Convolutional Neural Networks, Illumination and Contextual Colors

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering.

January 24th, 2022

Kevin Doran
5120FG12-6

Advisor: Professor Hiroshi Ishikawa
Research guidance: Research on Computer Vision

Abstract

This work investigates convolutional neural networks from the perspective of two aspects of human vision, chromatic adaptation to illumination and contextual colors. Differences rather than similarities are identified, with some of these differences revealing potential failure modes of convolutional neural network classification models. Some approaches to addressing these issues are investigated.

Acknowledgements

Under Professor Ishikawa, I was able to explore a wide variaty of disparate topics and would be guided back whenever I got lost. I am extremely grateful for his guidance. I thank the Japanese Govenment for helping me fund my studies. I thank Waseda University for it's care of its students during the tumultuous coronovirus pandemic.

January 24, 2022
Kevin Doran

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the space that the human brain builds to represent the surrounding world, the color the brain assigns to a surface of an object is dependent on many things *in addition* to the light coming from the corresponding surface in the real world. This is a known phenomenon, and one of the reasons people developed the field of color appearance models. In effect, the 3 dimensional color space is insufficient to determine the color experienced. With this in mind, it is interesting to ask: do modern neural networks for image recognition integrate the same information that humans do to identify certain colors? If not, there are likely conditions under which they perform poorly as a result. The above question can be played in reverse: if neural networks *do* model colors well, will inspecting the internals of a trained network help reveal how color is processed in human brains?

In this work, progress is made towards answering these questions by investigating how two vision phenomenon, color constancy and contextual colors, interact with a convolutional neural networks trained for image classification. The evidence collected highlight the discrepancies rather than similarities in how the neural network integrates color information compared to humans. These discrepancies highlight both dataset limitations and architecture limitations. Approaches to address these limitations are considered.

1.1 Motivation

The potential cross-pollination between computer vision and human vision was mentioned above. It is a relationship that is frequently utilized to guide research. In this work, neural networks are investigated to learn about the human perception of color, and color is investigated to improve neural networks.

A separate, less discussed reason to consider color in the context of neural networks is the need to learn representations that are comparable to those of human vision. For many tasks there is no requirement for a model to create any specific intermediate representation before outputting a result— standard object recognition tasks fall into this category. Tasks that ask questions about human perception; however, inherently require a degree of parity with human vision. If the prediction task is to determine the response of a human, say to a visual stimulus, then the nature of human vision, such as the representation of shape or color becomes relevant. For a specific example, consider a photo editing application taking instructions from a human like "Change the color of the boots to look more brown." For the application to succeed, it seems important for system to understand how to more strongly elicit the sensation of brown in humans.

A second benefit of models having a degree of parity with human vision is that it allows greater model interpretability and explainability. Consider a model whose decisions depend on fine texture detail imperceptible to humans; compare it to a model whose decisions depend on representations with a degree of human vision parity—the latter affords an easier exploration into the behavior of the model.

1.2 Outline

The four chapters from Chapter 2 to Chapter 5 each carry out a separate experiment. Chapter 2, investigates how the image classification accuracy of a pretrained neural network is affected by changes to scene illumination. Chapter 3 collects and analyses data on human perception of the color shift between orange and brown. Chapter 4 uses the data collected from Chapter 3 to revisit the investigation into neural networks for image classification. Chapter 5 draws on the findings from Chapter 2 and Chapter 4 to propose an improvement to the neural networks that were tested.

Chapter 2

Classification and out of distribution lighting

This chapter investigates the effects of illumination changes on the classification accuracy of two ImageNet trained neural network models, ResNet50 and ResNet18. Certain types of illumination are shown to have a strong effect on classification accuracy. A simple white balancing algorithm applied to images taken under these illuminations is shown to improve the accuracy of these pretrained models. The problem is revisited in Chapter 5 which considers an approach that modifies the models themselves.

2.1 Background

The color constancy phenomenon and the process of white balancing are introduced below. The ResNet neural network architecture is also mentioned; it used in the experiment in this and later chapters.

2.1.1 Color constancy

The degree to which a human's perception of a surface remains stable under changing illumination is often referred to as color constancy. The phenomenon allows for the effect shown in Fig. 2.1. Human visual perception can be thought of as estimating the illumination in a scene and then discounting this illumination as part of it's perception of surfaces in the scene.

Figure 2.1: Illusion, created by Kitaoka, demonstrating chromatic adaptation. [15] The color that appears red is isolated on the left and it's RGB values are listed.

2.1.2 White balancing

In photography, white balance is the attempt to discount the illuminant in order to estimate how an image would change if it was taken under different lighting conditions. There is a notable distinction between white balancing and the phenomenon of color constancy of human vision: despite surface appearance having a *degree* of stability under different illuminations, the illumination still has an effect on our perception of a scene; in contrast, white balancing seeks to *remove* the effect of an illuminant so that the scene appears lit by some other illumination. This distinction is why some authors, such as Fairchild, dislike the term "color constancy" and instead include the phenomenon under the broader term, chromatic adaptation. [8]. This distinction is being emphasized here as white balancing can be considered to be removing information from an image (information about the original illumination).

There exists a wide range and quickly growing number of algorithms which attempt to estimate the illumination of a scene from an image (the first step in white balancing). An extremely simple approach is to assume that mean chromaticity of an image is a good estimate of the scene's illumination. This approach relies on the assumption that the mean reflectance distributions of objects in scene is flat. An object with this reflectance distribution would appear gray, leading to this algorithm being called the grayworld algorithm.

The ideas behind the grayworld algorithm will be utilized in this chapter and in Chapter 5.

More involved methods include Barron's convolutional color constancy algorithm [3], and a similar but more performance approach called fast Fourier color constancy, also introduced by Barron [1]. The latter approach Barron developed to be used used by the Google Pixel phone. Both of these methods apply a pretrained filter to the histogram of an input image's chromaticities. Methods utilizing neural networks, such as [11] are also plentiful.

2.1.3 ImageNet

ImageNet is an evolving dataset of human labeled images, with new versions both adding images to and removing images from the dataset. [7] [18] The version labeled ILSVRC2012, created for the 2012 ImageNet Large Scale Visual Recognition Challenge is widely used for training neural networks on the task of image classification, and it is dataset used in this work. An important aspect of this dataset is it's collection process: the images were sourced from search engine results and labeled by humans through the crowdsourcing service, Mechanical Turk [7]. Consequently, conditions such as time of day, monitor being used, monitor distance or room lighting are not controlled.

2.1.4 ResNet

ResNet is name of a specific neural network architecture trained for the task of image classification on the ImageNet dataset. [10] It was chosen to be studied in this work due to its wide use and relatively simple architecture. Wide use has established high confidence in the expected accuracy bounds of a from-scratch trained network. High accuracy bounds allows changes to the model to be more confidently attributed as the cause of a change in accuracy. An additional benefit of the ResNet architecture is the existence of multiple scaled version such as ResNet18, ResNet26, ResNet34 and ResNet50. Two forms of the ResNet neural network architecture, ResNet50 and ResNet18, are studied in this and subsequent chapters. An important distinction between these models is the size of their receptive field, which is discussed in detail is Chapter 4. ResNet50 and ResNet18 models and pretrained weights were obtained from the Pytorch library, version 1.10.0 package. [13]

2.2 Experiment, part I

This experiment extracts a subset of data from the Multiple Light Source Dataset [16] to investigate how different illuminants affect classification accuracy. The dataset creation process, the experiment proceedure and the results are presented below.

2.2.1 Dataset

The Multiple Light Source Dataset includes images of 24 scenes taken under 18 lighting conditions. The 18 illuminant setups are a combination of halogen, tungsten and colored LED lights. Each setup, is given a label. 3 of the labels are explained below.

- 2HAL: 2 halogen lights.
- 2HAL_DESK: 2 halogen lights and a tungsten desk lamp.
- 2HAL_DESK_LED-RG025: 2 halogen lights, a tungsten desk lamp and two LEDs (red and green) set to 25% power.

16 of the illuminations are of the form given in the 3rd example above. The spectral information for for the illuminants and the camera sensitivities are provided in the dataset; however, this information is not needed in this experiment.

For the purposed of this work, the Multiple Light Source Dataset was first reduced to a subset in order to ensure every image contained a central object corresponding to one of the 1000 ImageNet classes. The chosen crops and labels for a single illumination are shown in Fig. 2.2. The labels were chosen by the author. For the two scenes labeled as "cup", both the classification output of "cup" and "coffee mug" were accepted as a correct classification (both "cup" and "coffee mug" are separate ImageNet classes). A single scene under all 18 illuminations is shown in Fig. 2.3. As a data augmentation step, for each image, a further 5 images were created by a second round of cropping: 4 corner crops and a center crop. In total, this process produced 990 images (11 scenes x 5 crops x 18 illuminations). The cropped images are RGB images with dimensions 224x224.

