HYPER-PARAMETER SELECTION ON CONVOLUTIONAL DICTIONARY LEARNING THROUGH LOCAL $\ell_{0,\infty}$ NORM

Gustavo Silva *Jorge Quesada* *Paul Rodriguez*

Electrical Engineering Department, Pontificia Universidad Católica del Perú, Lima, Peru
Email: {gustavo.silva, jorge.quesada, prodrig}@pucp.edu.pe

ABSTRACT

Convolutional dictionary learning (CDL) is a widely used technique in many applications on the signal/image processing and computer vision fields. While many algorithms have been proposed in order to improve the computational run-time performance during the training process, a thorough analysis regarding the direct relationship between the reconstruction performance and the dictionary features (hyper-parameters), such as the filter size and filter bank's cardinality, has not yet been presented.

As arbitrarily configured dictionaries do not necessarily guarantee the best possible results during the test process, a correct selection of the hyper-parameters would be very favorable in the training and testing stages. In this context, this works aims to provide an empirical support for the choice of hyper-parameters when learning convolutional dictionaries. We perform a careful analysis of the effect of varying the dictionary's hyper-parameters through a denoising task. Furthermore, we employ a recently proposed local $\ell_{0,\infty}$ norm as a sparsity measure in order to explore possible correlations between the sparsity induced by the learned filter bank and the reconstruction quality at test stage.

Index Terms— Convolutional sparse representation, convolutional dictionary learning, hyper-parameters

1. INTRODUCTION

Convolutional sparse representation (CSR) techniques have been shown to provide state-of-the-art results in a wide variety of signal and image processing and machine learning applications [1] . The CSR model consists in representing a given image as a sum over a set of convolutions between a set of dictionary filters $\mathbf{d}_k$ and their corresponding feature maps $\mathbf{x}_k$. The standard form of the Convolutional Sparse Coding (CSC) problem is given by the Convolutional Basis Pursuit Denoising (CBPDN) objective, namely[1]:

$$\underset{\{\mathbf{x}_k\}}{\arg\min} \frac{1}{2}\Big\| \sum_k \mathbf{d}_k * \mathbf{x}_k - \mathbf{s}\Big\|_2^2 + \lambda \sum_k \Big\|\mathbf{x}_k\Big\|_1 . \qquad (1)$$

Particularly, the presence of convolutions and sparsifying operations in this model has been commonly linked (both heuristically and theoretically) to the layers of Convolutional Neural Networks (CNN) [2]. These two fields have frequently borrowed ideas and tools from the other, cross-fertilizing the development of new approaches in both.

A natural observation in the CSR model is that the dictionary characteristics (such as cardinality and filter size) are of paramount importance for the quality of the representation, just as an appropriate layer choice is essential for a competitive CNN structure. However, while some studies have addressed the choice for filter bank (FB)'s cardinality and size of filter in convolutional layers of CNNs [3, 4], to the best of our knowledge there are currently no works in the field of CSR that attempt to find the optimal configuration for these values. Moreover, via the theoretical analysis of a local $\ell_{0,\infty}$ norm penalized version of (1), [5] was able to provide meaningful guarantees for the success of popular ℓ_1-norm penalized CSC algorithms, such as those based on the Alternating Direction Method of Multipliers (ADMM) [6] and Accelerated Proximal Gradient (APG) [7] frameworks.

In this context, the objective of this paper is to present a thorough evaluation on the effect of the choice of the mentioned hyper-parameters when performing convolutional dictionary learning (CDL). Our results favor the conclusion that the local $\ell_{0,\infty}$ norm of the feature maps obtained during training stage is closely related to the reconstruction performance (in terms of PSNR) during a denoising task, leading to a lower bound on the FB's cardinality. Furthermore, the experimental results also show that a possible lower bound for filter size can be 24×24, yielding results comparable to the optimal value.

The rest of this paper is organized as follows: Section 2 reviews technical details of the CDL problem and the local mixed $\ell_{0,\infty}$ norm. In Section 3, we report previous information about parameters and hyper-parameters selection. In Section 4, we present a thorough description and analysis of our experiments and results, respectively. In Section 5, we give our final remarks.

