A Theoretical Framework for Measuring the Performance of Deep Belief Networks

PE. Barbano, M.Scoffier, D. Healy, Y. LeCun

January 22, 2007

In many signal processing applications, the the set A, of signals we wish to recover, as well as the interference set B (usually referred to as *"Clutter"*) *are both sparse.* This implies that the resulting set $X = A + B$ of recorded signals must be sparse as well. The Detection/Estimation (DE) problem can be seen as a *classification problem* on this set X.

The work summarized in the present Abstract, illustrates how the *Convolutional Network* (CN, [1]), a successful approach to Supervised Learning, is in fact fundamentally related to the Mathematical Theory of *Compressed Sensing.* We show through a series of experiments that the *Convolutional Network* efficiently approximates the optimal polynomial time algorithm implied in the theory [1]. We present a theoretical foundation for a non probabilistic DE-theory for signals in clutter and argue in favor of the solution of such problems by means of CN-based strategies.

Resting on known results, but only recently starting to attract the efforts of the Scientific Community, *Discrete Compressed Sensing* ([2]) deals with the task of economically recording information about signals, viewed as a elements of a vector space $\mathcal{V} \simeq \mathbf{R}^N$. More specifically, n non-adaptive measurements $\{y_1, \ldots, y_n\} \in \mathbf{R}$ are made of $x \in \mathcal{V}$. Each measurement takes the form of a linear functional $\Phi_k : \mathcal{V} \to \mathbf{R}$ applied to x:

$$\Phi_k(x) \; = \; y_k \tag{1}$$

In general we can write:

$$\Phi \; = \; (\Phi_1, \ldots, \Phi_n)\,, \; n << N \tag{2}$$

So that, with obvious notation:

$$\Phi \; : \mathcal{V} \to \; \mathcal{W}. \tag{3}$$

We can interpret the series of measurements (in a given order) as having made the choice of a *compressive matched filter* ([3]) acting on the space $\mathcal{V}$ to acquire information about the object x.

[1] It has been proven that, under the assumption of sparsity of the sets X, such optimal *compressive matched filters* ([3]) can be found in polynomial time.

Once the information is collected (or *"encoded"*) by means of the matched filters, a non-linear mapping $\Delta_\Phi : W \rightarrow V$ is formed to reconstruct the original object x, and the error is computed with respect to some norm $N(\cdot) = \|\cdot\|$ of choice. Next we define the error:

$$E(x, \Phi, \Delta_\Phi) \; = \; \|\Delta_\Phi\left(\Phi(x)\right) - x\|. \tag{4}$$

The most relevant aspect of the Compressed Sensing framework is that the best possible performance of an encoder-decoder pair can be determined by a simple mathematical bound. In particular, if $X \subset \mathcal{V}$ is a sparse set, the error of Δ_Φ on X is the maximum error of any individual reconstruction of X:

$$E(X, \Phi, \Delta_\Phi) \; = \; sup_x\left(E(x, \Phi, \Delta_\Phi)\right), \tag{5}$$

and the error of the best decoder is :

$$E(X, N, n) \; = \; inf_\Phi\left(E(X, \Phi, \Delta_\Phi)\right). \tag{6}$$

It holds that, under reasonable regularity assumptions for X, the best possible decoder is bounded by:

$$d^n(X) \leq E(X, N, n) \leq C \cdot d^n(X). \tag{7}$$

where $d^n(X)$ is a number depending on the choice of the norm N called the *Gelfand n-width* of the set X^2. In particular, the minimum number M of measurements needed to reconstruct the geometry of X follows the bound:

$$M \; = \; C \cdot K \cdot log\left(\frac{N}{K}\right) \text{([3])} \tag{8}$$

Our works shows that the number of features needed by a CN-classifier follows this growth-law.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[2] David L. Donoho, "Compressed sensing.," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] Michael B. Wakin Mark A. Davenport and Richard G. Baraniuk, "The compressive matched filter," Technical Report TREE 0610, Rice University, Department of Electrical and Computer Engineering, November 2006.

[2] The Gelfand width of $X \subset \mathcal{V}$ is defined as the $inf_Y\{sup(\|x\|) \mid x \in X \cap Y\}$ where Y is a subspace of $\mathcal{V}$ of codimension n