Abstract

Brain-like Object Manifold Separation in Deep Neural Networks

Niranjan Rajesh

2023

The mechanisms by which the brain performs computations to solve visual object recognition is still not fully understood in neuroscience. A leading theory states that certain stimuli are encoded as manifolds by neuronal population activity of the ventral visual pathway. Literature suggests a key goal of the visual system is to disentangle these object manifolds by applying a series of space transformations – from a primarily entangled and linearly inseparable state of the manifolds to a space where the object manifold can be effectively separated by a hyperplane. This ability to separate while using a linear decision function towards the end of the pathway is proposed to be the remarkable and robust ability of object recognition.

Convolutional Neural Networks (CNN's), loosely based on the hierarchical visual system of mammals, could be used to verify the effects on linear separability of internal representations over the course of its layers. In other words, The deeper you go, the more linearly separable the neural activations. In this capstone project, I attempt to verify this by utilising a linear classifier on the representations of each layer for the forward pass of unseen data in a pre-trained CNN. Results show that there is a greater degree of separability in the layers of CNN in comparison to the pixel space, however, the increase in separability with increase in depth is not substantial.

Brain-like Object Manifold Separation in Deep Neural Networks

A Capstone Project
Presented to the Faculty of Computer Science
of
Ashoka University
in partial fulfillment of the requirements for the Degree of
Postgraduate Diploma in Advanced Studies and Research

by
Niranjan Rajesh

Advisor: Debayan Gupta

December, 2023

Contents

Chapter 1

Introduction

Our brains have the remarkable ability to recognise and classify objects despite variations in its appearance due to external stimuli like lighting, pose, perspective, etc. Once we have seen and learned how a cat looks like, we can look at a cat from any angle, orientation or even in different local brightness Moreover, our visual systems are able to generalise from very few data points. We encounter a species of a cat with a unique appearance for the first time, yet we are still able to confidently label them as a 'cat'. Thus, the ventral visual system is highly proficient at generalisable object-invariant recognition. These are traits that we aim to replicate in Artificial Intelligence (AI) through Deep Learning (DL) – a field initially inspired by neuroscience.

Early Neural Networks and Neuroscience Inspiration The first neural networks from the 1950s were multi-layer perceptrons [1] built with collections of computational units, 'neurons' in a hierarchical network that was directly inspired by the architecture of the brain. Later on in the 1980s, Convolutional Neural Networks (CNNs) [2,3] were designed, again with inspiration from the complex and simple cells discovered by Hubel and Wiesel [4]. The CNNs were the first Neural Networks specially designed to process visual images. After the inception of CNNs, it has only

been in the last decade CNNs have been adopted widely for object recognition tasks. This can be attributed to the hardwards developments in Graphical Processing Units (GPUs), the availability of large-scale image datasets [5] and the architectural designs of modern Deep CNNs [6, 7, 8, 9].

Despite the incredible advancements in the state of AI, many obstacles are left in the journey to bridge the gap of Artificial and Biological Intelligence. The state-of-the-art networks of today are vastly inefficient in terms of the data it requires to learn, especially in comparison to the rate of learning in humans. One significant requirement for large amounts of data is that these networks struggle with generalising beyond encountered training images. Appearance changes of an object cannot be robustly accounted for by a typical network unless the variability exists within these datasets or augmentations are introduced artificially [10]. This problem of invariability is solved elegantly by the brain - suggesting that the answers may lie in how the brain computes visual representations.

Further inspiration form Neuroscience through manifolds Neuroscience theory suggests that the variance in images can be represented as manifolds within larger spaces (pixels, neurons, etc.) where each point on the manifold is an instance of the variable object [11, 12]. Class manifolds [13] are structures made up of points that represent an instance of an object in a class. Object-invariant manifolds are similarly made up of points that represent an instance of the object under various physical transformations that vary its appearance. Images captured by a camera or a retina in the pixel or retinal space of objects produce highly entangled manifolds as seen in fig 1.1. These manifolds of objects are not linearly separable and hence, are difficult for classification. The theory states that the goal of the ventral visual stream is to transform its internal neural space to a 'good' neural space where these object or

class manifolds are linearly separable and easily classifiable as seen in fig1.2.

