Master's Thesis

Investigating Evidence for Hierarchical Color Learning in Convolutional Neural Networks

콘볼루션 신경망의 색상 위계 학습에 대한 탐구

2020 년 8 월

서울대학교 대학원

협동과정 인지과학전공

Chris Hickey

Investigating Evidence for Hierarchical Color Learning in Convolutional Neural Networks

콘볼루션 신경망의 색상 위계 학습에 대한 탐구

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2020 년 7 월

서울대학교 대학원

협동과정 인지과학전공

Chris Hickey

Chris Hickey의 공학석사 학위논문을 인준함

2020 년 6 월

위 원 장 ______김 홍 기______ (인)

부위원장 ______장 병 탁______ (인)

위 원 ______김 청 택______ (인)

Abstract

Investigating Evidence for Hierarchical Color Learning in Convolutional Neural Networks

Chris Hickey
Interdisciplinary Program in Cognitive Science
The Graduate School
Seoul National University

Empirical evidence suggests that color categories emerge in a universal, recurrent, hierarchical pattern across different cultures in the following order; *white, black* $<$ *red* $<$ *green, yellow* $<$ *blue* $<$ *brown* $<$ *pink, gray, orange, and purple.* This pattern is referred to as the "Color Hierarchy". Over two experiments, the present study examines whether there is evidence for such hierarchical color category learning patterns in Convolutional Neural Networks (CNNs). Experiment A investigates whether color categories are learned randomly, or in a fixed, hierarchical fashion. Results show that colors higher up the Color Hierarchy (e.g. red) are generally learned before colors lower down the hierarchy (e.g. brown, orange, gray). Experiment B examines whether object color affects recall in object detection. Similar to Experiment A, results show that object recall is noticeably impacted by color, with colors higher up the Color Hierarchy generally showing better recall. Additionally, objects whose color can be described by adjectives that emphasise colorfulness (e.g. vivid, brilliant, deep) show better recall than those which de-emphasise colorfulness (e.g. dark, pale, light). The effect of both color hue and adjective on object recall is still observable, even when controlling for contrast through grayscale images. These results highlight similarities between humans and CNNs in color perception, and provide insight into factors that influence object detection. They also show the value of using deep learning techniques as a means of investigating cognitive universalities in an efficient, unbiased, cost-effective way.

Keywords: color hierarchy; computer vision; convolutional neural network; cognitive universalities; faster R-CNN

Student Number: 2018-29483

Contents

List of Tables

List of Figures

Chapter 1

Introduction

How many colors are in the rainbow? In reality, the color solid is divisible into millions of infinitesimally small differences. However, out of an estimated 7,295,000 visible differences in the psychological color solid (Newhall, Nickerson, & Judd, 1943), most human languages use no more than 13 basic terms to conceptualize this vast color spectrum (Kay, Berlin, Maffi, Merrifield, & Cook, 2009). For example, an English speaker will subsume color differences into 11 basic color categories; white, black, red, green, yellow, blue, orange, brown, gray, pink, and purple. There are three physiological dimensions to color (see Figure 1.1). These are:

a) Hue The dominant wavelength

b) Saturation The dominance of the achromatic wavelength for a given hue

c) Brightness (or Luminance) The degree or intensity of lightness / darkness

However, differentiation of color categories occurs along the most salient

Figure 1.1: Three Physiological Dimensions of Color a) Hue b) Saturation c) Brightness

dimension of color: hue (i.e. red, yellow, blue) (Wyszecki & Stiles, 1982).

1.1 Is Color Categorization Random?

The emergence of distinct hue-based color categories presents a paradigmatic challenge across cognitive science and linguistic disciplines (Lakoff, 2008). This is mainly due to difficulties delineating the unique inter-play between perception, conceptualization, and language features involved in color categorization. As of yet no consensus theory has emerged to explain why these discrete categories of the continuous hue wavelength emerge. However, competing theories in this decades-old debate have largely centred around two schools of thought: Relativism versus Universalism. Relativist theorists argue that color categories are not innate or pre-ordained. Rather, they are based on culture, and therefore, the product of environmental factors and relatively idiosyncratic influences that differ across time (Casson, 1997) and linguistic groups (Wierzbicka, 2008).

Universalist theorists, on the other hand, claim that color contains inherent qualities that result in enduring, non-random patterns, consistently evident across environments and across time. One of the most enduring Universalist theories on how these hue-based categories emerge was first proposed by Brent Berlin and Paul Kay half a century ago (Berlin & Kay, 1969). Their cross-cultural research observed a fixed sequence according to which languages gain color terms over time. To account for this evolution, Berlin and Kay (1969) proposed the following two conjectures:

- there exists a limited set of “universal” cognitively-hardwired categories from which all languages draw their color lexicons, and
- languages “evolve” by adding color terms in a relatively fixed sequence, such that; *white, black* $<$ *red* $<$ *green, yellow* $<$ *blue* $<$ *brown* $<$ *pink,*

gray, orange, and purple

Therefore, if a language has a term for a given color in this inequality, that language will also likely have terms for all colors to the left of that term. This sequence has subsequently been referred to as the *"Color Hierarchy"* (See Figure 1.2).

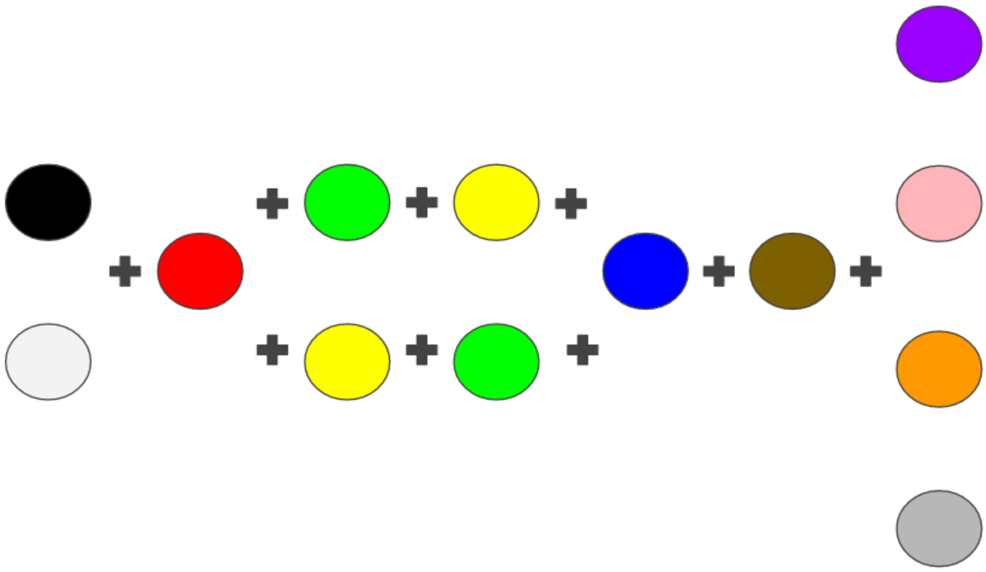


Figure 1.2: Color Hierarchy. In a language, if a term exists for a color in this figure, terms will also exist for all colors to the left of that color term.

