Learning only a handful of latent variables produces neural-aligned CNN models of the ventral stream

Learning only a handful of latent variables produces neural-aligned CNN models of the ventral stream

Yudi Xie, Esther Alter, Jeremy Schwartz, and James J. DiCarlo

Massachusetts Institute of Technology

February 29, 2024

Image-computable modeling of primate ventral stream visual processing has made great strides via brain-mapped versions of convolutional neural networks (CNNs) that are optimized on thousands of object categories (ImageNet), the performance of which strongly predicts CNNs' neural alignment. However, human and primate visual intelligence extends far beyond object categorization, encompassing a diverse range of tasks, such as estimating the latent variables of object position or pose in the image. The influence of task choice on neural alignment in CNNs, compared to CNN architecture, remains underexplored, partly due to the scarcity of large-scale datasets with rich known labels beyond categories. 3D graphic engines, capable of creating training images with detailed information on various latent variables, offer a solution. Here, we asked how the choice of visual tasks that are used to train CNNs (i.e., the set of latent variables to be estimated) affects their ventral stream neural alignment. We focused on the estimation of variables such as object position and pose, and we tested CNNs' neural alignment via the Brain-Score open science platform. We found some of these CNNs had neural alignment scores that were very close to those trained on ImageNet, even though their entire training experience has been on synthetic images. Additionally, we found training models on just a handful of latent variables achieved the same level of neural alignment as models trained on a much larger number of categories, suggesting that latent variable training is more efficient than category training in driving model-neural alignment. Moreover, we found that these models' neural alignment scores scale with the amount of synthetic data used during training, suggesting the potential of obtaining more aligned models with larger synthetic datasets. This study highlights the effectiveness of using synthetic datasets and latent variables in advancing image-computable models of the ventral visual stream.

Additional Details

Humans/primates have the behavioral capacity to infer object content beyond object category (aka "category orthogonal" latent variables). And electrophysiological evidence suggests that the ventral visual stream encodes rich category orthogonal information [1]. Both behavioral and neural evidence motivated us to investigate how the neural alignment of models is impacted when models are trained to estimate different sets of latent image variables, such as object position and pose. To investigate this question, we generated an ImageNet scale synthetic image dataset using ThreeDWorld (TDW) [2], a Unity-based 3D graphic engine (Fig. 1a). The entire image dataset contains 1.3 million images from 117 object categories made up of 587 specific object 3D mesh models (i.e. on average approximately 5 object instances per category). Each image in the dataset contains one object rendered with a

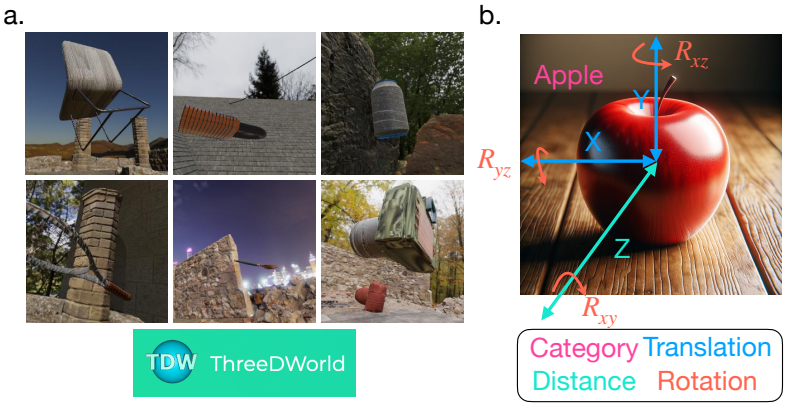

Figure 1: (a) Example images generated by TDW for training CNNs. Each image contains one object with varying position and pose against a randomly generated background. (b) An image containing one object. In addition to the object category (Apple), the set of latent variables we record are the following: translation (X, Y), distance (Z), and rotation (R_{xy}, R_{yz}, R_{xz}). (This image is for illustration only – it was not in the synthetic image set.)

a.

Number of outputs units	Training targets
1	Distance
2	Translation
3	Rotation or Distance + Translation
4	Distance + Rotation
5	Translation + Rotation
6	Distance + Translation + Rotation
117	Object category
587	Object identity
710	All cla. + all reg.
1000	ImageNet-1K

