UNSUPERVISED NEURAL NETWORK MODELS OF THE VENTRAL VISUAL STREAM

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, Daniel L. K. Yamins

CS 431/PSYCH 250 – Winter 2021

Sharan Ramjee

SUPERVISED VS SEMI-SUPERVISED VS UNSUPERVISED

	Overview	Process	Subtypes	Examples
Supervised Learning	Majority of algorithms. Machine is trained using **well-labeled data**; inputs and outputs are matched.	Mapping function takes inputs and matches to outputs, creating a target function.	Classification, Regression	Linear regression, Random forest, SVM.
Unsupervised Learning	**Unlabeled data** (inputs only) is analyzed. Learning happens without supervision.	Inputs are used to create a model of the data.	Clustering, Association.	PCA, k-Means, Hierarchical clustering.
Semi supervised	**Some data is labeled**, some not. Goal: better results than labeled data alone. Good for real world data.	Combination of above processes.	All the above.	Self training, Mixture models, Semi-supervised SVM

https://www.datasciencecentral.com/profiles/blogs/supervised-learning-vs-unsupervised-in-one-picture

SUPERVISED METHODS ARE IMPLAUSIBLE AS A MODEL OF THE DEVELOPMENT OF THE VENTRAL STREAM

Deep Convolutional Neural Networks (DCNNs) have had success in approximating the adult primate visual ventral stream and have yielded the most quantitatively accurate predictive models of the image-envoked population responses in early (V1), intermediate (V2, V3, V4), and higher (IT) cortical areas.

The need for unsupervised methods:

- Supervised methods involve enormous amounts of semantic labels, while infants do not have access to millions of category labels during development.
- Although supervised methods are predictively accurate for adult cortical neural representations, they cannot provide a correct explanation of how such representations are learned in the first place.

CONTRASTIVE EMBEDDING OBJECTIVES

Contrastive embedding objectives are a family of algorithms designed for metric learning, where the goal is to measure similarity/distance between a pair of objects in order to embed inputs into a lower-dimensional compact space.

$$f : \mathbb{R}^{k \times k} \rightarrow S^n$$

The goal is to make the embedding f(x) "unique" – far away in the embedding space from other stimuli but close to different views of the original stimulus.

An example of a supervised contrastive embedding objective is the Contrastive Loss:

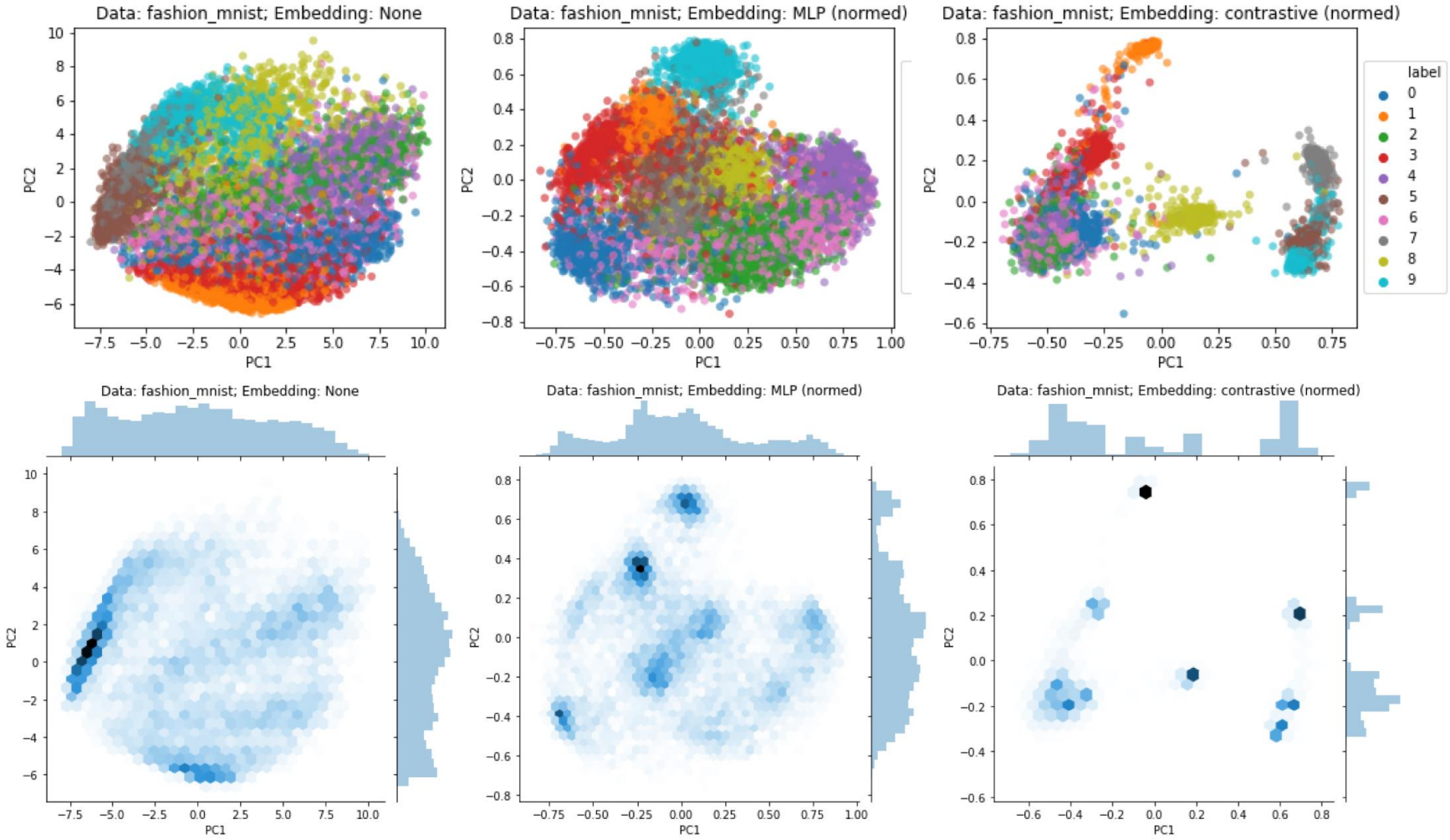
$$L_{contrast} = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\max(0, m - D_W)^2$$

Y is a binary indicator (Y = 0 if x_1 are x_2 are similar and Y = 1 otherwise)

