Sketch Me if You Can: Towards Generating Detailed Descriptions of Object Shape by Grounding in Images and Drawings

Ting Han
Artificial Intelligence Research Center
Tokyo, Japan
ting.han@aist.go.jp

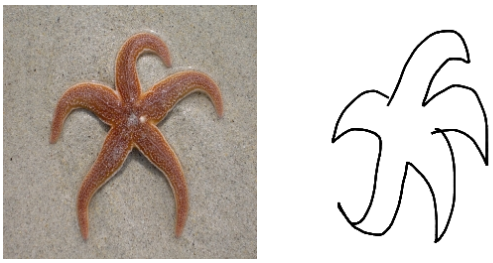
Sina Zarrieß
Friedrich-Schiller-Universität Jena
Jena, Germany
sina.zarriess@uni-jena.de

Abstract

A lot of recent work in Language & Vision has looked at generating descriptions or referring expressions for objects in scenes of real-world images, though focusing mostly on relatively simple language like object names, color and location attributes (e.g., brown chair on the left). This paper presents work on Draw-and-Tell, a dataset of detailed descriptions for common objects in images where annotators have produced fine-grained attribute-centric expressions distinguishing a target object from a range of similar objects. Additionally, the dataset comes with hand-drawn sketches for each object. As Draw-and-Tell is medium-sized and contains a rich vocabulary, it constitutes an interesting challenge for CNN-LSTM architectures used in state-of-the-art image captioning models. We explore whether the additional modality given through sketches can help such a model to learn to accurately ground detailed language referring expressions to object shapes. Our results are encouraging.

1 Introduction

Recent work in referring expression generation (REG) has focused more and more on large-scale image datasets (Kazemzadeh et al., 2014; Mao et al., 2016; Yu et al., 2016) and models that incorporate a state-of-the-art vision component (Mao et al., 2016; Yu et al., 2017; Zarrieß and Schlangen, 2018). As compared to traditional REG settings (Dale and Reiter, 1995; Krahmer and Van Deemter, 2012), these works have led to substantial advances in terms of the complexity of visual inputs that can be processed and the visual object categories that can be covered. At the same time, it is questionable whether these recent benchmarks for real-world REG constitute an equally big step forward in terms of the language that needs to be modeled. As noted by Achlioptas et al. (2019), the vocabulary and attributes learnt by state-of-the-art REG models is linguistically and lexically relatively constrained, and does not cover language that can be used to describe fine-grained differences between object parts and shapes (see Section 2) . Thus, Achlioptas et al. (2019) propose to go back to carefully designed datasets with graphical, abstract objects in order to elicit complex descriptions of object attributes and also to have access to more fine-grained representations of an object's geometry and topology.



(a) (b)

Figure 1: (a) Photo of a starfish; (b) Sketch of the starfish in (a). Starfish attribute description: *Top view, legs bend, thin legs, on sand.*

We explore modeling of fine-grained attribute descriptions of objects in real-world images, based on the *Draw-and-Tell* dataset introduced by Han and Schlangen (2017). This dataset was collected in a controlled procedure akin to traditional REG setups, resulting in fine-grained attribute descriptions and a rich vocabulary, see Figure 1 for an example. As the *Draw-and-Tell* data was originally designed for sketch-based image retrieval (Eitz et al., 2012; Sangkloy et al., 2016), each image is paired with hand-drawn sketches depicting the object in it. As illustrated in Figure 1(b), these sketches are somewhat distorted and abstract away from many visual properties of the complex real-world objects (e.g. colour). Yet they provide a clear outline of the object's shape. In this paper, we explore whether this

form of visual abstraction is useful for modeling and generating fine-grained attribute descriptions of objects. We investigate whether object sketches lead to improvements in neural generation of descriptions of real-world objects, especially for attributes related to shape and orientation.

