UNSUPERVISED REPRESENTATION LEARNING WITH PRIOR-FREE AND ADVERSARIAL MECHANISM EMBEDDED AUTOENCODERS

Xing Gao, Hongkai Xiong

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
{william-g,xionghongkai}@sjtu.edu.cn

ABSTRACT

Most state-of-the-art methods for representation learning are supervised, which require a large number of labeled data. This paper explores a novel unsupervised approach for learning visual representation. We introduce an image-wise discrimination criterion in addition to a pixel-wise reconstruction criterion to model both individual images and the difference between original images and reconstructed ones during neural network training. These criteria induce networks to focus on not only local features but also global high-level representations, so as to provide a competitive alternative to supervised representation learning methods, especially in the case of limited labeled data. We further introduce a competition mechanism to drive each component to increase its capability to win its adversary. In this way, the identity of representations and the likeness of reconstructed images to original ones are alternately improved. Experimental results on several tasks demonstrate the effectiveness of our approach.

Index Terms— Unsupervised representation learning, adversarial learning, autoencoders

1. INTRODUCTION

The choice of data representation primarily determines the performance of subsequent tasks, such as image classification [1] and object recognition [2, 3, 4], and the evolution of data representation largely contributes to the development of computer vision. Therefore, researchers are constantly committed to seeking and exploring effective representations for visual tasks. At the initial stage, researchers were dedicated to the design of preprocessing pipelines and transformations on the basis of considerable domain expertise and careful engineering, such as SIFT [2]. However, these hand-crafted features fail to capture high-level representation and thereby their performances are unsatisfactory. Recently, deep learning methods automatically learn multiple levels of representations from raw data, and drastically improve the performance in a series of tasks in computer vision [1, 3, 4]. However, the majority of deep learning algorithms belong to supervised learning paradigm, whose performances are highly dependent on large-scale labeled datasets which are not always available.

Compared to supervised learning methods, unsupervised learning techniques only employ available unlabeled data, and thereby are more flexible to handle. However, their performances are usually far inferior to their supervised counterparts. Recently, Generative Adversarial Nets (GAN) [5] substantially improves the quality of generated samples by introducing a discriminate model to combat with the generative model. However, most of the current work on GAN models focuses on image generation that maps latent variables to images, and only a few work involves representation extraction that projects image space to latent space. ALI [6] and BiGAN [7] introduce an extra network into GAN to learn the inverse mapping that projects data space to latent space. Although such methods achieve competitive performance when learned features are transferred to object detection and classification tasks, these generation based methods essentially extract co-occurrence statistics of features instead of distinctive features.

In this paper, we design an unsupervised representation learning framework to automatically learn a direct mapping from image space to representation space inspired by the adversarial mechanism of GAN models. On the one hand, we apportion the unsupervised representation learning task to two "pretext" tasks: reconstruction and discrimination. In addition to concentrating on low-level features driven by the reconstruction task as vanilla autoencoders [8] and regularized autoencoders, the proposed model further focuses on high-level representations, such as recognizable patterns, through the discrimination task. Achieving all of these, we only introduce a simple binary classifier to the framework of vanilla autoencoders. On the other hand, we introduce a competition mechanism to further enhance the distinction of representation and improve the resemblance of reconstructed image to original ones. Except that the encoder and the decoder are optimized in accordance with the reconstruction criterion, a discriminator composed of the encoder and the binary classifier competes with the decoder according to the discrimination criterion. Furthermore, we derive the compatibility of these two criteria in theory and valid the effectiveness of proposed model through a series of visual applications.

This work was supported in part by the NSFC Grants 61425011, 61720106001, and 61529101, and in part by the Program of Shanghai Academic Research Leader under Grant 17XD1401900. Thanks a lot.

978-1-5386-1737-3/18/$31.00

2. RELATED WORK

Representation learning focuses on extracting useful information of data to solve machine learning tasks, including classification and prediction. Unsupervised learning methods resort to available unlabeled data and adopt some "pretext" tasks to learn useful representations. One family of such methods are probabilistic models that find a parsimonious set of latent variables to describe the distribution of data, such as RBMs [9], but they become complicated and intractable with deep network architectures. Another category seek to directly learn a mapping between data space and representation space, represented by autoencoder family [8, 10]. However, such reconstruction-based methods tend to be trapped in low-level features and thereby fail to solve complex semantic tasks, such as object detection and image classification.