Berlin and Kay's initial hypothesis has gained strong support through *'The World Color Survey'* (Kay et al., 2009). This survey constituted a long-term study in which color names were obtained from informants of mostly unwritten languages spoken in pre-industrialized cultures that have had limited contact with modern, industrialized society. Through this survey, the evolution of color terms along the *Color Hierarchy* can be observed. For example, languages such as the Bolivian Amazonian language Tsimane' have only three words that categorize color. These correspond to black, white and red, the first three colors in the Hierarchy (Gibson et al., 2017). Languages with two additional terms, like Papua-New Guinean Berinmo, will generally have these additional terms correspond to the next two or three terms in the Hierarchy. In the case of Bermino, these two additional colors roughly correspond to green and blue (Kay & Regier, 2007).

Research into hierarchical preferences for certain colors extends beyond the emergence of color terms in language. Skelton, Catchpole, Abbott, Bosten, and Franklin (2017) found that when pre-verbal infants learned to categorize colors, they recognize hue categories corresponding to lexical terms towards the top of the *Color Hierarchy* fastest. Also, Tchernikov and Fallah (2010) found a similar hierarchical pattern in target selection, where participants presented with multiple moving targets on a screen, eye-tracked targets according to the following hierarchical color based bias; *red* > *green* > *yellow* > *blue*.

1.2 Modelling the Color Hierarchy

Efforts have been made to model the emergence of hierarchical color categorizations (Baronchelli, Gong, Puglisi, & Loreto, 2010; Loreto, Mukherjee, & Tria, 2012). Specifically Loreto et al. (2012) demonstrate that an approximation of Berlin and Kay's original Hierarchy emerges in a model where computational agents have human perceptual discrimination of color regions via a "Just Noticeable Difference" function. These agents agree on terms for colors on the color spectrum via a "Category Game". They found that the time needed for a population of artificial agents to reach consensus on a color name depends on the region of the hue in the overall color spectrum. These computational agents reached consensus on color terms from the spectrum in an order somewhat similar to Berlin and Kay's original color hierarchy; *red* $<$ *(magenta)-red* $<$ *violet* $<$ *green* $<$ *blue* $<$ *orange* $<$ *cyan.*

1.3 Convolutional Neural Networks and Color Learning

While various attempts have been made at modelling the emergence of color categories (also see Jameson and Komarova (2009)), few attempts have been made to investigate what contribution Convolutional Neural Networks (CNNs) can make to this question. CNNs have long been proposed as suitable frameworks to model biological vision (Rafegas & Vanrell, 2018). Deep convolutional feedforward networks for object recognition are not biologically detailed. Rather, they rely on nonlinearities and learning algorithms that may differ from those of biological brains. Nevertheless they learn internal representations that are highly similar to representations in the human and nonhuman primate Inferior Temporal cortex, indicating that CNNs may provide at least a partially faithful tool for modelling biological vision (Kriegeskorte, 2015).

In addition, CNNs have shown similarities to humans in the domain of color perception. Researchers have analysed trained CNN networks using physiologically inspired methodologies to understand how neurons codify color throughout different convolutional layers (Rafegas & Vanrell, 2017, 2018). Results found color representations for single hue colors, opponent colors and color shape entanglements for specific objects across all convolutional layers in a trained CNN. These representations, most notably the single and opponent selective neurons, show similarities to how color is encoded through the visual pathways in biological visual systems (Lim, Wang, Xiao, Hu, & Felleman, 2009; Conway, 2001).

Do neural networks faithfully imitate the human process and learn colors in an analogous, hierarchical way? Or rather, will neural networks learn color labels in a non-interpretable random way? As discussed above, parallels have already been drawn between the human visual system and CNNs in the domains of internal object representations (Kriegeskorte, 2015) and color perceptions (Rafegas & Vanrell, 2018). If hierarchical color learning patterns are observed in CNNs, this would serve as additional evidence to suggest that CNNs can offer at least partial fidelity to biological vision modeling. Additionally, if object detection models were shown to be influenced by color, this could shed valuable insight into factors that influence object detection performance.

1.4 Hypotheses

Over two experiments, the present study examines whether there is evidence for hierarchical color category learning patterns in CNNs. The study anticipates that CNNs can partially model human biological processes. Specifically, it anticipates that CNNs will show color learning that is different from what we would expect if it were truly learning colors randomly. By randomness, the present study means the network will not learn colors in any meaningfully interpretable way. This forms the basis of the core hypotheses under investigation. If there are no similarities between humans and neural network mechanisms, one would expect there to be no pattern in how CNNs learn color:

Null Hypothesis 1 *CNNs learn colors in a stochastic, random manner.*

However, following from previous research that finds parallels between CNNs and aspects of human visual perception, if we see noticable hierarchical patterns of color learning, the present study will support the alternative hypothesis:

Hypothesis 1 *CNNs learn colors according to consistent hierarchical patterns, analogous to the hierarchical color patterns seen in humans.*

Chapter 2

Datasets

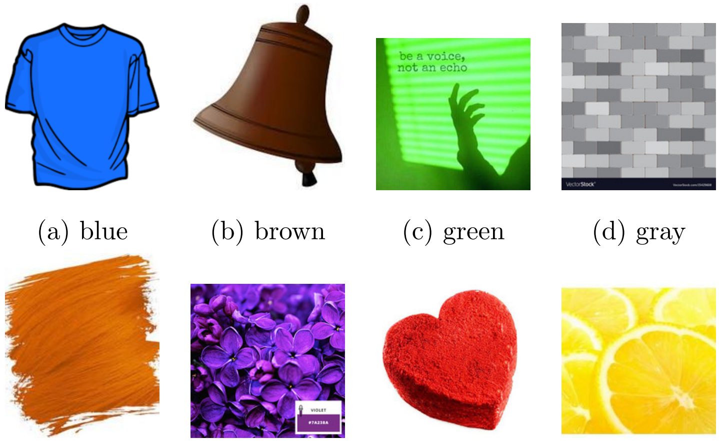
Two datasets were used in the present study. Experiment A used the *Basic Color Dataset*, while Experiment B used the *Modanet* dataset with added color annotations.