In the following, we present our ongoing work on generating detailed attribute descriptions of object by grounding in images and hand-drawn sketches. We first introduce the Draw-and-Tell dataset (Section 3), then describe a basic recurrent neural network architecture for generating attribution descriptions (Section 4). We carry out an automatic evaluation based on measures like BLEU (Papineni et al., 2002), vocabulary size, and the average length of generated descriptions. In addition, we provide a qualitative analysis and discussion on how incorporating sketches can benefit the task of generating fine-grained attribute descriptions.

2 Related Work

Visual language grounding and REG Foundational work in REG has often followed the well-known attribute selection paradigm established by (Dale and Reiter, 1995). Here, visual scenes have usually been carefully created and controlled so that the target and distractor referents and distractors would have similarities in their set of annotated attributes (e.g. type, position, size, color and so on), see Krahmer and Van Deemter (2012). In recently used image benchmarks for REG, the visual scene is typically given through a real-world image (Kazemzadeh et al., 2014; Yu et al., 2016), which makes it very difficult to systematically control the underlying attributes of a target referent and to what extent it resembles its distractors in the scene. At the same time, Yu et al. (2016) found that, in the standard version of the RefCOCO benchmark, many participants simply used location attributes like *left, right* relying on the 2D layout of the scene. As a remedy, they propose to introduce "taboo words" into the reference task in order to elicit "appearance-based" attributes. Achlioptas et al. (2019) adopt a different approach and suggest to collect data based on more abstract objects. They collect a dataset of referring expressions to chairs where various properties and parts of targets and distractors are controlled in terms of their visual similarity. Our work combines ideas from both paradigms: we use real-world images of objects paired with hand-drawn sketches, which allows us to integrate realistic and abstract visual inputs.

Multimodal Embedding Space For being able to model REG with multiple input modalities (images and sketches), we need to be able to represent these inputs as visual embeddings or features transferred from a CNN. Here, we rely on previous work that has mapped different modalities into joint vector spaces, as in text- or sketch-based image retrieval (Kiros et al., 2016; Sangkloy et al., 2016; Liu et al., 2017). We adopt (Sangkloy et al., 2016)'s Siamese network to project sketches and images into a joint space, and use the projections as inputs to a basic recurrent neural network for REG. It is noteworthy that this joint image-sketch space is designed to capture similarities across modalities, rather than complementary information expressed in different modalities. We leave the exploration of other modes of representation for future work.

3 The Draw-and-Tell Dataset

The **Draw-and-Tell** dataset (Han and Schlangen, 2017) includes 10,804 photographs of objects (referred to as target objects below), spread across 125 categories. Each image is paired with around 5 hand-drawn sketches and a description of the object's attributes, as shown in Figure 1.

The photos and sketches were selected from the Sketchy Database[1] (Sangkloy et al., 2016). Han and Schlangen (2017) augmented part of the Sketchy Database with object attribute descriptions which were collected from English speakers using a Crowdsourcing service. In each description task, workers were presented with 6 photos of objects from the same category. They were instructed to describe attributes of the target object, so that another person can distinguish the target object from distractor objects. Hence, this resembles classical settings in REG where distractors are controlled for being similar to the target reference. Attributes such as *shape*, *color* and *orientation* were suggested as examples to the workers, but they were also encouraged to list all attributes that they consider useful. Attribute phrases in the descriptions were typically separated by ",". As all the distractor images were in the same category and in separate images, workers were suggested not to use non-discriminative words such as category names or spatial relations in the descriptions.

[1] http://sketchy.eye.gatech.edu/

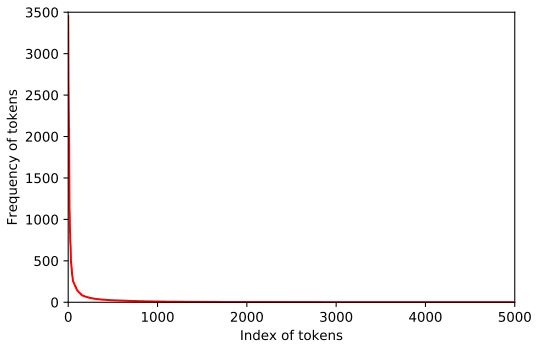


Figure 2: Word frequency in the corpus.