Figure 2.2: 198 images (11x18) are extracted from the Multiple Light Source Dataset: 11 separate scenes under 18 illuminations. The above 11 images are the 11 scenes shown for a single illumination. The same views are used for all 18 illuminations.

2.2.2 Method

The experiment procedure was to record, for each of the 990 images, the class predictions given by the pretrained ResNet50 and ResNet18 models. These models were introduced in 2.1.4. The 990 images were inputted into both ResNet models, and the class predictions were recorded.

2.2.3 Results

The results in for ResNet50 are presented in this section, and the results for ResNet18, which are similar, are included in the appendix. A condensed version of the classification results is shown in Fig. 2.4. The results are condensed by ignoring the scene-crop pairs that were correctly classified under every illumination. The complete version of results are shown in Fig. A.2. Summing

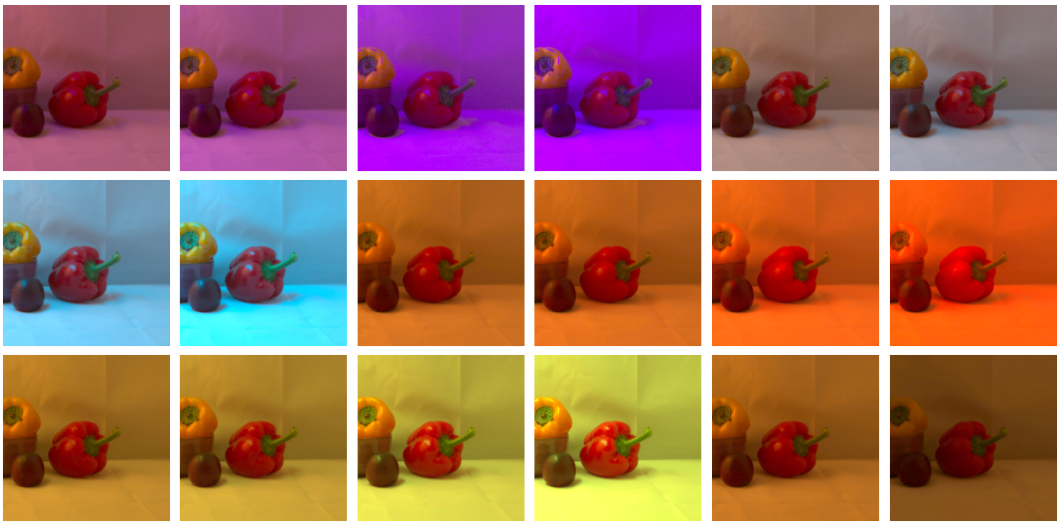


Figure 2.3: All 18 illuminations for the capsicum scene (before the 5-crop). Fig. A.1 in Appx. A contains the same images with an accompanying explanation of the illuminants and their labels.

correct classifications and normalizing for each illuminant produces the accuracy results shown in Fig. 2.5. The illuminant resulting in the most correct classifications was 2_HAL_DESK_LED-RG075 (2 halogen lights, a tungsten desk lamp and red and green LEDs set to 75% brightness). This illuminant will henceforth be referred to as the *best illuminant*. The illuminant resulting in the least correct classifications was 2_HAL_DESK_LED-B100 (2 halogen bulbs, a tungsten desk lamp and blue LEDs set to 100% brightness). This illuminant will henceforth be referred to as the *worst illuminant*. These two illuminants are shown in Fig. 2.6 for the capsicum scene. For these two illuminants, images for which classification was correct under the best illuminant and incorrect under the worse illuminant are listed in 2.7.

2.2.4 Discussion

The illumination has a significant effect on the classification accuracy. Most notably, classification suffers in the presence of predominantly blue illumination. Investigating specific examples where blue illumination caused a misclassification reveals that blue illumination causes the network to increase the probability assigned to object classes typically seen under water. This is evidence to suggest that ResNet trained for classification on ImageNet will struggle under illumination not encountered in the training set. When utilizing pre-trained networks via transfer learning, this limitation should be considered if networks trained on ImageNet are being used for tasks where images

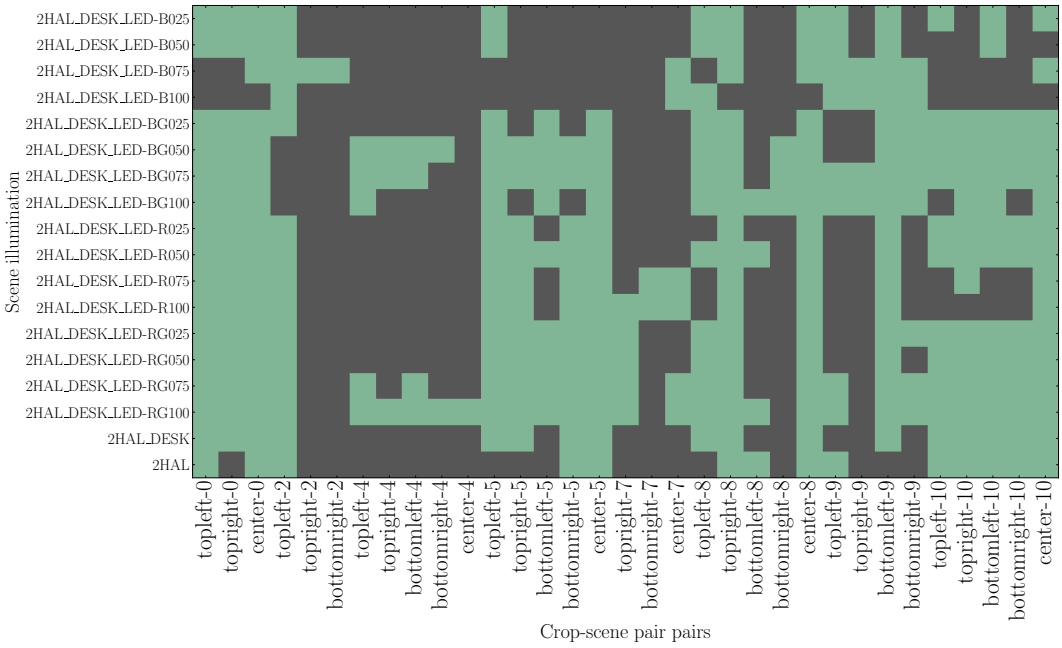


Figure 2.4: Classification results for for every scene-crop-illumination triplet. **Green: correct classification, black: incorrect classification**. The network is ResNet50. Illumination is given the whole y-axis, while the scenes and crops are flattened onto the x-axis. Scene-crop pairs that were correctly classified for every illuminant are excluded.

are taken under illumination not common in the ImageNet dataset.

2.3 Experiment, part II

The second part of the experiment attempts to improve the classification accuracy from part I. As certain illumination can negatively effect classification, adding a degree of illumination invariance to the input images is attempted.

2.3.1 Method

The experiment proceeds the same as in part I, except for one modification: before inputting the images to the ResNet model, the images are preprocessed with the grayworld white balancing algorithm, introduced in 2.1.2. 6 before-after examples are shown in Fig. 2.8. The 990 images are were inputted into both ResNet models, and the class predictions were recorded.

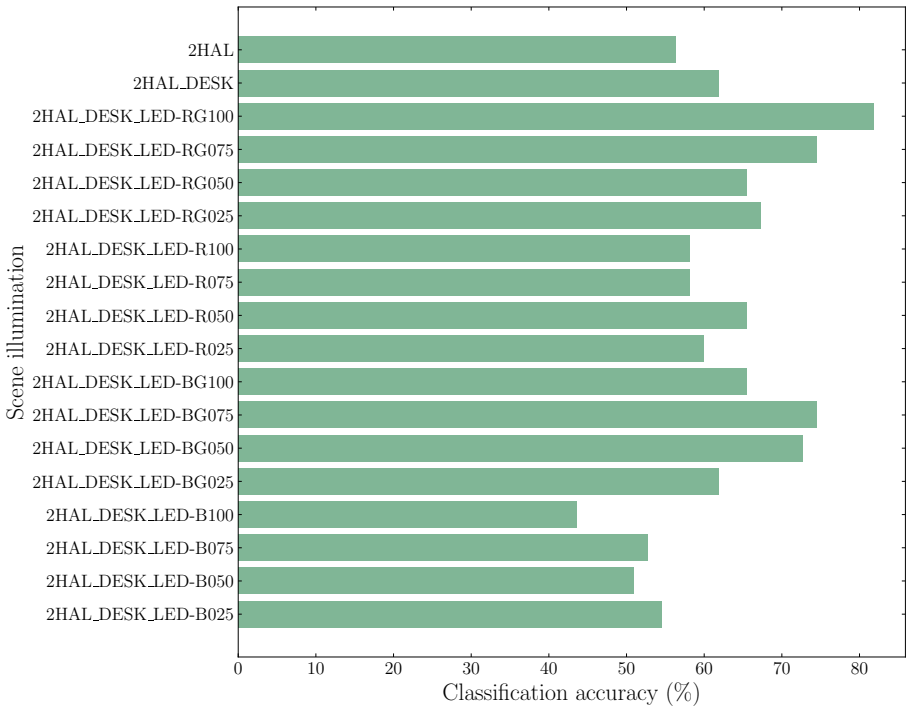


Figure 2.5: ResNet50's classification accuracy at classifying the 55 scene-crop pairs for each of the 18 illuminants.