Another thread of unsupervised representation learning is the generative model. For instance, GAN [5] and DCGAN [11] models generate high-quality images from latent space and have demonstrated the relationship between certain latent elements and gender. ALI [6] and BiGAN [7] further introduce an inference model to GAN framework to map image space back to latent space. However, all of these models essentially seek to learn a mapping that transforms a fixed simple probability distribution, such as Gaussian distribution or uniform distribution, to the implicit distribution of data, instead of attempting to extract representations of data. Even if the ALI and BiGAN models learn an extra inference network, the latent variable assigned to each data is randomly sampled from prior distribution during training phase so that these methods essentially learn co-occurrence statistics of features. The deviation between the priori distribution and the real distribution of latent variables, nevertheless, will lead to the inaccuracy of captured statistics characteristics. Furthermore, these statistics characteristics are not necessarily equivalent to distinctive and discernible features. Although we also utilize game playing of GAN, our method differs from them in: (1) our approach automatically extracts representations from raw data through classification and reconstruction tasks without any hypothesis about the priori distribution of representations, which is the same as supervised representation methods; (2) our method shares most of the architectures between the inference network and the discrimination network instead of using two separate networks, which enables the inference network to be driven by both criteria to focus on different aspects of representation and at the mean time saves a large number of free parameters to avoid overfitting.

3. ADVERSARIAL MECHANISM EMBEDDED AUTOENCODERS

In this section, we introduce the framework of adversarial mechanism embedded autoencoders (AME-AE), describe its learning procedure, and derive some theoretical results.

3.1. Framework

The framework of our approach, as shown in Fig.1 (a), consists of an encoder (E), a decoder (D), and a discriminator (Dis) composed of the encoder cascaded with a classifier (C). The encoder produces representation of input image $h = E(x)$, the decoder reconstructs image from the representation $\hat{x} = D(h)$, and the discriminator tells the original image x from the reconstructed one $\hat{x}$. Compared to autoencoder families, we introduce an extra classifier, based on which we share the encoder to construct the discriminator, as illustrated in Fig.1 (b). The introduction of the classifier brings in discrimination-based criterion in addition to reconstruction-based criterion, so as to prevent learned representations from being trapped in low-level phenomenon. Note that our approach is unsupervised, because the category of binary classification is just the original image as well as the reconstructed image which is the byproduct of decoder without any labeled information. Furthermore, the reuse of the encoder in the discriminator saves the free parameters and makes it possible to drive the encoder by both reconstruction and classification criteria to focus on different aspects of representations.

3.2. Training Criteria

To efficiently learn representations of data, our approach consists of two "pretext" tasks: reconstruction and discrimination. On the one hand, as with vanilla autoencoders, the encoder along with decoder is jointly optimized on the basis of reconstruction criterion. On the other hand, the discriminator is updated in line with discrimination criterion. In stead of utilizing cross-entropy loss, we introduce adversarial loss to make the discriminator compete with the decoder.

Reconstruction Criterion. We use the Mean Square Error (MSE) as reconstruction loss function:

$$\begin{aligned}\min_{E,D} L_{rec} &= E_{x\sim p_d(x)}\|x^i - \hat{x}^i\|_2^2 \\ &= E_{x\sim p_d(x)}\|x^i - D(E(x^i))\|_2^2.\end{aligned} \tag{1}$$

Discrimination Criterion. We make the discriminator compete with its adversary the decoder. Concretely, the discriminator is optimized to identify the category of original images or reconstructed images as accurately as possible, while the decoder is trained to improve the similarity between reconstructed images and original images in order to deceive the discriminator. Competition drives both models to improve their performances. In other words, they are optimized based on the following adversarial loss function:

$$\begin{aligned}\max_{E,C}\min_{D} L_{dis} =& E_{x\sim p_d(x)}[log(Dis(x^i)) \\ &+ log(1 - Dis(D(E(x^i))))],\end{aligned} \tag{2}$$

with $Dis \triangleq C \circ E$ and '$\circ$' meaning composite operation. For balancing the decoder and discriminator, we utilize non-

