Assessing brain-like activations in convolutional neural networks

Olivia Yoo
Neuroscience Concentration, Division of Life Sciences
Harvard College
Cambridge, MA 02138
oliviayoo@college.harvard.edu

1 Introduction

1.1 Question and hypothesis

I'm broadly interested in understanding the similarities and differences in information processing between biological and artificial vision. Towards that goal, I investigated which layers of convolutional neural networks (CNNs) have activations most similar to inferior temporal (IT) cortex activity in humans. I was particularly curious about the influence of network size and residual connectivity on a network's similarity to the regions of IT cortex that either recognize spatial layout or human faces. I also investigated how invariant the CNN activations are to various transformations of the visual stimuli, and what properties are preserved.

I hypothesized that the deepest convolutaionl layers of the largest neural networks would be the most similar to human IT activations. This similarity has been shown in some smaller hierarchical neural networks, and I expected that similar behavior would occur as the networks grow larger. These deepest layers would be mostly (though not completely) invariant to image transformations, and would be more invariant in the larger architectures. I expected that larger networks would be able to encode more complex features of the image, which may be more invariant to image transformations than the less complex features of smaller networks.

1.2 Significance

The goal of this project was to contribute to our understanding of how similar deep CNNs are to the human ventral stream by looking at much larger networks with residual connections. CNNs are a relatively biologically plausible model, as well as the residual connections – if I could identify similarities in activity in response to the same stimuli, CNNs and the brain might be doing similar kinds of computations. Exploring new kinds of networks would contribute to our understanding of biological object recognition and how it might emerge in artificial networks.

1.3 Background and literature review

Neuroscientists have long been interested in understanding the visual system. The visual cortex is hierarchically organized into modules with both feedforward and feedback connections. Within this hierarchical structure, visual information is processed into increasingly complex features. There are thought to be two different streams of visual processing in the brain. The dorsal ("what") stream is thought to be involved in processing an object's spatial location, and the ventral ("where") stream is involved with object identification and recognition. The ventral visual stream proceeds from the retina to V1, V2, V3, V4, and finally to the IT cortex. As Hubel and Wiesel showed in their seminal work, V1 neurons selectively respond to edges with a certain location and orientation, and receive input from many center-surround retinal ganglion cells that respond only to light in a very specific area. The increase in feature complexity from the retina to V1 is throught to occur through the rest of

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

the visual pathway. The IT cortex is a high-level region in the ventral visual pathway that is largely responsible for object recognition [Conway, 2018]. Within the IT cortex, the fusiform face area (FFA) is a region that has been shown to respond most strongly to faces compared to other kinds of objects [Kanwisher and Yovel, 2006]. The parahippocampal place area (PPA) is a region that responds mainly to the spatial layout of a place [Epstein and Kanwisher, 1998]. Both were investigated in this work. Since the IT cortex is later in the visual processing pathway and correspondingly has more complex features, it has been previously shown to be tolerant to many noisy parameters, like the object's position, scale, and background [Popivanov et al., 2015].

One approach to understand the brain is to construct models that attempt to explain a phenomenon, then compare the model's response to a human response. This can be a challenging task, given that human neural activity is often collected by fMRI or other techniques which don't neatly map towards outputs given by a computational model. Kriegeskorte et al. [2008] proposed a solution to this problem, called representational similarity analysis (RSA). Instead of trying to directly map brain or model activity onto a single metric, RSA instead compares the distance matrix of the response patterns elicited by the stimuli. This avoids dealing with the complicated and often non-analogous raw values from the data, instead constructing a representational dissimilarity matrix (RDM) to investigate the feature space of each subject/model. The RDM can then be much more easily compared between different subjects/models using a simple correlation metric. The advent of RSA enabled much more quantitatively rigorous comparisons of animal and computational neural activity, and it's what I used in this study to compare the neural network layers to human fMRI data.

Convolutional neural networks are a promising model for human visual system activity, due to their hierarchical structure. CNNs contain various layers of artificial neurons, whose activities are transformed between layers and manipulated to produce a prediction for a given task. In the last ten years, scientists have explored the realm of making CNNs to brain activity. Cadieu et al. [2014] compared the performance of CNNs to IT cortex on a visual recognition task, and discovered they are around 50-60% similar to IT cortex representations (and reach up to 80% after correcting for noise). These CNNs are significantly more similar to the IT cortex than the previous HMAX models. Yamins et al. [2014] demonstrated that hierarchical CNNs optimized for categorization performance can actually predict IT neural responses reasonably well. Kalfas et al. [2017] found that activity of middle superior temporal sulcus body patch (MSB) neurons was most responsive to shapes and outlines, and found that deep convolutional layers of CNNs were able to explain a median of 77% of the explainable variance of neuronal responses to image data. They found that this fit increased with the convolutional layers' hierarchy but was lower for the fully connected layers. Eickenberg et al. [2017] had similar findings for a different CNN (OverFeat) and demonstrated that some visual experimental paradigms were able to be generalized. Khaligh-Razavi and Kriegeskorte [2014] applied RSA to 37 different computational models, and discovered that the deep convolutional networks were best able to explain IT cortex. Kuzovkin et al. [2018] applied RSA to different timing and frequencies of human brain activity rather than spatial location within the brain. The overwhelming trend among different visual paradigms and higher ventral stream regions is that they are best represented by deeper convolutional layers of networks trained for object recognition.

