Attention biases neural representations of hierarchical visual features

Tomoyasu Horikawa (horikawa-t@atr.jp)
ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai
Seika, Soraku, Kyoto 619-0288, Japan

Yukiyasu Kamitani (kamitani@i.kyoto-u.ac.jp)
ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai
Seika, Soraku, Kyoto 619-0288, Japan
Graduate School of Informatics, Kyoto University, Yoshida-honmachi
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

Humans can voluntarily regulate how we perceive the external world by top-down mental processes. Visual attention highlights specific features of visual targets and enhances brain activity associated with the attended features. Previous studies demonstrated that the attentional modulation of brain activity allows decoding of attended features from the brain. However, it remains unclear whether and how brain activity associated with multiple levels of visual features can be affected by selective attention. Here, we quantified how hierarchical neural representations are modulated by object-based selective attention using fMRI and the brain decoding technique assisted by deep neural networks (DNNs). Using statistical models that decode fMRI activity into hierarchical DNN features, we decoded fMRI activity measured while subjects attended to one image of a superposition of two images. The decoded features were found to be biased to attended images over unattended images with greater effects for lower-/higher-level DNN features in lower-/higher-level visual areas. Furthermore, image reconstructions generated from the decoded features resembled attended images, demonstrating faithful reconstructions of mental images. Our analyses showed fine-grained attentional modulation for hierarchical visual features. The results also indicate that the attention-guided mental image reconstruction may provide a substrate for developing systems of neurofeedback and brain-machine interfaces.

Keywords: attention; decoding; deep neural network; fMRI; image reconstruction

Introduction

Visual selective attention is an internal mental process that facilitates behavioral performances (e.g., response speed and accuracy) and enhances brain responses in areas that encode information about attended targets (Serences et al., 2004). Previous studies that analyzed neural activity patterns during selective attention have shown that information about attended features can be decoded from the brain (Kamitani & Tong, 2005, 2006; Cerf et al., 2010). Using functional magnetic resonance imaging (fMRI) and statistical decoders trained to predict seen image features (edge orientations and motion directions), Kamitani and Tong (2005, 2006) have succeeded in reliably predicting attended features from visual cortical activity patterns. Cerf and his colleagues (2010) have also demonstrated that subjects can voluntarily control their attention to visualize attended images via a device for brain-machine interface that decode neural activity patterns in the medial temporal lobe.

As shown in previous studies, attention to specific visual features enhances neural activity associated with the attended feature representations, which enables decoding of attended information from brain activity patterns. However, little is known about whether or how brain activity associated with multiple levels of hierarchical visual features can be modulated by voluntary attention. Although until recently it had been difficult to thoroughly analyze neural representations of hierarchical visual features, recent developments of deep neural networks (DNNs) have enabled detailed analyses of hierarchical feature representations across visual cortical areas (Yamins & DiCarlo, 2016).

Here, we quantitatively evaluate how object-based selective attention modulates neural representations of hierarchical visual features in the human visual cortex. We adopted the DNN feature decoding methodology used in previous studies (Horikawa & Kamitani, 2017; Shen et al., 2019) to quantify how neural representations for multiple visual features in individual visual areas are modulated by selective attention. Using decoders trained to decode (translate) visual cortical activities into DNN features, we decode individual visual cortical activities, which were measured while subjects attended to one image of a superposition of two images, into DNN features, and compared the similarity between features decoded from brain activity patterns and those calculated from presented images (Figure 1). We show that the decoded features are biased to attended images in the analyses with most combinations of DNN layers and visual areas. We further perform the DNN-based visual image reconstruction analysis (Shen et al., 2019) to evaluate the effect of the attentional modulation, and test whether mental images of the subjects can be externalized via decoded DNN features.