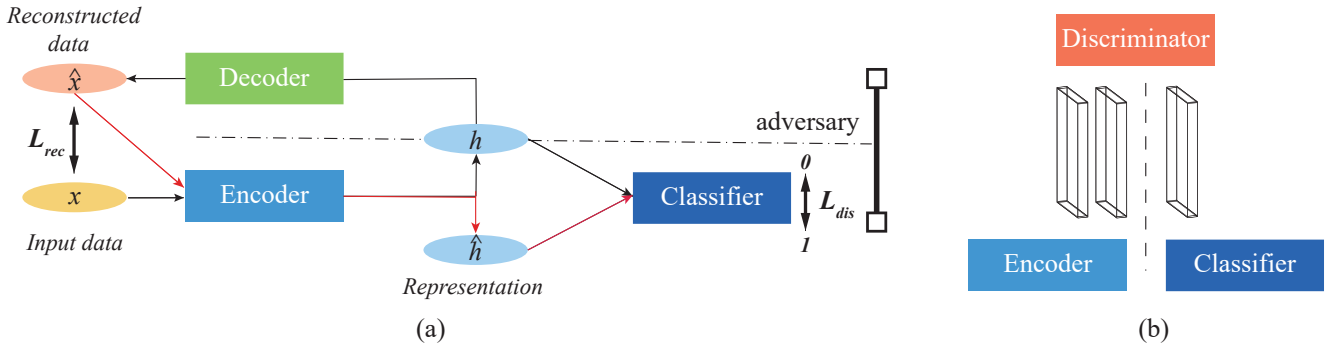


Fig. 1. (a) The framework of adversarial mechanism embedded autoencoders. The black arrow indicates the flow of input data, while the red arrow shows the flow of reconstructed data. (b) The relationship between the encoder, classifier, and discriminator.

saturating loss for the decoder in practice:

$$\max_{D} L_{dis} = E_{x\sim p_d(x)}[log(Dis(D(E(x^i))))]. \quad (3)$$

Joint Criteria. These losses are weighted to obtain the final cost function:

$$\min_{E,D,C} L(E,D,C) = -\lambda_{dis}L_{dis} + \lambda_{rec}L_{rec}. \quad (4)$$

We alternatively optimize the encoder together with classifier and the decoder on the basis of the joint criteria using gradient-based methods with minibatches of samples approximating expectation terms.

3.3. Optimality

The introduction of the classifier amortizes representation learning task over the reconstruction and discrimination "pretext" tasks to drive the encoder to focus on different aspects of features. Nonetheless, it is unclear whether the criteria corresponding to these two tasks are compatible. In this section, we theoretically derive the optimal solutions of these criteria and demonstrate that the optima solutions are consistent.

Proposition 1. *For any encoder and decoder, the optimal classifier C is*

$$C^*(h) = \frac{q_d(h)}{q_d(h) + q_r(h)}, \quad (5)$$

with q_d and q_r indicate the distributions of the representations of original data and the representations of reconstructed data, respectively.

Based on the optimal classifier, we further explore the characteristic of the encoder and decoder under the discrimination criterion.

Proposition 2. *Under an optimal classifier C^* and any encoder E, the objective of the decoder D $V_1(D) := \max_C L_{dis}(C,E,D) = L_{dis}(C^*,E,D)$ can be reformulated as the Jensen-Shanon divergence, and achieves its minimum if and only if $q_d = q_r$, and $D = E^{-1}$ is a minimum point.*

Note that $D = E^{-1}$ is the global minimum of the reconstruction criterion $L_{rec}(D,E)$. On this basis, we are able to obtain the constraints of optimal encoder and decoder under joint criteria.

Theorem 1. *Under an optimal classifier C^* and any given encoder E, the decoder D minimizes the loss function $V(D) := \lambda_{dis}V_1(D) + \lambda_{rec}V_2(D)$,with $V_2(D) := L_{rec}(E,D)$, and achieves its minimum if and only $D = E^{-1}$.*

Therefore, the reconstruction criterion and discrimination criterion are compatible and the constraints of optimal encoder and decoder are consistent that they are inverses. The proof is stated in the supplemental material.