2.1 Basic Color Dataset

The basic color dataset was made up of 880 images taken from Google images. Images were gathered by searching for specific colors, (e.g "red") and downloading images of either objects, wallpapers or backgrounds that best encapsulate this color. The total dataset was divided into a training dataset and a test dataset. The training dataset was comprised of 60 images for each of the eight colors. The test dataset was comprised of 50 images for each of the eight colors. For samples from the basic color dataset, please refer to Figure 2.1.

2.2 Modanet

Zheng, Yang, Kiapour, and Piramuthu (2018) created a large scale street fashion dataset with polygon annotations, containing 55,176 images. This dataset contains bounding box, polygon segmentation details, and labels for 13 clothing categories, namely: **bag, belt, boots, footwear, outer** (coat, jacket etc.),



(a) blue (b) brown (c) green (d) gray

(e) orange (f) purple (g) red (h) yellow

Figure 2.1: Samples from Basic Color Dataset used in Experiment A

dress, sunglasses, pants, top, shorts, skirt, headwear, scarfs (See Figure 2.2). Images in this dataset were given a train/test split of roughly 60/40, with image names ending in 0, 1, 3, 5, 7 and 9 being added to the training set, and the remaining images added to the test set.

2.2.1 Color Annotating Process

We then added color attributes to each of the objects in the dataset as follows: firstly, semantic segmentation was performed on each object. Then each item of clothing was assigned a color label from the ISCC-NBS color labelling system.

ISCC-NBS labels are a taxonomization of the Munsell color system developed by the Inter-Society Color Council (ISCC) and National Bureau of Standards (NBS) (Kelly & Judd, 1955). The purpose of this color naming scheme was to develop a *"designation to be sufficiently standardized as to be acceptable and usable by science, sufficiently broad to be appreciated and used by science, art, and industry, and sufficiently commonplace to be understood, at least in*

Figure 2.2: Sample images from the Zheng et al. (2018) Modanet dataset. The top row shows the original images. The second row shows polygon segmentation for each clothing item in the top row

a general way, by the whole public". In order to be intuitive to people, the ISCC-NBS system uses only common English colors and descriptive adjectives, excluding less commonly used color terms like teal, fuchsia, etc. Basic hues like red, yellow, blue or green are combined with descriptive adjectives like "brilliant", "vivid", "dark" or "pale" to produce 267 color categories; some examples are "light greenish gray" and "deep blue". Each category is a well-defined subset of the Munsell color system (Munsell, 1941), and every Munsell color belongs to a unique ISCC-NBS category.

In order to assign each item of clothing from the dataset a color label, the polygon information for each item of clothing was used to sample 500 random RGB pixels for each clothing item. The ISCC-NBS color label that was sampled

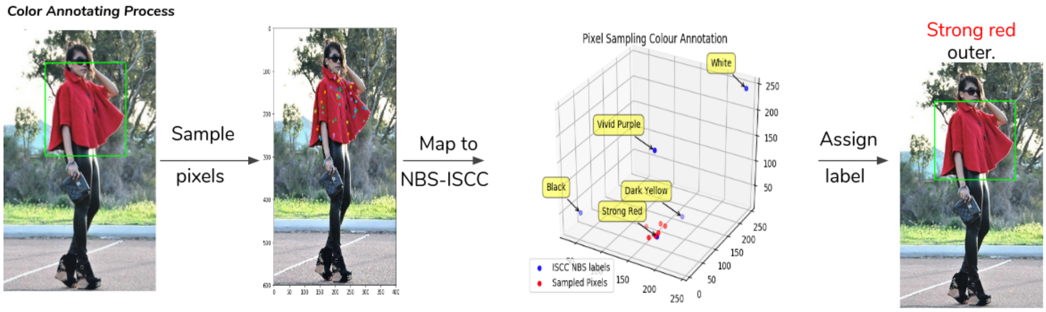


Figure 2.3: Samples of color attributes assigned to objects in the Modanet dataset using pixel sampling.

most often from each clothing polygon was then assigned as the label for that object. If no color label emerged for at least 10% of the labels for an object, that object was assigned a *null* color label. This process is outlined in Figure 2.3.

Chapter 3

Color Space

A color space is essentially an abstract mathematical model used to organise and describe colors through numbers. Color spaces are most commonly used to help us reproduce analog and digital representations of color.

Computer vision has improved rapidly over recent decades, notably in the domains of image classification and object detection. Most of the datasets used for image classification and object detection tend to be color images, represented through the default RGB color space. Most recent models developed for classification tend not to perform a color space transformation to the image, instead using the RGB or BGR[1] image directly for classification or object detection. In addition to RGB and BGR, the present study also investigates alternative color spaces inspired by aspects of human vision, as outlined below.

3.1 Opponent Color Space

Opponent color spaces are a family of physiologically-motivated color spaces inspired by the human visual system (Plataniotis & Venetsanopoulos, 2013). This inspiration stems from a theory that the human visual system perceives colors in terms of opponent hues, specifically yellow-blue and red-green (Hurvich & Jameson, 1957). Activation of one end of the axes inhibits activity in the

[1] BGR is simply a rearrangement of RGB where the red and blue color channels are switched

other. This explains why there generally is no concept for colors like "bluish yellow" or "reddish green". Opponent color spaces have been shown to give the best results in color-shape descriptors for object recognition tasks (Van De Sande, Gevers, & Snoek, 2009). The specific opponent space used in this paper was taken from Rafegas and Vanrell (2018).

RGB to Opponent Space Transformation

$$\begin{aligned} O_1 &\leftarrow (R + G + B - 1.5)/1.5 \\ O_2 &\leftarrow (R - G)/1.5 \\ O_3 &\leftarrow (R + G - 2B)/2 \end{aligned} \tag{3.1}$$

3.2 Luminance Color Spaces

Luminance based color spaces, such as YUV and YCrCb, encode color while taking the human retina into account, in order to allow for reduced bandwidth use of chrominance components. The Y channel carries information related to the luminance channel, which describes the intensity of light. This is similar to the function served by rod cells in the retina (Podpora, Korbas, & Kawala-Janik, 2014). The additional channels carry information on chrominance, similar to the function cone cells serve in the human eye. Previous research has indicated that luma-based color spaces, specifically YUV, may be better suited as input for computer vision tasks compared to the standard RGB input (Podpora et al., 2014).

RGB to YUV Transformation

$$
\begin{aligned}
Y &\leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \\
U &\leftarrow (B - Y) \cdot 0.492 + \delta \\
V &\leftarrow (R - Y) \cdot 0.877 + \delta
\end{aligned}
\tag{3.2}
$$

RGB to YCrCb Transformation

$$
\begin{aligned}
Y &\leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \\
Cr &\leftarrow (R - Y) \cdot 0.713 + \delta \\
Cb &\leftarrow (B - Y) \cdot 0.564 + \delta
\end{aligned}
\tag{3.3}
$$

Chapter 4

Experiment A: CNN Color Classification Recall Experiment

Experiment A[1] investigates how CNNs learn to classify basic categories. Specifically, Experiment A attempts to establish two things: Firstly, this experiment investigates whether all color categories are equally difficult to learn, or whether colors exhibit hierarchical patterns of learning. Secondly, this experiment investigates how color classification differs across five color spaces; OPP, RGB, BGR, YUV and YCrCb.

