Understanding Deep Image Representations by Inverting Them

Aravindh Mahendran[1], Andrea Vedaldi[1]
[1]Department of Engineering Science, University of Oxford

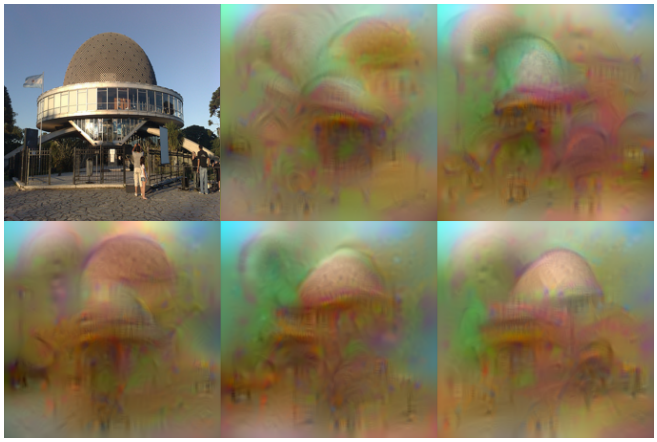

Figure 1: **What is encoded by a CNN?** The figure shows five possible reconstructions of the reference image obtained from the 1,000-dimensional code extracted at the penultimate layer of a reference CNN[3] (before the softmax is applied) trained on the ImageNet data. From the viewpoint of the model, all these images are practically equivalent. This figure is best viewed in color/screen.

Several image understanding and computer vision methods build on image representations such as textons [4], histogram of oriented gradients (SIFT [5] and HOG [2]), bag of visual words [1][8], sparse [12] and local coding [11], Fisher Vectors [6], and, lately, deep neural networks, particularly of the convolutional variety [3, 7, 13]. However, despite the progress in the development of visual representations, their design is still driven empirically and a good understanding of their properties is lacking. While this is true of shallower hand-crafted features, it is even more so for the latest generation of deep representations, where millions of parameters are learned from data.

In this paper we conduct a direct analysis of representations by characterising the image information that they retain (Fig. 1). We do so by modeling a representation as a function $\Phi(\mathbf{x})$ of the image $\mathbf{x}$ and then computing an approximated inverse ϕ^{-1}, *reconstructing* $\mathbf{x}$ *from the code* $\Phi(\mathbf{x})$. A common hypothesis is that representations collapse irrelevant differences in images (e.g. illumination or viewpoint), so that Φ should not be uniquely invertible. Hence, we pose this as a reconstruction problem and find a number of possible reconstructions rather than a single one. By doing so, we obtain insights into the invariances captured by the representation.

Our contributions are as follows. First, we propose a general method to invert representations, including SIFT, HOG, and CNNs. Crucially, this method **uses only information from the image representation** and a generic natural image prior, starting from random noise as initial solution, and hence captures only the information contained in the representation itself. We discuss and evaluate different regularization penalties as natural image priors. Second, we show that, despite its simplicity and generality, this method recovers significantly better reconstructions from HOG compared to recent alternatives [10]. As we do so, we emphasise a number of subtle differences between these representations and their effect on invertibility. Third, we apply the inversion technique to the analysis of recent deep CNNs, exploring their invariance by sampling possible approximate reconstructions. We relate this to the depth of the representation, showing that the CNN gradually builds an increasing amount of invariance, layer after layer (See Fig. 2). Fourth, we study the locality of the information stored in the representations by reconstructing images from selected groups of neurons, either spatially or by channel.

The neural network models for HOG and DSIFT and the MATLAB code for this paper are available from http://www.robots.ox.ac.uk/~vgg/research/invrep/index.htm. We use the MatConvNet toolbox [9] for implementing convolutional neural networks.

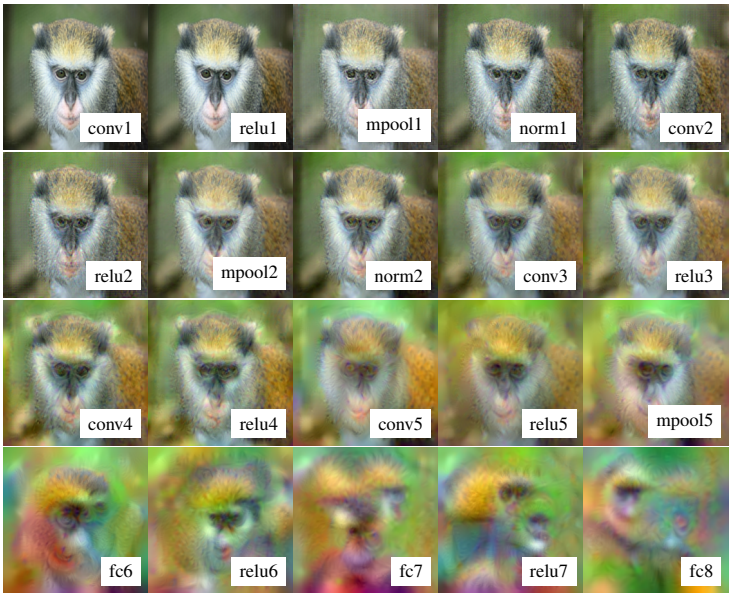

Figure 2: **CNN reconstruction.** Reconstruction of a test image from CNN features. This figure is best viewed in color/screen.

[1] G. Csurka, C. R. Dance, L. Dan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Stat. Learn. in Comp. Vision*, 2004.

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

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1), 2001.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004.

[6] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorizaton. In *CVPR*, 2006.

[7] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *CoRR*, volume abs/1312.6229, 2014.

[8] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[9] Andrea Vedaldi and Karel Lenc. MatConvNet: CNNs for MATLAB. http://www.vlfeat.org/matconvnet/, 2014.

[10] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: Visualizing object detection features. In *ICCV*, 2013.

[11] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *CVPR*, 2010.

[12] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *CVPR*, 2010.

[13] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

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