2. PREVIOUS RELATED WORK

Our primary interest lies in the branch of CSR that deals with estimating the optimal dictionary filters for a given image training set, termed Convolutional Dictionary Learning (CDL) and represented by the problem:

$$\underset{\{\mathbf{x}_{r,k}\}\{\mathbf{d}_k\}}{\arg\min} \frac{1}{2}\sum_r \Big\| \sum_k \mathbf{d}_k * \mathbf{x}_{r,k} - \mathbf{s}_r\Big\|_2^2 + \lambda \sum_r \sum_k \Big\|\mathbf{x}_{r,k}\Big\|_1 \quad \text{s.t.} \quad \|\mathbf{d}_k\|_2 = 1 \quad \forall k, \qquad (2)$$

where $\{\mathbf{x}_{r,k}\}$ represents the R sets of K feature maps (each one with $N_1 \times N_2$ samples), $\{\mathbf{d}_k\}$ a set of K, $L_1 \times L_2$ dictionary filters,

[1] Since the variables in (1) correspond to 2D signals (images), it is intuitive to calculate the fidelity term as a Frobenious norm. However, one can cast these variables as 1D signals without losing generality and treat them with the ℓ_2 norm, as it is done across the CSR literature.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

$\{s_r\}$ the R training images of size $N_1 \times N_2$, and λ denotes the regularization parameter. The norm constraint on the filter set is required to avoid scaling ambiguities.

Problem (2) has a non-convex geometry, which is usually minimized by alternating updates between two convex sub-problems: the feature update ($\{\mathbf{x}_{k,m}\}$) and the dictionary update ($\{\mathbf{d}_m\}$). In the literature, the latter has been most studied due to the high complexity related to the training set sizes.

The dictionary update sub-problem for (2) can be constructed by using the simplification $\sum_{k=1}^{K} \mathbf{x}_{r,k} * \mathbf{d}_k = X_r d$, which results in a convolutional variant of the Method of Optimal Directions (MOD) [8] that can be written as

$$\underset{\{\mathbf{d}\}}{\arg\min} \frac{1}{2} \sum_r \left\| \mathbf{X}_r * \mathbf{d} - \mathbf{s}_r \right\|_2^2 \quad \text{s.t} \quad \mathbf{d} \in C_{PN} \ . \tag{3}$$

where $C_{PN} = \{\mathbf{x} \in \mathbb{R}^N : (I - PP^T)\mathbf{x} = 0, \|\mathbf{x}\|_2 = 1\}$ is the constraint set for an adequate spatial support and normalized dictionary filters, and P represents the zero-padding projection operator.

Earlier algorithms [9],[10],[11] used to efficiently solve the CDL problem are based on ADMM approaches. Their main framework consists on handling the aforementioned problem in a transform plane such as the frequency domain, where the convolutions in the ℓ_2 fidelity term are replaced by simple Hadamard products. A more recent branch of research has focused on gradient descent (GD) formulations [12],[13],[14] which significantly improve the computational complexity and memory footprint with respect to their ADMM-based counterparts. One of the fastest GD methods is the accelerated proximal consensus approach [14], in which the solution is decoupled across the index of training images.

An alternative CDL stream is aimed at learning separable filter banks (instead of usual non-separable ones) in order to reduce the implicit cost of the convolution operations. [15] proposed to estimate the separable filters through independent ADMM formulations for vertical and horizontal filters. A more computationally efficient separable algorithm was proposed in [16], based on the Accelerated Proximal Gradient (APG) non-separable algorithm plus an additional rank-1 constraint. Recently, a combinatorial learning approach was proposed in [17] to exploit redundancy properties in the separable FB (using all possible combinations of the horizontal and vertical sets).