However, traditional CNNs are constrained in size. As the number of layers increases, performance increases (like in the eight-layer AlexNet from Krizhevsky et al. [2017]), but only up to a certain point, after which performance will steeply drop off. Residual network architectures get around this constraint by implementing a residual connection, also known as a "skip" connection. This is simply an identity mapping that adds the exact output from the previous layer to a deeper layer, forming what's known as a residual block. This was developed by Microsoft researchers and their residual network won first place in the ILSVRC 2015 image classification competition [He et al., 2015]. Their implementation, ResNet, can have sizes from 18 layers all the way up to 152, something that would have been unfathomable before. The layers within these larger networks with the residual connections had yet to be compared to human brain activity.

1.4 Key papers

The image and human fMRI data from my project was obtained in the Kriegeskorte et al. [2008] study that initially proposed the RSA method. I referenced that paper heavily, as well as the later Nili et al. [2014] paper from the same group to ensures that I was constructing and comparing the RDMs appropriately.

The final key paper I referenced was the Kuzovkin et al. [2018] work, where they used RSA to compare human EEG frequency data to the layers of AlexNet. I largely based my methods off their experimental paradigm, especially when it came to conducting a permutation test to determine if the correlations between the human and model RDMs were significant.

2 Methods

2.1 Experimental approach

To invetigate which layers and CNN architectures best resembled human IT cortex, I utilized RSA to compare human and model activations in response to viewing 92 natural images. I constructed RDMs of human fMRI responses to those images for four different subjects in four different regions of the IT cortex. I presented the same 92 images to six different model architectures (AlexNet, ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152) that were pre-trained for image classification on the ImageNet dataset. For each convolutional and fully connected layer in the network, I constructed RDMs based on the layer's activations in response to the images.

In this case, RDMs are matrices R where entry R_{ij} is (1 - Pearson correlation) between the response to image i and image j. The raw correlation values were binned into ten bins based on the percentile score of the dissimilarity value, and the percentile value was plotted [Kriegeskorte et al., 2008, BrainIAK, 2017b]. Pearson correlation was used because we are not typically concerned with the exact magnitude of activations for fMRI data. RDMs were compared by measuring the Spearman correlation coefficient between them. Relationships between features might not be precisely linear, so the Spearman rank correlation was used to not make any assumptions about the feature space.

To assess the invariance of the CNNs, I followed the same procedure but instead with randomly transformed (cropped and/or vertically/horizontally flipped) versions of those same 92 images. This generated another RDM for each of the layers in response to those modified images in each of the networks. I then compared the RDMs of the humans to the networks with the original and modified images to determine which layers in which networks had the highest correlation to human responses, and how invariant they were to the transformations.

2.2 Dataset

Kriegeskorte et al. [2008] conducted an fMRI experiment on four human subjects. 92 images were presented to the subjects from six different categories: human face, human body part, nonhuman face, nonhuman body part, natural inanimate, and artificial inanimate (Figure 1). fMRI data from the left and right fusiform face areas (lFFA and rFFA) and the left and right parahippocampal place areas (lPPA and rPPA) were collected. The data was preprocessed by Brain Image Analysis Kit (BrainIAK) which provided the raw images presented to the subjects, the fMRI data, and a tutorial notebook for processing the fMRI data and generating RDMs [BrainIAK, 2017a,b].

2.3 Processing human data

I utilized the BrainIAK package and tutorials to process the Kriegeskorte et al. [2008] images and data from the human subjects. Due to installation constraints, I conducted this analysis in a Google CoLab notebook called human_processing.ipynb.

2.3.1 Generating human RDMs

I first generated the RDMs for the true human responses to the 92 natural images. Each of the subjects had the same set of images presented to them in a random order. For the RDMs to be compared, each entry needed to correspond to the same stimulus. I ordered the images by category (human face, human body party, nonhuman face, nonhuman body part, natural inanimate, artificial inanimate) and then ordered the human fMRI responses appropriately. I then utilized the tutorial notebook code to generate the sixteen total RDMS - four patients, four regions each [BrainIAK, 2017b].

