LEARNING LOW RANK AND SPARSE MODELS VIA ROBUST AUTOENCODERS

Jie Pu[1]*, Yannis Panagakis*[1,2]*, and Maja Pantic* [1,2]

[1]Department of Computing, Imperial College London, UK
[2]Samsung AI Research, Cambridge, UK

ABSTRACT

Robust principal component analysis (RPCA), decomposes a data matrix into a superposition of a low-rank matrix and a sparse matrix under certain incoherent conditions. In this paper, we propose a nonlinear generalization of RPCA that uses two autoencoder networks to achieve such a decomposition, in which one autoencoder accounts for the low-rank component and the other for the sparse component. To this end, we provide a principled way of constructing these autoencoders for low-rank and sparse components. The generality of the proposed model is demonstrated by applying it onto three applications, namely 1) music/voice separation 2) image denoising and 3) video foreground separation. Experimental results indicate the effectiveness of the proposed model on these application domains.

Index Terms— Autoencoders, Low-rank, Sparsity

1. INTRODUCTION

Principal component analysis (PCA) is a simple yet widely used method for dimensionality reduction, where a low rank approximation to the input data matrix is carried out. PCA finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. Although this procedure is simple to implement, it is sensitive to the presence of outliers and noise in data. To increase its robustness to outliers and noise, robust principal component analysis (RPCA) [1] is proposed to remove sparse corruptions from input data and then obtain its low rank approximation. In other words, RPCA splits an input matrix X into two parts, $X = L + S$, where L is a low rank approximation and S contains the sparse outliers and noise. Given such decomposition, not only the recovery of low rank component gets improved, but the sparse part is extracted, which enables the development of a wide range of applications such as face recognition [2], background modeling [3] and singing-voice separation [4], to name but a few.

The key idea of RPCA for improving the quality of low rank representation, is to explicitly add one sparse component to model noise in original data. Autoencoder, as another way to improve the low rank representation, is to generalize linear mappings in PCA into nonlinear ones. As shown by Hinton et al. [5], the low rank representation learned by nonlinear autoencoder networks works much better than PCA. However, in the setting of standard autoencoder networks, the nonlinear low rank representation is learned from clean training data, since its training is sensitive to outliers and noise. A number of approaches for robustifying autoencoder have been explored and proposed over the past decade, including denoising autoencoder [6] [7] [8] [9] and maximum correntropy autoencoder [10] [11]. In this paper, we also investigate the problem of robustifying autoencoder, from a different perspective over previous methods.

As we state above, both RPCA and autoencoder can be viewed as an extension of PCA. Therefore, is it possible to combine the best of two methods, which remains the nonlinear representation ability as well as be robust to outliers and noise? This question motivates our research and leads to the proposed model *Robust Autoencoders*. Thus the proposed model can be viewed in two ways: 1) autoencoders with robustness to outliers and noise, by explicitly modeling a sparse corrupted component. 2) RPCA with nonlinearity, using autoencoder networks to parameterize nonlinear mappings. The key idea of the proposed model is, using two autoencoders to disentangle the low-rank component and the sparse component of a data matrix. It bridges the gap between RPCA and autoencoders, and seeks the best of "both worlds". The combination framework in this paper can be further extended to bridge neural networks with a general family of low-rank and sparse models.

Unlike standard autoencoders that learn low-rank representations via a set of clean training data, the proposed model obtains the low-rank representation via disentanglement of two autocoders. The disentanglement does not rely on clean training data and enables the model work in an unsupervised setting. Besides, the disentanglement explicitly separates out one sparse component, while standard autoencoders cannot. This separated sparse component can be the object of interests for some real-life applications, e.g. video foreground separation in Section 4.3.

