comp-syn: Perceptually Grounded Word Embeddings with Color

Bhargav Srinivasa Desikan
University of Chicago
Knowledge Lab
bhargav@uchicago.edu

Tasker Hull
Psiphon Inc, Toronto
t.mackersy@psiphon.ca

Ethan O. Nadler
Stanford University
KIPAC & Department of Physics
enadler@stanford.edu

Douglas Guilbeault
University of California, Berkeley
Haas Business School
douglas.guilbeault@berkeley.edu

Aabir Abubaker Kar
University of Chicago
Knowledge Lab
aabir@uchicago.edu

Mark Chu
Columbia University
mbc2165@columbia.edu

Donald Ruggiero Lo Sardo
Sony CSL Paris
losardor@gmail.com

Abstract

Popular approaches to natural language processing create word embeddings based on textual co-occurrence patterns, but often ignore embodied, sensory aspects of language. Here, we introduce the Python package comp-syn, which provides grounded word embeddings based on the perceptually uniform color distributions of Google Image search results. We demonstrate that comp-syn significantly enriches models of distributional semantics. In particular, we show that (1) comp-syn predicts human judgments of word concreteness with greater accuracy and in a more interpretable fashion than word2vec using low-dimensional word–color embeddings, and (2) comp-syn performs comparably to word2vec on a metaphorical vs. literal word-pair classification task. comp-syn is open-source on PyPi and is compatible with mainstream machine-learning Python packages. Our package release includes word–color embeddings for over 40,000 English words, each associated with crowd-sourced word concreteness judgments.

1 Introduction

The embodied cognition paradigm seeks to ground semantic processing in bodily, affective, and social experiences. This paradigm explains how sensory information contributes to semantic processing, either through metaphors involving references to sensory experience (Lakoff and Turner, 1989; Gallese and Lakoff, 2005) or through the simulation of sensory experience in mental imagery (Bergen, 2012). For example, color is pervasive in linguistic metaphors (e.g., "her bank accounts are in the *red*"), and primes a range of affective and interpretative responses (Mehta and Zhu, 2009; Elliot and Maier, 2014).

Extant methods in computational linguistics are limited in their ability to advance the embodied cognition paradigm due to their focus on text, which often precludes multi-modal analyses of images and other forms of sensory data—including color—involved in human meaning-making activities. *Distributional semantics* is a particularly prominent approach, wherein textual co-occurrence patterns are used to embed words in a high-dimensional space; the resulting "distance" between word embeddings correlates with semantic similarity (Landauer and Dumais, 1997). Although leveraging neural networks (Mikolov et al.,

Proceedings of the 28th International Conference on Computational Linguistics, pages 1744–1751
Barcelona, Spain (Online), December 8-13, 2020

Model	Embedding based on	Dimension	Concreteness prediction	Metaphor task
comp-syn	Perceptually uniform color distributions	8–16	Linear: $R^2 = 0.96$ Nonlinear: $\ln \mathcal{L} = 86$	92% test set accuracy
word2vec	Textual co-occurrence	~ 300	Linear: $R^2 = 0.17$ Nonlinear: $\ln \mathcal{L} = 76$	95% test set accuracy

Table 1: Comparison of comp-syn and word2vec and summary of our main results.

2013) and syntactic information (Levy and Goldberg, 2014) has led to recent progress, popular models like word2vec lack firm grounding in their use of sensory information. Moreover, high-dimensional word embeddings created by neural networks are difficult to interpret (Şenel et al., 2018) and require long training periods (Ji et al., 2019). Thus, it remains difficult to reconcile models of distributional semantics with theories of embodied semantics.

To address these limitations, *multi-modal* approaches to distributional semantics combine text and image data. However, these models continue to infer semantic associations using high-dimensional word-plus-image embeddings; their interpretability therefore remains an issue (Bruni et al., 2014; Socher et al., 2014; Lu et al., 2019). With respect to color, multi-modal models operate in standard colorspaces like RGB that are not perceptually uniform (International Commission On Illumination, 1978) and embed color in a manner that combines it with complex, multidimensional spatial information.