2.3.2 Results

The classification accuracy grouped by illuminant is shown in Fig. 2.9 for Resnet50, and in Fig. 2.10 for Resnet18. Image classification improves, especially for images taken under illuminants that led to poor accuracy in Experiment 1.

2.3.3 Discussion

The classification accuracy improvement suggests that removing the effects of illumination can be a useful preprocessing step for images that are taken under illumination which is out of distribution for ImageNet. It is interesting to consider *why* this approach might be effective. Two hypotheses are as follows.

1. The ImageNet dateset and classification task either doesn't allow or doesn't require models to learn representations that exhibit illumination invariance to the extent tested in this experiment.

2. The ResNet18 and ResNet50 architectures are not effective at learning representations of illumination.

The failure cases shown in Fig. 2.7 suggest that there is some truth to the first

Figure 2.6: *Left*: the capsicum scene under the best illumination. *Right*: the capsicum scene under the worst illumination.

Figure 2.7: Scene-crop pairs where classification is correct in the best illumination and incorrect in the worst illumination. Only 1 of the 5 crop is shown per scene. Some, but not all, of the other 4 crops experienced the same classification results.

hypothesis: the ResnNet50 model is capable of achieving 80% accuracy at the ImageNet classification task despite heavily weighting the presence of blue with classes associated with objects under water. With water more readily absorbing longer wavelength light, it is not surprising that the model has exploited this relationship. What is more interesting; however, is that the model does not seem to have been heavily penalized by forming such a rudimentary correspondence.

There is supporting evidence for the second hypothesis also. It is difficult to infer illumination information from a small patch of an image. The activations of the first few layers of the ResNet50 model have a very narrow receptive field. Indeed, for approximately the first half of ResNet50, no activation has a receptive field encompassing the whole 224x224 image. For ResNet18, the situation is even more pronounced with many of the activations leaving the final pool layer having a receptive field that doesn't include the whole image.

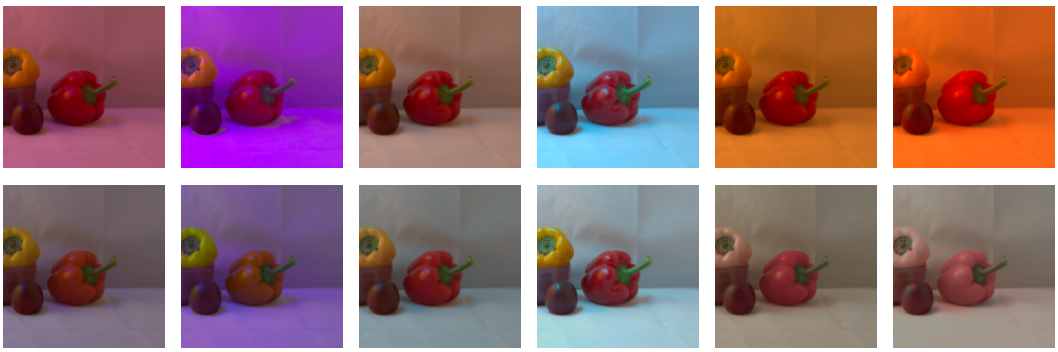


Figure 2.8: *Top row*: capsicum scene for 6 of the illuminants. *Bottom row:* the same 6 scenes after processing with the grayworld algorithm.

The narrow receptive fields of the early layers forces any robust representation of illumination to only be present in later layers of the network.

Both of these questions are taken up from a different perspective in the next chapter.

As discussed in Section 2.1.1, white balancing algorithms, such as the grayworld algorithm, seek to remove the effect of the original illumination, and as such, they are removing information from the original image. If for example, all objects in a scene were in fact diffuse reflectors with a flat reflectance distribution, then the grayworld algorithm would remove any trace of the original illumination. This highlights a major drawback of using any white balancing algorithm to preprocessing model inputs.

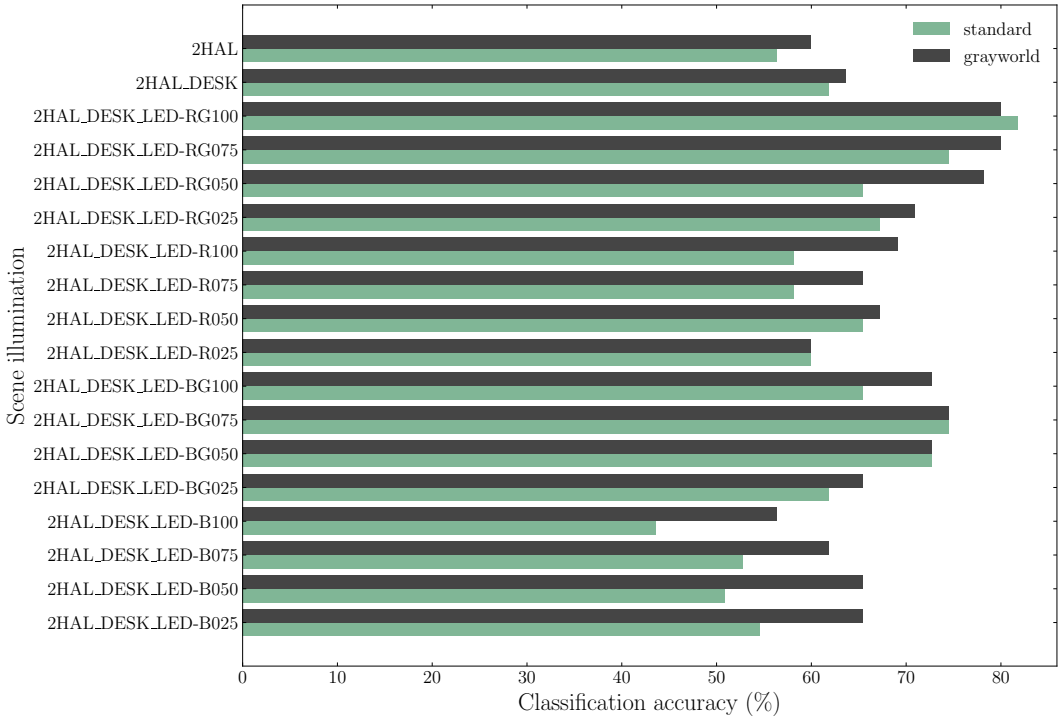


Figure 2.9: ResNet50's classification accuracy at classifying the 55 scene-crop pairs for each of the 18 illuminants.

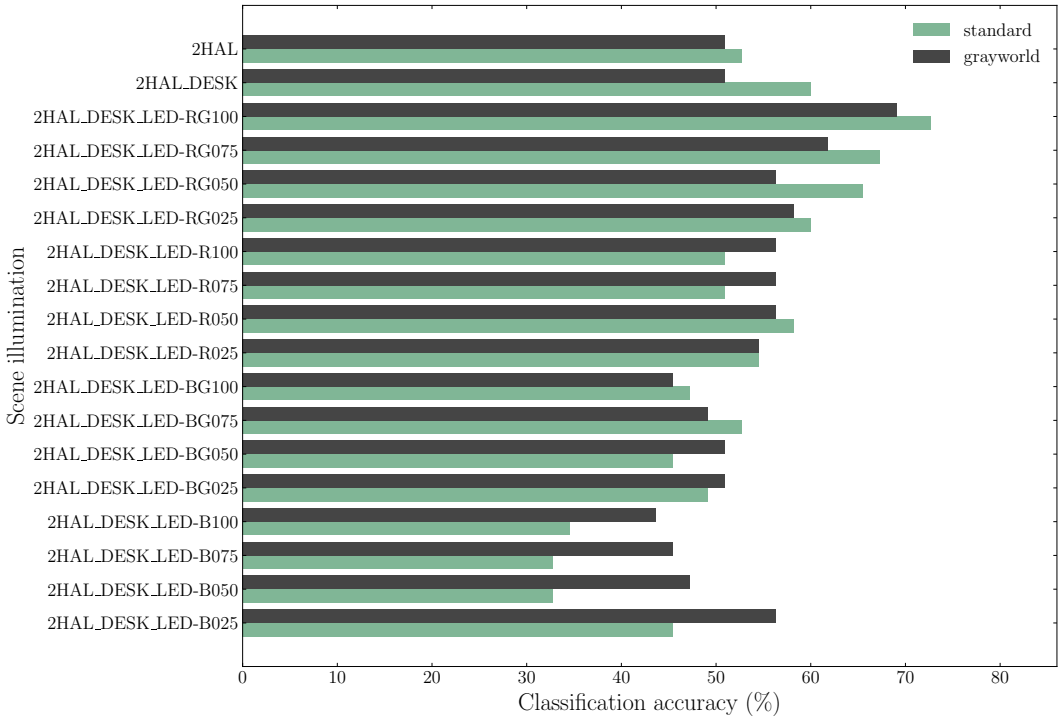


Figure 2.10: ResNet18's classification accuracy at classifying the 55 scene-crop pairs for each of the 18 illuminants.

