Deformable Part Models are Convolutional Neural Networks

Xuewen Yang

August 13 2018

Abstract

Deformable part models (DPMs) and convolutional neural networks (CNNs) are two widely used tools for visual recognition. They are typically viewed as distinct approaches: DPMs are graphical models (Markov random fields), while CNNs are black-box non-linear classifiers. In this paper, the authro show that a DPM can be formulated as a CNN, thus providing a synthesis of the two ideas. Our construction involves unrolling the DPM inference algorithm and mapping each step to an equivalent CNN layer. From this perspective, it is natural to replace the standard image features used in DPMs with a learned feature extractor. We call the resulting model a DeepPyramid DPM and experimentally validate it on PASCAL VOC object detection. We find that DeepPyramid DPMs significantly outperform DPMs based on histograms of oriented gradients features (HOG) and slightly outperforms a comparable version of the recently introduced R-CNN detection system, while running significantly faster.

1. Introduction

Part-based representations are widely used in visual recognition. In particular, deformable part models (DPMs) have been effective for generic object category detection. DPMs update pictorial structure models [2], which date back to the 1970s, with modern image features and machine learning algorithms.

Convolutional neural networks (CNNs) are another influential class of models for visual recognition. CNNs also have a long history, and have resurged over the last two years due to good performance on image classification, object detection, and more recently a wide variety of vision tasks (e.g., [3] [4]).

These two models, DPMs and CNNs, are typically viewed as distinct approaches to visual recognition. DPMs are graphical models (Markov random fields), while CNNs are black-box non-linear classifiers. Many people ask: Are these models actually distinct? To answer this question the author show that any DPM can be formulated as an equivalent CNN. In other words, deformable part models are convolutional neural networks. This construction relies on a new network layer, distance transform pooling, which generalizes max pooling.

DPMs typically operate on a scale-space pyramid of gradient orientation feature maps (HOG [1]). But the author now know that for object detection this feature representation is suboptimal compared to features computed by deep convolutional networks. As a second innovation, they replace HOG with features learned by a fully-convolutional network. This front-end network generates a pyramid of deep features, analogous to a HOG feature pyramid. They call the full model a DeepPyramid DPM.

2. DeepPyramid DPMs

A DeepPyramid DPM is a convolutional network that takes as input an image pyramid and produces as output a pyramid of object detection scores. Although the model is a single network, for pedagogical reasons we describe it in terms of two smaller networks, a feature pyramid frontend CNN and a DPM-CNNłtheir function composition yields the full network. A schematic diagram of the model is presented in Figure 1.

2.1. Feature pyramid frontend CNN

Objects appear at all scales in images. A standard technique for coping with this fact is to run a detector at multiple scales using an image pyramid. In the context of CNNs, this method dates back to (at least) early work on face detection, and has been used again in contemporary works, including OverFeat, DetectorNet, DenseNet, and SPP-net, a recently proposed method for speeding up R-CNNs. We follow this approach and use as our front-end CNN a network that maps an image pyramid to a feature pyramid. To do this, we use a standard single-scale architecture (Krizhevsky *et al.*) and tie the network weights across all scales.

2.2. Constructing an equivalent CNN from a DPM

In the DPM formalism, an object class is modeled as a mixture of components, each being responsible for modeling the appearance of an object sub-category (e.g., side

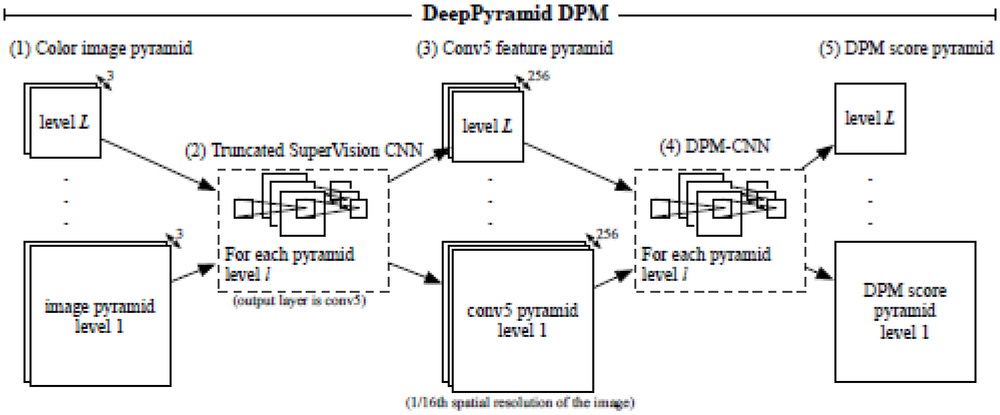

Figure 1. Schematic model overview. (1) An image pyramid is built from a color input image. (2) Each pyramid level is forward propagated through a fully-convolutional CNN (e.g., a truncated SuperVision CNN [27] that ends at convolutional layer 5). (3) The result is a pyramid of conv5 feature maps, each at 1/16th the spatial resolution of its corresponding image pyramid level. (4) Each conv5 level is then input into a DPM-CNN, which (5) produces a pyramid of DPM detection scores. Since the whole system is the composition of two CNNs, it can be viewed as a single, unified CNN that takes a color image pyramid as input and outputs a DPM score pyramid.

views of cars, people doing handstands, bi-wing propeller planes). Each component, in turn, uses a low-resolution global appearance model of the sub-type (called a root filter), together with a small number of higher resolution part filters that capture the appearance of local regions of the sub-type.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[2] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005. 1

[3] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 1

[4] J. Tompson, A. Jain, Y. Lecun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 1