Self-supervised Neural Network Models of Higher Visual Cortex Development

Chengxu Zhuang[1], **Siming Yan**[2], **Aran Nayebi**[1], **Daniel Yamins**[1]
1: Stanford University, 2: Peking University

Abstract

Deep convolutional neural networks (DCNNs) optimized for visual object categorization have achieved success in modeling neural responses in the ventral visual pathway of adult primates. However, training DCNNs has long required large-scale labelled datasets, in stark contrast to the actual process of primate visual development. Here we present a network training curriculum, based on recent state-of-the-art self-supervised training algorithms, that achieves high levels of task performance without the need for unrealistically many labels. We then compare the DCNN as it evolves during training both to neural data recorded from macaque visual cortex, and to detailed metrics of visual behavior patterns. We find that the self-supervised DCNN curriculum not only serves as a candidate hypothesis for visual development trajectory, but also produces a final network that accurately models neural and behavioral responses.

Keywords: Ventral visual stream; visual development; deep convolutional neural networks; self-supervised learning; semi-supervised learning

Introduction

Deep Convolutional Neural Networks (DCNNs) trained to solve high-variation object recognition tasks have been shown to be quantitatively accurate models of neural responses in the ventral visual stream of adult primates (Yamins et al., 2014; Kriegeskorte, 2015; Cadena et al., 2019). These same networks also generate visual behavior patterns that are consistent with those of humans and macaques (Rajalingham et al., 2018). Though this progress at the intersection of machine learning and computational neuroscience is intriguing, there is a fundamental problem confronting these approaches: training DCNNs has long used heavily supervised methods involving huge numbers of high-level semantic labels. For example, state-of-the-art DCNNs for visual object recognition are typically trained on ImageNet (Deng et al., 2009), comprising 1.2 million labeled images. When viewed as mere technical tools for tuning network parameters, such procedures can be acceptable, although they limit the purview of the methods to situations with large existing labelled datasets. However, as real models of biological learning, they are highly unrealistic, since humans and primates develop their visual systems with very few explicit labels (Braddick & Atkinson, 2011; Atkinson, 2002; Harwerth, Smith, Duncan, Crawford, & Von Noorden, 1986; Bourne & Rosa, 2005). This developmental difference significantly undermines the effectiveness of DCNNs as models of visual learning.

Motivated by the need for more label-efficient training procedures, deep learning researchers have been actively developing methods of self- and semi-supervised learning algorithms that train DCNNs with very few or no labels (Caron et al., 2018; Wu et al., 2018; Doersch et al., 2015; Zhang et al., 2016; Zhuang, Zhai, & Yamins, 2019; Tarvainen & Valpola, 2017; Zhuang, Ding, et al., 2019). Although the performance of these models is still below that of their supervised counterparts, recent progress with deep embedding methods has shown substantial promise in bridging this gap (Caron et al., 2018; Wu et al., 2018; Zhuang, Zhai, & Yamins, 2019; Zhuang, Ding, et al., 2019).

In this work, we show that the feature representations learned by these state-of-the-art self- and semi-supervised DCNNs model the adult ventral visual stream just as effectively as those of category-supervised networks. This, in turn, allows us to design a training curriculum that better simulates the real developmental trajectory of the visual system. More specifically, we first train DCNNs on self-supervised visual tasks, simulating the learning in early infancy when supervision is entirely absent (Cooper & Aslin, 1989). We find that the best of these "self-DCNNs" — based on the state-of-the-art Local Aggregation approach to training deep embeddings (Zhuang, Zhai, & Yamins, 2019) — predicts neural responses in primate V4 and IT areas with high accuracy, mapping these brain areas to anatomically reasonably intermediate and late hidden layers respectively. In fact, this self-DCNN has somewhat better V4 and IT neural prediction performance than its supervised counterpart. Unsurprisingly, however, the visual categorization behavior produced by the output layer of the task-generic self-DCNN is somewhat less consistent with that of humans and monkeys than that of network trained on categorization tasks. To address this gap, we then train the self-DCNNs using semi-supervised algorithms with a very small number of labelled data points, simulating late infancy when a modest amount of supervision occurs (Cooper & Aslin, 1989; Topping, Dekhinet, & Zeedyk, 2013). The resulting "semi-DCNNs" produce behavioral outputs that are substantially more similar to that of humans and monkeys than those of self-DCNNs, approaching the behavioral consistency levels of supervised networks. Taken together, our results suggest that this "self-then-semi-supervised" curriculum not only produces accurate models of the ventral visual stream, but also may serve as a candidate quantitative hypothesis for neural changes during visual development.