2.1. Accelerated Proximal Gradient Consensus

The consensus approach [6],[18] is a well-known strategy to decouple a problem through independent local variables, by imposing an equality constraint. Using a particular constrain set C_C, [14] proposed to rewrite the dictionary sub-problem (3) in a consensus-compatible form as

$$\underset{\{\mathbf{d}_r\}}{\arg\min} \frac{1}{2} \sum_r \left\| \mathbf{X}_r \mathbf{d}_r - \mathbf{s}_r \right\|_2^2 + \sum_r \iota_{C_{\text{PN}}}(\mathbf{d}_r) + \iota_{C_C}(\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_R) \ , \tag{4}$$

where $C_C = \{(\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_R) | \mathbf{d}_1 = \mathbf{d}_2 = \cdots = \mathbf{d}_R\}$ represents the constraint set that enforces equality among the local dictionaries, and ι_{C_C} and $\iota_{C_{PN}}$ are the indicator functions of the constraint sets C_C and C_{PN}, respectively. The proximal gradient derivation of (4) is given by

$$\mathbf{h}_r^{(i+1)} = \mathbf{d}_r^{(i)} - \alpha \nabla F_r(\mathbf{d}_r^{(i)}) \ , \tag{5}$$

$$\mathbf{g}^{(i+1)} = \text{Prox}_{\iota_{C_{\text{PN}}}} \left(\frac{1}{r} \sum_r \mathbf{h}_r^{(i+1)} \right) \ . \tag{6}$$

where g is a global consensus dictionary and α the step size. Likewise, computationally demanding components of the algorithm such as the gradient estimation of the ℓ_2 fidelity term are performed in the frequency domain.

2.2. Local mixed norms

The uniqueness of the solution and success of pursuit algorithms for the CSC (1) was analyzed in [5] via a reformulation based on the local $\ell_{0,\infty}$ pseudo norm, i.e.

$$(P_{0,\infty}): \qquad \min_{\mathbf{X}} ||\mathbf{X}||_{0,\infty} \qquad \text{s.t.} \qquad \mathbf{DX} = \mathbf{S}, \tag{7}$$

where local $\ell_{0,\infty}$ norm is defined in (8). The motivation for this metric is rooted on the theoretical analysis of the underlying solution for global ℓ_0 or ℓ_1 norms, in which X can have a moderate number of non-zero coefficients, might not be unique due to the sparsity condition (global number of non-zeros must be less than $\frac{1}{2}(1 + \frac{1}{\mu(D)})$, where $\mu(D)$ quantifies the mutual coherence of the dictionary [19]). [5] introduced the notion of a local measure of sparsity via the $\ell_{0,\infty}$ norm and corresponding problem $P_{0,\infty}$ applied to the global feature vector $\mathbf{X}$, namely

$$||\mathbf{X}||_{0,\infty} = \max_{i,j,r} ||\gamma_{i,j}^r||_0 \tag{8}$$

where $\gamma_{i,j}^r$ is a patch in the feature vector $\mathbf{X}$ that groups the elements that contribute to a specific region of the reconstructed 1D signal[2] $\hat{\mathbf{s}}_r = \mathbf{X}_r * \mathbf{d}$. This concept can be naturally extended for dealing with 2D images as depicted in Figure 1, where $\gamma_{i,j}^r$ is the portion of each feature map $\mathbf{x}_{r,k}$ that contributes to a given image patch $\beta_{i,j}^r$.

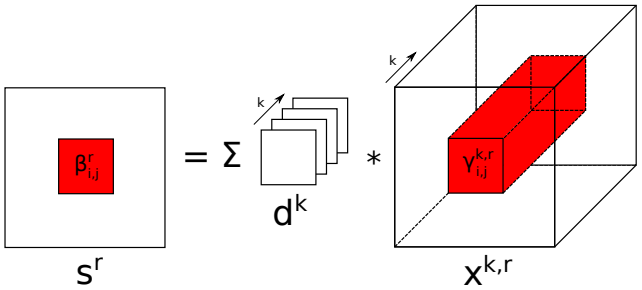

Fig. 1: The {i-th, j-th} patch $\beta_{i,j}^r$ of the global system $S^r = DX^r$, given by $\beta_{i,j}^r = D\gamma_{i,j}^r$.