Chapter 3

Distinguishing orange and brown

This chapter investigates the phenomenon of *contextual colors* through a case study on the perceptual distinction between orange and brown. The main purpose of the work in this chapter is to collect a dataset to be used in Chapter 4.

3.1 Background

Some colors perceived by humans can be described as *contextual colors*. Contextual colors are also called related colors. In "Color Appearance Models", Fairchild defines related and unrelated colors as follows [8]:

Unrelated color: colors perceived to belong to an area seen in isolation from other colors.

Related colors: colors perceived to belong to an area seen in relation to other colors.

In other words, a color is a contextual/related color if it can *only* be perceived when other certain colors are present in a vicinity. Brown and gray are two examples of contextual colors: in an otherwise dark environment, it is not possible for a light source to appear either brown or gray. [1]

[1] It is for this reason that Guinness failed to create a brown neon sign [8]. Signage for the Fukutoshin Metro Line in Tokyo is also affected: the train line is assigned the color brown, but signage often makes the line color appear orange or purple.

This phenomenon arises due to the human brain using some colors to characterize the reflectance properties of objects. To model the reflectance properties of an object, the spectral information of the light coming from the object *and* the light illuminating the object must be considered.

The description of color as tristimulus values, developed in the field of Colorimitry, is only adequate to describe human's perception of a single color in an otherwise dark environment. In attempting develop better models, the field of Color Appearance Models defined five perceptual dimensions that are necessary to specify a color appearance: hue, brightness, lightness, colorfulness and chroma. Consider the definition by Fairchild of brightness and lightness:

Brightness: Attribute of a visual perception according to which an area appears to emit, or reflect, more or less light.

Lightness: The brightness of an area judged relative to the brightness of a similarly illuminated area that appears to be white or highly transmitting.

The elaborate definition of lightness can be understood as relating to the perception of how much light a surface reflects. In the context of the five dimensions outline by Fairchild, gray can be understood to be separated from the color white along the dimension of lightness. Similarly, brown can transition to yellow, orange or red if it's lightness increases.

3.2 Related work

A number of studies have singled out brown for research. 1983's Fuld et al. "The Possible Elemental Nature of Brown" followed by Quinn et al.'s "Evidence that Brown is not an Elemental Color" helped establish brown's contentious position [9] [14]. In 2016, Steven Buck et al. published a flurry of work on brown [4] [6] [5] [12]. These studies all utilized a experiment setup of a yellow color being induced to brown by varying a background stimulus. The size, complexity and shapes of the center and surrounding stimulus were investigated in these experiments. The experiment presented in this chapter shares the same setup idea used in the above experiments.

Unlike the above experiments which investigated yellow and brown, the experiment in this chapter investigates the boundary between orange and brown. As noted in Morimoto et al., yellow transitions to brown through a color termed "butterscotch" [12]. The transition from orange to brown also encounters an

intermediate color where a binary classification is difficult; however, from experimentation not included in this report, this intermediate zone is narrower for the orange-brown transition in comparison to the yellow-brown transition. This narrower separation is hoped to strengthen the ability to establish an accurate linear model, which in turn can be used to establish higher lower-bounds for a neural network's classification accuracy.

This experiment in this work differs further through it's investigation of a range of variations to both hue and situation of the center stimulus in addition to the changes in relative brightness of the center and background stimuli. The extra space of colors places the dataset in 4 dimensions and allows the introduction of a third classification "neither" that is used to establish a boundary around the space of orange and brown colors.

A third major difference is this work's disregard for precise viewing conditions. With the goal of probing ImageNet trained image models, the lack of precise viewing conditions for the ImageNet dataset collection process reduce the benefit of establishing a precisely controlled viewing environment. In this regard, the works introduced above are far more thorough.

3.3 Experiment

The experiment in this chapter involves using a monitor to show a human participant (the author), a circle of a single RGB color against a background of a single RGB color. The participant then states whether they experience the circle to appear orange, brown, neither or both. This classification task is then repeated, each time changing the color of the circle and the color of the background.

The aim of this experiment is two fold:

1. Check that we are capable of eliciting the orange/brown perceptual phenomenon.

2. Gather some data that can be used to test neural networks.

3.3.1 Method

A web application was created to show a sequence of circles against a background with a single monochromatic color. The circles were positioned in the center of the screen with radius set to 30% of the height of a 1920x1080 resolution 24 inch Dell U2415 monitor, set to factory settings. The participant

viewed the monitor with eyes positioned 30 cm away from the center of the screen. Example stimuli are shown in Fig. 3.1. The monitor setup is shown in Fig. 3.2. While the experiment environment was not precisely controlled, it is nonetheless represents a real viewing condition, and one not unlike the environments of ImageNet labeling participants.

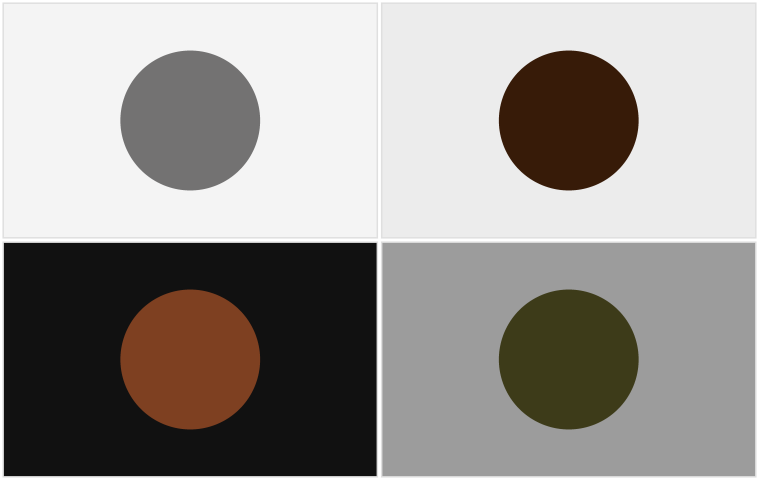


Figure 3.1: 4 separate stimuli examples. The appearance of these images printed in this manuscript will differ from their appearance as seen in the experiment.

The participant responded using keyboard arrow keys. The response was not time bound. There was a 2 second pause between stimuli, during which a black screen was shown. The pause was introduced to reduce the impact of one stimulus on the stimulus that followed. For some images, the circle neither appeared orange or brown and would be labelled “neither”. For other images the circle sufficiently orange and brown that a binary decision couldn’t be made, and would be labeled ”both”. In order to reduce the likelihood of errors, a speech API was used to create an audio confirmation of the participant’s classification choice after a key was pressed. In total 1517 classifications were made over two sittings, each lasting about 80 minutes. During each sitting, a 1-2 minute break was taken about every 10 minutes. Before the data was recorded, a shorter trial run was carried out. The data from the trial run was not included in the final dataset.

3.3.2 Results

The dataset created is a list of labeled points in 4 dimensions (3 dimensions for the circle color, and 1 dimension for the background value). The dataset is displayed in two dimensions in Fig. 3.3 and Fig. 3.4.

Figure 3.2: Data collection setup. This photo was taken during the day. The experiment was carried out at night, when the room was darker. It is interesting to note the monitor's backlight bleed, which is a factor likely to affect the experiment.

From the scatter plots in Fig. 3.4, it can be seen that rather than the circle's hue, it is the relation between the circle's brightness and the backgrounds brightness (the V values in HSV) that most clearly explain the classification results. This is evidence that we have recorded the phenomenon whereby orange transitions to brown based on comparative brightness with respect to surroundings. The results should make it clear how a 3 dimensional space is insufficient to describe human color perception. This dataset is used in the following two experiments.

3.3.3 Remarks

A number of observations were noted while collecting the data. Firstly, during the initial trial run, my perception of what constitutes brown or orange changed. As the trial continued, I became more aware of the distinction between browns and the colors that are described as "khaki". Similarly, I became more aware of the distinction between oranges and the colors that are described as "peach". I also became aware of my inability to label the circle color when it was sufficiently dark, despite it not appearing black. With these observations in mind, I decided to label any of these colors that are perceptually adjacent to orange and brown as the class "neither".