Results

Our curriculum approximates visual development using two stages: a *self-supervised stage* and a *semi-supervised stage* (see Fig. 1**A**). To evaluate the effectiveness of the trained DCNNs in modeling the ventral visual stream, we report their neural prediction performance for neurons in V4 and IT cor-

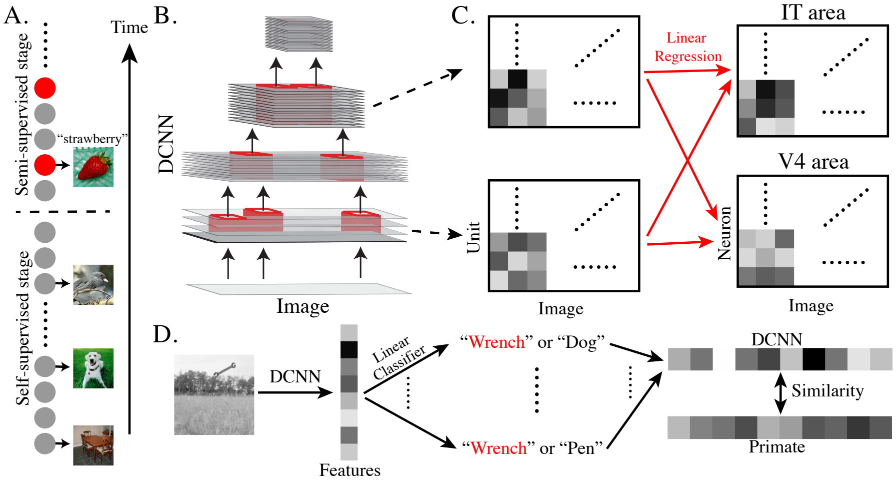

Figure 1: Illustration of the DCNN training curriculum and evaluation metrics for modeling the ventral visual stream. (**A**) Two-stage "self-then-semi-supervised" training curriculum. Gray dots represent unlabeled visual inputs, red dots labeled visual inputs. (**B,C**) To compute neural prediction performance for V4 and IT cortical areas, we first run the DCNNs on the stimulus for which neural responses were collected. DCNN unit activations from each convolutional layer are then used to predict the V4 and IT neural responses with regularized linear regression as in (Klindt et al., 2017). For each neuron, the Pearson correlation on held-out validation images between the DCNN-predicted responses and the recorded responses is computed. The medians of these Pearson correlations across neurons in each brain area are then reported as summary statistics. Neural data are from (Majaj et al., 2015). (**D**) To measure behavioral consistency, we follow the method of (Rajalingham et al., 2018; Schrimpf et al., 2018), in which linear classifiers are trained from each model's penultimate layer on a set of images from 24 common classes ("dog", "wrench", "pen", &c). The resultant image-by-category confusion matrix is compared to data from humans and monkeys performing the same alternative forced choice task. We report the "behavioral predictivity" metric in (Schrimpf et al., 2018).

tical areas (see Fig. 1**C**) and visual categorization behavioral consistency to humans and monkeys (see Fig. 1**D**).

Self-supervised stage. In this stage, we compare several representative self-DCNNs, which represent different hypotheses for visual learning in early infancy. These can be organized into two families: *single-image statistic* (SIS) algorithms (Fig. 2**A**) and *multi-image distribution* (MID) algorithms (Fig. 2**B**). SIS algorithms, including image colorization (Zhang et al., 2016) and surface-normals/depth estimation (Laina et al., 2016), supervise models with statistics either directly extracted from visual inputs or available from other biological sensory signals (see Fig. 2**A**). MID algorithms optimize visual representations so that the feature distribution across many inputs meets some specific requirement (see Fig. 2**B**). We find that SIS-based self-DCNNs accurately predict neural responses for V4 neurons (but not for IT neurons) while MID-based self-DCNNs show high levels of neural prediction performance for both V4 and IT neurons (see Fig. 2**C**). All self-DCNNs show a similar brain mapping correspondence, with mid-level representations best predicting V4 neural responses and high-level representations best predicting IT neural responses (see Fig. 2**D**). We find that the self-DCNN based on the Local Aggregation (LA), which has recently been shown to achieve state-of-the-art unsupervised visual recognition performance (Zhuang, Zhai, & Yamins, 2019), also achieves optimal neural prediction performance.