$D_W(x_1, x_2)$ is a learnable distance function parameterized by weights W

m > 0 is the margin, which defines a radius around the embedding space

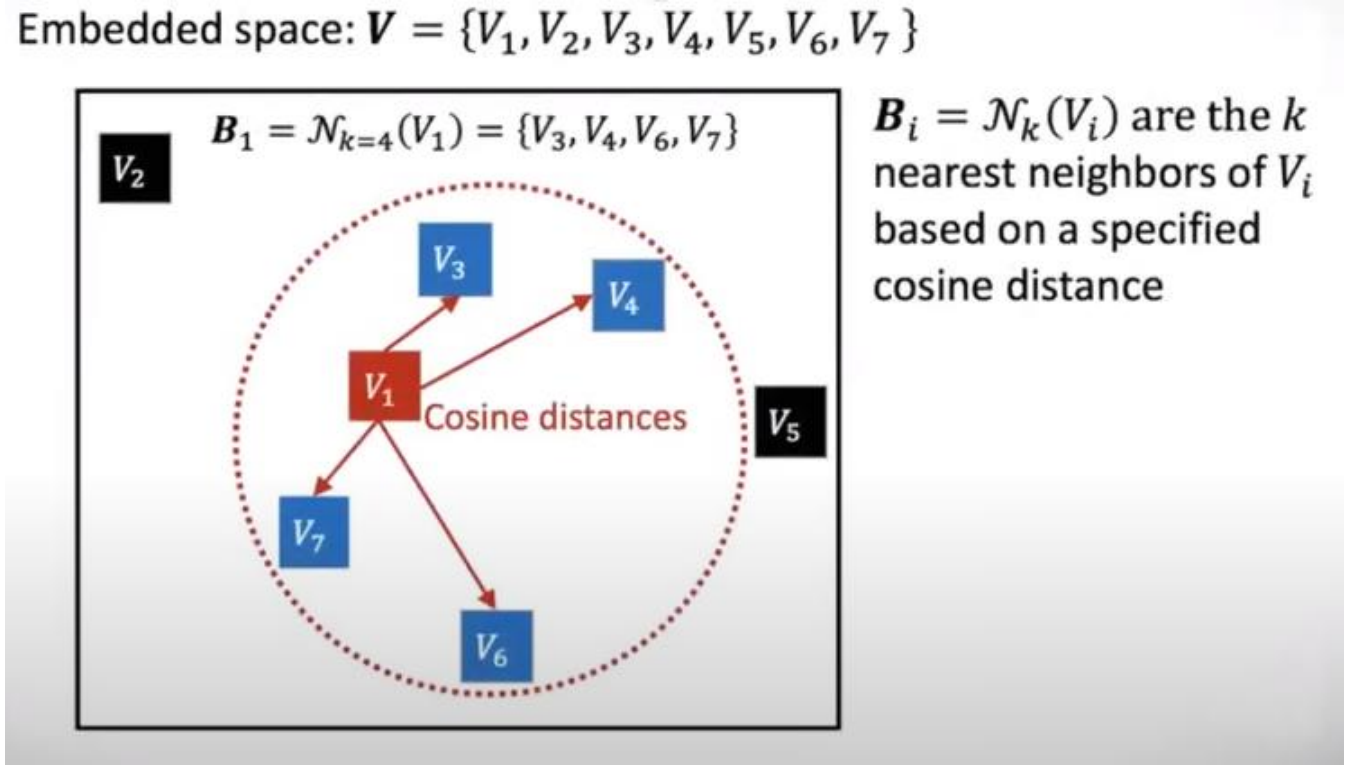
CONTRASTIVE LOSS FOR FASHION MNIST EMBEDDING



https://www.datasciencecentral.com/profiles/blogs/supervised-learning-vs-unsupervised-in-one-picture

UNSUPERVISED CONTRASTIVE EMBEDDING OBJECTIVES

Local Aggregation (LA) is an unsupervised contrastive embedding objective where uniqueness is encouraged by minimizing the distance (cosine similarity) to "close" embedding points and maximizing the distance to "further" points for each input.



$$L(\mathbf{C}_i, \mathbf{B}_i | \boldsymbol{\theta}, \mathbf{x}_i) = -\log \frac{P(\mathbf{C}_i \cap \mathbf{B}_i | \mathbf{v}_i)}{P(\mathbf{B}_i | \mathbf{v}_i)}$$

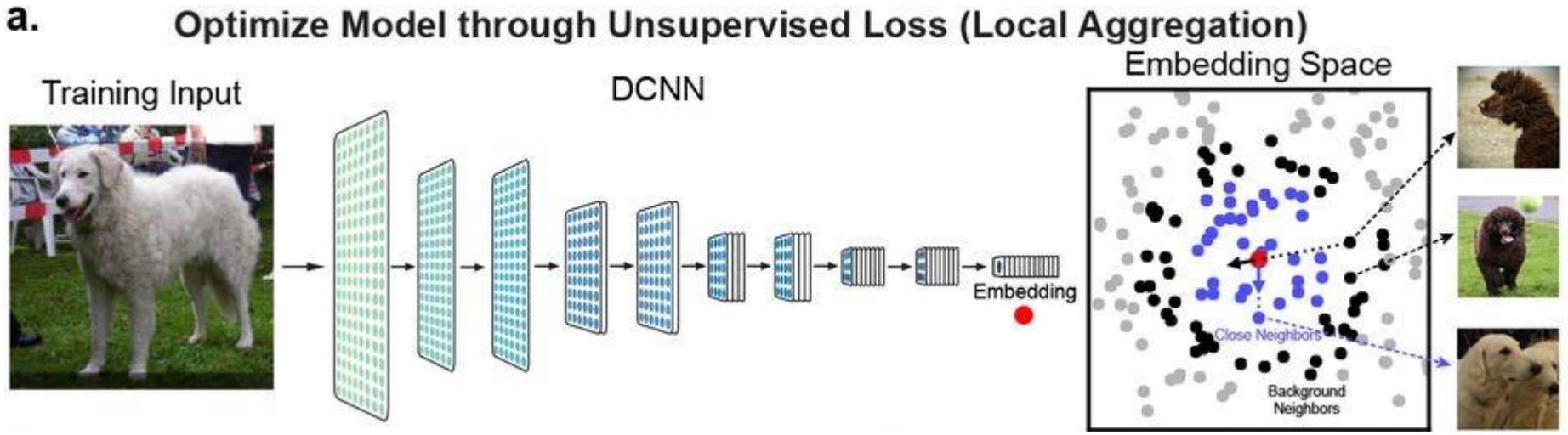
- $P(C_i)$ is the probability that the image is a closest neighbor.
- $P(B_i)$ is the probability that the image is a background neighbor.
- V_i is the image to be clustered.