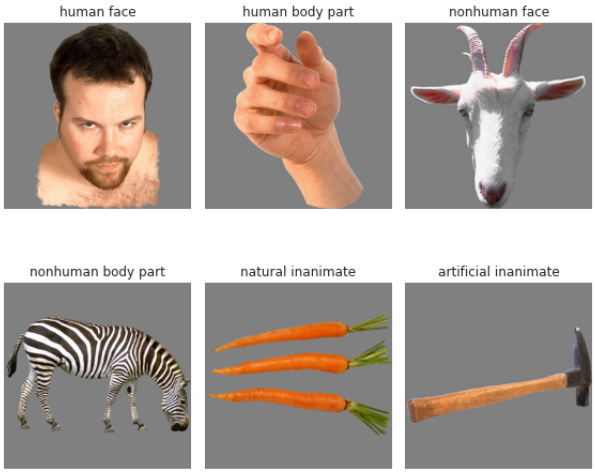

Figure 1: Sample images from each of the six categories of images.

2.3.2 Generating permuted RDMs

In order to determine whether a correlation between a model RDM and the true human RDM is significant, I created 10,000 randomized RDMs for each subject where there was no relationship between the stimuli and responses. To do so, I randomly permuted the response of each subject to the 92 image stimuli and constructed the RDMs.

I later created a null distribution of Spearman correlations between a layer RDM and the 40,000 randomized RDMs. I then compared the layer's correlation with the true human RDM to this distribution to determine whether the true human/model correlation was statistically significant ($p < 0.001$ compared to the null distribution). This threshold and method was chosen based on the Kuzovkin et al. [2018] experimental scheme.

2.4 Processing network data

2.4.1 Implementing networks

I investigated six different architectures in this study, all of which I accessed via PyTorch. I utilized ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 which were all pre-trained for image classification on ImageNet [team, 2023b]. I additionally used AlexNet, also pre-trained for image classification on ImageNet from Pytorch [team, 2023a]. I implemented all of these networks using a Google Colab notebook (with Colab Pro for more GPU resources) in cnns.ipynb.

2.4.2 Generating layer RDMs

My first step was to import the raw image data from Kriegeskorte et al. [2008] into the notebook and convert them to tensors using a custom PyTorch dataloader. I resized the images and subtracted out the means of the ImageNet dataset. I visualized the images afterward to ensure that I had loaded them in properly.

Then, for each of the networks, I downloaded the model from Pytorch and registered forward hooks for each of the convolutional models, utilizing code from an article by Kozodoi [2021]. These forward hooks capture the specific layer's activations upon a stimuli's forward pass through the network. I then set the model to validation mode, so that it did not learn from each stimulus that passed through, and fed the images in the correct order to the network. I then constructed the RDM for each of the layers in that network.

2.4.3 Modifying images

I built a separate dataloader for the images that I wanted to modify, and implemented random resized croppings, as well as random horizontal and vertical flips. These are transformations that the human visual system is very invariant to - humans can still recognize faces as such if they are upside down, and IT cortex has been shown to be invariant to these transformations [Popivanov et al., 2015]. I wanted to see if the network would be affected by these transformations at all. I then followed the same procedure in section 2.4.2 to construct RDMs based on responses to these images.

2.5 Human/model comparison

2.5.1 Constructing mapping matrices

In order to compare each network to the human IT responses, I constructed mapping matrices M for each network that were 4 (the number of IT regions) by the number of network layers. Matrix entry M_{ij} is the total sum of Spearman correlations between the human RDM for IT region i and the model RDM for layer j for each of the four human subjects, divided by the number of those correlations that were significant (as determined by the permutation test detailed in section 2.3.2). This matrix, when visualized, can identify which regions of the network are most similar to different IT cortex regions across individuals.

2.5.2 Individual layer activations

I was additionally interested in looking at how similar the layer activations were *within* the network. This would help determine which layers have similar feature representations and can also compare feature representations between the original and modified images. For each network, I constructed a square matrix L the size of the number of network layers. Entry L_{ij} was the Spearman correlation between the RDM for each layer. I did this to compare the network's response to the normal images to itself, the response to modified images to themselves, and the normal and modified images.

2.6 Codebase

All of the data, code, and figures for this project are located in my neuro140 GitHub repository. The code to load in the Kriegeskorte et al. [2008] data and functions to generate the RDMs in human_processing.ipynb was taken from BrainIAK [2017b], but I wrote the code to visualize the images and modified the data to generate both the ordered RDMs and permuted RDMs. The code to load in the CNNs came from team [2023a,b] and the code to implement the forward hooks was adapted from Kozodoi [2021]. I wrote the code to implement the image dataloaders and organized the layer response data to generate the RDMs. I wrote all of code for the data analysis and visualization myself in comparison.ipynb and figures.ipynb.