Even with this sharpened distinction, while carrying out the data collection,

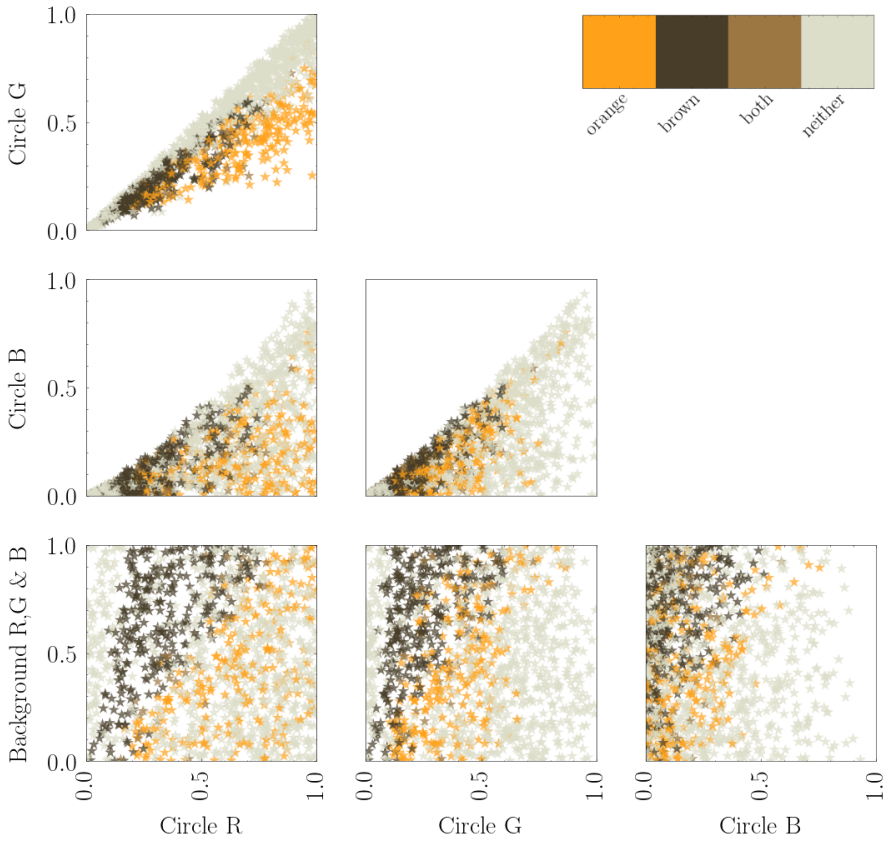


Figure 3.3: Circle color classifications. Each sub-figure plots one of the 4 dimensions of the dataset against another. The 4 dimensions being the R, G and B color components of the circle color, and the single dimension that determines the background RGB color. The data points are colored according to their classification.

there were many instances where it was very difficult to make a judgment call. If the data collection was to be repeated, I would break each of the 3 classes –orange, brown and neither–into the following 6 classes: "strongly orange", "weakly orange", "strongly brown", "weakly brown", "strongly neither" and "weakly neither". I think this granularity is more appropriate, and I think any further granularity is not useful. The nuance of the data collection emphasized the value of not averaging the results over many participants. It seems likely that different participants would come to decide on different color separations. Averaging over multiple participants might therefore suggest dynamics that exists in none of the participants.

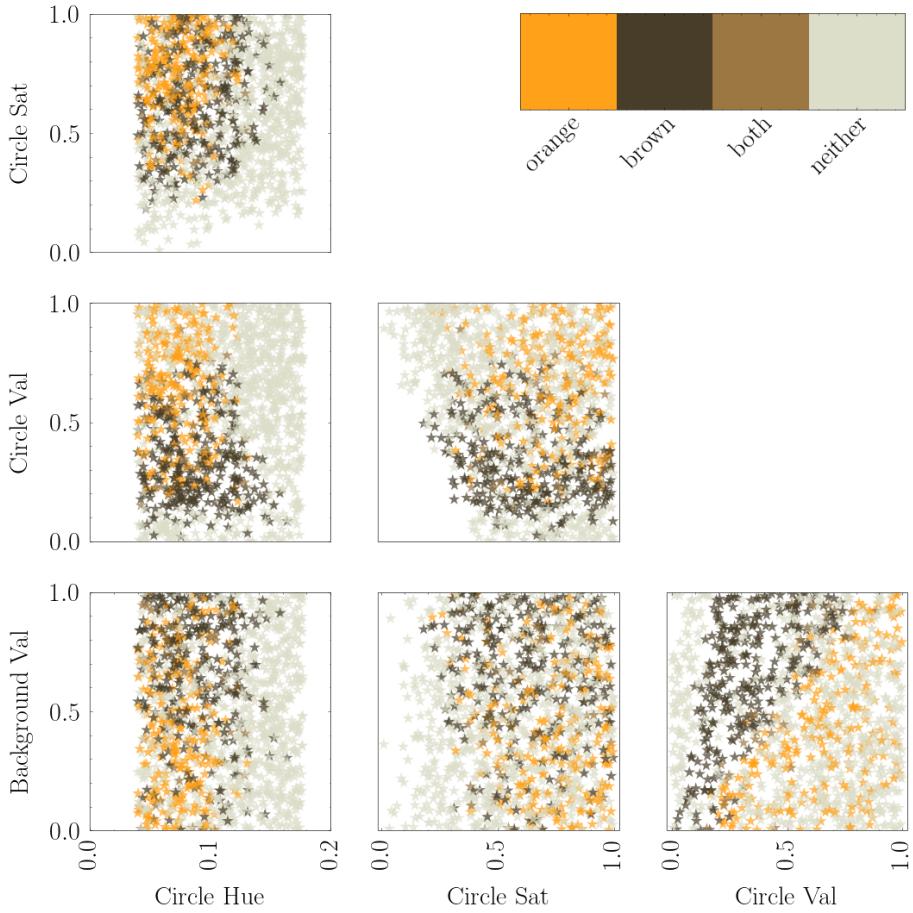


Figure 3.4: Circle color classifications. The circle and background colors have been transformed from RGB space to HSV space (hue-saturation-value). Each sub-figure plots one of the 4 dimensions of the dataset against another. The 4 dimensions being the H, S and V components of the circle color, and the single V dimension that determines the background color. The data points are colored according to their classification. Each HSV component has been normalized to the range $[0, 1]$.

3.4 Dataset splits

The dataset consisting of 1517 data points was split randomly into a train, validation and test set, with ratio 10:6:2, respectively. The small size of the dataset was considered and led to the test set being given a relatively large size. This dataset split was maintained for all subsequent experiments that used split data.

3.5 Model

In this section we investigate simple models to describe the data and to more precisely describe the distinction between orange and brown.

3.5.1 Orange vs. brown linear model

The hyperplane that best separates the orange and brown HSV data points with a mean-squared error (MSE) is given by the normalized vector and offset:

$$v = \begin{bmatrix} C_h \\ C_s \\ C_v \\ B_v \end{bmatrix} = \begin{bmatrix} 0.09 \\ -0.13 \\ -0.83 \\ 0.54 \end{bmatrix}, \quad b = 0.19$$

Ignoring the circle hue and saturation, and considering only the circle val and background val components, the line that best separates orange and brown HSV points with MSE is given by the vector and offset:

$$v = \begin{bmatrix} C_v \\ B_v \end{bmatrix} = \begin{bmatrix} -0.83, \\ 0.55 \end{bmatrix}, \quad b = 0.12$$

Thus, for viewing conditions of this experiment, we can roughly characterize the transition from orange to brown as being the point where the background HSV value component exceeds that of the center circle by 50%.

3.5.2 Orange vs. brown classification

Utilizing the train and test splits, train and test a binary classifier via logistic regression (in effect, a single perceptron with 4x1 weight matrix, length 1 bias vector, logistic loss function). The data was first transformed from RGB to HSV. The model achieves 96.6% orange-brown classification accuracy on the test set. The separation plane and zero-offset are defined by the vector and real:

$$v = \begin{bmatrix} C_h \\ C_s \\ C_v \\ B_v \end{bmatrix} = \begin{bmatrix} 0.464 \\ -0.752 \\ -6.261 \\ 3.945 \end{bmatrix}, \quad b = 1.460$$

The large values for the C_v and B_v weights corroborates the claim that, as

HSV color components, the circle V and background V components are most useful in separating orange and brown.

3.5.3 3 classes: orange, brown and neither

The classification is extended to include the third class, "neither", and a single layer neural network is trained on the RGB data via gradient descent of a softmax loss. In effect, the network is simply a 4x3 matrix and length 3 bias vector. We achieve a test set classification accuracy of 73.6%. Removing the background color from the input data causes the accuracy to drop to 71.7%, and instead removing the circle color and maintaining the background color causes the accuracy to drop to 57%. The distinction between the three classes is not expected to be effectively achieved by using a 4x3 perceptron; when the 4D data is reduced to 3D via principle-component-analysis, the data appears to be two adjacent clusters comprising the orange and brown data, and these clusters are surrounded by a hollow sphere of "neither" data points. Consequently, it is not expected that three hyperplanes could effectively be used to classify the data into the three classes. The motivation for inspecting this classification approach is to establish a lower bound on what can be considered a satisfactory classification accuracy. When investigating pre-trained neural networks in the next chapter, it is useful to keep in mind the 73.6% accuracy achieved by the single single layer network.

Chapter 4

Orange and brown in neural networks for vision

With the phenomenon of contextual colors in mind, it is interesting to ask whether neural networks trained on vision tasks make use of similar encodings. If they don't will there be conditions under which they perform poorly, or at least differently, compared to humans?

