Deformable Part Models are Convolutional Neural Networks

Xuewen Yang

August 23 2018

Abstract

For some images, descriptions written by multiple people are consistent with each other. But for other images, descriptions across people vary considerably. In other words, some images are specific C they elicit consistent descriptions from different people C while other images are ambiguous. Applications 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 (or reference description) of an ambiguous image, that query may be considered a decent match to the image. But if the image is very specific, a moderate similarity between the query and the reference description may not be sufficient to retrieve the image. In this paper, The author introduce the notion of image specificity. We present two mechanisms to measure specificity given multiple descriptions of an image: an automated measure and a measure that relies on human judgement. They analyze image specificity with respect to image content and properties to better understand what makes an image specific. They then train models to automatically predict the specificity of an image from image features alone without requiring textual descriptions of the image. Finally, they show that modeling image specificity leads to improvements in a text-based image retrieval application.

1. Introduction

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 C 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. The computer vision community has made tremendous progress on recognition problems such as object detection, image classification [1], attribute classification and scene recognition [3]. Various approaches are moving to higher-level semantic image understanding tasks. One such task that is receiving increased attention in recent years is that of automatically generating textual descriptions of images and evaluating these descriptions. However, these works have largely ignored the variance in descriptions produced by different people describing each image. In fact, early works that tackled the image description problem [2] or reasoned about what image content is important and frequently described claimed that human descriptions are consistent. They show that there is in fact variance in how consistent multiple human-provided descriptions of the same image are. Instead of treating this variance as noise, we think of it as a useful signal that if modeled, can benefit applications involving images and text.

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

They introduce the notion of image specificity which measures the amount of variance in multiple viable descriptions of the same image. Modeling image specificity can

benefit a variety of applications. For example, computer-generated image description and evaluation approaches can benefit from specificity. If an image is known to be ambiguous, several different descriptions can be generated and be considered to be plausible. But if an image is specific, a narrower range of descriptions may be appropriate. Photographers, editors, graphics designers, etc. may want to pick specific images C images that are likely to have a single (intended) interpretation across viewers.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[2] M. A. Sadeghi, M. A. Sadeghi, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: generating sentences from images. In *ECCV*, 2010. 1

[3] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1