Semi-supervised stage. Although the internal layers of the LA-DCNN show high accuracy in predicting V4 and IT neural responses, its final-layer outputs are less consistent with fine-grained metrics of visual categorization behavior in humans and monkeys than those of category-supervised DCNNs (see Fig. 3**C**). To fill this gap, we further train LA-DCNN using semi-supervision with 3% of labels in the ImageNet dataset — e.g. 36k distinct labelled datapoints, a number that is consistent with developmental measurements (Topping et al., 2013). We test two recent high-performing semi-supervised algorithms, including Mean Teacher (Tarvainen & Valpola, 2017) (MT, see Fig. 3**A**), and Local Label Propagation (Zhuang, Ding, et al., 2019) (LLP, see Fig. 3**B**). We find that LLP is significantly more behaviorally consistent than MT, and all the unsupervised networks. This result is consistent with the fact that LLP substantially outperforms MT on visual tasks, especially in the low-label regime (Zhuang, Ding, et al., 2019).

Discussion and Future Directions

While our results represent progress toward a plausible account of how un- and semi-supervised visual experience may shape higher visual representations, there is still a significant gap in behavioral consistency between the LLP-DCNN and the best supervised DCNNs (Rajalingham et al., 2018). We

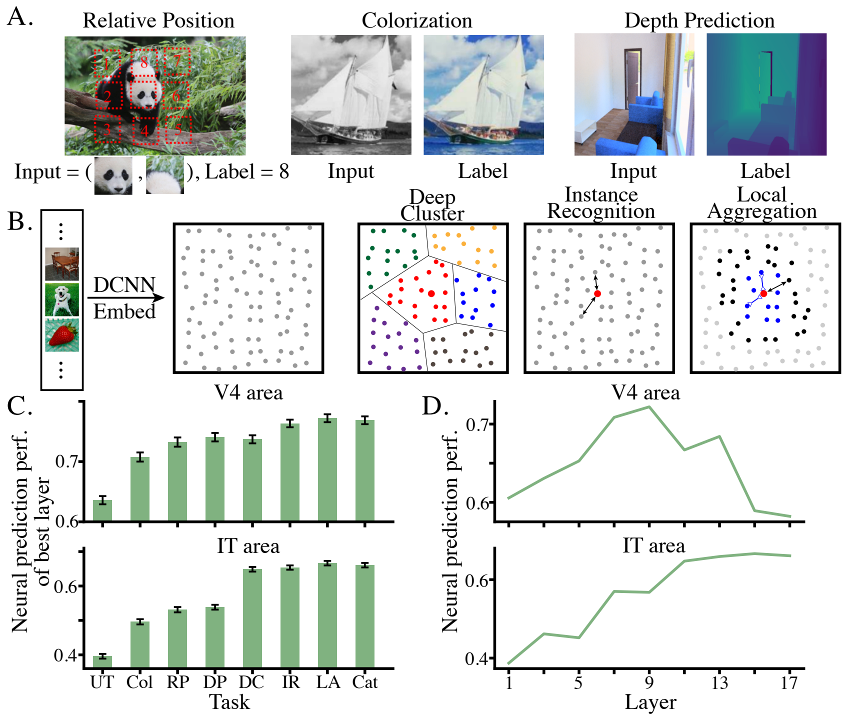

Figure 2: Self-supervised visual tasks and neural prediction. (**A**) Single-Image Statistics (SIS) tasks. For Relative Position (RP) (Doersch et al., 2015), the DCNN receives two patches from one image and classifies their spatial organization. For Colorization (Col) (Zhang et al., 2016), the DCNN receives a gray-scale image and predicts its per-pixel RGB values. For Depth Prediction (DP), the DCNN receives a RGB image and predicts the per-pixel relative depth generated by standardizing the depth across the whole image. The implementation follows (Laina et al., 2016). (**B**) Schematic for Multi-Image Distribution (MID) algorithms. All images are first embedded into a lower dimensional space using DCNNs. For Deep Cluster (DC) (Caron et al., 2018), K-means is applied to the embeddings and the cluster labels are used as category labels to train the DCNN. For Instance Recognition (IR) (Wu et al., 2018), the DCNN is optimized to maximize the distances between the embedding of the current input (red dot) and the embeddings of all the other images (gray dots). For Local Aggregation (LA) (Zhuang, Zhai, & Yamins, 2019), the DCNN is optimized to minimize the distance to “close” embedding points (blue dots) and to maximize the distance to the “further” embedding points (black dots) for the current input (red dot). (**C**) V4 and IT neural prediction performance of the best layer in DCNNs trained using different self-supervised tasks. “UT” represents an untrained ResNet-18. “Cat” represents a ResNet-18 trained on the ImageNet categorization task. Error bars represent standard deviation computed from 5 independent train-validation splits. (**D**) V4 and IT neural prediction performance of all layers for LA-trained DCNN. The x-axis represents the number of convolutional layers away from the inputs.

hope future semi-supervised algorithms can help bridge this gap. Moreover, the current DCNN training uses ImageNet images, whose statistics are likely to be meaningfully different from the visual inputs infants perceive during visual development. For example, DCNNs receive static and independent images during training, whereas infants receive a continuous stream of inputs that are temporally correlated and object-centric (Bambach, Crandall, Smith, & Yu, 2017). We hope to incorporate these ideas, and tighten the connection to developmental data, in future work.

