# Anomaly Detection for Images using Auto-Encoder based Sparse Representation

Qiang Zhao and Fakhri Karray

Center for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada
{qiang.zhao,karray}@uwaterloo.ca

**Abstract.** Anomaly detection is a pattern recognition task that aims to distinguish abnormal patterns from normal ones. In this paper, we propose a convolutional auto-encoder based model that aims to detect anomaly images by producing a sparse representation in the latent space. The proposed approach is able to represent the normal images using sparse encoding and the encoding can be well reconstructed by the decoder. However, the learned convolutional filters are not able to represent the abnormal images in a sparse way. Therefore, the decoder can not reconstruct the abnormal images with high quality. By assessing the reconstruction performance, we can distinguish the abnormal images from the normal ones.
The experimental results show the superiority of our proposed model over other variants of auto-encoder based anomaly detection models in terms of AUC. In addition, the result show that the sparse representations of normal images in a dimensional reduced space provide a better pattern recognition performance for image anomaly detection task.

**Keywords:** anomaly detection · auto-encoder · sparse representation.

## 1 Introduction

Anomaly detection is a pattern recognition task that aims to distinguish abnormal patterns from normal ones. However, some density based anomaly detection techniques such as clustering and nearest neighbor are based on the concept of local density that fail to apply to the data with high local fluctuations such as images. Based on that fact, more and more researchers have focused on exploring various generative models for pattern recognition purpose, for example, the variational auto-encoder[1] and the Generative Adversarial Networks (GANs)[2] provide the attractive alternatives to other maximum likelihood techniques.

Recently, anomaly detection has became an area of active research. Using deep learning methods for anomaly detection task has been extensively studied across a range of domains in [3]. The lack of generality to unseen anomalies is a major obstacle for anomaly detection task in images. Among these models, the auto-encoder inherent the feature extracting conception with dimensional reduction. In addition, what makes the auto-encoder a favorable model for image

anomaly detection is that it only depend on normal data(which is the common situation in anomaly detection tasks) during the training process. The proposed model exploring the potential of using auto-encoder based method to detect anomalies in images, that is to find a low-dimensional data representation in which normal images and abnormal images are expected to have significantly different expressions that could be leveraged to quantitative measurements.

When considering the architecture of the auto-encoder, the convolution networks are preferred. Over the last decade, deep convolutional neural networks have achieved remarkable success on various computer vision tasks such as image classification and object detection. On the one hand, the characteristics of the parameter sharing and local connectivity of the convolutional layers make it sensitive to unpredictable changes in local area of images. On the other hand, theoretical research[4] demonstrate that the hierarchical architectures achieve good performance on universal visual pattern recognition tasks. Therefore, a hierarchical convolution network is developed as a pattern encoder to obtain a robust latent representation for normal images.

The deep layered networks have a common problem of getting stuck in poor solution with random initialization. Especially, the encoding layers of the auto-encoder are developed to preserve quantity of information rather than quality of information. Thus, auto-encoder may lose critical information that most relevant to the image anomaly detection task. To address that problem, the related work[5] corrupting the particular feature and dividing the input data into the effective reconstructed part and the noise part. This work inspired by robust principal component analysis that fail to meet our expectation on generating the sparse feature representation in the latent space.

Another common challenge of using auto-encoder for feature representation learning in high dimensional data is to guarantee a robust reusable representation of the images in the latent space. Even further, the auto-encoder is trained to restore as much information as possible and not be able to determine what kind of information is relevant to the specific problem we are trying to solve during training process. In this case, another obstacle of using the auto-encoder method for image anomaly detection task is to avoid producing a perfect copy of input as output. Based on that fact, the problem of image anomaly detection is transferred to explore a criteria that could train the auto-encoder to obtain a latent representation that can effectively distinguish the abnormal images from normal ones. The related stacking denoising autoencoder[6] addresses this problem by forcing the auto-encoder to denoise the corrupted input images and [7] further improve the robustness of the feature representations by performing a layer-wise initialization and partially corrupting the input. However, both work take the risks of destroying the original normal patterns of images and increase the computational complexity for training the model. A novel method is proposed that care about the information loss between the normal images and their latent representations without corrupting the original features.