It is worth mentioning that Zhou et al. [12] developed a variant of robust autoencoder, which shares some similarity

This work was supported by the European Community Horizon 2020 under grant agreement no. 688835 (DE- ENIGMA). The work of Y. Panagakis was supported by EPSRC project (FACER2VM) under grant EP/N007743/1.

to the proposed model since they are also inspired by RPCA. However, the major difference between their model and the proposed model is: [12] uses one autoencoder to learn the low rank representation through a set of training data, while the proposed model obtain the low rank representation via disentanglement of two autoencoders.

2. ROBUST PRINCIPAL COMPONENT ANALYSIS

Robust principal component analysis (RPCA) is able to recover low rank representations of a data matrix even though a positive fraction of its entries are arbitrarily corrupted. This corrupted part of data matrices is explicitly modeled as a sparse component in RPCA. That is, $\mathbf{X} = \mathbf{L} + \mathbf{S}$, where $\mathbf{X}$ is the original data matrix, $\mathbf{L}$ has low rank and $\mathbf{S}$ is sparse. A natural estimator accounting for the sparsity of $\mathbf{S}$ is to minimize the number of nonzero entries, i.e. ℓ_0-norm. The following optimization problem is formulated:

$$\begin{aligned} &\min_{L,S} \ \text{rank}(\mathbf{L}) + \lambda \|\mathbf{S}\|_0 \\ &\text{s.t.} \quad \mathbf{X} = \mathbf{L} + \mathbf{S} \end{aligned} \tag{1}$$

Where λ is a positive regularization parameter to balance the significance of minimizing the sparsity compared to the minimization of rank. Although optimization problem (1) is straightforward to think, both rank and ℓ_0-norm minimization is NP-hard and thus intractable. Then, Candes et al. [1] use the nuclear norm $\|.\|_*$ and the ℓ_1-norm to serve as convex surrogates of the rank and ℓ_0-norm, respectively. Accordingly, the RPCA is defined as:

$$\begin{aligned} &\min_{L,S} \ \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \\ &\text{s.t.} \quad \mathbf{X} = \mathbf{L} + \mathbf{S} \end{aligned} \tag{2}$$

Where $\|.\|_*$ is the nuclear norm, i.e. the sum of singular values of the matrix. $\|.\|_1$ is the ℓ_1-norm, i.e. the sum of absolute values of entries. The optimization problem (2) is a convex relaxation of (1) and has been proved to reliably obtain the low-rank matrix $\mathbf{L}$ and the sparse matrix $\mathbf{S}$ under certain incoherent conditions [1].

3. ROBUST AUTOENCODERS

The proposed *Robust Autoencoder* (RAE) is essentially a combination of autoencoders and RPCA. It explicitly models the outliers in data matrices similar to RPCA (i.e. the sparse component to increase its robustness), and at the same time, has nonlinear capability similar to the autoencoder networks. Figure 1 shows the architecture of the proposed model.

In analogy to RPCA, the proposed model splits input data $\mathbf{X}$ into two parts $\mathbf{X} = \mathbf{L} + \mathbf{S}$, where each part is modeled by an autoencoder network. In particular, the low rank representation $\mathbf{L}$ is modeled by an under-complete (low-rank) autoencoder, while the sparse component $\mathbf{S}$ is modeled by a sparse autoencoder. That is,

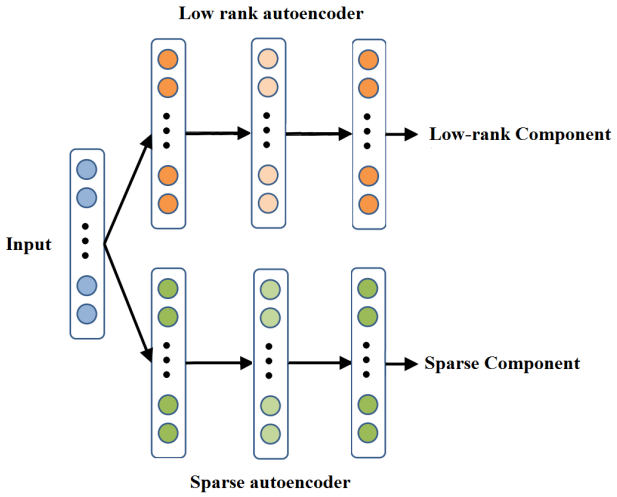


Fig. 1. Proposed robust autoencoders architecture. The low rank autoencoder contains an encoder $f_L(.)$ and a decoder $g_L(.)$. Similarly, the sparse autoencoder has an encoder $f_S(.)$ and a decoder $g_S(.)$.