3 Results

For the sake of length and clarity, I will largely be discussing my analysis of AlexNet, ResNet18, and ResNet152. This enables me to examine the impact of residual connections and network size. All of the figures generated for the remaining networks (ResNet34, ResNet50, and ResNet101) will be included in supplemental material.

3.1 RDMs

I generated 16 true human RDMs, one for each of the four subjects in the four IT cortex regions investigated with the fMRI data by Kriegeskorte et al. [2008]. Since the RDMs are organized by category, I could analyze the RDMs by eye to gain a sense of which regions were most similar/dissimilar to each other (Figure 2).

The true human RDM for the lFFA in subject BE shows low dissimilarity among the human face images, marked by the many dark blue boxes in the upper left corner of the RDM (Figure 2A). The human face responses appeared to be most dissimilar to the natural inanimate and artificial inanimate images, and moderately dissimilar to the nonhuman faces. The RDM is noisy overall, but does appear to have some trends that reflect the categories of stimuli presented to the subjects. The permuted

human RDM has been shuffled to destroy the relationship between the stimulus and response, and appears to be largely disorganized, which was to be expected (Figure 2B).

The RDM for the deepest convolutional layer of AlexNet showed a striking trend, with human faces and body parts being extremely similar to each other (Figure 2C). AlexNet seems to be representing the human characteristics very similarly, and differently from the other categories. This lends support to the idea that deeper convolutional layers can encode for higher level features, which again was expected based on the previous literature. The nonhuman body part stimuli also appear to be encoded similarly, and it appears that many of the different categories have dissimilar activations.

The RDM for the deepest convolutional layer of ResNet18 appears to be identifying the different object categories the best, as demonstrated by the striking pattern of similarities/dissimilarities that emerges (Figure 2D). Like AlexNet, it identifies the human body parts and faces very similarly, but also identifies the natural inanimate category as being very similar. It additionally identifies many categorical differences, marked by the large red areas belonging to intersections between two categories. The deepest convolutional layers belonging to ResNet50 and ResNet152 do not show the same level of categorical representation (Figure 2E, 2F). Both show similarities among human faces, but this similarity does not extend to human body parts, and there doesn't appear to be any strong categorical trends among the remaining stimuli. Contrary to my hypothesis, this suggests that perhaps the increase in convolutional network size doesn't improve feature identification. However, the human RDM doesn't exhibit as clear of a categorical RDM representation as AlexNet or ResNet18, and the larger networks seem to more closely resemble the noisy human response.

3.2 Layer maps

To compare the similarities among the layers in the network to the human IT regions across subjects, I visualized the mapping matrices (described in 2.5.1).

For the AlexNet model (Figure 3, the human lFFA and rFFA responses seemed to be much more highly correlated with the layers than the lPPA and rPPA responses. For the lFFA and rFFA responses, all the layers had very similar correlations, with the later layers in lFFA being slightly higher and the middle layers in rFFA being the highest. The higher correlation between the FFA regions makes sense, given that this region is more highly involved in recognizing faces and objects, while the PPA regions are more sensitive to object spatial location, but it does seem odd that the correlation might be negative. This correlation is most strongly negative in the first layers, and ends being fairly close to 0 (meaning that there is no correlation at all) by the final layers. PPA might be encoding completely separate information than my CNNs. The relative similarity between all layers was slightly different than I anticipated.

Like the AlexNet model, the ResNet18 model had similarities that were more positive for the FFA regions and more negative for the PPA regions. The negative correlations with PPA were concentrated within the first third of the layers, then evened off towards a correlation of zero by the end. The FFA regions were much more closely correlated with the ResNet18 layer activations, particularly in the highest convolutional layers - Layer 19 (the second-to-last convolutional layer) had the highest Spearman correlation (around 0.1) for both of the FFA regions. This lends support to my hypothesis that deeper convolutional layers are more similar to the IT cortex activations.

The much larger ResNet152 produced a confusing result (Figure 5). Like with the other networks, the FFA tends to overall have more positive correlations and the PPA negative ones. However, there appeared to be no real trend between the layer correlations and hierarchy of the network. The largest correlations of around 0.11 tended to be with the FFA regions in the middle of the network (layers 38, 84, 87, 105). The PPA also had some unexpectedly positive correlations with the network (layers 99, 100, 114, 115). Overall, the PPA appeared to have most of its negative correlations with ResNet152 in the earliest layers, and the FFA seemed to have most of its positive correlations in the middle and later layers. ResNet152 doesn't possess the same simple hierarchy of similarities that was shown in ResNet18, so a larger network may not be better at resembling the ventral visual stream.