Consider that the normal images are supposed to follow the consistent sparse representation in the latent space, that is, the feature representation of normal

images always satisfy the sparsity condition. We introduce the sparse coding to train the encoder learn a sparse representation in the latent space. Previous work like [8] apply sparse coding to detect the unusual event of the videos in a real time basis by continuously updating the learned dictionary. Also, the abnormality of the image is determined by proposed sparse reconstruction cost in [9]. However, both work only use sparse coding as a measurement of evaluating the normality of event rather than a standard criteria for representation learning.

Our work bridges the gap between the success of convolutional neural network for image feature learning and the consistency of the auto-encoder in learning a sparse representation in the latent space. By evaluating the reconstruction performance of each image, we can effectively distinguish the abnormal image from the normal one.

The rest of the paper is organized as follows: Sect. 2. introduces the proposed approach. The details of the experiment setup and results are described in Sect. 3. Finally, the conclusion is presented in Sect. 4.

## 2 Proposed Approach

Anomaly detection in images by sparse representation is based on the assumption that the abnormal images will show inconsistent embedded behavior such as lie separate from the normal samples in the latent space. Based on that assumption, we assume that the abnormal images can not reconstructed from the sparse representation with high quality. Thus, the reconstruction loss of normal images and abnormal images will have large differences and it is reasonable to treat the reconstruction loss as anomaly score for detecting the abnormal images.

### 2.1 Sparse Representation

Take the advantage of auto-encoder based model that it is not sensitive to the specific texture of the anomaly images. The proposed model benefits from[10] and extends the convolutional auto-encoder based model to learn the sparse representations that used to reconstruction the normal images.

The traditional training process of auto-encoder is to minimize the pixel-wise independent mean square error with respect to encoder parameters $\Theta$ and the decoder parameters $\Phi$, given by:

$$min_{\theta,\phi} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - g_{\Phi}(f_{\Theta}(x_i))||_2^2 \tag{1}$$

, where N is the number of training samples, $x_i$ is each normal image, $\Theta$ is the parameters learned during the convolutional mapping in encoder $f_{\Theta}(x)$ and the $\Phi$ is the parameters learned during the deterministic reconstruction process in decoder $g_{\Phi}(x)$.

A explicit criteria is proposed to force the model to learn the identity sparse representations of the normal images in the latent space. The formulated sparse

representation in the latent space is obtained by adding a constrained penalty term, that is, instead of using mean square error as loss function to optimize the reconstruction performance, the L1 norm constrained is added on the intermediate representations that trained to produce a sparse feature representation in the latent space. The novelty of the proposed model is based on the belief that the sparsity of the latent representation is the key to improve the accuracy of image anomaly detection task. The trade-off here is to force the model maintain the sparse representation required to reconstruct the input without holding redundant noise features. Also, the encoding convolutional filters are trained to produce a much divergent set of solutions that sensitive to abnormal patterns. The very first intuition of adding L1 norm as a regularisation term for training is to prevent overfitting and improve the generalisation ability of the model. But in this case, we introduce the idea of sparse coding as a "regularizer" to punish the complexity of the intermediate representations in the latent space. Therefore, by adding a parameterized L1 norm constrain on images' latent representations, the loss function is formulated that aims to train the proposed model both sensitive to the reconstruction performance and the representation sparsity, given by:

$$min_{\theta,\phi} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - g_{\Phi}(f_{\Theta}(x_i))||_2^2 + \lambda||f_{\Theta}(x_i)||_1 \tag{2}$$

, where the $\lambda$ is the tuned parameter that controls the degree of sparsity of the latent representation.

### 2.2 Anomaly Score

From anomaly detection perspective, the anomaly score evaluates how well a given test image follows the normal patterns. In the evaluating stage, we assume that the concatenated convolutional filters trained with normal images are not able to represent the anomaly images in a sparse way. Therefore, the decoder can not reconstruct the abnormal images with high quality. By comparing the reconstruction performance, we can distinguish the abnormal images from the normal ones. Based on above discussion, the reconstruction error (1) is adopted as the anomaly score to assess the abnormality extent of the test images.

