Learning Visually Grounded Representations with Sketches

Roma Patel Stephen Bach [1] Ellie Pavlick [1]

Abstract

We test whether visually grounded meaning representations for words can be improved by grounding to sketches rather than to natural images. Intuitively, sketches encode higher-level abstract representations of the concepts to which words refer. We test empirically whether such abstractions are beneficial for the purposes of grounded representation learning. We evaluate our representations in terms of correlations with human inferences about the semantic and visual similarity between concepts. Our results suggest that grounding to sketches yields better representations than does grounding to other visual representations.

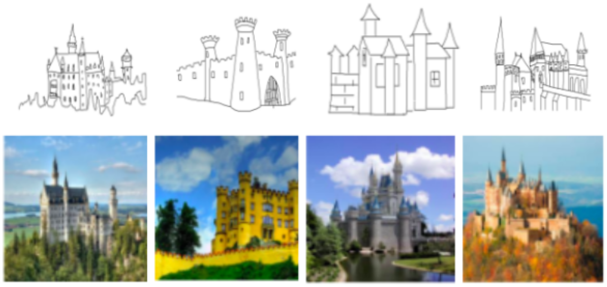


Figure 1. Figure shows sample photos paired with human-drawn sketches from the Sketchy dataset.

1. Introduction

Recent years have seen significant advancements in combining visual and textual information to learn grounded representations of concepts. The problem of learning to ground language is central to semantic understanding, and has been studied within areas such as semantic parsing (Zettlemoyer & Collins, 2012), video understanding (Feng & Lapata, 2010), and multimodal concept-learning (Johns & Jones, 2012). While there are several theories of what it means to know a concept (Laurence & Margolis, 1999), they are all unified in that they require some grounded notion of what the object is in the real world i.e., what it does, what it looks like, and what interactions it affords. Recent work in multimodal NLP, specifically targeted towards lexical semantics, has operationalised such notions with grounded distributional models (Bruni et al., 2014; Silberer & Lapata, 2014; Lazaridou et al., 2015) which have achieved measurable success when evaluated against human similarity judgements of concepts.

Most past work has represented visual information either in the form of manually specified attributes (represented as sparse binary feature vectors) or as densely pixelated, coloured photographs. In this paper, we propose to use human drawn sketches to learn grounded representations. Intuitively, sketches—in comparison to natural images—encode higher-level abstract representations of the essential components concepts to which words refer. From very young ages, children use drawings as graphical representations to encode their understanding of concepts in a visible format (Long et al., 2018). Furthermore, the extraneous background information and artifacts present in natural photos are often not required by humans for comprehension (Das et al., 2017) and can be exploited by statistical models (Agrawal et al., 2018; Tommasi et al., 2017) to improve performance without forming a satisfying representation of the concept in question.

This paper therefore addresses the following question: does grounding to sketches rather than other visual representations (e.g., natural photos or discrete visual attributes) yield better meaning representations of lexical concepts? We compare performance of our learned representations in terms of correlations with human judgments of visual and semantic similarity and in terms of categorisation of new visual input. We observe that tasks that condition on visual information in the form of sketches outperform those which use natural photos by more than 4 points on average when compared to human similarity judgements on a Spearman's correlation scale.

[*] Equal contribution [1] Department of Computer Science, Brown University. Correspondence to: Roma Patel <romapatel@brown.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

2. Experimental Design

The goal of our experiments is to test whether visually-grounded representations trained using sketch representations of concepts perform better than those trained using natural images or discrete visual attributes. To maintain a controlled comparison, we use the bimodal architecture introduced by silberer2014learning which receives two inputs: a visual component and a textual component. Our experiments vary the representation of the visual component in order to ascertain which of the three visual representations allow learning of the best meaning representations. Specifically, the visual component of the model receives natural photos as input and is trained to represent this photo such that it can predict one of three visual elements: a reconstruction of the natural photo itself, a sketch of the photo, or a vector of discrete attributes describing the photo.

2.1. Data

We use the Sketchy dataset (Sangkloy et al., 2016); currently the only dataset that contains natural photos paired with sketches drawn for each photo, allowing us to run closely controlled experiments. Other, larger sketch datasets (Ha & Eck, 2017; Eitz et al., 2012) do not contain one-to-one mappings between sketches and photos. To obtain attributes for concepts, we use the VisA dataset (Silberer & Lapata, 2012) which tags each concept with visual attributes from a taxonomy of 636 attributes (e.g., the concept *dolphin* has attributes such as *has_jaws* and *has_flippers*).

