Performance Optimization is Insufficient for Building Accurate Models for Neural Representation

Jonathan Yu (jy11@cs.princeton.edu)
Department of Computer Science, Princeton, NJ 08544

Qihong Lu (qlu@princeton.edu)
Princeton Neuroscience Institute and Department of Psychology, Princeton, NJ 08544

Uri Hasson (hasson@princeton.edu)
Princeton Neuroscience Institute and Department of Psychology, Princeton, NJ 08544

Kenneth A. Norman (knorman@princeton.edu)
Princeton Neuroscience Institute and Department of Psychology, Princeton, NJ 08544

Jonathan W. Pillow (pillow@princeton.edu)
Princeton Neuroscience Institute and Department of Psychology, Princeton, NJ 08544

Abstract

Convolutional neural networks, optimized for image classification, are state-of-the-art computational models for visual neural representation. Moreover, performance optimization often leads to better models of neural representation (Yamins et al., 2014). In this study, we investigate whether performance optimization always increases the similarity between the neural network and the brain, in terms of their representations. We compared AlexNet and a residual network on a recent human image-viewing fMRI dataset (Horikawa & Kamitani, 2017). The original study found a remarkable similarity between AlexNet and the brain (Horikawa & Kamitani, 2017). Although residual networks achieved better image classification performance, we found that the hidden representation of the residual network is much less similar to human brain representation, compared to AlexNet. This result suggests that performance optimization can eventually lead to systematic deviation from human brain representation. We conclude that additional neuroscience-inspired design is critical for building a better representation model of the brain.

Keywords: vision; neural representation; ConvNet; fMRI

Introduction

Convolutional neural networks one of the state-of-the-art "representational models" (Kriegeskorte & Kievit, 2013; Diedrichsen & Kriegeskorte, 2017) for vision. Many studies have shown that the hidden states between convolutional neural networks, typically trained on image classification, and the neural representation in the ventral visual pathway are highly similar (Cadena et al., 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Horikawa & Kamitani, 2017; Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann, McClure, & Kriegeskorte, 2017; Kriegeskorte, 2015; Seeliger et al., 2017; Shen, Horikawa, Majima, & Kamitani, 2017; Wen, Shi, Chen, & Liu, 2017; Yamins et al., 2014; Yamins & DiCarlo, 2016). Moreover, the learned feature detectors in some convolutional neural networks are qualitatively similar to findings from neurophysiology (Güçlü & van Gerven, 2014; Kriegeskorte, 2015).

Interestingly, it has been observed that performance optimization (e.g., image classification) can often lead to better models of neural representation (Seibert et al., 2016; Yamins et al., 2014; Yamins & DiCarlo, 2016). We believe that reason is the following: for any domain, there are relatively few ways to be optimal, whereas there are infinitely many ways to be suboptimal. Therefore, for a given domain, for which the brain is quite optimized, performance optimization should bring the computational model "closer" to the brain. If this explanation is correct, then performance optimization can eventually make the model better than the brain, which should make the model systematically different from how the brain works.

In the present study, we test whether performance optimization always leads to a greater similarity between neural networks and the brain, in terms of their hidden representations. In a recent functional magnetic resonance imaging (fMRI) study (Horikawa & Kamitani, 2017), researchers used the evoked fMRI responses to linearly predict the hidden state of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), a standard eight-layer convolutional neural network trained on ImageNet (Deng et al., 2009; Russakovsky et al., 2014). The result indicates that the hidden representation of AlexNet is highly similar to human brain representation. We repeated the analysis by Horikawa and Kamitani (2017) with a residual network with 50 layers (ResNet-50) (He, Zhang, Ren, & Sun, 2015). Although ResNet-50 is much better than AlexNet on the ImageNet classification task (Canziani, Paszke, & Culurciello, 2016), we found that the evoked fMRI responses are much less predictive for the hidden states of ResNet-50 (Figure 1), compared than AlexNet. This result suggests that performance optimization can eventually lead to systematic deviation from brain representations. We conclude that perfor-