Sets	Token No.	Vocab size	Overlap
Train	85974	4621	-
Validation	4698	1099	962
Test	9948	1601	1370

Table 1: Data statistics. **Overlap** column shows vocabulary overlaps between the training set and validation/test sets.

Data Statistics On average, each object description includes 3 attribute phrases. In total, there are 100620 tokens in all the descriptions. The vocabulary size is 4982. 2893 of all the words in the vocabulary appear less than 3 times, as shown in Figure 2. We split the dataset into train (9233), validation (500) and test (1071) sets. Table 1 shows the token numbers, vocabulary size of each data split, and vocabulary overlaps with the training set.

Image and Sketch Joint Embedding Space Along with the Sketchy Database, Sangkloy et al. (2016) published a Siamese network model that embeds images and sketches into a joint vector space. The Siamese network is composed of two separate networks with the same architecture, the sketch-net and the image-net. The image-net encodes photographs into image feature vectors. Similarly, the sketch-net encodes sketch images into feature vectors. The Siamese network was trained and optimized to project photo vectors as close as possible to corresponding sketch vectors, while in the mean time, as distinguishable as possible from other photo vectors. In this work, we used the pretrained models and took the output vector from the last fully connected layers as feature vectors (in 1024 dimension) to represent images and sketches. Next, we describe how we train a natural language generation model with the extracted vectors.

4 The RNN Caption Generator

Considering the small size of the *Draw-and-Tell* dataset, we built a basic Recurrent Neural Network model for the generation task (Tanti et al., 2017). The model takes a text vector and a visual feature vector as inputs, and predicts a sequence of tokens to describe the target object in the input visual vector. Therefore, the generated tokens are conditioned on the input visual feature vector.

The network includes an embedding layer, an LSTM layer, and a softmax layer. We encode word tokens as a one-hot vector, and concatenate the vector with image feature vectors (i.e., injecting image information into the network). An embedding layer takes the concatenated vector as input. The size of the LSTM layer is 512. The model was implemented using Tensorflow[2]. Training of the model is performed using the Adam optimizer and the cross entropy loss function.

When applying the model to generate object descriptions, the model first takes an image vector as input, then predicts each token conditioned on the image vector and previously predicted tokens until an end token EOS is predicted. We used a beam search method to predict the tokens. The bandwidth of the beam search algorithm is 3.

Note that in the current set-up, we do not include context or distractor images to generate discriminative descriptions, but focus on exploring the modality aspect in the generation task. See Zarrieß and Schlangen (2018) for a detailed discussion of the benefit of context features in image-based REG.

5 Experiments

We conducted ablation experiments by altering the input visual feature vectors. In the **Image Only** setup, we trained a model with *(object description, image feature)* pairs as input. This results in training set of 9233 description-image pairs. In the **Sketch Only** setup, we trained a model with *(object description, sketch feature)* pairs, using training data of the same size as in the Image Only setup. In the **Multimodal** setup, we use both *(object description, image feature)* and *(object description, sketch feature)* pairs to train the RNN. That is, we doubled the size of the training data. Therefore, the model does not only learn to generate expressions conditioned on image features, but also conditioned on sketch features.

