CAUSAL IMPORTANCE OF ORIENTATION SELECTIVITY FOR GENERALIZATION IN IMAGE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Although both our brain and deep neural networks (DNNs) can perform high-level sensory-perception tasks such as image or speech recognition, the inner mechanism of these hierarchical information-processing systems is poorly understood in both neuroscience and machine learning. Recently, Morcos et al. (2018) examined the effect of class-selective units in DNNs, i.e., units with high-level selectivity, on network generalization, concluding that hidden units that are selectively activated by specific input patterns may harm the network's performance. In this study, we revisit their hypothesis, considering units with selectivity for lower-level features, and argue that *selective units are not always harmful to the network performance*. Specifically, by using DNNs trained for image classification (7-layer CNNs and VGG16 trained on CIFAR-10 and ImageNet, respectively), we analyzed the *orientation selectivity* of individual units. Orientation selectivity is a low-level selectivity widely studied in visual neuroscience, in which, when images of bars with several orientations are presented to the eye, many neurons in the visual cortex respond selectively to a specific orientation. We found that orientation-selective units exist in both lower and higher layers of these DNNs, as in our brain. In particular, units in the lower layers become more orientation-selective as the generalization performance improves during the course of training of the DNNs. Consistently, networks that generalize better are more orientation-selective in the lower layers. We finally reveal that ablating these selective units in the lower layers substantially degrades the generalization performance, at least by disrupting the shift-invariance of the higher layers. These results suggest to the machine-learning community that, contrary to the triviality of units with high-level selectivity, lower-layer units with selectivity for low-level features can be indispensable for generalization, and for neuroscientists, orientation selectivity can play a *causally* important role in object recognition.

1 INTRODUCTION

Recognizing the natural world is a fundamental competency for animals and artificial intelligence. Although our cerebral cortex and recent deep neural networks (DNNs) both achieve high accuracy in sensory perception (Krizhevsky et al., 2012; Hinton et al., 2012), especially object recognition through natural vision, the rationale for this performance is not well understood, either in visual neuroscience or machine learning.

In machine learning, numerous studies have been conducted on why DNNs have good generalization ability (summarized in section 4.2). One recent interesting hypothesis proposed in Morcos et al. (2018) is that networks that rely on *single directions*[1] may generalize poorly. The authors performed ablation experiments to argue that networks that generalize poorly are sensitive to unit ablations and that selectively activated units are not important for generalization. However, the authors examined only *high-level* single directions; they performed ablation experiments on higher layers and analyzed the class selectivity of individual units. Their hypothesis must be evaluated for other directions, especially *low-level* directions (or "features"), in order to conclude that it is correct.

[1] "The activation of a single unit or feature map or some linear combination of units in response to some input" (Morcos et al., 2018).

In visual neuroscience, the most popular experimental setting is to present bar or grating images (Fig. A1) at several orientations to animals (e.g., cats, mice, or humans) while simultaneously recording the visually evoked neuronal responses. This paradigm is used to analyze the *orientation-selective* feature of many neurons in the visual cortex. This feature causes the responses evoked by a bar or grating image presented to an animal to be tuned to the orientation of the image (Hubel & Wiesel, 1959) (examples of this "orientation tuning curve" are shown in Fig. 1). Numerous neuroscience researchers have thus far analyzed orientation selectivity to investigate the functions of the visual cortex, especially primary visual cortex (V1), such as how neurons with similar properties are interconnected (Ko et al., 2011; Wertz et al., 2015), how visual discrimination tasks modulate neural coding (Schoups et al., 2001; Dragoi et al., 2002; Jehee et al., 2012), and how neural selectivity emerges and matures during development (Chapman & Stryker, 1993; White et al., 2001). However, there is a large gap between object recognition in a natural environment and orientation selectivity. Specifically, no researchers have thus far investigated whether orientation selectivity does have an important role in object recognition, or is merely a superficial byproduct of object recognition.

In this study, we address these two issues simultaneously by performing neuroscience-inspired experiments on DNNs trained for image classification. The use of DNNs to model the visual cortex is supported by several studies that suggested the hierarchical similarity of feature representations between DNN and the visual cortex (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cadieu et al., 2014; Güçlü & van Gerven, 2015; Horikawa & Kamitani, 2017). Just as in neuroscience experiments, we inputted grating images to the trained networks and analyzed the orientation selectivity of the units in the hidden layers to explore whether orientation selectivity contributes to the generalization performance in object recognition. The main findings in this paper can be summarized into four aspects.

- Orientation-selective units exist in all the layers of a DNN trained for image classification.
- During training, units in the lower layers become more orientation-selective in sync with the generalization performance.
- Networks that generalize better are more orientation-selective in the lower layers.
- Ablation experiments revealed that orientation-selective units in the lower layers play a *causal* role[2] in generalization, at least by introducing shift-invariance of the higher layers.

The major implications derived from these empirical analyses are:

- From the neuroscience perspective, orientation selectivity in the lower layers of hierarchical visual systems may be a causally crucial component and not be a trivial byproduct.
- From the machine-learning perspective, our results, in conjunction with very recent studies that argued the importance of class selectivity from different viewpoints (Liu et al., 2018; Zhou et al., 2018), suggest that the uselessness of single directions proposed in Morcos et al. (2018) is overstated. In addition, the significance of the orientation-selective units in the lower layers directly supports the assertion that Gabor feature representations in the lower layers are indispensable for the generalization of the DNNs.

2 METHODS

All analyses were performed on Keras (2.0.8) and Tensorflow (1.3.0). We primarily analyzed a 7-layer CNN (six convolutional (Conv) layers plus one fully connected (FC) layer) trained on the CIFAR-10 dataset (Krizhevsky & Hinton, 2009) (see section A.1 for hyperparameters and training details). To check the reproducibility of our results, we also analyzed i) another 7-layer CNN trained on CIFAR-10 with different initializations and ii) a VGG16 network (Simonyan & Zisserman, 2015), which is a 20-layer CNN trained on the ImageNet dataset (Russakovsky et al., 2015). We fed grating

[2] Determining causality between phenomena has been a central goal in natural science, including neuroscience: "Scientific work will always be the search for causal interdependence of phenomena" (Born, 1949), "Cracking the neural code–in other words, determining which spatiotemporal patterns of activity [...] causally drive behavior"(Packer et al., 2013).

images to these trained networks and analyzed the orientation selectivity of the hidden units[3] in exactly the same way as in visual neuroscience studies.

2.1 Grating images

A *grating image* is a two-dimensional gray-scaled sinusoidal wave $G(x, y)$, formulated as follows.

$$G(x, y) = \cos(k_0 y' + \tau) \quad y' = -(x - x_0)\sin\theta + (y - y_0)\cos\theta, \tag{1}$$

where k_0 is the spatial frequency (SPF), τ is the phase, θ is the orientation, and (x_0, y_0) is the center coordinate of the image. For the 7-layer CNNs, we prepared a total of 432 grating images using nine SPFs, 12 orientations (15° apart), and four phases (90° apart). For the 20-layer CNN, we prepared 2,736 grating images with 57 SPFs using the same 12 orientations and four phases.