4.1 Model

CNNs have already been proposed as a means by which partially faithful simulations of human vision can be modelled, specifically in the domain of color perception (Rafegas & Vanrell, 2018). The present study uses an architecture based on the CNN created by Rachmadi and Purnama (2015). This CNN architecture was chosen as it has already been proven to successfully learn to classify cars based on color from a car dataset taken from traffic camera images. Furthermore, this architecture consists of five convolutional layers. Five convolutional layer CNNs have been used in previous research to investigate similarities between human color perception and CNN internal color represen-

[1] All code associated with this experiment can be found at `https://github.com/chrishickey/color_hierarchy_experiment/tree/master/experiment1`

tations across convolutional layers (Rafegas & Vanrell, 2018).

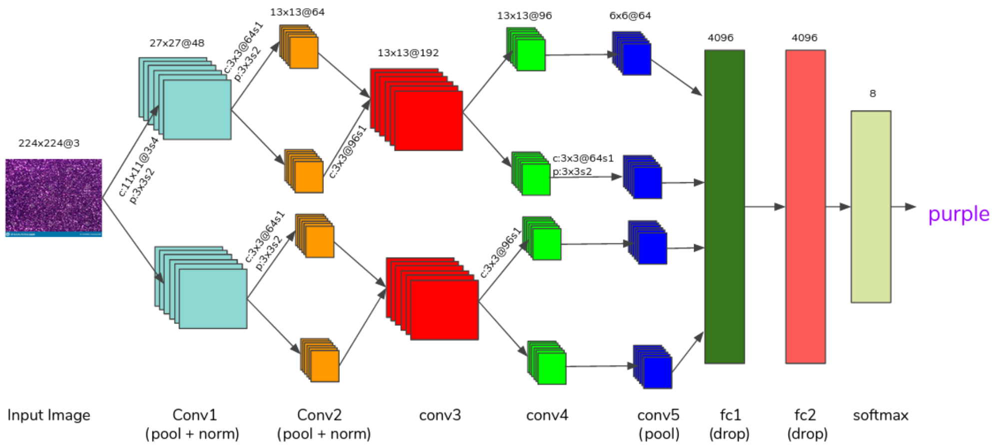


Figure 4.1: The CNN used in this study consists of two base networks and five convolutional layers. In addition to convolution, pooling was also performed in the first, fourth and fifth convolutional layers. Normalization was also performed on the first and second convolutional layers. Channel concatenation was performed before feeding forward to the fully connected layers.

The CNN architecture consists of two base networks, with five convolutional layers and two fully connected layers, each with 4,096 neurons (See Figure 4.1). ReLU Activation was used for all five convolutional and both fully connected layers. Additionally, pooling was performed in the first, fourth and fifth convolutional layers. Normalization was also performed on the first and second convolutional layers. Dropout of 0.6 was used in both fully connected layers. Softmax was used as the output layer to classify input images as one of the eight colors from the *Basic Color Dataset* described in the Datasets section. Stochastic gradient descent was used as an optimizer, with a learning rate of 0.001. In order to increase variety of the training data, zoom, shear and flip data augmentation were used in the training dataset. Mini batches of size 16 were also used during training. See Figure 4.1 for a visual representation of the model.

4.2 Method

For each of the five color spaces, 500 CNNs were trained to classify each of the eight colors shown in Figure 2.1. A total of 2,500 models were trained in this experiment. This large number of models was required in order to obtain a normally distributed sample of learning epochs for each color, suitable for Analysis of Variance (ANOVA) testing. For each model, the epoch in which a color was "learned" was recorded for each color. This experiment defines "learning" as achieving and maintaining *recall* on the test dataset for a color category, such that;

$$\frac{TruePositive}{TruePositive + FalseNegative} > 0.85 \tag{4.1}$$

Training stopped when all colors were successfully learned. ANOVA was then conducted to check for statistically significant differences in mean learning times between colors.

4.3 Results

Shapiro–Wilk analyses were carried out to establish if learning time for each color category in each color space was normally distributed. It was important to establish that data points for each color category in each color space were normally distributed, as this establishes whether or not the data is suitable for ANOVA. Results found that for all color categories across each color space, data were normally distributed with $p < 0.01$ certainty.

Next, ANOVA was used to investigate differences in learning time for each color category within each color space. If there were significant differences be-

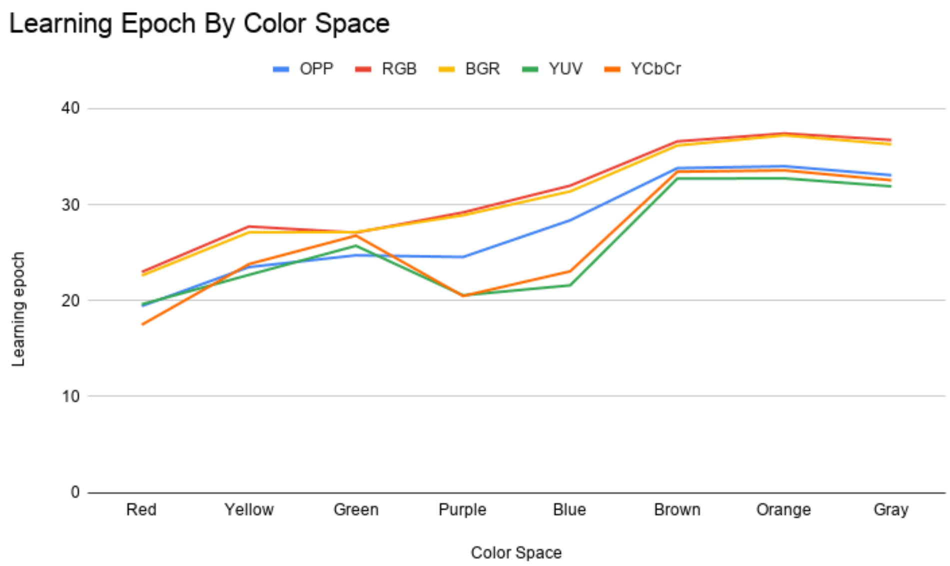


Figure 4.2: The mean time to learn each color across each color space in Experiment A