For our experiments, we take the intersection of classes contained in the VisA dataset and the Sketchy dataset, which gives us 69 categories and 299 total attributes. The Sketchy dataset gives us more than a 100 natural photos paired with 5 sketches per photo. To ensure all models are trained on the same amount of data, we sample one sketch per photo. We represent each image using the last 4096-dimensional fully connected layer of a CNN pretrained on ImageNet (we use VGG16 (Simonyan & Zisserman, 2014)). We hold out 5 images from each class to evaluate models on and train on all remaining images. As input to the textual component, we use fixed 300-dimensional GloVe representations for each class label. This means, that each class gets the same text input i.e., the pretrained GloVe vector for that class, but a unique visual input i.e., features for the image.

2.2. Model

We follow silberer2014learning and use a bimodal autoencoder to learn grounded representations. However, our setup differs from theirs in that our model takes as input a raw RGB image whereas their model takes as input a visual description of the image in the form of sparse attributes. In our setup, each training sample consists of a a 224 × 244 dimensional natural image and a label (e.g., one of the 69 classes) represented as a pretrained GloVe vector. These serve as input to the visual and textual components respectively.

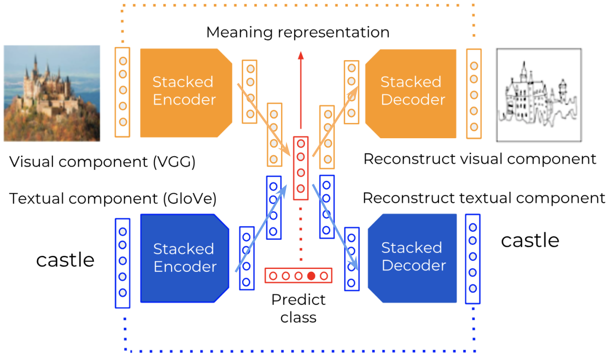


Figure 2. Our model architecture. Blue and orange correspond to textual and visual encoders respectively. Red depicts the *prediction loss* that predicts the class to which the latent representation belongs. This figure shows the **Sketch+Text** model, while the **Photo+Text** and **Attr+Text** models share the exact same architecture where the decoder (in orange) reconstructs either the photo or attribute vector.

The architecture of both the visual and textual components consists of 3 linear encoding layers followed by ReLU activations. We use a dropout of 0.2 at each encoding layer. We concatenate the visual and textual encoded representations and apply another linear transformation to obtain a latent representation of 69 dimensions. Each of these stacked layers form the visual and textual encoder respectively. Our decoders contain the exact same configuration as the encoders, in reverse order i.e., we apply a linear transform to the 69-dimensional latent representation, split it into visual and textual components and transform each with three decoding layers to reconstruct both visual and textual features. We use an L1 loss to compare the reconstructed output to the ground truth. We furthermore impose a supervised criterion on the latent representation as in (Silberer & Lapata, 2014) to classify the input, by applying a softmax layer over the 69 dimensional representation, to predict the class that the input image belongs to. The model therefore learns both from the feature representations of the visual and textual components that learns to reconstruct, as well as from the label that it tries to predict.

For our experiments, we perform comparisons of three settings (Figure 2). Each variant takes the same input (natural photo and word vector). We vary the visual component that the autoencoder aims to reconstruct, specifically: the natural photo (**Photo + GloVe**), the sketch (**Sketch + GloVe**), or the attributes (**Attribute + GloVe**). For comparison, we also evaluate unimodal variants of our model i.e., an autoencoder with only visual input in three settings (**Sketch, Photo, At-**

tributes). In all configurations, the latent representation serves as the meaning representation that we evaluate.

2.3. Evaluation

We evaluate the representations based on their correlation with human judgments of semantic and visual similarity between concepts. We use the dataset of human judgements collected by silberer2012grounded which are similarity ratings obtained for pairs of words on a scale of 1-5. For example, the concepts *chicken* and *owl* have a high semantic similarity of 4.25 but lower visual similarity of 3.00. We use the cosine distance as a measure of similarity between representations, and compute correlation with human judgments using Spearman's ρ. We compare these to correlations obtained from unimodal representations for text and images i.e., 300 dimensional word representations (GloVe), 299 dimensional k-hot attribute representations, 4096 dimensional image representations obtained from a pretrained CNN (VGG). It is important to note that the GloVe and attribute representations are not obtained over the same test set but are fixed representations for each concept. All other representations (e.g., *Sketch* + *Photo*) are obtained by using each trained model on test images and averaging representations from all images in the class.