2.2 Orientation selectivity

Similar to the neuroscience studies that present grating images to animals while recording their neuronal responses, the unit activations with respect to each grating image were collected. The activation matrix of each unit has a shape of $N_{ori} \times N_{SPF} \times N_{phase}$, where its $(k, l, m)^{th}$ element represents the activation with respect to the grating of the k^{th} orientation, l^{th} SPF, and m^{th} phase. After taking the maximum along the phase dimension, we extracted the activations on the SPF that yielded the highest activation, generating a vector of length N_{ori}. Using this vector, orientation selectivity was quantified by using the *global orientation selectivity index (gOSI)*, which is a popular metric in visual neuroscience (Wörgötter & Eysel, 1987). The gOSI of the i^{th} unit is formulated as

$$gOSI^i = \sqrt{\left(\sum_{k=1}^{N_{ori}} R_k^i \sin 2\theta_k\right)^2 + \left(\sum_{k=1}^{N_{ori}} R_k^i \cos 2\theta_k\right)^2} \bigg/ \sum_{k=1}^{N_{ori}} R_k^i, \tag{2}$$

where R_k^i is the activation of the i^{th} unit with respect to the grating image of the k^{th} orientation and θ_k is the degree of orientation (0–180°). This metric is between 0 and 1, and a higher value indicates higher orientation selectivity. Note that this metric is equivalent to 1 $-$ “circular variance.”

3 Results

3.1 Orientation selectivity in DNNs

To investigated whether the units in the DNNs are tuned to the orientations, as is the case in the visual cortex, we fed grating images to the trained networks and analyzed the evoked activations. Activations of 10 example units of a 7-layer CNN along each orientation are shown in Fig. 1. Most units, both in the lower layer and the higher layer, were selectively activated by specific orientations. We then quantified the degree of orientation selectivity by using the gOSI metric for all units in all layers. Interestingly, units with high gOSI (orientation-selective units[4]) existed in all layers (Fig. 2). Furthermore, this orientation selectivity was almost independent of the SPFs of the grating images (section A.2 and Fig. A2). Similar results were obtained with another 7-layer CNN with different initializations and a 20-layer CNN trained on ImageNet (Figs. A4 and A3), indicating the robust existence of orientation selectivity in all layers of the networks.

One can consider that the high orientation selectivity results from the sparsity of the hidden units. For example, a sparsely active unit, with its activation to the 0° grating being 1 and otherwise 0, has gOSI = 1.0. To investigate this possibility, we also computed gOSI after the activation matrix of each unit was randomly shuffled. The gOSI calculated in this way was much smaller than the

[3] The feature maps of convolutional layers and units of fully connected layers are both referred to as “unit” in this paper.

[4] In visual neuroscience, neurons with gOSI $>$ 0.33 are often considered to be orientation-selective (Piscopo et al., 2013; Kondo & Ohki, 2015).

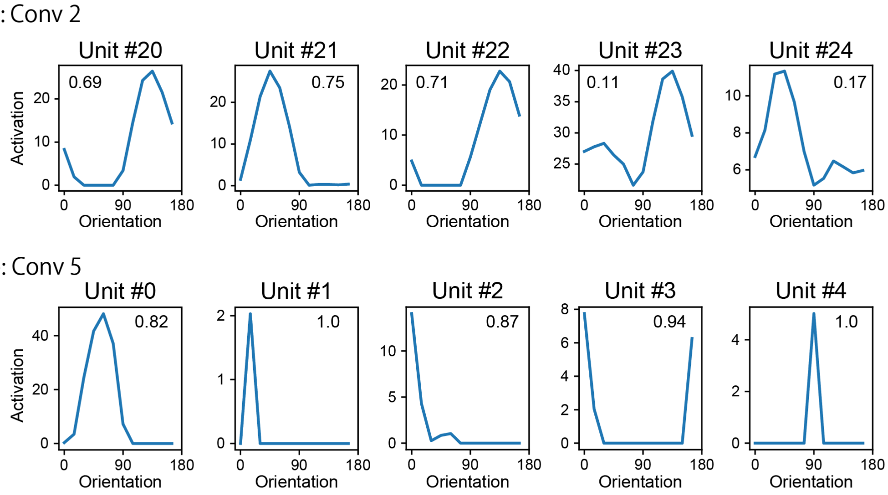

Figure 1: **Orientation-tuning curve of example units.** The gOSI value is shown in each panel.

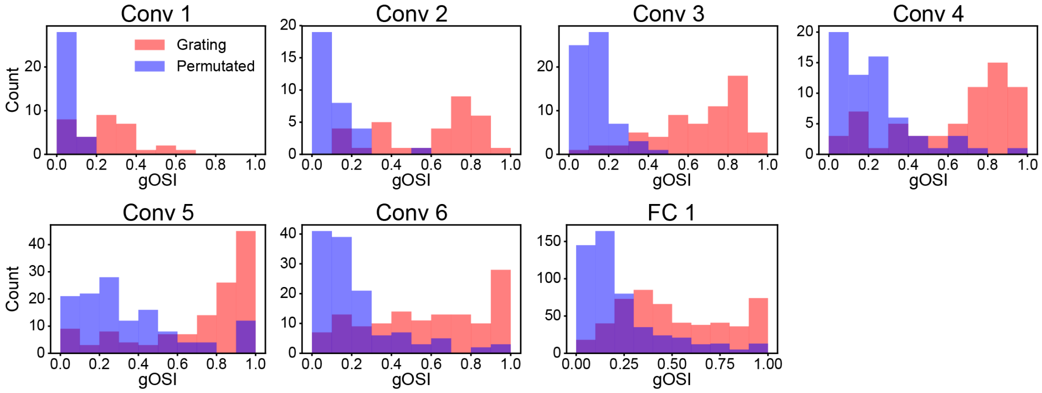

Figure 2: **Orientation-selective units exist in all hidden layers.** Histograms of gOSI for each layer are shown in red. Histograms of gOSI, after the activation matrix was shuffled, are shown in blue as a control.

original gOSI for most layers (Fig. 2), except for the high layers of the 20-layer CNN (Fig. A4), indicating that sparsity cannot fully explain the observed high gOSI.

3.2 Maturation of orientation selectivity during the course of training

A big challenge in visual neuroscience is to elucidate when and how orientation selectivity emerges and matures during development. To tackle this question, we analyzed the orientation selectivity during the course of training. Fig. 3a shows the average gOSI for each layer. The average gOSI of Conv 2–Conv 5 saturated at around 15 epochs, when validation loss also saturated, while that of Conv 6 and FC 1 saturated much earlier. To understand the relationship between the orientation selectivity and generalization, we then plotted the validation loss versus the average gOSI for each epoch (Fig. 3b), finding that the higher the average gOSI was, the less the validation loss was for the lower layers (Conv 1–Conv 5) (Fig. 3c). Similar results were obtained for a network with different initializations (Fig. A5). This strong correlation between gOSI and generalization suggests that, from the neuroscience perspective, orientation selectivity in the lower layers matures as the ability of object recognition matures, and from the machine-learning perspective, gOSI is a strong indicator for the generalization performance and is possibly a substitute signal for early stopping.

