Return of the Devil in the Details: Delving Deep into Convolutional Nets

Ken Chatfield
ken@robots.ox.ac.uk
Karen Simonyan
karen@robots.ox.ac.uk
Andrea Vedaldi
vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

The latest generation of Convolutional Neural Networks (CNN) have been shown to achieve impressive results in challenging benchmarks on image recognition and object detection, significantly raising the interest of the community in these methods. Nevertheless, it is still unclear how different CNN methods compare with each other and with previous state-of-the-art shallow representations such as the Bag-of-Visual-Words and the Improved Fisher Vector (IFV). This paper conducts a rigorous evaluation of these new techniques, exploring different deep architectures and comparing them on a common ground, identifying and disclosing important implementation details in a similar vein to our previous work on shallow encoding methods [1].

We identify several useful properties of CNN-based representations, including the fact that the dimensionality of the CNN output layer can be reduced significantly without having an adverse effect on performance. We also identify aspects of deep and shallow methods that can be successfully shared. In particular, we show that the data augmentation techniques commonly applied to CNN-based methods can also be applied to shallow methods, and result in an analogous performance boost.

Evaluation over multiple standard benchmark datasets is presented (PASCAL VOC 2007 and 2012, Caltech-101, Caltech-256 and ILSVRC-2012) and our best CNN-based method achieves performance comparable to state-of-the-art over all four (refer to Table 1). We also present a variety of other configurations, each striking a different trade-off in the balance between performance, computation speed and compactness.

As with our previous work, source code and CNN models to reproduce the experiments presented in the paper are available from the project webpage[1] to provide common ground for future comparisons, and good baselines for image representation research.

1 CNN-based Methods

Our **Fast (CNN-F)** method provides the fastest computation time, and is similar in architecture to the one used by Krizhevsky *et al*. [3], our **Medium (CNN-M)** method strikes balance between being relatively fast to compute and greater performance, being loosely based on the architecture of Zeiler and Fergus [7]. Finally, our **Slow (CNN-S)** method focuses on maximum performance, and is similar architecturally to the 'accurate' network from the OverFeat package [6]. We further investigate the impact of: (a) different data augmentation strategies, (b) reducing the output dimensionality of the output layer and (c) the performance boost (if any) possible by fine-tuning the networks to the target dataset.

2 Compared to Shallow Methods

By applying data augmentation techniques similar to with CNN-based methods to IFV, we obtain a performance boost to 68.0% on the PASCAL VOC 2007 benchmark. We further investigate the impact of: (a) different IFV normalisation and spatial information encoding strategies, (b) adding colour information to shallow features, or removing it from CNN-based methods and (c) combining IFV with CNN-based methods into a single fused representation.

3 Performace Evolution over PASCAL VOC 2007

A comparative plot of the evolution in the performance of the methods evaluated in this paper, along with a selection from our earlier review of shallow methods [1] is presented in Fig. 1. Classification accuracy over PASCAL VOC was 54.48% mAP for the BoVW model in 2008, 61.7% for the IFV in 2010 [1], and 73.41% for DeCAF [2] and similar [4, 5] CNN-based methods introduced in late 2013. Our best performing CNN-based method (CNN-S with fine-tuning) achieves 82.42%, comparable to the most recent state-of-the-art.

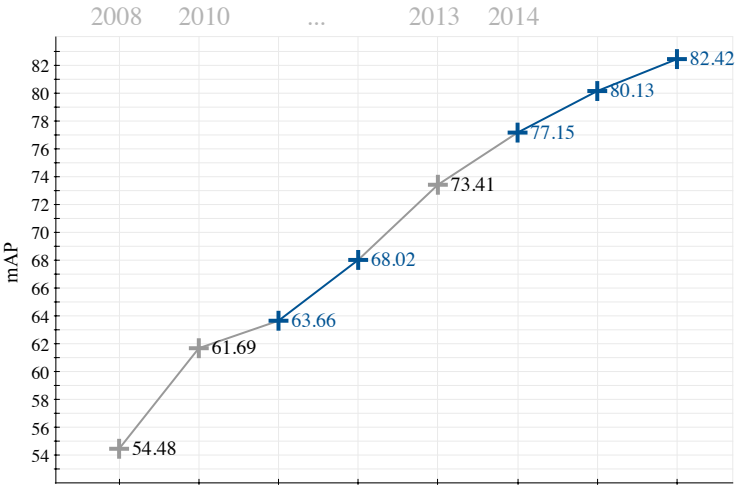

Figure 1: **Evolution of Performance on PASCAL VOC-2007 over the recent years.** Refer to Table 2 in the paper for details and references.

	ILSVRC-2012 (top-5 error)	VOC-2007 (mAP)	VOC-2012 (mAP)	Caltech-101 (accuracy)	Caltech-256 (accuracy)
FK IN	-	65.4	–	–	–
FK IN +aug	-	68.0	–	–	–
CNN F	16.7	77.4	79.9	–	–
CNN M	13.7	79.9	82.5	87.15 ± 0.80	77.03 ± 0.46
CNN S	**13.1**	79.7	82.9	87.76 ± 0.66	**77.61 ± 0.12**
CNN S TN	**13.1**	**82.4**	**83.2**	**88.35 ± 0.56**	77.33 ± 0.56

Table 1: **Sample of key results from the paper** on ILSVRC2012, VOC2007, VOC2012, Caltech-101, and Caltech-256. 'TN' – dataset-specific fine-tuning. For IFV, '+aug' indicates full data-augmentation.

[1] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. BMVC.*, 2011.

[2] 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. *CoRR*, abs/1310.1531, 2013.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[4] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proc. CVPR*, 2014.

[5] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *CoRR*, abs/1403.6382, 2014.

[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *Proc. ICLR*, 2014.

[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[1] http://www.robots.ox.ac.uk/~vgg/research/deep_eval/