Modeling Local and Global Deformations in Deep Learning: Epitomic Convolution, Multiple Instance Learning, and Sliding Window Detection

George Papandreou[1], Iasonas Kokkinos[2], Pierre-André Savalle[2]
[1]Google. [2]CentraleSupélec and INRIA.

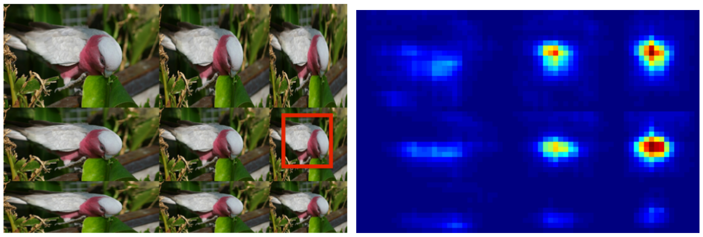

Figure 1: Image deformations can challenge high-level vision, but modeling their effects can lead to simple and accurate recognition algorithms. Here we show how our object detection system performs scale, position and aspect ratio search: scaled and squeezed versions of an image are fed to a fully convolutional DCNN, until at some point the object can be contained in a square of fixed size. At that point the detector's score (shown on the right) is maximized, providing a tight bounding box around the object. Our detector only has to consider normalized object instances.

Abstract Deep Convolutional Neural Networks (DCNNs) achieve invariance to domain transformations (deformations) by using multiple 'max-pooling' (MP) layers. In this work we show that alternative methods of modeling deformations can improve the accuracy and efficiency of DCNNs. First, we introduce *epitomic convolution* as an alternative to the common convolution-MP cascade of DCNNs, that comes with the same computational cost but favorable learning properties. Second, we introduce a *Multiple Instance Learning* algorithm to accommodate global translation and scaling in image classification, yielding an efficient algorithm that trains and tests a DCNN in a consistent manner. Third we develop a DCNN sliding window detector that explicitly, but efficiently, searches over the object's position, scale, and aspect ratio.

Key results We demosnstrate competitive image classification and localization results on the ImageNet dataset, achieving 10.0% top-5 classification error rate. We also report competitive object detection results on the Pascal VOC2007 dataset, achieving 58.6% mAP, using a sliding window scheme which is significantly simpler faster than state-of-art systems based on object proposals.

Methods Over the last few years Deep Learning has been the method of choice for image classification [6, 9, 10], while a host of other works have shown that the features learned by deep neural networks can be successfully employed in other tasks [1, 3, 7, 8, 11].

In this work we aim at enriching the set of tools used to model deformations in DCNNs, by exploiting established computer vision techniques, such as image epitomes, multi-scale pyramids and Procrustes analysis. We combine these techniques with ideas from machine learning (back-propagation, multiple-instance learning), object recognition (image patchworks, sliding window detection), and signal processing (the à trous algorithm) and develop algorithms of higher accuracy and/or efficiency than the ones obtained using more standard deformation modeling tools.

First, we deal with the modeling of **local deformations in image classification**. For this we introduce the epitomic image representation [5] into the setting of DCNNs. While originally developed for generative image modeling, we show here that epitomes can also be used to train DCCNs discrimatively. Epitomic convolution comes at the exact same computation cost but exhibits faster convergence during training and higher classification accuracy when compared to standard max-pooled convolution.

Second, we address the modeling of **global transformations in image classification**. Our goal is to explicitly deal with object scale and position when applying DCNNs to image classification. While a standard practice is to fuse classification results extracted from multiple image windows, we show that by using a principled Multiple Instance Learning (MIL) framework we obtain substantially larger gains. An algorithmic contribution is that we show how MIL can be efficiently implemented for fully-convolutional DCNNs by compacting an image pyramid into the patchwork data structure of [2].

Finally, we turn to the modeling of **global transformations in object detection**. Rather than using region proposals to come up with candidate object boxes, we explicitly search over positions, scales, and aspect ratios, as illustrated in Fig. 1. This can be understood as a variant of Procrustes analysis, where global deformations are first discarded before performing a finer deformation modeling. We show that by performing this explicit search over position, scale, and aspect ratios we can obtain results that are comparable to the current-state-of-the-art while being substantially simpler and easier to train, as well as six times faster, thanks to the sharing of computation during convolutions. An algorithmic contribution that we introduce in this context is that we accelerate sliding window detection by using the à trous (with holes) algorithm to reduce the effective size and receptive field of a DCNN pre-trained on ImageNet.

We have implemented the proposed methods using Caffe [4]. Code and models reproducing the results in this paper are made publicly available from `http://cvn.ecp.fr/iasonas/deepdet`.

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arXiv, 2014.

[2] Charles Dubout and François Fleuret. Exact acceleration of linear object detectors. In *Computer Vision–ECCV 2012*, pages 301–311. Springer Berlin Heidelberg, 2012.

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

[4] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding, 2013.

[5] N. Jojic, B. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *Proc. ICCV*, pages 34–41, 2003.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2013.

[7] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *Proc. ICCV*, 2013.

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. 2014.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. http://arxiv.org/abs/1409.1556/.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[11] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. arXiv, 2013.

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