# Analyzing Convolutional Neural Networks Through the Eyes of Sparsity

Vardan Papyan*
Department of Computer Science
Technion - Israel Institute of Technology
Email: vardanp@campus.technion.ac.il

Yaniv Romano*
Department of Electrical Engineering
Technion - Israel Institute of Technology
Email: yromano@tx.technion.ac.il

Michael Elad
Department of Computer Science
Technion - Israel Institute of Technology
Email: elad@cs.technion.ac.il

***Abstract*—It is becoming increasingly difficult to ignore the remarkable results of Convolutional Neural Networks (CNN), and the need for its theoretical analysis. In this work, we aim to alleviate this gap by proposing a novel model – the multi-layer convolutional sparse coding (ML-CSC). This defines a set of signals for which the forward pass of CNN is nothing but a thresholding pursuit. Leveraging this connection, we are able to attribute to the CNN architecture theoretical claims such as uniqueness of the representations (feature maps) throughout the network and their stable estimation, all guaranteed under simple local sparsity conditions. Sitting on these theoretical grounds, we propose a better pursuit that is shown to be theoretically superior to the forward pass.**

## I. Introduction

Deep learning [1], and in particular CNN [2]–[4], has gained a copious amount of attention in recent years as it has led to many state-of-the-art results spanning through many fields. In the core of CNN is the ubiquitous forward pass algorithm, which is a multi-layer scheme that provides an end-to-end mapping, from an input signal to some desired output. Each layer consists of two steps: The first convolves the input $\mathbf{X}$ with a set of learned filters, and the second applies a point wise non-linear function, e.g. ReLU, on the resulting response maps summed with a bias. The output of this layer is then fed into another one, thus forming the multi-layer structure. For two layers, this can be summarized in the following equation

$$\text{ReLU}\Bigg( \mathbf{W}_2^T \ \text{ReLU}\bigg( \mathbf{W}_1^T\mathbf{X} + \mathbf{b}_1 \bigg) + \mathbf{b}_2 \Bigg), \tag{1}$$

where $\mathbf{W}_i$ is the matrix containing the different filters shifted spatially in all locations, and $\mathbf{b}_i$ are the corresponding biases.

A seemingly unrelated paradigm, that has also led to remarkable results, is the sparse representation concept [5]–[9]. When handling natural signals, this model has been commonly used for modeling local patches extracted from the global data mainly due to the computational difficulties related to the task of learning the dictionary [5], [6], [9]–[11]. However, in recent years an alternative to this patch-based processing has emerged in the form of the convolutional sparse coding (CSC) model [12]–[16]. Indeed, the convolutional extension was extensively analyzed in a recent work [17], [18], shedding light on its theoretical aspects and prospects of success. Interestingly, while the CSC is a global model, its analysis relied on local properties such as (i) the $\|\cdot\|_{0,\infty}^{\mathbf{S}}$ norm, which is defined as the maximal number of non-zeros in a local patch representation extracted from a global sparse vector, and (ii) the $\|\cdot\|_{2,\infty}^{\mathbf{P}}$ norm that measures the maximal $\ell_2$ norm of a patch extracted from a global signal.

## II. From Atoms to Molecules: Multi-Layer Convolutional Sparse Model

Convolutional sparsity assumes an inherent structure for natural signals. Similarly, the representations themselves could also be assumed to have such a structure. In what follows, we propose a novel layered model that relies on this rationale.

*Definition 1:* For a global signal $\mathbf{X}$, a set of convolutional dictionaries $\{\mathbf{D}_i\}_{i=1}^K$, and a vector $\boldsymbol{\lambda}$, define the deep coding problem DCP$_{\boldsymbol{\lambda}}$ as:

$$\begin{aligned}(\text{DCP}_{\boldsymbol{\lambda}}): \quad \text{find} \quad \{\mathbf{\Gamma}_i\}_{i=1}^K \quad \text{s.t.} \quad \mathbf{X} &= \mathbf{D}_1\mathbf{\Gamma}_1, & \|\mathbf{\Gamma}_1\|_{0,\infty}^{\mathbf{S}} &\leq \lambda_1 \\ \mathbf{\Gamma}_1 &= \mathbf{D}_2\mathbf{\Gamma}_2, & \|\mathbf{\Gamma}_2\|_{0,\infty}^{\mathbf{S}} &\leq \lambda_2 \\ &\vdots & &\vdots \\ \mathbf{\Gamma}_{K-1} &= \mathbf{D}_K\mathbf{\Gamma}_K, & \|\mathbf{\Gamma}_K\|_{0,\infty}^{\mathbf{S}} &\leq \lambda_K,\end{aligned}$$

where the scalar $\lambda_i$ is the $i$-th entry of $\boldsymbol{\lambda}$.
Intuitively, given a signal $\mathbf{X}$, this problem seeks for a set of representations, $\{\mathbf{\Gamma}_i\}_{i=1}^K$, such that each one is locally sparse.