CS	Colors							
	Red	Yellow	Green	Purple	Blue	Brown	Orange	Gray
OPP	19.44 (9.87)	23.50 (11.5)	**24.73** (6.56)	24.55 (8.80)	28.38 (9.65)	33.81 (6.45)	34.01 (7.23)	33.08 (6.80)
RGB	22.99 (11.30)	27.73 (13.14)	27.09 (8.27)	29.20 (9.77)	31.99 (10.93)	36.60 (8.10)	37.43 (8.98)	36.75 (8.62)
BGR	22.62 (10.83)	27.12 (12.49)	27.15 (7.84)	28.90 (9.54)	31.39 (10.7)	36.17 (7.74)	37.23 (8.69)	36.30 (8.37)
YUV	19.62 (8.87)	**22.69** (11.23)	25.72 (6.59)	20.56 (6.84)	**21.60** (11.77)	**32.72** (5.99)	**32.75** (6.86)	**31.92** (6.50)
YCbCr	**17.49** (8.36)	23.81 (11.71)	26.78 (6.88)	**20.49** (7.78)	23.06 (11.36)	33.44 (6.16)	33.58 (6.97)	32.55 (6.46)

($n = 500$ for each color)

Table 4.1: Means (and standard deviations) of color learning time across each color space in Experiment A

tween any learning epochs within a color space, post-hoc analyses were conducted in order to establish which color category learning times were significantly different from each other. ANOVA analyses found statistically significant differences between the number of epochs required to learn colors in all five color spaces. Post-hoc Games-Howell analyses found hierarchical learning patterns in all five color spaces, with OPP, RGB, BGR and YUV all having four hierarchical levels, and YCbCr having five hierarchical levels. Some similarities are observable in the hierarchical patterns found in all five color spaces. For example, each hierarchy learned red in its first hierarchical layer. Additionally, each hierarchy learned brown, gray and orange in its final hierarchical layer. Results of ANOVA and Games-Howell analyses are summarized in Table 4.2, where the inequality sign $<$ denotes that the color to the left of the inequality was learned significantly faster than the color to the right of the inequality. An $=$ sign indicates that there was no statistically significant differences between learning epoch for two color categories.

CS	F	η^2	Epochs to Learn
OPP	206***	.27	$red < yellow = green = purple < blue < brown = gray = orange^{***}$
RGB	141***	.20	$red < yellow = green = purple < blue < brown = gray = orange^{***}$
BGR	151***	.21	$red < yellow = green = purple < blue < brown = gray = orange^{***}$
YUV	232***	.29	$red = blue = purple < yellow < green < brown = gray = orange^{***}$
YCbCr	270***	.32	$red < purple < yellow = blue < green < brown = gray = orange^{**}$

Table 4.2: The results of ANOVA and post-hoc analyses on differences in learning epoch between color categories across all color spaces. 500 learning epochs per color category were recorded for 500 trained networks. This was done for all 5 color spaces, resulting in 2,500 trained networks in total. F is the F-test statistic and η^2 is the effect size from ANOVA; in addition, there were 7 degrees of freedom between color categories for all color spaces ($\mathrm{df}_b^+ = 7$) and 3,984 within color categories ($\mathrm{df}_b^+ = 43,984$). The "Epochs to Learn" category details the results of post-hoc analyses. The inequality ($<$) denotes a significant difference at the $p < .01$ level, with the color to the left of the inequality being learned significantly faster than the color to the right. Equality ($=$) denotes the opposite. ** $p < .01$. *** $p < .001$.

Chapter 5

Experiment B: Faster R-CNN Colored Clothing Recall Experiment

Experiment B[1] examines how object color affects object detection in Faster R-CNNs (Ren, He, Girshick, & Sun, 2015). Specifically, does object color affect how successfully Faster R-CNNs are able to detect an object? And if so, which colored objects are more likely to be detected?

5.1 Model

Experiment B trained a Faster R-CNN (Ren et al., 2015) with a Resnet-50 backbone and a Feature Pyramid Network (Lin et al., 2017) to detect 13 clothing categories; **bag, belt, boots, footwear, outer, dress, sunglasses, pants, top, shorts, skirt, headwear, scarfs**. Horizontal flip data augmentation was applied to the training data in order to make the training set more diverse. Batches of 8 were used in training. Stochastic gradient descent was used as an optimizer, with Nesterov momentum of 0.9 and a learning rate of 0.005. In addition to this architecture being a state-of-the-art model for object detection, researches have started to observe similarities between this model and certain aspects of human vision. Specifically, Rosenfeld, Zemel, and Tsotsos (2018) highlight how parallels can be drawn between *feature interference* seen

[1] All code associated with this experiment can be found at `https://github.com/chrishickey/color_hierarchy_experiment/tree/master/experiment2`

in these Faster R-CNN architectures, and the *"binding problem"*, seen in biological vision. Feature interference refers to the phenomenon where Faster R-CNN's sometimes classify the same object differently based on background information. Similarly, the "binding problem" refers to how human vision sometimes leads to feature integration through illusory conjunctions, where features from two different objects are combined into one (Treisman, Schmidt, et al., 1982).

5.2 Method

Training continued until mAP on the test dataset plateaued at around 70%[2]. Next, within all clothing categories, objects were further sub-classified based on two criteria; basic color (e.g. strong red shirts, moderate red shirts etc. were all categorised as "red" shirts), and descriptive adjective, (e.g. strong yellowish pink shirts, strong red shirts etc, were all categorised as "strong" colored shirts). Two things should be noted about these classifications. Firstly, ambiguously colored clothing were not assigned color adjectives (e.g. grayish greenish yellow shirts). Secondly, the adjectives analysed in this experiment can be summarized into two groups; adjectives which emphasise chromatic hue saturation (e.g. brilliant, vivid, deep, strong) and adjectives which de-emphasise hue saturation (e.g. light, pale, dark, moderate).

Recall values were then calculated for each subcategory; i.e. out of all successfully recalled shirts, what percentage of red shirts were recalled. Object detections were prioritised based on the models certainty. The maximum number of object detections allowed per image equalled the number of ground truth objects in that image. If a subcategory contained less than 50 instances (e.g. if there were only 20 red belts in the test set), it was excluded from analy-

[2]mAP scores cited in the original paper (Zheng et al., 2018) were achievable when using a pre-trained backbone. However the purpose of this experiment was to investigate color learning patterns, not to maximize model performance.

ses, as denoted by the '-' in Tables 5.1 and 5.2. If a clothing category had less than three subcategories of objects (e.g. only brown and green boots met the 50 instance subcategory threshold), this clothing category was excluded from analyses. This is because a diverse range of subcategory colors and descriptive adjectives per object category are required to accurately and robustly assess the impact of colors and adjectives on object detection. All color and descriptive adjective recall scores across subcategories which met the threshold were then averaged out to produce a mean recall score for both colors and descriptive adjectives.

