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

Ben Peloquin

Abstract

How does the language you speak impact the way you see the world? We explore this question in computers -- how does jointly training on linguistic and image data impact latent representations in a neural model? We train a multimodal variational autoencoder on a color reference game dataset, learning a joint distribution over colors and color descriptions. We show that the model can successfully learn a joint distribution over modalities, and perform a series of experiments on the learned representations. We show that joint training with language gives rise to interesting effects, including inducing more categorical clustering in latent space.

Introduction

How does the language you speak impact the way you see the world? The Sapir-Whorf hypothesis presents a strong claim of linguistic relativism -- that language *determines* thought and that cognitive categories are primarily derived from linguistic categories (Kay & Kempton, 1984). While controversial, this hypothesis has inspired work for the last five decades and remains a current area of research (Berlin & Kay, 1994; Gibson, et al. 2017; Regier & Xu, 2017). Recent computational formulations of linguistic relativism have found success in taking a moderated approach -- perhaps language impacts cognition some of the time, but not always (Regier & Xu, 2017). We adopt this moderated view and apply it it to computers. How does learning a joint distribution over perceptual and linguistic modalities impact representations in neural models? We train a multimodal variational autoencoder (MVAE) to learn to joint distribution over colors (image modality) and color-descriptions (language modality). We propose a set of experiments which interrogate various aspects of the learned embeddings. As a sanity check we first show (1) that MVAEs can learn joint distributions over image and language modalities when the latent space is not overly constrained. Next we perform a series of experiments (2) examining learning trajectories for basic color terms, grounding the analysis in previous color theory, (3) assess the degree to which we can begin to measure semantics as transformations in latent space and finally (4) attempt to provide qualitative and quantitative characterizations of the impact of language on latent representations of images.

Dataset

Monroe et. al (2017) introduce a color reference game dataset using a online behavioral experiment, which pairs workers on mechanical turk (using a framework introduced by Hawkins, 2015). During the reference game, workers are split into speaker/listener pairs. During each trial the participants view a set of three colors. One of the colors is assigned as the target. This information is made known only to the speaker. The speaker's job is to provide a description of the target color so that the listener can pick it out of the set of colors. The experimenters manipulated task difficulty in three conditions by varying the distance between the colors. Intuitively, in contexts in which multiple colors were "closer" in perceptual space, descriptions are required to be longer to differentiate among them, whereas we expect the opposite when the colors are distinct. Rather than focus on condition-level effects, we train our model only on the target RGB color and accompanying description. (Future work will attempt to capture the pragmatic effects of context.) This dataset is amenable to modeling with MVAE models in that the image and language modalities are both sufficiently complex to provide interesting variation in the data while also being fairly simple. (A known issue with VAEs is blurriness in modeling complex images.) The image modality data consists of RGB pixels while the descriptions are typically between 2-7 words (max 18 words).

Model

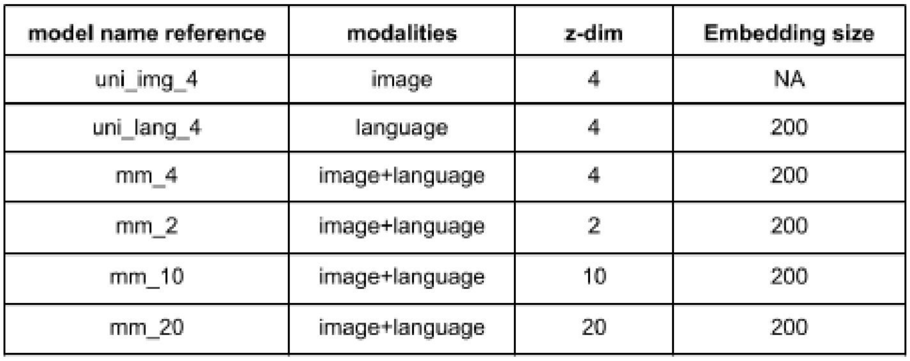
Variational autoencoders (VAEs) are latent variable generative models which capture $p_\theta(x, z) = p_\theta(x|z)p(z)$, where $p(z)$ is a prior over the latent variable z and $p_\theta(x|z)$ is a "decoder" model, typically implemented as a neural network with parameters θ . The objective, to maximize the marginal likelihood of the data $p(x)$, is intractable, so the evidence lower bound (ELBO) objective is optimized instead. The ELBO is defined using an inference network $q_\Phi(z|x)$:

$$E_{q_\Phi(z|x)}[\lambda log(p_\theta(x \mid z)] - \beta KL[\, q_\Phi(z|x) \,||\, p(z)]$$

Typically $\lambda = 1$ and β is annealed to 1 while training. Multimodal variational autoencoders extend the basic VAE framework to n modalities, which are conditionally independent given the latent variable z . Wu & Goodman (2018) present a version of this framework which uses a product of experts as the approximating distribution for the joint posterior.

