Auto-Encoder based Deep Representation Model for Image Anomaly Detection

by

Qiang Zhao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

In this dissertation work, two related papers were published:

1. **Q. Zhao**, and F. Karray, "Anomaly Detection for Images using Auto-Encoder based Sparse Representation," in *17th International Conference on Image Analysis and Recognition*, Jun. 2020.

2. C. Ou, **Q. Zhao**, F. Karray and A. E. Khatib, "Design of an End-to-End Dual Mode Driver Distraction Detection System," in *16th International Conference on Image Analysis and Recognition*, Aug. 2019.

In paper 1, I was responsible for the literature review, data processing, building the model, and running all the experiments included in this paper. The daytime driving distraction dataset that paper 1 use comes from paper 2. Part of the work in Section 3.1.1 is from paper 1. For paper 2, I was responsible for data collection, data pre-processing, building, and running part of deep learning models and then assembling the results. Dr. Chaojie Ou (a co-author of the article) was responsible for literature review, project coordination, and most of the deep learning models included in this paper. One of the datasets of this dissertation work comes from paper 2.

Abstract

Image anomaly detection is to distinguish a small portion of images that are different from the user-defined normal ones. In this work, we focus on auto-encoders based anomaly detection models, which assess the probability of anomaly by measuring reconstruction errors. One of the critical steps in image anomaly detection is to extract robust and distinguishable representations that could separate abnormal patterns from normal ones. However, current auto-encoder based methods fail to extract such distinguishable representations because their optimization objectives are not tailored for this specific task. Besides, the architectures of those models are unable to capture features that are robust to irrelevant distortions but sensitive to abnormal patterns.

In this work, two auto-encoder based models are proposed to address the aforementioned issues in optimization objectives and model architectures, respectively. The first model learns to extract distinct representations for abnormal patterns by imposing sparse regularizations on the latent space during the optimization process. This sparse regularization makes the extracted abnormal features unable to be represented as sparse as the normal ones. The second model detects abnormal patterns using Asymmetric Convolution Blocks (ACB), which strengthens the crisscross part of the convolutional kernel, making the extracted features less sensitive to geometric transformations.

The experimental results demonstrate the superiority of both proposed models over other auto-encoder based anomaly detection models on popular datasets. The proposed methods could also be easily incorporated into most anomaly detection methods in a plug-and-play manner.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Fakhri Karray, for his continuous guidance and patience throughout my master's study. Under his supervision, I utilized pattern recognition techniques to solve real-life problems. His professional suggestions help me to conduct in-depth researches, and his generous advice makes this work possible. It is a great honor for me to complete my study as one of his students.

I also want to express my thanks to the committee members, Dr. Kumaraswamy Ponnambalam and Dr. Andrew Morton, for taking their time to read my thesis and provide me with professional suggestions.

I would like to express my gratitude to the financial support from the Electrical and Computer Engineering department of the University of Waterloo and Natural Sciences and Engineering Research Council of Canada.

Last but not least, I would like to express my sincere gratitude to my family for providing their selfless support and love in the way of pursuing my master's degree. I would like to express my deepest appreciation for their continuous sacrifices through good and bad times. I would not have been able to complete this work without the support of my family.

Dedication

This dissertation work is a special dedication to my family, a sincere gratitude to my mom and dad who have always encouraged me through the hard times. I hope that this accomplishment will make you all proud.

Table of Contents

List of Figures

List of Tables

Abbreviations

ACB Asymmetric Convolution Blocks iv, viii, x, xi, 5, 6, 15, 21, 23–25, 35–42

AUC Area Under the Receiver Operating Characteristic Curve xii, 21, 30, 32, 33, 35, 36, 38, 39, 41

CAE Convolutional Auto-encoder 5, 15, 16, 23, 25, 29, 30, 33, 36, 38, 41

CNN Convolutional Neural Network 3, 4, 21–23, 31, 43

DBM Deep Boltzmann Machine 12

GDA Gaussian Discriminant Analysis 8

LDA Linear Discriminant Analysis 8

LLE Locally Linear Embedding 9

LPP Locality Preserving Projections 9

PCA Principal Component Analysis 8

RBM Restricted Boltzmann Machine 7, 11, 13

VAE Variational Auto-encoder xi, 17, 30, 33, 36–39, 41

Chapter 1

Introduction

In this chapter, the problem of anomaly detection is introduced first. Then, the challenges and drawbacks of current anomaly detection methods are pointed out. Finally, we demonstrate the motivations for using the two proposed methods.

