Journal Club

**Editor's Note: These short reviews of recent *JNeurosci* articles, written exclusively by students or postdoctoral fellows, summarize the important findings of the paper and provide additional insight and commentary. If the authors of the highlighted article have written a response to the Journal Club, the response can be found by viewing the Journal Club at www.jneurosci.org. For more information on the format, review process, and purpose of Journal Club articles, please see http://jneurosci.org/content/preparing-manuscript#journalclub.**

# Can Deep Learning Model Perceptual Learning?

**Shahab Bakhtiari**
Integrated Program in Neuroscience, Montreal Neurological Institute, McGill University, Montreal, Quebec H3A 2B4, Canada

Review of Wenliang and Seitz

For a radiologist to be able to detect abnormal tissue in an x-ray image, years of deliberate practice is needed. This long-lasting, acquired expertise in visual perception as a result of experience is known as visual perceptual learning (VPL). What kind of neuroplasticity explains this behavioral phenomenon?

To answer this question, researchers have used psychophysical experiments to better understand the behavioral characteristics of VPL (Dosher and Lu, 2017). As an example, in a simple orientation discrimination task, two static sinusoidal grating patterns are presented sequentially, and subjects indicate whether the second pattern (target stimulus) is rotated clockwise or counterclockwise compared with the first one (reference stimulus). The separation angle of the two gratings determines the required precision of the task. After days of training, humans and nonhuman animals improve on the task, and their visual sensitivity increases (Dosher and Lu, 1998; Schoups et al., 2001).

One important but limiting aspect of this behavioral improvement is its specificity: the learned expertise is limited to the visual features of the training task. In the same orientation discrimination task, changing the retinal location of the stimulus or the orientation of the reference grating causes the subjects' performance to drop back to baseline. In other words, there is limited or no transfer of the learned expertise (Fiorentini and Berardi, 1980; Fahle, 2005). This specificity was the basis of the conjecture that plasticity in primary visual cortex, where neuronal receptive fields are small and narrowly tuned, underlies VPL. In the orientation discrimination task, according to this hypothesis, the training retunes only those neurons in the primary visual cortex that are activated by the training stimulus. The resulting improvement in behavioral sensitivity is therefore limited to tasks requiring the activated neurons (Fiorentini and Berardi, 1981).

Contrary to this hypothesis, however, a seminal article by Ahissar and Hochstein (1997) drew a more complex picture and showed that the specificity of the VPL depended on the visual precision of the training task. An easy, coarse orientation discrimination (i.e., large separation angles between the target and the reference stimuli) led to more transfer of the learning effect to new locations and stimuli compared with a difficult, fine orientation discrimination (Ahissar and Hochstein, 1997). Based on these observations, Ahissar and Hochstein (2004) proposed the reverse hierarchy theory (RHT) as the mechanism of VPL. According to this theory, learning starts in higher visual areas and moves down to lower visual areas only if higher areas are unable to perform the task. Thus, training on less precise visual tasks modifies the neural representation only in higher visual areas, which can discriminate relatively large changes in stimuli. In the orientation discrimination task, for example, when the reference and the target stimuli have large angle separations (e.g., 45°), the two stimuli activate two separate subpopulations of neurons in higher visual areas, which is sufficient for solving the task. Training on these tasks therefore modifies the neural representation in the higher visual areas. However, when the angle separation is small, the two stimuli activate the same subpopulation of higher visual areas. These more precise and fine-scale visual tasks need neurons with narrower tuning and smaller receptive fields, which are present only in early visual areas. The reduced specificity of low precision training, therefore, mirrors the large receptive fields of higher visual areas, as the same neurons respond to a wider range of stimuli, and performance on untrained stimuli will also be enhanced. In contrast, the high specificity of high-precision training mirrors the smaller receptive fields of primary visual areas that respond only to a narrow range of stimuli and is not responsive to untrained stimuli (Ahissar and Hochstein, 2004).

The RHT does not suggest any mechanism for the training process that would cause selective plasticity dictated by task precision, but in a recent study, Wenliang and Seitz (2018) showed that the properties of the RHT emerge from a deep neural network (DNN), with appropriate train-

Received Aug. 28, 2018; revised Oct. 30, 2018; accepted Nov. 1, 2018.

I thank Dr. Christopher C. Pack for comments and guidance in the preparation of the manuscript.

