Deconvolutional Networks for Feature Learning

Matt Zeiler, Dilip Krishnan, Graham Taylor, Rob Fergus

Courant Institute of Mathematical Sciences, New York University

{*zeiler,dilip,gwtaylor,fergus*}*@cs.nyu.edu*

Introduction

Building robust low-level image representations, beyond edge primitives, is a long-standing goal in vision. In its most basic form, an image is a matrix of intensities. How we should progress from this matrix to stable mid-level representations, useful for high-level vision tasks, remains unclear. Popular feature representations such as SIFT or HOG spatially pool edge information to form descriptors that are invariant to local transformations. However, in doing so important cues such as edge intersections, grouping, parallelism and symmetry are lost.

The primal sketch scheme of Marr [1] proposed that image representations should be built in stages, starting with image primitives such as edges. These would then be grouped using geometric constraints to form "tokens" and thence to larger-scale tokens that capture the high-level content of images such as objects. While this idea has a certain elegance, the difficulties lie in its instantiation: What should the tokens be? How should the grouping occur? Which geometric constraints are important and which can be ignored? Robust schemes that address these issues have proved to be difficult to formulate.

In this paper we propose a new architecture, called the Deconvolutional Network, that permits the automatic construction of hierarchical image representations which embody the key elements of Marr's primal sketch. Each level of the hierarchy groups information from the level beneath to form more complex tokens that exist over a larger scale

Figure 1: **(a)**: "Tokens" from Fig. 2-4 of *Vision* by D. Marr [1]. These idealized local groupings are proposed as an intermediate level of representation in Marr's primal sketch theory. **(b)**: Selected filters from the 3rd layer of our hierarchical model, trained in an unsupervised fashion on real-world images using convolutional sparse image decomposition.

in the image. Our grouping cue is sparsity: by encouraging parsimonious representations at each level of the hierarchy, tokens will naturally group into more complex structures. However, as we demonstrate, sparsity itself is not enough – it must be deployed within the correct architecture to have the desired effect. We adopt a convolutional approach since it provides stable latent representations at each level which preserve locality and thus facilitates the grouping mechanism.

A key feature of our approach is that it is entirely unsupervised. At each level we solve a sparse, over-complete decomposition of the latent representation from the layer beneath. Using the same parameters for learning each layer, we can automatically build rich features that correspond to concepts such as edge junctions, parallel lines, curves and basic geometric elements, such as squares. Remarkably, some look very similar to the tokens posited by Marr in his description of the primal sketch (see Fig. 1). A technical contribution of our paper concerns the convolutional sparse decomposition. This is a challenging optimization problem that must be effectively solved if useful features are to spontaneously emerge. Standard techniques perform poorly and we propose an alternative scheme that works well in practice.

Model

The Deconvolutional Network is top-down in nature as it decomposes an image into a set of feature maps with no mechanism for generating the feature maps directly from the input such as the sparse auto-encoder approaches of Ranzato et al. [2], or DBNs [3]. In this model, the feature maps are inferred while jointly learning a set of filters from which the previous layer feature maps (or the input image planes in the case of the first layer) can be reconstructed by convolving over the feature maps and summing their contributions. Shown in Fig. 2 is a single layer architecture with the layer-l feature maps, $z_{i,l}$ convolved with layer-l filters, $f_{i,j}^l$ to reconstruct layer-$(l-1)$ feature maps below.

Learning proceeds by minimizing a simple cost function enforcing the reconstruction of the input maps from the top-down convolutions while imposing sparsity over the feature maps. This sparsity takes the form of an L^p-norm where p is allowed to be ≤ 1. Using the minimization technique proposed by Krishnan and Fergus [4], inference is fast.
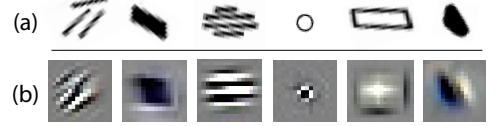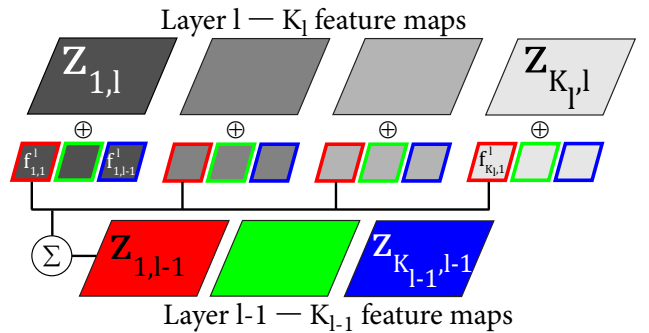

Figure 2: A single layer of convolutional sparse decomposition (best viewed in color). For clarity, only the connectivity for a single input map is shown. In practice the first layer is fully connected, while the connectivity of the higher layers is specified by a map g^l, which is sparse.