1.1 Problem

Anomaly detection is a pattern recognition task that distinguishes abnormal patterns from the normal ones. Normal patterns are usually defined by humans in a specific use case and abnormal ones are samples that deviate significantly from the normal ones. Anomalies could be indicative of previously unknown mechanisms or could be generated by underlying processes of manual designs [43]. Anomalies are also referred to as deviants, outliers, irregularities, and abnormalities in the statistics literature [11]. The main difference between anomaly detection and traditional classification tasks is that the characteristics of anomalies are diverse and hard to summarize into simple rules. This requires anomaly detection

methods to capture abnormal patterns that do not appear in the model's development. However, classification tasks need clear definitions of all classes and cannot be generalized to unseen data.

Due to its robustness and simplicity in use that does not require human-labeled data, anomaly detection has been widely used in a variety of research fields and industrial applications in recent years.

1.2 Challenges

Detecting abnormal patterns in large scale and high dimensional data is still a big challenge because most methods heavily rely on handcraft features obtained from feature engineering, which is not comprehensive and may miss critical information for detecting anomalies. Feature engineering requires domain knowledge and is labor-intensive. The diversity and high dimensionality of real-life data also make it computationally expensive.

Except for this general challenge, there are many other specific ones for different anomaly detection methods. Most abnormal detection techniques are unsupervised such as distribution-based, clustering-based, depth-based, density-based, and distance-based. For distribution-based models, Otey et al. [32] presented a tunable algorithm for distributed anomaly detection in dynamic mixed-attribute datasets. For classification-based models, Kingdon [57] proposed an abnormal transaction detection method using a single-class support vector machine with a Radial Basis Function (RBF) kernel. In terms of clustering-based approaches, the work presented in [81] projects the customer's transaction onto the timeline to form a histogram. Anomaly detection is then performed on the clustered data according to the segmentation of this histogram. For entropy-based ones, Armin et al. [12]

present a semi-supervised anomaly detection algorithm to deal with extremely unbalanced data.

Although the methods mentioned above have advantages, they also have critical shortcomings. For density-based models such as k-nearest neighborhood, one disadvantage is that they rely on the local density assumption, which fails to apply to data having big local fluctuations such as natural images.

1.3 Motivations

Recent years have witnessed an unprecedented development of deep neural networks that achieve good performance learning the representations for high dimensional data. With the proliferation of diverse architectures of deep neural networks, deep learning surpasses traditional machine learning methods in many aspects. Speech translation, image classification, object recognition, and anomaly detection, are cases in point. As proposed in [91], using deep learning methods for detecting anomaly patterns has been extensively studied across a wide range of domains. The notion of feature re-use is one of the key advantages of deep learning, which is constructing hierarchical levels of features of the data. As shown in the works of [52] [114] and [110] deep representations are significantly more efficient than the ones that are insufficiently re-used. Also, a deep architecture could produce abstract representations. For example, the Convolutional Neural Network (CNN) proposed in [116] produces the abstract representations by the down-sampling mechanism.

Among these deep learning models, CNN has shown great promise in various computer vision tasks such as image classification and object detection. By taking advantage of the parameter sharing and local connectivity characteristics of convolution, CNN can capture geometrical transformation invariant features. Theoretical research [62] also demonstrates

that the hierarchical architectures of CNN help to capture both structural and semantic features. Therefore, a deep convolution network is a good candidate for extracting feature representations of high dimensional data such as images.

Among the CNN models, auto-encoders are good candidates for recognizing anomalies for high-dimensional data due to their dimension reduction capabilities. Besides, the training process of auto-encoder based models only requires normal samples, which are easily accessible. Collecting abnormal samples is labor-intensive and time-consuming due to the sporadic nature of abnormal samples and the need of advice from human experts.

