Using deep neural network features to predict voxelwise activity in ultra-high field fMRI

Rebekka Heinen (rebekka.heinen@ruhr-uni-bochum.de)
Dept. of Neuropsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum
Universitätsstrasse 150, 44801 Bochum, Germany

Lorena Deuker (lorena.deuker@ruhr-uni-bochum.de)
Dept. of Neuropsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum
Universitätsstrasse 150, 44801 Bochum, Germany

Thomas Naselaris (tnaselar@musc.edu)
Medical University of South Carolina, Department of Neuroscience, 173 Ashley Avenue
Charleston, SC 29425, USA

Nikolai Axmacher (nikolai.axmacher@ruhr-uni-bochum.de)
Dept. of Neuropsychology, Institute of Cognitive Neuroscience, Ruhr University Bochum
Universitätsstrasse 150, 44801 Bochum, Germany

Abstract

Deep neural network features can be used to train encoding models that accurately predict brain activity from the visual cortex. Using these features together with ultra-high field fMRI could open a new set of opportunities ranging from human vision to areas such as learning and memory consolidation. Is it possible to apply encoding models based on deep neural network features to high-resolution fMRI data? We investigated this using the feature-weighted receptive field (fwrf) model on ultra-high field fMRI during a natural image viewing task. Applying the fwrf model to our data we were able to predict brain activity along the ventral visual stream (VVS). In line with previous studies, we found a shift from low to high network layers while predicting brain activity in early visual areas compared to higher regions of the VVS. We conclude that encoding models based on neural network features can be applied to ultra-high field fMRI data, suggesting similar processing of visual scenes in neural networks and the human visual association cortex. Our results suggest that these models cannot only be used to study vision but other processes such as memory and imagination.

Keywords: fMRI; ultra-high field; encoding models; deep neural networks

Introduction

Every day we experience constant visual input until we go to sleep. While receiving this input we need to process it, react to it and/or store it into our memory. One question in cognitive neuroscience is to explore how the brain is able to perform this continuous task and specifically how visual input is represented on a neural level.

The first step to investigate this question would be to predict how the brain will react to visual input. Over the past years encoding models emerged as new tools to study human vision and the underlying neural processes by enabling researchers to predict brain activity based on stimulus features (Kay, Naselaris, Prenger, & Gallant, 2008). Additionally, neural networks from computer vision provide insight into how visual input is processed as they show a similar processing hierarchy as visual processing along the visual pathway (Güçlü & van Gerven, 2015). St-Yves and Naselaris (2018) developed a method that combines neural networks from vision with brain data from functional MRI into one model. Using their feature-weighted receptive field model (fwrf) they were able to accurately predict brain activity based on encoding models trained on brain activity and neural network features.

Currently, the main focus of neural network based encoding models lies in vision neuroscience. To facilitate their use in other fields of research such as memory or imagination, the use of higher MRI field strength for higher spatial resolution and higher sensitivity may have strong advantages compared to lower field strengths (Dumoulin et al., 2018). In addition, these advantages enable the use of larger field of view sizes, covering nearly the whole brain while keeping spatial resolution high. Thus, in order to gain new opportunities to apply deep neural network (DNN) based encoding models outside of vision research and outside of visual cortical areas, there is need for an implementation for ultra-high field MRI. Though there are studies applying encoding models to data from higher field strength of 4-7 Tesla (Naselaris et al., 2015; Nishimoto et al., 2011), these studies did not use neural network based encoding models and only covered parts or the occipital cortex.

We therefore aimed to investigate whether the fwrf model can be applied to ultra-high field fMRI data. Our second aim was to test whether - depending on the region of interest - different DNN layers contribute differently to the model's prediction accuracy.

