Deformable Part Models are Convolutional Neural Networks

Ross Girshick[1] Forrest Iandola[2] Trevor Darrell[2] Jitendra Malik[2]
[1]Microsoft Research [2]UC Berkeley

Part-based representations are widely used in visual recognition. In particular, deformable part models (DPMs) [4] have been effective for generic object category detection. DPMs update pictorial structure models, which date back to the 1970s [5], 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 [6, 9, 10], and have resurged over the last two years due to good performance on image classification [8], object detection [7], and more recently a wide variety of vision tasks.

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. In this paper, we ask: Are these models actually distinct? To answer this question we show that any DPM can be formulated as an equivalent CNN (see: Figure 1, Figure 2). In other words, deformable part models *are* convolutional neural networks. Our 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 [3]). But we now know that for object detection this feature representation is suboptimal compared to features computed by deep convolutional networks [7]. As a second innovation, we 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. We call the full model a *DeepPyramid DPM*.

We experimentally validate DeepPyramid DPMs by measuring object detection performance on PASCAL VOC. Since traditional DPMs have been tuned for HOG features over many years, we first analyze the differences between HOG feature pyramids and deep feature pyramids (see: Figure 3). We then select a good model structure and train a DeepPyramid DPM that significantly outperforms the best HOG-based DPMs. While we don't expect our approach to outperform a fine-tuned R-CNN detector [7], we do find that it slightly outperforms a comparable R-CNN (specifically, an R-CNN on the same conv_5 features, without fine-tuning), while running about 20x faster (0.6s *vs.* 12s per image).

Our experiments also shed some light the relative merits of region-based detection methods, such as R-CNN, and sliding-window methods like DPM. We find that region proposals and sliding windows are complementary approaches that will likely benefit each other if used in an ensemble. This makes sense; some object classes are easy to segment (*e.g.*, cats) while others are difficult (*e.g.*, bottles, people).

Interpreted more generally, this paper shows that sliding-window detectors on deep feature pyramids significantly outperform equivalent models on HOG. While not surprising, the implementation details are crucial and challenging to pin down. As a result, HOG-based detectors are still used in a wide range of systems, such as recent hybrid deep/conventional approaches [2], and especially where region-based methods are ill-suited (poselets [1] being a prime example). We therefore believe that the results presented in this paper will be of broad practical interest to the visual recognition community. An open-source implementation will be made available on the first author's website, which will allow researchers to easily build on our work.

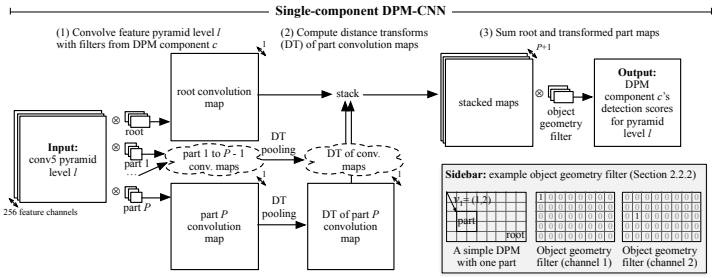

Figure 1: CNN equivalent of a single-component DPM.

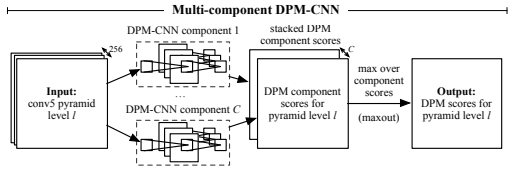

Figure 2: Multi-component DPMs are implemented by maxout units.

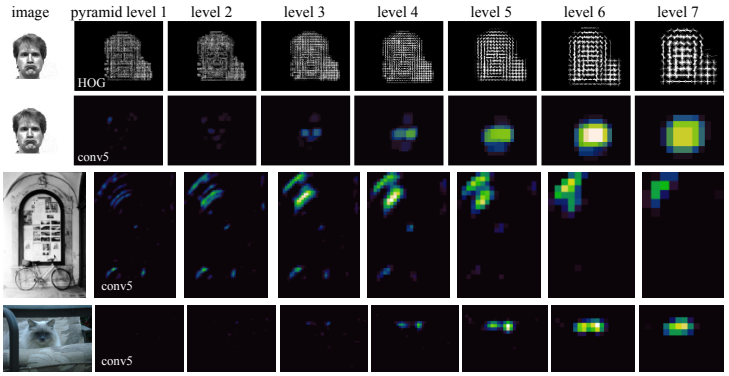

Figure 3: HOG versus conv_5 feature pyramids. In contrast to HOG features, conv_5 features have sparse activations in position and scale, much like part detectors. Each conv_5 pyramid shows 1 of 256 feature channels.

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[2] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014.

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

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010.

[5] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computer*, 1973.

[6] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 1980.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[9] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comp.*, 1989.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. *Parallel Distributed Processing*, 1:318–362, 1986.

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