Hierarchical Color Learning in Convolutional Neural Networks

Chris Hickey
Seoul National University
chickey@bi.snu.ac.kr *

Byoung-Tak Zhang
Seoul National University
btzhang@bi.snu.ac.kr

Abstract

Empirical evidence suggests that color categories emerge in a universal, recurrent, hierarchical pattern across different cultures. 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 investigated whether color categories are learned randomly, or in a fixed, hierarchical fashion. Results show that colors higher up the Color Hierarchy (e.g. red) were generally learned before colors lower down the hierarchy (e.g. brown, orange, gray). Experiment B examined whether object color affects recall in object detection. Similar to Experiment A, results found that object recall was 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). These results highlight similarities between humans and CNNs in color perception, and provide insight into factors that influence object detection.

1. Introduction

The human eye can see 7,000,000 colors. However, from these millions of colors, most languages have no more than 13 basic terms to conceptualize this vast color spectrum[2]. This segmentation in human language occurs along the most salient dimension of color: hue (i.e. red, yellow, blue). One of the most enduring theories regarding how these few, hue-based categories emerge was first proposed by Berlin & Kay half a century ago [1]. Their cross cultural research observed a fixed sequence according to which languages gain color terms over time.

To account for this evolution, Berlin & Kay [1] put forward two suppositions: (i) color lexicons for all languages are drawn from a set of fixed "universal" categories, and (ii) languages add color terms in a relatively fixed sequence, such that; *white, black* $<$ *red* $<$ *green, yellow* $<$ *blue* $<$ *brown* $<$ *purple, gray, orange, and pink*. Therefore, if a language has a term for a given color in this inequality, it will also have terms for all colors to the left of that given color. This sequence is referred to as the *"Color Hierarchy"*. The endurance of this theory is bolstered by its replicability in both human studies [2], and computational studies where computational agents negotiate color terms for regions of the hue spectrum via a "Category Game" [5].

Over two experiments, the present study aims to investigate color learning patterns in neural networks. 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? Parallels have already been drawn between the human visual system and CNNs in the domains of internal object representation [4] and color perception [7]. 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.

2. Experiment A: CNN Color Classification Recall Experiment

Experiment A examines how CNNs learn to classify basic color categories. Specifically, are color categories equally difficult to learn, or are color categories learned in some predictable, hierarchical order similar to humans?

2.1. Dataset

Basic Color Dataset: The dataset for this experiment was made up of 880 images taken from Google Images. Images were gathered by searching for specific colors, (e.g "red") and downloading images that best encapsulate this color (See Figure 1). Eight of the eleven English basic color terms were used in this study. Black and white were ex-

*This work was partly supported by the Korean government (2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind , P0006720-GENKO).

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

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

Figure 1: Samples from the Basic Color Dataset used in Experiment A

cluded because these terms can correspond to brightness rather than hue [9]. Pink was also excluded, due to pink being considered a shade of red in many languages [1]. 60 images for each color were assigned to the training set. The remaining 50 images were assigned to the test set.

2.2. Method

This experiment used the CNN architecture outlined in [6]. This CNN architecture was chosen as it has been proven to successfully learn to classify colors from a dataset taken from traffic camera images. Relu Activation was used for all convolutional and fully connected layers. Softmax was used for output. Stochastic gradient descent was used as an optimizer, with a learning rate of 0.001. Zoom, shear and flip data augmentation, and mini batches of size 16 were also used in training. The epoch in which a color was "learned" was recorded for each color. Training was stopped when all colors were successfully learned. This experiment defines "learning" as achieving and maintaining *recall* on the test dataset for a color category, such that;

$$\frac{TruePositive}{TruePositive + FalseNegative} > 0.85 \qquad (1)$$

Five color space image inputs were investigated in this experiment; OPP, RGB, BGR[1], YCbCr and YUV. For each of the five color spaces, 500 CNNs were trained to classify each of the eight colors shown in Figure 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. ANOVA analyses was then conducted to check for statistically significant differences in learning times between colors.

[1]BGR is just a rearrangement of RGB, effectively acting as a control in this study

Color Space	Red	Yellow	Green	Purple	Blue	Brown	Orange	Gray
OPP	19.44	23.50	**24.73**	24.55	28.38	33.81	34.01	33.08
RGB	22.99	27.73	27.09	29.20	31.99	36.60	37.43	36.75
BGR	22.62	27.12	27.15	28.90	31.39	36.17	37.23	36.30
YUV	19.62	**22.69**	25.72	20.56	**21.60**	**32.72**	**32.75**	**31.92**
YCbCr	**17.49**	23.81	26.78	**20.49**	23.06	33.44	33.58	32.55

Table 1: Mean number of epochs taken to learn each color for all color spaces. The minimum number of epochs required to learn each color is highlighted in bold.

2.3. Results

The average number of epochs taken to learn each color category is summarised in Table 1. It is noteworthy that different color spaces exhibited different hierarchical learning patterns. Green was learned fastest using OPP color space images as input. Red and purple were learned fastest using YCbCr color space as input. Finally, yellow, blue, brown, orange and gray were learned fastest using YUV color space.

