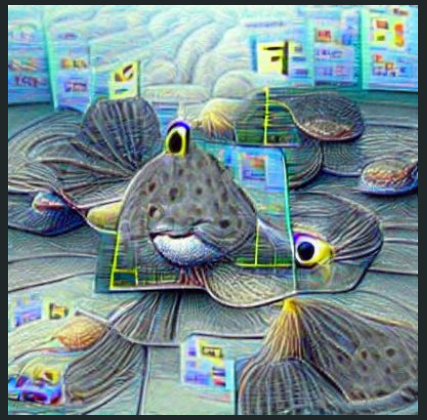


Are DNN representations always similar to cortical representations?

By: Kushaan Gupta, Sophia Lazarova, Arshed Nabeel, Harshit Pateria
TA: Mina Rezaie; Project TA: Rutvi Prajapati

neuromatch academy

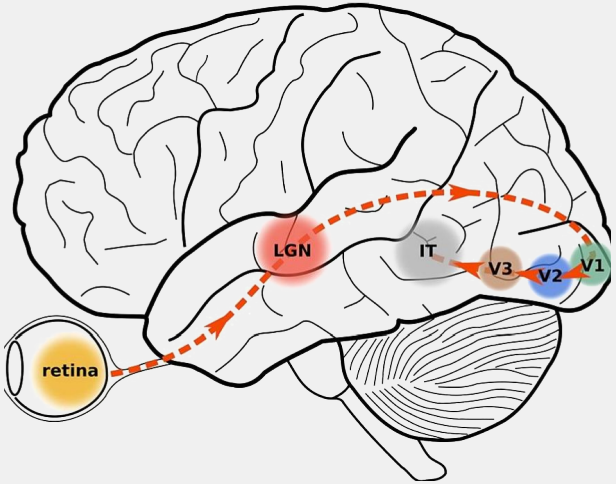
Team: Flounder Vision
Pod: Nice Flounders

Representations

Introduction

- Early layers in a deep convolutional neural network are shown to have representations similar to the early visual hierarchy of the brain. Similarly, later layers have representations more similar to the higher visual regions*.

- Studies like this usually use *deep neural networks* ***trained on a classification task (task-optimized neural networks)***.



* Brain-like functional specialization emerges spontaneously in deep neural networks. Katharina Dobs, Julio Martinez, Alexander J.E. Kell, Nancy Kanwisher. bioRxiv 2021.07.05.451192; DOI: 10.1101/2021.07.05.451192

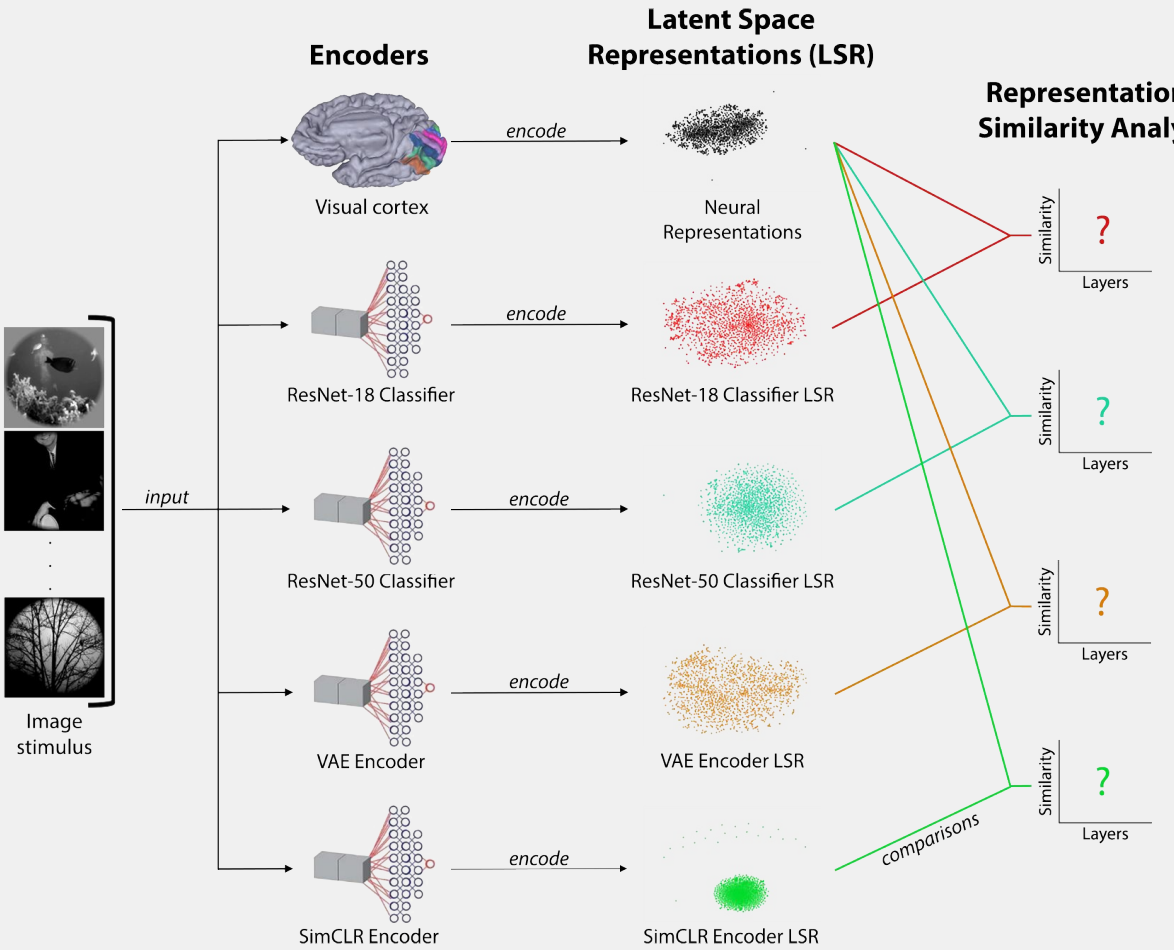
We are asking..

