Spring 5-4-2021

Translating Convolutional Neural Networks Approach to the Ventral Pathway

Victoria Zhang
Washington University in St. Louis

Washington University in St. Louis

McKelvey School of Engineering

Department of Computer Science

Thesis Examination Committee:

Dr. Chien-Ju Ho

Dr. Ulugbek Kamilov

Dr. Carlos Ponce

Translating Convolutional Neural Networks Approach to the Ventral Pathway

By

Victoria Zhang

A thesis presented to the McKelvey School of Engineering

of Washington University in St. Louis in partial fulfillment of the

requirements for the degree of

Master of Science

May 2021

St. Louis, Missouri

Dedication

I dedicate this thesis to my dear parents,

who support me in every way.

Acknowledgments

I would like to express many thanks to my principal investigator Dr. Carlos Ponce for leading me into the amazing world of computational neuroscience and guiding my growth from a student to a scholar. I also would like to offer my deep appreciation to my advisors Dr. Ulugbek Kamilov and Dr. Caitlin Kelleher, and great appreciation to my fellow labmates in Ponce lab, Binxu Wang, Mary Burkemper, Dr. James Johnson, Olivia Rose, and Katie Mueller, and our collaborator, Dr. Margaret Livingstone and Dr. Till Hartmann for offering great insights and collecting data. Finally, thank my parents Hairong Gao and Kaijun Zhang for their consistent support in my education. Special thanks to my boyfriend Ben Miao for his daily support. This work could not have been completed without the continuous support and encouragement coming from such a supportive community.

Victoria Zhang

Washington University in St. Louis

May 2021

Table of Contents

List of Figures

Abstract

Do artificial neurons in CNNs learn to represent the same visual information as the biological neurons in primate brains? Previous studies have shown that the visual recognition pathway (ventral stream) in humans and monkeys increasingly represents animate objects [16]. We used a heatmap attribution technique borrowed from convolutional neural networks to generate biological feature maps identifying regions in scenes that elicit responses from neurons along the ventral stream (V1/V2, V4, and IT). Biological feature maps were then compared to activation maps produced by units in convolutional neural networks. We found that image regions containing animals elicited increasingly larger responses along the ventral stream, while such animacy features are not represented in artificial neural networks.

Chapter 1. Introduction

The domains of biological vision and machine vision are two binary stars orbiting around the fundamental principles of visual recognition in natural environments. The convolutional neural networks (CNNs), historically inspired from the brain neural networks and now becoming the best approximation for the visual system, share similarities in many ways with the brain ventral visual pathway. Do the visual neural networks in primate brains learn and encode visual information the same way as CNNs? If not, how do they differ? To benefit applications in society such as healthcare, how can we further improve CNNs to make them more brain-like? Answering these questions would take us to the next level of primate visual system understanding, and these findings from biology would further guide the future development of brain-like machine vision.

Overview of the Ventral Visual Pathway

The ventral visual pathway, characterized as the "what" pathway, is a visual information processing stream in the brain originating in the primary visual cortex V1, going through extrastriate visual cortical areas V2, V4, and extending to the inferotemporal cortex (IT) [15]. V1 neurons encode information about local orientation, spatial frequency, and color (REF). V2 and V4 neurons are tuned moderately complex patterns. Inferotemporal cortex (IT) neurons organize their responses as a function of semantic category [16] or as a function of common category shape [2] and were first known to respond to complex visual objects like hands [6] and faces [5]. Converting streams of rich and complex visual information carried by light into neural signals,

the visual system allows us to recognize objects and scenes experienced in our daily lives. However, the principles of visual recognition in natural environments remain unknown. The classic majority of visual neuroscience studies have yielded enormous insights about the type of objects that neurons encode along the ventral pathway, but it has limitations. Some have relied on highly simplified stimuli like dots, lines, and colors, which are parameterizable but too different from natural images. Other investigations have relied on behavioral tasks that animals do not experience in the wild, such as discriminating the orientation of lines.

Overview of Convolutional Neural Networks

Convolutional neural networks (CNNs) are historically inspired by the view of the brain as a set of individual simple cells interacting to give rise to high-level processes [9]. Akin to visual areas in the brain, a CNN is usually formed by sets of functions hierarchically arranged as layers, with each layer performing a simple and biologically plausible operation. Image data, represented as matrices of RGB color channels, are fed into CNNs as the input. Features in the image are then discovered by convolving filters (artificial neurons) systematically across the entire image through a convolution-rectification (Conv-ReLu) layer. The dimension of these features is reduced by choosing the maximum values among the feature patches through a max-pooling layer. Iterations of computations in these convolution, rectification, and pooling layers make it possible for CNNs to extract information from high-dimensional image data without losing any features and further use them to make good predictions. Because of the similarities in architecture with the biological visual networks and the outstanding performance in image

recognition tasks, CNNs have become the state-of-the-art models of the visual system in the past decade [19].

