‘Lighter’ Can Still Be Dark: Modeling Comparative Color Descriptions

Olivia Winn
Computer Science Department
Columbia University
olivia@cs.columbia.edu

Smaranda Muresan
Data Science Institute
Columbia University
smara@columbia.edu

Abstract

We propose a novel paradigm of grounding comparative adjectives within the realm of color descriptions. Given a reference RGB color and a comparative term (e.g., ‘**lighter**’, ‘**darker**’), our model learns to ground the comparative as a direction in the RGB space such that the colors along the vector, rooted at the reference color, satisfy the comparison. Our model generates grounded representations of comparative adjectives with an average accuracy of 0.65 cosine similarity to the desired direction of change. These vectors approach colors with Delta-E scores of under 7 compared to the target colors, indicating the differences are very small with respect to human perception. Our approach makes use of a newly created dataset for this task derived from existing labeled color data.

1 Introduction

Multimodal approaches to object recognition have achieved a degree of success by grounding adjectives and nouns from descriptive text in image features (Farhadi et al., 2009; Lampert et al., 2009; Russakovsky and Fei-Fei, 2010; Lazaridou et al., 2015). One limitation of this approach, particularly for fine-grained object recognition, is when objects are differentiated not by having unique sets of attributes but by a difference in the strengths of their shared attributes (Wang et al., 2009; Duan et al., 2012; Maji et al., 2013; Vedaldi et al., 2014). In text, this difference is described using comparative adjectives. For example, the sexual dimorphism of the American black duck is described with the phrase “females tend to be ***slightly paler*** than males, with ***duller olive*** bills”.[1]

(a) The grounding of ‘**darker**’ trained on teal data, applied to a teal sample

(b) The grounding of ‘**darker**’ trained on ***pink*** data, applied to a teal sample

Figure 1: Grounding ‘darker’

In a recent study of pragmatic referring expression interpretation in the context of color selection, Monroe et al. (2017) found that speakers almost always used comparative adjectives when the target color was very similar to a distractor, rather than using multiple positive form adjectives to create a highly specific description of the color independent of its surroundings. Though color has been studied in terms of its contextual dependence and vagueness in grounding (Egré et al., 2013; McMahan and Stone, 2015; Monroe et al., 2016, 2017), no approaches have focused explicitly on learning to ground comparative adjective; in this work we focus on comparative color descriptions.

The presence of distractors in the Monroe et al. (2017) study is important - comparatives describe a change in a feature *with respect to a reference point*. While the description ***light blue*** can be understood to represent a particular subset of colors in RGB, for example, neither ‘**lighter**’ nor ‘**lighter blue**’ have explicit representations; it is only with a reference that we can image what color either might refer to. If the reference color is a deep navy blue, then we imagine the target to be much closer to navy than, for example, a sky blue.

We propose a new paradigm of learning to ground comparative adjectives within the realm of color descriptions: given a reference RGB color and a comparative term (e.g. ‘**lighter**’, ‘**paler**’),

[1] https://www.allaboutbirds.org/guide/American_Black_Duck/id

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 790–795
Melbourne, Australia, July 15 - 20, 2018.

our deep learning model learns to ground the comparative as a direction in RB space such that the colors along the vector, rooted at the reference color, satisfy the comparison (Section 3). The reference color does more than quantify the specific RGB values to apply the comparative to: it also affects the grounding of the comparative. For example, '**darker**' might seem like a simple change - simply reduce the values of all color channels equally towards 0. But as Fig 1 shows, '**darker**' refers to a different direction in RGB space depending on the reference color, and thus we need a reference-dependent approach.

Our approach makes use of a newly created dataset for this task derived from an existing labeled color dataset (McMahan and Stone, 2015) (Section 2). Our results in Section 5 show that our model generates grounded representations of comparative adjectives with an average accuracy of 0.65 cosine similarity to the desired direction of change. These learned vectors approach colors with Delta-E scores of under 7 compared to the target colors, indicating the differences are very small with respect to human perception.

2 Data

We utilize the labeled RGB color data originally collected by Munroe (2010), through an online survey asking participants to provide free-form labels to various RGB samples. This data was then cleaned by McMahan and Stone (2015)[2]. The cleaned data contains 821 color labels, averaging 600 RGB datapoints per label. These labels do not contain comparative adjectives, but many start with adjectives in the positive form (e.g., ***dusty***, ***bright***). As Lassiter and Goodman (2017) write, "Vague terms ... are generally thought in linguistic semantics to rely on a free threshold variable: 'heavy' is interpreted as 'heaver than θ'." Coming back to the example of ***light blue***, implicit in the term is the assumption that there is a reference ***blue***, such that ***light blue*** is understood as '**lighter**' than this reference. By representing this referential ***blue*** with the ***blue*** RGB samples from the data, we can assume the ***light blue*** RGB samples are '**lighter**' than these references, giving us a quantitative θ in which to ground '**lighter**'. Applying this process to the rest of the labels, we convert the original dataset into 415 tuples (reference color label, comparative adjective, and target color label), such as (***blue***, '**lighter**', ***light blue***), where each color label is a set of RGB datapoints as in McMahan and Stone (2015). Note that not all labels containing quantifiers could be utilized in this manner; one does not consider ***cobalt blue*** to be '**more cobalt**' than the average ***blue***. The new dataset of 415 tuples contains 79 unique reference color labels and 81 unique comparatives and is made available online.[3]