$$\begin{aligned} &\min_{f_L, g_L, f_S, g_S} \ \text{rank}(\mathbf{L}) + \lambda \|\mathbf{S}\|_0 \\ &\text{s.t.} \quad \mathbf{X} = \mathbf{L} + \mathbf{S} \\ &\qquad \mathbf{L} = g_L(f_L(\mathbf{X})) \\ &\qquad \mathbf{S} = g_S(f_S(\mathbf{X})) \end{aligned} \tag{3}$$

Where $f_L(.), g_L(.)$ are the encoder and decoder functions for $\mathbf{L}$. $f_S(.), g_S(.)$ are the encoder and decoder functions for $\mathbf{S}$.

Herein, we use ℓ_1-norm to serve as convex surrogates of ℓ_0-norm, as Candes et al. [1] suggested. For the rank minimization, we adopt ideas from under-complete autoencoders [13], which implicitly force the rank of encoding representation be small by setting the number of neurons in the middle layer much less than the input layer. This relaxed rank constraint can be described as $rank(f_L(\mathbf{X})) < rank(\mathbf{X})$. Therefore, the complete optimization problem of the proposed robust autoencoders can be formulated as following:

$$\begin{aligned} &\min_{f_L, g_L, f_S, g_S} \ L\big(\mathbf{X},\ g_L(f_L(\mathbf{X})) + g_S(f_S(\mathbf{X}))\big) + \lambda \|g_S(f_S(\mathbf{X}))\|_1 \\ &\text{s.t.} \quad rank(f_L(\mathbf{X})) < rank(\mathbf{X}), \end{aligned} \tag{4}$$

where L is a loss function to measure the difference, and λ is a positive parameter to balance the significance of minimization. In particular, a small value of λ will encourage more data to be isolated into the sparse component $\mathbf{S} = g_S(f_S(\mathbf{X}))$ and obtain a better reconstruction of $\mathbf{X}$, while a large λ works the other way around.

It is worth mentioning that the sparse autoencoder we used in Figure 1 is different from traditional sparse autoencoders [14] [15]. Traditional sparse autoencoders often have

an over-complete encoding representation in the middle layer (i.e. $rank(f_S(\mathbf{X})) > rank(\mathbf{X})$) and a sparsity constrain on the encoding representation (i.e. $\|f_S(\mathbf{X})\|_1$). They are resembling of *sparse coding* [16] and seek to learn useful sparse representation. However, our goal is to disentangle a sparse component in original data space, in analogy to RPCA. Thus our sparse autoencoder has the same dimension of the encoding representation (i.e. $rank(f_S(\mathbf{X})) = rank(\mathbf{X})$), and put the sparsity constrain on the output layer (i.e. $\|g_S(f_S(\mathbf{X}))\|_1$).

4. EXPERIMENTS

In this section, we present an experimental evaluation of the proposed model in practical applications. Three sets of experiments are conducted which are summarized as follows:

- As a novel approach for low-rank and sparse modelling, the proposed model is evaluated on the music/voice separation task with a comparison against other unsupervised state-of-the-art methods.
- As a robust extension of standard autoencoders, the proposed robust autoencoder is evaluated on an image denoising task with a comparison against standard autoencoders.
- As a nonlinear variant of RPCA, the proposed robust antoencoder is evaluated on a video foreground separation task with a comparison against RPCA.

Our implementation is built on the machine learning library, tensorflow [17]. We use two layers of autoencoders, and each layer of encoder $f(.)$ and decoder $g(.)$ has the activation function *ReLu*. Note, all experiments are conducted in an unsupervised manner, i.e. no training data is provided and we directly optimize the problem in (4).