Experiments

In our experiments, we analyzed the filters learned using two datasets of 100×100 images, one containing natural scenes of fruits and vegetables and the other consisting of scenes of urban environments. The first layer had 9 feature maps fully-connected to the input image planes. The second layer had 45 maps: 36 of which were connected to pairs of maps in the first layer, and the remainder were singly-connected. In Fig. 3 we show the filters that spontaneously emerge, projected back into pixel space.

(a) Layer-1 Natural Images

(b) Layer-1 City Images

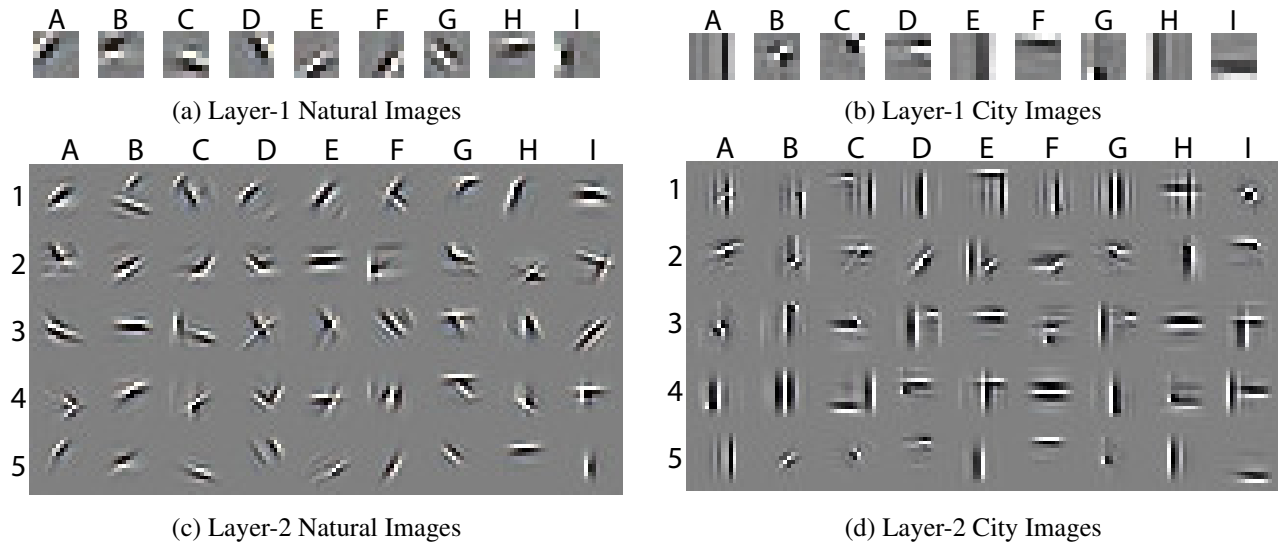(c) Layer-2 Natural Images

(d) Layer-2 City Images

Figure 3: Filters from each layer in our model, trained separately on scenes of natural images and city images. Note the rich diversity of filters and their increasing complexity with each layer. Note the predominance of horizontal and vertical structures in the city image filters whereas the natural image filters are more evenly distributed over orientation.

We evaluated our model in a recognition setting on the Caltech 101 dataset. Using the inferred features maps of both the first and second layers, we trained an SVM classifier using a spatial pyramid matching kernel. As with previous work [5], the dataset was split such that 30 images per category were reserved for training while the remaining images (up to a maximum of 50 per category) were used for testing. These splits were conducted 5 times and the average performance and standard deviation are reported below. Using a single layer model with 8 feature maps (for direct comparison to the SIFT spatial pyramid of [5]) we report a recognition rate of $62.8 \pm 1.1\%$ comparable to the SIFT performance of $64.6 \pm 0.7\%$. Alterations to that direct comparison easily increased our one layer recognition performance to $66.2 \pm 1.2\%$ and adding the second layer further increased performance to $67.0 \pm 1.0\%$ surpassing the performance of convolutional DBNs [6] ($65.5 \pm 0.5\%$) and the SVM-KNN of Zhang et. al. [7] ($66.2 \pm 0.5\%$).

Conclusion

We have presented the Deconvolutional Network, a conceptually simple framework for learning sparse, over-complete feature hierarchies. Applying this framework to natural images produces a highly diverse set of filters that capture high-order image structure beyond edge primitives. These arise without the need for hyper-parameter tuning or additional modules, such as local contrast normalization, max-pooling and rectification [8]. Our approach relies on robust optimization techniques to minimize the poorly conditioned cost functions that arise in the convolutional setting. While our algorithm is slow compared to approaches that use bottom-up encoders, heavy use of the convolution operator makes it amenable to parallelization and GPU implementations which could give significant speed gains.

References

[1] D. Marr. *Vision*. Freeman, San Francisco, 1982.
[2] M. Ranzato, Y-L. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS*. MIT Press, 2008.
[3] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.
[4] D. Krishnan and R. Fergus. Analytic Hyper-Laplacian Priors for Fast Image Deconvolution. In *NIPS*, 2009. To appear.
[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[6] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, pages 609–616, 2009.
[7] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Disciminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.
[8] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *ICCV*, 2009.

Topic: visual processing and pattern recognition   Preference: oral