3.4. Discussion

Our approach takes the advantages of reconstruction criterion and discrimination criterion and makes them complement each other. On the one hand, the reconstruction criterion leads the encoder to extract low-level features, such as texture and brightness, because the MSE loss implicitly concentrates on features responding to significant brightness variations. On the other hand, the discrimination criterion induces the encoder to focus on recognizable patterns. According to [12], generative models trained with adversarial loss can even generate elaborate structured patterns composed of a small number of pixels without significant intensity variations well, such as ears, which is impossible under MSE loss. For this reason, it is reasonable to infer that that the adversarial loss implicitly specifies structured patterns as salient features.

Furthermore, according to the training procedure of adversarial based methods, it is through the discriminator that the adversarial loss acts on the generator. In other words, the adversarial loss resorts to the discriminator to extract and select features. Therefore, the adversarial loss implicitly leads the discriminator to respond to structured pattern features. The framework of our approach reconciles the MSE loss and the adversarial loss and enables us to directly apply them to

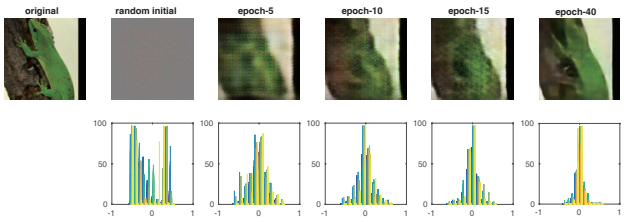
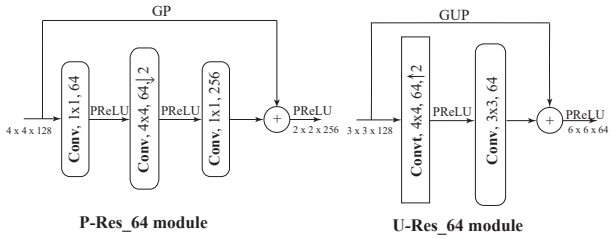
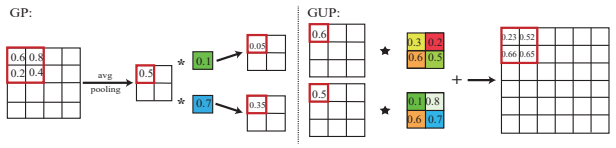


Fig. 2. Visualization of the convergence process of training phase. The first row presents an original image and several reconstructed images at different epoches, while the second row shows the corresponding histograms of the RGB three-channel distributions of pixel-wise differences between the original image and reconstructed ones.



(a) Building blocks: residual modules.



(b) Illustration of operations of group shortcuts.

Fig. 3. Building blocks: residual modules and shortcuts. The '$*$' and 'Conv' mean convolution operation, and the '$\star$' as well as 'Convt' indicate transposed convolution operation. For each layer, the formation is operation, filter kernel size, output channels, (stride, default =1).

the encoder, which drives the encoder to extract local low-level features as well as global high-level structured patterns. In addition, the adversarial loss alone is unstable and sensitive to hyperparameters in practice, the introduction of MSE loss alleviates this problems to some extent.

Similar to the ALI and BiGAN models, it is hard to provide theoretical analysis of convergence due to complex dependency. We utilize experiment results to illustrate the proposed algorithm approximately converges to the optimal solution. According to Fig.2, the RGB three-channel distributions of pixel-wise differences between the original image and reconstructed images gradually concentrate to 0, which indicates the reconstructed image converging to the original image and thereby the constraint of optimal solution is satisfied.

4. EVALUATION

4.1. Network Architecture and Configuration

Since the encoder and decoder are both deep networks and composed of considerable parameters, it is a challenge to ensure the effective propagation of information and gradient through these two networks. To this end, we propose two kinds of residual building blocks: the P-Res module and U-Res module, as shown in Fig.3. The P-Res module is a modified version of the "bottleneck" building block, which consists of three convolution layers and a group projection (GP) shortcut composed of average pooling and group convolution. Inspired by image super-resolution technology [13], the U-Res module is designed to reconstruct image through step by step improvement. Concretely, a low-resolution feature map f is mapped to a rough one $\hat{F}$ as the size of high-resolution one through the group up projection (GUP) shortcut, at the same time the residual or fine details $F - \hat{F}$ is ameliorated through a transposed convolution layer along with a convolution layer. The GUP shortcut contains few parameters and consists of group transposed convolution. Note that P-Res and U-Res these two building blocks provide a highway to propagate information and gradient, and at the same time reduce the parameters to a great extent.

