Draw and Tell: Multimodal Descriptions Outperform Verbal- or Sketch-Only Descriptions in an Image Retrieval Task

Ting Han and **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University
firstname.lastname@uni-bielefeld.de

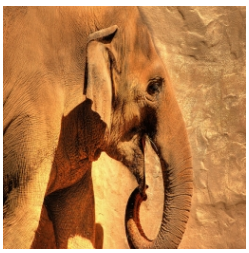
Abstract

While language conveys meaning largely symbolically, actual communication acts typically contain iconic elements as well: People gesture while they speak, or may even draw sketches while explaining something. Image retrieval *prima facie* seems like a task that could profit from combined symbolic and iconic reference, but it is typically set up to work either from language only, or via (iconic) sketches with no verbal contribution. Using a model of grounded language semantics and a model of sketch-to-image mapping, we show that adding even very reduced iconic information to a verbal image description improves recall. Verbal descriptions paired with fully detailed sketches still perform better than these sketches alone. We see these results as supporting the assumption that natural user interfaces should respond to multimodal input, where possible, rather than just language alone.

1 Introduction

In natural interactions, descriptions are typically multimodal: Someone explaining a route might point at visible landmarks while talking, or gesture them into the air, or may sketch a route on a piece of paper, if they have one handy (Emmorey et al., 2000; Tversky et al., 2009).

Especially descriptions of visual objects or situations can be supported by the iconic mode of reference provided by gestures or sketches, that is, reference via similarity rather than via symbolic convention (Pierce, 1867; Kendon, 1980; McNeill, 1992; Beattie and Shovelton, 1999). A technical task that is a direct, but controlled model of this is the task of *image retrieval*, that is, the task of retrieving one out of many photographs, based on a description of it.[1] In current work, these descriptions are typically 'monomodal', either purely verbal descriptions (Schuster et al., 2015; Hu et al., 2016),[2] or via hand-drawn sketches (Sangkloy et al., 2016; Qian et al., 2016; Yu et al., 2016).



Elephant, trunk coiled towards mouth, facing right

Figure 1: A photograph; a verbal description of its content; and a sketch.

In this work, we were interested in combining these modalities for image retrieval. We collected verbal descriptions of images (as shown in Figure 1), where the images were taken from an existing collection that provides for each image a matching sketch (Sangkloy et al., 2016). We trained "words-as-classifiers" models (Kennington and Schlangen, 2015; Schlangen et al., 2016) on the verbal descriptions to match these with images, and used the "triplet network" introduced by Sangkloy et al. (2016) to extract embeddings for the sketches. These models provide comparison scores for descriptions and candidate images, and can be combined into a joint score for a multimodal description (Section 3). We experiment with reduced sketches containing only a certain amount of the strokes from the full sketch, and

[1] Note that we use *strict retrieval* here, where a single, known image is to be retrieved, rather than an arbitrary one that fits a general description.

[2] As implemented and in commercial use on popular internet search engines.

Figure 2: Example of the crowd-worker task. *Provide a description that identifies the left-most image within this set:* [the elephant] facing right, trunk coiled toward mouth

show that adding even very reduced iconic information to the verbal image description improves recall (Section 4). Verbal descriptions paired with fully detailed sketches still perform better than these sketches alone. Using reduced sketches allows us to quantify redundancy between modalities, and it also makes it possible to explore how information becomes available incrementally, as sketch and utterance progress. We see our results as supporting the claim that natural user interfaces should respond to multimodal input—sketches, or, going beyond that, iconic gestures—, where possible, rather than just language alone.

2 Data

Collecting verbal descriptions of images The starting point is a collection of 10,805 real-life photographs taken from ImageNet (Russakovsky et al., 2015), as selected by the Sketchy corpus of sketches (see next section; we sampled from all categories). We collected a verbal description for each photograph from English speakers using the Crowdflower service.[3] Workers were asked to list attributes of the target object so that another person would be able distinguish the described photograph from 6 distractors from the same image category. (See example in Figure 2.) Using attributes such as orientation, colour or shape was suggested, however, workers were encouraged to list any attribute values that might help, separating them with commas. As the image category was already known, workers were only asked to provide attributes.

To evaluate the quality of the descriptions, we randomly selected 100 descriptions and conducted an image selection task. A different set of workers was presented with 7 images in the same category, one target and 6 distractors. Workers correctly selected 71% of the photographs, which shows that some of the collected verbal descriptions did not refer unambiguously.

[3] http://www.crowdflower.com

In total, we collected 10,805 object descriptions (100,620 tokens altogether). After spell checking, the vocabulary size is 4,982 (type/token ratio of 0.5). On average, each object was annotated with 3 attributes, while each attribute on average spans over 4.6 words. There were 29,234 different types of (potentially multi-word) attributes. To reduce this variability and ease the learning (described below), we devised a rule-based normalisation that mapped constructions such as "facing to the left", "facing left", "looking to the left" to the same attribute type (*facing-left*), leaving us with 18,673 different attribute types.