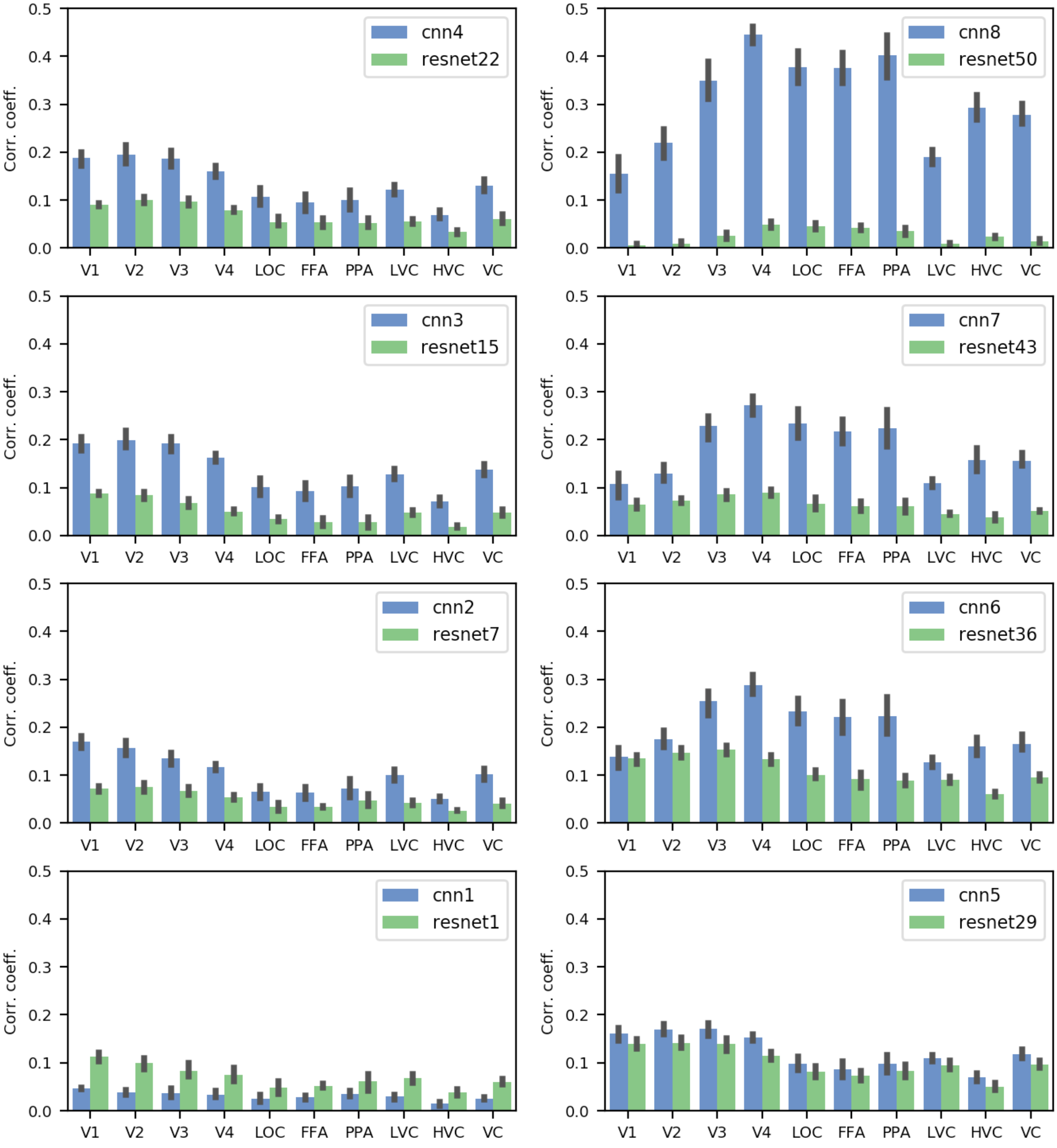

Figure 1: The correlation between human fMRI data and neural network activity patterns for AlexNet (cnn) and a residual network with 50 layers (resnet). All error bars indicate 95% bootstrapped confidence intervals. The indices represent the layer of the respective neural network. AlexNet has eight layers: five convolutional layers, two densely connected layer and an output layer, denoted by cnn1 to cnn8. For the residual network, we chose 8 layers from ResNet-50, roughly evenly spaced across the entire architecture. The last layer (marked as “resnet50”) is the output layer of the residual network, chosen to match the output layer of AlexNet (cnn8). The rest of the residual network layers are convolutional. The Region of interests (ROIs) include V1, V2, V3, V4, lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA). The low-level visual cortex (LVC) includes V1, V2, V3; The high-level visual cortex (HVC) includes V4, LOC, PPA, FFA; Visual cortex (VC) includes all ROIs.

mance optimization only is insufficient for building accurate computational models for the brain.

Methods and Results

fMRI experiment (Horikawa & Kamitani, 2017)

In the fMRI experiment, participants performed a one-back task for static images chosen from ImageNet. There were two image sets. The first image set contains 1200 images, chosen from 150 categories from ImageNet, 8 images per category. We used this set as the train set for the later linear model. The second image set contains 50 images, and each image was presented 50 times to the human participant. The average evoked fMRI response across 50 repetitions provides a more stable estimation of the fMRI response.

Deep residual network (He et al., 2015)

In this study, we used a residual network with 50 layers (ResNet-50) (He et al., 2015). Different from a standard convolutional neural network, residual networks have residual blocks. Each residual block consists of three layers with a "skip connection" that directly sends the raw activity vector from the first layer to the third layer. These skip connections allow gradients to propagate more easily, which made training very deep networks possible.

ResNet-50 is much more optimized for image classification than the AlexNet. On the ImageNet classification task, the top-one classification accuracy for ResNet-50 is above 75%, and AlexNet's accuracy is slightly less than 55% (Canziani et al., 2016). The chance level of this classification task is 0.1%. In the present study, We used the ResNet-50 implemented in Keras (Chollet, n.d.; Chollet & Others, 2015) and pre-trained on ImageNet.