3. Results

Table 2 shows correlation coefficients of model predictions against human visual and semantic similarity ratings. For comparison, we list the results reported in silberer2014learning, although we note that their models were trained and tested on different input and are not directly comparable. Consistent with prior work, we see that the bimodal autoencoders consistently outperform the unimodal. More interestingly, our results show that in both cases (unimodal and bimodal), models trained using sketches outperform those that use natural photos or attributes in terms of semantic similarity. In the bimodal case, sketches outperform photos on visual similarity as well, and perform equally to attributes. While the text-only representations (GloVe and Attributes) are more highly correlated with human judgements, they are not directly comparable — their use is limited in learning grounded representations that taken in different images as input every time.

While there is only a slight gain in performance from the model that conditions on sketches, qualitative analysis of the predictions made by models on test images gives us useful insight into the differences in the three visual elements. These predictions are obtained by applying a softmax activation over the latent representation and then ranking the classes. Figure 3 shows examples of the top three predicted categories for the each of the three bimodal autoencoders. We see that in unusual cases (e.g., an airplane that is not outdoors, but in a museum) the *photo* model conditions heavily on artifacts and misclassifies the image (as e.g., saxophone, trumpet) possibly by virtue of the colour and background of the image, while the *sketch* and *attribute* models predict more sensible classes (e.g., helicopter, airplane). We note that while sketches are harder to obtain than natural images, related work that explores generative models to draw sketches (Ha & Eck, 2017) can be used to augment data to allow building of better visually grounded representations. A direction to explore in future work, is whether automatically-generated sketches can substitute for human-drawn ones.

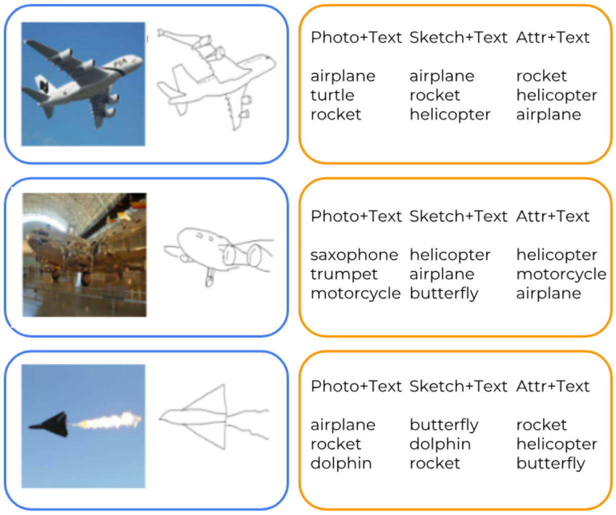
Input	**Output**	**Sem**	**Vis**
Photo + Text	Sketch + Text	0.71	0.69
Photo + Text	Photo+ Text	0.67	0.64
Photo + Text	Attr. + Text	0.68	0.69
SAE	Attr. + Text	*0.70	*0.64
Photo	Sketch	0.44	0.40
Photo	Photo	0.39	0.42
Photo	Attributes	0.43	0.40
VGG		0.51	0.54
Word (GloVe)		0.79	0.64
Attributes		0.78	0.69

Table 1. Correlation with human judgments of semantic and visual similarities. Top section shows bimodal models. SAE results are copied from those reported in silberer2012grounded and are not directly comparable. Second section shows unimodal (vision-only) models; VGG and Glove represent SOTA unimodal representations. Attributes (bottom row) can be viewed as a weak upper bound, as they reflect similarities derived from explicit, human-coded attributes of concepts that our learned representations will implicitly capture.

4. Related Work

Previous work on grounded representations has used natural images (Lazaridou et al., 2015) or discrete attributes (Silberer & Lapata, 2012) to learn representations of concepts using both visual and textual information. Specifically, autoencoders have been used to derive representations for concepts (Silberer & Lapata, 2014) by combining visual and textual modalities. We build directly on the work by silberer2014learning, which uses stacked bimodal autoencoders.