In our experiments, we use the local $\ell_{0,\infty}$ norm (of the feature maps $X_{k,r}$ obtained during the training stage) as a proxy to evaluate the quality of given learn convolutional dictionary. Our results show that there exists a high correlation between this local measure of sparsity and the reconstruction performance of the associated dictionaries in a denoising task.

3. PARAMETERS AND HYPER-PARAMETERS

The CDL problem (2) only presents one parameter to configure, namely λ which controls the sparsity level of the feature maps. In the literature, this value goes from 0.05 to 0.2. However, depending on the solution method of each CDL sub-problem, which are most

[2]For a more thorough explanation on the structure of the $\gamma_{i,j}^r$ patch for 1D signals, see [5]

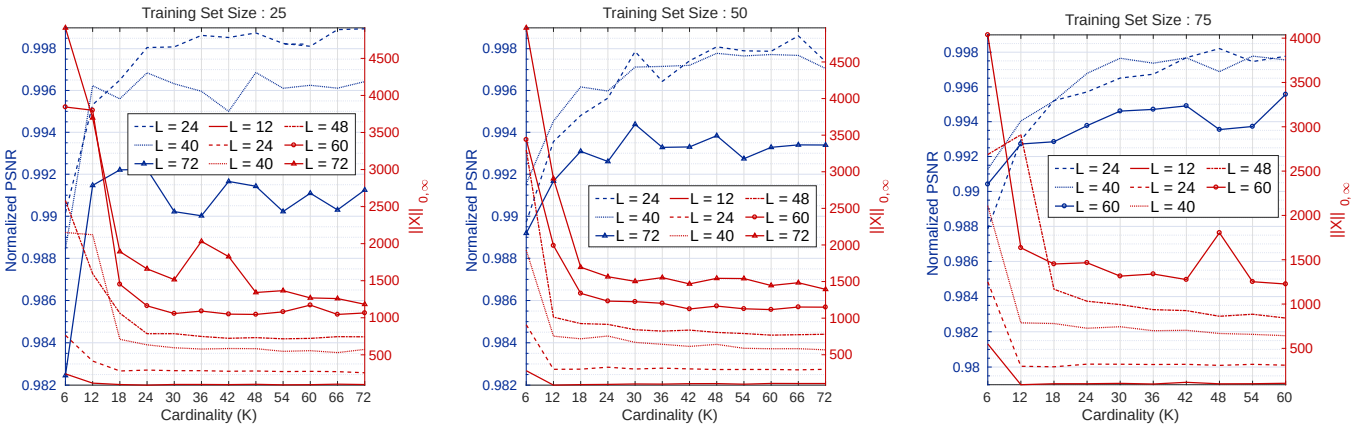

Fig. 2: From left to right using training set sizes of 25, 50 and 75 images. In blue lines, we plot the average normalized PSNR for some values of filter size on a denoising task (only three L values are selected in order to avoid cluttering in the graphs), whereas in red lines we plot the corresponding $\ell_{0,\infty}$ norm of the feature maps (obtained once the training stage is finished) for several more values of L when varying the FB's cardinality on the training stage.

commonly the ADMM and APG frameworks, additional parameters that impact the convergence ratio come into consideration.

In the ADMM case, [9] empirically established a relation for the penalty parameters of feature and dictionary sub-problems with respect to λ and the training set size, respectively. Furthermore, most of the works [11],[13],[14] selected the optimal parameters via grid search. In the APG case, on the other hand, [12] introduced a closed form solution for the optimal step size α at each iteration as a function of the fidelity term and the feature maps. All these approaches have addressed the search for the optimal algorithm parameters corresponding to each framework's domain; however, to the best of our knowledge, no works have attempted to study the impact of hyper-parameters such as filter size or FB's cardinality on the performance of the resulting dictionary filters in the context of convolutional sparse coding.