References

Atkinson, J. (2002). The developing visual brain.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2017). An egocentric perspective on active vision and visual object learning in toddlers. In *2017 icdl-epirob* (pp. 290–295).

Bourne, J. A., & Rosa, M. G. (2005). Hierarchical development of the primate visual cortex, as revealed by neurofilament immunoreactivity: early maturation of the middle temporal area (mt). *Cerebral cortex*, *16*(3), 405–414.

Braddick, O., & Atkinson, J. (2011). Development of human visual function. *Vision research*, *51*(13), 1588–1609.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

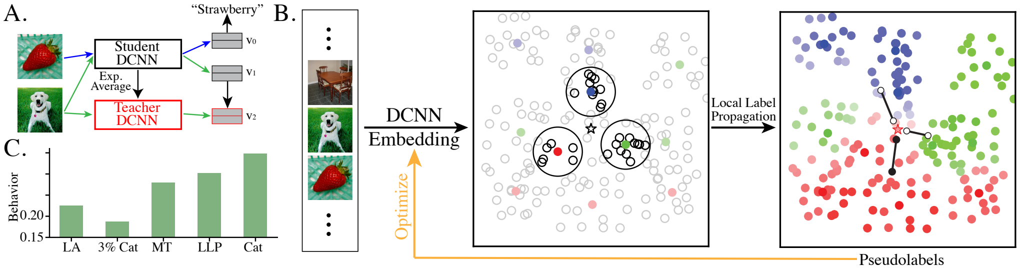

Figure 3: Semi-supervision and behavioral consistency. (**A**) Schematic for the Mean Teacher (MT) (Tarvainen & Valpola, 2017) algorithm. During training, the "student DCNN" is optimized to correctly categorize the labeled images and to produce similar predictions to another "teacher DCNN" whose weights are exponential averages of the weights of the student DCNN. (**B**) Schematic for Local Label Propagation (LLP) (Zhuang, Ding, et al., 2019) algorithm. All images are first embedded into a lower dimensional space (middle panel, colored points represent labeled images and unfilled point represent unlabeled images). For each unlabeled image (⋆ in the middle panel), a group of labeled images nearby in embedding space is identified (highlighted colored points) and a pseudolabel is inferred by weighting the labels of these nearby images according to their distances to ⋆ as well as their local neighbor densities (highlighted unfilled points around these colored points). The inferred pseudolabels are used to optimize the DCNN. (**C**) behavioral consistency of DCNNs trained by different tasks. "3% Cat" represents a ResNet-18 trained on ImageNet with only 3% of images labeled. The standard deviation of these measures across multiple runs is typically 0.001.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Eccv* (pp. 132–149).

Cooper, R. P., & Aslin, R. N. (1989). The language environment of the young infant: Implications for early perceptual development. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *43*(2), 247.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database.

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Iccv* (pp. 1422–1430).

Harwerth, R. S., Smith, E. L., Duncan, G. C., Crawford, M., & Von Noorden, G. K. (1986). Multiple sensitive periods in the development of the primate visual system. *Science*, *232*(4747), 235–238.

Klindt, D., Ecker, A. S., Euler, T., & Bethge, M. (2017). Neural system identification for large populations separating what and where. In *Advances in neural information processing systems* (pp. 3506–3516).

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *2016 fourth 3dv* (pp. 239–248).

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems* (pp. 1195–1204).

Topping, K., Dekhinet, R., & Zeedyk, S. (2013). Parent–infant interaction and childrens language development. *Educational Psychology*, *33*(4), 391–426.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Cvpr* (pp. 3733–3742).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *Eccv* (pp. 649–666).

Zhuang, C., Ding, X., Murli, D., & Yamins, D. (2019). Local label propagation for large-scale semi-supervised learning. *arXiv preprint arXiv:1905.11581*.

Zhuang, C., Zhai, A. L., & Yamins, D. (2019). Local aggregation for unsupervised learning of visual embeddings. *arXiv preprint arXiv:1903.12355*.