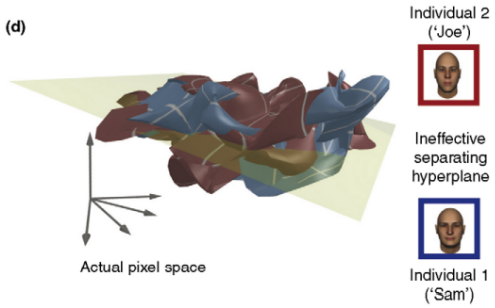

Figure 1.1: A 3D rendition of the pixel space from [12]

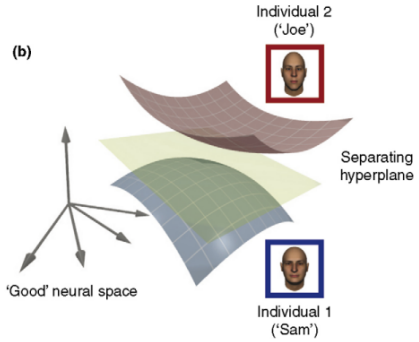

Figure 1.2: A 3D rendition of a 'good neural space' from [12]

The scope of this project This project is concerned with the notion of manifold separation in CNNs. I verify the hypothesis that CNN's are able to powerfully classify objects as it learns how to transform its internal neural space in order to make the manifolds separable. I attempt to achieve this goal by using a pre-trained network to extract layer-wise activations for images and determine the degree of linear separability among these activations for a binary classification problem. The test for linear separability simply involves measuring the test accuracy after fitting a linear classifier to the representations. The results do show evidence of increasing separability in the network's internal representations compared to the raw pixel space.

Chapter 2

Methodology

In this chapter, I discuss the methodology used to implement the investigation of object manifold separability in CNNs. The code for this implementation can be found in **my github repository for the project**.

2.1 The "Blessing" of Dimensionality

Manifolds and neural representation spaces are very high dimensional structures and spaces respectively. Since the project is concerned with verifying linear separability of structures in these high dimensional spaces, it is important to consider the combinatorics of linear separation in high dimensional spaces. Simply put, if there are N data points, they must necessarily be linearly separable in an $N-1$-dimensional space [14]. If there is an underlying structure among the data points, the number of dimensions that guarantees linear separability further decreases.

CNN layer activation spaces are very high dimensional in nature – a consequence of the high dimensional images these networks take as input. To verify separability of activations in the context of a binary classification problem, extra care and attention was given to the choice of the dataset and CNN architecture so as to not fall prey to the "blessing" of dimensionality.

2.2 Architecture and Data

CNN Architecture MobileNetv2 [15] was chosen due to its compact size and relatively low number of filters per layer that limits the activation space dimensionality. The model is able to perform well with such a limited size due to its architectural design, as seen in fig 2.1, of inverted residual blocks and thin bottleneck layers that forces the model to learn meaningful representations that are also very low-dimensional.

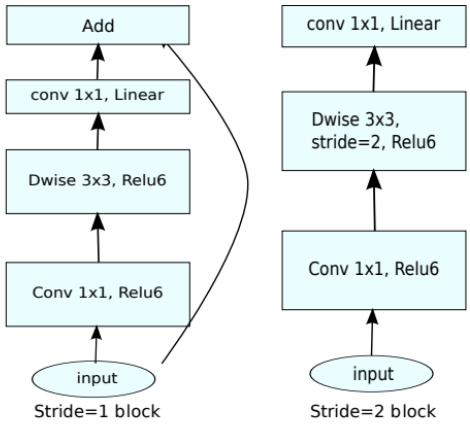

Figure 2.1: The MobileNetv2 block architecture figure from its paper [15]