4.1. Music/voice separation

We now evaluate the proposed model in the task of music/voice separation, which aims to separate the singing voice and musical background from a monaural recording. This task is very challenging when no prior training or particular features are provided, i.e. in an unsupervised manner. RPCA has been shown to provide the state-of-the-art results [4].

Since music instruments can reproduce the same sounds each time they are played and music has, in general, an underlying repeating musical structure, we can think of music as a low-rank signal. Singing voices, on the contrary, have more variation (higher rank) but are relatively sparse in the time and frequency domains. We can then think of singing voices as components making up the sparse matrix. In both RPCA and the proposed model, the low-rank component $\mathbf{L}$ is expected to contain music accompaniment and the sparse component $\mathbf{S}$ to contain vocal signals.

The separation performance is evaluated on the MIR-1K dataset [18], containing 1000 Chinese karaoke clips performed by amateur singers. Voice and music will be mixed at 0 dB. The experimental settings closely followed that of [4]. For the evaluation criteria, we report the global normalized source-to-distortion ratio (GNSDR), global source-to-interference ratio (GSIR) and global source-to-artifacts ratio (GSAR) as [4], [19]. The most important measure is GNSDR as it measures the overall performance.

Results for the music channel of MIR-1K dataset are shown in Table 1, with comparison to REPET-SIM [20], RPCA [4] and RPCAs [19]. As you can see, the proposed robust autoencoder achieved the best separation results for the music channel of MIR-1K dataset, in terms of the overall criterion GNSDR. On the other hand, the proposed model performs slightly worse than RPCA for separating the voice channel of MIR-1K dataset, where the proposed RAE obtained GNSDR $= 1.63$ dB as comparison to GNSDR$= 1.68$ dB of RPCA with binary mask [4].

Methods	GNSDR	GSIR	GSAR
REPET-SIM [20]	2.83	4.55	9.82
RPCA [4]	3.32	5.41	9.20
RPCAs [19]	4.52	6.48	**10.4**
RAE	**5.99**	**7.25**	6.48

Table 1. Separation quality in dB for the music channel of MIR-1K dataset. RPCAs is the RPCA with vocal/non-vocal masks [19]. RAE is the proposed robust autoencoder.

4.2. Image denoising

Denoising one noisy image without information of any clean images, is challenging. Given a noisy image $\mathbf{X}$, our goal is to represent the image using two disjoint parts, $\mathbf{L}$ and $\mathbf{S}$. We aim $\mathbf{L}$ to contain all the information relevant for the clean part, and $\mathbf{S}$ to contain only the corrupted noise. This is a much harder task than the traditional denoising task performed by autoencoders [6] [8] [9] [12], since it has no separated training process on clean images and thus requires the model be much more robust to noise.

We analyzed 10 gray-scale images typically used to benchmark image denoising methods. "Salt and pepper" noise is added to each image with a noise density = 0.1 and 0.2 (this effects 10% and 20% of pixels, respectively). Then each corrupted image is used as the input to the proposed model. By learning the optimization problem in (4), the proposed model obtains the disentanglement of each corrupted image $\mathbf{X}$ into two parts, low-rank component $\mathbf{L}$ and sparse component $\mathbf{S}$.

The key insight here which enables our disentanglement is, natural images have regularities and can be effectively compressed in low dimensions [21], while noise and outliers are often incompressible. Thus we expect the low-rank

component $\mathbf{L}$ to capture the clean part of $\mathbf{X}$ and be similar to the original clean image. The incompressible noise occupies only a fraction of images (10% or 20%) thus can be effectively modeled by a sparse component $\mathbf{S}$.