Here, we introduce a novel word embedding method based on color that is explicitly interpretable with respect to theories of embodied cognition. We build on work which shows that color distributions in online images reflect both affective and semantic similarities among words in abstract domains (e.g., in the domain of academic disciplines), while also characterizing human judgments of concept concreteness (Brysbaert et al., 2014; Guilbeault et al., 2020). We present the Python package comp-syn[1], which allows users to explore word–color embeddings based on the perceptually uniform color distributions of Google Image search results. We provide embeddings for a set of 40,000 common English words, and we benchmark the performance of our model using crowd-sourced human concreteness ratings (Brysbaert et al., 2014). We show that comp-syn complements the performance of text-based distributional semantics models by providing an interpretable embedding that both (1) predicts human judgments of concept concreteness, and (2) distinguishes metaphorical and literal word pairs.

2 Python Package: comp-syn

comp-syn is an open-source Python package available on GitHub and downloadable through PyPi. The code follows Python best practices and uses industry standard packages for scientific computing, facilitating easy integration with Python code bases; we provide complete details in the Supplementary Material. comp-syn creates word–color embeddings by computing the perceptually uniform $J_z A_z B_z$ (Safdar et al., 2017) color distributions of their corresponding top 100 Google Image search results. We represent these distributions by their mean and (optionally) standard deviation in 8 evenly-segmented $J_z A_z B_z$ bins, yielding 8 to 16-dimensional word–color embeddings. The details of our search and our method for generating word–color embeddings are described in Appendix A.

We use Google rather than a curated image database such as ImageNet (Deng et al., 2009) because popular datasets are heavily biased toward concrete objects, which limits their applicability in abstract semantic domains. Moreover, Google Image search results reflect content that users interact with most frequently (Jing and Baluja, 2008), underscoring their relevance for connecting distributional properties of words and images to human semantic processing. To ensure that our approach is entirely unsupervised, we do not select particular features of images when measuring their color distributions. This approach avoids importing pre-determined semantic notions into our analysis and connects our method to cognitive theories that attribute aesthetic relevance to recognizable features in both the background and foreground of images (Riley, 1995; Elliot and Maier, 2014; Guilbeault et al., 2020).

[1] Short for "Computational Synaesthesia."

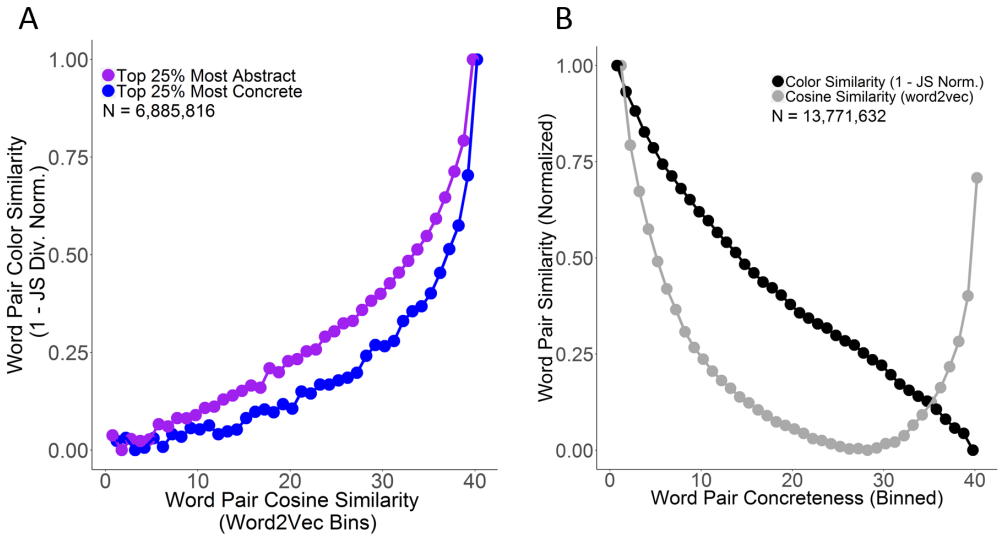

Figure 1: The semantic coherence of comp-syn. (A) Word pair color similarity in comp-syn, measured using the Jensen-Shannon divergence between their perceptually uniform color distributions, correlates with word2vec cosine similarity ($N > 10^6$ pairs). This holds for both concrete (blue) and abstract (magenta) word pairs, rated according to crowd-sourced human judgments. Data represent average color similarity in 40 equally-sized bins of word2vec similarity. (B) Word pair similarity vs. human concreteness judgments ($N > 10^7$ pairs). comp-syn similarity monotonically decreases with word pair concreteness, while word2vec similarity is a nonlinear function of word pair concreteness. Data represent average word pair similarity in word2vec (gray) and comp-syn (black) in 40 equally-sized bins of summed word pair concreteness.

