Character Sequence Models for Colorful Words

Kazuya Kawakami [♠], **Chris Dyer** [♠♣] **Bryan R. Routledge** [♢] **Noah A. Smith** [♡]
[♠]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[♣]Google DeepMind, London, UK
[♢]Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA
[♡]Computer Science & Engineering, University of Washington, Seattle, WA, USA
{kkawakam,cdyer}@cs.cmu.edu, routledge@cmu.edu, nasmith@cs.washington.edu

Abstract

We present a neural network architecture to predict a point in color space from the sequence of characters in the color's name. Using large scale color–name pairs obtained from an online color design forum, we evaluate our model on a "color Turing test" and find that, given a name, the colors predicted by our model are preferred by annotators to color names created by humans. Our datasets and demo system are available online at http://colorlab.us.

1 Introduction

Color is a valuable vehicle for studying the association between words and their nonlinguistic referents. Perception of color has long been studied in psychology, and quantitative models linking physical stimuli and psychological perception have been in place since the 1920s (Broadbent, 2004). Although perceptually faithful color representations require only a few dimensions (§2), linguistic expressions of color often rely on association and figurative language. There are, for example, 34,000 examples of "blue" in our data. The varieties of blue range can be emotional, descriptive, metaphoric, literal, and whimsical. Consider these examples (best viewed in color): murkey blue, blueberry muffin, greeny blue, and jazzy blue.

This rich variety of descriptive names of colors provides an ideal way to study linguistic creativity, its variation, and an important aspect of visual understanding. This paper uses predictive modeling to explore the relationship between colors (represented in three dimensions) and casual, voluntary linguistic descriptions of them by users of a crafting and design website (§3).[1]

In this dataset's creative vocabulary, word-level representations are so sparse as to be useless, so we turn to models that build name representations out of *characters* (§4). We evaluate our model on a "color Turing test" and find that, given a name, it tends to generate a color that humans find matches the name better than the color that actually inspired the name. We also investigate the reverse mapping, from colors to names (§5). We compare a conditional LSTM language model used in caption generation (Karpathy and Fei-Fei, 2014) to a new latent-variable model, achieving a 10% perplexity reduction.

We expect such modeling to find purchase in computational creativity applications (Veale and Al-Najjar, 2015), design and marketing aids (Deng et al., 2010), and new methods for studying the interface between the human visual and linguistic systems (Marcus, 1991).

2 Color Spaces

In electronic displays and other products, colors are commonly represented in RGB space where each color is embedded in $\{0, \ldots, 255\}^3$, with coordinates corresponding to red, green, and blue levels. While convenient for digital processing, distances in this space are perceptually non-uniform. We instead use a different three-dimensional representation, *Lab*, which was originally designed so that Euclidean distances correlate with human-perceived differences (Hunter, 1958). *Lab* is also continu-

[1] http://www.colourlovers.com

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1949–1954,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

	Number of pairs	Unique names
Train	670,032	476,713
Dev.	53,166	52,753
Test	53,166	52,760
ggplot2	66	66
Paint	956	956

Table 1: Datasets used in this paper. The train/dev./test split of the COLOURlovers data was random. For ggplot2 and Paint, we show the number of test instances which are not in Train set.

Model	Test	ggplot2	Paint
Unigram	1018.35	814.58	351.54
Bigram	977.46	723.61	364.41
RNN	750.26	431.90	305.05
1-layer LSTM	664.11	355.56	303.03
2-layer LSTM	652.49	343.97	274.83

Table 2: MSE in *Lab* space on held-out datasets.

ous, making it more suitable for the gradient-based learning used in this paper. The transformation from RGB to *Lab* is nonlinear.

3 Task and Dataset

We consider the task of predicting a color in *Lab* space given its name. Our dataset is a collection of user-named colors downloaded from COLOURlovers,[1] a creative community where people create and share colors, palettes, and patterns. Our dataset contains 776,364 pairs with 581,483 unique names. Examples of the color/name pairs from COLOURlovers are the following: Sugar Hearts You, Vitamin C, Haunted milk.