4.1 Choice of experiment

The high dimensionality of information flowing between neural network layers in a model such as ResNet means that when considering the search for a latent encoding that might represent a contextual color such as orange, both searching to confirm the presence or searching to confirm the absence of such an encoding is difficult. The strategy taken in this work is to focus on the layers towards the very end of the network. This strategy takes advantage of the low resolution of activations in later layers.

Without carrying out any experiments, it is possible to place some limits on what can be expected of earlier layers of the network. The next two sections explore this idea.

4.2 The boundary of orange-brown

To elicit the perception of brown in humans, an otherwise orange or yellow area must be surrounded by a color of sufficient relative brightness. As shown by DeLawyer et al., the boundary between the target area and it's immediate surroundings are insufficient alone to determine the perceived color, as the

brightness from more distant areas of an image can override any effect of the target's immediate boundary. [6] Fig. 4.1 demonstrates this effect. From experiments not recorded here, depending on the brightnesses of the 3 areas, the size of the dark middle ring can can extend a considerable extent around the center circle and still fail to induce the perception of orange, despite inducing orange in the absence of the white background.

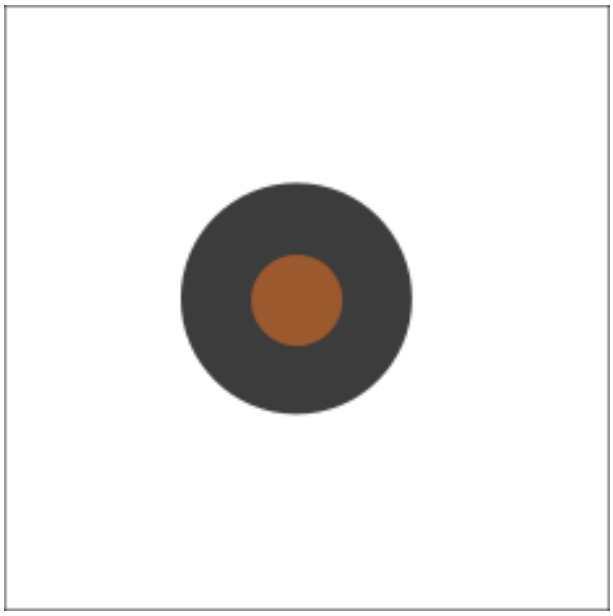


Figure 4.1: The white surround is able to induce the center circle to be perceived as brown, dispite the dark middle ring having a relative brightness that would otherwise induce orange.

This phenomenon when considered with respect to neural network receptive fields, places limits on where in the network a representation that distinguishes orange from brown could be found. This idea is considered in the next section.

4.3 Receptive fields

The *receptive field* of an activation in some part of a neural network is defined as the region of the input that can contribute to determining this activation. For convolutional neural networks, the height and width of the receive field grows as one considers activations further from the input image. When considering the receptive field of the final layer, a positive relationship between recetive field size and ImageNet top-1 accuracy has been observed. [2] This relationship must be viewed with scrutiny; however, as it is difficult to disentagle receptive field sizes from parameter count or network depth. For

the purposes of this work, receptive field sizes place restrictions on the size of an area in the input image that can be identified as either brown or orange. For a given activation in the network, the size of an orange/brown area must be smaller than the receptive field of an activation in order for the activation to be able to classify the color. As explained in the previous section, in some circumstances, it may be necessary for the receptive field to be considerable larger than the target area in order to make an accurate distinction between contextual colors such as orange and brown. For ResNet50, the activations at roughly 1/4, 2/4, 3/4 and 4/4 through the network have a receptive field of approximately 35, 99, 291 and 483. [2] For ResNet18, the equivalent field sizes are approximately 43, 99, 211 and 435.

The issue of the size of receptive fields in comparison to the size of an orange/brown stimulus raises many questions as to how small or big, or with what resolution it is possible for a layer in a neural network to be capable of making an orange-brown distinction.

With these restrictions already in place for earlier layers of a convolutional neural network, the main experiment of this chapter proceeds to focus on the very *last* layer of the ResNet architecture. With the resolution of this layer's output being just 1x1, it is more feasible to investigate if sufficient information is present in this layer to distinguish between a simple orange or brown stimulus.

4.4 Testing the final layer

This experiment seeks to determine if sufficient information reaches the output end of the ResNet50 model used in Chapter 2 in order to distinguish orange and brown. The dataset created in chapter 3 is used to design a classification task, on which the pretrained ResNet50 model will have it's last layer retrained while all other weights are kept fixed.

The two next section, 4.4.1 and 4.4.3, covers the setup in more detail. The results are presented in section 4.4.4 and a discussion follows in section 4.4.5.

4.4.1 Orange-brown dataset

From the orange-brown dataset created in Chapter 3, a new dataset of labeled *images* was created. For every labeled 4D data point from the orange-brown dataset, a 244x244 RGB image array was created and given the same label. In effect, an image was created that was similar to that shown to the

participant who labeled the orange-brown dataset from chapter 3. Only the classes "orange", "brown" and "neither" were included. The very few "both" data points were excluded due to how few there were. Thus, this dataset implies an image classification task with 3 classes.

The correspondence between the two datasets deserves more attention. The data collection process in Chapter 3 kept the circle and background sizes constant. In addition, viewing distance was kept constant. All three variables can affect the elicited circle color.[5] [17]. The pretrained networks that will be investigated in this chapter were trained for classification on the ImageNet dataset. The data collection process for this dataset did not control for aspects of viewing conditions such as viewing distance, background illumination, screen resolution or monitor properties. This raises the question as to what size is appropriate to set for the foreground circle; the model might achieve better or worse results depending on the circle size. To address this issue, instead of a single dataset being created, multiple datasets will be created—each a variation that differs by the size and position of the foreground circle. 28 different circle radii and 14 different circle positions are tested, resulting in a total of 392 separate datasets, each with 1517 image-label pairs. The circle radii range from 6 to 168 pixels in steps of 6, with an additional radius at 1 pixel. The first circle position is placed at the center of the image, and subsequent circle positions are offset from the center along the diagonal towards the bottom right in steps of 8, with a maximum offset of 112 pixels. The variation in circle parameters is demonstrated in Fig. 4.2.

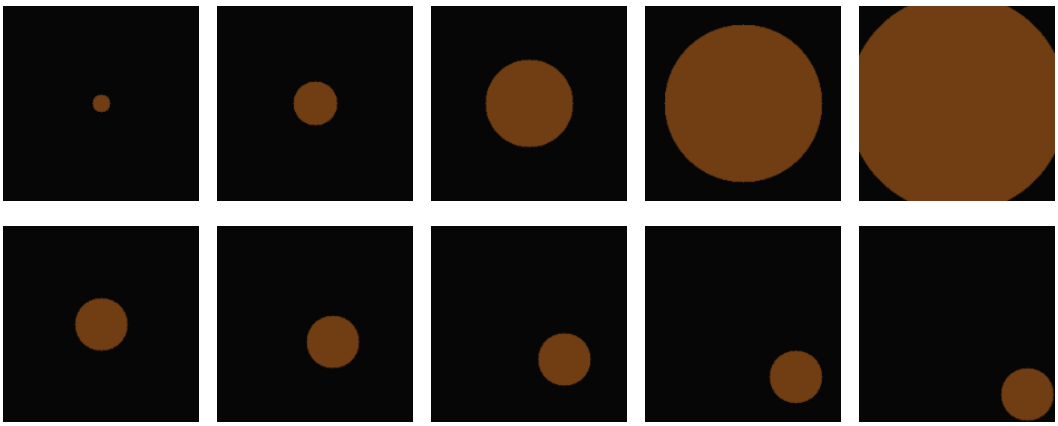


Figure 4.2: 8 samples; 1 sample each from 8 dataset variations (8 out of 392). *Top*: variation in center circle size. *Bottom*: variation in distance from the center.

4.4.2 Red-green dataset

To place results in perspective, a second set of datasets were created. This second set was designed to duplicate the data described in the previous section, the only difference being the 3 colors being classified. Instead of orange, brown or neither, the classes are red, green and neither. These colors cannot be are far less dependent on the background illumination. The aim of this dataset is to create a classification problem that is easier in the sense that the classification can be achieved based on the circle color alone, without regard to the comparative brightness's of the circle and background. The dataset generation process was to generate a random color and bin the color according to the ranges in Table 4.1. Colors were ignored if either they did not match any of these categories or if the category had collected sufficient samples. The color counts were chosen to match those of the orange-brown-neither data. The ranges in Table 4.1 were chosen based on experimentation in the same viewing conditions as in Chapter 3.

Color class	Hue range	Sat range	Val range	Count
red	$\geq$350, or [0, 8]	$\geq$ 50	$\geq$ 30	277
green	[70, 140]	$\geq$ 50	$\geq$ 50	306
neither	[18, 60] or [150, 340]	$\geq$ 50	$\geq$ 30	891

Table 4.1: Ranges for HSV hue sat and val components of the red, green and neither classes. The hue, sat and val components are not normalized to [0, 1], and instead are listed with the standard ranges [0, 360], [0, 100] and [0, 100] respectively.