While it is reasonable to believe that the comparative adjective describes the relationship between the colors in general, individual pairs of colors from the data may not display the appropriate θ. Thus, we make the assumption that the comparison holds true for the *average* of the target ***light blue*** samples, and use the average as our ground truth given the ***blue*** reference colors and the comparative adjective '**lighter**'.

3 Method

We have chosen to represent comparative adjectives in RGB space as directions, such that given ain input RGB reference color r_c and a comparative adjective w our model outputs a vector $\vec{w}_g$ pointing from r_c in the direction of change in RGB, which in training is measured against the direction towars a target color t_c. Fig 1 is a good indication for why this representation is appropriate; our output $\vec{w}_g$ corresponds to the rate of change across the color bar, indicating the direction along with the degree of the compared property increases. All points along this line are representations of w in respect to r_c.

The network architecture consists of two fully connected layers, shown in Fig 2. The comparative is represented as a bi-gram to account for comparatives which necessitate using '**more**' (e.g. "**more electric**"); single-word comparatives are preceded by the zero vector. We used the Google pre-trained word embeddings[4] with d=300 (Mikolov et al., 2013a,b). As these vectors are two orders of magnitude larger than the reference RGB color, we input the reference directly into both layers of the network, helping to mitigate the loss of

[2] A few of the labels (such as '**horrible**') were manually discarded, as the corresponding set of colors were too widely spread across RGB space for the label to be considered as describing a distinct color.

[3] https://bitbucket.org/o_winn/comparative_colors

[4] https://code.google.com/archive/p/word2vec/

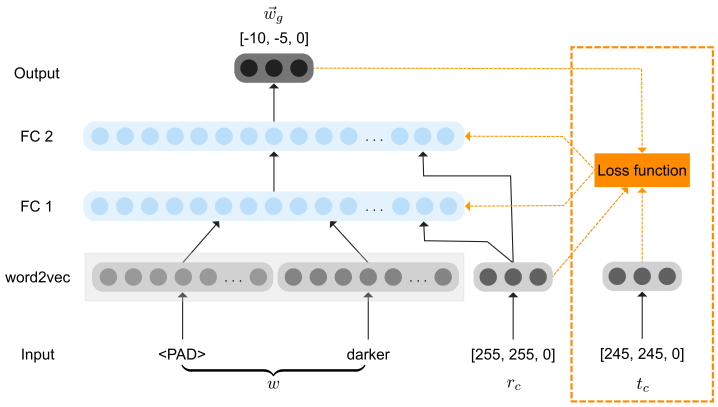

Figure 2: Network Architecture

Data	# Tuples	# Dtpts
Training	271	15.3M
Test (Seen Pairings)	271	2.4M
Test (Unseen Pairings)	29	0.29M
Test (Unseen Ref. Color)	63	2.4M
Test(Unseen Comparative)	41	0.38M
Test (Fully Unseen)	11	58k

Table 1: Data Split

information this dichotomy in size would otherwise produce (an empirical study of various input configurations determined inputting the color into only one of the layers to be insufficient). The output of the first hidden layer has d=30; each layer reduces the dimension of the output by an order of magnitude.

The loss function of the model has two metrics. The first is the cosine similarity between $\vec{w}_g$ and the vector from r_c to t_c. To restrain the length of $\vec{w}_g$, the second metric is the distance between t_c and the result of $\vec{w}_g$ + r_c. Training the length of $\vec{w}_g$ to roughly match the distance between r_c and t_c helps it to capture that the difference should be small enough to warrant a comparison rather than separate descriptors, while still representing enough of a difference to be comparable.

4 Experimental Setup

Table 1 shows the data split between training and testing both in terms of tuples (#Tuples column) and in terms of the actual datapoint instances (#Dtpts column) for our experiments. To properly measure the accuracy of our model, our test set covers five input conditions:

- *Seen Pairings.* The reference color label, the comparative adjective and their pairing have been seen in the training data.

- *Unseen Pairings.* The reference color label and the comparative adjective have been seen in the training data, but not their pairing.

- *Unseen Ref. Color.* The reference color label, and thus all the corresponding RGB color datapoints, have not be seen in training, while the comparative has been seen in the training data.