We considered two held-out datasets from other sources; these do not overlap with the training data.
ggplot2: the 141 officially-named colors used in ggplot2, a common plotting package for the R programming language (e.g., MidnightBlue. MediumSeaGreen),[2]
Paint: The paint manufacturer Sherwin Williams has 7,750 named colors (e.g., Pompeii Red, Butter Up).[3]

4 Names to Colors

Our word-to-color model is used to predict a color in *Lab* space given the sequence of characters in a color's name, $\boldsymbol{c} = \langle c_1, c_2, \ldots, c_{|\boldsymbol{c}|} \rangle$, where each c_i is a character in a finite alphabet. Each character c_i is represented by learned vector embedding in $\mathbb{R}^{300}$. To build a color out of the sequence, we use an LSTM (Hochreiter and Schmidhuber, 1997) with 300 hidden units. The final hidden state is used as a vector representation $\mathbf{h} \in \mathbb{R}^{300}$ of the sequence. The associated color value in *Lab* space is then defined to be $\hat{\mathbf{y}} = \sigma(\mathbf{W}\mathbf{h} + \mathbf{b})$, where $\mathbf{W} \in \mathbb{R}^{3\times 300}$ and $\mathbf{b} \in \mathbb{R}^3$ transform $\mathbf{h}$.

This model instantiates the one proposed by Ling et al. (2015) for learning word embeddings built from representations of characters.

To learn the parameters of the model (i.e., the parameters of the LSTMs, the character embeddings, and $\mathbf{W}$ and $\mathbf{b}$), we use reference color labels $\mathbf{y}$ from our training set and minimize squared error, $||\mathbf{y} - \hat{\mathbf{y}}||^2$, averaged across the training set. Learning is accomplished using backpropagation and the Adam update rule (Kingma and Ba, 2014).

4.1 Evaluation

We evaluated our model in two ways. First, we computed mean-squared error on held-out data using several variants of our model. The baseline models are linear regression models, which predict a color from a bag of character unigrams and bigrams. We compare an RNN and LSTMs with one and two layers. Table 2 shows that the two-layer LSTM achieves lower error than the unigram and bigram baselines and an RNN. We see the same pattern of results on the out-of-domain test sets.

The Color Turing Test. Our second evaluation attempts to assess whether our model's associations are human-like. For this evaluation, we asked human judges to choose the color better described by a name from one of our test sets: our model's predicted color or the color in the data. For each dataset, we randomly selected 20 examples. 111 judges considered each instance.[4] Judges were presented instances in random order and forced to make a choice between the two and explicitly directed to

[2] http://sape.inf.usi.ch/quick-reference/ggplot2/colour

[3] http://bit.ly/PaintColorNames

[4] We excluded results from an additional 19 annotators who made more than one mistake in a color blindness test (Oliver, 1888).

Preference	Test	ggplot2	Paint
Actual color	43.2%	32.6%	31.0%
Predicted color	56.7%	67.3%	69.0%

Table 3: Summary of color Turing test results.

make an arbitrary choice if neither was better.[5] The test is shown at `http://colorlab.us/turk`.

Results are shown in Table 3; on the ggplot2 and Paint datasets, our prediction is preferred to the actual names in a majority of cases. The Test dataset from COLOURlovers is a little bit challenging, with more noisy and creative names; still, in the majority of cases, our prediction is preferred.

4.2 Visualization and Exploration

To better understand our model, we provide illustrations of its predictions on several kinds of inputs.

Character by character prediction. We consider how our model reads color names character by character. Fig. 1 shows some examples, such as *blue*, variously modified. The word *deep* starts dark brown, but eventually modifies *blue* to a dark blue. Our model also performs sensibly on colors named after things (*mint*, *cream*, *sand*).