The goal is to minimize the probability that the image is a closest neighbor (P(Ci))

https://arxiv.org/abs/1903.12355

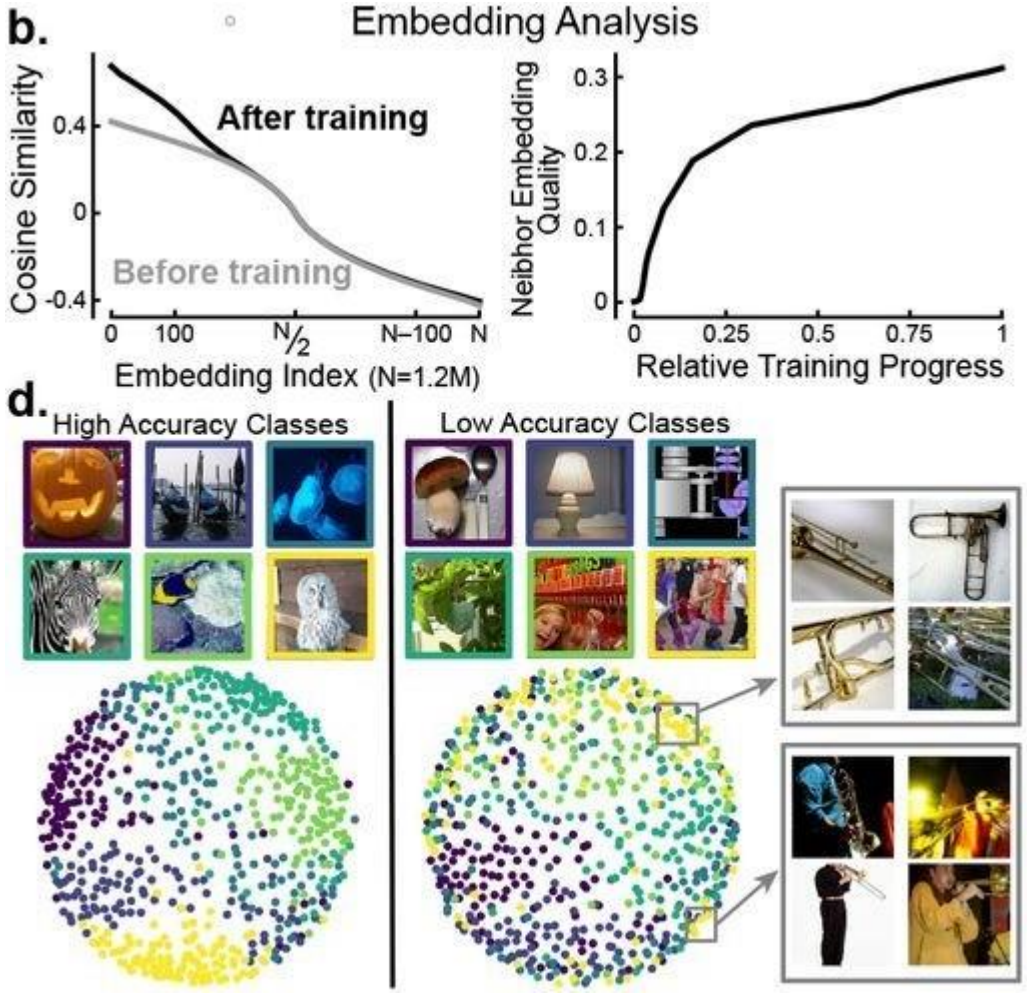
LA EXPLICITLY SEEKS TO CREATE FEATURES THAT **GENERICALLY** REFLECT *ANY* RELIABLE NATURAL STATISTIC DISTINGUISHING BETWEEN SETS OF INPUTS

LA embeds all the images into a lower dimensional space using a DCNN (ResNet-18), which is optimized to minimize the distance to "close" embedding points (blue) and to maximize the distance to "background" points (black).



https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

OPTIMIZATION ENCOURAGES LOCAL CLUSTERING IN THE EMBEDDING SPACE, WITHOUT AGGREGATING EVERYTHING



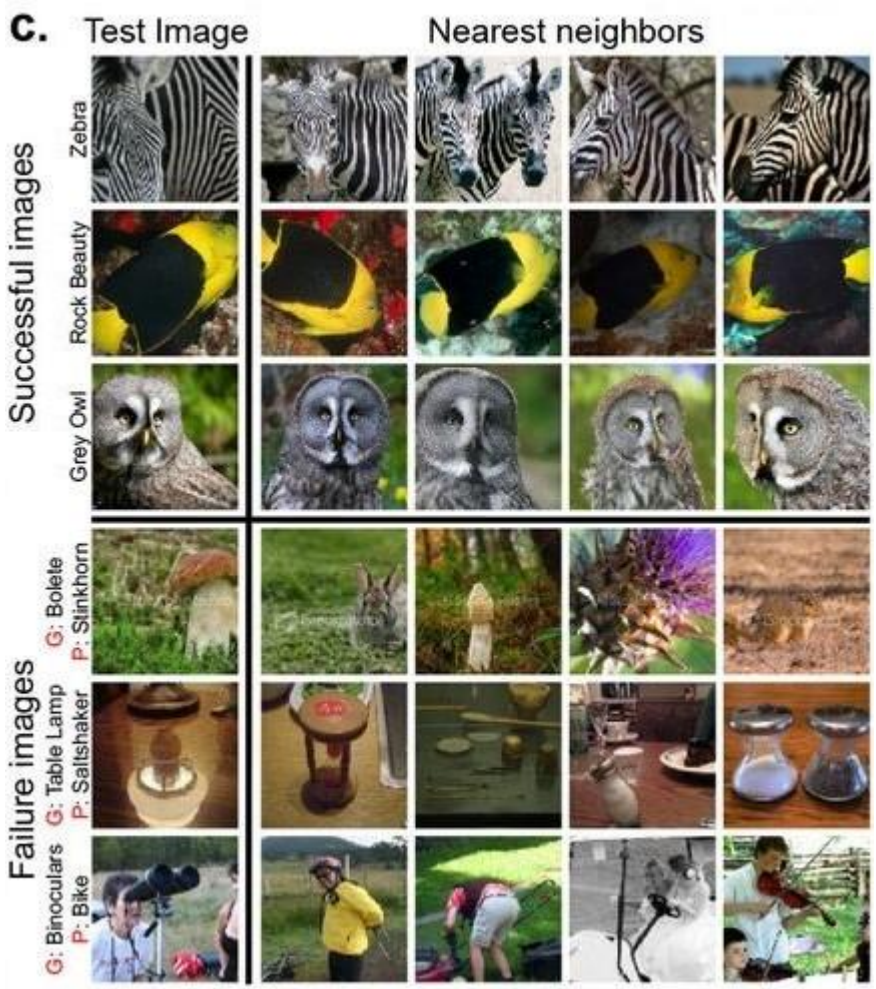
The average neighbor embedding "quality" increases as training progresses. Here, "quality" is defined as the fraction of 10 closest neighbors of the same ImageNet class label.