4. RESULTS

The experiments consisted in evaluating the performance of different learned dictionaries in terms of the normalized PSNR metric, described in Section 4.1, for the denoising task. These experiments were carried out on a standard desktop computer equipped with an Intel i7-7700K CPU (4.20 GHz, 8MB Cache, 32GB RAM).

• **Learning stage:** Non-separable dictionaries, with different filter sizes $\{L \times L\colon\ 8 \leq L \leq 72\}$ and FB's cardinalities $\{K\ :\ 6 \leq K \leq 72\}$, were estimated using three training set sizes, sparsity parameter λ = 0.1 and 500 iterations. The training sets used for these experiments contained batches of 25, 50 and 75 gray-scale images of size 256×256 pixels, cropped and re-scaled from a set of images obtained from the MIRFFLICKR-IM dataset [20]. For learning non-separable dictionary filters, we used the APG consensus based algorithm [14] which is publicly available in [21].

• **Testing stage:** We used eight standard images (such as Mandrill, Peppers, Barbara, etc.) that were corrupted with AWGN with variance $\sigma^2 = 0.04$. The denoising algorithm corresponds to the ADMM-based MATLAB code of the SPORCO library available in [22]. In order to ensure a fair comparison, since the testing algorithm has a sparsity-regularizing parameter λ, a search grid over $\lambda \in [0.001\ ,\ 0.95]$ was used to find the optimal value that provides the best PSNR for each learned dictionary. Furthermore, we performed 5 realization per evaluated case, and averaged the scores to obtain the final PSNR.

4.1. Evaluated metric

In the denoising process, the computed PSNR values differ in scale among the evaluated images, as can be observed in Table 1. The PSNR values estimated using Mandrill image range from 20.76 dB to 21.2 dB, while using the Pepper image the values are a little higher, ranging from 21.41 dB to 25.41 dB approximately. In order to provide a global metric across the test images, we normalized the PSNR by dividing the resulting values obtained from different filter sizes and FB's cardinalities by the maximum one attained for each image. This new metric is represented as

$$\text{Normalized PSNR}(i,l,k) = \frac{\text{PSNR}(i,l,k)}{\max\{\text{PSNR}(i,:,:)\}} \quad , \qquad (9)$$

where (i,l,k) are the image, filter size and candinality index.

Furthermore, we use the local mixed $\ell_{0,\infty}$ norm of the feature maps obtained once the training stage is finished, detailed in Section 2.2 as a measure of sparsity, since it can naturally provide a more meaningful insight of the sparse structure of the reconstructed result than standard global norms.

		PSNR(i,l,k)				
Image	**L\K**	**6**	**12**	$\cdots$	**48**	**72**
Mandrill	**8**	20.76	20.79	$\cdots$	20.76	20.76
	12	20.90	20.97	$\cdots$	21.01	21.01
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	72	20.94	21.06	$\cdots$	21.20	21.20
Peppers	**8**	24.41	24.42		24.40	24.40
	12	24.90	24.91	$\cdots$	24.91	24.91
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	72	25.12	25.30	$\cdots$	25.41	25.41

Table 1: PSNR comparison using different learned non-separable dictionaries for the denoising task, where L correspond to the filter size and K to the FB's cardinality.