Translation Between CNNs and the Ventral Visual Pathway

The hierarchically arranged Conv-ReLu, Max-Pooling layers schematic in CNNs characterize the cluster of cortical areas along the ventral visual pathway by design, making them ideal models for us to study and compare. How CNNs code and decode feature information of a natural image gives us insights to translate the geometry to visualize the representation of neurons along the ventral visual pathway. Conversely, the implications from electrophysiological activities along the ventral visual pathway can be used to validate if CNNs replicate the representation processing in the visual system.

To find what types of objects that neurons encode, previous study used genetic algorithms to allow neurons to "build" their preferred stimuli [26], synthetic novel objects [20] or the systematic removal of object parts to identify key shape primitives [13, 14]. It has shown that the ventral pathway in humans and monkeys increasingly responds to animate objects [16]. In this study, we used a complementary electrophysiological design to identify the preferences of neurons on the ventral pathway. In particular, we used various approaches from machine learning, including one called *attribution* [22], borrowed from CNNs. Attribution is a technique that highlights the parts of an image that best activate a hidden unit. We used it to identify the parts of an image that cause a neuron to respond, signaling the presence of features that the neuron has learned to encode. In CNNs, convolving different locations in an image with the

same filter is called *weight sharing* and is an efficient way to minimize the number of model parameters. The consequent output convolution is called a *feature map*, which has the highest values at locations with patterns corresponding to that of the filter.

To translate this approach to the brain, we reversed the geometry: we replicated the weight sharing operation using actual receptive fields (RFs) from multiunit populations in V1 (five monkeys), V4 (two monkeys), posterior IT (two monkeys), and central IT (one monkey). Instead of moving the filter around the image, we had the animals perform a fixation task while we moved the picture relative to the stationary RFs. We used 36 large scene photographs containing a variety of natural objects and textures. The resulting biological feature maps revealed the neurons' preferred shapes in the scenes, automatically segmenting the specific parts they were most activated by, with minimal investigator bias and zero stimulus pre-processing (see Methods).

Our results show that our feature map approach is able to localize key objects in the natural images. We interpreted these maps using prediction feature maps generated by various algorithms: (1) semantic hypotheses and (2) convolutional neural networks (CNNs). Finally, this approach can validate if CNNs, as a class of individual units, play the same role as neurons in the primate brain ventral pathway in learning and encoding visual information in the natural environment.

Chapter 2. Methods

Animals and Behavior

Eight monkeys from Washington University School of Medicine and Harvard School of Medicine implanted with chronic microelectrode arrays were used in this experiment. Two monkeys were implanted with 64-channels in three chronically implanted electrode arrays along the ventral stream (V1/V2, V4, IT). Three monkeys were implanted with floating arrays (Microprobes for Life Sciences, Gaithsburg FL) in the IT cortex. Arrays PIT1 and PIT2 were implanted in two different animals in the right-hemisphere posterior inferotemporal cortex (PIT, immediately anterior to the inferior occipital sulcus, representing the left visual field). We recorded neuronal activity in a third animal from the right-hemisphere central inferotemporal cortex, above the posterior middle temporal sulcus (array CIT). Three monkeys were implanted with Utah arrays (Blackrock Microsystems) in the right hemisphere operculum (arrays V11-V13). The task for all monkeys was to maintain their gaze on a red fixation target (<0.2° radius) while large pictures were flashed at different locations on the screen. If the animals maintained their gaze within a 2.2° radius of the fixation target (1° for V1-array monkeys), they received a juice reward.

Receptive Field Mapping

Each recording site's receptive field was mapped by moving a picture of a grayscale cartoon face (Fig. 1d) in a 16°x16° grid at different locations relative to the fixation point with 2° spacing. The picture was flashed on for 100 ms and left off for 200 ms. The picture evoked firing rate responses from each array channel; in V1, responses were defined as the mean firing rate over 50-200 ms after picture onset minus the mean firing rate over the first 30 ms after picture onset; in IT, responses were defined as the mean rate over 70-190 ms minus the first 30 ms. The subsequent 9 x 9 response matrices from every array channel were interpolated over 250 x 250 points. Each interpolation was used to compute a receptive field, defined by regions where the mean firing rate exceeded the 99^{th} percentile of all responses in the map. The largest region exceeding this threshold was identified using the Matlab function `regionprops.m`, and its centroid location was used as the channel's RF location. We defined the population RF as the mean array response.

To characterize the reliability of individual RF estimate, we tested whether the site responded differentially to each position using a one-way ANOVA (picture position as level, $P < 0.05$ after false discovery rate correction). If the channel was reliable, we used the channel-specific RF estimate. Otherwise, we assigned it the population estimate. Since the RF center of multiunit sites in each array channel does not always overlap with the center of the heatmap, we averaged the RF location per channel, transformed it into a 2D vector in Cartesian space, and used it as a translation vector to shift the heatmap using the Matlab function `imtranslate.m`.

