Image Specificity

Mainak Jas[1], Devi Parikh[2]
[1]Aalto University. [2]Virginia Tech.

Consider the two photographs in Figure 1. How would you describe them? For the first, phrases like "people lined up in terminal", "people lined up at train station", "people waiting for train outside a station", *etc*. come to mind. It is clear what to focus on and describe. In fact, different people talk about similar aspects of the image – the train, people, station or terminal, lining or queuing up. But for the photograph on the right, it is less clear how it should be described. Some people talk about the the sunbeam shining through the skylight, while others talk about the alleyway, or the people selling products and walking. In other words, the photograph on the left is *specific* whereas the photograph on the right is *ambiguous*.

In this paper, we introduce the notion of image specificity. Given multiple human-provided descriptions of an image, we define the specificity of the image to be the average similarity score between pairs of sentences describing that image. We measure this similarity via human annotations as well as automatic text similarity measures (e.g., cosine similarity).

Various computer vision applications, particularly those involving images and text can benefit from an understanding of which images are specific and which ones are ambiguous. For instance, consider text-based image retrieval. If a query description is moderately similar to the caption of an ambiguous image, that query may be considered a decent match to the image. But if an image is very specific, a moderate similarity between the query and reference descriptions may not be sufficient to retrieve the image. This concept can be applied to improve image descriptions approaches (specific images will have consistent descriptions but ambiguous images will have more varied descriptions), evaluation of image description approaches (perhaps ambiguous images should be penalized less for not matching reference descriptions as compared to specific images), and image tagging (specific images should have fewer tags compared to ambiguous images).

We experiment on three different datasets: MEM-5S (888 images, 5 sentences/image), ABSTRACT-50S (500 images, 50 sentences/image) and PASCAL-50S (1000 images, 50 sentences/image). We show that specificity is a well-defined property of images and characterize specificity in terms of the visual properties of the image. We find that images with people tend to be specific, while mundane images of generic buildings or blue skies do not tend to be specific. In fact, we show that it is possible to predict specificity using image features with a rank correlation of up to 0.35 with ground-truth specificity. We encourage the reader to explore the dataset browser available on the authors' webpages to understand better the underlying factors affecting specificity. The reader may note that specificity is different from image properties such as importance [1, 4], memorability [2] and saliency [3]. In fact, many of these works [1] actually claim that image descriptions are consistent but we show this is not the case.

We also demonstrate the benefit of this notion of specificity on text-based image retrieval. Let us say the user is looking for an image from a database of images. We call this image the target image. The user inputs a query sentence that describes the target image. Every image in the database is associated with a single reference sentence. This can be, for example, the caption in an online photo database such as Flickr. The goal is to sort the images in the database according to their relevance score from most to least relevant, such that the target image has a low rank.

The baseline approach automatically computes a similarity between the query and reference sentence, and sorts the images in descending order using the similarity score. In the proposed approach, instead of sorting just based on similarity score, we model how well this score fits into the distribution of similarity scores for that image by fitting a Logistic Regression (LR) model for every image. The outputs of these LR models will indicate how well the query matches the reference sentence. Figure 2 shows how the learnt LR models improve the ranking accuracy as the number of training sentences increase, indicating that specificity indeed captures the variation in multiple sentences describing an image. We extend this approach to work

Figure 1: Some images are *specific* – they elicit consistent descriptions from different people (left). Other images (right) are ambiguous.

when multiple sentence descriptions are not available to estimate specificity (often encountered in practical applications). We learn a mapping from the image features to the LR model parameters using Support Vector Regressors (SVRs). These predicted LR models are in turn used to estimate how well the query matches the reference sentence, and thus rank the images. All code and data are publicly available on the authors' webpages.

[1] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3562–3569, 2012.

[2] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011.

[3] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[4] Merrielle Spain and Pietro Perona. Measuring and predicting importance of objects in our visual world. 2007.

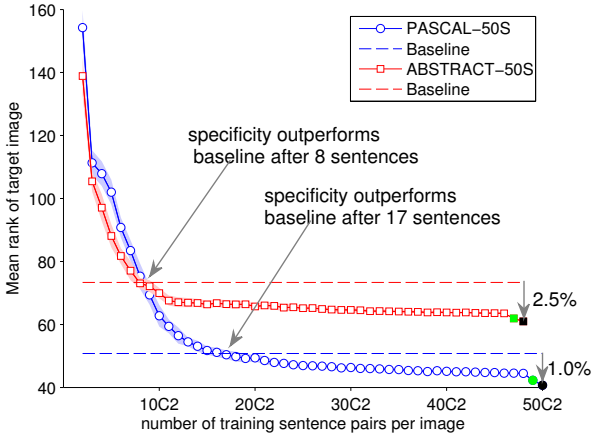

Figure 2: Image retrieval results: Increasing the number of training sentences per image improves the mean target rank obtained using ground-truth specificity. The green dot and black dots correspond to including the reference and query sentences in the training.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.