- *Unseen Comparative.* The comparative adjective has not been seen in training, but the reference color label has been seen.

- *Fully Unseen.* Neither the comparative adjective nor the reference color have been seen in the training.

For the conditions where the reference color label has been seen in training, the actual RGB reference color datapoints associated with the labels were different from the ones used in training: 15% of the datapoints from each training reference color label were set aside for testing, providing RGB values close but not equivalent to those seen in training. 10% of the reference color labels were set aside for testing, as were 10% of the comparative words; this amounted to 8 reference colors and 8 comparatives. The number of tuples and actual datapoints instances for each test condition is given in Table 1.

The network was trained at a 0.001 learning rate for 800 epochs, with the output of the first layer of dimension d=30.

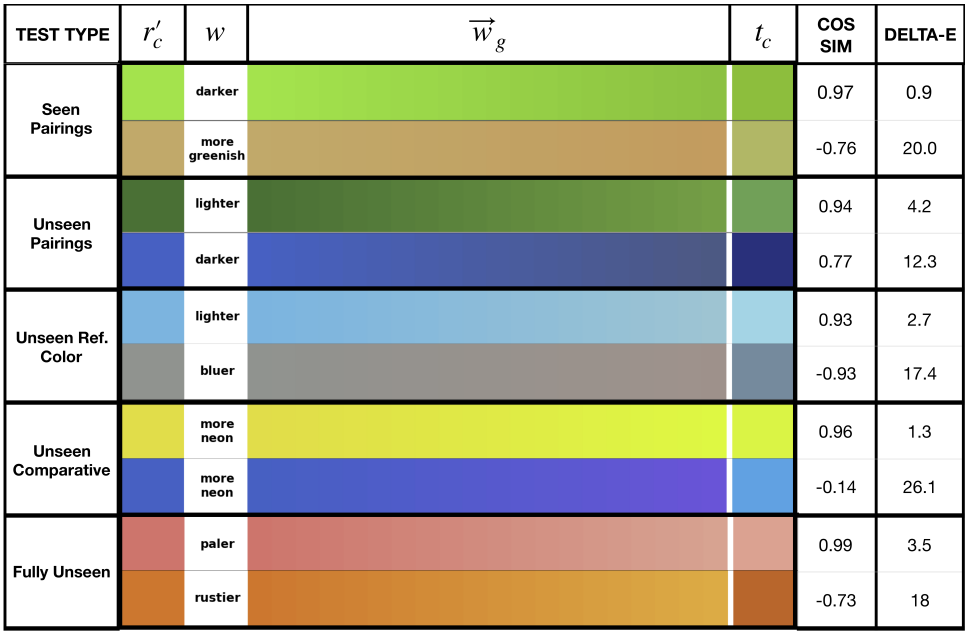

Figure 3: Examples of learned comparatives for each test condition

5 Results

Figure 3 shows examples of learned groundings of comparatives for each of the five test conditions (Test Type column). It shows the reference RGB color datapoint $r_c^{'}$ (always unseen), the comparative word w, the learned grounding vector $\vec{w}_g$, the target color t_c, and two scores: cosine similarity and Delta-E. The upper sample for each test type is an example of a highly accurate result, while the lower sample exemplifies failure.

Delta-E is a metric for understanding how the human eye perceives color differences (Table 2). This is a useful metric as distances in RGB space are not perceived linearly. Figure 4 shows two example pairs of colors which are spaced equally in terms of distance in RGB, but in terms of the Delta-E metric the green colors are closer together.

As seen in Figure 3, grounding comparatives in directional vectors over RGB allows them to capture a full range of modification of the reference color. Even for some of the error cases the resulting outputs tend to capture directions which are reasonable illustrations of the color the comparative described. Though the '**darker**' grounding example from unseen pairings is incorrectly de-saturating the reference color, it is also in fact making the color darker. Most impressive is the '**paler**' example at the bottom, which is able to capture the direction of the comparative almost perfectly. Regarding failures, we see that they tend

Delta-E	Perception
≤ 1.0	Imperceptible
1 - 2	Requires close observation
2 - 10	Perceivable
11 - 49	More similar than opposite
100	Exact opposites

Table 2: Delta-E Ranges

to be of comparatives words that relate to a different color, such as '**more greenish**' and '**bluer**', rather than comparatives such as '**lighter**'.

Table 3 provides quantitative results in terms of average cosine similarity and average Delta-E. Overall, the average cosine similarity is 0.65, with an average Delta-E of 6.8. Separating the performance by test condition, we see that the conditions where the reference and comparatives were both seen perform the best (independent of whether the pairing was seen in training); again 'seen reference' refers only to the label being seen and not the reference color datapoint itself. The fully unseen case performs the worst by far with respect to cosine similarity, though it is not as deviant in Delta-E. It is again apparent that the performance of the model drops when given comparatives which refer to another color.

