Single-Image Crowd Counting via Multi-Column Convolutional Neural Network

Liangjie Cao

Jun 14, 2018

Abstract

*Since scenes are composed in part of objects, accurate recognition of scenes requires knowledge about both scenes and objects. In this paper they address two related problems: 1) scale induced dataset bias in multi-scale convolutional neural network (CNN) architectures, and 2) how to combine effectively scene-centric and object-centric knowledge (*i.e. *Places and ImageNet) in CNNs. An earlier attempt, Hybrid-CNN [5], showed that incorporating ImageNet did not help much. Here they propose an alternative method taking the scale into account, resulting in significant recognition gains. By analyzing the response of ImageNet-CNNs and Places-CNNs at different scales they find that both operate in different scale ranges, so using the same network for all the scales induces dataset bias resulting in limited performance. Thus, adapting the feature extractor to each particular scale (*i.e. *scale-specific CNNs) is crucial to improve recognition, since the objects in the scenes have their specific range of scales. Experimental results show that the recognition accuracy highly depends on the scale, and that simple yet carefully chosen multi-scale combinations of ImageNet-CNNs and Places-CNNs, can push the stateof-the-art recognition accuracy in SUN397 up to 66.26% (and even 70.17% with deeper architectures, comparable to human performance).*

1. Introduction

*State-of-the-art in visual recognition is based on the successful combination of deep representations and massive datasets. Deep convolutional neural networks (CNNs) trained on ImageNet (*i.e. *ImageNet-CNNs) achieve impressive performance in object recognition, while CNNs trained on Places (Places-CNNs) do in scene recognition [5, 1]. However, CNNs also have limitations, such as the lack of invariance to significant scaling. This problem is particularly important in scene recognition, due to a wider range of scales and a larger amount of objects per image.*

*In this paper we will study these two problems (*i.e. *dataset bias in patch-based CNNs under different scaling conditions, and how to effectively combine Places and ImageNet) and will see that they are related. Torralba and Efros [2] studied the dataset bias as a cross-dataset generalization problem, in which the same classes may have slightly different feature distributions in different datasets. In our particular case, this bias in the feature distribution is induced by scaling the image. If the scaling operation is considerable, the characteristics of the data may change completely, switching from scene data to object data. Understanding and quantifying this bias can help us to design better multi-scale architectures, and even better ways to combine object and scene knowledge. In particular, the authors propose multi-scale architectures with scale-specific networks as a principled way to address scale-related dataset bias and combine scene and object knowledge (*i.e. *Places and ImageNet).*

2. Objects in object datasets and scene datasets

The knowledge learned by CNNs lies in the data seen during training, and will be of limited use if tested in a different type of data. Thus, CNNs trained with ImageNet are limited when used for scene recognition due to this training/test bias, while Places-CNNs perform better in this task. While this is essentially true, objects and scenes are closely related, so knowledge about objects can be still helpful to recognize scenes, if used properly.

To evaluate the dataset bias we use SUN397 [4] as target dataset. Since Places contains scene data, with 205 scene categories overlapping with SUN397, and significantly more data, we can expect a low dataset bias. Thus we focus on ImageNet (in particular ILSVRC2012), which contains mostly object data. Fortunately, both ImageNet and SUN have a fraction of images with region annotations and labels, so we can collect some relevant statistics and compare their distributions (we used the LabelMe toolbox [3]). Since we will use this information to interpret the variations in recognition accuracy in next experiements.

Fig. 1a shows the distribution of object sizes, and Fig. 1c some examples of objects of different normalized sizes. They normalized the size of the object relative to the equivalent training crop. While objects in ImageNet are mostly large, often covering the whole image, objects in SUN397

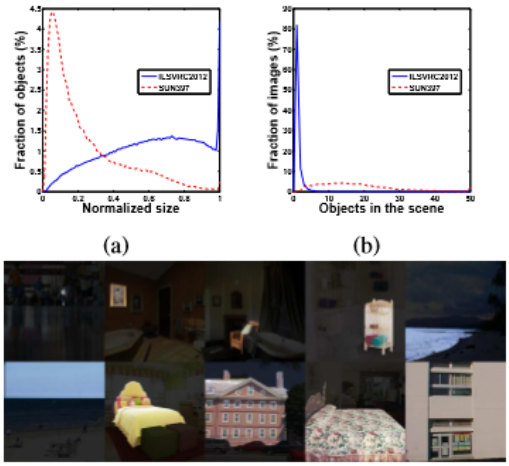

Figure 1. **Characteristics of objects in ILSVRC2012 (object data) and SUN397 (scene data): (a) distribution of objects sizes (normalized), (b) number of objects per scene, and (c) examples of objects by increasing normalized size**

are much smaller, corresponding to the real distribution in scenes. Thus Fig. 1a shows an obvious mismatch between both datasets.

3. Dataset bias in object recognition

In order to study the behaviour of ImageNet-CNNs and Places-CNNs in object recognition, they need object data extracted from scenes datasets. The authors selected 100 images per category from the 75 most frequent object categories in SUN397, so they can have enough images to train SVM classifiers. They took some precautions to avoid selecting too small objects. In contrast to most object and scene datasets, in this case they have the segmentation of the object within the scene, so they can use it to create variations over the same objects. Then they created four variations (see Fig. 2): original masked, original with background, canonical masked and canonical with background. In particular, to study the response to different scaling, the canonical variant is scaled in the range 10%-100%. Note how scaling the variant with background shifts progressively the content of the crop from object to scene.

4. Discriminability and redundancy

Accuracy provides a good indirect measure of the utility of the feature for a given target task (e.g. *scene recognition) via a classifier* (e.g. *SVM). Here they also consider two information theoretic metrics measuring directly the discriminability and redundancy of the deep feature. They define the discriminability of a feature x = (*x_1, ,x_4096*) with respect to a set of classes C = 1, ,M*

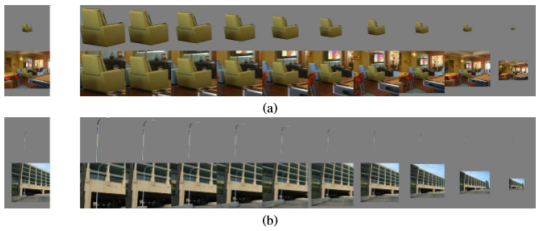

Figure 2. **The two variants used in the object recognition experiments: object masked (top row) and object with background (bottom row) with two examples of (a) armchair and (b) streetlight. Left crops show the object in the original scale in the scene. Right crops show the object scaled progressively from the canonical size (100%) down to 10%. All the images are centered in the object of interest.**

$$D(x, C) = \frac{1}{\mid C \mid\mid S \mid} \sum_{c \in C} \sum_{x_i \in X} I(x_i; c) \qquad (1)$$

where $I(x_i; c)$ *is the filter* x_i *and the class c. In order to evaluate how redundant is the feature (compared with other filters), they use the redundancy of a feature* x*, defined as*

$$R(x) = \frac{1}{\mid S \mid_2} \sum_{x_i \in x} \sum_{x_j \in x} I(x_i; x_j) \qquad (2)$$

References

[1] *J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: a deep convolutional activation feature for generic visual recognition. In* ICML*, 2013.*

[2] *A. Torralba and A. A. Efros. Unbiased look at dataset bias. In* CVPR*, 2011.*

[3] *A. Torralba, B. C. Russell, and J. Yuen. LabelMe: online image annotation and applications.* Proceedings of the IEEE*, 2010.*

[4] *J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: largescale scene recognition from abbey to zoo. In* CVPR*, 2016.*

[5] *B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In* NIPS*, 2014.*