Object Detectors Emerge from Training CNNs for Scene Recognition

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bolei,khosla,agata,oliva,torralba}@mit.edu

Abstract

With the success of new computational architectures for visual processing, such as convolutional neural networks (CNN) and access to image databases with millions of labeled examples (e.g., ImageNet, Places), the state of the art in computer vision is advancing rapidly. One important factor for continued progress is to understand the representations that are learned by the inner layers of these deep architectures. Here we show that object detectors emerge from training CNNs to perform scene classification. As scenes are composed of objects, the CNN for scene classification automatically discovers meaningful objects detectors, representative of the learned scene categories. With object detectors emerging as a result of learning to recognize scenes, our work demonstrates that the same network can perform both scene recognition and object localization in a single forward-pass, without ever having been explicitly taught the notion of objects.

1. Introduction

Learning to classify scenes (i.e., classifying an image as being an office, a restaurant, a street, etc) using the Places dataset [8] gives the opportunity to study the internal representation learned by a CNN on a task other than object recognition. In the case of scenes, the representation is clearer. Scene categories are defined by the objects they contain and, to some extent, by the spatial configuration of those objects. For instance, the important parts of a bedroom are the bed, a side table, a lamp, a cabinet, as well as the walls, floor and ceiling. Objects represent therefore a distributed code for scenes (i.e., object classes are shared across different scene categories). Importantly, in scenes, the spatial configuration of objects, although compact, has a much larger degree of freedom. It is this loose spatial dependency that, we believe, makes scene representation different from most object classes (most object classes do not have a loose interaction between parts). In addition to objects, other feature regularities of scene categories allow for other representations to emerge, such as textures [6], GIST [4], bag-of-words [11], part-based models [5], and ObjectBank [3]. While a CNN has enough flexibility to learn any of those representations, if meaningful objects emerge without supervision inside the inner layers of the CNN, there will be little ambiguity as to which type of representation these networks are learning.

The main contribution of this work [7] is to show that object detection emerges inside a CNN trained to recognize scenes, even more than when trained with ImageNet. This is surprising because our results demonstrate that reliable object detectors are found even though, unlike ImageNet, no supervision is provided for objects. Although object discovery with deep neural networks has been shown before in an unsupervised setting [2], here we find that many more objects can be naturally discovered, in a supervised setting tuned to scene classification rather than object classification.

2. Uncovering the CNN Representations

The deep features from Places-CNN tend to perform better on scene-related recognition tasks compared to the features from ImageNet-CNN. For example, as compared to the Places-CNN that achieves 50.0% on scene classification, the ImageNet-CNN combined with a linear SVM only achieves 40.8% on the same test set. The performance of scene recognition using Places-CNN is quite impressive given the difficulty of the task.

To understand why Places-CNN works so well and the nature of the representation that the network is learning, we take several approaches: Firstly, we simplify the input images by removing some image parts to see how the prediction of CNN various. This simplified image (named minimal image representation) allows us to highlight the elements that lead to the high classification score. We found that objects seem to contribute important information for the network to recognize the scene. For instance, in the case of bedrooms these minimal image representations usually contain the region of the bed, or in the art gallery category, the regions of the paintings on the walls. Secondly, we investigate the shape and size of the receptive fields (RFs) of the various units in the CNNs. A data-driven approach to
estimate the learned RF of each unit in each layer, then we could use the estimated RF to segment the images. We find that the units at different layers have different activation regions, while the activation regions tended to become more semantically meaningful with increasing depth of layers. To further understand and quantify the precise semantics learned by each unit, we ask workers on Amazon Mechanical Turk (AMT) to identify the common theme or concept that exists between the top scoring segments for each unit. For each unit, after shown the image segments worker is asked to write a word to describe the most possible description for the segments, cross out the segments which don’t correspond to the description, then categorize the word to one of 6 semantic groups ranging from low-level to high-level.

Fig. 1 reveals how different levels of abstraction emerge in different layers of both networks. The vertical axis indicates the percentage of units in each layer assigned to each concept category. ImageNet-CNN has more units tuned to simple elements and colors than Places-CNN while Places-CNN has more objects and scenes. ImageNet-CNN has more units tuned to object parts (with the maximum around conv4). It is interesting to note that Places-CNN discovers more objects than ImageNet-CNN despite having no object-level supervision.

Fig. 2 shows some units from the Places-CNN grouped by the type of object class they seem to be detecting. Each row shows the top five images for a particular unit that produce the strongest activations. The segmentation shows the regions of the image for which the unit is above a certain threshold. Each unit seems to be selective to a particular appearance of the object. For instance, there are 5 units for lightning, each unit detecting a particular type of lightning devices providing finer-grained discrimination; there are 9 units selective to people, each one tuned to different scales or people doing different tasks.

From our analysis above, many of the units in the inner layers could perform interpretable object localization. Thus we could use this single Places-CNN with the annotation of units to do both scene recognition and object localization in a single forward-pass. We would explore other potential application of the CNN trained for scene recognition in the future work.

References