[2]https://www.tensorflow.org/

top view legs bend thin legs on sand

Brown and white in color laying on the sand
Red and yellow in color has green leaves
Central position brown color green grass

(a)

this chicken is back and have a colorful head in a forest background

Red and white chicken facing left red comb
Black and white in color facing left
Black and white in color facing to the left

(b)

white stripes facing right black dot on the back

Black and white in color facing to the right
Black and white in color facing left
White and black in color facing left

(c)

weathered wood many plans with no spaces facing forward

A wooden bench in a park with a park
Light brown wooden bench in front of desk
Light brown wood facing left side view

(d)

trunk decorated ears pierced curled trunk

brown elephant facing right head down
facing right full body shot eating grass
gray elephant facing to the right

(e)

brown leaning down face to the right in long gray grass

facing right full body shot looking forward
black and white in color facing right
facing right head down looking forward

(f)

Figure 3: Samples of generated descriptions. Grey utterances are attribute annotations from humans; Red utterances are generated using both sketch and image features. Blue utterances are from the sketch feature only model; Green utterances are from the image feature only model.

During training, the data was randomly shuffled, with a mini batch size of 50. The maximum epochs is 100. Words appearing less than 3 times were removed.

All the models were evaluated with image features in the test set as input to resemble the task of object description from images. Figure 2 shows the evaluation results.

Metrics We evaluated the generated attribute descriptions with **BLEU1** score. We also report vocabulary size, average number of tokens in each generated object description to show the word capacity of the models.

Table 2 shows the evaluation results. The *Multimodal* model achieved a slightly higher BLEU1 score. It also results in a slightly larger vocabulary of the generated descriptions.

As sketch vectors are expected to encode iconic information, we analyze *color*, *shape*, and *orientation words* in the generated descriptions in the Qualitative analysis section below.

5.1 Qualitative Analysis

Figure 3 shows some examples for generated referring expressions. We observed that, given only sketch vectors in the training data, the model still generates color words, but less accurate than Image Only and Multimodal models. For instance, in Figure 3 (a), the Sketch Only model describes a *brown* starfish as *red and yellow*. As shown in Figure 3 (b), the Multimodal model used *red comb* to describe the chicken, while the other two models only used color and orientation words. We conjecture that this could be due to the combination of sketch and image features in the training set. In Figure 3 (e) and (f), the descriptions of the Sketch Only model correctly describe the directions, while missing other attributes such as color or adding wrong attributes (e.g., grass).

Experiments	BLEU1	BLEU2	Vocab. size	Token Number
Sketch only	0.58	0.43	310	8.24
Image only	0.61	0.43	325	8.42
Multimodal	**0.64**	0.44	331	8.90

Table 2: BLEU scores, vocabulary size and average length of object descriptions in the experiments.

6 Conclusion & Discussion

We have presented our ongoing work on generating fine-grained attribute descriptions of objects in

real-life images by grounding in images and hand-drawn sketches. Given a medium-sized dataset with many low frequency words, we deployed sketch vectors from a joint sketch-image embedding space to improve the generation results. We show that by training a basic recurrent neural network with both sketch and image features, the model is able to capture more fine-grained attribute descriptions. Moreover, even when training only with sketch feature vectors, the model still achieves a satisfactory performance according to automatic evaluation with a BLEU.

In future work, we plan to add human evaluation results to show how humans perceive the generated descriptions in terms of naturalness and accuracy. We plan to further explore the multimodal joint embedding space for fine-grained object description generation tasks such as shape and orientation description generation.

Acknowledgments

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Panos Achlioptas, Judy Fan, X.D. Robert Hawkins, D. Noah Goodman, and J. Leonidas Guibas. 2019. ShapeGlot: Learning language for shape differentiation. *CoRR*, abs/1905.02925.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10.

Ting Han and David Schlangen. 2017. Draw and tell. multimodal descriptions outperform verbal-or sketch-only descriptions in an image retrieval task. In *The 8th International Joint Conference on Natural Language Processing. Proceedings of the Conference. Vol. 2: Short Papers*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2016. Unifying visual-semantic embeddings with multimodal neural language models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2871.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*.

Marc Tanti, Albert Gatt, and Kenneth Camilleri. 2017. What is the role of recurrent neural networks (rnns) in an image caption generator? In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 51–60.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing, Cham.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.