Unsupervised Learning of Sparse and Invariant Features Hierarchies

Marc'Aurelio Ranzato[1], **Y-Lan Boureau**[1], **Fu Jie Huang**[1] **and Yann LeCun**[1]

[1] The Courant Institute of Mathematical Sciences - New York University

Unsupervised learning methods are commonly used to produce feature extractors in image analysis systems. A challenging question is whether these methods can learn *invariant hierarchies* of features. This would make much easier the problem of extracting useful information from very high dimensional datasets with few labeled samples, as it is often the case in many object recognition tasks in computer vision.

The feed-forward, multi-stage Hubel and Wiesel architecture $[1, 2, 3, 4, 5]$ stacks multiple levels of alternating convolutional feature detectors, and local pooling of feature maps using some weighted average of units within a neighborhood. These 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 $[2, 3]$. In all these models, invariance is never taken into account while learning the features, but might be achieved after training by using the pooling layers $[6]$.

We propose a fully unsupervised algorithm for learning hierarchies of sparse and locally shift-invariant features. At each level, there are multiple convolution filters followed by a max-pooling layer within a spatially local neighborhood, and a sigmoid non-linearity. Training is performed level by level, separately. At each stage, a single module is coupled with a feed-back layer whose role is to reconstruct the input of the module from its output (see fig. 1 for details). These coupled layers are trained simultaneously to minimize the average reconstruction error. The output of a layer is a sparse overcomplete representation of its input, similarly to $[6]$. However, this representation is also locally shift-invariant thanks to the winner-take-all operation performed by the max-pooling layer. The next stage of convolutional and max-pooling layers is trained in an identical fashion on the outputs of the first layer $[7]$, resulting in higher level, more invariant representations (see for an exmple fig. 2), that are then fed to a supervised classifier. Such a procedure produces features that yield 0.64% error rate 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.

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.

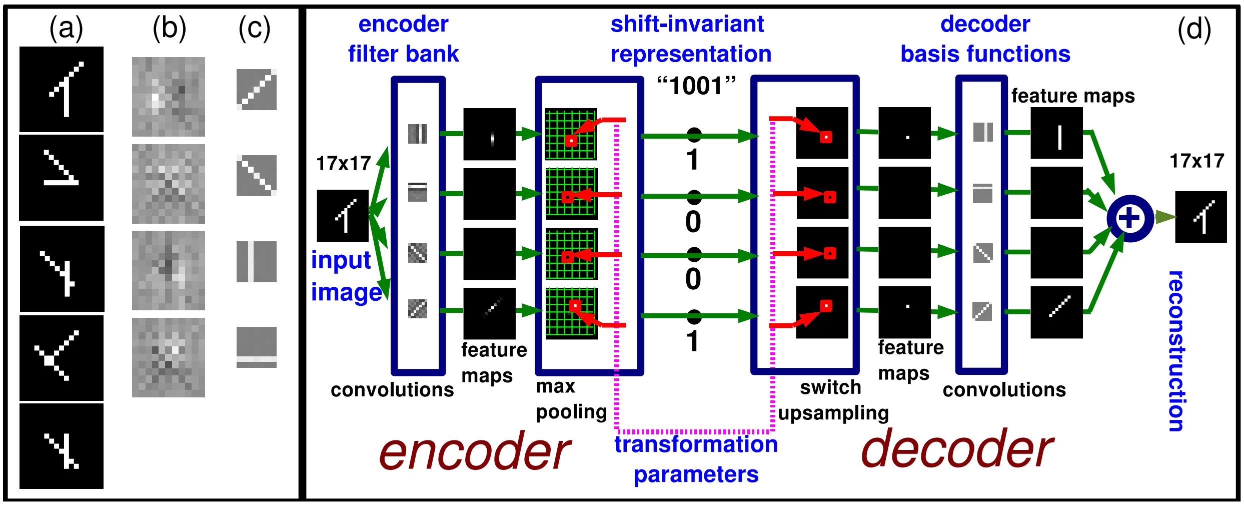

Figure 1: Left Panel: (a) sample images from the a toy dataset. Each sample contains two intersecting segments at random orientations and random positions. (b) Non-invariant features learned by an auto-encoder with 4 hidden units. (c) Shift-invariant decoder filters learned by the proposed algorithm. The algorithm finds the most natural solution to the problem. Right Panel (d): architecture of the shift-invariant unsupervised feature extractor applied to the two bars dataset (just a single module). The encoder convolves the input image with a filter bank and computes the max across each feature map to produce the invariant representation. The decoder produces a reconstruction by taking the invariant feature vector (the "what"), and the transformation parameters (the "where"). Reconstructions is achieved by adding each decoder basis function (identical to encoder filters) at the position indicated by the transformation parameters, and weighted by the corresponding feature component.

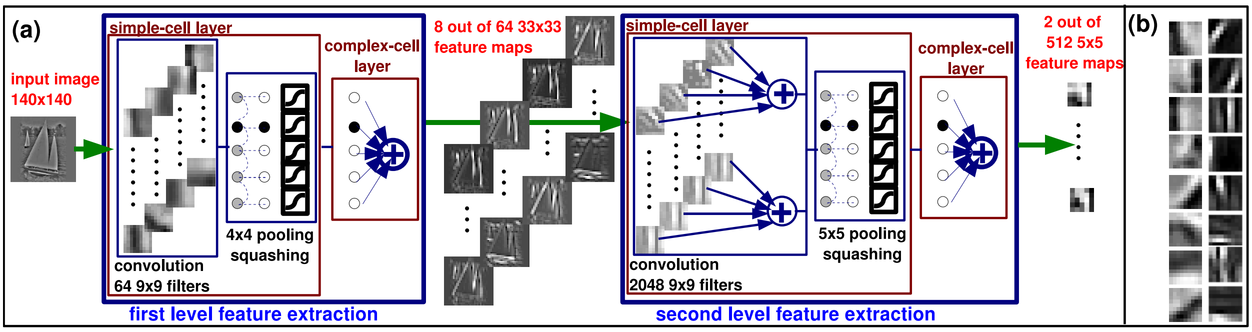

Figure 2: (a) Hierarchical architecture used for recognition. (b) Some filters in the first and second module of the feature extractor learned on the Caltech 101 dataset.