We use recovered peak signal-to-noise ratio (R-PSNR) to measure the denoising performance. In particular, R-PSNR is defined as: $\text{R-PSNR}(\mathbf{L}, \mathbf{X}) = \text{PSNR}(\mathbf{L}, \mathbf{I}) - \text{PSNR}(\mathbf{X}, \mathbf{I})$, where $\mathbf{I}$ is one original image, $\mathbf{X}$ is the noisy image and $\mathbf{L}$ is the denoised image. Higher values of R-PSNR indicate better performance. Experimental results of the proposed model is shown in Table 2, with comparisons against PCA and a standard autoencoder. It is easy to see the superior performance of the proposed model thanks to its effective disentanglement.

Images	$\sigma = 0.1$			$\sigma = 0.2$		
	PCA	AE	RAE	PCA	AE	RAE
baboon	2.04	3.58	**4.01**	2.29	3.80	**6.25**
hill	2.60	3.49	**11.55**	2.65	3.51	**12.60**
barbara	2.46	3.38	**6.82**	2.52	3.70	**6.89**
boat	2.57	4.47	**8.77**	2.61	4.43	**10.65**
camera man	0.93	1.81	**3.04**	1.04	2.09	**3.80**
couple	2.61	5.79	**9.58**	2.62	6.39	**7.85**
house	0.95	3.57	**10.94**	1.05	3.19	**7.35**
lena	2.64	4.57	**6.59**	2.64	4.98	**9.15**
man	2.56	3.03	**5.86**	2.61	3.61	**6.83**
pepper	0.92	3.51	**9.53**	1.02	3.25	**6.37**

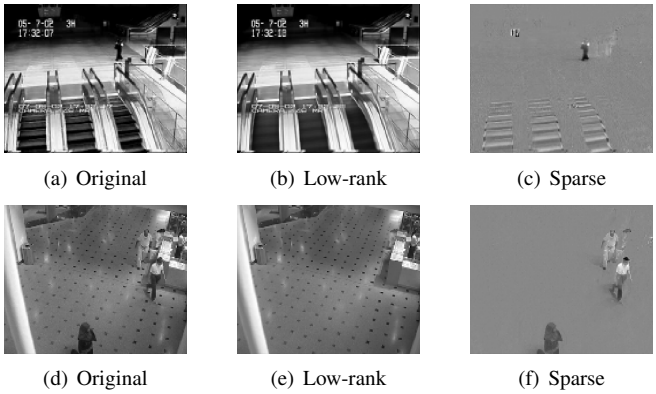
Table 2. Denoising results of PCA, autoencoder (AE) and the proposed robust autoencoder (RAE), with the noise density σ = 0.1 and 0.2. Performance is measured as R-PSNR in dB.

4.3. Video foreground separation

To further investigate the disentangle ability of the proposed model, we apply it to the problem of foreground separation in videos. The previous experiment of image denoising focus on the low-rank component $\mathbf{L}$, but here the object of interest is the sparse component $\mathbf{S}$. The observed video is formed as a matrix $\mathbf{X}$ by vectorizing each frame and stacking them column-wise. We assume the background in a video is static and hence forms a low-rank component $\mathbf{L}$, while the foreground is a dynamic but sparse perturbation.

Two benchmark datasets [3], named *Escalator* and *ShoppingMall*, is used to evaluate the separation performance. The *Escalator* dataset has 3417 frames at a resolution of 160×130, and the *ShoppingMall* dataset has 1286 frames at 320×256. The separated foreground of these two videos is shown in Figure 2. As we can see, the proposed autoencoder networks effectively separate out the dynamic but sparse foreground, which is the pedestrians in *ShoppingMall*, the moving escalator and one walking woman in *Escalator*.