3 Results

Table 1 summarizes our main results, which we now describe in turn.

3.1 Comparison to word2vec Similarity

We begin by demonstrating that pairwise word similarity is highly correlated in comp-syn and word2vec, illustrating the semantic coherence of our color embeddings. We calculate pairwise distances between each of the 40,000 words in our dataset and 500 randomly-selected words from the same set, yielding over 10^7 distinct pairs. We compare the cosine similarity between word pairs in the Mikolov et al. (2013) word2vec model with the Jensen-Shannon (JS) divergence between $J_z A_z B_z$ distributions in comp-syn. Fig. 1A shows that JS divergence in comp-syn is significantly correlated with cosine similarity in word2vec ($p < 0.00001$, JT $= 773$, Jonckheere-Terpstra test). This implies that our low-dimensional, interpretable embedding captures aspects of the key information contained in a widely-used distributional semantics model.

3.2 Relation to Human Concreteness Judgments

Next, we examine the relation between our word–color embeddings and human concreteness judgments (Brysbaert et al., 2014). We label word pairs with summed concreteness ratings in the highest (lowest) quartile of our data as "concrete" ("abstract"). Fig. 1A shows that abstract word pairs are more similar in colorspace than concrete word pairs, even at fixed word2vec similarity ($p = 0.001$, DF $= 77$, Dickey-Fuller test). Moreover, as shown in Fig. 1B, comp-syn captures concreteness judgments in a more interpretable fashion than word2vec. In particular, because abstract words are more similar in colorspace (on average), there is a monotonic relationship between color similarity and concreteness; indeed, a linear model of comp-syn similarity accounts for nearly all of the variance in word pair concreteness ($R^2 = 0.96$). On the other hand, word2vec similarity is a nonlinear function of word pair concreteness. Although nonlinear functions of word2vec similarity also predict concreteness

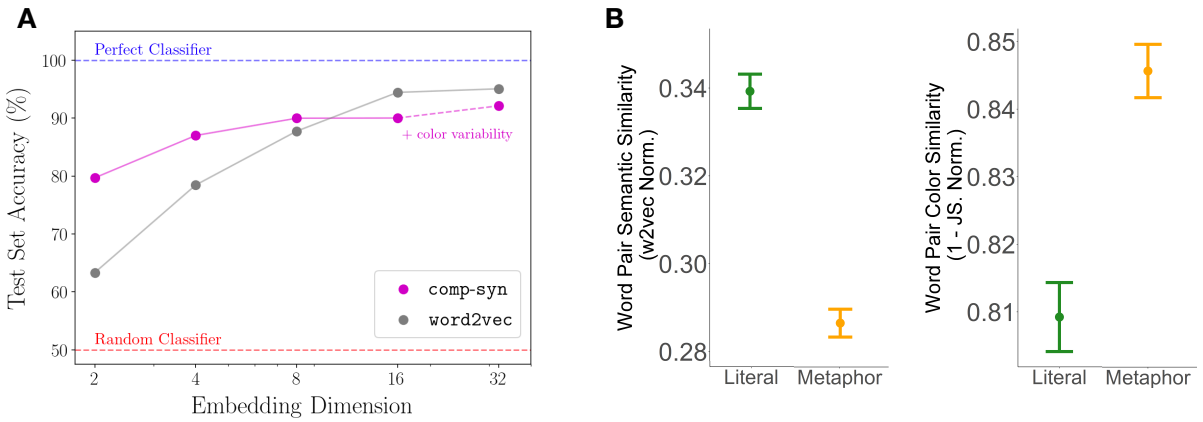

Figure 2: Classifying metaphorical vs. literal adjective-noun pairs with comp-syn. (A) Test set classification accuracy for comp-syn (magenta) and word2vec (gray) as a function of PCA embedding dimension. (B) Average similarity of metaphorical and literal adjective-noun pairs in word2vec (left panel) and comp-syn (right panel). Error bars indicate 95% confidence intervals.

well ($R^2 > 0.99$), comp-syn versions of the same nonlinear models are significantly more accurate (log-likelihood difference $\Delta \ln \mathcal{L} \sim 10$ in favor of comp-syn) and less complex (Bayesian information criterion $\Delta\text{BIC} \sim 20$, also in favor of comp-syn).