Samples from one of the red-green-neither datasets are shown in Fig. 4.3.

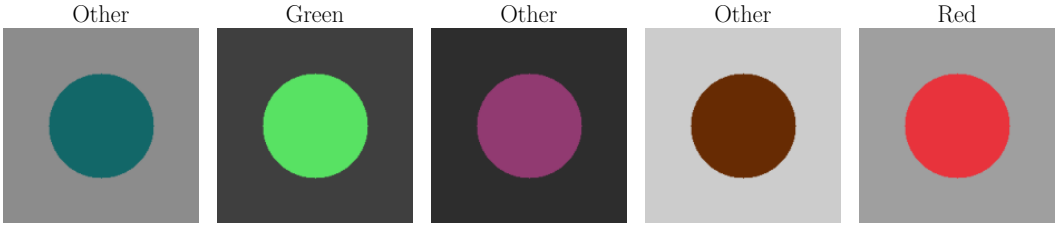


Figure 4.3: 5 samples from 1 dataset variation (centered circle with radius 60 pixels).

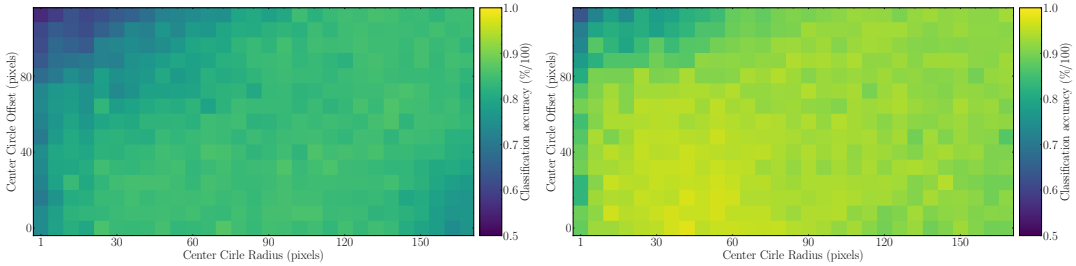
4.4.3 Method

For the two ResNet models, and for all of the 2x392 datasets (both orange-brown and red-green) a standard classification experiment was run. A model

was trained using the training split and tested on the testing split. The validation split was used to select the final model to be tested after 20 epochs. The final fully-connected layer of the model was trained via gradient, while all other weights of the model were kept fixed.

4.4.4 Results

For ResNet50, the average accuracy classification accuracy over all 392 orange-brown runs is 81% compared to 98% for the red-green runs. The accuracy details are displayed in more detail in the heatmaps of Fig. 4.4. The results for ResNet18 are similar are are included in Fig. A.3.



(a) Orange-brown datasets

(b) Red-green datasets

Figure 4.4: *Left*:Accuracy for each of the 392 orange-brown dataset variations, as a heatmap. The average and variance over all runs is 0.81 and 0.003. *Right*: Accuracy for each of the 392 red-green dataset variations, as a heatmap. The average and variance over all runs is 0.977 and 0.002.

4.4.5 Discussion

As can be seen in 4.4 (and Fig. A.3), the retrained ResNet50 architecture (and ResNet18) is more capable at distinguish the colors in the red-green dataset compared to the colors in the orange-brown dataset. A shift from a red to green circle produces a change in the final layer feature vector from which a linear correlation to the circle color can be established (to a greater degree than the orange-brown case). This hints at the possibility that there is less information available to distinguish orange from brown compared to information available to distinguish green from red. Two caveats must be mentioned before such a statement can be strengthened. Firstly, a non-linear classifier (such as multi-layer perceptron) may be able to classify the orange, brown neither classes with high accuracy. Secondly, the red-green-neither dataset may have some accidental quality making it easy to classify. Proceeding to test

other color pairs, such as colors which are closer in hue, may help to address this concern.

Chapter 5

Grayworld illumination prior

Chapter 2 and Chapter 4 both highlight between human vision and the ResNet models investigated. Chapter 2 highlighted specific modes of failure, such as failure to classify under strong blue light. While an approach to mitigate the issue was suggested, the approach had many drawbacks. In this section an alternative attempt is outlined to address the poor illumination invariance. The idea of applying the grayworld algorithm is the same; however, this time, instead of preprocessing the input to undo the illumination, we modify the ResNet. The alternation to ResNet is described in Section 5.1. This investigation is currently ongoing, so results are not available to include included in this report.

5.1 ResNet modification

The forward pass of ResNet is modified as shown in 5.1. Before the first layer of the ResNet architecture, the mean of the RGB values of the input were calculated. The input image, of dimension 224x224x3 was then expanded to 224x224x6, with each of the extra 3 dimensions being filled by one of means just calculated. Thus, for one of the extra dimensions, all of it's values are equal. The 224x224x6 tensor is then passed to the first layer as before. The code change is carried out on the Pytorch implementation of ResNet and is shown in the code listing 5.1.

Listing 5.1: Modification to the beginning of the Pytorch implementation of the ResNet architecture found in the Pytorch Image Models (TIMM) Python package. Three extra dimensions are added before the first trainable layer. Each extra dimension is filled with a single value, which is the average of one of the 3 RGB channels of the input.

```
def forward_features(self, x):
  # [modification start]
  mean = torch.mean(x, dim=(2, 3), keepdim=True)
  mean_as_channel = mean.expand(-1, -1, *x.shape[2:])
  x = torch.cat((x, mean), dim=1)
  # [modification end]
  x = self.conv1(x)
  x = self.bn1(x)
  x = self.act1(x)
  x = self.maxpool(x)
  x = self.layer1(x)
  x = self.layer2(x)
  x = self.layer3(x)
  x = self.layer4(x)
  return x
```

5.2 Illumination and receptive fields

The hypothesis behind the modification to the ResNet model is that, as explained in Section 4.3, it is many layers into the ResNet model before an activation has a receptive field large enough to make accurate predictions about the scenes illumination. Chapter 2 demonstrated how awareness of the whole image per-channel RGB means allows for a simple yet effective illumination estimation. Consequently, providing this information to the beginning of the network may enable the earlier layers of the model to account for scene illumination despite a narrow receptive field.

Chapter 6

Conclusion

This work identified differences rather than similarities between human and machine vision. The effect on classification accuracy of illumination on two ImageNet trained ResNet models was demonstrated. Evidence for the comparative difficult this model has in distinguishing orange and brown was also presented. Both of these issues have a connection to the human vision phenomenon such as chromatic adaptation and contextual colors, where a single point in our visual field is affected by a larger area of the visual field. It is hypothesized that giving summary information about the whole image, or a large part of the image, to the input of a convolutional neural network may help the network improve it's performance. This might also help the network more easily represent images with human-like representations.

Bibliography

[1] Anastasiia Vlasiuk and Hiroki Asari. Feedback from retinal ganglion cells to the inner retina. *PLOS ONE*, 16(7), July 2021. 5

[2] A. Araujo, W. Norris, and J. Sim. Computing receptive fields of convolutional neural networks. *Distill*, 2019. 24, 25

[3] J. T. Barron. Convolutional Color Constancy. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 379–387, Dec. 2015. 5

[4] S. L. Buck and T. DeLawyer. A new comparison of brown and yellow. *Journal of Vision*, 12(14):9–9, Dec. 2012. 15

[5] S. L. Buck, A. M. Shelton, B. Stoehr, V. Hadyanto, M. Tang, T. Morimoto, and T. DeLawyer. Influence of surround proximity on induction of brown and darkness. *Journal of The Optical Society of America A-optics Image Science and Vision*, 33(3), Mar. 2016. 15, 26

[6] T. DeLawyer, T. Morimoto, and S. L. Buck. Dichoptic perception of brown. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 33(3):A123–128, Mar. 2016. 15, 24

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 5

[8] M. D. Fairchild. *Color Appearance Models*. John Wiley & Sons, Aug. 2013. 4, 14

[9] K. Fuld, J. S. Werner, and B. Wooten. The possible elemental nature of brown. *Vision Research*, 23(6):631–637, Jan. 1983. 15

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 5

[11] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2373–2382, Honolulu, HI, July 2017. IEEE. 5

[12] T. Morimoto, E. Slezak, and S. L. Buck. No effects of surround complexity on brown induction. *Journal of The Optical Society of America A-optics Image Science and Vision*, 33(3), Mar. 2016. 15

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[14] P. C. Quinn, J. L. Rosano, and B. R. Wooten. Evidence that brown is not an elemental color. *Attention Perception & Psychophysics*, 43(2):156–164, Feb. 1988. 15

[15] A. Shapiro, L. Hedjar, E. Dixon, and A. Kitaoka. Kitaoka's Tomato: Two Simple Explanations Based on Information in the Stimulus. *i-Perception*, 9:204166951774960, Feb. 2018. v, 4