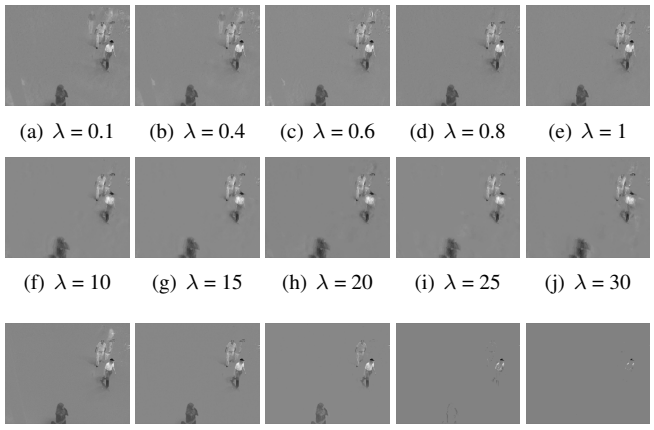
The nonlinear property of the proposed model is demonstrated with a comparison against RPCA. As shown in Figure 3, the proposed model produce a more stable foreground separation than RPCA when the regularization parameter λ varies. The λ of the proposed model varies from 0.1 to 30, which is 300 times bigger, but still produces reasonably good results. In contrast, the λ of RPCA varies from 0.001 to 0.03 (30 times), while its performance dramatically decreases.



(a) Original (b) Low-rank (c) Sparse

(d) Original (e) Low-rank (f) Sparse

Fig. 2. Foreground separation results of the proposed model. (a)-(c) from *Escalator* video, and (d)-(f) from *ShoppingMall*.



(a) $\lambda = 0.1$ (b) $\lambda = 0.4$ (c) $\lambda = 0.6$ (d) $\lambda = 0.8$ (e) $\lambda = 1$

(f) $\lambda = 10$ (g) $\lambda = 15$ (h) $\lambda = 20$ (i) $\lambda = 25$ (j) $\lambda = 30$

(k) $\lambda = 0.001$ (l) $\lambda = 0.005$ (m) $\lambda = 0.01$ (n) $\lambda = 0.02$ (o) $\lambda = 0.03$

Fig. 3. Foreground separation results of the frame 1715 in *ShoppingMall*. (a)-(j) are results of the proposed model with varying values of λ. (k)-(o) are results from RPCA.

5. CONCLUSION

In this paper, we have shown how to disentangle the low-rank component and the sparse component of a data matrix using two autoencoders. The proposed model can be viewed as a combination of RPCA and autoencoders, which seeks to combine the best of "both worlds". It remains the nonlinear capability as autoencoders, and at the same time be robust to outliers and noise as RPCA. The advantages of the proposed model have been demonstrated in three practical applications: music/voice separation, image denoising and video foreground separation.

6. REFERENCES

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[2] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[3] Liyuan Li, Weimin Huang, Irene Yu-Hua Gu, and Qi Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1459–1472, 2004.

[4] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 57–60.

[5] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[7] Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3377–3381.

[8] Lingheng Meng, Shifei Ding, and Yu Xue, "Research on denoising sparse autoencoder," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 5, pp. 1719–1729, 2017.

[9] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, 2012, pp. 341–349.

[10] Yu Qi, Yueming Wang, Xiaoxiang Zheng, and Zhaohui Wu, "Robust feature learning by stacked autoencoder with maximum correntropy criterion," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6716–6720.

[11] Weifeng Liu, Puskal P Pokharel, and Jose C Principe, "Correntropy: A localized similarity measure," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 4919–4924.

[12] Chong Zhou and Randy C Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 665–674.

[13] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[14] Christopher Poultney, Sumit Chopra, Yann L Cun, et al., "Efficient learning of sparse representations with an energy-based model," in *Advances in neural information processing systems*, 2007, pp. 1137–1144.

[15] Y-lan Boureau, Yann L Cun, et al., "Sparse feature learning for deep belief networks," in *Advances in neural information processing systems*, 2008, pp. 1185–1192.

[16] Bruno A Olshausen and David J Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607, 1996.

[17] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[18] Chao-Ling Hsu and Jyh-Shing Roger Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[19] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 718–722.

[20] Zafar Rafii and Bryan Pardo, "Music/voice separation using the similarity matrix.," in *ISMIR*, 2012, pp. 583–588.

[21] Gregory K Wallace, "The jpeg still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.