To qualitatively explore the relationship between our word–color embeddings and human concreteness judgment, Fig. 3A shows the perceptually uniform color distributions associated with some of the most and least concrete words in our corpus. These *colorgrams* illustrate that concrete words (e.g., "pyramid") are often associated with color distributions that are peaked in specific regions of colorspace, while abstract words (e.g., "concept") feature more variegated color distributions. The spatial and textural features of the images reflect these properties, and exploring the relationship between these aspects of *colorgrams*, is an interesting avenue for future study.

3.3 Metaphor Pair Analysis

The results above suggest that our word–color embeddings encode complementary information about concept concreteness relative to purely textual embeddings. This raises the question of what can be learned from cases in which word2vec and comp-syn provide conflicting similarity predictions when evaluated on the same pair of words. Here, we address this question by demonstrating that comp-syn significantly enriches metaphorical word pair classification, which often requires extensive manual tagging due to subtle uses of both sensory and abstract features (Lakoff and Johnson, 2008; Bethard et al., 2009; Indurkhya and Ojha, 2013; Dodge et al., 2015; Winter, 2019).

We trained a gradient-boosted tree classifier implemented via XGBoost (Chen and Guestrin, 2016) to label adjective-noun pairs as either metaphorical or literal, using over 8000 word pairs from Tsvetkov et al. (2014) and Gutiérrez et al. (2016). This dataset encompasses a statistically representative range of metaphorical and literal contexts for each adjective (Gutiérrez et al., 2016). To compare embeddings, we compressed word2vec's 300-dimensional word vector differences using PCA to match the dimensionality of comp-syn, following Bolukbasi et al. (2016). Fig. 2A shows that a classifier trained using only word2vec achieves a limiting test set accuracy of 95%, compared to 92% for comp-syn. Importantly, comp-syn outperforms word2vec at low embedding dimensions, indicating that it captures semantic content in an interpretable fashion. This analysis does not demonstrate either model's best-case performance on this task; rather, it highlights the complementary information provided by comp-syn.

Strikingly, word2vec and comp-syn distances provide qualitatively different information when distinguishing literal and metaphorical adjective-noun pairs. Fig. 2B shows that literal adjective-noun pairs are more similar than metaphorical pairs in word2vec ($p < 0.001$, Wilcoxon rank sum); the

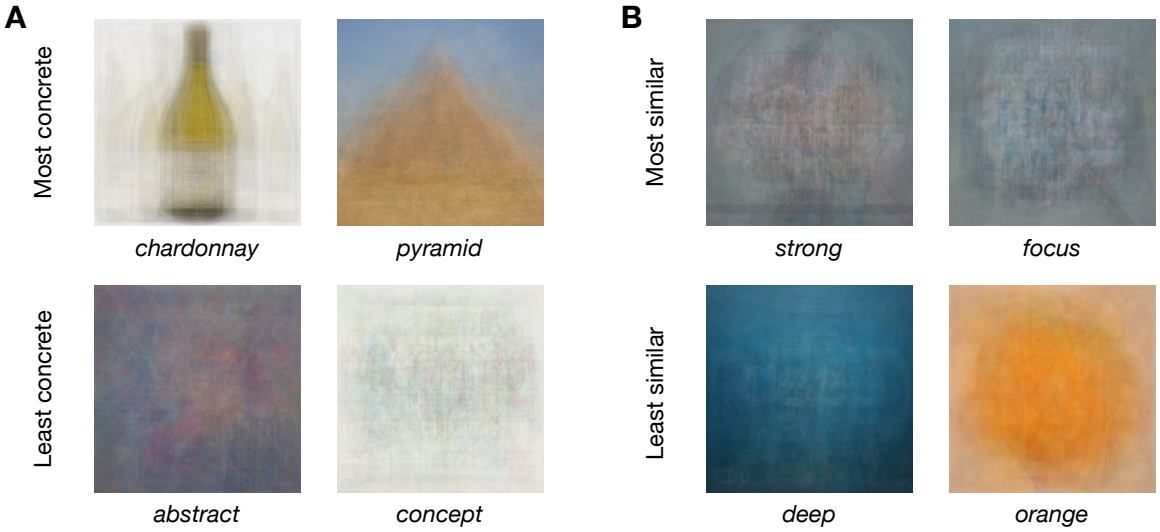

Figure 3: (A) Examples of the most and least concrete terms, visualized using our word–color embedding method. (B) Examples of the most and least similar adjective-noun pairs according to `comp-syn` word–color embeddings. We represent each term using a *colorgram*, i.e., a composite image produced by averaging the perceptually uniform colors of pixels across Google Image search results.