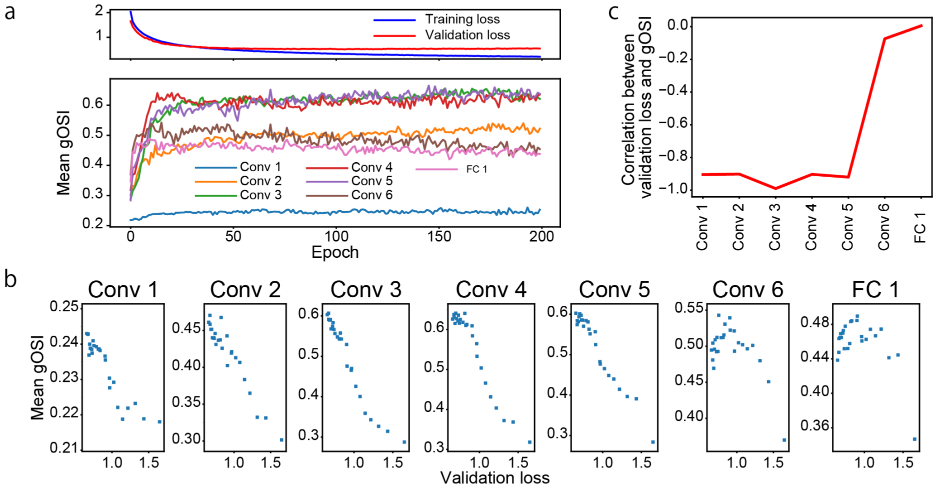

Figure 3: **Units in the lower layers become more orientation-selective in sync with generalization performance during the course of training.** (a) Evolution of loss (top) and average gOSI (bottom). (b) Relationship between the validation loss and average gOSI during the first 50 epochs. Each point indicates one epoch. (c) Spearman correlation coefficient between the validation loss and average gOSI during the first 50 epochs.

3.3 Relationship between orientation tuning and generalization performance

Section 3.2 triggers an intriguing question, viz., is orientation selectivity a correlate of generalization performance? We studied this question by comparing many networks trained in a different manner. We prepared four different settings of data augmentation (no augmentation; horizontal and vertical shift of the images; horizontal flip of the images; and horizontal shift, vertical shift, and horizontal flip were all incorporated). For each setting, we trained 50 models with independent initializations, resulting in generation of 200 networks with different generalization performances. All the networks had seven layers and were trained on CIFAR-10. When the generalization performance and gOSI of the 200 models were compared (Fig. 4a), we found a strong correlation between gOSI and generalization performance in the lower layers (Fig. 4b), except for Conv 1. Similar results were obtained for a set of 200 CNNs with different initializations (Fig. A6a and b).

A scenario of concern is that since the training loss and the test loss are correlated, orientation selectivity might be important only for the training loss, not for the test loss, and we are observing a side effect. To examine this possibility, we additionally trained six networks on CIFAR-10, whose labels were corrupted at different rates (p) from 0.0 (no labels were corrupted) to 1.0 (all labels were random) (Zhang et al., 2017). In this experiment, data augmentation was not incorporated so that networks were able to memorize the dataset. The training accuracy was nearly 100% for all models, whereas the test accuracy decreased as the corruption rate increased; test accuracy was approximately 70% for $p = 0.0$ and 10% for $p = 1.0$. We then compared the gOSI of the six networks (Fig. 4c), revealing that networks with lower test accuracy (higher label corruption rates) had lower gOSI in the low layers (especially in Conv 2–Conv 4; note that the y-axis range is different among the layers). Similar results were obtained for a set of CNNs with different initializations (Fig. A6c). These results indicate that orientation selectivity in the lower layers does correlate with the generalization performance.

3.4 Causal role of orientation tuning for generalization

Thus far, we have shown that networks that generalize better are more orientation-selective in the lower layers. We further investigated the *causal* relationship between orientation selectivity and generalization performance through *ablation* experiments (Morcos et al., 2018). The use of ablation experiments was inspired by several neuroscience studies in which the causal role in some behavior of a set of neurons is investigated by lesioning or inactivating them (e.g., using optogenetics).

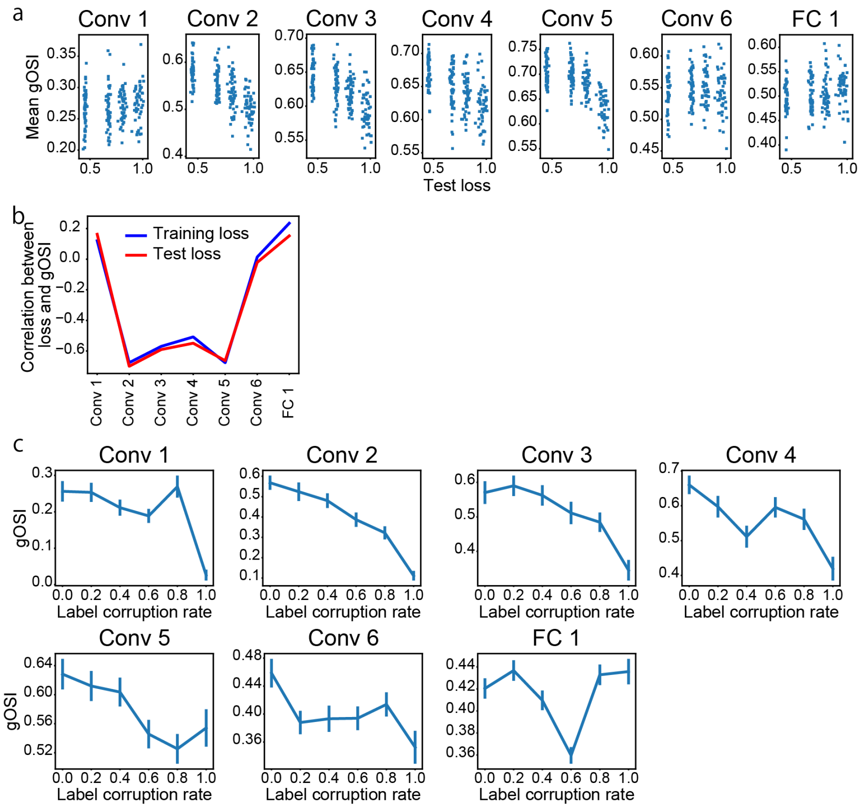

Figure 4: **Networks that generalize better are more orientation-selective in the lower layers.** (a) Relationship between the test loss and average gOSI. Each dot indicates one of the 200 trained models. (b) Spearman correlation coefficient between the loss and gOSI for each layer. (c) Relationship between the label corruption rate and gOSI. The higher the label corruption rate, the worse is the generalization accuracy (Zhang et al., 2017).