Stimuli

The pictures were 16° x 16° natural scenes photographs embedded in a 30° x 30° brown-noise image. Each photo included natural (animals in the wild) and artificial settings (humans in laboratory settings) as well as computer-generated textures (bar fields). The pictures were presented in a 9° x 9° grid (spacing of 2°) with the center position located at the center of the population receptive field. The pictures were shown for 100-ms ON and 100 to 200-ms OFF, 3-14 flashes between juice rewards. Different pictures were interleaved within a trial. Every picture was presented at every position 30-37 times (V1 experiments) or 11-13 times (IT experiments). Each image was later segmented independently using an automated semantic image segmentation model (DeepLab with ResNet-101 backbone [21], trained on COCO-Stuff 164k dataset [7]) and a human-based segmentation task (Amazon Mechanical Turk, MTurk). All natural-scene photos were obtained from www.pexels.com and were free of copyright restrictions (one picture was provided by the Biomedical Primate Research Centre, Rijswijk, The Netherlands).

Neural Feature Maps

We convolved large photographs of natural scenes with receptive fields (RFs) from microelectrode array sites in V1/V2, V4, and IT (Fig. 1h). Convolution was done by randomly flashing each natural scene image at different positions relative to the center of the population RF, while the monkeys performed a fixation task (Fig.1a-c). Observed feature maps were matrices of spike rates. Each spike rate was the response of an individual channel to each position in the 16° x 16° natural image (e.g., the top left matrix entry corresponds to the activity

evoked when the top left part of the image fell on the receptive field). The evoked response was defined as the mean firing rate in 10-ms windows after image onset (1-10, 11-20, …,191-200 ms) minus the mean firing rate during the first 20 ms after image onset. There were 81 different positions (nine horizontal and nine vertical) (Fig. 1b), and thus the raw data were 9 x 9 x 20 tensors per multiunit. For time-averaged feature maps, we averaged each 9 x 9 x 20 tensor over the third dimension. To define a null hypothesis, for every time-averaged feature map, we scrambled the position information for every picture 999 times while keeping the array responses the same across channels and across time.

The convolutional output comprised the evoked firing rates arranged in a 2-D matrix, where each row and column corresponded to the part of the image that was inside the population RF (Fig. 1d, e). We called these 9 x 9 matrices biological feature maps in analogy to the convolutional outputs of filter weight sharing in convolutional neural networks. These biological feature maps could be resized to the original stimulus images (16° x 16°) and superimposed for visual inspection (Fig. 1f-g).

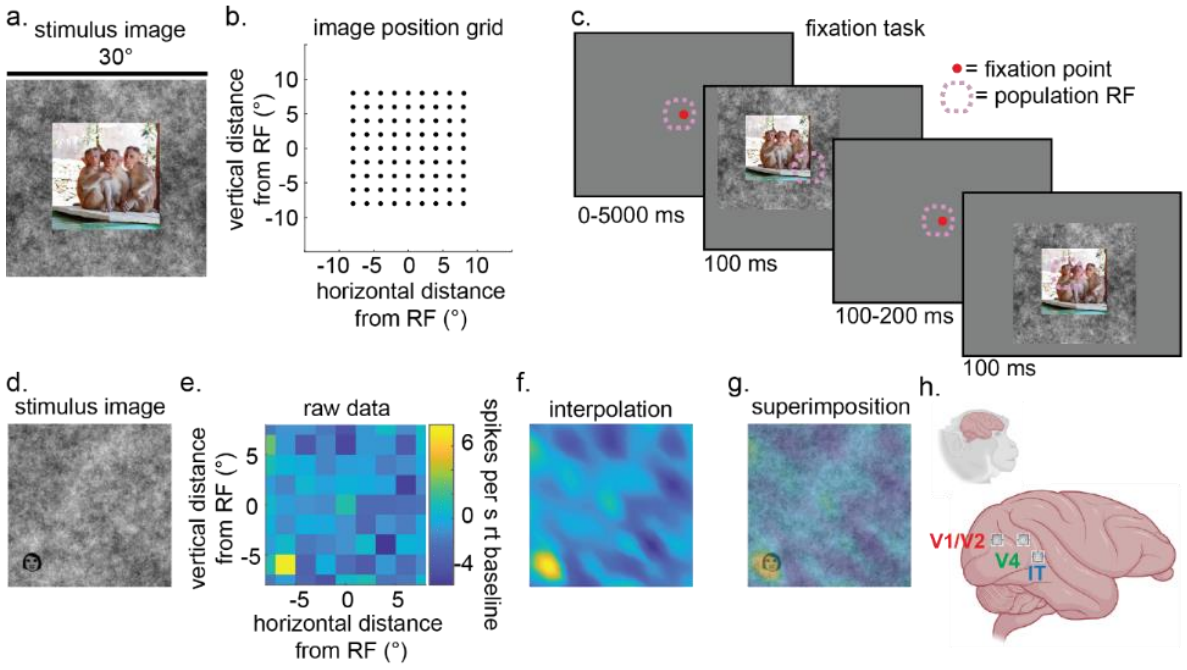

Figure 1. Experiment Setting

Prediction Space

Prediction feature maps were generated for every picture using deep convolutional neural networks (CNNs), and semantic segmentation approaching including automated (DeepLab [18]), and human-based (MTurk) semantic image segmentation models.

