Modeling human visual responses with a U-shaped deep neural network for motion flow-field estimation

Atsushi Wada[1,2] (a-wada@nict.go.jp), Satoshi Nishida[1,2] (s-nishida@nict.go.jp), Hiroshi Ando[1,2] (h-ando@nict.go.jp), Shinji Nishimoto[1,2,3] (nishimoto@nict.go.jp)

[1]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Yamadaoka 1-4, Suita, Osaka 565-0871, Japan

[2]Graduate School of Frontier Biosciences, Osaka University, Yamadaoka 1-3, Suita 565-0871, Japan

[3]Graduate School of Medicine, Osaka University, Yamadaoka 2-2, Suita 565-0871, Japan

Abstract:

Deep Neural Networks (DNNs) for visual classification have recently been shown to exhibit representations homologous to those observed in the human visual system. For further exploring the relationship between artificial and natural neural networks, we focus on a U-shaped contraction-expansion architecture for DNNs, which recursively refine the network output for solving pixel-wise 2D-prediction tasks. By using FlowNet (Dosovitskiy et al., 2015), a U-shaped DNN for motion flow-field estimation, we show that the DNN-features extracted from FlowNet accurately predict human visual responses to natural movie stimuli. We further present that the features from the expansion compared to contraction layers yield higher encoding performance for the mid-level regions in the dorsal visual pathway. Our results may support the notion of the information integration between the early processing stages preserving fine spatial information and the downstream processing stages providing global contextual cues for motion estimation in the human visual system.

Keywords: optic-flow; natural stimuli; encoding; fMRI; CNN

Introduction

Recent advances in Deep Neural Networks (DNNs) (LeCun, Bengio, & Hinton, 2015) have demonstrated human-competitive performance in multiple domains. Despite the simplification of many biological details in DNNs, the features acquired by DNNs have shown to predict neural responses to natural visual stimuli with the state-of-the-art performance (Güçlü & van Gerven, 2015; Wen et al., 2017), which opens up a novel approach to quantitatively investigate the hierarchical processing in the visual cortex (Yamins & DiCarlo, 2016).

Here, we focus on FlowNet (Dosovitskiy et al., 2015), a DNN that predicts 2D-motion flow-field (i.e., optical-flow) given a pair of temporally adjacent video frames (Figure 1A) instead of a standard DNN for image classification, which simply predict a category label. FlowNet utilizes a distinct U-shaped contraction-expansion architecture that implements recursive refinement steps though the expansion network, where each refinement step integrates global contextual cues with local spatial features (Figure 1B). By using FlowNet as an encoding model, we show that the DNN features emerged from such integration of information across distant processing stages well explain the fMRI responses in the mid-level regions in the human dorsal visual pathway.

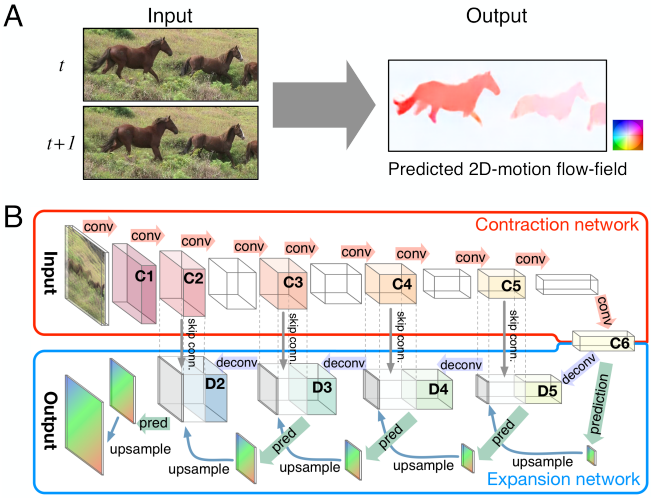

Figure 1: DNN for motion flow-field estimation. (A) Problem illustration. (B) FlowNet architecture.

