A Deep Analysis Dictionary Model

Jun-Jie Huang and Pier Luigi Dragotti
Department of Electrical and Electronic Engineering, Imperial College London, UK

I. Introduction

Deep Neural Networks (DNNs) are computational models which are composed of multiple layers of linear transforms and point-wise non-linearities. Back-propagation algorithm [1] is usually applied to optimize this highly non-linear and non-convex system. However, a clear understanding of the functioning of the linear transforms and the non-linearites is still lacking. Contrary to DNNs, the sparse representation theory [2] is much more established. Therefore building a deep model [3]–[6] using sparse representation and redundant dictionaries could be a way to facilitate the interpretation of DNNs.

II. Proposed Deep Dictionary Model

We propose a deep dictionary model [7], [8] for regression tasks. The proposed L-layer deep dictionary model is composed of $L-1$ layers of analysis dictionary and soft-thresholding pairs $\{\boldsymbol{\Omega}_i \in \mathbb{R}^{d_i \times d_{i-1}}, \boldsymbol{\lambda}_i \in \mathbb{R}^{d_i}\}_{i=1}^{L-1}$ and a synthesis dictionary $\boldsymbol{D} \in \mathbb{R}^{d_L \times d_{L-1}}$. We assume that the analysis dictionaries are overcomplete, that is $d_{i-1} < d_i$ for $1 \leq i \leq L-1$. Therefore the forward model can be expressed as:

$$\boldsymbol{y} = \boldsymbol{D}\mathcal{S}_{\boldsymbol{\lambda}_{L-1}}(\boldsymbol{\Omega}_{L-1}\mathcal{S}_{\boldsymbol{\lambda}_{L-2}}(\cdots\boldsymbol{\Omega}_2\mathcal{S}_{\boldsymbol{\lambda}_1}(\boldsymbol{\Omega}_1\boldsymbol{x})\cdots)), \qquad (1)$$

where $\boldsymbol{x} \in \mathbb{R}^{d_0}$ and $\boldsymbol{y} \in \mathbb{R}^{d_L}$ are the vectorized input and estimated signal, respectively, and $\mathcal{S}_{\boldsymbol{\lambda}}(\cdot)$ is the soft-thresholding operator.

We note that the soft-thresholding thresholds can not all be too large. Let us consider a single layer model $\boldsymbol{y} = \boldsymbol{D}\mathcal{S}_{\boldsymbol{\lambda}_1}(\boldsymbol{\Omega}_1\boldsymbol{x})$, if all thresholds are sufficiently large (i.e. $\lambda_j > \delta$ for $j \in [1, d_1]$), there exists a convex polyhedron in the feature space in which non-zero input $\boldsymbol{x}$ has $\mathcal{S}_{\boldsymbol{\lambda}_1}(\boldsymbol{\Omega}_1\boldsymbol{x}) = \boldsymbol{0}$ and thus $\boldsymbol{y} = \boldsymbol{0}$. The information within $\boldsymbol{x}$ will then be completely lost. There should be at least k analysis-thresholding pairs for information preserving if the input data spans a k dimensional subspace of the output data.

In order not to lose essential information, each analysis dictionary $\boldsymbol{\Omega}_i$ is designed to consist of two sub-dictionaries $\boldsymbol{\Omega}_i = [\boldsymbol{\Omega}_{\mathrm{I}i}; \boldsymbol{\Omega}_{\mathrm{C}i}]$. The information preserving analysis dictionary (IPAD) $\boldsymbol{\Omega}_{\mathrm{I}i}$ aims at passing key information from its previous layer and is associated with relatively small thresholds $\boldsymbol{\lambda}_{\mathrm{I}i}$. The clustering analysis dictionary (CAD) $\boldsymbol{\Omega}_{\mathrm{C}i}$ with its thresholds $\boldsymbol{\lambda}_{\mathrm{C}i}$ is to facilitate the separation of key feature in the signal. The threshold $\boldsymbol{\lambda}_{\mathrm{C}i}$ can be relatively large. Fig. 1 shows the analysis and the thresholding operations at layer i.