CNN Feature Predictions

CNN predictions were defined as the activation maps from every hidden unit in five convolutional layers of three deep convolutional networks: AlexNet [17], VGG-16 [10], and ResNet-50 [8]. There were selected because they are among the best performing CNNs [9], are well-known in the visual neuroscience literature, and are available in Matlab (`alexnet.m`, `vgg16.m`, and `resnet50.m`). AlexNet has five convolutional layers (‘conv1’ to ‘conv5’), and we used all hidden units in these layers. VGG-16’s architecture has 16 weight layers and uses sequential 2-3 convolution-rectification layer motifs, which allows for a deeper trainable network [24]. We used all units in the last convolutional layer of each motif: 'conv1_2', 'conv2_2', 'conv3_3', 'conv4_3', and 'conv5_3.' ResNet-50 has a 50-weight layer architecture that used parallel or "skip" pathways to transfer activations from early layers to deeper non-sequential layers [3]. The network is divided into five building blocks compressing repeating layer motifs, so we used on convolutional layer per building block: 'conv1', 'res2c_branch2c', 'res3d_branch2c', 'res4f_branch2c', and 'res5c_branch2c'. Every CNN hidden unit feature map was subsampled in an evenly spaced 9 x 9 grid and then normalized to the range of 0-1.

Semantic Map Segmentation

Each image used was segmented independently using both automated and human-based segmentation approaches. To get segmented images from DeepLab [18], we adapted the unofficial PyTorch implementation of DeepLab [21] by merging the ten animal labels and one person labels from COCO-Stuff 164k dataset [7] labels into a single animacy class and use it as a mask to segment the natural images.

In the human-based approaches, we first used Google Cloud Vision to detect and extract ten labels that can identify each natural image. We removed general labels such as “organism” as we a priori assumed these labels would not give us specific item segmented. We also replaced redundant and overlapping labels like “snout” with either “human,” “monkey,” or “animal,” and added, “eyes,” “nose,” “mouth/beak,” and “hand” accordingly. After revising the image labels, there were 3 - 6 labels per image, and we published five semantic segmentation tasks per image using these labels. Subjects on MTurk were asked to trace polygons around the border of revised labels for an image randomly shuffled from our 36 natural images. Combinations of the previous maps were used to create higher-order categories such as “face,” “animacy,” and “environment.” There were a total of 21 possible semantic maps per image (Fig. 2). Each map was downsampled to a 9 x 9 grid and then upsampled back to the original image size.

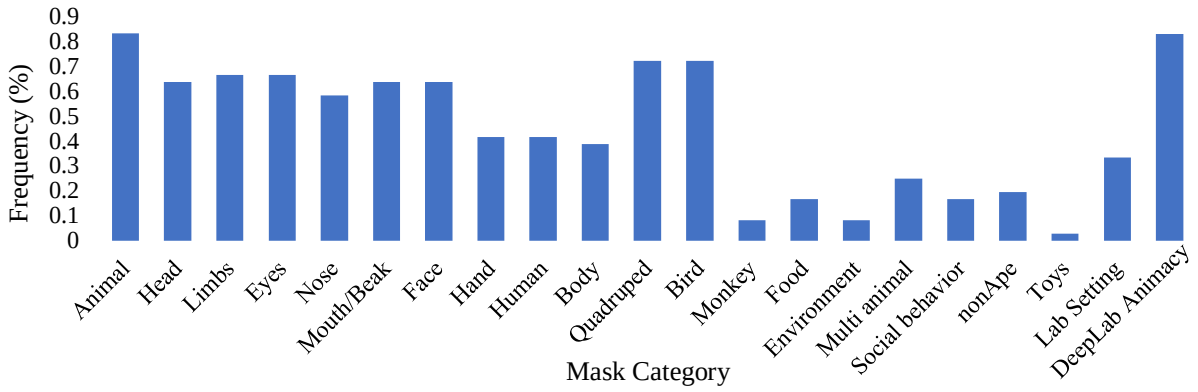

Figure 2. Semantic Mask Category

Comparing Neuronal Feature Maps to CNN Prediction Space

To quantify the similarity between observed maps and the prediction maps in CNNs, we randomly shuffled 32 CNN feature maps 200 times per picture and measured the Pearson correlation coefficient between each multiunit feature map and every CNN feature map. As cross-validation, responses to every picture flash were randomly assigned to one of the two sets: one dataset was used to identify the best prediction matched (per Pearson correlation), and the second data set to quantify the correlations associated with each paired match. This was done for every test and scrambled feature map. For a given unit's feature map, a prediction map was designated as a statistical match if its correlation value exceeded 99% of the correlation values observed for the shuffled distribution – this was done using the selection set; we then saved the corresponding correlation value from the independent quantification set.

Visualizing Input Through Synthesis of Artificial Images in GAN

To visualize the input features that maximize the activations of individual units in CNNs, we randomly selected 91 CNN activation maps that best matched observed biological feature maps and used a generative adversarial network [1] to reconstruct images based on these activations. Then these reconstructed images were fed into Google Cloud Vision to extract the ten labels that can be used to explain the images. All of the labels were then classified into five categories based on their hypernyms in WordNet: “Artificial objects,” “Shape, pattern and color,” “Natural objects,” “Animal face and bodies,” and “Human face and bodies.”