However, auto-encoder is trained to reconstruct the data as close as possible to the original one and cannot pinpoint what kinds of features are distinguishable from the anomalies. Also, the auto-encode is designed to preserve the quantity rather than the quality of the information in the data. Therefore, the auto-encoder based models may lose information that is relevant to detecting the anomalies. To unravel the above two problems, we need to explore an optimization objective that makes the auto-encoder produce the feature representations that are robust for normal images but distinguishable for abnormal ones. The related work proposed in [86] and [87] try to solve the problems by de-noising the partially corrupted versions of input images. However, both works pay a heavy price by altering the original normal patterns of images in the quest of obtaining the feature representations. Besides, the partially corrupted input may be accidentally treated as anomalies by the model and the de-noising operation increases the computational complexity.

A possible solution for the disadvantage of auto-encoders is to incorporate entropy-based methods. In classical statistical mechanics, entropy is the measure of uncertainty or *mixedupness* according to Gibbs [22]. The entropy measures the degree of probability of the given system spreading out to all possible set of microstates[1]. This measure is governed

[1]A microstate is a set of microscopic configurations of a thermodynamic system.

by Gibbs entropy formula[2] shown in Equation 1.1, where k_B is known as Boltzmann's constant and p_i is the probability that the system is in the i-th microstate:

$$S = -k_B \sum_i p_i \ln p_i \tag{1.1}$$

However, since the macroscopic properties are hard to define for natural images, we instead seek help from the sparse coding theorem. As mentioned in [8], for most of the high dimensional data, only a small part of the underlying factors are relevant to the problem that the model aims to solve and the data could be represented by a vector with mostly zero initial values. From the data representation learning perspective, the auto-encoder and sparse coding can represent up to $\mathcal{O}(2^k)$[3] dimension of data by only $\mathcal{O}(N)$ parameters.

The first model proposed in Section 3.1.1 is based on a Convolutional Auto-encoder (CAE) and incorporates sparse coding strategies to detect anomalies in images. This method imposes sparse regularizations on the latent features of auto-encoder, instead of adding penalties which is unsuitable for natural images, such as the derivatives of the Jacobian matrix[4] of the features. As a result, the proposed method can obtain low-dimensional sparse feature representations for normal images while failing to do so for abnormal images. In this way, the proposed model can differentiate abnormal images from normal ones.

The second method proposed in Section 3.1.2 is also based on a CAE. The proposed model adopts ACB to strengthen the weights of the central part of the traditional convolution kernels, making the model invariant to geometric transformations such as rotation

[2]The entropy of this distribution is given by the Gibbs entropy formula that state the macroscopic of a system which is characterized by a distribution on the microstates

[3]k is the number of non-zero features in a representation vector

[4]Jacobian matrix of a vector-valued function is the matrix of its first-order partial derivatives.

and translation. In most anomaly detection tasks, these geometric transformations are usually considered noise and irrelevant to abnormal patterns. Traditional convolutional kernels fail to extract such robust features and may not be able to deal with unexpected geometric transformations.

1.4 Summary

This chapter introduces the problem of image anomaly detection, including the problem formulation, challenges, and the motivations of the two proposed methods. The first model learns to extract distinct representations for abnormal patterns by imposing sparse regularizations on the latent space. The second model detects abnormal patterns using ACB, making the extracted features less sensitive to geometric transformations.

Chapter 2

Related Work

The difficulty of establishing an appropriate target for training the model inspired researchers to find innovative ways to obtain optimal feature representations. The work presented in this chapter is consists of three aspects:

- Global feature learning methods and their extensions.
- Manifold learning methods that are motivated by the geometrical structure of the data.
- Some salient approaches of deep learning representation:
    1. Inference learning of probabilistic models, including the indirected types such as Restricted Boltzmann Machine (RBM) and the directed types such as sparse coding.
    2. Reconstruction based approaches such as auto-encoders.

2.1 Global Feature Learning

The most widely used unsupervised feature learning method Principal Component Analysis (PCA) [89] aims at projecting the high dimensional data to a low dimensional feature space. PCA has been widely used as a dimension reduction tool [100] because of its simplicity. Linear Discriminant Analysis (LDA) is another well known supervised global feature learning method LDA [33]. PCA and LDA have a number of extensions in later works [15] and [98]. There are several versions of the PCA: the kernel PCA [98], the probabilistic PCA [31], the probabilistic relational PCA [71], the sparse PCA [123], the GPLVM [69] [105] and so on.