The i^{th} layer IPAD and threshold pair $(\boldsymbol{\Omega}_{\mathrm{I}i}, \boldsymbol{\lambda}_{\mathrm{I}i})$ passes the key information from layer $i-1$ to layer $i+1$. IPAD $\boldsymbol{\Omega}_{\mathrm{I}i}$ is obtained by applying an extension of the geometric analysis operator learning method [7], [9] with its input training data $\boldsymbol{X}_i$. The learned $\boldsymbol{\Omega}_{\mathrm{I}i}$ is able to sparsify while preserve information from its input data. As $\boldsymbol{\Omega}_{\mathrm{I}i}\boldsymbol{X}_i$ can be well characterized by an i.i.d. zero-mean Laplacian distribution, the threshold $\boldsymbol{\lambda}_{\mathrm{I}i}$ is set to be proportional to the inverse of the standard deviation $\boldsymbol{\sigma}$ of the Laplacian distribution. The soft-thresholding $\mathcal{S}_{\boldsymbol{\lambda}_i}(\boldsymbol{\Omega}_i\boldsymbol{x})$ can be interpreted as a denoising operation.

The CAD and threshold pair $(\boldsymbol{\Omega}_{\mathrm{C}i}, \boldsymbol{\lambda}_{\mathrm{C}i})$ is to generate discriminative features for the representation of its input. The hope is that each pair of atom and threshold $(\boldsymbol{\omega}_{\mathrm{C}ij}, \lambda_{\mathrm{C}ij})$ can identify a cluster of data in its input feature space and thus facilitate prediction. The key idea of learning $(\boldsymbol{\Omega}_{\mathrm{C}i}, \boldsymbol{\lambda}_{\mathrm{C}i})$ is to model its input training data as a mixture of Gaussians [10] and learn $(\boldsymbol{\omega}_{\mathrm{C}ij}, \lambda_{\mathrm{C}ij})$ pairs which can identify data belonging to different Gaussian models. With mixture of Gaussian modelling, the training data has been clustered with labels. The learning objective of an atom $\boldsymbol{\omega}_{\mathrm{C}ij}$ is formulated as maximizing the absolute value of the inner product between it and the data belonging to one Gaussian model while minimizing the absolute value of the inner product between it and the data belonging to other Gaussian models. Fig. 2 shows the histogram of the absolute value of the inner product with a learned atom for two groups of data. The threshold $\lambda_{\mathrm{C}ij}$ is set to achieve the maximum separation between these two groups of data.

The multi-layer analysis dictionary and threshold pairs $\{(\boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i)\}_{i=1}^{L-1}$ are learned in a greedy manner. Although the decision boundary of analysis and thresholding at each layer is linear, it will introduce non-linear decision boundaries if there are more than one layer and leads to discriminative representation. The synthesis dictionary $\boldsymbol{D}$ is obtained using least squares.

III. Simulation Results

In this section, we report the simulation results of our proposed method against alternative approaches. For image super-resolution task, the standard 91 training images [11] are applied for training and *Set14* [12] is used for evaluation. The up-scaling factor is set to 2. The low-resolution and high-resolution patch size is 3×3. The input low-resolution feature is the raw pixel values with removed mean. For comparison, DNNs with the same structure are learned using the same training data. Let us denote DNN-R and DNN-S as the DNN with ReLU non-linearity and soft-thresholding non-linearity, respectively. The implementation is based on Pytorch with Adam optimizer, batch size 256, initial learning rate 0.01, learning rate decay step 100, and decay rate 0.1. The total number of epoch for training is 500.

Table I reports the PSNR (dB) of different methods. Our proposed DDM method achieves a similar average PSNR when compared to the DNN-R method. This validates the effectiveness of our proposed deep dictionary model and shows that the simultaneous information preserving and clustering model could be a good interpretation of the workings of DNNs. The DNN-S method achieves the highest average PSNR which is around 0.2 dB higher than that of the DDN-R method and our proposed DDM method. This suggests that DNNs with soft-thresholding as non-linearity is more effective for image enhancement applications. The result also indicates that our DDM method can be further improved. In particular, an optimization strategy needs to be devised to determine the ratio between the number of information preserving atoms and the number of clustering atoms.