## 3 Experiment

We test the effectiveness of our model on two well-known image anomaly detection scenarios respectively. Also, to emphasize the effectiveness of the proposed model, we compare it with several variants of auto-encoder models in terms of AUC score in both scenarios.

As shown in Fig.1 that the encoder has four successive convolutional sections. The proposed model's architectural based on the several developed changes to the convolutional neural networks. [11] suggests that replacing the downsampling layer with consequent strided convolutional layer will not loss accuracy for feature selection. And each convolutional layer followed by the batch

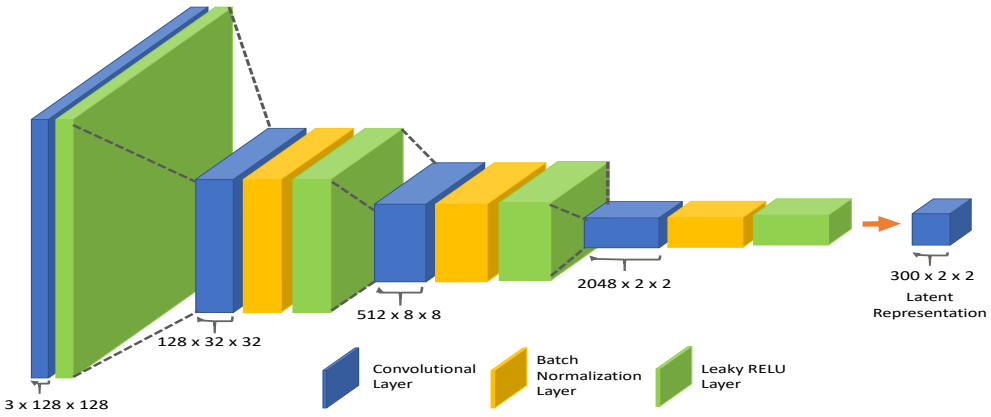



**Fig. 1.** The architecture of the encoder part of the proposed convolutional auto-encoder.

normalization layer that helps to stabilize the distribution of each hidden unit during training. In particular, the batch normalization[12] address the problem of decaying learning rate in a deep neural network and therefore accelerate the training process with less care to put on the model initialization. However, to prevent the model oscillation caused by applying batch normalization to every convolutional layer, we omit the batch normalization at the input layer of encoder and the output layer of decoder as [13] suggests. The last convolutional layer in encoder and decoder is sent to the Sigmoid function and the Tanh function respectively. Furthermore, [14] suggest that a non-zero slope for negative part in rectified linear unit could preserve the information when data transfer through deep layers. Therefore, we adopt the leaky RELU activation layer after each convolutional section and we found it improve the performance of proposed model when training epoch is relatively small.

### 3.1 Data set

**The HAM10000 data set** The HAM10000 data set is a multi-source dermatoscopic images that consists several pigmented skin lesions[15]. The detail information of the data set shown in Table 1. We separate the normal skin images(labeled as "NV") into 100 test set and 6605 training set. During the testing stage, we randomly select 100 images from each class of diseases to calculate the anomaly score using (1) and compare it with the anomaly score of the images in normal test set.

**Daytime Driving Distraction data set** The daytime distraction driving image data set[16] contains three different distraction behaviors (talking, texting and focusing on the GPS near the gear stick of the simulation while driving) as

**Table 1.** The HAM10000 data set:

| **Disease** | ***NV*** | ***AKIEC*** | ***BCC*** | ***BKL*** | ***DF*** | ***MEL*** | ***VASC*** |
|---|---|---|---|---|---|---|---|
| Number | 6705 | 327 | 514 | 1099 | 115 | 1113 | 142 |

well as normal driving behavior. The detail information of the data set shown in Table 2 and some images are deleted due to the unexpected disturbances occur during the image collecting process. The integral view of 25 drivers' upper body movements is captured while they are driving in a simulated environment. For training the model, we randomly pick 16 drivers' normal driving images and testing on the distraction driving images from the remaining 9 drivers.