A known difficulty training VAE models is that the regularizing KL term can be overly influential early in training. This can lead to learning a degenerate joint distribution highly concentrated around modal values in the dataset. Standard solutions include annealing the KL-term during early epochs (Bowman et. al, 2015). We faced a similar problem. Given our training paradigm, which used a single pixel representation, our reconstruction loss was about an order of magnitude smaller than the regularizing term leading to degenerate latent representations. Previous work with MVAEs on image data used images that were 64x64, leading to higher reconstruction losses. Increasing the image size to 64x64 or, equivalently, scaling the reconstruction loss by the constant factor 4096 solved the problem of only learning modal colors.

In the following experiments, we compare the latent embeddings in several models under different parameters settings. Of particular importance is the size of the latent dimension (*z-dim)* and the number of modalities (multimodal vs unimodal). Our primary interest is in comparing the learned representations of uni-modal models to multimodal models, particular the uni-modal image to multimodal language and image model. Table 1 contains information about the trained models for reference.



model name reference	modalities	z-dim	Embedding size
uni_img_4	image	4	NA
uni_lang_4	language	4	200
mm_4	image+language	4	200
mm_2	image+language	2	200
mm_10	image+language	10	200
mm_20	image+language	20	200

Table 1. Summary of trained models parameterizations.

Experiments

Experiment 1: Sanity check - learning a joint distribution over color/color-descriptions.

Because MVAEs model a joint distribution over modalities we can sample from the latent embedding space. As qualitative check we can examine the color/color-descriptions generated from this space. Figure 1. Shows a set of sample image reconstructions given various descriptions of the color “blue.” Clearly the model is learning reasonable associations between colors and images capturing semantics of various linguistic devices such as word order (e.g. green-blue is different from blue-green), compositionality (e.g. green-ish blue), even novel descriptions (e.g. “robin’s egg blue”)

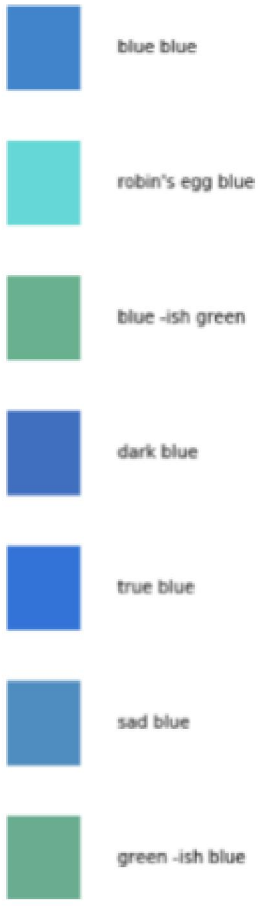


Figure 1. Image reconstructions of color descriptions for different blues.

Experiment 2: Basic color learning trajectories

Berlin & Kay (1969) is seminal work on the psycholinguistics of color. While their broader goal was to propose a set of universals in linguistic description of color space, they also proposed an ordering for learned color terms. Under their proposal “white” and “black” will be learned first, then “red”, then “green”, then “yellow”. While “white” and “black” were not colors in our dataset, we can examine basic color learning trajectories for the other colors in our models. For each epoch we encode the basic color terms “red”, “blue”, “green”, “brown”, “grey”, “yellow”, “purple” and “pink” and reconstructed the RGB values the model has learned (data shown uses *mm_4)*. We can inspect the color reconstructions at each epoch (figure 2 left facet) as well as the distance from the final color the model converges to at the final epoch (figure 2 right facet). Qualitative analysis indicates that the colors “green”, “blue”, and “red” appear to learned slightly earlier (around epoch 20) and vary less from this point on, compared to the colors “yellow”, “brown” and “grey.” This may be an artifact of our use of RGB to describe the color space. Future work should experiment with additional color description such as HSL and CIELAB space.