Correspondence should be addressed to Shahab Bakhtiari, 3801 University Street, Room 896, Montreal, QC H3A 2B4, Canada. E-mail: shahab.bakhtiari@mail.mcgill.ca.

https://doi.org/10.1523/JNEUROSCI.2209-18.2018

ing. Inspired by visual cortex, DNNs are made of layers of neuron-like feature detectors that are organized in a strictly hierarchical fashion, with every layer receiving its input from its preceding layer, and only the last layer projecting to a decoder. DNNs are end-to-end trainable and can model complex input–output associations. Importantly, they can reach human-level accuracy in natural image classification (Krizhevsky et al., 2012). Once trained on natural image classification, the learned feature detectors mimic the increasing complexity of visual cortex with the early layers tuned to simple features, such as orientation, and deeper layers selective for more complex visual forms (Zeiler and Fergus, 2014). Their similarity to visual cortex, and especially their hierarchical structure, makes DNNs a natural way to model VPL, particularly as envisioned by the RHT.

By training a DNN on an orientation discrimination task, Wenliang and Seitz (2018) showed that important behavioral and physiological findings in the VPL literature can be reproduced by the model (Wenliang and Seitz, 2018). To show this, they used a DNN architecture (known as AlexNet) pretrained on an image classification task (Krizhevsky et al., 2012). They retrained the network on an orientation discrimination task similar to the one described above. They hypothesized that first, the output behavior of the trained DNN would replicate the behavioral features of the VPL; and, second, the training-induced changes in the artificial neurons would be qualitatively similar to the neurophysiological changes reported in nonhuman primates and follow a layer-selective pattern that would be determined by the task precision, as predicted by the RHT.

To simulate the VPL paradigm, Wenliang and Seitz (2018) trained the DNN on five different precision levels of the orientation discrimination task (0.5–10° separation angle) and measured the learning effect with the training stimulus as well as other spatial frequencies and orientations that were not shown to the DNN during the training. Consistent with experimental observations (Ahissar and Hochstein, 1997; Jeter et al., 2009), the DNN showed less transfer of improved sensitivity to new stimuli when trained on the high-precision version of the task and more transfer with less precise training. Moreover, the DNN reached the maximum accuracy faster with the low-precision task compared with the higher-precision task. Like in humans, learning a higher-precision task was more difficult for the DNN and led to less transfer to new stimulus conditions.

In the DNN model, the highest precision version of the task required more changes in the first layer of the network compared with the other layers (Wenliang and Seitz (2018), their Fig. 3*D*). The first and second layers of the DNN resemble the primary visual cortex in terms of visual feature selectivity (Khaligh-Razavi and Kriegeskorte, 2014). Therefore, the modifications in this layer were specific to the training stimulus, and the learned improvement did not transfer to other test stimuli (e.g., with different orientations), which was consistent with the RHT prediction. Interestingly, the changes in the tuning curves of the DNN neurons also qualitatively followed the pattern of changes that previous electrophysiology studies had shown in nonhuman primates (e.g., increase in the slope of the tuning curves) and could reconcile apparent inconsistencies, such as variable tuning changes reported by different electrophysiology studies (Schoups et al., 2001; Ghose et al., 2002).

The DNN model by Wenliang and Seitz (2018) and some previous models attributed the behavioral improvement by VPL to retuning of sensory neurons (Goldstone, 1998). However, the empirically observed improvement by VPL is not necessarily caused by retuning of sensory neurons. Models that associate the behavioral effect of VPL with plasticity in sensory representations, such as the DNN model, assume that, before training, the information represented by the sensory areas are optimally pooled and read out by areas downstream of visual cortex, and the sensory representation is the limiting factor for the behavior. In the orientation discrimination task, for example, an optimal decoder reads out information from all sensory neurons with different tuning properties, and neurons with similar tunings to the task reference orientation contribute the most in the decision. In this scenario, VPL improves behavior by changing the tuning of the sensory neurons to become more informative for the task. This is true for the DNN, but not necessarily for the visual system. When the readout is not optimal (Parker and Newsome, 1998), that is, when the most informative neurons are not contributing the most in the decision, performance improvement by VPL mainly involves optimizing the readout from these neurons based on the training task. The VPL models proposed before the study by Wenliang and Seitz (2018) suggested a similar reweighting mechanism in which the learning process corresponds to searching for the most informative neurons to read out from, with no change in their tuning (Jacobs, 2009; Dosher et al., 2013).

Retuning the sensory neurons (representation model) and reweighting the readout (reweighting model) can both generate aspects of the VPL behavioral features, but not with the same mechanism. In the reweighting model, unlike the DNN, the final decoder neuron has direct access to all the neurons in different layers. The reweighting model assigns a larger weight to early visual areas in the high-precision task, instead of reparameterizing their tuning curves, which leads to more specificity (Dosher and Lu, 1998). However, as Wenliang and Seitz (2018) showed, in the strictly hierarchical structure of the DNN, even without the decoder having direct access to the early areas, the selective tuning of early layers affects the learning. It is indeed crucial to understand which one of these two mechanisms mainly explains the VPL specificity.

