Weak Sapir-Whorf for computers:
Exploring the impact of language on color representation in multimodal variational autoencoders

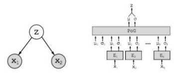
Ben Peloquin
Collaborators: Mike Wu

Introduction

How does the language you know impact the way you perceive the world? The Sapir-Whorf hypothesis argues that linguistic knowledge broadly impacts cognition. We test a weak version of this hypothesis exploring the effect of language supervision on learned representations for color data in multimodal variational autoencoders (MVAEs). MVAEs can learn a joint distribution over separate modalities (e.g. visual and linguistic) using a product-of-experts inference network and optimizing the ELBO objective (see model).

Model

MVAEs are latent variable generative models of the form p_{theta}(x, z) = p_{theta}(x|z)*p(z) with a neural decoder p_{theta}(x|z), spherical gaussian prior p(z).



The target of training is to maximize the evidence lower bound objective, which avoids the intractable marginal likelihood of the data.

$$\mathbb{E}_{q_\phi(z|x)}[\lambda p_\theta(x|z)] - \beta KL[q_\phi(z|x)||p(z)]$$

Data

Monroe et al. (2017) collected data from a color-reference game experiment. In each trial speakers and listeners were presented with three colors. The speaker's goal to produce a referring expression so that the listener could select the correct target color. There were several conditions varying difficulty via the similarity of the distractor colors. We use the target colors and their descriptions for the current analysis. See below for an example from their data set.

Context			Utterance
🟦	🟦	🟦	darker blue

Experiments

1. Can we learn a joint distribution over colors and utterances?
2. What do the learning trajectories look like? (think Berlin & Kay, 1969)
3. Can we explore semantics as transformations in latents space? E.g. terms like "-ish", "-er" and "-est"?
4. Does language "supervision" give rise to different clusters? tSNE and clustering statistics.

References

Berlin & Kay (1969). Basic Color Terms. University of California Press, Berkeley.

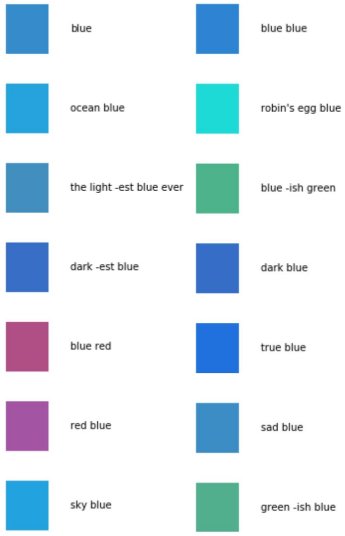
Monroe, W., Hawkins, R., Goodman, N., Potts, C., (2017). Color in Context: A pragmatic Neural Model for Grounded Language Understanding. *TACL*.

Wu & Goodman (2018). Multimodal Generative Models for Scalable Weakly-Supervised Learning. arXiv preprint arXiv:1802.05335

Zhao, S., Song, J., Ermon, S. (2017) "Towards deeper understanding of variational autoencoders". arXiv preprint arXiv: 1702.08658,2017

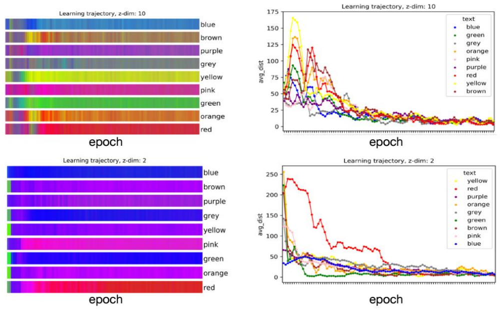
Experiment 1. Sampling from the joint distribution over modalities — words for "blue"

The model learns a joint distribution over descriptions (language) and colors (RGB). As a first test we can examine color reconstructions for novel descriptions. The model is sensitive to word order (blue-green vs green-blue) and compositional structure (blue-ish vs blue).



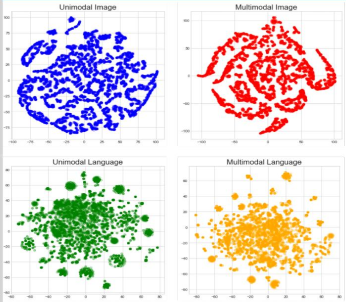
Experiment 2. Learning basic color terms

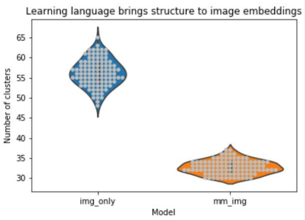
[LEFT column] What happens when we constrain the latent space dimensionality? The model can't represent the full color space with a latent dimension of two. **[RIGHT column]** Primary colors like red, blue and green appear to be learned first with less variance.



Experiment 4. Inspecting the latent space

[Left quadrant] tSNE plots of the latent space. Knowing language information brings more images together, but knowing color (image) information has lesser impact on language clusters. The linguistic data appears to be provide more statistical structure. **[Right quadrant]** Measuring the effect of language — clustering statistic — we cluster colors and descriptions using non-parametric Dirichlet Process Gaussian Mixture Model allowing us to infer the optimal number of clusters. We bootstrap this measure n=100 times using the number of clusters as test statistic — learning a joint distribution over language and images brings more structure to image clusters.

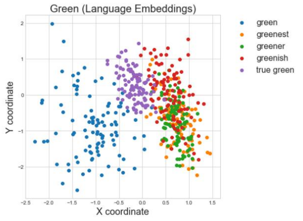


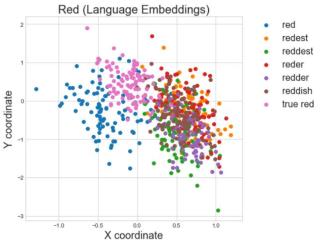


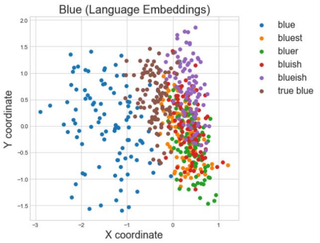
t(198) = 61.6, p < 1e-100

Experiment 3. Semantics in the latent space

We can explore the semantics of morphemes like "-ish", "-est" and "-er" as well as compositionality like "true X." Plots: we sample many embeddings for the color descriptions (e.g. 'green', 'true green') and plot their principle components (with a pre-trained PCA on a dev set). Note the shifts induced by "green"-> "true green" -> "greenest" and other colors. Note, plotted colors **DO NOT** correspond to reconstructed RGB values for readability.


Green (Language Embeddings)
Y coordinate
X coordinate
green
greenest
greener
greenish
true green


Red (Language Embeddings)
Y coordinate
X coordinate
red
redest
reddest
reder
redder
reddish
true red


Blue (Language Embeddings)
Y coordinate
X coordinate
blue
bluest
bluer
bluish
blueish
true blue