ANOVA analyses found statistically significant differences between the number of epochs required to learn colors for all 5 color spaces. Post-hoc Games-Howell analyses found hierarchical learning patterns in all 5 color spaces, with OPP, RGB, BGR and YUV all having 4 hierarchical levels, and YCbCr having 5 hierarchical levels. Some similarities are observable in the hierarchical patterns found in all 5 color spaces. Each hierarchy learned red in its first hierarchical layer, and learned brown, gray and orange in its final hierarchical layer. A similar pattern is also seen in the *Color Hierarchy*[1] observed in humans, as outlined in the Introduction. Results of ANOVA and Games-Howell analyses are summarized in Table 2.

3. Experiment B: Faster R-CNN Colored Clothing Recall Experiment

Experiment B examines how object color affects object detection in Faster R-CNNs[8]. Specifically, does object color affect how successfully Faster R-CNNs are able to detect an object?

3.1. Dataset

Modanet: Zheng et *al*. [10] created a large scale street fashion dataset with polygon annotations, containing 55,176 images. 13 categories are labeled in this dataset: **bag, belt, boots, footwear, outer** (coat, jacket etc.), **dress, sunglasses, pants, top, shorts, skirt, headwear, scarfs**. We then added color attributes to each of the objects in the dataset as follows: firstly, semantic segmentation was performed on each object. Then 500 random RGB pixels were sampled from this segmentation and mapped to their nearest NBS-ISCC color label [3]. NBS-ISCC color labels aim to be as commonly understandable as possible, using only

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 2: Results of ANOVA and post-hoc Games-Howell analyses on differences in epochs required to learn each color. 500 results per color category from 500 networks were analyzed for each color space. F is the F-test statistic and η^2 is the effect size. The "epochs to learn" column describes 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 faster than the color to the right. Equality denotes the opposite. ** $p < .01$. *** $p < .001$.

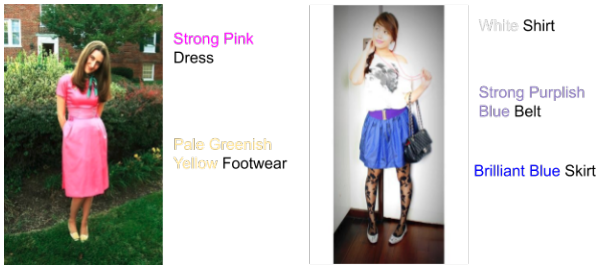

Figure 2: Samples of color attributes assigned to objects in the Modanet dataset using pixel sampling.

13 basic color terms and a handful of adjectives to create 267 unique color labels. The NBS-ISCC color label sampled most often from the object pixel sampling process was assigned as the label for that object (See Figure 2). If no color label emerged for at least 10% of the labels for an object, that object was assigned a *null* color label. Data was 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.

3.2. Method

Experiment B trained a Faster R-CNN with a Resnet-50 backbone and a Feature Pyramid Network to detect the 13 clothing categories outlined above. Batches of 8 were used in training. Stochastic gradient descent was used as an optimizer, with a learning rate of 0.005. Training continued until mAP on the test dataset plateaued at close to 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). 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 analyses, as denoted by the '-' in tables 3 and 4. If a clothing category had less than 3 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.

[2]mAP scores cited in the original paper[10] 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.

3.3. Results

Following other studies on color perception in computer vision [7], the results outlined in this section were obtained using OPP color space images as input. However a similar hierarchical pattern was also found for RGB color space inputs. Table 3 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.

Table 4 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.

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 3: 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	.730	.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**	.720	.778
Headwear	-	-	.703	-	**.711**	.641	.686	.617
Mean	**.762**	.698	.671	.667	.648	.620	.610	.594

Table 4: 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.

4. Conclusion and Future Work

Results from both experiments suggest hierarchical color learning patterns in CNNs, similar to humans. Results from Experiment A show that color categories are learned in a hierarchical pattern, regardless of color space input type. Results from Experiment B show that color has a noticeable impact on recall for object detection in Faster R-CNNs. Recall for red clothing items was on average 10% higher than recall for pink, brown, orange or gray clothing items. Additionally, "Brilliant", "Vivid" and "Deep" colors which emphasise colorfulness show noticeably better recall compared to "Pale", "Dark" and "Light" colors. Across both experiments, colors higher up the *Color Hierarchy*, (e.g. red, green) showed faster learning and better recall than colors lower down the Hierarchy. Future work should explore how these hierarchical learning patterns can be exploited to improve model performance. For example, certain color space image inputs may be more suitable, depending on what color objects are being detected in object detection tasks.

References

[1] Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. 1969. 1, 2

[2] Paul Kay, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook. *The world color survey*. CSLI Publications Stanford, CA:, 2009. 1

[3] Kenneth Kelly and Deane Judd. *The ISCC-NBS method of designating colors and a dictionary of color names*. Number 553. US Government Printing Office, 1955. 2

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

[5] Vittorio Loreto, Animesh Mukherjee, and Francesca Tria. On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, 109(18):6819–6824, 2012. 1

[6] Reza Fuad Rachmadi and I Purnama. Vehicle color recognition using convolutional neural network. *arXiv preprint arXiv:1510.07391*, 2015. 2

[7] Ivet Rafegas and Maria Vanrell. Color encoding in biologically-inspired convolutional neural networks. *Vision research*, 151:7–17, 2018. 1, 3

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[9] Anna Wierzbicka. Why there are no 'colour universals' in language and thought. *Journal of the Royal Anthropological Institute*, 14(2):407–425, 2008. 2

[10] Shuai Zheng, Fan Yang, M Kiapour, and R Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of 26th ACM international conference on Multimedia*, pages 1670–1678, 2018. 2, 3