The Sketchy Corpus We profited from the availability of a dataset that pairs individual images (from ImageNet, using 100 images each from 125 different categories) with sketch representations of their content, the *sketchy database* (Sangkloy et al., 2016). These sketches were drawn from memory, but were validated to represent specifically the given image and not just its semantic category. Figure 1 above gave an example of such a sketch.

The sketches are stored as SVG files containing the start and end times of strokes, which allowed us to construct reduced versions containing only the first $n\%$ of strokes. Figure 3 shows some examples of such reductions. This gives us a rough approximation to an importance ordering of details in the sketch, under the assumption that the most salient features of the image might be drawn first. (We will further explore this assumption in future work.)

In the experiments reported below, we follow the training/test split used by Sangkloy et al. (2016). As we used the pre-trained sketch-image retrieval model from the Sketchy Database, we follow the train-test split setup of the corpus. In total, there are 9,734 unique photographs in our training set, and 1,071 photographs and 5,371 sketches in the test set (that is, for most images there are 5 different sketches). That is, there are

5371 sketch/photograph ensembles in our image retrieval evaluations. Chance level recall of the image retrieval task @K=1 is 0.093% (@K=10: 0.93%).

3 Models

The retrieval combines separate word/image and sketch/image models, which will be described here.

Grounding verbal descriptions to images To judge how well a verbal description fits with a photograph, we trained logistic regression classifiers for all category words and attribute types (following the "words-as-classifiers" (WAC) approach, (Kennington and Schlangen, 2015; Schlangen et al., 2016)).[4] The classifiers take a feature representation of an image (extracted by the convolutional neural-network described below) and produce for each word an "appropriateness score". To train for example the classifier for the word "elephant", we selected all photographs which were annotated with the category word "elephant" as positive training examples, then randomly selected the same amount of photographs that are not annotated as elephant as negative examples. (Similarly for the attribute types.) We trained classifiers for words or (normalised) attribute types which occurred more than 10 times in the corpus.

Given an image description D : $w_{a_1}, \cdots, w_{a_n}, w_c$, where w_{a_i} indicates an attribute word, and w_c indicates a category word, we compute a score for a given photograph **P** and using the word/image classifiers $s_w(\cdot)$ as follows:

$$s_D(D, \mathbf{P}) = s_{w_c}(\mathbf{P}) \times \sum_{i=1}^{n} s_{w_{a_i}}(\mathbf{P}) \qquad (1)$$

(That is, attribute contributions are combined additively and then multiplicatively with the category. Attributes for which no classifier could be trained were left out of the composition.)

Comparing Sketches with Images For the comparison of the sketches with the images, and the extraction of image features, we used the "tripled network" model devised and trained by (Sangkloy et al., 2016). This model is composed of two GoogLeNet networks (Szegedy et al., 2015), one for sketches and one for images. It is trained with a ranking loss function, with input tuples of the form (S, I+, I-) corresponding to a sketch, a matching image and a non-matching image. As a result, the network has a set of parameters for the sketch-network and a set of parameters for the photo-network. It learns a joint 1024 dimensional embedding space of sketches and photographs. The vector distance between a sketch and an image indicates their visual similarity (please refer to the original paper for more details for model structures). We used the reciprocal of the distance as the score to measure the fitness between a sketch and a photograph:

$$s_{sk}(\mathbf{S}, \mathbf{P}) = d(\mathbf{S}, \mathbf{P})^{-1} \qquad (2)$$

where **P** indicates the feature vector of the photograph, derived with the image network, while **S** indicates the feature vector of the sketch, derived with the sketch network.

Multimodal Fusion We adopt a late fusion approach, and combine the scores as follows:

$$s_{sk+cat+att} = s_{sk}(\mathbf{P}, \mathbf{S}) \times s_d(\mathbf{P}, \mathbf{d}) \qquad (3)$$

4 Results

We evaluate the performance of verbal descriptions alone, and verbal descriptions with various levels of sketch detail added, with the results shown in Table 1, and procedures explained in the following.

Metric Following the convention of image retrieval tasks evaluation, we measure the photograph retrieval performances by average recall @K. For a given photo query, recall @ K is 1 if the corresponding photograph is among the top K retrieved results and 0 otherwise. We average over all test queries to produce average recalls. We report the average recall @K=1 and @K=10.

Mono-modal descriptions First of all, we evaluated the image retrieval performance only with verbal descriptions. Using just attributes (*att*), we achieve an average recall (@1) of 0.03, which is not surprising, given that attributes such as "facing left" can potentially describe many images. Giving the category alone (*cat*) gives an average recall of 0.12 (@1) and 0.9 (@10), respectively. This shows that the category classifiers perform well in detecting the right category (there are 8.57 images on average from each category in the test

[4] Using $\ell 2$ regularisation, liblinear optimizer, regularisation strength 1.0.