Results

We first trained decoders that predict DNN features of viewed natural images from brain activity patterns in visual cortex (VC

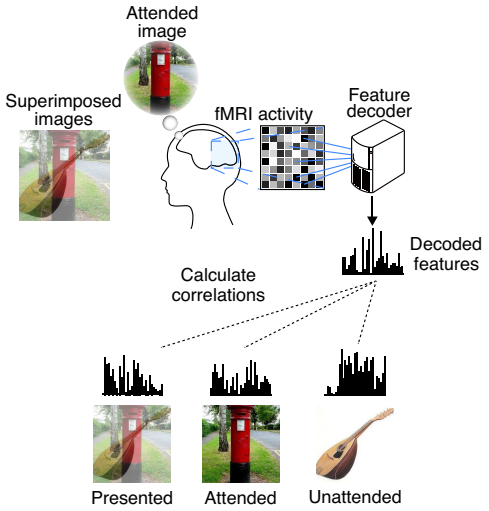

Figure 1: Overview. DNN features were decoded from brain activity patterns measured during object-based selective attention. Similarities between a decoded DNN feature pattern and feature patterns calculated from a presented image or each of its component images (attended and unattended images) were evaluated.

that consisted of V1-V4 and higher visual cortex [HVC]) following the procedures of Horikawa and Kamitani (2017) and Shen et al. (2019). The trained decoders were then used to predict DNN features from brain activities measured in an independent attention experiment. In the attention experiment, subjects were presented with image sequences, each of which consisted of two spatially superimposed images, and were asked to attend to one image of a superposition of two images while ignoring the other. For evaluation, first, DNN feature patterns for each stimulus were separately calculated from a presented image (spatially superimposed two images) and each of its component images (attended and unattended images). Then, Pearson correlation coefficients were calculated between a feature pattern decoded from brain activity induced by a stimulus and feature patterns calculated from presented, attended, and unattended images. These procedures were repeated for all combinations of DNN layers (DNN1–8, AlexNet) and visual areas used for decoder training.

The correlation coefficients for all combinations of DNN layers and visual areas are shown in Figure 2. First, the decoded features were positively correlated with the features calculated from presented images for all combinations of DNN layers and visual areas, suggesting that the trained decoders were able to accurately translate visual cortical activities into hierarchical DNN features. Meanwhile, comparisons between the correlations for attended and unattended images showed higher correlations for attended images than unattended images with most combinations of layers and areas (a total of 213 out of 240 pairs, 88.75% for combinations of eight DNN layers and six visual areas from five subjects), demonstrating that selective attention reliably induced attentional modulation and biased hierarchical visual feature representations in individual visual areas to attended images. Intriguingly, although overall correlations for attended images were slightly lower than those for presented images, those correlations for attended images tended to be almost equivalent to the correlations for presented images specifically with the decoded features predicted from the HVC activity. This may indicate the susceptibility of the higher visual areas by the top-down attentional modulation.

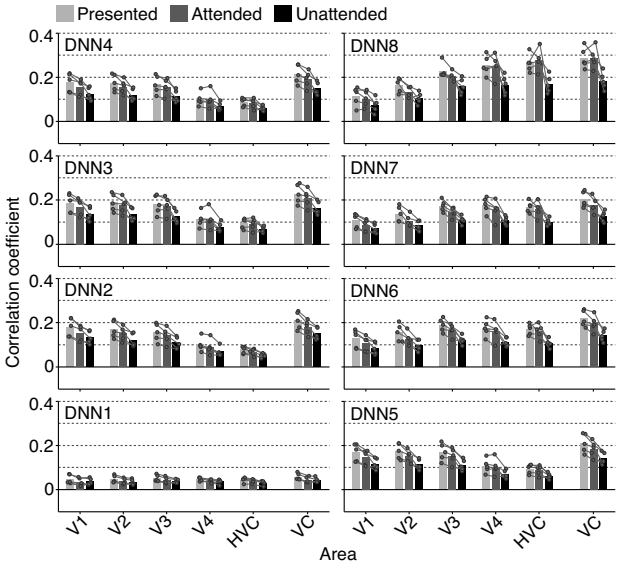

Figure 2: Correlation coefficient between DNN features decoded and calculated from presented, attended, and unattended images. Bars indicate mean correlations across five subjects. Dots indicate results of individual subjects.

To investigate the feature specificity of the attentional modulation in individual visual areas, we calculated the differences of correlation coefficients for attended and unattended images for each visual area, and checked trend differences across DNN layers (Figure 3). The peaks of DNN layers with highest correlation differences tended to shift gradually across visual areas (e.g., peaked at DNN3 for V2 to DNN8 for HVC), showing larger differences for lower- and higher-level DNN features in lower- and higher-level visual areas (two-way ANOVA, interaction between layers and areas, $p = 0.018$). The results revealed the feature specificity of attentional modulation tied to the hierarchical correspondence between the brain and the DNN.