Chapter 3. Results

As discussed above, feature maps (computed from baseline-subtracted spike rates measured over 200 ms from picture onset) allowed us to define the representations encoded by multiunits using a large hypothesis space, testing hypotheses considered to be at odds with one another. If neurons across the ventral pathway encode objects based on a top-down organization of object categories in the activity space, the neural response within these categories in the biological feature maps will increase from V1/V2 to V4 and IT. Further, if artificial units in CNNs encode the same visual information as neurons along the ventral pathway, the feature maps generated from CNN units and neurons would localize the same features.

Regions Showed Increasing Animacy Representations

We found that along the ventral pathway, there was an increasing response to signaling animal features, such as monkey and human faces, hands, and bodies in the natural images (Fig. 3a). The neuronal populations in the primary visual cortex responded across the whole image, including both low-level features like color and orientation and high-level features like hand (Fig. 3b). This is consistent with previous research that dorsal cortical layers in V1 encode not just orientation, but also more complex visual features [25]. V4 populations, while focusing on the same features as V1, started to include more features that are related to an animal (Fig. 3c). Finally, in IT, the populations narrowed their strongest responses to features that were almost only related to an animal (Fig. 3d).

We further examined the neural response within the “animacy” category by applying the pre-segmented “animal” mask (Fig 4a) over the biological feature map and compared the relative response within/outside the mask across V1/V2, V4, and IT relative to time (Fig. 4b). After image onset, the difference between within and outside the mask in response was the largest in IT (Fig. 4b, column 3), and then V4 (Fig. 4b, column 2), and the smallest difference was in V1. We further repeated the same analysis for all monkeys and all images, and the peak difference in response increased from V1/V2 to V4 and IT (Fig. 4c), arguing that population neurons along the ventral pathway respond more preferably to animacy features.