Multi-Dimensional Scaling (MDS) is used to visualize the LA embedding space.

For the low accuracy classes (right), we can observe the existence of two (or more) distinct clusters for the same class (trombone), which leads to low accuracies on downstream tasks.

https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

SUCCESS AND FAILURE CASES OF LA



https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

A weighted K-Nearest-Neighbor (KNN) classifier in the embedding space (K = 100) is used to classify the images.

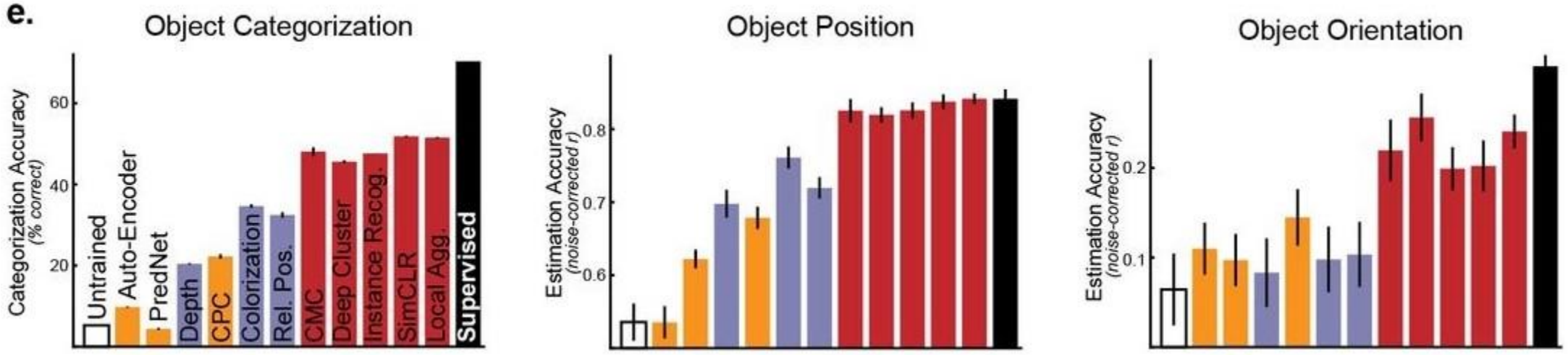
Even when uniform distance in the unsupervised embedding does not align with ImageNet class, nearby images in the embedding are nonetheless related in semantically meaningful ways.

UNSUPERVISED NETWORKS ACHIEVE GENERAL IMPROVEMENT IN THE QUALITY OF THE VISUAL REPRESENTATION

Once the embeddings are obtained using unsupervised contrastive embedding methods (LA, in this case), a supervised linear readout can be used to assess transfer performance.

Red – Contrastive embedding tasks
Orange – Predictive coding methods and Auto-Encoder
Black – Model supervised on ImageNet category labels

Blue – Self-supervised tasks
White – Untrained model



https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

RECENT UNSUPERVISED MODELS CAPTURE NEURAL RESPONSES THROUGH-OUT VENTRAL VISUAL CORTEX

A regularized linear regression model is fit from network activations of each unsupervised model to neural responses collection from array electrophysiology experiments in the macaque ventral visual pathway.

For the neural response no of one layer whose output shape is [sx, sy, c], a spatial mask ms of shape [Sx, Sy] and a channel mask mc of shape [c] are fit for each neuron to predict its response r:

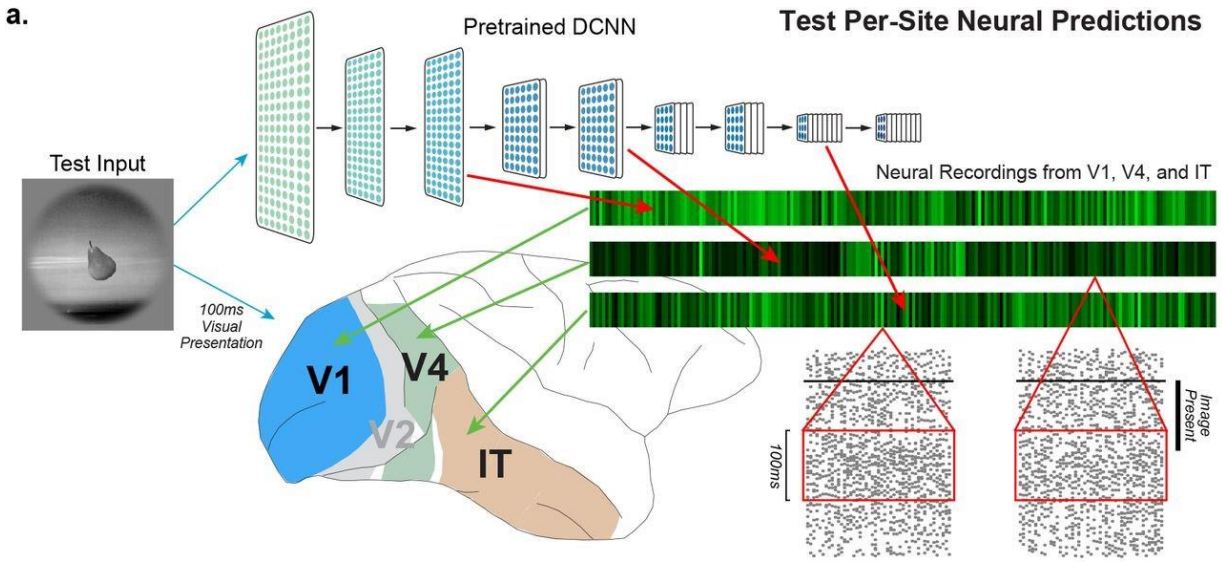
$$\hat{r} = \sum_{i=1}^{s_x} \sum_{j=1}^{s_y} \sum_{k=1}^{c} m_s[i,j] m_c[k] n_o[i,j,k]$$