Assume we are given a signal $\mathbf{X}$ and our goal is to find its underlying representations, $\{\mathbf{\Gamma}_i\}_{i=1}^K$. Tackling this problem by recovering all the vectors at once might be computationally and conceptually challenging; therefore, we propose the *layered thresholding algorithm* that gradually computes the sparse vectors one at a time across the different layers. Denoting by $\mathcal{S}_\beta^+(\cdot)$ the soft nonnegative thresholding operator with a threshold $\beta$; we commence by computing $\hat{\mathbf{\Gamma}}_1 = \mathcal{S}_{\beta_1}^+(\mathbf{D}_1^T\mathbf{X})$, which is an approximation of $\mathbf{\Gamma}_1$. Next, by applying another thresholding algorithm, however this time on $\hat{\mathbf{\Gamma}}_1$, an approximation of $\mathbf{\Gamma}_2$ is obtained, $\hat{\mathbf{\Gamma}}_2 = \mathcal{S}_{\beta_2}^+(\mathbf{D}_2^T\hat{\mathbf{\Gamma}}_1)$. This process is iterated until the last representation $\hat{\mathbf{\Gamma}}_K$ is acquired.

Assuming two layers for simplicity, the layer thresholding algorithm can be summarized as follows

$$\hat{\mathbf{\Gamma}}_2 = \mathcal{S}_{\beta_2}^+\Bigg( \mathbf{D}_2^T \ \mathcal{S}_{\beta_1}^+\bigg( \mathbf{D}_1^T\mathbf{X} \bigg) \Bigg).$$

Comparing the above with Equation (1), we conclude that *the aforementioned pursuit and the forward pass of the CNN are equal*! Building on this observation, we present in this work a line of theorems providing theoretical insights for CNN in the view of sparsity. Next, we present one of these.

*Theorem 2:* (Stability of the layered soft thresholding): Suppose a clean signal $\mathbf{X}$ has a decomposition

$$\mathbf{X} = \mathbf{D}_1\mathbf{\Gamma}_1, \ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2, \ \cdots, \ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K,$$

and that it is contaminated with noise $\mathbf{E}$ to create the signal $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. Denote by $\mu(\mathbf{D}_i)$ the mutual coherence of the convolutional dictionary $\mathbf{D}_i$, and by $|\Gamma_i^{\min}|$ and $|\Gamma_i^{\max}|$ the lowest and highest entries in absolute value in the vector $\mathbf{\Gamma}_i$, respectively. Let $\{\hat{\mathbf{\Gamma}}_i\}_{i=1}^K$ be the set of solutions obtained by running the layered soft thresholding algorithm with thresholds $\{\beta_i\}_{i=1}^K$, i.e. $\hat{\mathbf{\Gamma}}_i = \mathcal{S}_{\beta_i}(\mathbf{D}_i^T\hat{\mathbf{\Gamma}}_{i-1})$ where $\hat{\mathbf{\Gamma}}_0 = \mathbf{Y}$. Assuming that $\forall\ 1 \leq i \leq K$

a) $\|\mathbf{\Gamma}_i\|_{0,\infty}^{\mathbf{S}} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)}\frac{|\Gamma_i^{\min}|}{|\Gamma_i^{\max}|}\right) - \frac{1}{\mu(\mathbf{D}_i)}\frac{\epsilon_{i-1}}{|\Gamma_i^{\max}|}$; and
b) The threshold $\beta_i$ is proportional to $\epsilon_i$ (defined below),

then

1) The support of the solution $\hat{\mathbf{\Gamma}}_i$ is equal to that of $\mathbf{\Gamma}_i$; and
2) $\|\mathbf{\Gamma}_i - \hat{\mathbf{\Gamma}}_i\|_{2,\infty}^{\mathbf{P}} \leq \epsilon_i$,

where $\epsilon_i = \sqrt{\|\mathbf{\Gamma}_i\|_{0,\infty}^{\mathbf{P}}}\left(\epsilon_{i-1} + \mu(\mathbf{D}_i)\left(\|\mathbf{\Gamma}_i\|_{0,\infty}^{\mathbf{S}} - 1\right)|\Gamma_i^{\max}| + \beta_i\right)$ for $i > 0$, and $\epsilon_0 = \|\mathbf{E}\|_{2,\infty}^{\mathbf{P}}$.

*The authors contributed equally to this work.

The research leading to these results has received funding from the European Research Council under European Union's Seventh Framework Programme, ERC Grant agreement no. 320649.

# References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] B. B. LeCun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*. Citeseer, 1990.

[3] 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, 1998.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries." *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[6] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. on Image Process.*, vol. 20, no. 7, pp. 1838–1857, 2011.

[7] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.

[8] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 1697–1704.

[9] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *arXiv preprint arXiv:1411.3230*, 2014.

[10] Y. Romano and M. Elad, "Patch-disagreement as a way to improve K-SVD denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1280–1284.

[11] J. Sulam and M. Elad, "Expected patch log likelihood with a sparse prior," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2015, pp. 99–111.

[12] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 391–398.

[13] B. Kong and C. C. Fowlkes, "Fast convolutional sparse coding (fcsc)," *Department of Computer Science, University of California, Irvine, Tech. Rep*, 2014.

[14] B. Wohlberg, "Efficient convolutional sparse coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7173–7177.

[15] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1823–1831.

[16] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 5135–5143.

[17] V. Papyan, J. Sulam, and M. Elad, "Working locally thinking globally-part I: Theoretical guarantees for convolutional sparse coding," *arXiv preprint arXiv:1607.02005*, 2016.

[18] ——, "Working locally thinking globally-part II: Stability and algorithms for convolutional sparse coding," *arXiv preprint arXiv:1607.02009*, 2016.