Similar to Morcos et al. (2018), after ablating a set of units by clamping their activations to zero, we computed the difference between the test losses before and after ablation. As shown in Fig. 5, for the lower layers (Conv 2 and Conv 3; indicated as asterisks), ablating the units with the top 50% gOSI values caused a more than twofold impact on test loss than ablating the units with the bottom 50% gOSI values. This difference in ablation impact cannot be explained by the possibility that the units with the bottom 50% gOSI values are more silent because the average activity of the units with the top 50% gOSI values in response to the grating images was lower than that of the units with the bottom 50% gOSI values, both for Conv 2 and Conv 3. Similar results were obtained for a network with different initializations and a 20-layer CNN trained on ImageNet[5] (Fig. A7; Conv 2 and Conv 3 in the 7-layer CNN; and block1_conv2, block1_pool, and block2_conv1 in the 20-layer CNN). These results suggest that orientation-selective units in the lower layers are causally important for the generalization of the networks.

3.5 A POSSIBLE MECHANISM OF ORIENTATION-SELECTIVE UNITS FOR GENERALIZATION

Finally, we examined how the orientation-selective units in the lower layers contribute to the overall generalization observed thus far. Because in neuroscience, orientation-selective filters ("simple cells" in V1) have been hypothesized to be combined to create a shift-invariant filter ("complex cells") (Movshon et al., 1978), we hypothesized that orientation-selective units in the lower layers of DNNs contribute to the invariance of the higher layers with respect to parallel shifts of the in-

[5] Impact on loss was evaluated on the 50,000 validation images of ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC2014). http://image-net.org/challenges/LSVRC/2014/download-images-5jj5.php

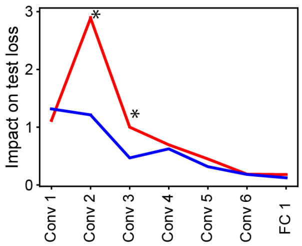

Figure 5: **Importance of orientation-selective units in the lower layers for generalization revealed through ablation experiments.** The difference between test losses before and after ablation is shown for each layer. Red: units with the top 50% gOSI values were ablated. Blue: units with the bottom 50% gOSI values were ablated. Asterisks indicate layers for which ablating units with the top 50% gOSI values yielded more than twofold impact than ablating units with the bottom 50% gOSI values.

put images. Here, we examined the degree of shift-invariance of each unit[6] in all layers using the following approach, which was inspired by Goodfellow et al. (2009) but slightly modified. After collecting activations of the target unit with respect to test images that activate the unit near maximally ($> 90\%$) and their shifted images[7], we computed how much the unit activations are influenced by the shift by using the coefficient of variation (CV) metric, which was afterward averaged along with the test images. We first confirmed that this metric decreases as the layer becomes deeper, indicating that unit activations of higher layers are more shift-invariant as suggested in (Goodfellow et al., 2009). When CVs computed in this way were compared between the vanilla network and the network where orientation-selective units of Conv 2 or Conv 3 were ablated, as we did in section 3.4, we found that units in the fully connected layer of the ablated network had significantly higher variances than those of the vanilla network (Fig. 6a), indicating that ablating orientation-selective units in the lower layers significantly disrupts the shift-invariance of the fully connected layer. Similar results were obtained with another 7-layer CNN with different initializations (Fig. 6b). These observations imply that orientation-selective units in the lower layers produce a part of shift-invariance of the fully connected layer, thereby contributing to the generalization performance.

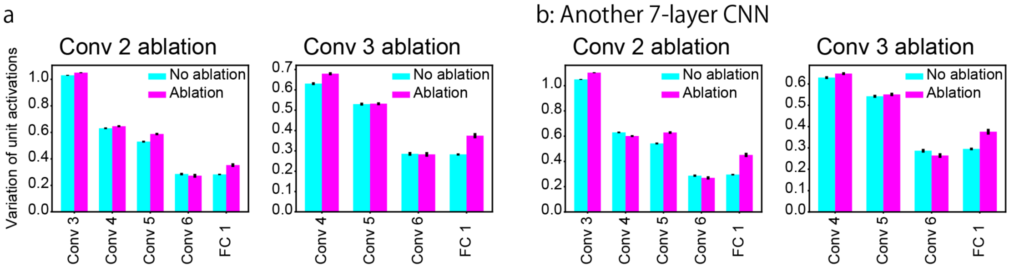

Figure 6: **Orientation-selective units are important for the shift-invariance of the fully connected layer.** Coefficients of variation of the unit activations with respect to the parallel shift of the test images are shown for each layer (mean ± standard error).

4 Related work and discussion

Our results suggest that, for the several network architectures we explored, the orientation selectivity in the lower layers of DNNs is causally indispensable for object recognition, not a superficial byproduct of object recognition. In addition, these representations are not necessary for memoriza-

[6] In this part, “unit” refers to each component of the feature maps of convolutional layers and unit of fully connected layers.

[7] we shifted the images in three ways; 10% vertical shift, 10% horizontal shift, or 10% horizontal and vertical shifts.

tion (Fig. 4c) but are important for generalization, at least by introducing shift-invariance of the fully connected layer.

4.1 Discussion from the neuroscience perspective

In neuroscience, orientation selectivity has been extensively analyzed in numerous papers since Hubel & Wiesel (1959) introduced the concept that neurons in V1 selectively respond to a specific orientation of bars. Nevertheless, whether the orientation-selective property does contribute to object recognition in natural scenes has not been examined thus far, possibly due to experimental limitations. Comparing the performance of object recognition with the degree of orientation selectivity among numerous well-trained animals (as we did in section 3.3) or inactivating orientation-selective neurons alone in the visual cortex (as we did in section 3.4) is experimentally very difficult or impossible with the current biotechnology. In this study, we tackle this issue by using DNN to model the brain. With this approach, we can obtain comprehensive data on neural activity, neural connectivity, and developmental process with infinite signal-to-noise ratio, at single-cell resolution, and chronically.

We also found that orientation-selective units exist in all layers of the DNNs. This is consistent with neuroscience studies where orientation-selective neurons exist not only in V1 but also in higher layers of the visual hierarchy, such as V4 (Desimone & Schein, 1987) and middle temporal (MT) area (Albright, 1984). However, as with the DNNs shown in this study, these orientation-selective neurons in the higher visual cortex might not encode the orientation per se.

We also analyzed orientation selectivity during the course of training and revealed that orientation selectivity in the lower layers matures as the ability of object recognition saturates. Again, this finding has not been proven in neuroscience, primarily due to experimental difficulty. Corresponding neuroscience experiments might involve the chronic recording of neurons during development and longitudinal comparison between the performance of object recognition and orientation selectivity, which is very difficult with current biotechnology.