Previous work dealing with sketches has largely been focused on tasks like sketch-based image retrieval (Eitz et al., 2012; Sangkloy et al., 2016), 3-D shape-retrieval (Wang et al., 2015) and cognitively motivated analyses of human-drawn sketches (Long et al., 2018; Eitz et al., 2012). Other work has focused on training recurrent neural networks to draw stroke based representations of categories (Ha & Eck,



Photo+Text	Sketch+Text	Attr+Text
airplane	airplane	rocket
turtle	rocket	helicopter
rocket	helicopter	airplane

Photo+Text	Sketch+Text	Attr+Text
saxophone	helicopter	helicopter
trumpet	airplane	motorcycle
motorcycle	butterfly	airplane

Photo+Text	Sketch+Text	Attr+Text
airplane	butterfly	rocket
rocket	dolphin	helicopter
dolphin	rocket	butterfly

Figure 3. Example ranked predictions made by models. Top: all models perform well. Middle: photo model is biased by background information. Bottom: sketch model to be biased by the sketch outline.

2017). Our work draws on the core idea of using sketches and combines this with methods that attempt to learn semantic meaning representations. Specifically, we attempt to map between visual information in the form of sketches and textual information in the form of dense word vectors to learn grounded meaning representations for words.

Psychological work on inferring representations from behavior has concluded that human similarity judgments capture stimulus generalization behavior (Shepard, 1987) and have been shown to encode the complex spatial, hierarchical (Peterson et al., 2018), and overlapping (Shepard & Arabie, 1979) structure of human representations, around which numerous models of categorization and inference are built (Goldstone, 1994; Nosofsky et al., 1992; Nosofsky, 1987). If we can capture similarity judgments, we will have obtained a considerably high-resolution picture of human psychological representations. Our approach draws from such techniques and uses human similarity judgements to evaluate representations to test the information they encode.

5. Conclusion

We presented a method that learns visually grounded representations of concepts by making use of human-drawn sketch representations. Our evaluation shows that an approach that requires a model to reconstruct visual information (as sketches) and textual information (as dense word embeddings) allows learning of representations that are more correlated with human judgements. This outperforms a model that replaces sketches with natural photos, highlighting the information gained from sketch representations.

Looking forward, we believe that sketch representations are worth studying both qualitatively (e.g., analysing renderings of different concepts from different humans) and quantitatively (e.g., analysing the effect of training models to condition on such representations). Our analysis and evaluation on a small dataset of photos paired with sketches suggests that this is a promising direction for future exploration.

References

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Bruni, E., Tran, N.-K., and Baroni, M. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

Eitz, M., Hays, J., and Alexa, M. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44–1, 2012.

Feng, Y. and Lapata, M. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 91–99. Association for Computational Linguistics, 2010.

Goldstone, R. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4):381–386, 1994.

Ha, D. and Eck, D. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

Johns, B. T. and Jones, M. N. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4 (1):103–120, 2012.

Laurence, S. and Margolis, E. Concepts and cognitive science. *Concepts: core readings*, pp. 3–81, 1999.

Lazaridou, A., Pham, N. T., and Baroni, M. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

Long, B., Fan, J. E., and Frank, M. C. Drawings as a window into developmental changes in object representations. 2018.

Nosofsky, R. M. Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1):87, 1987.

Nosofsky, R. M., Kruschke, J. K., and McKinley, S. C. Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2): 211, 1992.

Peterson, J. C., Soulos, P., Nematzadeh, A., and Griffiths, T. L. Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *CoRR*, abs/1805.07647, 2018. URL http://arxiv.org/abs/1805.07647.

Sangkloy, P., Burnell, N., Ham, C., and Hays, J. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.

Shepard, R. N. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

Shepard, R. N. and Arabie, P. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87, 1979.

Silberer, C. and Lapata, M. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1423–1433. Association for Computational Linguistics, 2012.

Silberer, C. and Lapata, M. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 721–732, 2014.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

Tommasi, T., Patricia, N., Caputo, B., and Tuytelaars, T. A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, pp. 37–55. Springer, 2017.

Wang, F., Kang, L., and Li, Y. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1875–1883, 2015.

Zettlemoyer, L. S. and Collins, M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012.