3.3 Within-network similarities

.

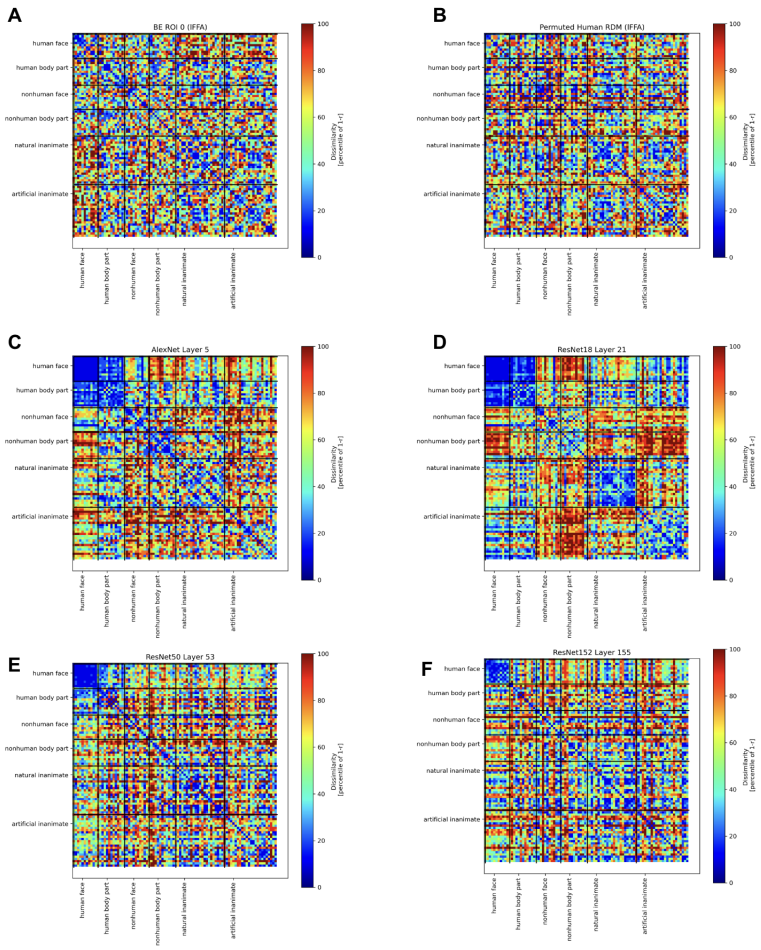

Figure 2: RDMS. A) True RDM of human responses from individual BE in the lFFA. B) One of 40,000 permuted RDMs, C) RDM of layer activations from the deepest convolutional layer of AlexNet. D) RDM of layer activations from the deepest convolutional layer of ResNet18. E) RDM of layer activations from the deepest convolutional layer of ResNet50. F) RDM of layer activations from the deepest convolutional layer of ResNet152.

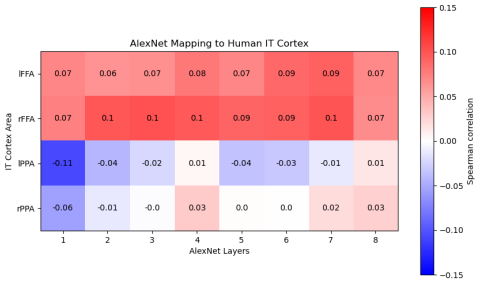

Figure 3: Mapping matrix of similarities between AlexNet layer activations and human IT cortex responses to the original 92 images.

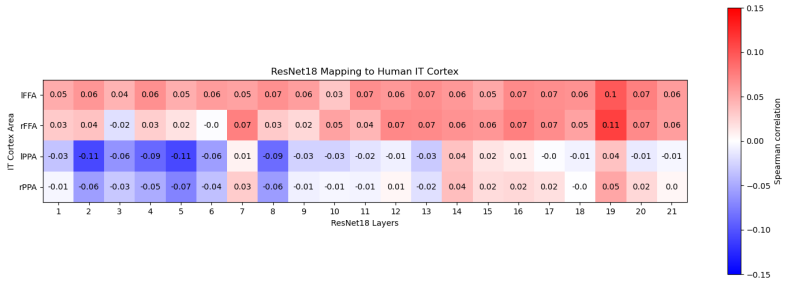

Figure 4: Mapping matrix of similarities between ResNet18 layer activations and human IT cortex responses to the original 92 images.