5.3 Results

Following other studies on color perception in computer vision (Rafegas & Vanrell, 2018), the results outlined in this section were obtained using OPP color space images as input. However hierarchical color patterns were also found using other color spaces as model inputs. Table 5.1 shows recall per object category based on color. Similar to Experiment A, red is the best recalled color subcategory across most clothing categories, with colors lower down the *Color Hierarchy* such as orange, brown and pink showing the worst recall.

Category	Red	Green	Blue	Purple	Yellow	Pink	Brown	Orange	Gray
Outer	**.725**	.680	.669	.667	.647	.557	.571	.650	.696
Skirt	**.819**	.673	.732	.727	-	.62	.654	-	-
Bag	**.752**	.646	.669	.656	.675	.66	.694	.702	-
Footwear	**.807**	.784	.805	.724	.752	.621	.740	.698	.755
Belt	.657	-	.456	.517	-	.462	**.660**	.481	-
Top	.629	.661	.614	.632	.619	**.736**	.580	.626	.423
Dress	.702	**.718**	.698	.690	-	.660	.500	-	-
Pants	**.914**	.859	.911	.878	-	.830	.713	-	-
Mean	**.751**	.717	.694	.686	.673	.643	.639	.631	.625

Table 5.1: Recall for each clothing category based on color in Experiment B. The best recall score within each clothing category is highlighted in bold.

Category	Brilliant	Vivid	Deep	Strong	Dark	Moderate	Light	Pale
Outer	**.725**	.671	.703	.614	.715	.564	.602	.663
Skirt	.805	**.813**	.758	.767	.739	.652	.633	.610
Boots	-	-	.556	**.559**	.497	.421	.364	.265
Bag	**.790**	.742	.710	.716	.675	.663	.632	.636
Footwear	**.807**	.789	.755	.774	.73	.718	.733	.697
Belt	-	.563	**.639**	.599	.577	.561	.502	.402
Top	**.722**	**.722**	.621	.665	.575	.579	.644	.654
Dress	.720	**.726**	.636	.697	.664	.647	.659	.663
Pants	-	**.902**	.872	.831	.857	.815	.845	.843
Scarf	-	.296	**.390**	.351	.330	.382	.305	.303
Shorts	-	.759	.709	.763	.707	**.795**	.72	.778
Headwear	-	-	.703	-	**.711**	.641	.686	.617
Mean	**.762**	.698	.671	.667	.648	.620	.610	.594

Table 5.2: Recall for each clothing category based on descriptive adjective used to describe the color in Experiment B. The best recall score within each clothing category is highlighted in bold.

Table 5.2 shows recall per object category based on descriptive adjective that was used to describe clothing color. Adjectives which emphasise higher levels of chromatic hue colorfulness, such as “brilliant”, “vivid” and “deep”, show best recall by the Faster R-CNN model. Conversely, adjectives which de-emphasise colorfulness, such as “dark”, “light” and “pale”, show notably worse recall performance across most clothing categories.

CNNs have color selective neurons similar to those in biological vision (Rafegas & Vanrell, 2017, 2018). In order to isolate the specific augmenting effect of both color and descriptive adjective on object detection recall caused by these color selective neurons, a second Faster R-CNN model was trained using colorless, grayscale images as input (See Figure 5.1). This second model acts as a baseline against which we can assess the effect of color on object recall. Recall for each subcategory from this control model trained on grayscale “colorless” image input was then subtracted from the original recall scores. These recall scores are intended to show the impact of color selective neurons on ob-

(a) Original (b) Grayscale

Figure 5.1: Comparing original and grayscale images in Modanet dataset used in Experiment B

ject recall and control for non-color related factors such as background color contrast or other features. These results are summarized in Tables 5.3 and 5.4

Category	Red	Green	Yellow	Blue	Purple	Pink	Orange	Gray	Brown
Outer	.143	.100	**.235**	.094	.093	.016	.117	.071	-.025
Skirt	**.256**	.200	-	.249	.234	.185	-	-	.179
Bag	**.151**	.108	.078	.047	.000	.072	.000	-	.062
Footwear	**.056**	.005	.028	.025	-.012	.000	-.018	.047	.039
Belt	**.163**	-	-	.029	-.017	.058	.135	-	.113
Top	-.011	-.011	-.082	-.023	-.029	**-.010**	-.077	-.058	-.058
Dress	.061	.028	-	.034	**.086**	.000	-	-	-.042
Pants	**.062**	.056	-	.038	-.027	.000	-	-	-.137
Mean	**.110**	.069	.065	.062	.041	.040	.031	.020	.017

Table 5.3: Recall for each clothing category based on color in Experiment B, after subtracting recall values from the control grayscale model. The best recall score within each clothing category is highlighted in bold.

Table 5.3 shows that for all color categories, the presence of color led to better recall on average than the absence of color. However, certain colors showed noticeable improved recall over the baseline grayscale images when compared

Category	Brilliant	Vivid	Strong	Deep	Light	Moderate	Dark	Pale
Outer	**.147**	.132	.083	.068	.110	.048	.058	.126
Skirt	**.317**	.192	.302	.258	.181	.172	.199	.185
Boots	-	-	.044	-.022	-.019	-.055	-.013	-.059
Bag	.099	**.166**	.069	.086	.019	.052	.024	.031
Footwear	.037	.050	.044	.051	.013	.020	.025	-.007
Belt	-	.019	.099	.112	**.119**	.050	.037	.039
Top	.017	**.026**	-.003	.008	-.016	-.066	-.017	-.038
Dress	**.037**	.037	.010	-.006	-.037	-.009	-.006	-.015
Pants	-	**.105**	.063	.016	.024	-.050	-.017	-.009
Scarf	-	.111	.070	.104	.031	**.124**	-.026	.040
Shorts	-	.019	.034	.06	.037	**.091**	.059	.074
Headwear	-	-	-	**.117**	.069	.109	.085	.028
Mean	**.109**	.086	.074	.071	.044	.040	.034	.033

Table 5.4: Recall for each clothing category based on descriptive adjective used to describe the color in Experiment B, after subtracting recall values from the control grayscale model. The best recall score within each clothing category is highlighted in bold.

to others. Similar to Table 5.1, red clothing items generally showed the most improved recall across most categories. For 7 out of 8 clothing categories, red showed greater than 5% improvement over the baseline grayscale images. Compare this to pink, where only 3 of the 8 clothing categories showed greater than 5% improvement, or brown, where half the clothing categories showed dis-improved recall when compared to the baseline. It is noteworthy that color appeared to improve recall much more for smaller items such as bags, skirts and belts as opposed to larger items like dresses, or pants.