Experiment 3: Semantics in the latent space

While the previous analysis focused on single word descriptions of color, these types of responses were actually infrequent. More common were rich descriptions employing a variety of linguistic devices (e.g. compositionality, comparatives, superlatives, negation, etc.). Among the most common devices were suffixes such as “-ish”, “-er” and “-est” as in “blue-ish”, “blue-er” and “blue-est”. How does appending the suffix “-ish” impact the models encoding of “blue” in the latent space? We show an initial description of semantics in latent space in *Figure 3*. Of note, using the adjective “true” appears to describe the central point in latent space for the colors “green”, “blue” and “red”. Future analyses will extend this description of latent-space semantics.

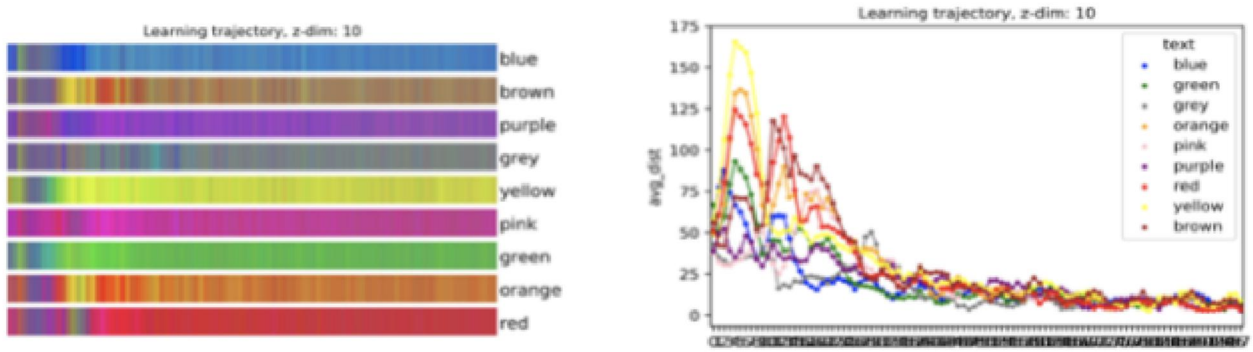


Figure 2. Basic color term learning trajectories. Left facet: horizontal axis is training epoch (1-100), vertical axis is basic color term. Colors are current RGB reconstructions for the basic color term. Right facet: horizontal axis is training epoch (1-100). Vertical axis is euclidean distance in RGB space from final color converged to.

Experiment 4: Quantifying the impact of linguistic knowledge on image representations

Given our framing of this project in terms of the Sapir-Whorf hypothesis -- that language has some impact on cognition -- we’d like to quantify the extent to which *knowing* language impacts what you know in other domains, or at least, how that knowledge is structured. *Figure 4* shows tSNE plots comparing the latent representations for image encodings in a unimodal image model (*uni_img_4*) compared to image-only encodings in a multimodal model which has learned a joint distribution over colors and images (*mm_4*). For comparison, we show the tSNE plots of language encodings in a unimodal language model (*uni_lang_4*) with language-only encodings in a multimodal model (*mm_4*). Two things are of note. First, the tSNE plot of *image clusters* trained with the multimodal model appear to show more categorical structure (more well-defined clusters). Intuitively, this appears to indicate that the linguistic knowledge is providing additional structure over and above what is learned with image-data alone. The same does not appear to hold in the case of language -- adding knowledge of colors does not appear to significantly alter the clusters learned for language. To test the first observation we can examine whether there is a higher degree of clustering in the image embeddings. We use a non-parametric clustering algorithm -- Bayesian Gaussian Mixture Model with Dirichlet Process prior, which allows us to infer an approximate distribution over the parameters of a Gaussian mixture distribution. This formulation allows us to infer the number of components from the data. We run this clustering algorithm for 100 simulations on our embeddings from the *uni_img_4* and *mm_4*. These simulations give us a distribution over the number components. We use a test-statistic, comparing the number of components using an independent samples t-test. Results indicate that the training with language results in clusters that have significantly more categorical structure, reducing the number of components by almost half. *Figure 5* shows the results of this analysis.