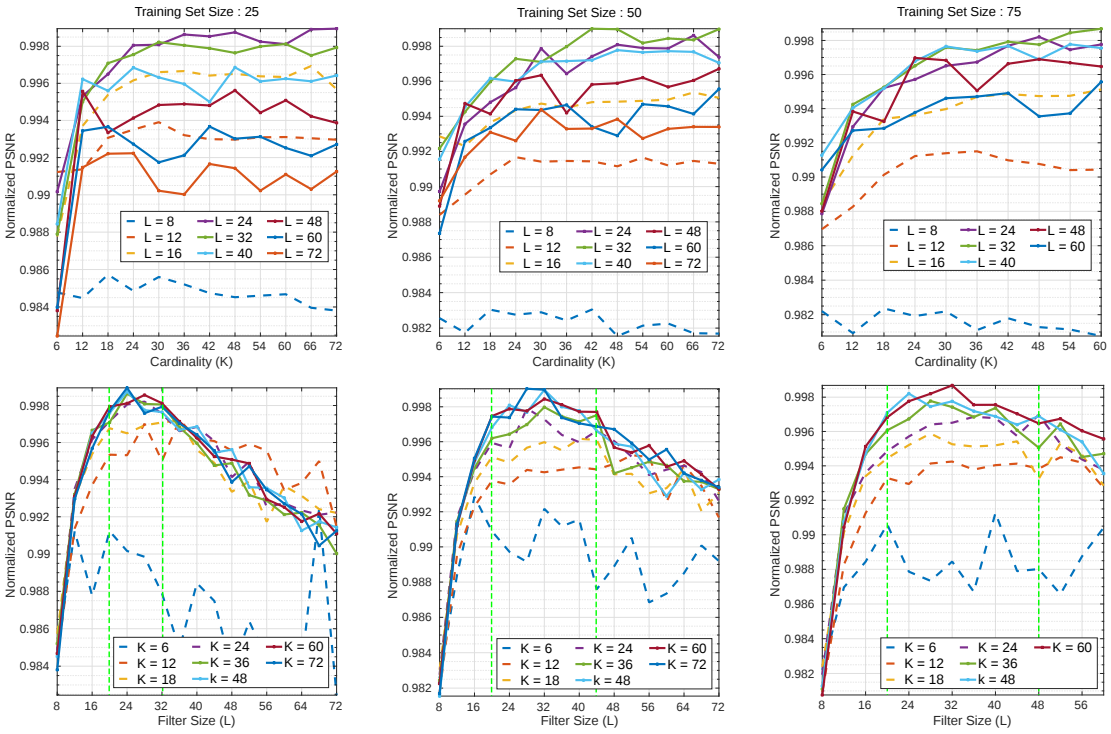

Fig. 3: Comparison of the average normalized PSNR scores (denoising task) w.r.t. the FB's cardinality (top row) and the filter size (bottom row) when using training set sizes of 25, 50 and 75 to learn the convolutional dictionaries. It is worth noting that the resulting plots of the average normalized PSNR scores w.r.t. the filter size are akin to a positive skew bell shape curve, where their skewness depends on the training set size; in green dash line, we remark a region (approximately) greater or equal to the median of the skew bell shape curve.

4.2. General analysis

It is worth mentioning that we report the average normalized PSNR scores for the eight test images in our computational results. In Figure 2, we jointly plot the average normalized PSNR scores (denoising task) along with the corresponding local $\ell_{0,\infty}$ norm of the feature maps obtained once the training stage is finished for the selected values of filter sizes and FB's cardinality. We can observe that going above a certain cardinality value ($18 - 24$ filters) has a negligible effect on the sparsity (measured by the $\ell_{0,\infty}$ norm) of the solution. We note that this value is remarkably close to the bound in which the average normalized PSNR scores stop rapidly increasing, which contributes to establishing the existence of a lower bound for FB's cardinality.

In Figure 3, we assess more broadly the changes of the average normalized PSNR scores through the FB's cardinality and the filter size. By observing the top row, we can appreciate, in most of the curves, a steady increment in terms of normalized PSNR up to 30 filters, after which they are almost constant. In this context, we note that a filter bank composed of 30 filters requires a smaller degree of computational resources and processing time with respect to larger ones that attain comparable performance.

Moreover, in Figure 3, bottom row, it can be observed that when the filter size is varied, the curves describe a (positive) skew bell shape, whose skewness increase with the training set size; vertical green dash lines are used it highlight a region (approximately) where the average normalized PSNR scores are greater o equal to their corresponding median value. For small training sets such as 25 images, the curve width that limits the best normalized PSNR values is narrow. This range becomes wider when increasing the training set size. Particularly, small filter sizes such as 8×8 and 12×12 that have been used in several works [9],[10],[12],[13], in which the main objective was to provide new computationally efficient algorithms with minimal loss in performance. However, a remarkable benefit in terms of performance can be achieved when using the filter size of 24×24 (lower bound). Although not the optimal size, the difference in performance when increasing the filter size above this lower bound is negligible.