[16] A. Smagina, E. Ershov, and A. Grigoryev. Multiple light source dataset for colour research. *arXiv preprint:1908.06126*, 2019. 6

[17] H. Uchikawa, K. Uchikawa, and R. M. Boynton. Influence of achromatic surrounds on categorical perception of surface colors. *Vision Research*, 29(7):881–890, Jan. 1989. 26

[18] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 547–558, New York, NY, USA, Jan. 2020. Association for Computing Machinery. 5

Appendix A

Additional figures

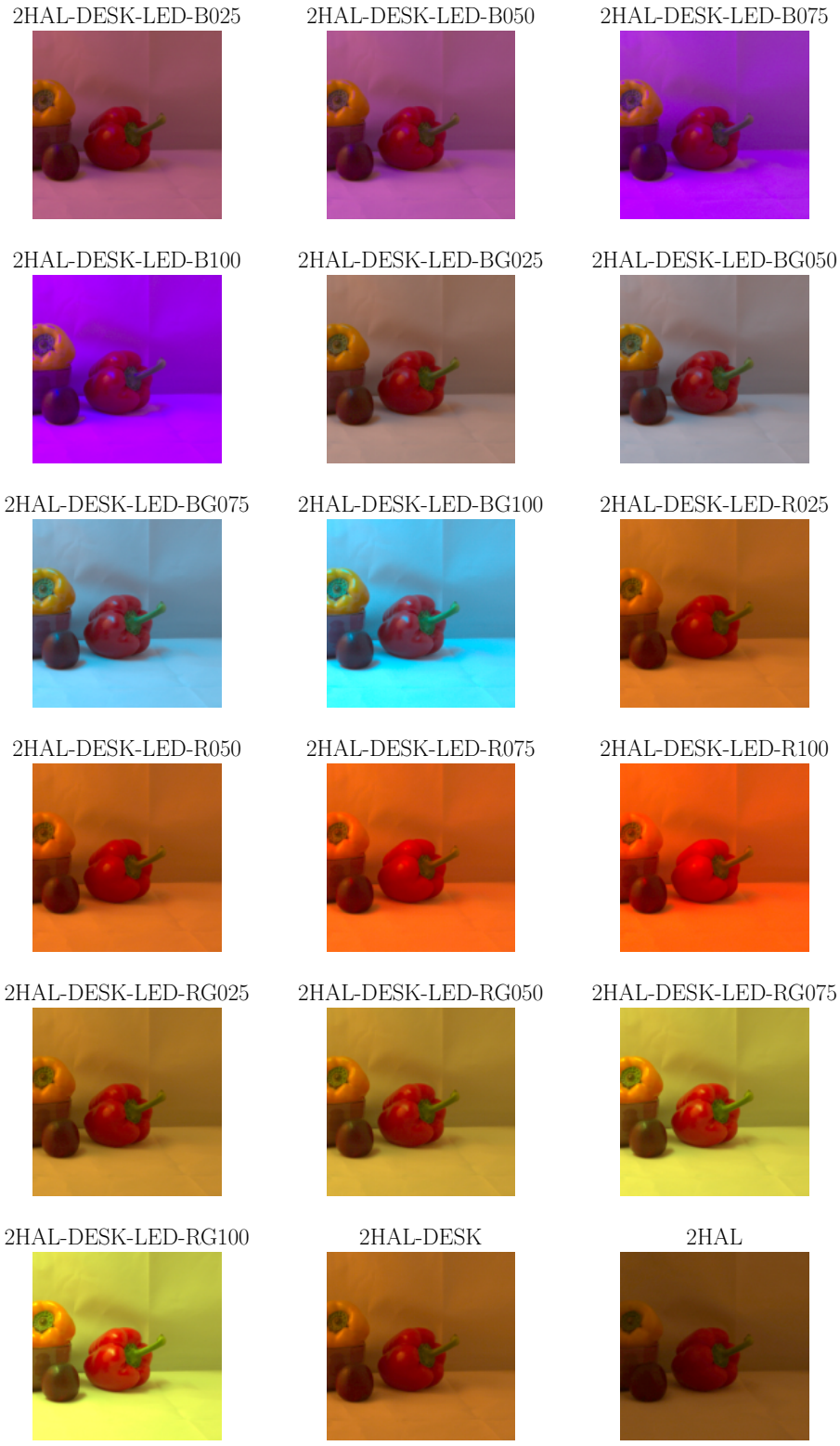


Figure A.1: All 18 illuminations for the capsicum scene, with illuminants labeled.

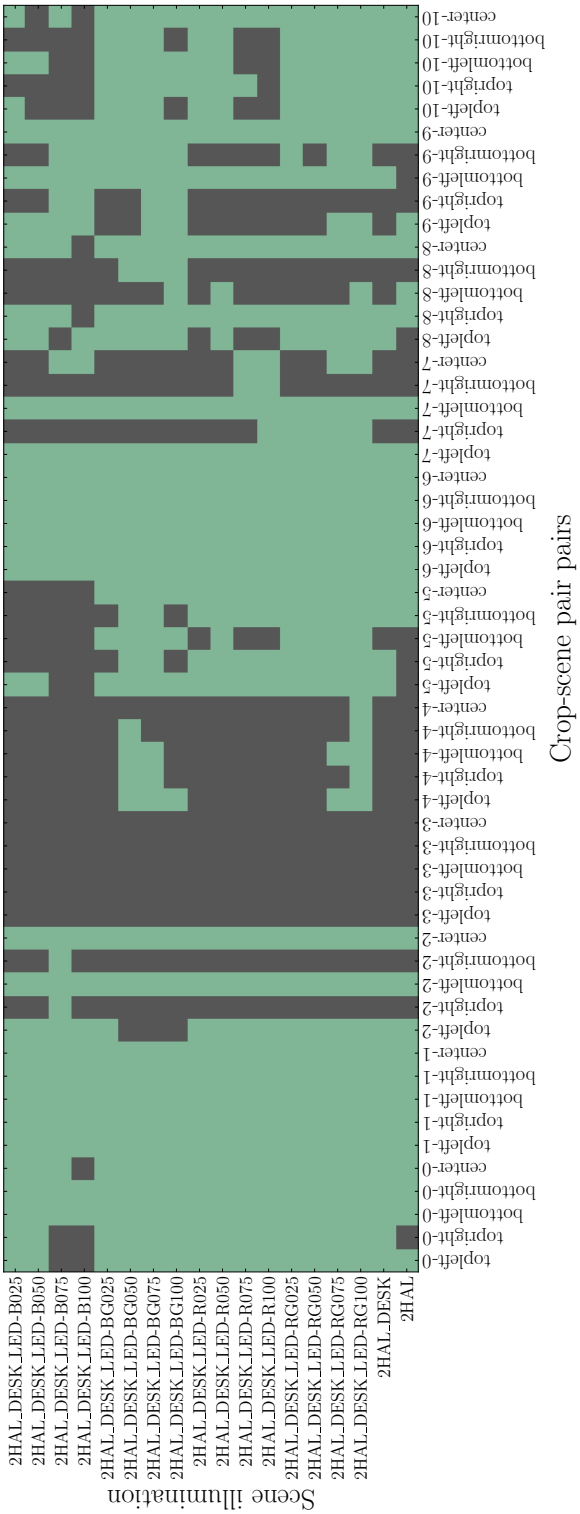
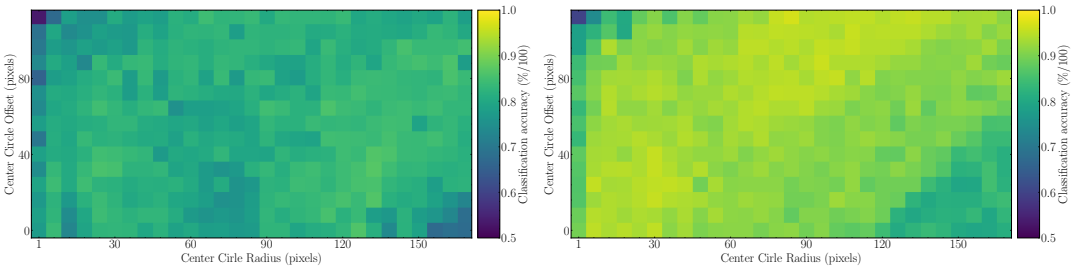


Figure A.2: Classification results for for every scene-crop-illumination triplet. **Green: correct classification, black: incorrect classification**. The network is ResNet50. Illumination is given the whole y-axis, while the scenes and crops are flattened onto the x-axis.



(a) Orange-brown datasets

(b) Red-green datasets

Figure A.3: ResNet18 results, equivalent to ResNet50 results in Fig. 4.4. *Left*:Accuracy for each of the 392 orange-brown dataset variations, as a heatmap. The average and variance over all runs is 0.81 and 0.001. *Right*: Accuracy for each of the 392 red-green dataset variations, as a heatmap. The average and variance over all runs is 0.967 and 0.002.