Conclusion

Jointly training a neural model on visual and linguistic data gives rise to a set of interesting changes in the learned representations. In particular, models that had linguistic knowledge (jointly trained on language- and image-data) induced more categorical structure when we examined embeddings for images alone. The same did not appear to hold for embeddings of language alone. This appears to indicate a possibly privileged role that language plays in providing categorical scaffolding in other domains. We present these findings in light of the Sapir-Whorf hypothesis which predicts that language uniquely defines conceptual structure in humans. Future work will extend the analyses started here, in particular, examining the extent to which we can characterize the *kind of categorical structure induced by language* on other domains, like the visual domain of color perception. As a by-product of this framework we can also examine semantics in the latent space, how linguistic devices give rise to regular transformations in embeddings.

Collaborators

This project is a joint collaboration with my labmate Mike Wu. We worked on the modeling and experiments together.

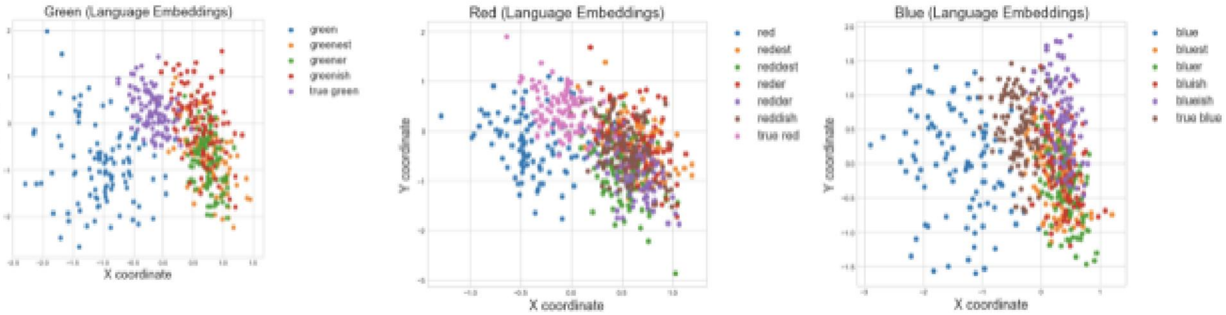


Figure 3. Semantics in the latent space. Axes are first two principal components from PCA trained on full dataset. Points are samples from latent space for color terms in varying semantic contexts. We see some consistencies, for example the description “True X” typically results locations *between* “X” and “X-er”.

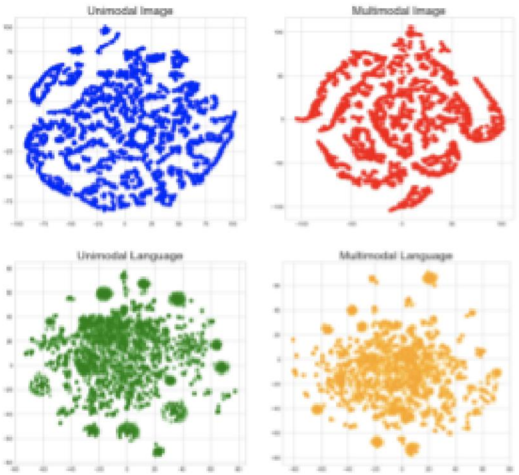


Figure 4. Language data adds more structure to image embeddings (top right compared to top left) while image data impacts language embeddings less (bottom row) visualized via tSNE plots.

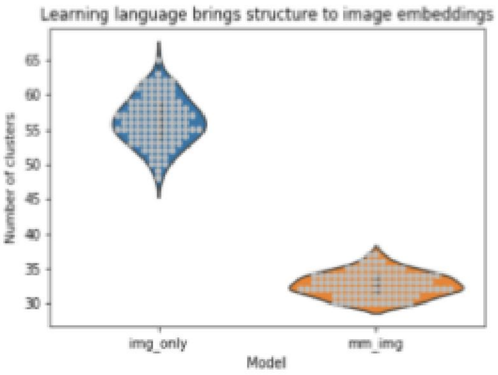


Figure 5. Jointly training with language leads to more categorical structure in image embeddings. Results of running n=100 simulated non-parametric (Bayesian Gaussian Mixture Model with Dirichlet Process prior) to infer number of clusters. *t(198) = 61.6, p < 1e-100* .

References

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

Bowman S., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S. (2016). Generating Sentences from a Continuous Space. *Proceedings of CoNLL*

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S., Conway B. (2017). Color naming across languages reflects color us. *PNAS*.

Kay & Kempton (1984). What is the Sapir-Whorf Hypothesis. *The American Anthropologist.*

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

Regier, T, & Xu, Y., (2017). The Sapir-Whorf hypothesis and inference under uncertainty. *Science*. e1440.

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