From the neurophysiology perspective, the observed changes in the tuning curves (Schoups et al., 2001) are not easily explained by the reweighting model. Therefore, one can argue that the DNN model is better suited for describing the VPL as it can reproduce the tuning curve changes. However, the changes in the tuning curves, although replicated in few studies, are not able to explain the large behavioral effects of VPL (Schoups et al., 2001). In the study by Wenliang and Seitz (2018), since the deep pure hierarchical architecture is not able to fully implement the reweighting model (see below), the major part of the improvement is caused by the tuning curve changes across different layers. Adding connections to the DNN from all the layers to the final decoder layer (called "skip connections") can potentially combine the ideas of the reweighting model and the representation model in one single framework, by which we can evaluate the relative contribution of retuning and reweighting to behavioral improvement. The existence of these skip connections in the visual cortex is also supported by anatomical studies (Felleman and Van Essen, 1991).

A method that can potentially differentiate between the two models is an inactivation experiment. In a recent study, Liu and Pack (2017) showed that inactivation of middle temporal area (MT) in the visual dorsal pathway influenced motion perception thresholds only after a monkey was trained on a stimulus that better stimulated MT (Liu and Pack, 2017). This training, according to the reweighting model, increases the readout weight of MT neurons as the most informative neurons for the task. Before the training, the other, lower areas had larger readout weights, and inactivating MT did not change motion perception. Without the skip connection in the DNN, regardless of the training task, inactivation of a deep layer blocks the path of visual information propagating up to the final decoder layer. Therefore, in the architecture used by Wenliang and Seitz (2018), the DNN is not able to explain this inactivation result. However, in an extended DNN with skip connections, the inactivation experiment can be simulated and compared with the empirical findings (Liu and Pack, 2017).

In summary, Wenliang and Seitz (2018) demonstrated that the DNN provides a theoretical explanation for many aspects of VPL. However, more elaborate comparisons between different network architectures and human/animal data are needed to map out the limitations of the DNN in explaining this phenomenon.

## References

Ahissar M, Hochstein S (1997) Task difficulty and the specificity of perceptual learning. Nature 387:401–406. CrossRef Medline

Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. Trends Cogn Sci 8:457–464. CrossRef Medline

Dosher BA, Lu ZL (1998) Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proc Natl Acad Sci U S A 95:13988–13993. CrossRef Medline

Dosher B, Lu ZL (2017) Visual perceptual learning and models. Annu Rev Vis Sci 3:343–363. CrossRef Medline

Dosher BA, Jeter P, Liu J, Lu ZL (2013) An integrated reweighting theory of perceptual learning. Proc Natl Acad Sci U S A 110: 13678–13683. CrossRef Medline

Fahle M (2005) Perceptual learning: specificity versus generalization. Curr Opin Neurobiol 15:154–160. CrossRef Medline

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47. CrossRef Medline

Fiorentini A, Berardi N (1980) Perceptual learning specific for orientation and spatial frequency. Nature 287:43–44. CrossRef Medline

Fiorentini A, Berardi N (1981) Learning in grating waveform discrimination: specificity for orientation and spatial frequency. Vision Res 21:1149–1158. CrossRef Medline

Ghose GM, Yang T, Maunsell JH (2002) Physiological correlates of perceptual learning in monkey V1 and V2. J Neurophysiol 87:1867–1888. CrossRef Medline

Goldstone RL (1998) Perceptual learning. Annu Rev Psychol 49:585–612. CrossRef Medline

Jacobs RA (2009) Adaptive precision pooling of model neuron activities predicts the efficiency of human visual learning. J Vis 9(4):22.1–15. CrossRef Medline

Jeter PE, Dosher BA, Petrov A, Lu Z-L (2009) Task precision at transfer determines specificity of perceptual learning. J Vis 9(3):1.1–13. CrossRef Medline

Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10:e1003915. CrossRef Medline

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Paper presented at Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV, December.

Liu LD, Pack CC (2017) The contribution of area MT to visual motion perception depends on training. Neuron 95:436–446.e3. CrossRef Medline

Parker AJ, Newsome WT (1998) Sense and the single neuron: probing the physiology of perception. Annu Rev Neurosci 21:227–277. CrossRef Medline

Schoups A, Vogels R, Qian N, Orban G (2001) Practising orientation identification improves orientation coding in V1 neurons. Nature 412: 549–553. CrossRef Medline

Wenliang LK, Seitz AR (2018) Deep neural networks for modeling visual perceptual learning. J Neurosci 38:6028–6044. CrossRef Medline

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. Paper presented at ECCV ’14: European Conference on Computer Vision, Zurich, Switzerland, September.