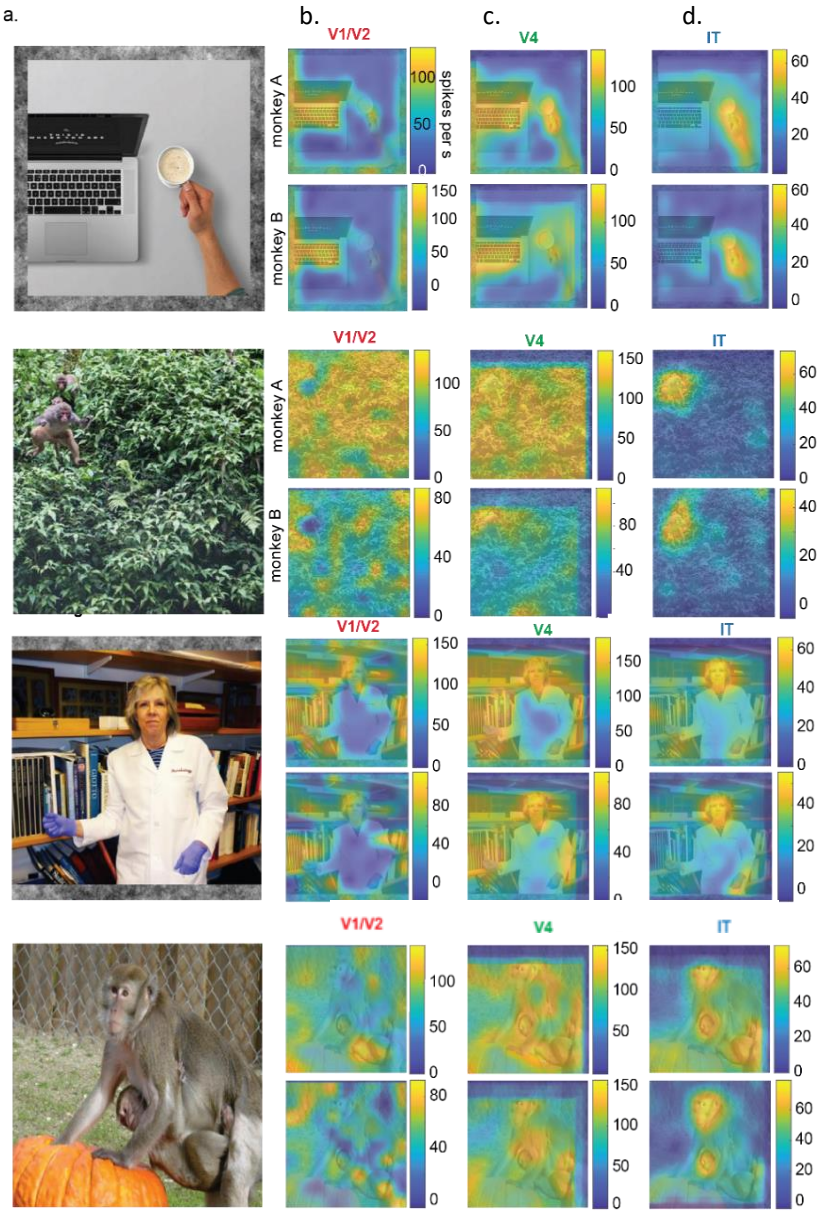

Figure 3. Biological Feature Maps along the Ventral Pathway

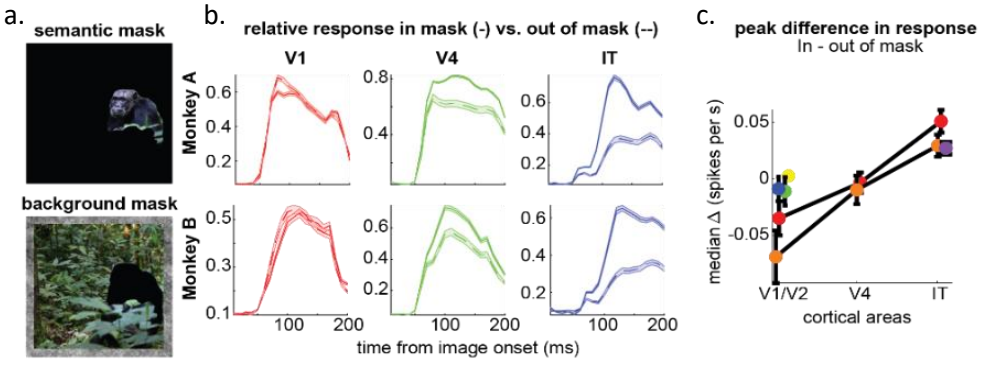

Figure 4. Neuronal Response within/outside of Semantic Category

CNN Prediction Space Showed Similarity with Biological Feature Maps in Regions of Animacy

Similarly, this attribution approach can weigh the hypotheses that visuo-cortical neurons need only be modeled by hidden units in artificial deep neural networks. We did a parallel study between *in silico* experiments and CNNs, comparing the biological feature maps (Fig. 5a-b) from neurons along the ventral pathway with feature maps from units from 3 convolutional layers in different CNNs (Fig. 5c, column 1) in terms of Pearson correlation coefficient. The reconstructed stimuli from the best matching CNN unit activations are interpreted as high-level features, including animal-related labels like "dog," while reconstructed stimuli generated from worst matching activations are described by low-level features such as "rectangle" (Fig. 5c, column 2). However, even for the best matching stimuli, only 12% labels could be classified as animal face and bodies and 2% as human face and bodies (Fig. 6).