LDA is a linear feature learning method which could obtain new features used for recognition tasks [16]. Another related work proposed in [100] learned features from gray-scale images and applied the features to a supervised classifier to recognize anomalies. Gaussian Discriminant Analysis (GDA) is an alternative method to analyze the data distribution. However, the generalized GDA is not the optimal solution to the original trace ratio problem[1] as mentioned in [106]. Therefore, the Newton-Raphson method and relational Fisher analysis are introduced to GDA as mentioned in [58] and [121] respectively.

2.2 Manifold Learning

The manifold learning approaches are geometrically based methods that aim at discovering the embedded structure of the high dimensional data. For example, the Isomap has been proposed in [104] using Floyd-Warshall algorithm ([34]) to compute pair-wise distances

[1]The ratio problem is generally in the form of $Tr(W^T S_a W)/Tr(W^T S_b W)$, W is the desired transformation matrix, S_a and S_b are constant positive semidefinite matrices.

among data points. And the Locally Linear Embedding (LLE) proposed in [94] encodes the data information with local neighborhoods. Inspired by the LLE, the laplacian eigenmaps method is proposed in [17] that based on the relationship between the Laplace Beltrami operator and graph laplacian. Another linear version of the Laplacian eigenmaps algorithm is proposed in [45], the Locality Preserving Projections (LPP) could be used to obtain the low dimensional features in recognition tasks [46]. Also, the local tangent space alignment method that represents the manifold by tangent space is proposed in [118]. Furthermore, the work proposed in [120] combined the Laplacian eigenmaps algorithm with a local tangent space alignment method that computes the similarity of data in tangent space and also uses the Laplacian eigenmaps algorithm to learn data embedding. The success of manifold learning methods in applications involving shape based recognition as mentioned in [21]. But manifold learning methods fail to apply to the data without low dimensional manifolds.

2.3 Deep Representations

As mentioned in [70] and [99], the development of the deep learning theory brings research interests to a wide range of domains. The breakthrough of deep learning in feature representation field is extensively discussed in [107] and [19]. The key point proposed in [30] [77] [113] and [42] indicates that the abstract representation of data is composed by the hierarchy learned features from previous layers. In particular, there are several ways to evaluate the quality of the representations learned from deep neural networks. For example, the work proposed in [23] measures the quality of representations by classification error, the work in [50] evaluates the representation quality by the invariance properties of the learned features, and as mentioned in [93] the representation ability of the model depends on the

quality of the samples generated from the proposed model.

Another alternative approach to train the deep neural networks is proposed in [54], where it jointly trains every layer without explicit latent variables and iteratively constructs the free energy function to achieve better results. However, the problem is how to define the objective criteria for training the model with a free energy function? There are many options proposed in the following work: the hybrid Monte Carlo[2] proposed in [76], the score matching methods proposed in [4] and [6], the ratio matching methods as the extension of the score matching proposed in [5], the contrastive divergence proposed in [29] and the noise contrastive estimation proposed in [74].

2.3.1 Probabilistic Models

From an inference learning perspective, the feature learning can be regarded as finding the joint probability distribution over the observed data. And the training process can be interpreted as obtaining the parameters that maximize the likelihood of the joint distribution of latent variables when giving the existing data samples. There are two paradigms of the probabilistic models, one is direct and the other is indirect.

The theory behind the direct models are shown in Equation 2.1, which is adapted from relevant research work as proposed in [75] [9] [31] and [49].

$$p(x, h) = p(h)\, p(x|h) \tag{2.1}$$

[2]The Monte Carlo method is a computational algorithm that relies on repeated random sampling to obtain numerical results, that is to use randomness to solve problems that might be deterministic in principle.

where $p(x|h)$ is the conditional likelihood, $p(h)$ is the prior to construct joint distribution of $p(x, h)$.

In particular, the sparse coding is the most popular research areas of the probabilistic models. From the indirect inference learning perspective, the sparse coding can be regarded as restoring the feature vector of the input x from Equation 2.2:

$$f(x) = h^* = \arg\min_h ||x - Wh||_2^2 + \lambda||h||_1 \tag{2.2}$$

where h is the representation coding and the W is the matrix of mapping.