Sketch Detail	10%		30%		50%		70%		90%		100%	
Recall	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10
sk	0.01	0.06	0.07	0.27	0.17	0.55	0.25	0.70	0.31	0.79	0.35	0.84
att	0.03	0.23	0.03	0.23	0.03	0.23	0.03	0.23	0.03	0.23	0.03	0.23
cat	0.12	**0.90**	0.12	**0.90**	0.12	0.90	0.12	0.90	0.12	0.90	0.12	0.90
cat+att	0.14	0.83	0.14	0.83	0.14	0.83	0.14	0.83	0.14	0.83	0.14	0.83
sk+att	0.03	0.16	0.09	0.39	0.20	0.64	0.28	0.76	0.33	0.83	0.37	0.87
sk+cat	0.12	0.76	0.20	0.85	0.28	**0.92**	0.34	**0.94**	**0.38**	**0.96**	**0.41**	**0.96**
sk+cat+att	**0.15**	0.81	**0.21**	0.87	**0.30**	**0.92**	**0.35**	**0.94**	**0.38**	0.95	**0.41**	**0.96**

Table 1: Average recall at K=1 and10, at different levels of sketch detail. Highest number in column in bold. Numbers for language-only conditions do not change with level of sketch detail.

set). Combining *cat* and *att* improves performance somewhat @1, but even has a negative impact @10, indicating that the attributes can "override" the category and push images that are appropriate for the attributes, but not the category, into the top 10.

We also show results for the sketches alone, at various levels of detail of the sketch. (E.g., "10%" only contains the first 10% of strokes, etc. The 100% condition is the one reported by (Sangkloy et al., 2016), our results are within 0.01 of the ones reported there.)

cat+att	30% *sk+ cat+att*	30% *sk*	100% *sk*
chicken, can see head only, head is mainly red skin			
Rank=1	Rank=1	Rank=27	Rank=1
camel, light brown, laying down, head on right, has blanket to ride on			
Rank=3	Rank=1	Rank=29	Rank=1
butterfly, facing left, white			
Rank=3	Rank=1	Rank=32	Rank=1

Figure 3: Retrieval with verbal description only (1st column), verbal description plus 30% sketch (2nd column), 30% sketch (3rd column) and 100% sketch (4th column).

Multimodal descriptions As Table 1 (first column) shows, combining even the very reduced sketch information at a 10% detail level improves results @1 compared to language-only (if only marginally). The improvement increases with the level of sketch detail, and reaches at 70% sketch detail a level at which the multimodal ensemble performs as well as the full sketch (0.35 @1), improving 0.16 points over the language-only baseline. The fullest combination (full utterance, 100% sketch) improves over the full sketch by 0.06 points (0.41 vs. 0.35).

Figure 3 shows some selected examples with sketches at various detail settings.

5 Conclusions

This paper introduced a corpus of natural language attribute descriptions of images taken from a corpus that paired these images with sketches. We showed that a model of grounded word meaning trained on these data can be combined with an existing model of sketch/image relation, where the combination improves retrieval performance relative to the separate models. Specifically, the model profited even from small amounts of iconic information (sketches reduced to 30% of their strokes). We draw from these results the tentative conclusion that it can be advantageous to add modalities other than language (and hence allow reference other than through symbols, namely through iconic similarity relations) for certain tasks.

In future work, we plan to directly train a joint model that directly processes language and iconic input. Our ultimate goal is to allow *gestural* iconic input, which can be expected to also provide only a reduced level of detail, in a setting where real-world locations (rather than images of objects) are to be described. How comparable this is to the

reduced sketches used here is an exciting question to explore next.

We have made the image descriptions of the corpus publicly available in Bielefeld University PUB system (Han and Schlangen, 2017). The code of the image retrieval models is available on GitHub https://github.com/TINGH/multimodal-object-description

Acknowledgments

This work was supported by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). The first author would like to acknowledge the support from the China Scholarship Council (CSC).

References

Geoffrey Beattie and Heather Shovelton. 1999. Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of language and social psychology* 18(4):438–462.

Karen Emmorey, Barbara Tversky, and Holly a. Taylor. 2000. Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation* 2(3):157–180.

Ting Han and David Schlangen. 2017. Draw and Tell: a Corpus of Multimodal Object Descriptions. https://doi.org/10.4119/unibi/2913193.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 4555–4564.

Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25(1980):207–227.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*.

D McNeill. 1992. Hand and Mind: What Gestures Reveal About Thought. *What gestures reveal about* pages 1–15.

Charles Sanders Pierce. 1867. On a new list of categories. In Charles Hartshorne and Paul Weiss, editors, *C.S. Pierce: The Collected Papers*, Harvard University Press, Cambridge, M.A., USA.

Xueming Qian, Xianglong Tan, Yuting Zhang, Richang Hong, and Meng Wang. 2016. Enhancing sketch-based image retrieval by re-ranking and relevance feedback. *IEEE Transactions on Image Processing* 25(1):195–208.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35(4):119.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*. Berlin, Germany.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*. volume 2.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 1–9.

Barbara Tversky, Julie Heiser, Paul Lee, and MariePaule Daniel. 2009. Explanations in Gesture, Diagram, and Word. In Kenny R. Coventry, Thora Tenbrink, and John Bateman, editors, *Spatial Language and Dialogue*, Oxford University Press, pages 119–131.

Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. 2016. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 799–807.