Dataset For the problem at hand, the CIFAR-10 [16] was chosen primarily because of its ubiquity in classification benchmarking and also its small size. CIFAR-10 consists of 60,000 32x32 color images in 10 different classes (see 2.2), making it a challenging benchmark for object recognition tasks. The dataset includes a diverse range of objects and complex backgrounds, requiring a robust model for accurate classification. The variety and complexity of CIFAR-10 enable a investigation into the network's ability to disentangle object manifolds through different layers. The compact image size also allows for overcoming the problem in 2.1, making CIFAR-10 a well-suited dataset. Pairwise binary classification of the CNN on CIFAR-10 dataset

can be inspected in Fig 2.3. With the combination of the CIFAR-10 Dataset and

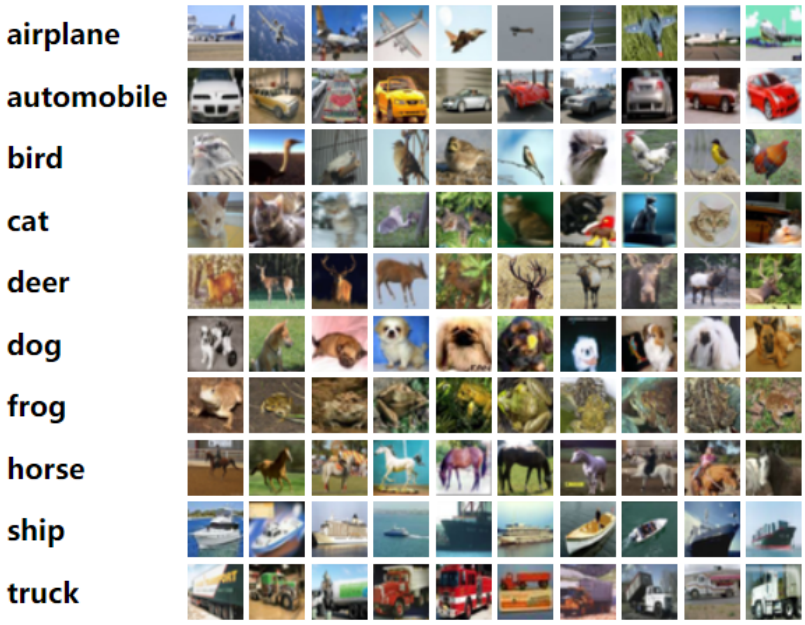

Figure 2.2: The CIFAR-10 classes figure from [16]

the MobileNet Layers, we were able to beat the dimensionality problem (sec 2.1). The maximum dimensionaity in layers of interest was 8192 for the first convolution layer. Since we are testing for separability in a binary class setting, we have 10,000 data-points which does not guarantee linear separability. This can be verified in fig 2.4.

2.3 Verifying Linear Separability

In order to verify linear separability over the course of the layers in the CNN, the following procedure was followed:

1. MobileNetv2 is trained on CIFAR-10 training data until convergence
2. Choose two classes from the dataset for the binary classification-based separability test.