From this interpretation, learning dictionaries in sparse coding is a process of maximizing the data likelihood when only given the "code" (h^*). The advantages of the sparse coding, as proposed in the work of [2] and [7] significantly outperforms other encoding schemes in object classification tasks. Comparing with general probabilistic learning models, the computationally efficient property of the sparse coding makes it one of the popular feature encoding methods. The related applications including the visual cortex proposed in [9], the image modelling mentioned in [2] [66] [63], the natural language processing proposed in [1] and audio recognition mentioned in [92]. Also, the sparsity training criteria can penalize the active features when other features are relatively small. As a result, the generalized features are close to zero.

As for the indirect probabilistic models, the RBM [84] is the most popular one. In the last few years, several improvements have been proposed to help RBM better capture the real value data. The model proposed in [79] has been used to synthesize natural images and [53] has been used to model the natural textures. To improve the ability of the RBM model to learn the statistical information of natural images, [78] introduced two statistical concepts (mean and covariance) to the RBM. In addition, the trained layers

of deep generative models could be combined with Deep Boltzmann Machine (DBM) as proposed in [93]. However, it is unclear how to train the generative model to approximate the maximum likelihood of the data. The algorithm proposed in [30] introduced one potential way to solve this problem.

2.3.2 Reconstruction based model

Since the inference learning in probabilistic models is always associated with the latent variables, an alternative non-probabilistic approach that directly parametrizes the feature representations is the auto-encoder framework proposed in [39]. The general auto-encoder training criteria is to find parameter sets Φ, Θ that minimize the reconstruction error as shown in Equation 2.3. In case the input data have a binary nature, the binary cross-entropy loss as shown in Equation 2.4 is adopted.

$$\jmath_{AE}(\Phi, \Theta) = \sum_{i} L(x^i, g_\Phi(f_\Theta(x^i))) \tag{2.3}$$

where x_i is each normal image, $f_\Theta(x)$ represents the encoding process and $g_\Phi(x)$ represents reconstruction process in decoder.

$$L(x^i, x^j) = -\sum_{i=1}^{d_x} x^i \log(x^j) + (1 - x^i) \log(1 - x^j) \tag{2.4}$$

where x_i and x_j are a pair of data samples.

The reconstruction mechanism of auto-encoder also indicates the obstacle of obtaining a good generalization ability of the auto-encoder during the training process. This is achieved by various methods in the different forms of auto-encoder. For example, a study presented in [113] explores the possibility of building deep networks by using auto-encoder

rather than RBM for training. And another comparative study of the gradient of the auto-encoder reconstruction error is presented in [109].

Based on this fact, recent research has made "constraints" on the latent representations to form the so-called regularized auto-encoder. The effect of the regularization to the auto-encoder is that it cannot reconstruct everything well, even though it performs well in training samples, it will ultimately fail in generalizing test samples. By assuming that only a few configurations of the input are needed, the work proposed in [80] uses a sparsity penalty to avoid the magnitude of the hidden configurations. The sparsity penalty can be used in hidden unit as proposed in [77] [42] [40] [50] or the product of the hidden unit activation as shown in [80] [117] [90]. However, a comprehensive analysis about taking risks of penalizing the output of the hidden unit to compensate the data variants is lacking.

An alternative regularized auto-encoder which is designed to improve the robust ability of the model is called the denoise auto-encoder. As proposed in [87] and [86], the denoise auto-encoder aims to reconstruct the original input with the latent representation that learned from the synthetical corrupted input. As shown in Equation 2.5, the denoise auto-encoder is trained to capture the underlying explanatory factors of the data and optimally cancel the impact of the corruption process on the joint data distribution.

$$\jmath_{DAE} = \sum_{i} E_{q(\tilde{x}|x^i)}\, L(x^i, g_\Phi(f_\Theta(\tilde{x}))) \tag{2.5}$$

where $\tilde{x}$ is the corrupted version of the data sample x^i, $E_{q(\tilde{x}|x^i)}$ take average of all corrupted samples generated form the $q(\tilde{x}|x^i)$ corrupted process. Work done in [35] further generalizes the denoise auto-encoder and the denoise auto-encoder with small Gaussian noise could estimate the derivative of data log-density. The features learned by denoise auto-encoder improved the performance of classification compared to the features learned from RBM as

presented in [73]. Also, the work proposed in [65] has shown the advantages of applying denoising to efficiently implement the score matching.