reverse holds in `comp-syn`, where literal pairs are significantly *less* similar than metaphorical pairs ($p < 0.001$, Wilcoxon rank sum). This is a consequence of the fact that images associated with concrete words are more variable in colorspace (Guilbeault et al., 2020). In this way, `comp-syn` reveals differences in color similarity between literal and metaphorical adjective-noun pairs that are of interest for cognitive theory (Indurkhya and Ojha, 2013). Particularly, our findings suggest that metaphors can exploit color similarities between words that are dissimilar in textual embeddings, which may help facilitate cognitive processing of semantic relations among concepts from distinct domains (Guilbeault et al., 2020).

Qualitative inspection of specific adjective-noun word pairs highlights some notable differences between textual and word–color embeddings. Fig. 3B provides visual representations of the color distributions for word pairs in our metaphorical vs. literal dataset that are most and least similar in `comp-syn`. Interestingly, while metaphorical pairs are more similar than literal pairs in `comp-syn`, the *least* similar metaphorical pairs explicitly invoke color, e.g., "deep orange". Algorithmically, this is due to the fact that `comp-syn` embeddings associated with color terms are unusually coherent. On the other hand, the color distributions associated with the most similar pairs in `word2vec` (e.g., "bushy beard") often noticeably contrast, while `word2vec`'s least similar pairs (e.g., "rough customer") do not strongly invoke color. These findings point to an important direction for future research enabled by the grounded nature of `comp-syn`: how do linguistic metaphors leverage sensory information to characterize colorspace itself (e.g., in the use of spatial information in the popular metaphor "deep purple")?

4 Conclusion and Future Work

We have presented `comp-syn`, a new Python package that provides perceptually grounded color-based word embeddings. These embeddings are interpretable with respect to theories of embodied cognition because (1) `comp-syn` represents color in a fashion that emulates human perception, and (2) `comp-syn` leverages Google Image search results that human users dynamically interact with and produce. By linking `comp-syn` with human concreteness judgments for 40,000 common English words, our package provides a multi-modal playground for exploring grounded semantics. We demonstrated that `comp-syn` enriches popular distributional semantics models in both word concreteness prediction and metaphorical word-pair classification. A myriad of `comp-syn` applications await, including color-based classification of text genres and the characterization of sensory imagery in everyday language.

References

Benjamin K Bergen. 2012. *Louder than words: The new science of how the mind makes meaning*. Basic Books.

Steven Bethard, Vicky Lai, and James Martin. 2009. Topic model analysis of metaphor frequency for psycholinguistic stimuli. *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 9–16.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: Deep semantic automatic metaphor analysis. *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49.

Andrew Elliot and Markus Maier. 2014. Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual Review of Psychology*, 65:95–120.

Vittorio Gallese and George Lakoff. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, 22(3-4):455–479.

Douglas Guilbeault, Ethan O Nadler, Mark Chu, Donald Ruggiero Lo Sardo, Aabir Abubaker Kar, and Bhargav Srinivasa Desikan. 2020. Color associations in abstract semantic domains. *Cognition*, 201:104306.

E. Dario Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, Berlin, Germany, August. Association for Computational Linguistics.

Bipin Indurkhya and Amitash Ojha. 2013. An empirical study on the role of perceptual similarity in visual metaphors and creativity. *Metaphor and Symbol*, 28.

International Commission On Illumination. 1978. Recommendations on uniform color spaces, color-difference equations, psychometric color terms. *Color Research & Application*, 15.

S. Ji, N. Satish, S. Li, and P. K. Dubey. 2019. Parallelizing word2vec in shared and distributed memory. *IEEE Transactions on Parallel and Distributed Systems*, 30(9):2090–2100.

Yushi Jing and Shumeet Baluja. 2008. Pagerank for product image search. In *Proceedings of the 17th international conference on World Wide Web*, pages 307–316.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

George Lakoff and Mark Turner. 1989. *More than cool reason: A field guide to poetic metaphor*. University of Chicago press.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Ravi Mehta and Rui Zhu. 2009. Blue or red? exploring the effect of color on cognitive task performances. *Science*, 323:1226–1229.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Charles Riley. 1995. *Color Codes: Modern Theories of Color in Philosophy, Painting and Architecture, Literature, Music, and Psychology*. UPNE.