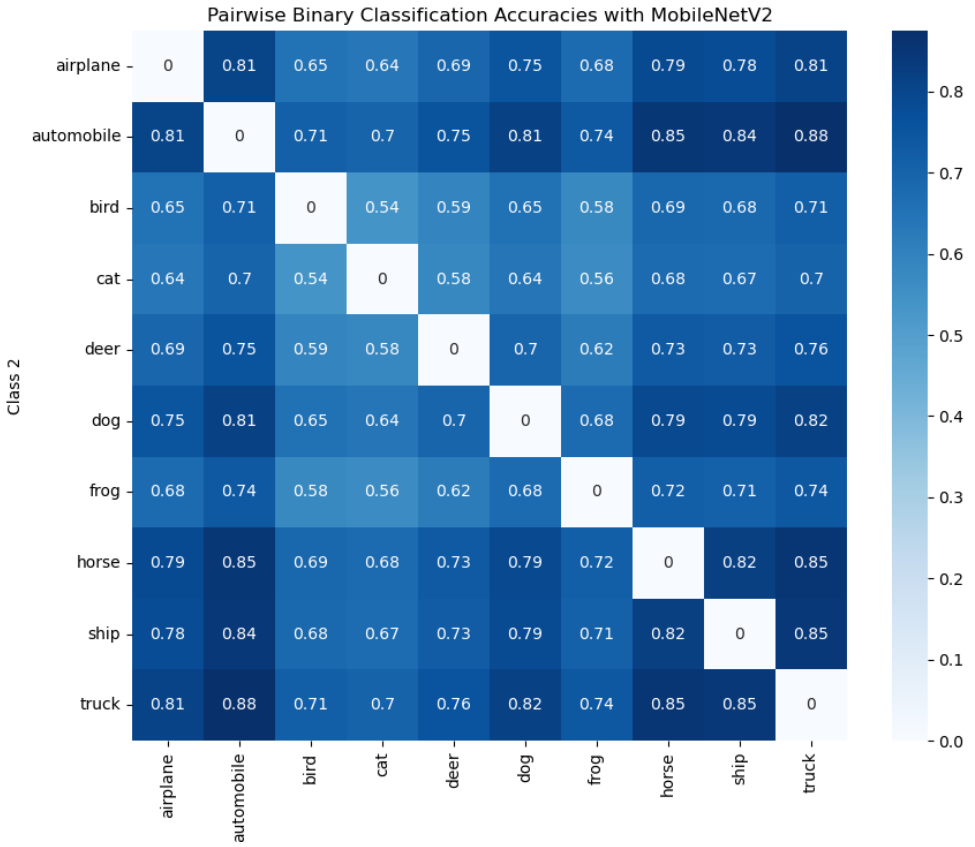

Figure 2.3: Pairwise Model Accuracy on CIFAR-10 Binary Classification

3. Determine linear separability of the images from the two cases in the pixel space. This is done by fitting the linear classifier with the flattened training images and testing it (identical to layer-wise examination below).

4. For each layer of interest in the trained CNN:

   (a) Activations are obtained for each data point in the training set and test. The training set consists of the CIFAR-10 training data for the two classes whereas the test set consists of the unseen data-points for the two classes. A new training and test activation data is obtained this way.

   (b) A linear classifier is fit on the training data activations until convergence or max_iterations

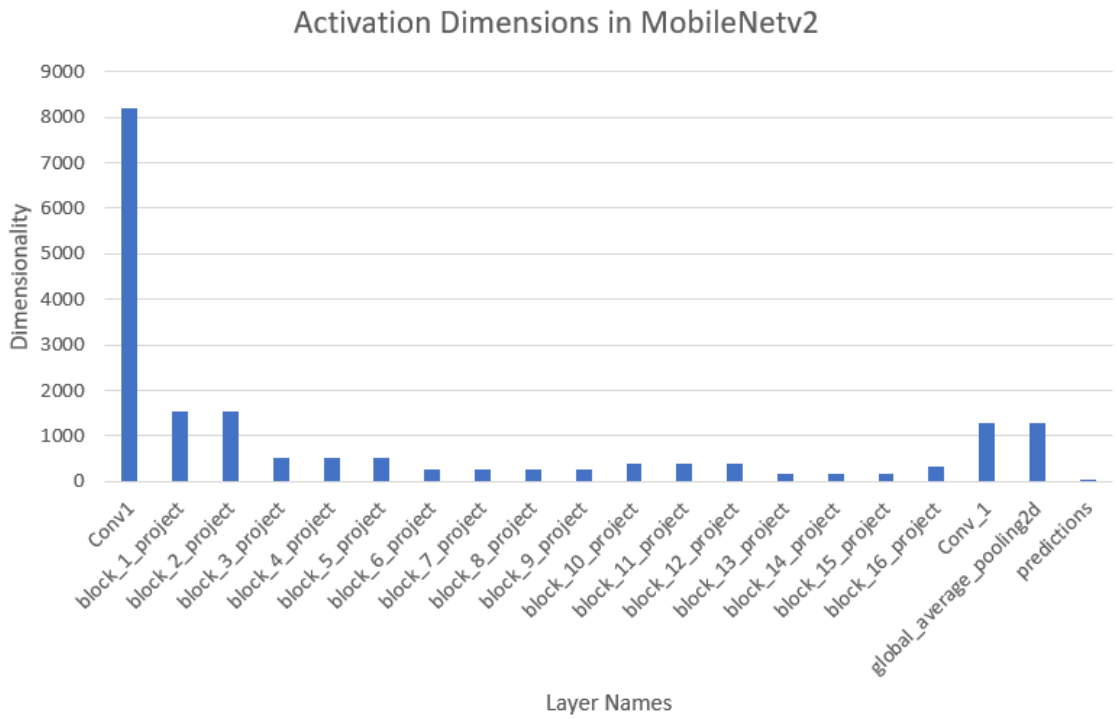

Figure 2.4: The dimensions of flattened activations from each layer of interest in the network. Layers are in increasing order of depth.