The optimized loss is:

$$L = (\hat{r} - r)^2 + w(\|m_s\|_2^2 + \|m_c\|_2^2)$$

DEEP CONTRASTIVE EMBEDDING MODELS EVIDENCE THE MODEL-LAYER-TO-BRAIN CORRESPONDENCE THROUGH THE COMPARISON OF LAYER REPRESENTATIONS WITH NEURAL RESPONSES

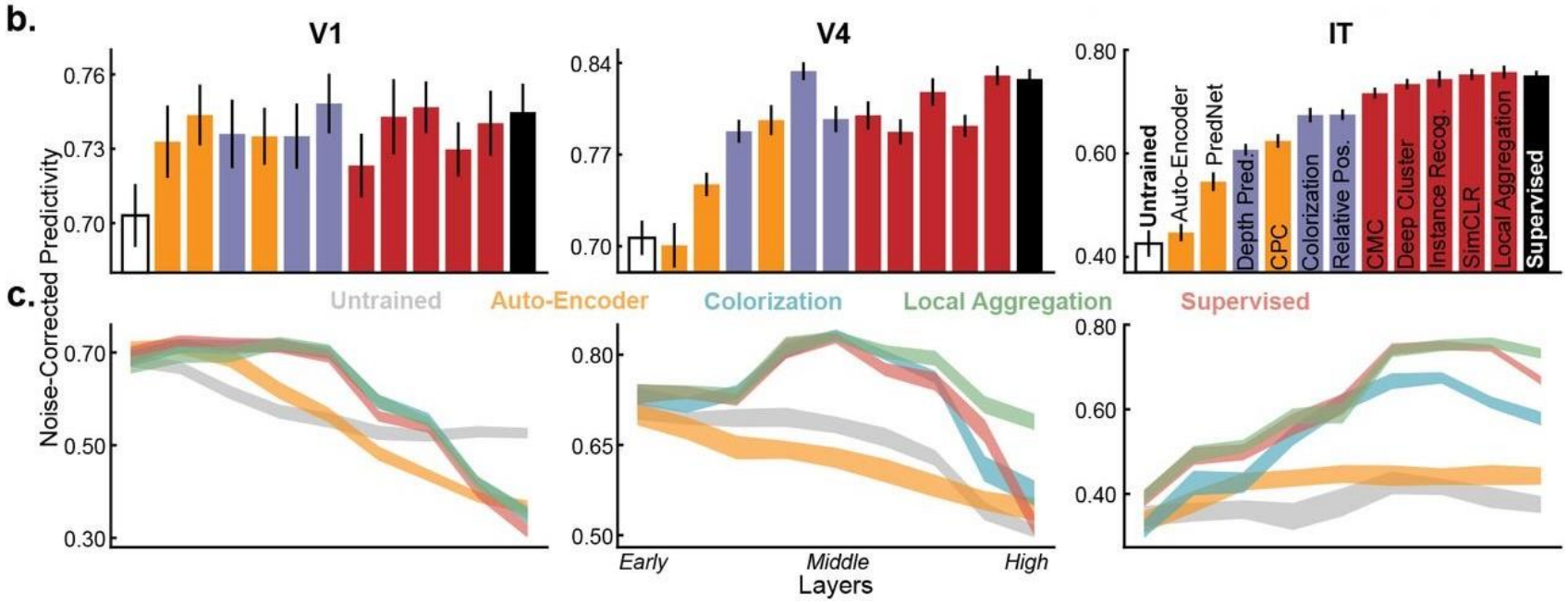
Network (trained using unsupervised objectives and run on stimuli with known neural responses) unit activations from each convolutional layer were used to predict V1 (early-layers), V4 (mid-layers), and IT (higher-layers) neural responses using regularized linear regression.



https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

THE LOCAL AGGREGATION (LA) MODEL ACHIEVES ACCURATE NEURAL PREDICTIVITY

For each neuron, the Pearson correlation between the predicted responses and the recorded responses was computed and then corrected by the noise ceiling of that neuron. The median of the noise-corrected correlations across neurons were then reported.



https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

Pearson correlation is a measure of linear correlation between two sets of data.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