Muhammad Safdar, Guihua Cui, Youn Jin Kim, and Ming Ronnier Luo. 2017. Perceptually uniform color space for image signals including high dynamic range and wide gamut. *Optics express*, 25(13):15131–15151.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland, June. Association for Computational Linguistics.

Bodo Winter. 2019. *Sensory Linguistics*. John Benjamins Publishing Company.

L. K. Şenel, İ. Utlu, V. Yücesoy, A. Koç, and T. Çukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779.

A Methods

A.1 Google Image Search

To generate word–color embeddings, we collect the top 100 Google Image search results for each of the 40,000 terms in our analysis. The Google Image searches were run from 10 servers running in a commercial datacenter in New York, USA. These servers were created and used only for this experiment, so that results would not be overly-personalized. Additional search parameters were included as query strings `safe=off&site=&tbm=isch&source=hp&gs_l=img`.

A.2 Word–Color Embeddings

We use the `PIL` Python module to convert each image into an $m \times n \times 3$ array of sRGB values, where m and n are the intrinsic image dimensions. For computational efficiency, we then compress each image into an anti-aliased $300 \times 300 \times 3$ array. Next, we transform sRGB pixel values into their counterparts in the perceptually uniform $J_z A_z B_z$ colorspace. Unlike in standard colorspace, Euclidean distances in $J_z A_z B_z$ coordinates linearly correspond to differences in human color perception (Safdar et al., 2017). Moreover, our use of the $J_z A_z B_z$ colorspace rather than a standard colorspace like RGB increases the semantic coherence of our word–color embeddings (Guilbeault et al., 2020).

We measure the color distribution of each image in 8 evenly-segmented $J_z A_z B_z$ subvolumes spanning the range of $J_z A_z B_z$ coordinates that maps to all possible RGB tuples: $J_z \in [0, 0.167]$, $A_z \in [-0.1, 0.11]$, $B_z \in [-0.156, 0.115]$. Next, we average $J_z A_z B_z$ distributions over all 100 images for each term to obtain an aggregate, 8-dimensional color mean embedding. We also compute the standard deviation over the 100 images in each $J_z A_z B_z$ subvolumes to obtain an aggregate, 8-dimensional color variability embedding. These color mean and variability embeddings can be concatenated to create a 16-dimensional embedding. The details of our compression, binning, and averaging steps do not affect our results (Guilbeault et al., 2020).

To compare word–color embeddings, we use the Jensen-Shannon (JS) divergence to measure the similarity of aggregate $J_z A_z B_z$ distributions. In particular, for $J_z A_z B_z$ distributions C_i and C_j associated with terms i and j, the JS divergence is given by

$$D_{\mathrm{JS}}(C_1 \,||\, C_2) \equiv \frac{1}{2}\left[D_{\mathrm{KL}}(C_1 \,||\, \bar{C}_{12}) + D_{\mathrm{KL}}(C_2 \,||\, \bar{C}_{12})\right], \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence and $\bar{C}_{12} \equiv (C_1 + C_2)/2$. The JS divergence is a measure of the distance between two color distributions, such that lower values correspond to more similar distributions in perceptually uniform colorspace. We choose this metric because it is a well-defined distance measure that satisfies the triangle inequality and allows us to avoid undefined values associated with empty $J_z A_z B_z$ bins. Terms with relatively high mutual JS divergences usually exhibit *colorgrams* with perceptibly different average colors.

B Best Practices for Usage

We caution that our word–color embeddings are *distributions* rather than *vectors*. Thus, their components are positive semidefinite, and different embeddings must be compared using similarity measures designed for distributions such as JS divergence. On the other hand, `word2vec` embeddings are vectors with components that can be positive or negative, and are compared using similarity measures designed for vectors such as cosine similarity. Importantly, unlike `word2vec` embeddings, our word–color embeddings cannot be composed by vector addition. We are exploring algebraic techniques for word composition in colorspace; these techniques must respect the underlying mathematical structure of $J_z A_z B_z$ (and other) colorspaces, which are not closed under standard binary operations like addition or multiplication.