4.2 Discussion from the machine-learning perspective

Generalization of DNN itself is intriguing and has been investigated in many papers, especially after Zhang's work (Zhang et al., 2017). Several researches have attributed the high generalization ability of DNNs to the convergence into flat minima (Keskar et al., 2017; Neyshabur et al., 2017). Wilson et al. (2017) proposed that stochastic gradient descent has an advantageous effect, and Ulyanov et al. (2018) proposed that the network structure itself is important. Although the relationship between these theories and orientation selectivity would be part of a future study, in this study we provide empirical evidence on the contribution of orientation-selective units in the lower layers to the overall generalization, partly via producing shift-invariance of the fully connected layer.

The role of orientation-selective units in the lower and higher layers might be different. In the lower layers (e.g. Conv 2), a Gabor filter is the optimal stimulus that activates the unit most strongly (Erhan et al., 2009; Krizhevsky et al., 2012; Zeiler & Fergus, 2014). This is also verified in neuroscience; Ukita et al. (2018) recently performed unbiased analyses to reveal that Gabor-like images indeed activate V1 neurons most strongly. On the other hand, Gabor filters might be suboptimal stimuli for units in the higher layers considering more elaborated features coded in the higher layers (Le et al., 2012; Simonyan et al., 2014). This might explain why orientation-selective units in the higher layers do not contribute to generalization as shown in the sections from 3.2 to 3.4. Examining the detailed role of orientation-selective units in the higher layers would be part of a future study.

When orientation selectivity is regarded as one of the low-level single directions, the results are opposite to those of a recent study (Morcos et al., 2018). Therefore, low-level single directions (e.g., orientation selectivity) and high-level single directions (e.g., class selectivity) might contribute differently to generalization. While selectivity to high-level directions might discourage distributed coding, selectivity to low-level directions might be important for embedding natural images. Further studies should investigate this discrepancy in more detail.

REFERENCES

Thomas D. Albright. Direction and orientation selectivity of neurons in visual area MT of the macaque Direction and Orientation Selectivity of Neurons in Visual Area MT of the Macaque. *Journal of Neurophysiology*, 52(6):1106–1130, 1984.

Inbal Ayzenshtat, Jesse Jackson, and Rafael Yuste. Orientation Tuning Depends on Spatial Frequency in Mouse Visual Cortex. *eNeuro*, 3(5), 2016.

Max Born. *Natural philosophy of cause and chance*. Oxford University Press, 1949.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12):e1003963, 2014.

Barbara Chapman and Michael P. Stryker. Development of orientation selectivity in ferret visual cortex and effects of deprivation. *Journal of Neuroscience*, 13(12):5251–5262, 1993.

Robert Desimone and Stanley J. Schein. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835–868, 1987.

Valentin Dragoi, Jitendra Sharma, Earl K. Miller, and Mriganka Sur. Dynamics of neuronal sensitivity in visual cortex and local feature discrimination. *Nature Neuroscience*, 5(9):883–891, 2002.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical report, University of Montreal*, pp. 1–13, 2009.

Ian J Goodfellow, Quoc V Le, Andrew M Saxe, Honglak Lee, and Andrew Y Ng. Measuring Invariances in Deep Networks. In *Advances in Neural Information Processing Systems 22*, pp. 646–654, 2009.

Umut Güçlü and Marcel A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(May):15037, 2017.

David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148(12):574–591, 1959.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, 2015.

Janneke F. M. Jehee, Sam. Ling, Jascha D. Swisher, Ruben S. van Bergen, and Frank Tong. Perceptual Learning Selectively Refines Orientation Representations in Early Visual Cortex. *Journal of Neuroscience*, 32(47):16747–16753, 2012.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations (ICLR)*, 2017.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11): e1003915, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

Ho Ko, Sonja B. Hofer, Bruno Pichler, Katherine A. Buchanan, P. Jesper Sjöström, and Thomas D. Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011.

Satoru Kondo and Kenichi Ohki. Laminar differences in the orientation selectivity of geniculate afferents in mouse primary visual cortex. *Nature Neuroscience*, 19(2):316–319, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Machine Learning (ICML)*, 2012.

Kairen Liu, Rana Ali Amjad, and Bernhard C. Geiger. Understanding Individual Neuron Importance Using Information Theory. *arXiv:1804.06679*, 2018.

Ari S. Morcos, David G. T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations (ICLR)*, 2018.

J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Receptive field organization of complex cells in the cat's striate cortex. *Journal of Physiology*, 283:79–99, 1978. ISSN 0022-3751. doi: VL-283.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems 30*, pp. 5947–5956, 2017.

Adam M. Packer, Botond Roska, and Michael Häusser. Targeting neurons and photons for optogenetics. *Nature Neuroscience*, 16(7):805–815, 2013.

Denise M. Piscopo, Rana N. El-Danaf, Andrew D. Huberman, and Cristopher M. Niell. Diverse Visual Features Encoded in Mouse Lateral Geniculate Nucleus. *Journal of Neuroscience*, 33(11): 4642–4656, 2013.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Aniek Schoups, Rufin Vogels, Ning Qian, and Guy Orban. Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412(6846):549–553, 2001.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR) Workshop*, 2014.

Jumpei Ukita, Takashi Yoshida, and Kenichi Ohki. Characterization of nonlinear receptive fields of visual neurons by convolutional neural network. *bioRxiv*, 2018.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Michael A Webster and Russell L De Valois. Relationship between spatial-frequency and orientation tuning of striate-cortex cells. *Journal of the Optical Society of America A*, 2(7):1124–1132, 1985.

Adrian Wertz, Stuart Trenholm, Keisuke Yonehara, Daniel Hillier, Zoltan Raics, Marcus Leinweber, Gergely Szalay, Alexander Ghanem, Georg Keller, Balzs Rózsa, Karl-klaus Conzelmann, and Botond Roska. Single-cellinitiated monosynaptic tracing reveals layer-specific cortical network modules. *Science*, 349(6243):70–74, 2015.

Leonard E. White, David M. Coppola, and David Fitzpatrick. The contribution of sensory experience to the maturation of orientation selectivity in ferret visual cortex. *Nature*, 411(6841):1049–1052, 2001.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In *Advances in Neural Information Processing Systems 30*, pp. 4148–4158, 2017.

F. Wörgötter and U. Th Eysel. Quantitative determination of orientational and directional components in the response of visual cortical cells to moving stimuli. *Biological Cybernetics*, 57(6): 349–355, 1987.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (23):8619–8624, 2014.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the Importance of Individual Units in CNNs via Ablation. *arXiv:1806.02891*, 2018.

A Appendix

A.1 Architecture and training detail of the 7-layer CNN

We analyzed a 7-layer CNN trained using the CIFAR-10 dataset. The network consisted of the input layer, six convolutional layers (Conv 1–Conv 6), one fully-connected layer (FC 1), and the output layer. For the convolutional layers, 32, 32, 64, 64, 128, 128 filters were used, in this order. The sizes of the filters were (3, 3). The stride was (1, 1), (2, 2), (1, 1), (2, 2), (1, 1), (2, 2), in this order. The fully connected layer had 512 units. ReLU activation and batch normalization (Ioffe & Szegedy, 2015) were used in both convolutional and fully connected layers.