Table 5.4 shows that for all color adjectives, the presence of color led to better recall on average than the absence of color. However similar to Table 5.2, adjectives which emphasise higher levels of chromatic hue colorfulness, such as “brilliant”, “vivid” and “deep”, showed the greatest improvement in recall when compared to the baseline grayscale images. Conversely, adjectives which de-emphasise colorfulness, such as “dark”, “light” and “pale”, showed the lowest level of improved recall when compared to the baseline grayscale images.

Chapter 6

Discussion and Conclusion

Experiment A Experiment A investigated whether color categories are learned in a random order by CNNs, or whether meaningful hierarchical patterns can be observed when CNNs learn specific color categories. Results found that while there is large variation between individual CNN models, CNNs learn color categories along a normally distributed number of epochs, with statistically significant differences observable between the mean learning time for each color category. These statistically significant differences were found across five different color spaces. These results suggest that conventional color categories as understood by humans are not randomly learned by CNNs, and that certain color categories are learned faster than others in a statistically significant way.

Experiment B Experiment B builds on the first experiment by investigating whether the hierarchical color category learning patterns seen in Experiment A affected performance in practical CNN based computer vision tasks, specifically object detection. Results found that for Faster R-CNNs, object recall on a dataset of fashion clothing items was impacted by object color, with recall being noticeably better for certain colored clothing items compared to others. For example, recall for red clothing items was on average 10% higher than recall for pink, brown, orange or gray clothing items. These differences persisted, even when controlling for factors such as background color contrast through models trained on grayscale images. In addition to color category, descriptive adjectives for colors also appeared to noticeably impact clothing item recall. Adjectives which emphasised color chromatic hue such as "brilliant", "vivid" and "strong" colors appear to show noticeably better recall when compared to "moderate", "dark" and "pale" colors that de-emphasise hue. These differences also persisted when controlling for other factors through grayscale image trained models.

Experiments A and B In both experiments A and B, hierarchical color learning patterns were observed, giving support to the alternative hypothesis outlined in the introduction. Some notable similarities can be seen between the hierarchical learning patterns that emerge in both experiments. In both experiments, red shows both the fastest learning speed and best recall. Green, yellow and blue generally followed red in showing the next best learning speed and recall. Finally, brown, orange and gray displayed consistently slower learning speed and worse recall scores when compared to other colors. This hierarchical pattern observed in both experiments mimics the general hierarchical patterns observed in the human *Color Hierarchy.*

Contributions and Weaknesses The present study contributes to literature on the use of neural networks and deep learning methodologies as a means of investigating the existence of cognitive universalities. Similar to methodologies used to investigate the universality of cognitive arithmetic difficulty (Cho, Lim, Hickey, Park, & Zhang, 2019; McClelland et al., 2010), the present study uses neural networks as a means of investigating the universality of hierarchical color patterns. This investigation is done using partially faithful models of human vision (CNNs) and biologically-inspired color space images as input

(OPP, YCbCr). Results found support for Berlin and Kay (1969) and their original *Color Hierarchy*. Additionally, results from the present study give tentative support to studies that claim that models inspired by the human visual system can at least partially simulate the emergence of such hierarchical color category patterns (Baronchelli et al., 2010; Regier, Kay, & Khetarpal, 2007). This is distinct from other research which suggests cultural (Wierzbicka, 2008) or evolutionary psychological (Tchernikov & Fallah, 2010) explanations for the emergence of hierarchical color patterns. In reality, the fact that color hierarchical patterns in language show recurrent, cross-cultural similarities while not being strictly uniform across languages suggests that some confluence of both universalities and cultural factors lead to the emergence of the *Color Hierarchy*.

For the purpose of this study, color categories were generally considered as both discrete and unrelated. However, as Gärdenfors (2004) outlines, this is actually not the case. Color categories are a prime example of something which can be represented through conceptual space - a geometric structure formed by a set of quality dimensions. Given that color can be represented via a geometrical structure along three quality dimensions (hue, chromaticness / saturation, and brightness), distance between separate color categories as well as an area occupied by any single color category are both tractable and calculable. Therefore we can not assume that the emergence of hierarchical color preferences are purely a function of biological vision. In order to conduct a truly thorough investigation into what causes the emergence of hierarchical color patterns, examining how hierarchical color patterns relate to geometric distance between conventional color categories should be considered.

Additionally, the present study labels color categories either through the visual assessment of a native English speaker, or by using the ISCC-NBS color labeling system which is very much based on common English language understanding of colors. A more robust investigation of hierarchical color preferences could explore if hierarchical patterns emerge when labelling of separate color categories is considered from the perspective of different languages, similar to the precedent set in studies like Taft and Sivik (1997), and Kay et al. (2009). If the present study were to be replicated and evidence for hierarchical color patterns in CNNs was found using labelling systems from multiple languages, this would add significant weight to the strength of conclusions that could be drawn from the present study. Additionally, if results from this experiment were duplicated using different recall values and numbers of convolutional layers for Experiment A, and different object detection models such as YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) for Experiment B, this would also serve to further enhance the robustness of the conclusions that could be drawn from the present study.

Future Study Firstly, the present study focuses solely on investigating whether training CNN based models on color data could lead to the emergence of hierarchical color patterns. Therefore, future studies should aim to better understand how and why these patterns emerge in CNNs. Previous research has outlined a methodology for investigating which single, opponent and non-opponent color selective neurons are present at each convolutional layer in a trained CNN (Rafegas & Vanrell, 2017, 2018). Investigating how these color selective neurons emerge and evolve over the course of training should shed light on how color hierarchies emerge, at least in the context of CNNs. Additionally, seeing how color space image input affects the emergence of color selective neurons could offer insight as to why different color spaces (e.g luminance color spaces, opponent color spaces, etc.) lead to the emergence of different hierarchical color patterns in CNNs, as was found in Experiment A.

Secondly, in addition to finding hierarchical color patterns in CNNs, the present study also found evidence to suggest that different color spaces learn certain colors better than others (Experiment A). Furthermore, high color saturation (e.g. “brilliant”, “vivid” colors) lead to better object recall than less colorful objects (Experiment B). It could be possible to leverage this information in order to improve performance for object detection tasks. For example, given that different color spaces may be more optimized for certain colors, using multiple color space images as input to a model, similar to the methodology outlined by Gowda and Yuan (2018), may lead to improved object detection performance. Additionally, artificially enhancing color saturation of images, thus making them more colorful, may also improve object detection performance based on the results from Experiment B.