D	A	M	G	R
De	Aq	Mi	Go	Ro
Dee	Aqu	Min	Gol	Ros
Deep	Aqua	Mint	Gold	Rosy
Deep	Aqua	MintC	Gold	Rosy
Deep B	Aqua B	MintCr	Gold S	Rosy P
Deep Bl	Aqua Bl	MintCre	Gold Sa	Rosy Pi
Deep Blu	Aqua Blu	MintCrea	Gold San	Rosy Pin
Deep Blue	Aqua Blue	MintCream	Gold Sand	Rosy Pink

Figure 1: Visualization of character-by-character prediction.

Genre and color. We can use our model to investigate how colors are evoked in text by predicting the colors of each word in a text. Fig. 3 shows a colored recipe. Noting that many words are rendered in neutral grays and tans, we investigated how our model colors words in three corpora: 3,300 English poems (1800–present), 256 recipes from the CURD dataset (Tasse and Smith, 2008),[6] and 6,000 beer reviews.[7] For each corpus, we examine the distribution of Euclidean distances of $\hat{\mathbf{y}}$ from the *Lab* representation of the "middle" color RGB (128, 128, 128). The Euclidean distances from the mean are measuring the variance of the color of words in a document. Fig. 2 shows these distributions; recipes and beer reviews are more "colorful" than poems, under our model's learned definition of color.

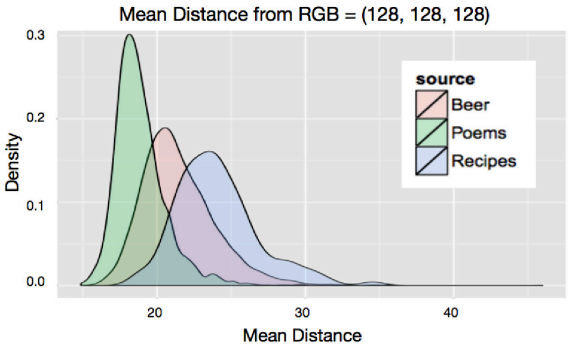
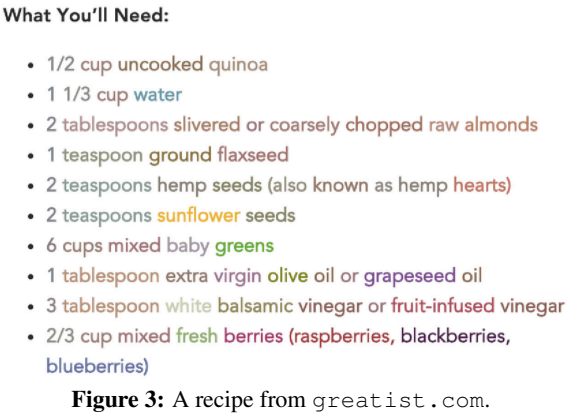


Figure 2: Distribution of Euclidean distances in *Lab* from estimated colors of words in each corpus to RGB (128, 128, 128).


What You'll Need:

- 1/2 cup uncooked quinoa
- 1 1/3 cup water
- 2 tablespoons slivered or coarsely chopped raw almonds
- 1 teaspoon ground flaxseed
- 2 teaspoons hemp seeds (also known as hemp hearts)
- 2 teaspoons sunflower seeds
- 6 cups mixed baby greens
- 1 tablespoon extra virgin olive oil or grapeseed oil
- 3 tablespoon white balsamic vinegar or fruit-infused vinegar
- 2/3 cup mixed fresh berries (raspberries, blackberries, blueberries)

Figure 3: A recipe from `greatist.com`.

5 Generating Names from Colors

The first of our two color naming models generates character sequences conditioned on *Lab* color representations, following other sequence-to-sequence approaches (Sutskever et al., 2014; Karpathy and Fei-Fei, 2014). The transformation is as follows: First, a linear transformation maps the color vector into 300 dimensions, together comprising the initial

[5] A preliminary study that allowed a judge to say that there was no difference led to a similar result.