DEEP CONTRASTIVE LEARNING CAN LEVERAGE NOISY REAL-WORLD VIDEO DATASTREAMS

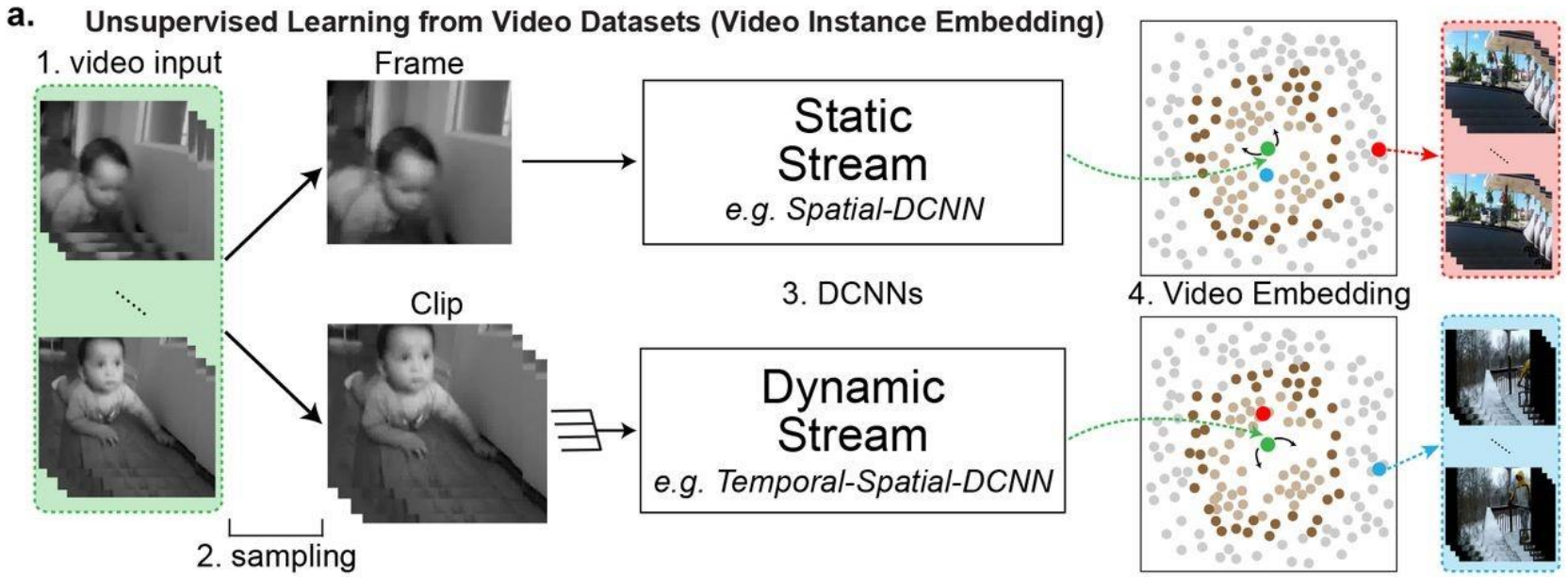
The underlying dataset (ImageNet) used to train deep contrastive embedding models diverge significantly from real biological datastreams:

- Infants receive images from a much smaller set of object instances than ImageNet that are viewed under much noisier conditions.
- Infants receive continuous stream of temporally correlated inputs whereas ImageNet consists of statistically independent static frames.

The authors introduce a novel extension of LA, the Video Instance Embedding (VIE) algorithm that learns representations on the SAYCam dataset that are highly robust and approach neural predictivity of those trained on ImageNet.

THE TEMPORALLY-AWARE VIE-TRAINED REPRESENTATION WAS BETTER THAN A PURELY STATIC NETWORK TRAINED WITH LA ON SAYCAM

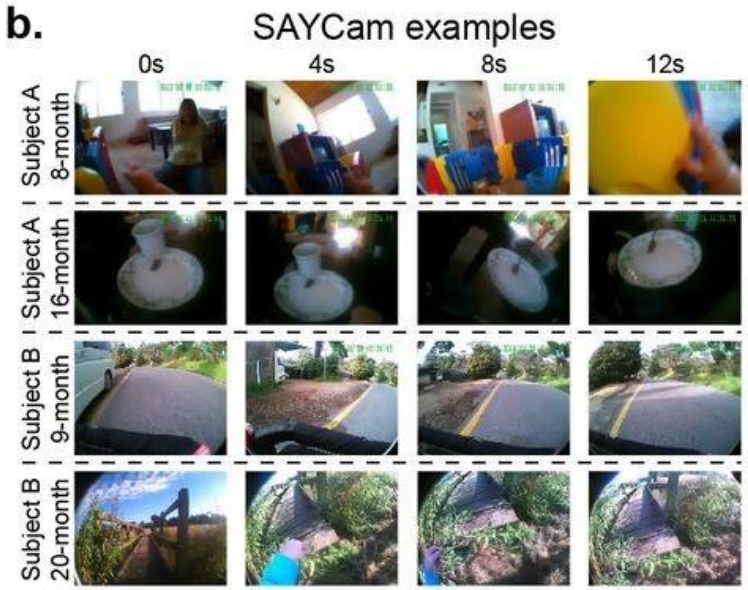
Frames were sampled into sequences of varying lengths and temporal densities. They were then embedded into lower-dimensional space using a static pathway (ResNet-18) for single images and a dynamic pathway (3D-ResNet-18) for multi-images (16 consecutive frames).



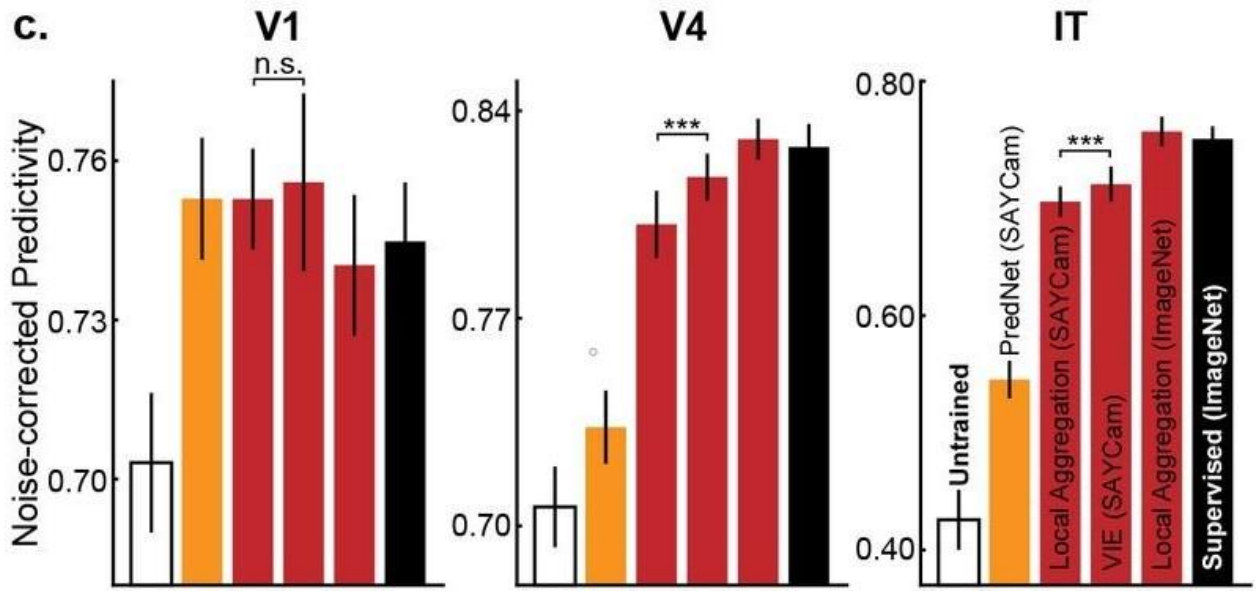
https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

These pathways were optimized to aggregate the resulting embeddings and their "close" neighbors (light brown) and to separate the resulting embeddings and their "background" neighbors (dark brown).