How robust are these findings:

- Across different network architectures/depths?
- For networks trained in a task-agnostic manner (autoencoders, self-supervised learning)?

Approach

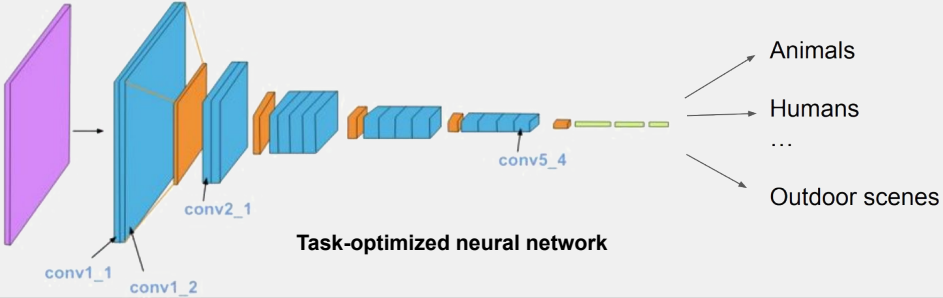
We did the comparison using **Representational Similarity Analysis (RSA):** A quantitative measure of representational similarity.

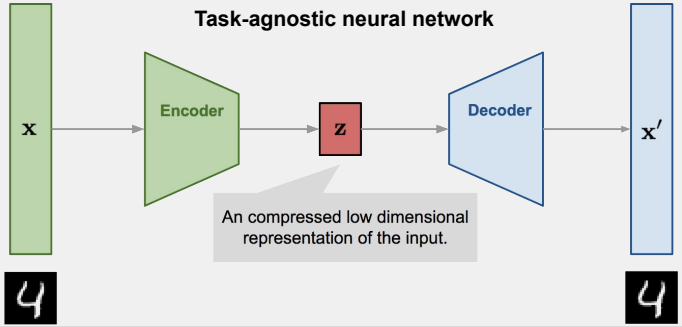


Approach

Compare the representations in the visual cortex (V1, Lateral Occipital) to representations in:

- **Task-optimized neural networks** of different depths: *ResNet-18, ResNet-50*.
- **Task-agnostic neural networks**:
  - A ***variational autoencoder*** (VAE) optimized to represent images using a latent representation.
  - A ***self-supervised neural network*** trained using a contrastive loss function (SimCLR).