We implement the proposed AME-AE framework on Matconvnet [14], and apply it to two datasets: MNIST and STL-10. The encoder and the decoder are respectively designed as a residual convolution network composed of the P-Res modules and a residual deconvolution network consisting of the U-Res modules, and the classifier is a shallow neural network. We take ADAM as solver for optimization, with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and initial learning rate $1e-4$. Hyper-parameters λ_{dis} and λ_{rec} vary with tasks. More details about the network architectures are stated in supplemental material.

4.2. Image Classification

Since classification is the most basic and widely used application, we firstly evaluate the performance of learned representation on image classification. We take the encoder unsupervisedly trained on STL-10 unlabeled set as a fixed feature extractor, and apply a 4-quadrant max-pooling on each feature map of the encoder to obtain representations of data. Then a linear L_2 regularized SVM is trained on such representations. We report the results using standard training and testing protocols of STL-10 dataset in Table 1. We compare our approach with various regularized autoencoders, classical unsupervised learning methods, and a supervised network. For fair comparison, all of the autoencoders and the supervised network share the same encoder architecture as the proposed AME-AE.

On the one hand, according to the first group of Table 1, the proposed AME-AE framework overwhelms all the other common regularized autoencoders with about 5 percentage and more. Furthermore, we take three kinds of autoencoders

Table 1. Image classification on the STL-10 dataset.

method	accuracy(%)
AE	49.5 ± 0.7
DAE	49.7 ± 0.4
Sparse AE	53.8 ± 0.6
AME-AE	58.1 ± 0.6
Sparse AME-AE	58.2 ± 0.8
Denoising AME-AE	$\mathbf{60.7 \pm 0.7}$
K-means [15]	51.5 ± 1.7
ICA [16]	52.9
Sparse filtering [17]	53.5 ± 0.5
SC [18]	59.0 ± 0.8
DLIF [19]	61.0
EPLS [20]	61.0 ± 0.6
DCGAN [11]	$\mathbf{63.8 \pm 0.5}$
Supervised net with same architecture	52.2 ± 1.6

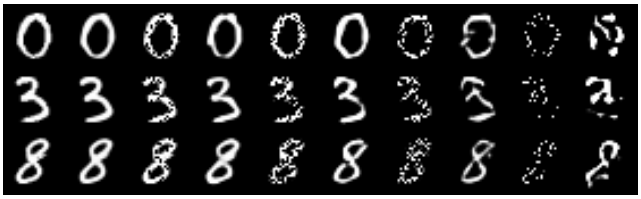
and place them into the framework of AME-AE to further validate the effect of discrimination criterion. Vanilla autoencoder (AE) and AME-AE, denoising autoencoder (DAE) and Denoising AME-AE, Sparse AE and Sparse AME-AE are such three pairs that one of each pair is a certain autoencoder and another is its AME-AE counterpart. For all these three pairs, the proposed framework respectively improves accuracy performances by 8.65, 11.18, and 4.45 percentage. Therefore, the cooperation of discrimination and reconstruction tasks enhances the capability of the encoder to extract high-level features for classification.

On the other hand, the proposed method achieves competitive performance with other unsupervised learning methods, as shown in the second group of Table 1. Although the DCGAN model surpasses the proposed model, the number of its features are 1.3 times larger than the proposed model due to network architecture differences and thereby it can contain more information. In addition, the performance of supervised net is inferior to the proposed AME-AE due to limitation of labeled data, which further validates the effect of the AME-AE framework in the case of limited labeled data.

4.3. Image Reconstruction

In this subsection, we investigate the reconstruction performance. On the STL-10 dataset, we compare reconstruction performance with the DeconvNet [21] which takes the MSE reconstruction loss as target function. As shown in Fig.4 (b), the AME-AE can accurately reconstruct the input images except for certain details even from the last (11th) layer of the encoder, while the images reconstructed by the DeconvNet even from the first layer are mottled and distorted and become worse with the increment of the depth of layer.