(c) Training and testing accuracy are measured for the classifier and stored

The layers of interest in this particular experiment were the convolutional layers at the end of each block. They were chosen to get a direct result of the operations from each block. They serve as ideal candidates to measure the degree of linear separability of the network's block.

2.3.1 Linear Classifiers

A brief bench-marking of linear classifiers was conducted on the activations of the initial convolutional layer of the CNN model. The flattened activations output for the layer had 8192 dimensions and 5000 images per class was used for training. The models used were a linear SVM, logistic regression, SGD-boosted SVM and an SGD-boosted logistic regression. As simple, yet powerful linear classifying models, only variations of SVM and Logistic Regression were chosen. As seen in Fig 2.5, the

simple Logistic Regression model fared best on the training task and was chosen for the downstream task of measuring linear separability.

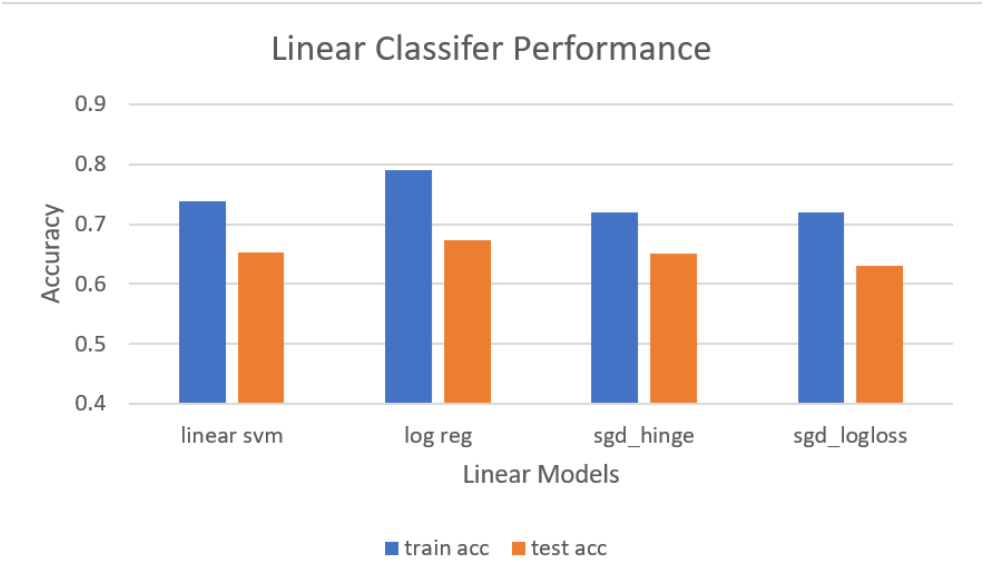

Figure 2.5: Performance of Linear Classifiers (Linear SVM, logistic regression, SGD-boosted SVM and an SGD-boosted logistic regression) on the extracted first convolutional layer activations ($10,000 \times 8192$ - train and 2000×8192 test) data

Chapter 3

Linear Separability Results

Layer-wise separability of Dogs and Cats

The procedure from 2.3 was followed to first investigate the degree of linear separability displayed by the images themselves in the pixel space and the corresponding representations of these images in the network. The results are displayed in fig 3.1. Firstly, the images themselves were only about 58% linearly separable. This is an almost-naive fact that is being reinforced by the separability test. Images are complex data-forms which are feature rich and highly variant. The $10,000$ training data-points of dogs and cats are sparsely distributed in the $32 \times 32 \times 3 = 3072$ dimensions of the CIFAR-10 image space. This is further extended by the fact that the objects to be classified within these datasets are quite similar in structure to one another.

The first key observation is that the first few layers of the network immediately improve the degree of separability of the dataset. The projection layers at the end of blocks 1 and 2 improve the classifier test accuracy by 13.7% and 15.7% respectively from the pixel space classifier. This suggests that the early convolutional layers perform operations to the internal neural space of the network that make the data

more readily linearly-separable. This observation hints that the class manifold within the pixel space is entangled and the early convolutional layers and their blocks are responsible for the initial disentanglement.