[6] `http://www.cs.cmu.edu/~ark/CURD/`

[7] `http://beeradvocate.com`

hidden and memory vectors. Next a character LSTM is iteratively applied to the hidden, memory, and next-character vectors, and the next character produced by applying affine and then softmax functions to the hidden vector. The model is trained to maximize conditional likelihood of each character given its history. We used 300 dimensions for character embeddings and recurrence weights. The output vocabulary size was 98 without lowercasing.

We also propose a model to capture variations in color description with latent variables by extending the variational autoencoder (Kingma and Welling, 2013) to a conditional model. We want to model the conditional probability of word $\mathbf{y}$ and latent variables $\mathbf{z}$ given color $\mathbf{x}$. The latent variable gives the model capacity to account for the complexity of the color–word mapping. Since $p(\mathbf{y}, \mathbf{z} \mid \mathbf{x}) = p(\mathbf{z})p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$, the variational objective is:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log q_\phi(\mathbf{z} \mid \mathbf{x}) + \log p_\theta(\mathbf{y}, \mathbf{z} \mid \mathbf{x})] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log q_\phi(\mathbf{z} \mid \mathbf{x}) + \log p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z})p(\mathbf{z})] \\
&\simeq -D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \mid\mid p(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^l)
\end{aligned}$$

The first term regularizes the shape of posterior, $q(\mathbf{z} \mid \mathbf{x})$, to be close to prior $p(\mathbf{z})$ where it is a Gaussian distribution, $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The second term is the log likelihood of the character sequence conditioned on color values. To optimize θ and ϕ, we reparameterize the model, we write $\mathbf{z}$ in terms of a mean and variance and samples from a standard normal distribution, i.e., $\mathbf{z} = \mu + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We predict mean and log variance of the model with a multi-layer perceptron and initialize the decoder-LSTM with $\mathbf{h}_0 = \tanh(\mathbf{W}\mathbf{z} + \mathbf{b})$. We trained the model with mini-batch size 128 and Adam optimizer. The sample size L was set to 1.

Evaluation. We evaluated our models by estimating perplexity on the Test set (Table 1). Our baseline is a character-level unconditional LSTM language model. Conditioning on color improved per-character perplexity by 7% and the latent variable gave a further 3%; see Table 4.

A second dataset we evaluate on is the Munroe Color Corpus[8] which contains 2,176,417 color description for 829 words (i.e., single words have multiple color descriptions). Monroe et al. (2016) have developed word-based (rather character-based) recurrent neural network model.

[8] https://blog.xkcd.com/2010/05/03/color-survey-results/

Model	Perplexity
LSTM-LM	5.9
VAE	5.9
color-conditioned LSTM-LM	5.5
color-conditioned VAE	**5.3**

Table 4: Comparison of language models.

Our character-based model with 1024 hidden units achieved 12.48 per-description perplexity, marginally better than 12.58 obtained with a word-based neural network model reported in that work. Thus, we see that modeling color names as sequences of characters is wholly feasible. However, since the corpus only contains color description for 829 words, the model trained on the Munroe Color Corpus does not provide suitable supervision for evaluation on our more lexically diverse dataset.

6 Related Work and Discussion

Color is one of the lowest-level visual signals playing an important role in cognition (Wurm et al., 1993) and behavior (Maier et al., 2008; Lichtenfeld et al., 2009). It plays a role in human object recognition: to name an object, we first need to encode visual information such as shape and surface information including color and texture. Given a visual encoding, we search our memory for a structural, semantic and phonological description (Humphreys et al., 1999). Adding color information to shape significantly improves naming accuracy and speeds correct response times (Rossion et al., 2004).

Colors and their names have some association in our cognition. The Stroop (1935) effect is a well-known example showing interference of colors and color terms: when we see a color term printed in a different color—blue—it takes us longer to name the word, and we are more prone to naming errors than when the ink matches—blue (De Houwer, 2003).

Recent evidence suggests that colors and words are associated in the brain. The brain uses different regions to perceive various modalities, but processing a color word activates the same brain region as the color it denotes (del Prado Martín et al., 2006; Simmons et al., 2007).