Furthermore, we investigate the robustness of representation to noises through reconstruction on the MNIST dataset.



(a) Reconstruction on the MNIST dataset.The odd columns are (corrupted) input images with 0%, 20%, 40%, 60%, 80% pixels dropped out from left to right, while the even columns are corresponding reconstructed ones.

(b) Reconstruction on the STL dataset. The first to the last rows present original images, reconstructed images from proposed AME-AE, reconstructed ones from the first and second layers of the DecovNet [21] in turn.

Fig. 4. Results of image reconstruction.

Fig. 5. Digits and images obtained by decoding the linearly interpolated representations. The left and right columns are original pairs and middle columns are interpolated ones.

We sample several handwritten digits and corrupt them by randomly setting some pixels to 0, which are then compressed and reconstructed through the encoder and decoder. According to Fig.4 (a), the reconstructed images are almost perfect even with 40% of pixels off. Note that there is not any corruption operations on input data during training phase. These results demonstrate that learned representations focus on the structure patterns rather than brightness factors of images and thereby are robust for the heavy noises.

4.4. Manifold Learning

We further explore and interpret the characteristic of learned representations from the perspective of manifold learning. First, we explore representation space through interpolation in representation space and reconstruct images from the interpolated representations. As shown in Fig. 5, digit 1 is smoothly transformed to digit 7, and a fighter under the blue sky is gradually transited to an airliner under gray sky. These semantic variations indicate that the encoder learns relevant high-level representations and the mapping between representation space and data space is sensitive to variations in representa-

tion space. On the other hand, the robustness of reconstructed images to noise, as shown in Fig.4 (a), reveals that the added noises are orthogonal to representation space. In summary, the mapping between representation space and data space is sensitive to variations in representation space but insensitive to noises orthogonal to representation space, which indicates the representations of AME-AE capture the manifold.

5. CONCLUSION

We have proposed a novel unsupervised learning framework that brings in competition and discrimination mechanisms. The joint training criteria of reconstruction and discrimination substantially enhances the ability of encoder to extract global high-level representations. Furthermore, we obtain a theoretical conclusion about the consistency of these training criteria. The effectiveness of learned representation is validated and analyzed through image classification and manifold learning. In the future, the framework could be further extended to a multi-modal one.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] David G Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE ICCV*. IEEE, 1999, vol. 2, pp. 1150–1157.

[3] Ross Girshick, "Fast R-CNN," in *Proc. IEEE ICCV*, 2015, pp. 1440–1448.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[6] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville, "Adversarially learned inference," in *Proc. ICLR*, 2017.

[7] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell, "Adversarial feature learning," in *Proc. ICLR*, 2017.

[8] Quoc V Le, "Building high-level features using large scale unsupervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8595–8598.

[9] Yoshua Bengio, Aaron C Courville, and James S Bergstra, "Unsupervised models of images by spike-and-slab rbms," in *Proc. ICML*, 2011, pp. 1145–1152.

[10] 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," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[11] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016.

[12] William Lotter, Gabriel Kreiman, and David Cox, "Unsupervised learning of visual structure using predictive generative networks," *arXiv preprint arXiv:1511.06380*, 2015.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[14] Andrea Vedaldi and Karel Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proc. ACM MM*. ACM, 2015, pp. 689–692.

[15] Adam Coates, Andrew Ng, and Honglak Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. AISTATS*, 2011, pp. 215–223.

[16] Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *NIPS*, 2011, pp. 1017–1025.

[17] Jiquan Ngiam, Zhenghao Chen, Sonia A Bhaskar, Pang W Koh, and Andrew Y Ng, "Sparse filtering," in *NIPS*, 2011, pp. 1125–1133.

[18] Adam Coates and Andrew Y Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *Proc. ICML*, 2011, pp. 921–928.

[19] Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng, "Deep learning of invariant features via simulated fixations in video," in *NIPS*, 2012, pp. 3203–3211.

[20] A. Romero, P. Radeva, and C. Gatta, "Meta-parameter free unsupervised sparse feature learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1716–1722, Aug 2015.

[21] Matthew D Zeiler, Graham W Taylor, and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. IEEE ICCV*. IEEE, 2011, pp. 2018–2025.