Predicting neural network activity patterns with fMRI responses

In the original paper (Horikawa & Kamitani, 2017), for a given image, researchers used the evoked fMRI response to predict the activity pattern of AlexNet. Specifically, they randomly chose 1000 units from each layer of AlexNet as the input. For each unit, they fitted a linear model that maps the evoked fMRI response to the hidden state of that unit. These linear models were trained on 1200 images from 150 categories and subsequently evaluated on the 50 test images. The average correlations across five subjects are shown in Figure 1. The result indicates that the hidden representation of AlexNet is sufficiently similar to the human brain, in the sense that there exists a linear mapping that "connect" them reasonably well. For a more detailed description of the data, please refer to the original paper by Horikawa and Kamitani (2017).

We repeated this analysis with a residual network (He et al., 2015). To compare with AlexNet, we chose eight layers from ResNet-50, roughly evenly spaced across the entire network. The last chosen layer is also the output layer of ResNet-50. To closely match the original AlexNet experiment (Horikawa & Kamitani, 2017), we also randomly selected 1000 units from each chosen layer as the input to the linear models.

Figure 1 shows that the linear prediction performance with ResNet-50 is significantly worse than AlexNet (except for the comparison of their first layers, resnet1 and cnn1, respectively). For example, in the AlexNet experiment, the output layer (cnn8), which should strongly encode category distinction, achieved the best performance. In comparison, the linear model performance on the output layer of the ResNet-50 is very low. A similar pattern holds for most other comparisons. This result suggests that ResNet-50 is an inferior model for visual neural representation.

Conclusion

It has been observed that performance optimization can lead to better representation models for vision (Yamins & DiCarlo, 2016; Yamins et al., 2014; Seibert et al., 2016). In the present study, we found that the hidden states of a residual network are much less similar to human fMRI responses, compared to AlexNet, even though the residual network is much more optimized (Canziani et al., 2016). The result suggests that neural networks can "go too far" with performance optimization.