**Table 2.** Daytime Distraction Driving data set:

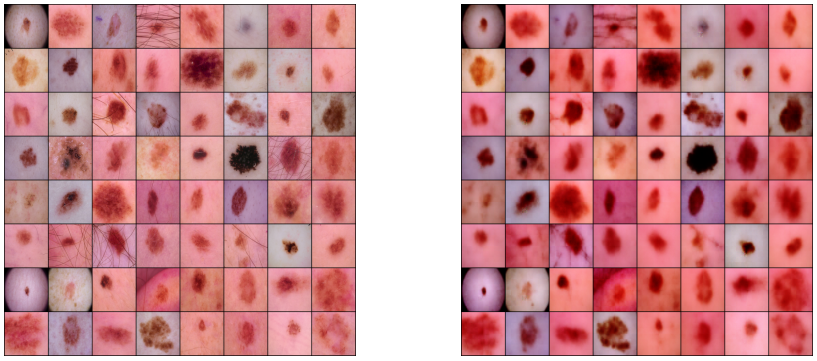
| **Behavior** | ***Normal*** | ***Talking*** | ***Texting*** | ***GPS*** |
|---|---|---|---|---|
| Number | 4993 | 4921 | 4991 | 4926 |

### 3.2 Experiment Setting

As shown in Fig. 1, the encoder part of the proposed model have four convolutional blocks. Similarly, the decoder has four symmetrical deconvolutional blocks. During the training stage, we feed the model with normal samples $\{x_1, x_2, ..., x_n\}$ for each $x \in R^{d \times d}$ that we considered to be i.i.d. sample from the unknown prior distribution.

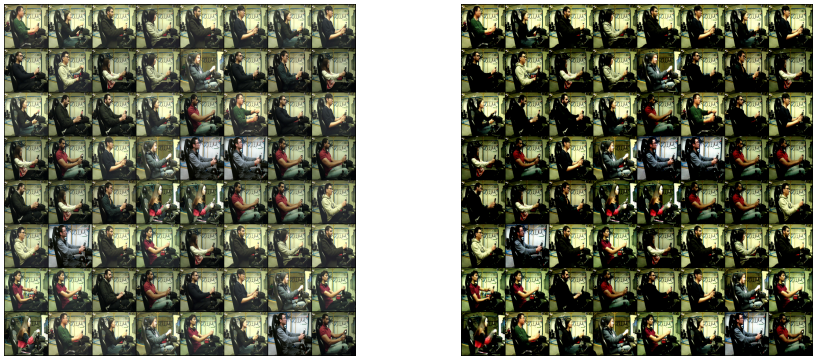
To show the competent representation learning ability of the proposed model, we compare its performance with several variants of auto-encoder based models, including the baseline auto-encoder model trained by binary cross entropy loss, the VAE[1] and the baseline model without doing sparse representation. In order to maintaining the reliability of the results, the architecture of each model is constructed to have similar convolutional blocks as our proposed model. We set the prior to be gaussian distribution during training process of the VAE in both scenarios.

- The same data pre-processing is performed for both data sets, that is to transform each image to the size of 128 x 128 and normalize each channel of images to the range of -1 to 1.
- The dimension of the latent space is set to 300 and the sparse representation parameter $\lambda$ is set to 0.002 for HAM10000 data and 4e-4 for driving distraction data.
- The batch size for training both data sets is set to 64 and the model is trained for 70 epoch each run, optimized with the Adam optimizer[17], the learning rate is set to 1e-4 with the weight decay of 1e-5.



(a) original image (b) reconstructed image

**Fig. 2.** The HAM10000 images and its corresponding reconstructed images



(a) original image (b) reconstructed image

**Fig. 3.** The driving distraction images and its corresponding reconstructed images

### 3.3 Experiment Results

The original images and its corresponding reconstructed ones are shown in Fig.2 and Fig.3. As we can see, the reconstructed image is blurred but also highlight the object and in a sense clear out the interference from other backgrounds.

As shown in Table 3 and Table 4, our proposed model achieves overall better results compare to the list variants of auto-encoder based models in both data sets. Table 3 shows the proposed model with sparse encoding improves the anomaly detection performance on skin disease data, especially for those types of anomaly diseases that perform relatively poor in the base model. For example, the proposed model greatly improve the detection of "AKI" disease from 0.59 AUC to 0.80 AUC and "VAS" disease from 0.48 AUC to 0.66 AUC. This also illustrates that the proposed model could effectively represent the normal images in a sparse way. As a result, the proposed model trained with normal images can not represent the abnormal images in a sparse way. Therefore, the anomaly images can not be well reconstructed from the latent sparse represen-

tation and result in large reconstruction error. In this way, the proposed model performs well on detecting any kind of anomaly case. Same with distraction driving scenario, the proposed model relatively improve the detection performance on talking and texting anomaly images.