DEEP SPATIOTEMPORAL CONTRASTIVE LEARNING CAN TAKE ADVANTAGE OF NOISY AND LIMITED NATURAL DATASTREAMS TO ACHIEVE PRIMATE-LEVEL REPRESENTATION LEARNING



https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

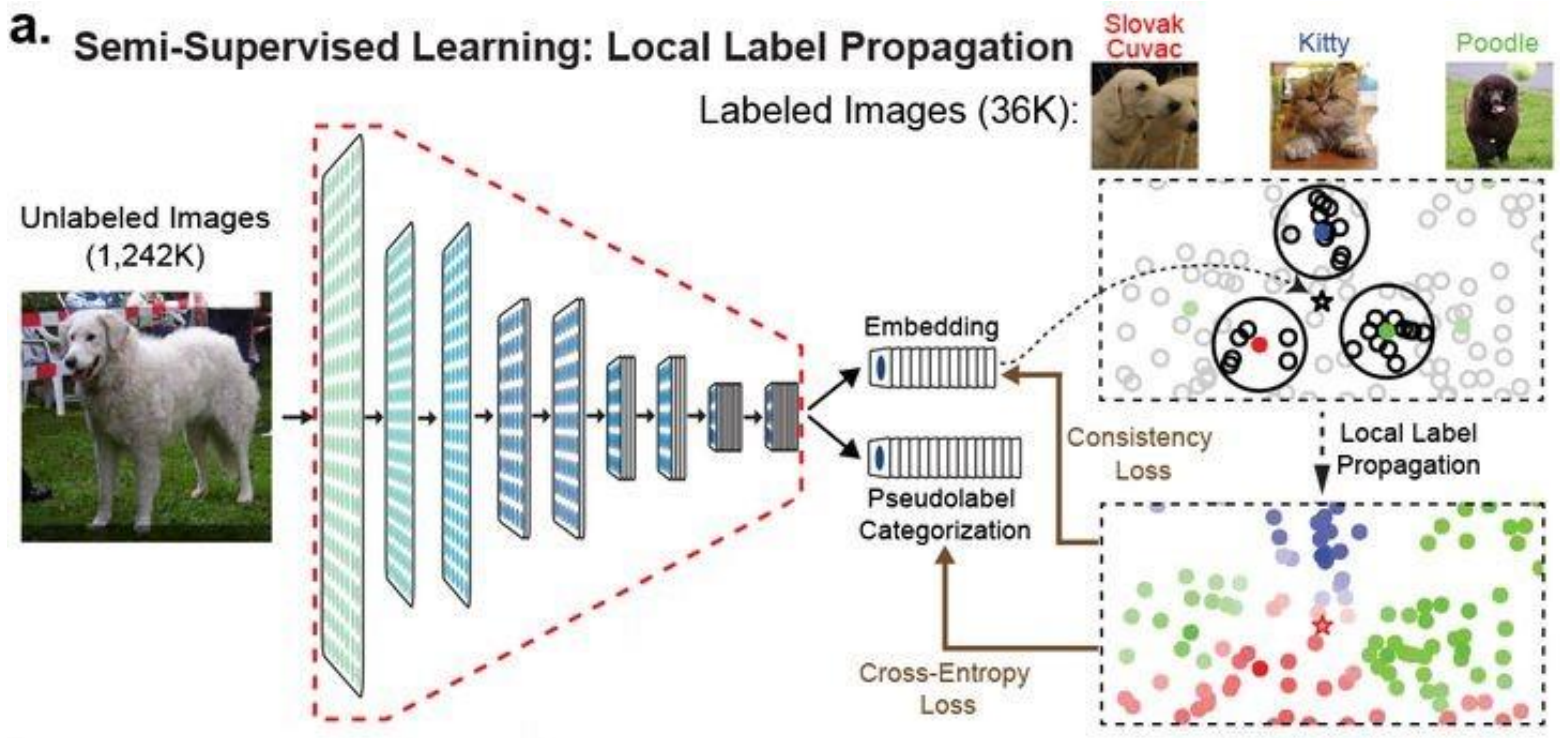


SAYCam dataset contains head-mounted video camera data from three children (collected 2 hours/week spanning ages 6-32 months).

A small but statistically significant gap between the SAYCam-trained and ImageNet-trained networks remains, possibly due either to limitations in the dataset or in VIE itself.

PARTIAL SUPERVISION IMPROVES BEHAVIORAL CONSISTENCY

While infants do not receive a lot of semantic labels during development, they do receive some labels, either from parental instruction or from environmental rewards.