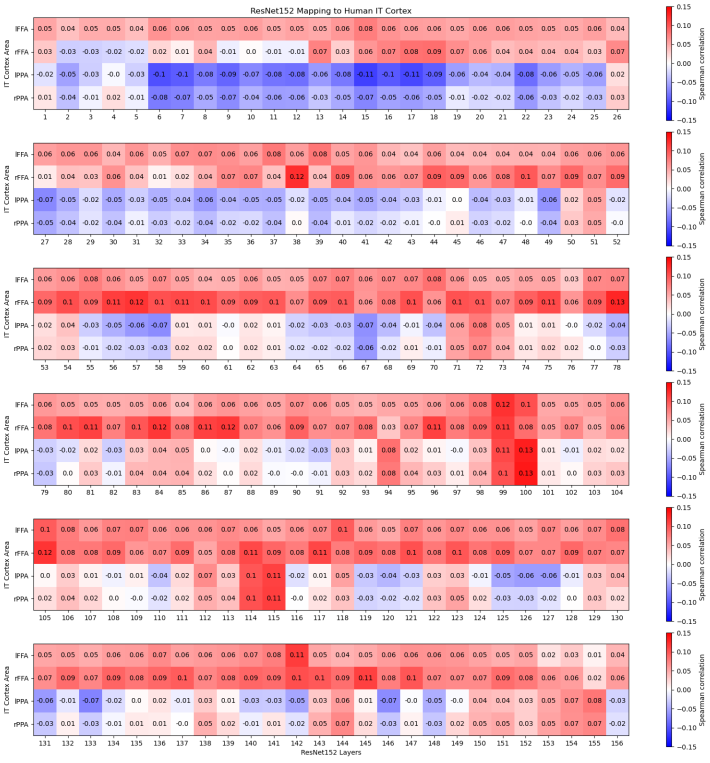

Figure 5: Mapping matrix of similarities between ResNet152 layer activations and human IT cortex responses to the original 92 images.

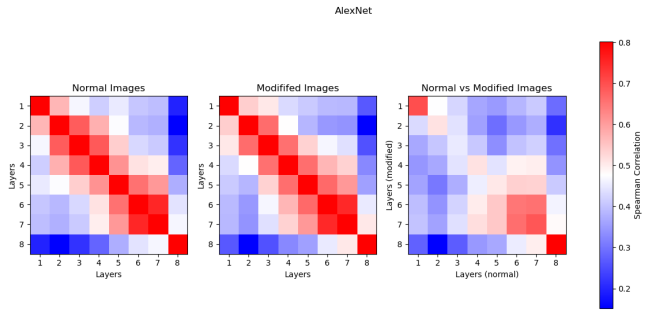

Figure 6: Similarities between the RDMs of layer activations for AlexNet in response to both the normal and modified natural images. Left: a similarity matrix between the normal responses to itself. Middle: a similarity matrix between the modified responses to itself. Right: a similarity matrix comparing the responses to the normal and modified images.

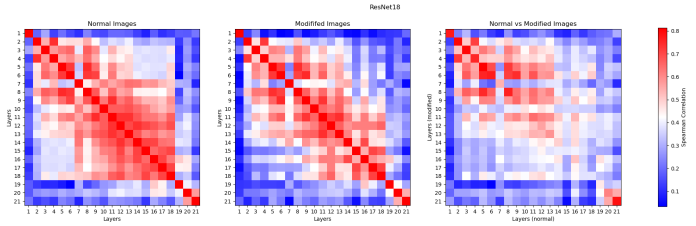

Figure 7: Similarities between the RDMs of layer activations for ResNet18 in response to both the normal and modified natural images. Left: a similarity matrix between the normal responses to itself. Middle: a similarity matrix between the modified responses to itself. Right: a similarity matrix comparing the responses to the normal and modified images.

To investigate how similar the networks responded to the normal and the modified images, I generated individual layer activations to determine how similar the feature representations in the networks were (as described in section 2.5.2).

AlexNet's layer activations in response to the normal images demonstrated that layers were most similar to the layers immediately above and below them, as might be expected due to the natural progression of transformed features throughout the network (Figure 6, left). This trend was also seen in the response to the modified images (Figure 6, middle). When comparing the layer activations for the normal and modified images, the first and last layers were the most similar (Figure 6, right). This makes sense, given that the first layer is extracting very similar low-level features of the image (like edges), and the last layer is the fully-connected layer designed to perform the image classification task. What's interesting here is that layers 6 and 7 are very correlated between networks. This suggests that AlexNet is encoding similar higher-level features of the images, despite being transformed.

Comparing the similarity of layer activations for ResNet18 demonstrates that the first layer and last three layers are very dissimilar from the others, but the remaining layers in the middle are relatively similar to each other (Figure 7, left and middle). This again, makes sense given the logical hierarchical transformation of features, but I am still curious as to why the first and last layers are so separate from the rest, yet similar between the two responses - perhaps the first layer is still doing the low level feature identification and the last few doing more of the classification task. When comparing the layers' response to the different stimuli, it appears that ResNet18 has more similar lower-level layers, and more dissimilar higher-level layers. This is different than what I see with AlexNet, and suggests that ResNet18 might be doing a different kind of feature encoding at a high level.