Results

We collected human visual responses from three subjects using fMRI while they viewed a set of color natural movies presented in wide-view (82°×52°) with their eyes fixated on a stationary central cross. The same movie stimuli were fed into a pre-trained FlowNet to compute layer-wise DNN activation patterns. The activation pattern for each of the ten representative layers in FlowNet (labeled C1 to C6 and D2 to D5 in

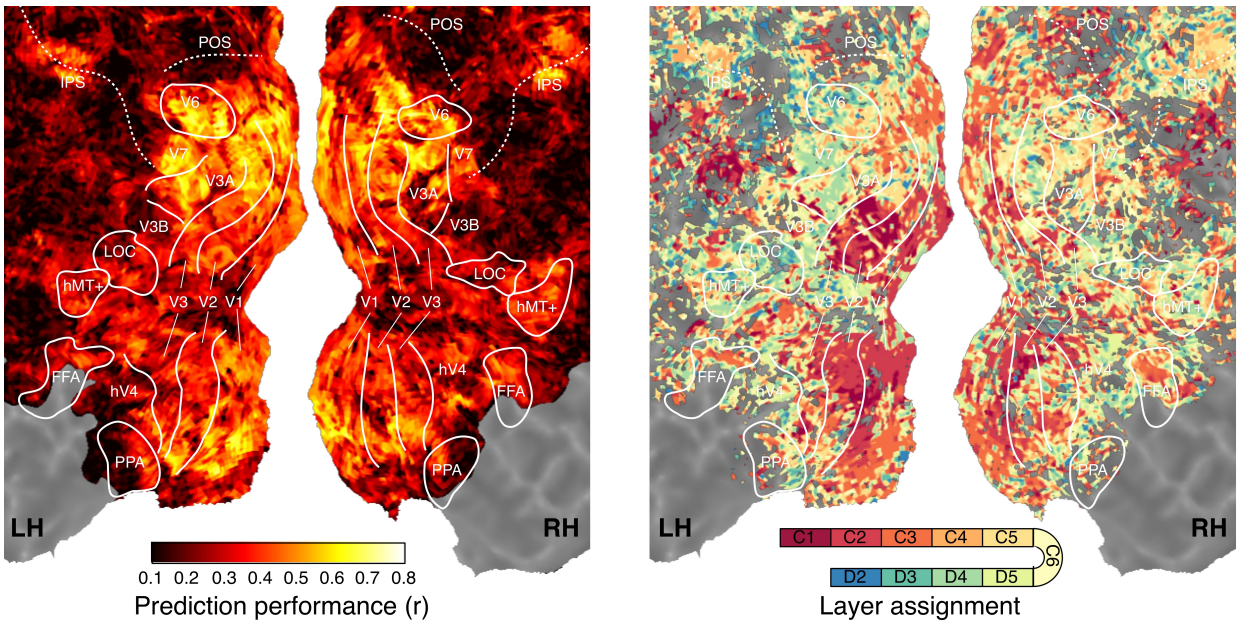

Figure 2: Voxel-wise modeling results with layer-wise FlowNet DNN features. (left) Prediction accuracies max-pooled across layers (right) the optimal layer assignments mapped onto the cortex.

Figure 1B) was reduced its dimension to 3,600 by PCA, and then temporally down-sampled to the fMRI sampling rate (0.5 Hz). Using the processed DNN features for each layer, we modeled voxel-wise brain responses (Naselaris et al., 2011) by performing regularized linear regression (Nishimoto et al. 2011).

Our results showed that the model acquired using DNN features can accurately predict the held-out brain responses from the held-out natural movies widely over the visual cortex (Figure 2, left). By further assessing which DNN layer exhibited the best prediction accuracy for each voxel (Figure 2, right), we found that the layers in the extraction network (labeled D2 to D5) showed better prediction performance compared to the layers in the contraction network (labeled C1 to C6) in the mid-level regions along the dorsal visual pathway, including areas V6 and V7. As the expansion layers are trained to integrate the global contextual cues from the preceding adjacent layer and the fine spatial information branched from the early processing stages, our result may implicate a similar information integration in the human visual system.

We believe that our study is the first of its kind to apply a U-shaped, contraction-expansion DNN to model and predict visual responses to natural movies. Further examinations may provide insights into how and where global contextual cues might be integrated with information in distant processing stages to achieve spatially fine-grain contextual representations.

Acknowledgments

This research was supported by JSPS KAKENHI Grant number JP15H05311, JP16K16081 and JP18K18141.

References

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., et al. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. *IEEE Intl Conf Comp Vis (ICCV),* 2758–2766.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J Neurosci*, *35*(27), 10005–10014.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

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 Biol*, *21*(19), 1641–1646.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2017). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cereb Cortex*, *99*, 1–25.

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