Finally, investigating cognitive universalities - defined as core mental attributes shared by humans everywhere - is a notoriously difficult task, given the various costs, linguistic issues, and other difficulties inherent in conducting unbiased, cross-cultural studies (Norenzayan & Heine, 2005). Partially biologically faithful deep learning models offer an economical way to test the existence of universalities in a comparatively controlled setting, as was the aim of the present study. If parallel universalities can organically emerge from neural network models, this lends credence to the biological faithfulness of that model, as was seen in the arithmetic difficulty modelling study conducted by Cho et al. (2019). Continuing to investigate universalities through deep learning (for example, difficulty in cognitive object rotation), could serve to advance a multitude of academic disciplines, including AI, cognitive science and psychology.

References

Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, *107*(6), 2403–2407.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Univ of California Press.

Casson, R. W. (1997). Color shift: evolution of english color terms from brightness to hue. *Color categories in thought and language*, 224.

Cho, S., Lim, J., Hickey, C., Park, J., & Zhang, B.-T. (2019). Simulating problem difficulty in arithmetic cognition through dynamic connectionist models. *arXiv preprint arXiv:1905.03617*.

Conway, B. R. (2001). Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (v-1). *Journal of Neuroscience*, *21*(8), 2768–2783.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought.* MIT press.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., . . . Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790.

Gowda, S. N., & Yuan, C. (2018). Colornet: Investigating the importance of color spaces for image classification. In *Asian conference on computer vision* (pp. 581–596).

Hurvich, L. M., & Jameson, D. (1957). An opponent-process theory of color vision. *Psychological review*, *64*(6p1), 384.

Jameson, K. A., & Komarova, N. L. (2009). Evolutionary models of color categorization. i. population categorization systems based on normal and dichromat observers. *JOSA A*, *26*(6), 1414–1423.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey.* CSLI Publications Stanford, CA:.

Kay, P., & Regier, T. (2007). Color naming universals: The case of berinmo. *Cognition*, *102*(2), 289–298.

Kelly, K., & Judd, D. (1955). *The iscc-nbs method of designating colors and a dictionary of color names* (No. 553). US Government Printing Office.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Lakoff, G. (2008). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Lim, H., Wang, Y., Xiao, Y., Hu, M., & Felleman, D. J. (2009). Organization of hue selectivity in macaque v2 thin stripes. *Journal of neurophysiology*, *102*(5), 2603–2615.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2117–2125).

Loreto, V., Mukherjee, A., & Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, *109*(18), 6819–6824.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348–356.

Munsell, A. H. (1941). *A color notation. baltimore, md.: Munsell color co.* Inc.

Newhall, S. M., Nickerson, D., & Judd, D. B. (1943). Final report of the osa subcommittee on the spacing of the munsell colors. *josa*, *33*(7), 385–418.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological bulletin*, *131*(5), 763.

Plataniotis, K. N., & Venetsanopoulos, A. N. (2013). *Color image processing and applications*. Springer Science & Business Media.

Podpora, M., Korbas, G. P., & Kawala-Janik, A. (2014). Yuv vs rgb-choosing a color space for human-machine interaction. *FedCSIS Position Papers*, *18*, 29–34.

Rachmadi, R. F., & Purnama, I. (2015). Vehicle color recognition using convolutional neural network. *arXiv preprint arXiv:1510.07391*.

Rafegas, I., & Vanrell, M. (2017). Color representation in cnns: parallelisms with biological vision. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 2697–2705).

Rafegas, I., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision research*, *151*, 7–17.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*,

104(4), 1436–1441.
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).
Rosenfeld, A., Zemel, R., & Tsotsos, J. K. (2018). The elephant in the room. *arXiv preprint arXiv:1808.03305*.
Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences*, *114*(21), 5545–5550.
Taft, C., & Sivik, L. (1997). Salient color terms in four languages. *Scandinavian Journal of Psychology*, *38*(1), 29–34.
Tchernikov, I., & Fallah, M. (2010). A color hierarchy for automatic target selection. *PLoS One*, *5*(2).
Treisman, A., Schmidt, H., et al. (1982). Illusory conjunctions in the perception of objects. *1982*, *14*, 107–141.
Van De Sande, K., Gevers, T., & Snoek, C. (2009). Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *32*(9), 1582–1596.
Wierzbicka, A. (2008). Why there are no 'colour universals' in language and thought. *Journal of the Royal Anthropological Institute*, *14*(2), 407–425.
Wyszecki, G., & Stiles, W. S. (1982). *Color science* (Vol. 8). Wiley New York.
Zheng, S., Yang, F., Kiapour, M., & Piramuthu, R. (2018). Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of 26th acm international conference on multimedia* (pp. 1670–1678).

국문초록

경험적으로 색상 계열은 보편적으로 다양한 문화권에 걸쳐 순환적이고 위계적인 패턴을 나타내며 그 순서는 다음과 같이 나타난다; *검은색 < 붉은색 <녹색 < 노란색 < 파란색 < 갈색 < 분홍색, 회색, 주황색, 보라색*. 이러한 경향은 “색상 위계 (Color Hierarchy)”라 불린다. 소개될 두 가지 실험을 통해 본 연구는 콘볼루션 신경망의 경우에도 색상 위계 순서에 따른 색상계열 학습이 진행되는지 알아본다. 실험 A는 색상 계열이 무작위로 학습되는지 위계적인 순서를 통해 학습되는지 알아본다. 실험의 결과를 통해 색상 위계상으로 더 상위의 색상(예: 붉은색)은 일반적으로 하위의 색상들 (예: 갈색, 주황색, 회색) 보다 앞서 학습이 이뤄짐을 볼 수 있다. 실험 B는 색상 위계에 따른 학습 편차가 객체인식 학습의 재현률(recall)에도 영향을 끼치는지 알아본다. 실험 A에서와같이 색상 위계는 객체인식 재현률에도 큰 영향을 끼친다. 추가적으로 색상을 강조하는 부사(예: 선명한, 눈에 띄는, 짙은)와 함께 묘사된 객체의 경우에는 반대로 색상을 억제하는 부사(예: 어두운, 옅은, 엷은)와 함께 묘사된 객체들보다 재현률이 높게 나타난다. 부사와 색상의 효과는 흑백 이미지들에 대해서도 여전히 관측된다. 이와 같은 결과들은 사람과 콘볼루션 신경망의 색상 지각과정의 유사성을 보여주며 객체 인식에 영향을 주는 요인들에 대한 통찰력을 제공한다. 또한 이 결과들은 딥러닝 방법이 인지과정의 보편성을 살피는 데에 효율적이고, 치우치지 않으며, 경제적인 방법임을 지시한다.

주요어: 색상 위계; 컴퓨터 비전; 콘볼루션 신경망; 인지과정의 보편성; 물체 탐지
학번: 2018-29483

Acknowledgements

I would like to thank Professor Zhang and everyone involved in the SNU Bio Intelligence Lab and Cognitive Science Department for being so kind to work with over the last two years. I would also like to say a special thanks to Paula Higgins for always being there to help me day or night whenever I needed it.