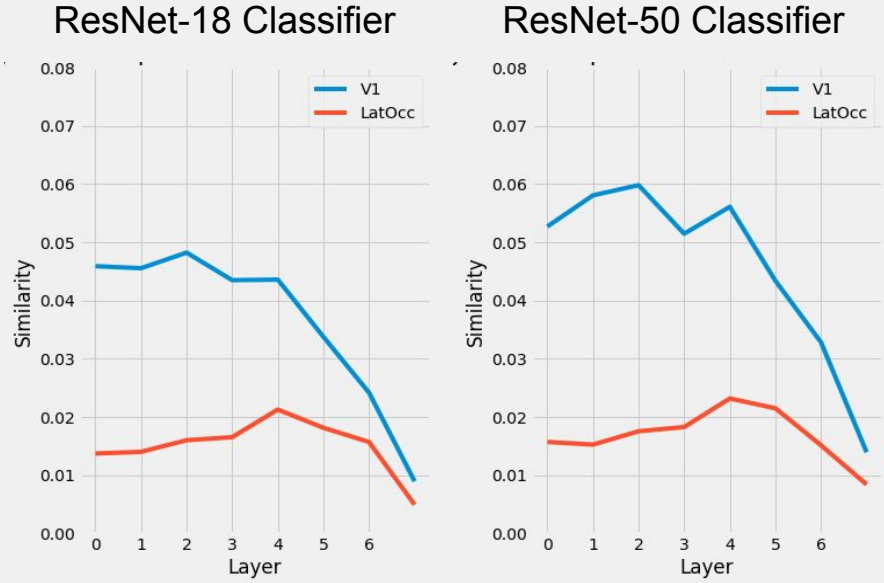


Results

Task-optimized (image classifier)

The trends in representational similarity does not depend on the architecture (depth) of the network.

- Lower visual regions (e.g. V1) show higher similarity with the earlier layers of the neural network.
- Higher visual regions (e.g. Lateral Occipital Region) show higher similarity with middle layers of the network.

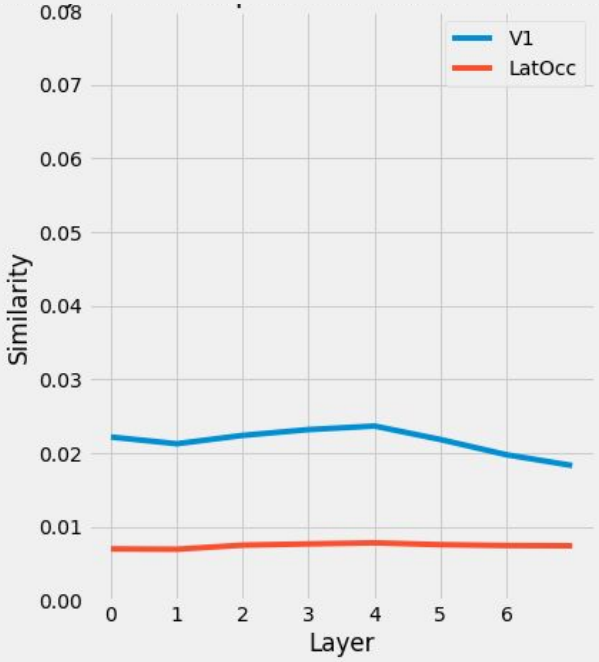


Results

Task-agnostic (generative models)

VAEs do not learn representations similar to the visual cortex.

- The similarity is low for both lower and higher visual regions.
- There are no trends in similarity across the network.



Results

Task-agnostic (self-supervised)

Self-supervised learning (SimCLR) results in representations similar to the visual cortex.

- The trends mirror those observed in task-optimized neural networks.

Conclusions

- Our results suggest that comparisons between DNNs and the brain need to take into account details like the task and loss-function on which the neural network was trained on.
- Generative models like VAEs (task-agnostic) represent image features in a manner dissimilar to how the human brain represents image features.
- Self-supervised learning learns invariant representations that are similar to the representations in the brain, despite being trained in a task-agnostic way.

Caveats/Future Work

- We use pre-trained models from multiple sources. Verify results with a unified dataset and training pipeline.
- Other ways to quantify representational similarity than RSA: *noise-corrected predictivity**
- Brain-inspired architectures and training: can we incorporate activations from brain regions into the training process?

* 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
Proceedings of the National Academy of Sciences Jan 2021, 118 (3) e2014196118; DOI: 10.1073/pnas.2014196118