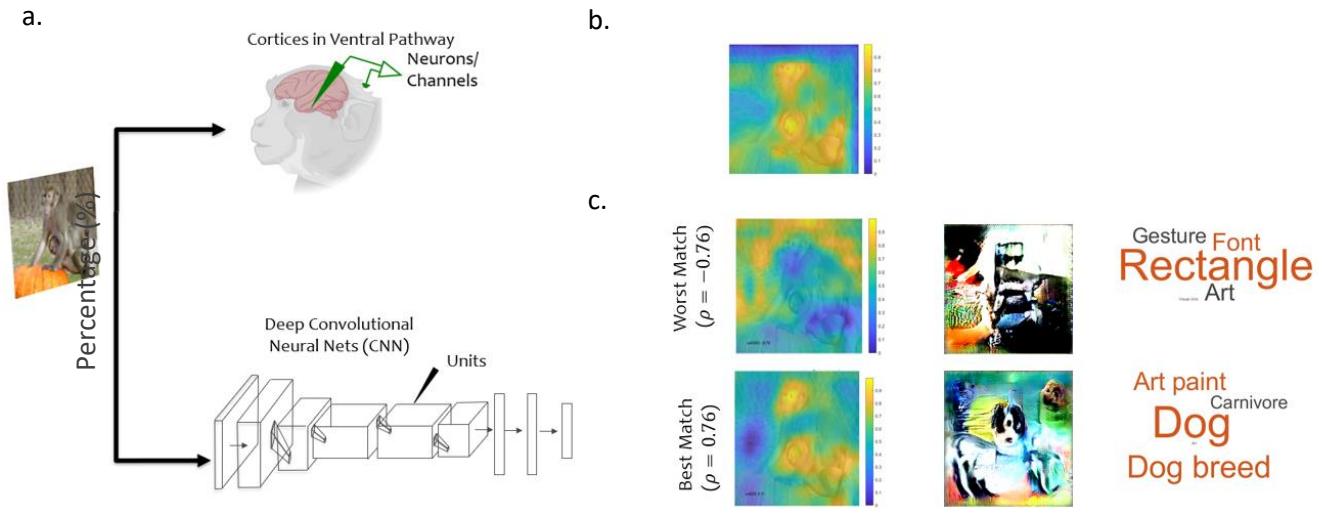

Figure 5. Parallel Experiment Between *in silico* and CNNs

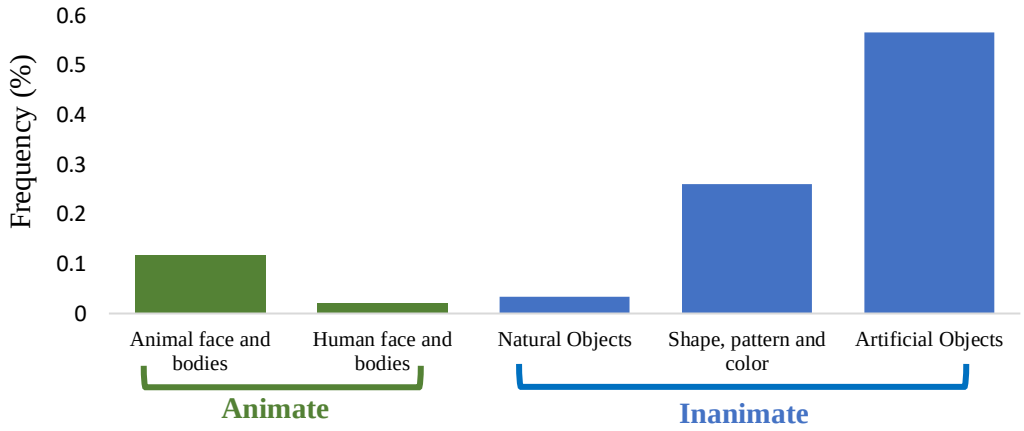

Figure 6. Percentage Categories of Semantic Labels

We further found that the best matching feature map from the first convolutional layer in AlexNet [16], VGG-16 [10], and ResNet-50[8] generally resembled the biological feature maps of the primary visual cortex (Fig. 3b, Fig. 7b). For the images that did not exclusively include an animal face (Fig. 7, first two images), the best matching feature maps from deeper convolutional layers in these CNNs were similar to the biological feature maps from later cortical areas (V4 and IT) in the ventral pathway. This is consistent with studies which show that V1-like filters for edges and corners emerge from algorithmic constraints for sparseness [23] or information maximization [4]. However, for the images that clearly show a monkey or human face (Fig. 7, last two images), the biological feature maps localized the human and monkey heads, while the face features were not specifically highlighted in the feature maps of the later convolutional layers. These results suggest that only a few filters from CNN could be a model of natural visual system responding to animate objects, consistent with the hypothesis that artificial neurons in CNNs do not learn to represent the same visual information as the biological neurons in primate brains.