The same matrices for ResNet152 appear to show the same trend as that of ResNet18 (and is one I see across the other residual networks). The activations in response to the normal images are more similar to each other - perhaps given the more uniform size and location of the images - than the modified images. Additionally, the layer activations in response to the normal and modified images are more similar for lower-level layers than higher-level ones. This suggests to me that the residual

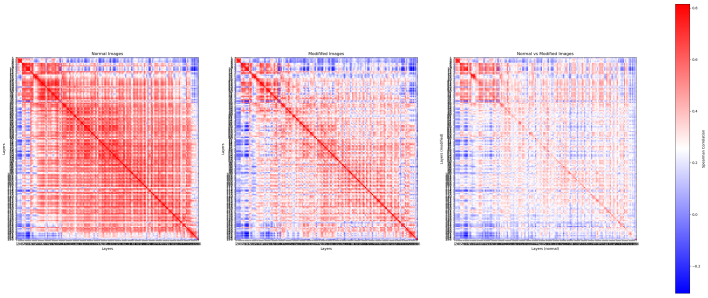

Figure 8: Similarities between the RDMs of layer activations for ResNet152 in response to both the normal and modified natural images. Left: a similarity matrix between the normal responses to itself. Middle: a similarity matrix between the modified responses to itself. Right: a similarity matrix comparing the responses to the normal and modified images.

architectures may not be as invariant to image transformations as the simpler hierarchical models such as AlexNet.

4 Discussion

4.1 Human/CNN comparison

When comparing the similarity of CNN layer activations to the activity of four human subjects in four different IT cortex regions, I found that there are mainly positive correlations with the left and right fusiform face areas, and mainly negative or approximately zero correlations with the left and right parahippocampal place areas. This makes sense, given that the FFA has been found to be associated with object detection (particularly for faces) and the PPA has been found to respond better to the spatial location of an object. The PPA response therefore might not correlate well with the activations of these networks that have been trained for object recognition - it might be more correlated with the same networks trained on a spatial recognition task. It is interesting and an area for future investigation why the PPA is not just uncorrelated, but rather negatively correlated with the responses of these networks. I therefore focus on the FFA results in my discussion, which will hopefully provide more insight given that the networks were developed for object recognition.

In AlexNet, I found that many of the layers had a positive correlation with the FFA activity, which was very similar across layers but was highest among the fully connected layers for the lFFA and the convolutional layers for the rFFA. This was a little bit unexpected, given that previous research has demonstrated that the final convolutional layer (layer 5) is the most closely associated with IT response. However, I don't have much information about the fMRI data, and it's possible that this is an artifact of the noisy data from only four individuals.

In ResNet18, I found that the highest correlation to the FFA activity was in the penultimate convolutional layer, with a Spearman correlation around 0.1 (Figure 4). This similarity was also found in the ResNet34 mapping (not shown here, in supplement). However, as the network grows larger, the trend of increasing similarity with depth seems to fade. The most similar layers begin to appear in the middle of the hierarchy for ResNet50, ResNet101 (both not shown here, in supplement), and ResNet152 (Figure 5).

Though I'm mainly interested in looking at the relative correlations between the different regions and layers, it's important to note here that none of the correlations were larger than 0.13. Given previous work, I anticipated that the Spearman correlations would have been larger. Again, it's possible that this is due to a high amount of noise in the fMRI dataset.

4.2 CNN layer comparison

In AlexNet, in response to both the normal and modified images, the layers of the network were most similar to the previous and next layers (6). The last layer was distinct from the rest. The first and last layers were had the most similar responses when presented with both the normal and modified

images - likely because the first layer was identifying very similar low-level features and the last layer was performing the classification. But the first two fully connected layers were similar whether the network was presented with the modified or normal images, suggesting that this network may be representing the high-level object features similarly.

For all of the residual networks, the layers within the network were overall more similar to each other when the normal images were presented, and less similar to each other when the modified images were presented (Figure 7, 8). This suggests that the network might not be as invariant to the change in size/orientation as AlexNet. Additionally, the earlier layers respond the most similarly to both the normal and modified images, suggesting that the residual networks encode the low-level features similarly and the high-level ones differently.

4.3 Limitations, future directions

One limitation which I've touched on already is that the human data is relatively limited. Having more subjects might be able to increase the amount of signal relative to noise in the fMRI data and help identify the true significance in the dataset more clearly. The same goes for the number of images presented - having more responses to different categories may help shed some light on the similarities between the FFA, PPA, and these networks.