The second key observation is that the network's convolutional layer activations across its depth show a general increase in the degree of linear separability. Despite the peculiar drop in classifier accuracy (even below its pixel space counterpart), the layers display a rise in separability. This is the first piece of evidence that shows that the class manifolds of the networks begin to separate over the course of the layers. It is important to note that the training accuracy of the linear classifier is increasing at a more significant rate than the test accuracy. This could imply that there is a lack of generalisability within the layer activations which for the same class of objects which further suggests distinct visual representations being learned within the same class. The uncharacteristic drop in accuracy at the 'block14' projection-convolution layer is fascinating – especially since it is being mirrored in the training and test accuracy. A possible explanation could be one that involves a change in strategy within the network. CNNs typically start with low-level feature learning in the early layers and integrate more complex and spatial features towards the end. Block14 in this case could mark the beginning of the transition to these complex, spatial features. The representations being learned in this depth does not seem to be very effective but the network is able to recover swiftly and maintain its general rise in accuracy for the rest of the layer - perhaps, when the network has successfully learned how to integrate the complex features into its decisions. This is a premature hypothesis and further testing is required to corroborate it.

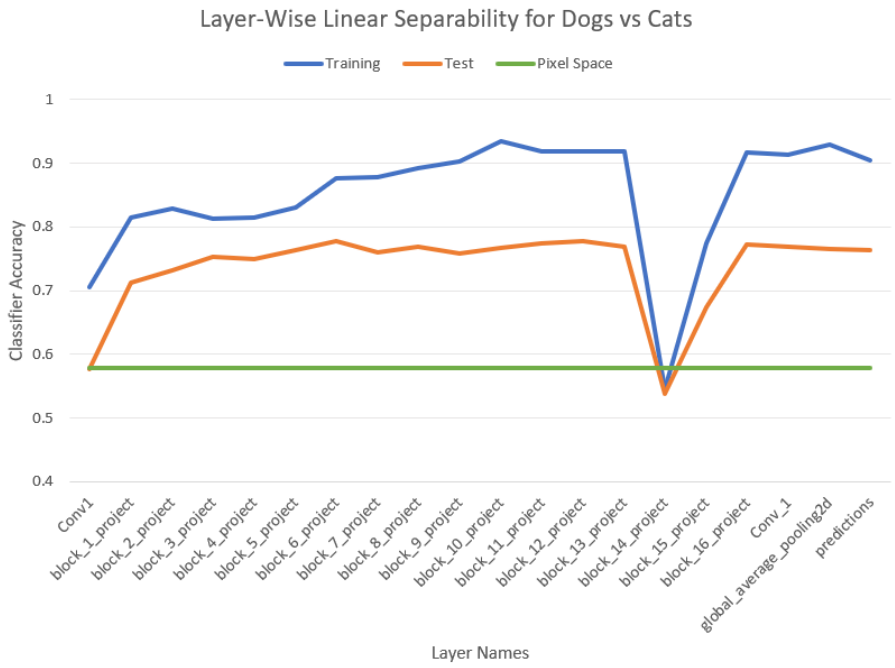

Figure 3.1: The linear classifier accuracy on the pixel space as well as the layer-wise activations for the dogs-cat binary classification subset of CIFAR-10 Dataset

Separability of Different Classification Pairs

The same procedure was repeated for multiple classification pairs in the second experiment. The results for this experiment are displayed in fig 3.2. The purpose of this experiment was to see the differences in the performance of the linear classifier for different class manifold learning problems. Three class pairs were chosen upon examination of the results from the pairwise classification scores from fig 2.3:

- Dog-Cat: A hard binary classification problem
- Ship-Frog: An easy binary classification problem
- Truck-Automobile: An unexpectedly simple binary classification problem

The Dog-Cat dataset was an obvious choice for testing due to the high level of structural similarity in the physical appearance of the class subjects. The difficult problem should have a lower degree of separability overall. In direct contrast, I chose to also

test the separability program on the Ship-Frog dataset which have significantly different structural visual features which should make their class manifolds be disentangled easier. A curious case was the Truck-Automobile dataset. Despite looking like identical classes, the model was able to achieve a relatively high performance on the pair. It would be interesting to see the separability of these two class manifolds.

Figure 3.2 supports my above hypothesis that the visually distinct classes are more 'easily' separable. The Ship-Frog dataset was already highly separable in the pixel space and in the early layers. After a slight boost in the degree of separability, the classifier accuracy stabilises quite fast. The Dog-Cat dataset which is harder to tell apart due to structural similarities within the data points does generally have a lower degree of separability.