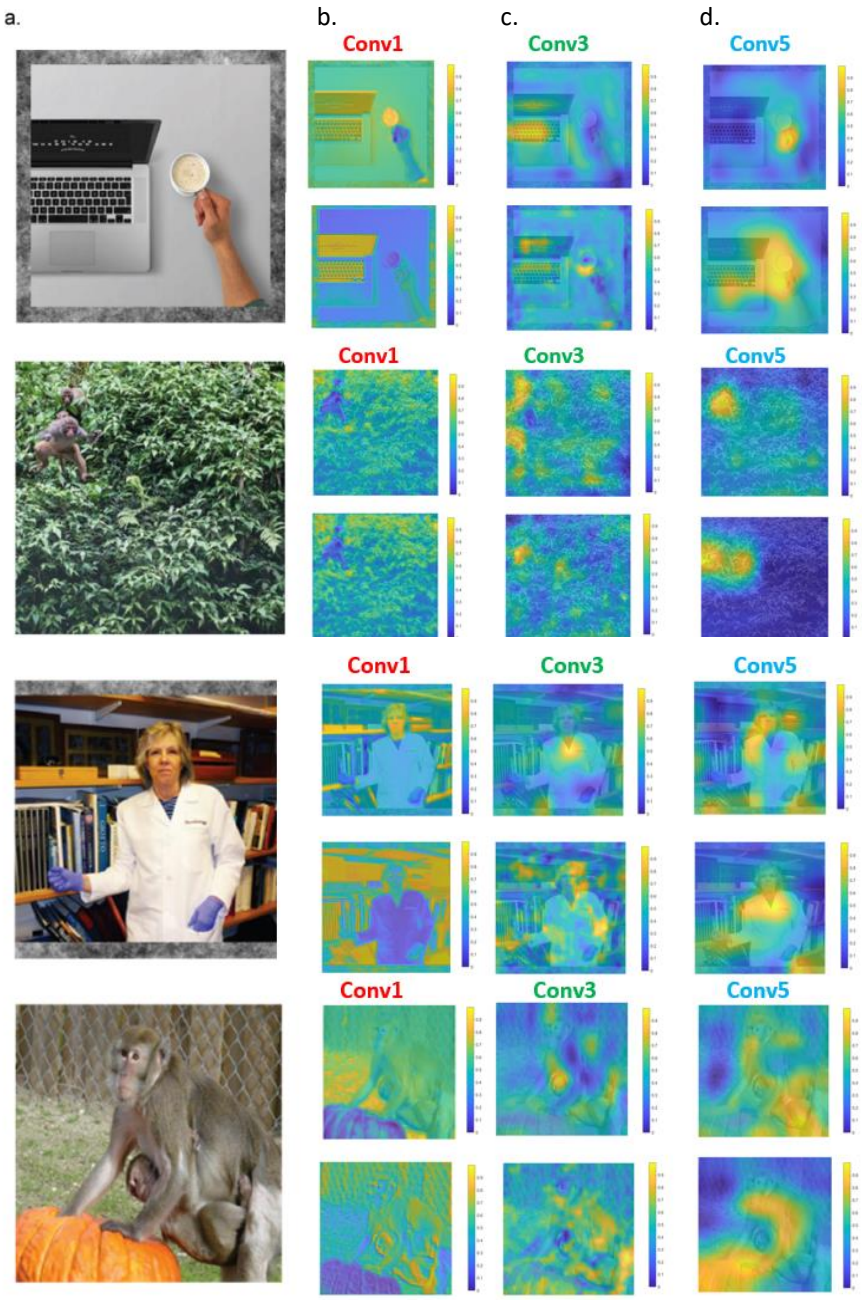

Figure 7. CNN Feature Maps of Convolutional Layers

Chapter 4. Discussion

Neural networks were historically inspired by the view of the brain as a set of individual units interacting to give rise to high-level processes and as long as this view holds, it is difficult to entertain an alternative, especially given compelling demonstrations that some CNNs may converge to the same solutions as the brain [12,27]. We used the CNN-related technique of attribution, where a neuronal receptive field is convolved with a larger visual scene to highlight pixel regions containing preferred shape configurations and give rise to a feature map. Our results show that our feature map approach is able to localize key objects in the natural images. Specifically, we have discovered a focusing response to signaling animal features, such as monkey and human faces, hands, and bodies across the ventral pathway. These results argue that animacy information, transmitted along the ventral pathway, characterizes the fundamental principle of visual recognition of primates in the natural environment.

To validate if such learning and computation principles are also reflected in individual units in CNNs, we further identified the filters in CNNs whose activations are most similar to the biological feature maps. We took the activation of these filters to generate an image capable of maximally activating that unit and found that in few cases, the synthesized image has been recognized as an animal. This suggests that even though CNNs are becoming the best visual recognition models of the primate visual system, they cannot fully represent the primate visual system. To create machine vision models that are more like the visual system in the brain, we need to take further inspirations from biology, such as including more filters responding to animate objects and adding recurrent connections [11].

Chapter 5. References

1 Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In Advances in neural information processing systems, pp. 658–666, 2016.

2 Baldassi, C. et al. Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. PLoS Comput. Biol. 9, e1003167 (2013).

3 Bell, A. H. et al. Relationship between Functional Magnetic Resonance Imaging-Identified Regions and Neuronal Category Selectivity. J. Neurosci. 31, 12229–12240 (2011).

4 Bell, A. J. & Sejnowski, T. J. The "independent components" of natural scenes are edge filters. Vision Res. 37, 3327–38 (1997).

5 Bruce, C., Desimone, R. & Gross, C. G. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. 46, 369–84 (1981).

6 Gross, C. G., Bender, D. B. & Rocha-Miranda, C. E. Visual receptive fields of neurons in inferotemporal cortex of the monkey. Science 166, 1303–6 (1969).

7 H. Caesar, J. Uijlings, V. Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. In CVPR, 2018.

8 "He, Zhang, Ren & Sun, “Deep Residual Learning for Image Recognition,” in arXiv preprint. arXiv:1512.03385"

9 Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–54 (1962).

10 "K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in arXiv

preprint. arXiv:1409.1556, 2014."

11 Kar, K., & DiCarlo, J. J. (2020). Fast recurrent processing via ventral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. https://doi.org/10.1101/2020.05.10.086959

12 Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput. Biol. 10, e1003915 (2014).

13 Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol. 71, 856–67 (1994).

14 Kovács, G., Vogels, R. & Orban, G. A. Selectivity of macaque inferior temporal neurons for partially occluded shapes. J. Neurosci. 15, 1984–97 (1995).

15 Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends in Cognitive Sciences, 17(1), 26–49. https://doi.org/10.1016/j.tics.2012.10.011

16 Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–41 (2008).

17 Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 1097–1105 (2012).

18 L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE TPAMI, 2018.

19 Lindsay, G. W. (2020). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. Journal of Cognitive Neuroscience, 1–15. https://doi.org/10.1162/jocn_a_01544

20 Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. Curr. Biol. 5, 552–63 (1995).

21 Nakashima, deeplab-pytorch, (2017), GitHub repository, https://github.com/kazuto1011/deeplab-pytorch

22 Olah, C., Mordvintsev, A. & Schubert, L. Feature Visualization. Distill 2, e7 (2017).

23 Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996).

24 Pasupathy, A. & Connor, C. E. Responses to contour features in macaque area V4. J. Neurophysiol. 82, 2490–502 (1999).

25 Tang, S. et al. Complex Pattern Selectivity in Macaque Primary Visual Cortex Revealed by Large-Scale Two-Photon Imaging. Curr. Biol. 28, 38–48.e3 (2018).

26 Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z. & Connor, C. E. A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nat. Neurosci. 11, 1352–1360 (2008).

27 Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A. 111, 8619–24 (2014).