5. CONCLUSIONS

We have presented a careful analysis of how hyper-parameters such as filter size and filter bank's cadinality affect the reconstruction performance in terms of PSNR for a denoising task in the context of convolutional dictionary learning. In contrast to the small filter sizes commonly used in the literature, we have observed a remarkable benefit in terms of performance when using the moderate filter size of 24×24 (lower bound). We can also conclude that there is a direct relationship between the $\ell_{0,\infty}$ norm of the feature maps induced by the FB's cardinality and the reconstruction metric used in the denosing task, which indicate the existence of a lower bound for the cardinality at approximately 24 to 30 filters.

6. REFERENCES

[1] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[2] J. Sulam, A. Aberdam, A. Beck, and M. Elad, "On multi-layer basis pursuit, efficient algorithms and convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[3] Shizhong Han, Zibo Meng, James O'Reilly, Jie Cai, Xiaofeng Wang, and Yan Tong, "Optimizing filter size in convolutional neural networks for facial action unit recognition," *CoRR*, vol. abs/1707.08630, 2017.

[4] Saleh Albelwi and Ausif Mahmood, "A framework for designing the architectures of deep convolutional neural networks," *Entropy*, vol. 19, no. 6, pp. 242, 2017.

[5] Vardan Papyan, Jeremias Sulam, and Michael Elad, "Working locally thinking globally-part ii: Stability and algorithms for convolutional sparse coding," *arXiv preprint arXiv:1607.02009*, 2016.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[7] Yurii Nesterov, "A method for solving the convex programming problem with convergence rate o$(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, pp. 543–547, 1983.

[8] Kjersti Engan, Sven Ole Aase, and Jhon Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. IEEE, 1999, vol. 5, pp. 2443–2446.

[9] Brendt Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 301–315, Jan. 2016.

[10] Michal Ŝorel and Filip Ŝroubek, "Fast convolutional sparse coding using matrix inversion lemma," *Digital Signal Processing*, vol. 55, pp. 44 – 51, 2016.

[11] Cristina Garcia-Cardona and Brendt Wohlberg, "Subproblem coupling in convolutional dictionary learning," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sept. 2017, pp. 1697–1701.

[12] Gustavo Silva and Paul Rodriguez, "Efficient convolutional dictionary learning using partial update fast iterative shrinkage-thresholding algorithm," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 4674–4678.

[13] Cristina Garcia-Cardona and Brendt Wohlberg, "Convolutional dictionary learning: A comparative review and new algorithms," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366–381, Sep. 2018.

[14] Gustavo Silva and Paul Rodríguez, "Efficient algorithm for convolutional dictionary learning via accelerated proximal gradient consensus," in *IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 3978–3982.

[15] Jorge Quesada, Paul Rodriguez, and Brendt Wohlberg, "Separable dictionary learning for convolutional sparse coding via split updates," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4094–4098.

[16] Gustavo Silva, Jorge Quesada, and Paul Rodríguez, "Efficient separable filter estimation using rank-1 convolutional dictionary learning," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP*, 2018, pp. 1–6.

[17] J. Quesada, G. Silva, P. Rodriguez, and B. Wohlberg, "Combinatorial separable convolutional dictionaries," in *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, April 2019, pp. 1–5.

[18] Neal Parikh, Stephen Boyd, et al., "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[19] David Donoho and Michael Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[20] Mark J Huiskes, Bart Thomee, and Michael S Lew, "New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative," in *Proceedings of the international conference on Multimedia information retrieval*. ACM, 2010, pp. 527–536.

[21] Gustavo Silva, "APG consensus based code," *Matlab and Python library available from goo.gl/1rvfDq*, 2018.

[22] Brendt Wohlberg, "Sparse optimization research code (SPORCO)," *Matlab and Python library available from goo.gl/BjVgH5*, 2017.