Another limitation, specifically for the PPA analysis is that these networks were pretrained for object recognition. I did not have the compute resources to be able to train these networks, so I was only able to work with the networks pretrained for object recognition on ImageNet. I'd be interested to investigate the results of this analysis with a network trained for a different purpose.

I'd also be interested to see how correlated the responses were to the different categories of data. In particular, I'm wondering if the FFA data would be very highly correlated with the layer responses to faces, but not to other image stimuli. This might help us get a sense of whether the networks are better at detecting certain kinds of objects than others.

Additionally, I think it would be useful to repeat this analysis with an untrained network (with completely randomized weights) to gain a baseline understanding for how a neural network transforms the data. This might help put some of the correlation statistics into context, and would be a useful next step in analyzing the significance of these results.

4.4 Conclusions

In summary, the deeper convolutional layers in the ResNet architectures tended to have a higher similarity to the human FFA activations, but this trend disappeared as the networks grew larger than ResNet34. The residual networks appeared to represent the original and modified images differently, though had shared lower-level features. AlexNet appeared to represent the original and modified images very similarly, and had particularly shared higher-level features. This reinforces our understanding of the traditional hierarchical model AlexNet for biological vision, and suggests that small residual networks (ResNet18 and ResNet34) might warrant future investigation as models of biological vision. The larger residual networks might be too unwieldly or require further modification in order to resemble the human pathway.

References

BrainIAK. Brain imaging analysis kit tutorials and data downloads, 2017a. URL https://brainiak.org/tutorials/#data-downloads.

BrainIAK. 06 - rsa, 2017b. URL https://brainiak.org/tutorials/06-rsa/.

C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10:e1003963, 12 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003963. URL https://dx.plos.org/10.1371/journal.pcbi.1003963.

B. R. Conway. The organization and operation of inferior temporal cortex. *Annual Review of Vision Science*, 4(1):381–402, 2018. doi: 10.1146/annurev-vision-091517-034202. URL https://doi.org/10.1146/annurev-vision-091517-034202. PMID: 30059648.

M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 5 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.10.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811916305481.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature 1998 392:6676*, 392:598–601, 4 1998. ISSN 1476-4687. doi: 10.1038/33402. URL https://www.nature.com/articles/33402.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

I. Kalfas, S. Kumar, and R. Vogels. Shape selectivity of middle superior temporal sulcus body patch neurons. *eNeuro*, 4:113–130, 5 2017. ISSN 2373-2822. doi: 10.1523/ENEURO.0113-17.2017. URL https://www.eneuro.org/content/4/3/ENEURO.0113-17.2017https://www.eneuro.org/content/4/3/ENEURO.0113-17.2017.abstract.

N. Kanwisher and G. Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361:2109–2128, 12 2006. ISSN 09628436. doi: 10.1098/RSTB.2006.1934. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2006.1934.

S. M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 10:e1003915, 11 2014. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1003915. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003915.

N. Kozodoi. Extracting intermediate layer outputs in pytorch, 2021. URL https://kozodoi.me/python/deep%20learning/pytorch/tutorial/2021/05/27/extracting-features.html.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 11 2008. ISSN 16625137. doi: 10.3389/NEURO.06.004.2008/BIBTEX.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL https://doi.org/10.1145/3065386.

I. Kuzovkin, R. Vicente, M. Petton, J. P. Lachaux, M. Baciu, P. Kahane, S. Rheims, J. R. Vidal, and J. Aru. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications Biology 2018 1:1*, 1:1–12, 8 2018. ISSN 2399-3642. doi: 10.1038/s42003-018-0110-y. URL https://www.nature.com/articles/s42003-018-0110-y.

H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10:e1003553, 2014. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1003553. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003553.

I. D. Popivanov, J. Jastorff, W. Vanduffel, and R. Vogels. Tolerance of macaque middle sts body patch neurons to shape-preserving stimulus transformations. *Journal of Cognitive Neuroscience*, 27:1001–1016, 5 2015. ISSN 0898-929X. doi: 10.1162/JOCN_A_00762. URL https://direct-mit-edu.ezp-prod1.hul.harvard.edu/jocn/article/27/5/1001/28350/Tolerance-of-Macaque-Middle-STS-Body-Patch-Neurons.

P. team. Alexnet, 2023a. URL https://pytorch.org/hub/pytorch_vision_alexnet/.

P. team. Resnet, 2023b. URL https://pytorch.org/hub/pytorch_vision_resnet/.

D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111:8619–8624, 6 2014. ISSN 10916490. doi: 10.1073/PNAS.1403112111/SUPPL_FILE/PNAS.201403112SI.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.