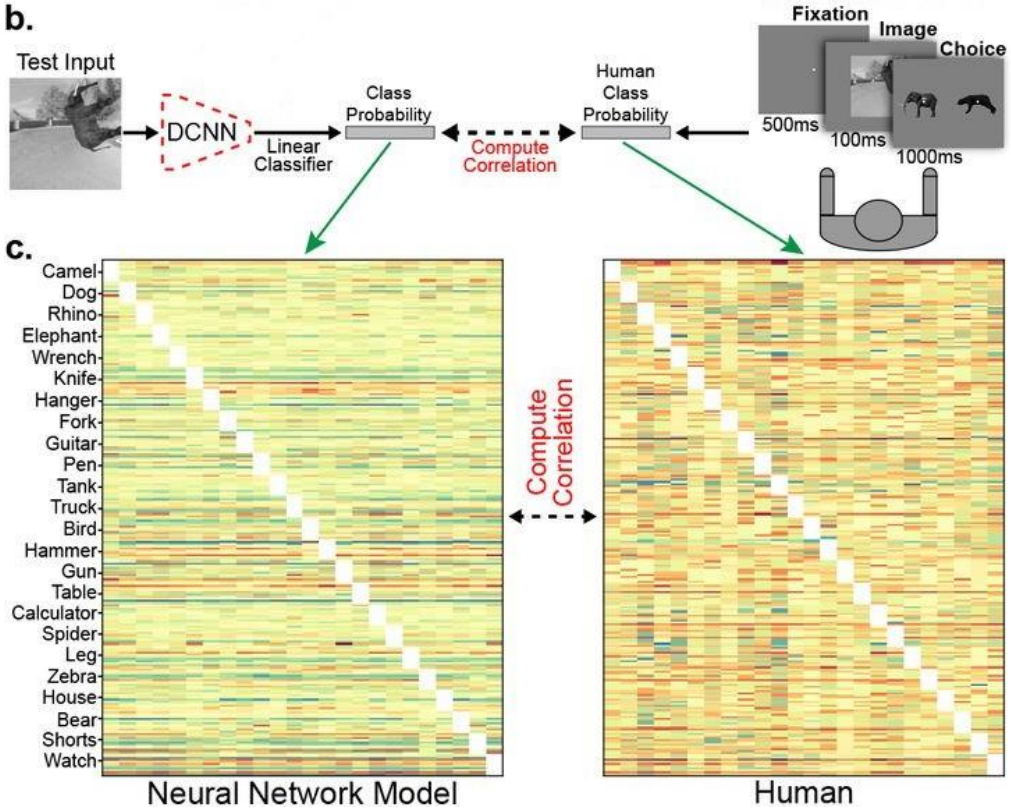


Local Label Propagation (LLP) is a semi-supervised learning algorithm that builds on contrastive embedding methods. Here, the weighted (by distance and density) labels from labelled neighboring embeddings (colored points) are propagated to generate pseudo-labels for unlabeled embeddings (star point).

https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

THAT SEMI-SUPERVISED LEARNING METHODS CAPTURE FEATURES OF REAL VISUAL LEARNING THAT BUILDS ON, BUT GOES BEYOND, TASK-INDEPENDENT SELF-SUPERVISION

To measure behavioral consistency, linear classifiers were trained from each model's penultimate layer on a set of images from 24 classes.

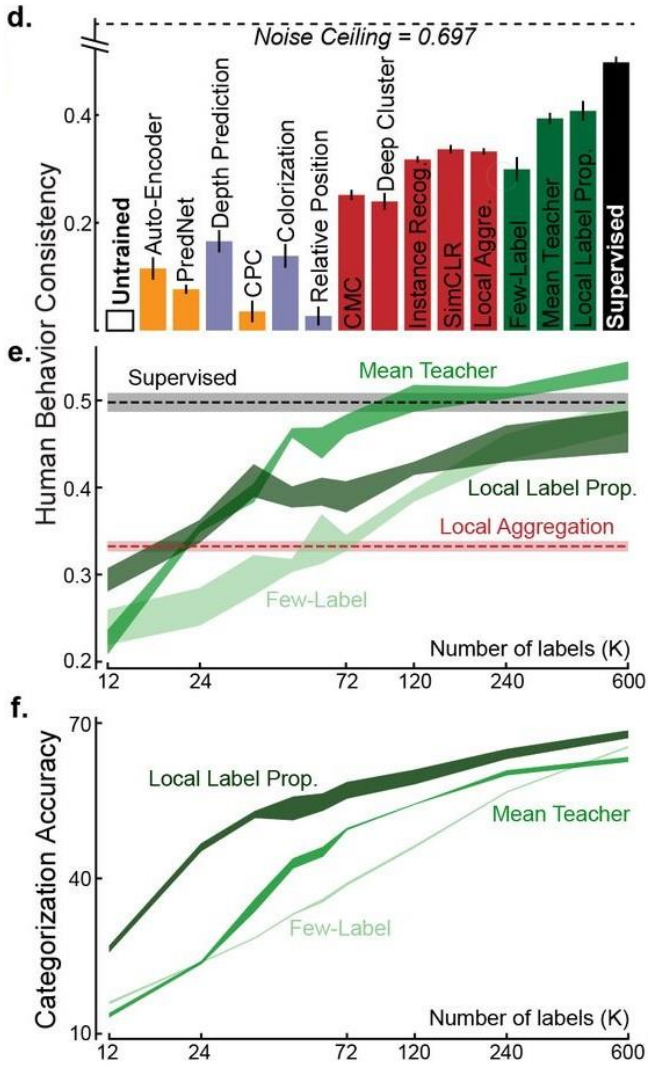


https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

The resulting image-by-category confusion matrix was compared to data from humans performing the same alternative forced choice task, where each trial started with a 500ms fixation point, presented the image for 100ms, and required the subject to choose from the true and another distractor category shown for 1000ms.

Finally, the Pearson correlation corrected by the noise ceiling was computed.

COMPARISON TO UNSUPERVISED AND SUPERVISED METHOD



Apart from Local Label Aggregation, Mean-Teacher (MT) and Few-Label (FL) semi-supervised methods were also implemented using a ResNet-18.

semi-supervised deep contrastive embeddings can leverage small numbers of labelled examples to produce representations with substantially improved error-pattern consistency to human behavior

Even with 36k labels (3% supervision) on ImageNet, both LLP and MT lead to representations that are more behaviorally consistent than purely unsupervised methods, though there is a gap to the supervised models.

https://www.biorxiv.org/content/10.1101/2020.06.16.155556v1.full

DRAWBACKS OF UNSUPERVISED METHODS IN PREDICTING AND EXPLAINING NEURAL RESPONSES IN VISUAL CORTICAL AREAS

- The same deep feed-forward neural network architecture is used, and this is insufficient to describe the response dynamics of real neurons.

- The standard backpropagation learning rule is used for optimization, and backpropagation has several features that make it unlikely to be implementable in real organisms.

- Although SAYCam is more realistic than ImageNet, it still lacks in utero retinal waves, the long period of decreased visual acuity, and non-visual modalities that are likely to strongly self-supervise visual representations during development.

SUMMARY

- Deep unsupervised contrastive embedding methods achieve neural prediction accuracy in multiple ventral visual cortical areas that equals or exceeds that of models derived using today's best supervised methods.
- The mapping of these neural network models' hidden layers is neuroanatomically consistent across the ventral stream.
- These methods produce brain-like representations even when trained on noisy and limited data measured from real children's developmental experience.
- Semi-supervised deep contrastive embeddings can leverage small numbers of labelled examples to produce representations with substantially improved error-pattern consistency to human behavior.