 7T-fMRI brain data acquired at different resolutions to investigate visual pathways and V1 layers.

3-Tesla Eighteen healthy volunteers with normal or corrected-to-normal vision participated in this study. Twenty-four real-world scenes from six categories (beaches, buildings, forests, highways, industry, and mountains), from the dataset in (Walther, Caddigan, Fei-Fei, & Beck, 2009)

were shown to participants. Images were displayed in grayscale on a rear-projection screen using a projector system. Stimuli spanned $19.5^\circ \times 14.7^\circ$ of visual angle, and were presented with the lower-right quadrant occluded by a white box (occluded region spanned $\approx 9^\circ \times 7^\circ$). A centralised fixation checkerboard marked the centre of the scene images. Over the course of the experiment, each image was presented 16 times.

Functional scanning is conducted at the Centre for Cognitive Neuroimaging, at University of Glasgow. We used EPI sequences to acquire partial brain volumes aligned to maximise coverage of the visual pathway (18 slices; voxel size: 3mm, isotropic; 0.3mm interslice gap; TR $= 1000$ms; TE $= 30$ms; matrix size $= 70 \times 64$; FOV $= 210 \times 192$mm).

7-Tesla Three healthy volunteers with normal or corrected-to-normal vision participated in this study. 384 real-world scenes were chosen from the SUN database (Xiao et al., 2010). One set was presented with the lower-right quadrant occluded by a white box, one set was presented without occlusion, and one set was presented with and without occlusion. Images were displayed in grayscale, matched for global luminance, and masked with a circular aperture which linearly faded to the background (mean grayscale across scenes) from 4.9° to 5.15° visual angle (Kay et al., 2008). Stimuli were presented on a rear-projection screen using a projector system and spanned $10.38^\circ \times 10.38^\circ$ visual angle. A centralised fixation checkerboard marked the centre of the scene images.

MRI data were collected at the University of Maastricht, Netherlands using a research-dedicated 7T Magnetom MRI system with a 32-channel head coil. High-resolution functional images were obtained using a T2*-weighted gradient echo EPI with the following parameters: echo time (TE) $= 25$ms, repetition time (TR) $= 2000$ms, iPAT-factor $= 3$, multi-band factor $= 2$, flip angle $= 75$, number of slices $= 56$, matrix $= 186 \times 186$, voxel size $= 0.8$mm isotropic. The field-of-view included occipital early visual cortex, centered on the calcarine sulcus.

Discussion

Visual stream The first goal we aim to achieve next is to replicate the study of Cichy et al. (2016) in comparing spatial visual brain representations with representations in an artificial deep neural network (DNN) through RSA. Relating 3T data (with occluded images only) and the encoder branch of the network, we want to see how DNN layers are similar to visual stream areas. The key result we hope to achieve is to reveal the hierarchy of human visual processing in space from early visual areas towards the dorsal and ventral streams, in line with results in (Cichy et al., 2016).

Our approach expands Cichy et al.'s work by using RSA to test, which DNN branch - that is, encoder or decoder - has stronger similarity with brain data representations. The semi-partial correlation may help to discriminate between contribution from one and the other branch (unique information each branch contributed to voxel prediction). In this context, an interesting question is to understand if cortical processing of the occluded area is more similar to the encoder or the decoder stream, bringing more hints on how we should interpret the predictive coding theory. How CNN layers resembles representations in Inferior Temporal (IT) cortex in humans was already shown in (Khaligh-Razavi & Kriegeskorte, 2014); our work will extend theirs to also other areas of visual pathway. In addition, using regression methods we can build encoding models of voxel responses, to evaluate the ability of the two branches in reconstructing voxel betas.

V1 layer specificity Using 7T data, we will focus attention on layer-specific processing in the visual cortex. This data was collected using static images with and without occlusion. In this context, we define feedforward models as the responses to scenes as they were presented (i.e., with the occlusion); instead feedback models are defined as model responses as they would be predicted (i.e., with no occlusion). Both datasets allow us to build encoding models that describe feedforward and feedback streams: feedforward models use occluded images and feedback models using non-occluded images. The encoding will be conducted at the voxel level, modelling voxel-specific predictors through ridge regression with regularisation (Formisano, De Martino, & Valente, 2008). Understanding how every DNN layer contributes to the prediction of layer-wise activity, analysing encoding and decoding branch contributions, will give us an insight into the cortex organisation at the layer-wise level. In terms of similarities, RSA analyses can give us additional information on the same questions.

Network architecture An additional analysis will compare different architectures of the trained artificial network. In this study, we trained a network to reconstruct images using skip connections. These connections guarantee a better reconstruction of output image, but it is not trivial to evaluate what would happen in the prediction of voxel responses. The previous analyses can therefore be conducted analysing different results using different network architectures: with and without skip connections.

Acknowledgement

This work was supported by the Human Brain Project (EU grant 604102) and the European Research Council (ERC StG 2012_311751-BrainReadFBPredCode), both awarded to L.M.

References

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fmri time series: classification and regres-

sion of brain responses using machine learning. *Magnetic resonance imaging*, *26*(7), 921–934.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2672–2680). Curran Associates, Inc.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*(27), 10005-10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, *10*(11), e1003915.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., ... Yacoub, E. (2015). Contextual feedback to superficial layers of v1. *Current Biology*, *25*(20), 2690–2695.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400 - 410. (Multivariate Decoding and Brain Reading) doi: http://dx.doi.org/10.1016/j.neuroimage.2010.07.073

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, *10*(4), e1003553.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2536–2544).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences*, *107*(46), 20099–20103. Retrieved from http://www.pnas.org/content/107/46/20099 doi: 10.1073/pnas.1000233107

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of neuroscience*, *29*(34), 10573–10581.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (cvpr), 2010 ieee conference on* (pp. 3485–3492).

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.