The network was trained with the following setup: parameters were updated using the Adam optimizer (Kingma & Ba, 2014), the batch size was 32, and the total number of epochs was 200. The CIFAR-10 dataset was divided into 80% for training, 10% for validation, and the remaining 10% for the test; the validation accuracy was monitored for every epoch and the parameters with the highest accuracy for the validation set were used in the subsequent quantitative analyses. Data augmentation (horizontal shift, vertical shift, and horizontal flip of the images) was used, unless otherwise stated. In section 3.2, the network was trained with a small learning rate (0.0001) so that epochwise evolutions could be visualized clearly.

A.2 Minimal dependence of orientation selectivity on spatial frequencies

We investigated whether the orientation selectivity is influenced by SPFs. For an ideally orientation-selective neuron, the preferred orientation should be identical for all SPFs. In neuroscience, however, preferred orientations of V1 neurons depend partially on the SPFs, although the extent of dependence varies among the reports (Webster & De Valois, 1985; Ayzenshtat et al., 2016).

For each orientation-selective unit (gOSI > 0.33), we first took the maximum of the activation matrix along the phase dimension, which resulted in a new activation matrix with a shape of $N_{ori} \times N_{SPF}$. We then collected preferred orientations on some SPFs whose maximum activations were more than 50% of the overall maximum activations. We finally computed the range of these preferred orientations by the circular difference between the maximum degree and the minimum degree. We used this range as the metric of dependence of preferred orientations on SPFs. Note that this range is zero for an ideally orientation-selective unit. Interestingly, this range is zero for most orientation-selective units (Fig. A2). Although this range is large for some units, especially in higher layers, it is small considering the data of real V1 neurons (Fig. 2 of Ayzenshtat et al. (2016)). Collectively, these results indicate that orientation selectivity has minimal dependence on SPFs.

A.3 Orientation-selectivity of a random network

An interesting finding in visual neuroscience is that the initial formation of orientation selectivity does not require visual experiences (Chapman & Stryker, 1993; White et al., 2001). On the basis of these evidences, we also quantified the orientation selectivities on a randomly weighted network that was not trained using CIFAR-10. Surprisingly, but consistent with the neuroscience findings, we observed that a small number of units were orientation-selective in this random network (Fig. A8), although the selectivity was much weaker than that of well-trained networks.

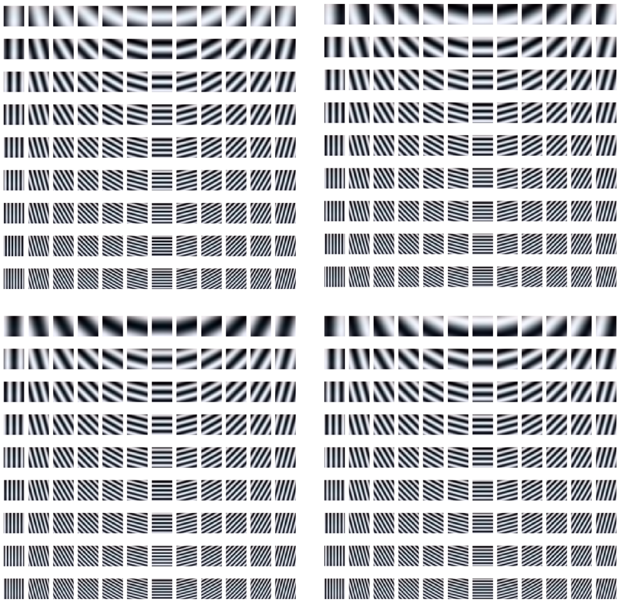

Figure A1: **All grating images presented to the 7-layer CNNs trained on CIFAR-10.**

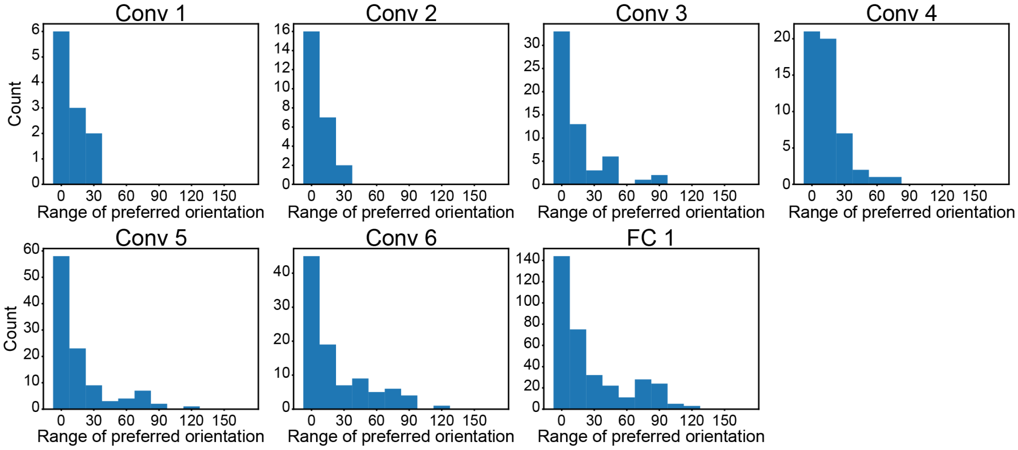

Figure A2: **Minimal dependence of orientation selectivity on spatial frequencies.** The range of preferred orientations among various spatial frequencies was computed for each orientation-selective unit (gOSI > 0.33) and their distribution is plotted.