Finally, we have qualitatively examined the effect of the attentional modulations on decoded DNN features using the DNN-based visual image reconstruction method (Shen et al., 2019). The method optimizes pixel values of an input image so that the DNN features calculated from the input image become closer to target DNN features, in this case, decoded

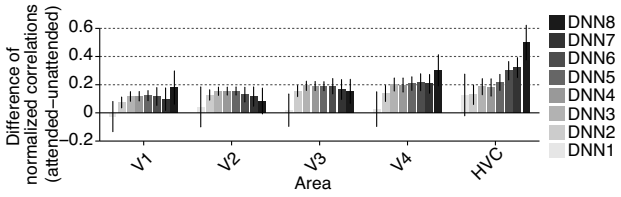

Figure 3: Differences of normalized correlations between attended and unattended images. Correlations for unattended images were subtracted from those for attended images for each combination of layers and areas. To eliminate baseline differences across layers and areas, correlation coefficients for attended and unattended images were normalized before the subtraction using DNN feature decoding performance separately evaluated with another fMRI dataset. Error bars indicate 95% confidence intervals across test samples (five subjects pooled).

features predicted from brain activity during attention. The reconstructed images are shown in Figure 4. Although the reconstruction quality remains room for improvement, the generated images appeared to resemble attended images rather than unattended images. Notably, even when the same superimposed images were presented, the appearances of the reconstructed images were biased according to the images to be attended. These results demonstrated the feasibility of a method of mental image reconstructions guided by visual attention.

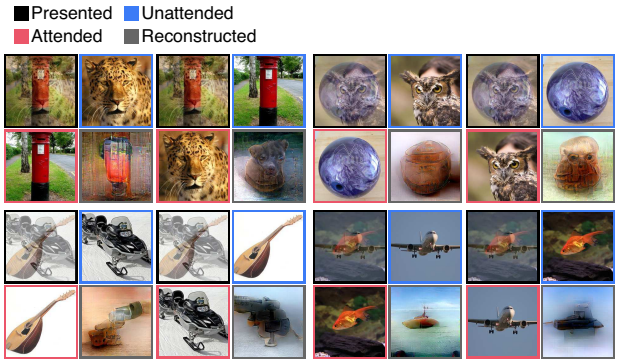

Figure 4: Reconstruction results. Images reconstructed from decoded DNN features are shown (VC, DNN1–8, VGG19). Images with black, red, blue, and gray frames indicate presented, attended, unattended, and reconstructed images, respectively.

Discussion

We have quantified the effect of attentional modulation for multiple levels of visual features in individual visual areas using the fMRI decoding of DNN feature representations. We showed that the object-based selective attention reliably biased the neural representations of hierarchical visual features to attended images with the feature specificity that reflected the hierarchical correspondence between the visual areas and DNN layers. We have also demonstrated that images attended by the subjects can be reconstructed based on the biased neural representations. These results highlight the potency of the internal mental processes to regulate bottom-up sensory inputs by top-down voluntary attention. Additionally, we have provided a proof-of-concept demonstration for the attention-guided mental image visualization, which could be a powerful tool for neurofeedback and brain-machine interface technologies.

Acknowledgments

This research was supported by grants from the New Energy and Industrial Technology Development Organization (NEDO), JSPS KAKENHI Grant number JP15H05710, JP15H05920, and JP17K12771, JST CREST Grant Number JPMJCR18A5, and JST PRESTO Grant Number JPMJPR185B Japan.

References

Cerf, M. et al. (2010). On-line, voluntary control of human temporal lobe neurons. *Nature*, *467*, 1104–1108.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.*, *8*. doi: 10.1038/ncomms15037.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective content of the human brain. *Nat. Neurosci.*, *8*, 679–685.

Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion direction from activity in the human visual cortex. *Curr. Biol.*, *16*, 1096–1102.

Serences, J. T. et al. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, *14*, 1346–1357.

Shen, G. et al. (2019). Deep image reconstruction from human brain activity. *PLOS Comput. Biol.*, *15*, e1006633.

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