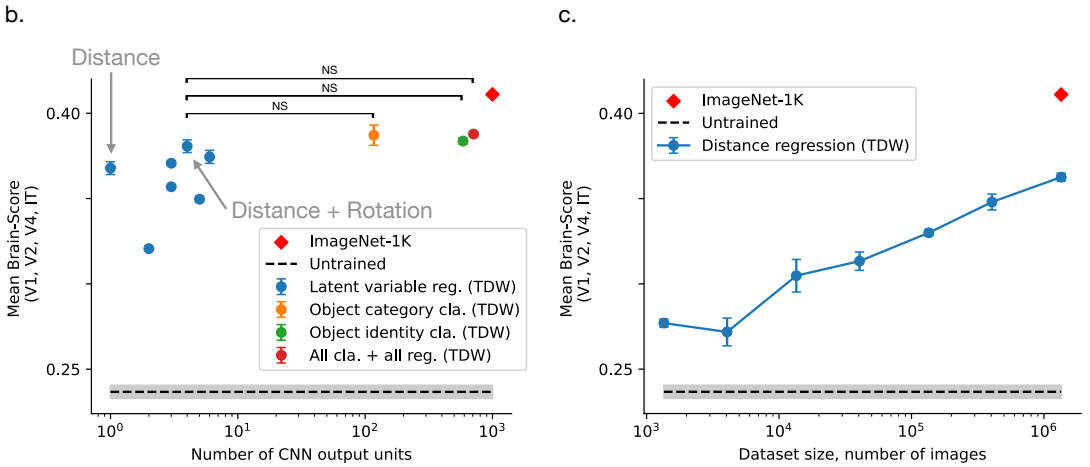

Figure 2: Neural alignment of a fixed CNN architecture trained on different sets of latent variables. (a) The latent variables use to optimize the CNN weights and the corresponding number of CNN output units that receive supervision. Distance, Translation, and Rotation are regression tasks. Object category and object identity are classification tasks. "All cla. + all reg." = all classification and regression tasks listed above. (b) The neural alignment score of models trained on different objectives (x-axis, see table (a)), compared with that of ImageNet-trained and untrained models. Error bars or shaded regions show the standard deviation across multiple random model initializations (N=3 per model). reg. = regression, cla. = classification. There is no significant score difference between the "Rotation + Distance" model and the "Object category classification" model (Mann-Whitney U test, p-value: 0.2), the "Object identity classification" model (p-value: 0.4), the "all cla. + all reg." model (p-value: 0.1). (c) Neural alignment of models trained on distance estimation (y-axis) as a function of the number of images in the image dataset (x-axis).

random position and pose on a random background. We record the ground truth information of a set of latent variables for each image (Fig. 1b); these include the x and y positions of the object, its distance to the camera, and the three Euler angles quantifying the object's rotation related to a pre-defined orientation of each object model. All these latent variables are in the camera reference frame. We also record each image's broader object category and the specific object model category in addition to the latent image variables.

All tests here were carried out using a ResNet-18 CNN architecture. Using supervised stochastic gradient descent, we optimized the weight parameters of the CNN to predict a specific subset of latent image variables (see below). We sort these models by their number of latent variables they were optimized to predict, which is equivalent to the number of output units (Fig. 2a); for example, a model trained to estimate object distance has only one output unit, while a model trained to perform object categorization has 117 output units. After training each model, we assigned the best predictive layers in the model to each of four ventral stream areas (V1, V2, V4, IT) and measured each model's neural alignment with all four areas using Brain-Score's public benchmarks [3]. This measures the ability of an assigned model layer representation to predict neural responses in the corresponding brain region on held-out images.

We found that models optimized only on the synthetic image dataset reached nearly the same level of neural alignment as the same CNN architecture trained on the ImageNet (1000 categories, photographs). For example, models optimized on all categories and category orthogonal latent variables in the synthetic dataset reached 94% of the neural alignment score (mean score 0.388, cf. mean score of 0.411 for ImageNet-1K training) (Fig. 2b).

We further investigated the neural alignment of CNNs trained to predict different sets of latent variables on the synthetic dataset. We found that many of the CNNs trained to estimate a very small number of latent variables can achieve surprisingly good neural alignment scores, comparable to models trained on more than one hundred object categories or more than five hundred object identity (Fig. 2b). The CNNs trained to estimate the distance and three rotation parameters achieved the highest score among these CNNs trained to estimate a small number of latent variables (Distance + Rotation, mean score 0.381). CNNs trained on estimating the distance of the object to the camera, which receives supervision from only one output unit, have surprisingly good results (Distance regression, mean score 0.368), achieving 95% of the neural alignment score of the CNNs receiving supervision on all classification tasks and the full set of latent variables (718 output units, All classification + all latent regression, mean score 0.388).

Finally, we investigated how the neural alignment of these latent trained CNNs scales with the size of the dataset used for training. We found that the neural alignment score of CNNs trained to estimate object distance

increases logarithmically with dataset size (Fig. 2c). This trend persists without an apparent plateau, even at the scale of image datasets akin to ImageNet. Given this scaling pattern, coupled with the capability of 3D graphics engines to generate arbitrarily large datasets, we anticipate that, by leveraging even larger synthetic datasets (e.g. 10 to 100 million images), we might surpass the neural alignment of ImageNet pre-trained models.

References

1. Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience* **19,** 613–622 (2016).
2. Gan, C. *et al.* Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954* (2020).
3. Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv,* 407007 (2018).