

# Learning Sparse and Invariant Features Hierarchies

Y-Lan Boureau<sup>1</sup>, Marc'Aurelio Ranzato<sup>1</sup>, Fu Jie Huang<sup>1</sup> and Yann LeCun<sup>1</sup>

<sup>1</sup> The Courant Institute of Mathematical Sciences - New York University

Understanding how the visual cortex builds invariant representations is one of the most challenging problems in visual neuroscience. The feed-forward, multi-stage Hubel and Wiesel architecture [1, 2, 3, 4, 5] stacks multiple levels of alternating layers of simple cells that perform feature extraction, and complex cells that pool together features of a given type within a local receptive field. These computational models have been successfully applied to handwriting recognition [1, 2], and generic object recognition [4, 5]. Learning features in existing models consists in handcrafting the first layers and training the upper layers by recording templates from the training set, which leads to inefficient representations [4, 5], or in training the entire architecture supervised, which requires large training sets [2, 3]. We propose a fully unsupervised algorithm for learning sparse and locally invariant features at all levels. Each simple-cell layer is composed of multiple convolution filters followed by a winner-take-all competition within a local area, and a sigmoid non-linearity. For training, each simple-cell layer is coupled with a feed-back layer whose role is to reconstruct the input of the simple-cell layer from its output. These coupled layers are trained simultaneously to minimize the average reconstruction error. The output of a simple-cell layer can be seen as a sparse overcomplete representation of its input. The complex cells add the simple cell activities of one filter within the area over which the winner-take-all operation is performed, yielding representations that are invariant to small displacements of the input stimulus. The training procedure is similar to [6], but the local winner-take-all competition ensures that the representation is spatially sparse (and the complex-cell representation locally invariant). The next stage of simple-cell and complex-cell layers is trained in an identical fashion on the outputs of the first layer of complex cells [7], resulting in higher level, more invariant representations, that are then fed to a supervised classifier. Such a procedure yields 0.64% error on MNIST dataset (handwritten digits), and 54% average recognition rate on the Caltech-101 dataset (101 object categories, 30 training samples per category), demonstrating good performance even with few labeled training samples.

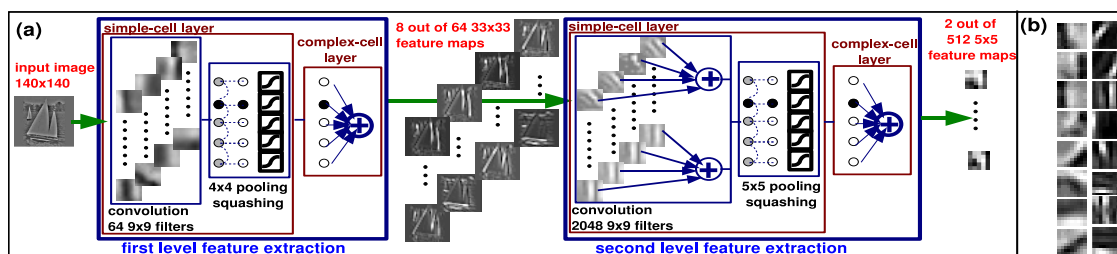


Figure (a) Architecture for recognition. (b) Some learned filters in the 1st and 2nd stages of simple-cell layers.

## References

- [1] Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. K. Fukushima, S. Miyake, Pattern Recognition 1982.
- [2] Gradient-Based Learning Applied to Document Recognition. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, IEEE 1998.
- [3] Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting, Y. LeCun, F.-J. Huang, CVPR 04.
- [4] Object Recognition with Features Inspired by Visual Cortex. T. Serre, L. Wolf, T. Poggio, CVPR 05.
- [5] Multiclass Object Recognition with Sparse, Localized Features. J. Mutch, D. Lowe, CVPR 06.
- [6] Efficient Learning of Sparse Representations with an Energy-Based Model. M. Ranzato, C. Poultney, S. Chopra, Y. LeCun, NIPS 06.
- [7] Reducing the dimensionality of data with neural networks. G.E. Hinton and R.R. Salakhutdinov, Science 06.