Methods

First, a pre-trained deep neural network (Krizhevsky, Sutskever, & Hinton, 2012) for object recognition was used to process 1,400 colored natural images. Images represented seven distinct categories (animals, objects, city scenes, faces, people, nature landscapes, buildings) and consisted of 200 images per category. DNN activations for each network layer were extracted for all images. Next, two subjects underwent a natural image viewing task during functional MRI recordings (1,5x1,5x1,5 mm, TR=2 sec., whole brain) in a Siemens 7 Tesla Magnetom. Time courses for all voxels were extracted as input for the encoding model. In the last step, feature-weighted receptive field models, as implemented by St-Yves and Naselaris (2018), were trained based on extracted voxel time courses and DNN unit activations for each subject. Trained fwrf models were then used to predict the voxel time course of a new set of images from validation fMRI runs.

Results

Implementing the fwrf model for ultra-high field fMRI

Our first goal was to implement the fwrf model for our 7 Tesla fMRI dataset. Using the methods described above we were able to predict voxel activity for areas along the ventral visual stream (VVS) based on the trained encoding models (Figure 1).

Whereas voxels outside the VVS mostly showed very low correlations (Pearson), correlations within the VVS reached values of up to r = 0.68 (subject 2: r = 0.75), demonstrating the link between the visual pathway and DNN features. Overall, 37 of 80 images (subject 2 = 26/80) showed a maximum correlation coefficient on the diagonal (Figure 2), indicating that the predicted voxel activity correlates highest with the real activity for these images, i.e. these images were identified correctly. The last finding concerning the correlation between predicted and real activity patterns showed that there are higher correlations for images of the same category (Figure 2) causing some of the images to be misclassified as another similar image from the same category (e.g. cat and cheetah) than for images of different categories.

Shift in DNN layer contribution along the VVS

Our second analysis tested for a hypothesized shift in prediction accuracy depending on the DNN layer and the region of interest along the VVS. We investigated this by comparing the percentage of layer contribution to prediction accuracy for the new set of images (Figure 3). In both subjects we found that layers that process low visual features such as spatial frequencies indeed showed higher contributions to accuracy in early visual areas like V1 and V2. Reversely, DNN layers that process higher visual features (e.g. faces or object classes) showed higher contributions to prediction accuracy in higher visual processing areas, e.g. the lateral occipital cortex (LOC) and the fusiform gyrus.

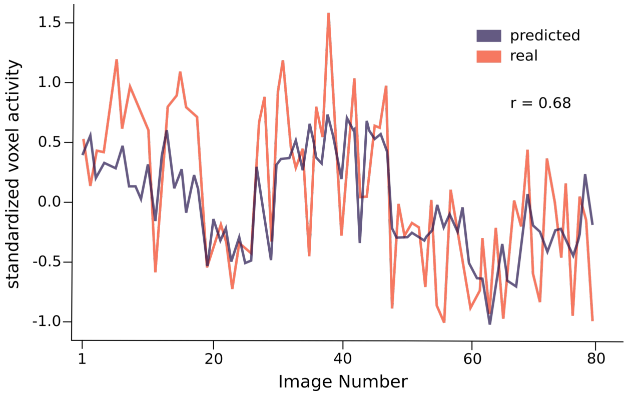

Figure 1: Predicted and real activity time course for the validation set from one voxel within the VVS (subject one). The validation set consisted of 80 images from all 7 image categories. Overall 5,744 voxels for subject one and 9,841 voxels for subject 2 showed Pearson correlation coefficients higher than r = 0.2.

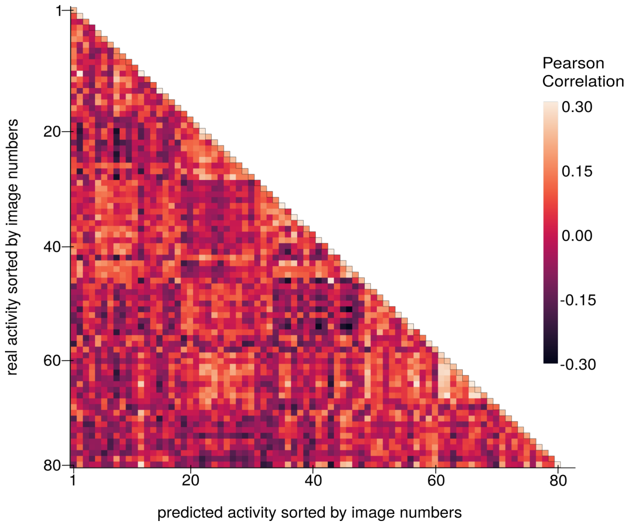

Figure 2: Pearson correlation coefficients of predicted and real voxel time course using the 5,744 voxels showing Pearson correlation coefficients higher than r = 0.2 (subject one). The figure shows (1) high correlation coefficients on the diagonal as well as (2), higher within-category than between-category correlations on the off-diagonal.