**Table 3.** The AUC Results of the HAM10000 data set.

| | AUC | | | | | | |
|---|---|---|---|---|---|---|---|
| | **MEL** | **BCC** | **AKI** | **BKL** | **DF** | **VAS** | **ALL** |
| **Baseline** | 0.60 | 0.57 | 0.59 | 0.71 | 0.57 | 0.48 | 0.59 |
| **VAE** | 0.78 | 0.57 | 0.68 | 0.60 | 0.56 | 0.59 | 0.63 |
| **CAE** | 0.80 | 0.65 | 0.76 | 0.69 | 0.60 | 0.60 | 0.68 |
| **CAE + sparse** | 0.79 | 0.74 | 0.78 | 0.70 | 0.65 | 0.66 | 0.72 |

**Table 4.** The AUC Results of the Driving Distraction data set.

| | AUC | | | |
|---|---|---|---|---|
| | **TALK** | **TEXT** | **GPS** | **ALL** |
| **Baseline** | 0.56 | 0.67 | 0.63 | 0.62 |
| **VAE** | 0.61 | 0.64 | 0.75 | 0.67 |
| **CAE** | 0.61 | 0.66 | 0.82 | 0.70 |
| **CAE + sparse** | 0.64 | 0.70 | 0.83 | 0.72 |

## 4 Conclusion

In this paper, we attempt to answer the question: what makes a good representation for anomaly detection task? We explore the potential of combining the sparse coding with the auto-encoder and detect anomaly images by representing the images in a sparse way in the latent space. The experiment results show that our work have a better overall performance than other variants of the auto-encoder based models on image anomaly detection task. Take the advantage of auto-encoder based model that only depend on normal images for training, the proposed model can apply to a quite common situation when only few labeled normal images is available but large amount of unlabeled data still left to deal with. Also, the proposed model can potentially be extended to other anomaly detection tasks such as the abnormal event detection in videos. The future work should focus on obtaining the identical sparse representation that not sensitive to the possible changes occuring in normal images.

## References


1. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. The 2nd International Conference on Learning Representations (ICLR) (2014)
2. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. NIPS (2014)
3. Chalapathy, R., Chawla,S.: Deep Learning for Anomaly Detection: A Survey. arXiv preprint arXiv:1901.03407 ( 2019)
4. Fukushima, K.: Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural networks, vol. 1, pp. 119-130 (1988)
5. Zhou, C., Paenroth, R.C.: Anomaly Detection with Robust Deep Autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 665–674 (2017)
6. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. The Journal of Machine Learning Research, pp. 3371–3408 (2010)
7. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and Composing Robust Features with Denoising Autoencoders. The 25th International Conference on Machine Learning, pp. 1096–1103 (2008)
8. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In Computer vision and pattern recognition (CVPR), pp. 3313–3320 (2011)
9. Cong, Y., Yuan, J., Liu, J.: Sparse Reconstruction Cost for Abnormal Event Detection. In Computer vision and pattern recognition (CVPR), pp. 3449-3456 (2011)
10. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. Advances in neural information processing systems(NIPS), pp. 801–808 (2007)
11. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
12. Ioffe, s., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv preprint arXiv:1502.03167 (2015)
13. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
14. Xu, B., Wang, N., Chen, T., Li, M.: Empirical Evaluation of Rectified Activations in Convolution Network. arXiv preprint arXiv:1505.00853 (2015)
15. Philipp, T.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions (2018) [Online]. Available: https://doi.org/10.7910/DVN/DBW86T
16. Ou, C., Zhao, Q., Karray, F., Khatib, A.E.: Design of an End-to-End Dual Mode Driver Distraction Detection System. In: Campilho, A., Karray, F., Yu, A. (eds.) ICIAR 2019, LNCS, vol. 11663, pp. 199-207. Springer, Cham (2019)
17. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)