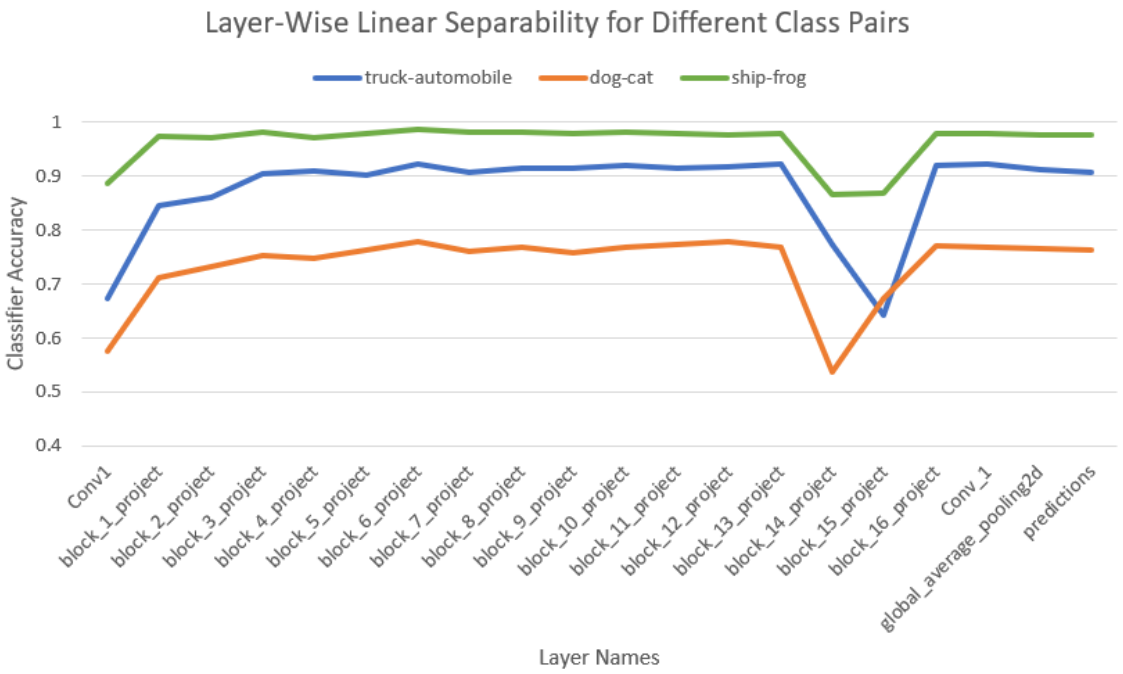

Figure 3.2: The linear classifier accuracy on the pixel space as well as the layer-wise activations for the three different binary classification subsets of CIFAR-10 Dataset

.

The third key observation follows that classes that are more difficult to classify has a lower degree of manifold separability over the course of the network layers. Interestingly, the Truck-Automobile dataset should also be fairly difficult to learn due to the images in both classes being nearly identical. The degree of manifold separability is closer to the Ship-Frog dataset than the Dog-Cat dataset. This implies that model has gotten quite good at separating these two classes despite challenges barring any potential hidden characteristics of the classes that make it easy to learn to separate the class manifolds.