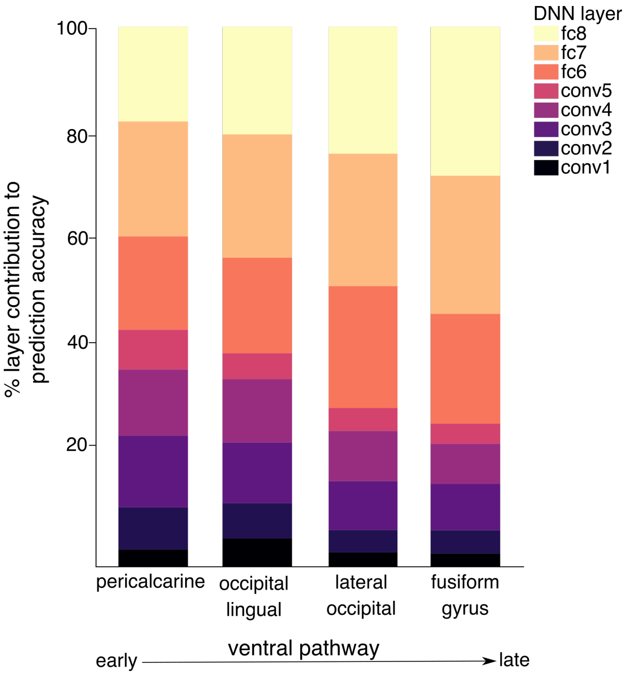

Figure 3: DNN layer contribution to prediction accuracy of the encoding model. While early DNN layer (conv1-5) showed higher contributions to prediction accuracy for early visual cortex, contribution of higher DNN layers (fc6-8) increased further along the visual pathway (results from subject one).

Conclusion

Our results indicate that the fwrf model by St-Yves and Naselaris (2018) can be applied to ultra-high field MRI data. Using this model, we were able to predict brain activity based on image features from a DNN. Higher similarities within as compared to between image categories across the VVS demonstrate the networks ability to grab visual categorical information, not only basic visual features.

We were also able to replicate the parallel processing hierarchy of stimulus features, starting with low DNN layer features in early visual areas ranging to categorical features for high DNN layers in later areas along the VVS, as proposed by Güçlü and van Gerven (2015) and Cichy et al. (2016). Using ultra-high field fMRI data, we could show a step-like shift of layer contribution along fine-graded regions of the VVS even into regions as late as the fusiform gyrus.

After predicting brain activity from stimulus features, the next step would be to reconstruct stimuli based on fMRI data. While reconstructing images from brain activity using neural networks can already be achieved (Seeliger et al., 2018; Shen, Horikawa, Majima, & Kamitani, 2019), the use of ultra-high field fMRI data might improve this reconstruction. Given the high spatial resolution, one might be able to reconstruct even complex, categorical-based image features. This could get us closer to the ultimate goal of reconstructing mental images.

Taken together our findings suggest that deep neural network based encoding models can be applied to ultra-high field high-resolution datasets, covering more than the occipital cortex. This in turn shows that these models could have a larger field of application, using the advantages of voxel wise models, neural network features and parallel visual processing in men and machine to investigate other questions such as how neural representations are changed during learning and consolidation processes.

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*, 27755.

Dumoulin, S. O., Fracasso, A., van der Zwaag, W., Siero, J. C., & Petridou, N. (2018). Ultra-high field mri: Advancing systems neuroscience towards mesoscopic human brain function. *Neuroimage*, *168*, 345–357.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage*, *105*, 215–228.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & Van Gerven, M. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS computational biology*, *15*(1), e1006633.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.