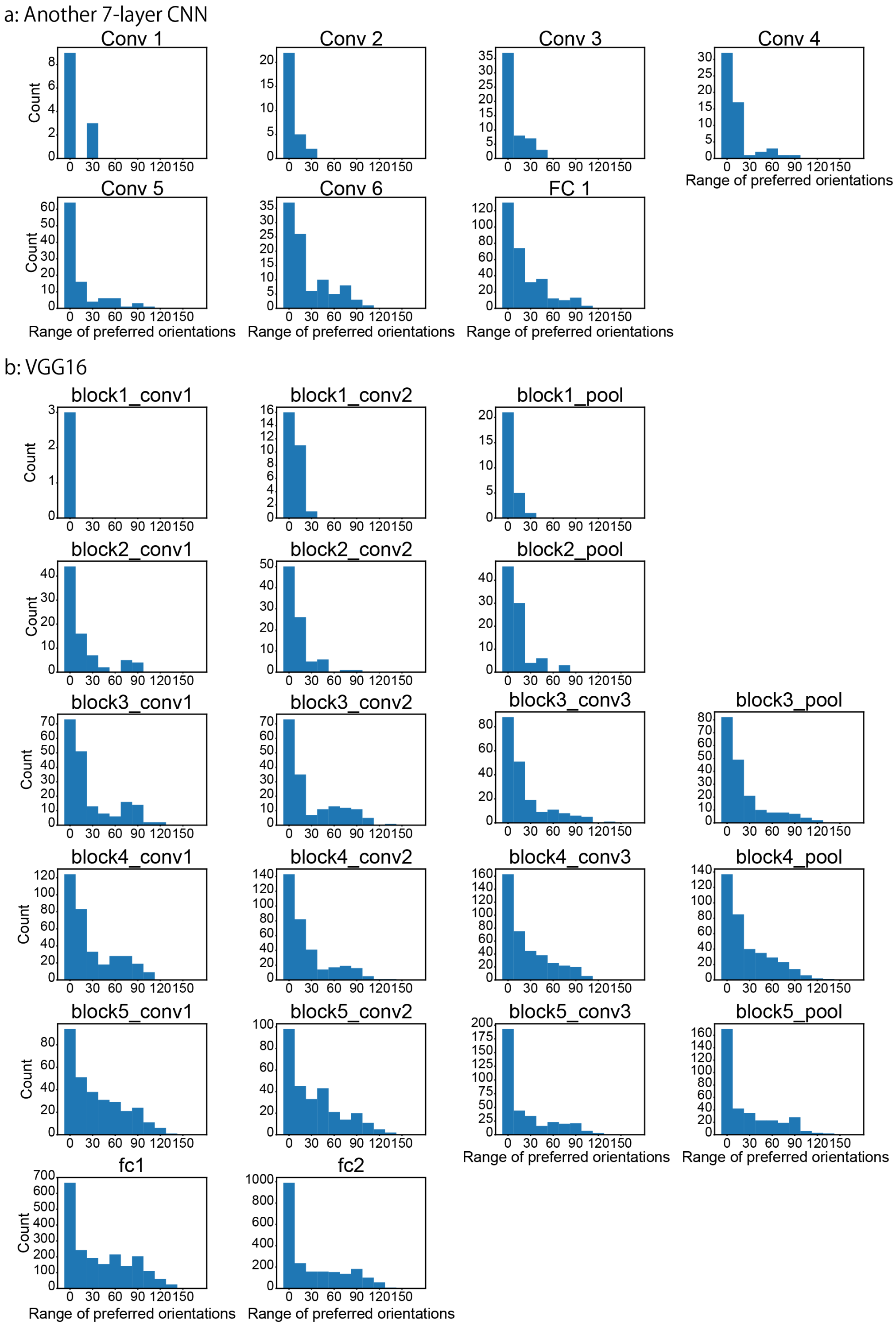

Figure A3: **Reproduction of Fig. A2 using (a) a 7-layer CNN with different initializations and (b) a 20-layer CNN trained on ImageNet.** The range of preferred orientations among various spatial frequencies was computed for each orientation-selective unit (gOSI > 0.33) and their distribution is plotted.

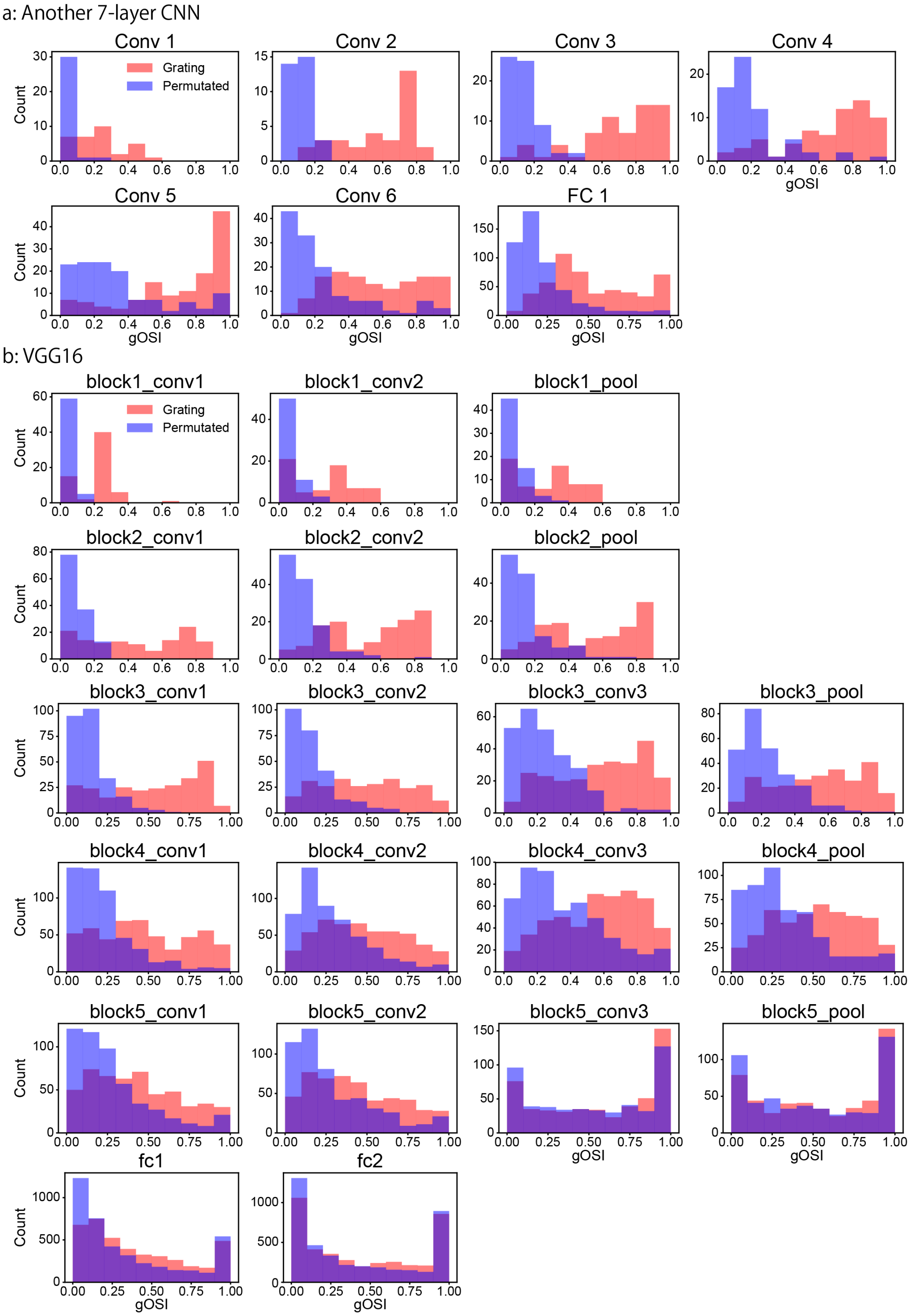

Figure A4: **Reproduction of Fig. 2 using (a) a 7-layer CNN with different initializations and (b) a 20-layer CNN trained on ImageNet.** Histograms of gOSI for each layer are shown in red. Histograms of gOSI, after the activation matrix was shuffled, are shown in blue as a control.