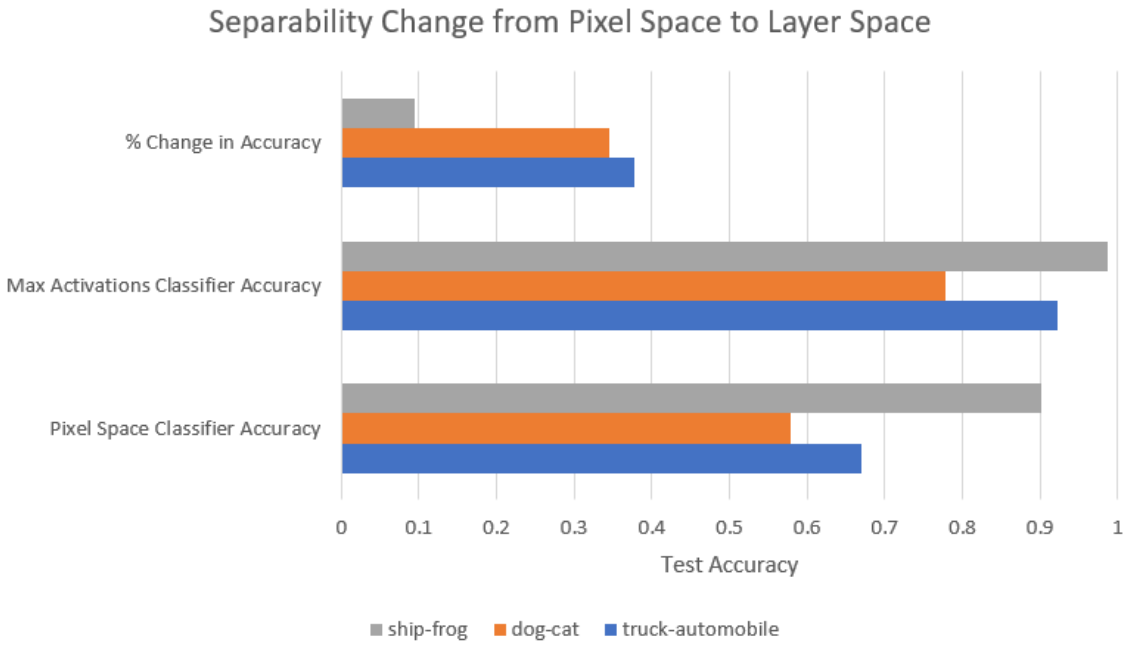

Figure 3.3:
.

The base and max classification accuracy values for each binary pairs of data are looked at more closely in fig 3.3. The base accuracy denoting the separability at the pixel space and the max accuracy being the layer at which maximum linear separability of activations was achieved. The figure shows that despite the high accuracy values of the Ship-Frog dataset, it was the Truck-Automobile dataset that saw the biggest jump between base and max classification. This, paired with the unexpect-

edly high classification accuracy in the same dataset. An explanation could be that the model plays more importance in learning the decision boundary, or equivalently learning to separate the class manifolds, for the Truck-Automobile dataset.

Chapter 4

Conclusion

Over the course of the project, I have developed a framework for identifying and verifying the class manifold hypothesis from the field of visual recognition. The hypothesis states neural population behaviour encodes object and class manifolds of visual representations.The points on these manifolds are made up of representations for a stimulus that is a variation within the object (external stimuli causing appearance changes) or within the class itself (multiple objects belonging to the same class). In the raw visual or pixel space, these manifolds are highly disentangled such that a linear decision function is not able to separate them but over the course of the layers, these manifolds are disentangled to a state such that they can be linearly separated by a decision function. This boils down the role of visual object recognition in the ventral visual pathway to a series of transformations aimed to disentangling these manifolds. In my project, I design a framework to verify this phenomenon in CNNs. I achieve this by training MobileNetV2 on CIFAR-10 and extracting layer-by-layer activations for each train and test image. These train and test image representations are then learned by a linear logistic regression model to investigate the degree of linear separability in the representations. In my work, I present three main findings:

- The first few layers of the CNN immediately improve the degree of separability

in comparison to the images' separability in the pixel space.

- The deeper you go in the CNN, the more linearly separable the representations become
- Classes that are difficult to distinguish suffer from lower manifold separability in the pixel space and in the activation spaces. This, however, could be mitigated by the model prioritising learning the difficult classes.

In this work, I add to the manifold hypothesis of visual recognition and my contributions imply that learning decision boundaries between classes is equivalent to the model learning to separate the class manifolds.

4.1 Future Work

A lot of work remains. The investigation in my project solely focused on class manifolds. Studies could be conducted on object-invariant manifolds that are constructed by the same object but captured under varying physical conditions that affect its appearance. The framework can be experimented on other CNNs, especially those that are deeper. It would be interesting to see how architectural changes in the model could affect the degree of manifold separability.

Bibliography

[1] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[2] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[4] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[10] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

[11] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[12] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[13] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

[14] MLS (https://stats.stackexchange.com/users/11915/mls). Is it true that in high dimensions, data is easier to separate linearly? Cross Validated. URL:https://stats.stackexchange.com/q/33441 (version: 2012-07-31).

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.