Figure 5 shows the comparative '**electric**' applied to colors outside of our dataset. With no known t_cs we cannot quantitatively measure the accuracy, but we can qualitatively assess the re-

Figure 4: Same RGB distance, different Delta-E

Test Condition	Avg Cos	Avg Delta-E
Seen Pairings	0.68	6.1
Unseen Pairings	0.68	7.9
Unseen Ref. Color	0.40	11.4
Unseen Comparative	0.41	10.5
Fully Unseen	-0.21	15.9
Overall	**0.65**	**6.8**

Table 3: Results

sults as plausible.

We also examined whether the model could generate plausible comparative terms given a r_c and t_c. All of the comparatives in the model's vocabulary were applied to r_c, and the corresponding $\vec{w}_g$ were sorted by cosine similarity to given reference-target direction. When given a green reference and a dark green target (both sampled from the test data), the model outputs '**truer**', '**deeper**', and '**darker**' as the closest comparatives.

In Figure 6, given a reference sampled from '**purple**' and a target sampled from '**soft purple**', the model outputs the 5 most plausible comparatives - '**softer**' was the 9^{th} closest. They are presented in descending order by distance between the target color and its projection on the modifying vector. We see that the comparatives the model returns are semantically very similar, as are their corresponding $\vec{w}_g$ vectors.

6 Related Work

Though color has been studied in terms of its contextual dependence and vagueness in grounding

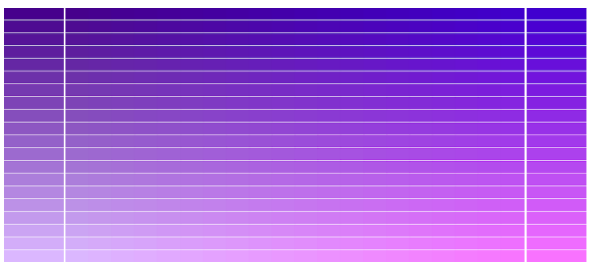

Figure 5: Groundings for '**more electric**'

Figure 6: Top comparatives generated by the model

(Egré et al., 2013; McMahan and Stone, 2015; Monroe et al., 2016, 2017), no approaches have focused explicitly on learning to ground comparatives. Related to this work is that of image ranking, which is inherently a form of comparison (Parikh and Grauman, 2011; Yu and Grauman, 2014). However, ranking methods do not ground the comparatives themselves in image features. Besides the fact that no ranked color data exists, ranking methods are not flexible enough to handle the high dependence of color comparatives on the individual reference color.

7 Conclusion

We propose a new paradigm of grounding comparative adjectives describing colors as directions in RGB space such that the colors along the vector, rooted at the reference color, satisfy the comparison. We introduce a new methodology for transforming labeled color data into comparative color data, and propose a simple but effective learning model that is able to accurately modify unseen colors and comparatives. With respect to the desired output, the representations have an average accuracy of 0.65 cosine similarity, and average Delta-E scores of under 7. Our model can also provide plausible descriptions of the difference between a given reference and target pair, as well as the grounded representations of the comparatives generated, providing an explanation for the model decision. This model is the first step towards fine-grained object recognition through comparative descriptions, providing a way to utilize relational descriptive text. This approach could be extended to other properties such as size, texture, or curvature. It could also be used to aid in zero-shot learning from text sources, generating human-understandable explanations for categorization of similar objects, or providing descriptions of new, unknown objects with respect to known ones.

References

Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. 2012. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3474–3481. IEEE.

Paul Egré, Vincent De Gardelle, and David Ripley. 2013. Vagueness and order effects in color categorization. *Journal of Logic, Language and Information*, 22(4):391–420.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE.

Daniel Lassiter and Noah D Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194(10):3801–3836.

Angeliki Lazaridou, Georgiana Dinu, Adam Liska, and Marco Baroni. 2015. From Visual Attributes to Adjectives through Decompositional Distributional Semantics. *Transactions of the Association for Computational Linguistics*, 3(0):183–196.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association of Computational Linguistics*, 3(1):103–115.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Randall Munroe. 2010. Color survey.

Devi Parikh and Kristen Grauman. 2011. Relative Attributes. *ICCV*, pages 503–510.

Olga Russakovsky and Li Fei-Fei. 2010. Attribute learning in large-scale datasets. In *European Conference on Computer Vision*, pages 1–14. Springer.

Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. 2014. Understanding objects in detail with fine-grained attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3622–3629.

Josiah Wang, Katja Markert, and Mark Everingham. 2009. Learning models for object recognition from natural language descriptions. In *Proceedings of the 20th British Machine Vision Conference (BMVC2009)*.

Aron Yu and Kristen Grauman. 2014. Fine-Grained Visual Comparisons with Local Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.