It is reasonable to expect some performance optimization can bring a computational model "closer" to the brain because there are very few ways to be optimal, whereas there are potentially infinitely many ways to be suboptimal. However, we believe that performance optimization can eventually lead to systematic deviation from brain representations, especially when the model exceeds human-level performance. Indeed, a recent study showed that many highly optimized neural networks are not predictive of the behavioral performance of primates on some vision tasks (Rajalingham et al., 2018). Our result also confirms that performance optimization by itself is insufficient for building accurate computational models of the brain. We believe that additional constraints informed by visual neuroscience are critical for building better computational models. For example, a recent study shows that networks with recurrent connections are better at predicting human fMRI responses evoked by dynamic natural stimuli (Shi, Wen, Zhang, Han, & Liu, 2017).

Future directions

This is an ongoing project. Notably, a previous study found that ResNet-50 outperformed AlexNet on voxel-wise modeling (Wen et al., 2017). Our result is inconsistent with this finding, and we are still actively investigating this inconsistency.

However, in theory, we think performance optimization should eventually lead to systematic deviation from how the brain works. For example, if we measure the similarity between the brain and neural networks using representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008), while varying the performance of the network, we expect to see an inverted-U shaped relation. Initially, brain-network similarity should be positively correlated with performance (of the network), but this correlation should eventually become negative. To test this hypothesis, we plan to investigate other convolutional neural networks with varying image classification performance.

Supplement

The fMRI data is obtained from OpenNeuro:
`https://openneuro.org/datasets/ds001246/versions/00002`
We used the code from Horikawa and Kamitani (2017):
`https://github.com/KamitaniLab/GenericObjectDecoding`
Our code is available at:
`https://github.com/aerrowfinn72/ResNet-Image-Decoding`
`https://pypi.org/project/qmvpa/`

Acknowledgments

We thank Horikawa and Kamitani (2017) for sharing their data and code.

References

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2017, October). *Deep convolutional models improve predictions of macaque V1 responses to natural images.*

Canziani, A., Paszke, A., & Culurciello, E. (2016, May). An analysis of deep neural network models for practical applications.

Chollet, F. (n.d.). *deep-learning-models.*

Chollet, F., & Others. (2015). *Keras.*

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016, June). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.*, *6*, 27755.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009, June). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Diedrichsen, J., & Kriegeskorte, N. (2017, April). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.*, *13*(4), e1005508.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017, May). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, *152*, 184–194.

Güçlü, U., & van Gerven, M. A. J. (2014, November). Deep neural networks reveal a gradient in the complexity of neural representations across the brain's ventral visual pathway.

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December). Deep residual learning for image recognition.

Horikawa, T., & Kamitani, Y. (2017, May). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.*, *8*, 15037.

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

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017, May). *Deep neural networks in computational neuroscience.*

Kriegeskorte, N. (2015, November). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu Rev Vis Sci*, *1*(1), 417–446.

Kriegeskorte, N., & Kievit, R. A. (2013, August). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.*, *17*(8), 401–412.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008, November). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, *2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018, January). *Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2014, September). ImageNet large scale visual recognition challenge.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & van Gerven, M. A. J. (2017, July). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *Neuroimage*.

Seibert, D., Yamins, D. L., Ardila, D., Hong, H., DiCarlo, J. J., & Gardner, J. L. (2016, January). *A performance-optimized model of neural responses across the ventral visual stream.*

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2017, December). *Deep image reconstruction from human brain activity.*

Shi, J., Wen, H., Zhang, Y., Han, K., & Liu, Z. (2017, August). *Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision.*

Wen, H., Shi, J., Chen, W., & Liu, Z. (2017, June). *Deep residual network reveals a nested hierarchy of distributed cortical representation for visual categorization.*

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

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, June). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, *111*(23), 8619–8624.