IV. Conclusion

In this paper, we proposed a novel method to learn a pair of analysis dictionary and soft-threshold which is used to construct the deep dictionary model for regression tasks. The learned analysis dictionaries together with the corresponding soft-thresholds can simultaneously preserve important information from the previous layer as well as facilitate discrimination of key features. Simulation results show that our proposed deep dictionary model achieves comparable performance with DNNs.

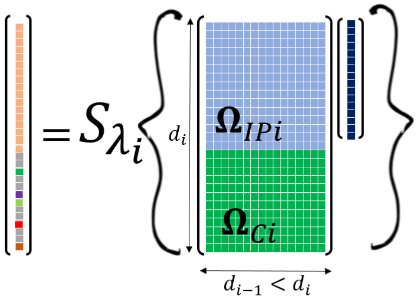

Fig. 1. The layer i of the proposed deep dictionary architecture which consists of an analysis dictionary followed by soft-thresholding. The analysis dictionary $\boldsymbol{\Omega}_i$ consists of an information preserving dictionary $\boldsymbol{\Omega}_{\mathrm{I}i}$ and a clustering dictionary $\boldsymbol{\Omega}_{\mathrm{C}i}$. The soft-thresholds corresponding to $\boldsymbol{\Omega}_{\mathrm{C}i}$ are much higher than those used for $\boldsymbol{\Omega}_{\mathrm{I}i}$.

Image	Bicubic	DNN-R	DNN-S	DDM
baboon	24.86	25.46	25.48	25.42
barbara	27.88	28.41	28.41	28.43
bridge	26.62	27.37	27.45	27.40
coastguard	29.26	30.17	30.21	30.17
comic	24.63	27.28	27.45	27.19
face	34.73	35.33	35.42	35.37
flowers	30.20	31.72	31.97	31.73
foreman	35.21	37.36	38.11	37.56
lenna	34.57	35.87	36.04	35.86
man	29.16	30.16	30.29	30.15
monarch	32.77	35.12	35.67	35.25
pepper	34.98	36.23	36.50	36.28
ppt3	24.66	28.31	28.47	28.12
zebra	28.03	32.61	32.84	32.59
Average	29.83	31.53	31.74	31.54

TABLE I

PSNR (dB) by different methods evaluated on Set 14 [12].

References

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[2] M. Elad, “Sparse and redundant representations: From theory to applications in signal and image processing,” 2010.

[3] V. Papyan, Y. Romano, and M. Elad, “Convolutional neural networks analyzed via convolutional sparse coding,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2887–2938, 2017.

[4] J. Sulam, V. Papyan, Y. Romano, and M. Elad, “Multi-layer convolutional sparse modeling: Pursuit and dictionary learning,” *arXiv preprint arXiv:1708.08705*, 2017.

[5] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa, “Deep dictionary learning,” *IEEE Access*, vol. 4, pp. 10 096–10 109, 2016.

[6] S. Mahdizadehaghdam, A. Panahi, H. Krim, and L. Dai, “Deep dictionary learning: A parametric network approach,” *arXiv preprint arXiv:1803.04022*, 2018.

[7] J.-J. Huang and P. L. Dragotti, “A deep dictionary model for image super-resolution,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’18)*, March 2018.

[8] ——, “A deep dictionary model to preserve and disentangle key features in a signal,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’19)*, May 2019.

[9] S. Hawe, M. Kleinsteuber, and K. Diepold, “Analysis operator learning and its application to image reconstruction,” *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2138–2150, 2013.

[10] K. P. Murphy, “Machine learning: a probabilistic perspective,” 2012.

[11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

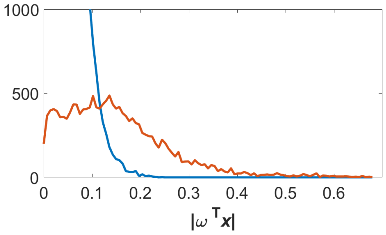

Fig. 2. The histogram of $|\boldsymbol{\omega}^T\boldsymbol{x}|$ for the data from a Gaussian (orange) and from other Gaussians (blue).

[12] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.