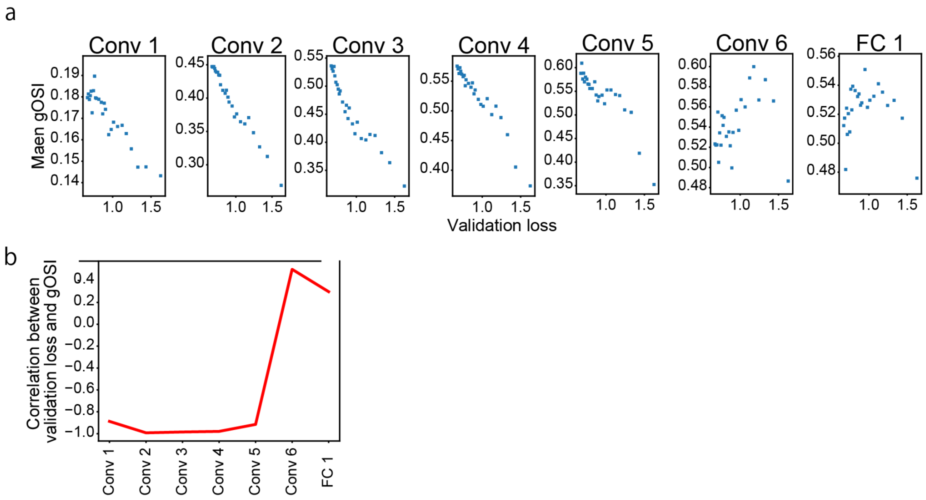

Figure A5: **Reproduction of Fig. 3 using a 7-layer CNN with different initializations.** (a) Relationship between the validation loss and average gOSI during the first 50 epochs. Each point indicates one epoch. (b) Spearman correlation coefficient between the validation loss and average gOSI during the first 50 epochs.

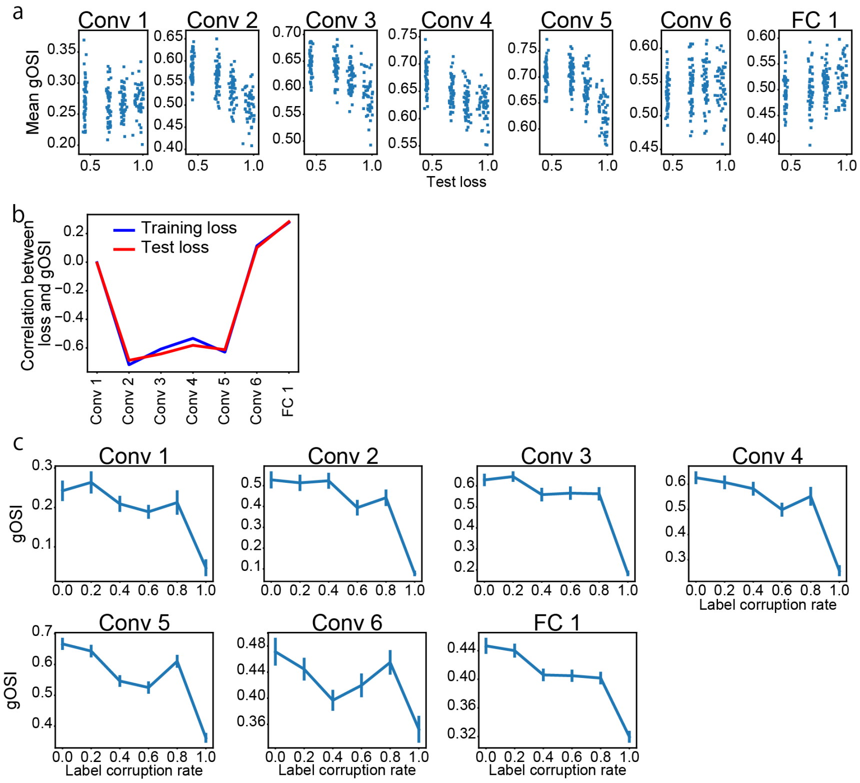

Figure A6: **Reproduction of Fig. 4 using 7-layer CNNs with different initializations.** (a) Relationship between the test loss and average gOSI. Each dot indicates one of the 200 trained models. (b) Spearman correlation coefficient between the loss and gOSI for each layer. (c) Relationship between the label corruption rate and gOSI. The higher the label corruption rate, the worse is the generalization accuracy (Zhang et al., 2017).

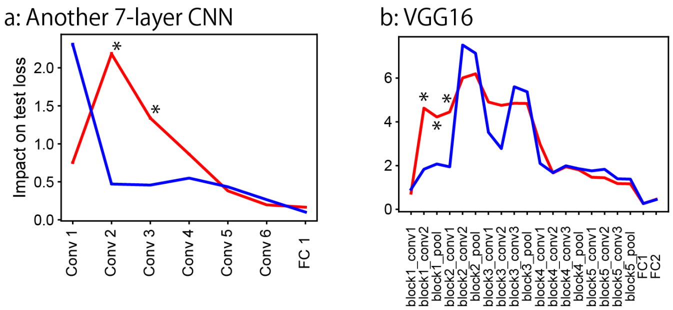

Figure A7: **Reproduction of Fig. 5 using (a) a 7-layer CNN with different initializations and (b) a 20-layer CNN trained on ImageNet.** The difference between test losses before and after ablation is shown for each layer. Red: units with the top 50% gOSI values were ablated. Blue: units with the bottom 50% gOSI values were ablated. Asterisks indicate layers for which ablating units with the top 50% gOSI values yielded more than twofold impact than ablating units with the bottom 50% gOSI values.

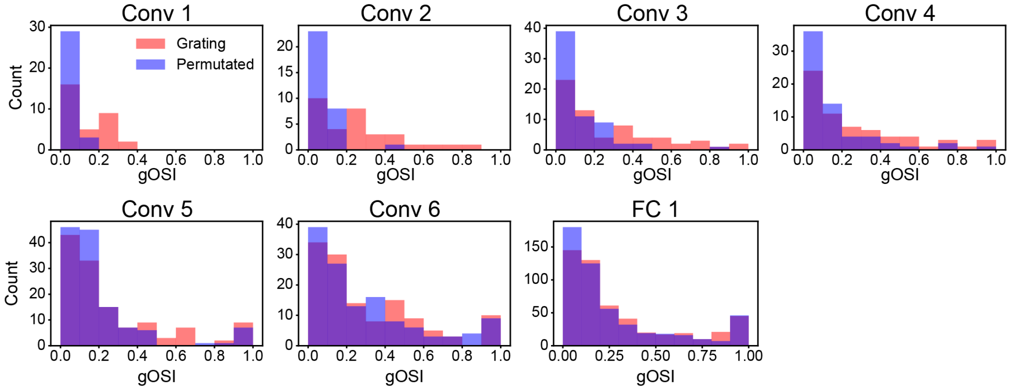

Figure A8: **A small number of units is already orientation-selective before training.** Orientation selectivity was analyzed on a network with random weights.