Closer to NLP, the relationship between visual

stimuli and their linguistic descriptions by humans has been explored extensively through automatic text generation from images (Kiros et al., 2014; Karpathy and Fei-Fei, 2014; Xu et al., 2015). Color association with word semantics has also been investigated in several previous papers (Mohammad, 2011; Heer and Stone, 2012; Andreas and Klein, 2014; McMahan and Stone, 2015).

7 Conclusion

In this paper, we introduced a computational model to predict a point in color space from the sequence of characters in the color's name. Using a large set of color–name pairs obtained from a color design forum, we evaluate our model on a "color Turing test" and find that, given a name, the colors predicted by our model are preferred by annotators to color names created by humans. We also investigate the reverse mapping, from colors to names. We compare a conditional LSTM language model to a new latent-variable model, achieving a 10% perplexity reduction.

Acknowledgments

We thank Lucas Beyer for very helpful comments and discussions, and we also appreciate all the participants of our color Turing test.

References

Jacob Andreas and Dan Klein. 2014. Grounding language with points and paths in continuous spaces. In *CoNLL*, pages 58–67.

Arthur D. Broadbent. 2004. A critical review of the development of the CIE1931 RGB color-matching functions. *Color Research & Application*, 29(4):267–272.

Jan De Houwer. 2003. On the role of stimulus-response and stimulus-stimulus compatibility in the Stroop effect. *Memory & Cognition*, 31(3):353–359.

Fermín Moscoso del Prado Martín, Olaf Hauk, and Friedemann Pulvermüller. 2006. Category specificity in the processing of color-related and form-related words: An erp study. *Neuroimage*, 29(1):29–37.

Xiaoyan Deng, Sam K. Hui, and J. Wesley Hutchinson. 2010. Consumer preferences for color combinations: An empirical analysis of similarity-based color relationships. *Journal of Consumer Psychology*, 20(4):476–484.

Jeffrey Heer and Maureen Stone. 2012. Color naming models for color selection, image editing and palette design. In *Proc. CHI*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Glyn W. Humphreys, Cathy J. Price, and M. Jane Riddoch. 1999. From objects to names: A cognitive neuroscience approach. *Psychological Research*, 62(2-3):118–130.

Richard S. Hunter. 1958. Photoelectric color difference meter. *Josa*, 48(12):985–993.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603.

Stephanie Lichtenfeld, Markus A. Maier, Andrew J. Elliot, and Reinhard Pekrun. 2009. The semantic red effect: Processing the word red undermines intellectual performance. *Journal of Experimental Social Psychology*, 45(6):1273–1276.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *CoRR*, abs/1511.04586.

Markus A. Maier, Andrew J. Elliot, and Stephanie Lichtenfeld. 2008. Mediation of the negative effect of red on intellectual performance. *Personality and Social Psychology Bulletin*.

Aaron Marcus. 1991. *Graphic design for electronic documents and user interfaces*. ACM.

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

Saif Mohammad. 2011. Colourful language: Measuring word-colour associations. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–106. Association for Computational Linguistics.

Will Monroe, Noah D. Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proc. EMNLP*.

Charles A Oliver. 1888. Tests for color-blindness. *Transactions of the American Ophthalmological Society*, 5:86.

Bruno Rossion, Gilles Pourtois, et al. 2004. Revisiting snodgrass and vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *PERCEPTION-LONDON-*, 33(2):217–236.

W. Kyle Simmons, Vimal Ramjee, Michael S. Beauchamp, Ken McRae, Alex Martin, and Lawrence W. Barsalou. 2007. A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, 45(12):2802–2810.

J. Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Dan Tasse and Noah A Smith. 2008. Sour cream: Toward semantic processing of recipes. Technical report, Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA.

Tony Veale and Khalid Al-Najjar. 2015. Unweaving the lexical rainbow: Grounding linguistic creativity in perceptual semantics.

Lee H. Wurm, Gordon E. Legge, Lisa M. Isenberg, and Andrew Luebker. 1993. Color improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human perception and performance*, 19(4):899.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.