As with the denoise auto-encoder, the contractive auto-encoder proposed in [97] is also designed for training the robust representations. The contractive auto-encoder achieves this capability by penalizing the variation of the input as shown in Equation 2.6.

$$\jmath_{CAE} = \sum_{i} L(x^i, g_\Phi(f_\Theta(\tilde{x}))) + \lambda ||J(x^i)||_F^2 \tag{2.6}$$

where λ is the parameter that controls the extent of regularization and $||J(x^i)||_F^2$ is the Frobenius norm of the Jacobian Matrix[3] of an encoder.

The connection between denoise auto-encoder and contractive auto-encoder is studied in [35]. Compared to denoise auto-encoder, the contractive auto-encoder has several advantages: it introduces an analytic penalty rather than a random one and introduces the λ that controls the extent of the regularization. As mentioned in [36], the denoise auto-encoder and the contractive auto-encoder both win the final stage of the transfer learning contest.

2.4 Summary

This chapter describes the related work of three type of feature learning methods. Besides the global feature learning and the manifold learning techniques, the deep representation methods are among the most important approaches to deal with the issues raised in this work. The proposed models in Chapter 3 are based on the theoretical development of deep representation models.

[3]The Jacobian Matrix is a matrix contains the first-order partial derivatives of a vector-valued function.

Chapter 3

Proposed Method

In this chapter, two CAE based methods are proposed in Section 3.1.1 and 3.1.2 with detailed descriptions of the drawbacks of existing methods and corresponding solutions. The first model described in Section 3.1.1 learns to extract distinct representations for abnormal patterns by imposing sparse regularizations on the latent space during the optimization process. The second model described in Section 3.1.2 detects abnormal patterns using ACB, which could extract features less sensitive to geometric transformations.

3.1 Proposed Models

3.1.1 Auto-Encoder with Sparse Representation

Problem Formulation

As mentioned in [20], one of the common challenges for feature representation by using an auto-encoder is to produce a robust representation of the data in the latent space. This

is a tough problem for high dimensional data with different local features such as natural images since auto-encoders are designed to reconstruct the input rather than detecting abnormal patterns.

Based on this premise, the key point of image anomaly detection is to explore the optimization strategies of the auto-encoder to produce a robust latent representation that can effectively detect abnormal patterns in images. The most common issue for training auto-encoders is that they will end up producing output that is the same as the input. The work in [86] addresses this problem by training the model with corrupted versions of input images, but still aiming at reconstructing the pristine input images. Another work in [87] performs layer-wise initialization with partially corrupted input and produces a robust feature representation. This method further improves the robustness of the feature representation of the data. However, it is worth noting that corrupted images may destroy the original normal patterns and adding uncertainty in anomaly detection. Therefore, the above two methods fail to meet the expectation of obtaining a robust representation of images. So we need to explore an alternative way to achieve this goal.

We presume that the normal patterns of images are supposed to have consistent feature encoding in learned latent feature spaces. According to the sparse coding theorem, the normal patterns have low entropy and they could be encoded into sparse representations. However, for abnormal patterns, the entropy will be high and the corresponding representations will not be as sparse as the normal ones. Inspired by the sparse coding theorem, sparse regularization could be imposed on the CAE to produce the sparse representations for normal images and non-sparse ones for abnormal images. So we can detect the abnormal images by measuring the sparsity of representations together with the reconstruction loss.

Previous work in [61] incorporates sparse coding in a different way. It first encodes the

images using a pre-trained VAE [27]. Then it follows the traditional sparse coding strategy that encoding the features with a learned dictionary. Finally, sparsity is measured based on the encoded coefficients. The drawbacks of this model are two fold. First, the model is not end-to-end optimized, making the extracted features sub-optimal for the anomaly detection task. Second, pre-defined thresholds are needed to control the model's sensitivity, which needs manual tweaking for different datasets.

Another work in [14] aims at solving the real-time anomaly event detection task. This method adopts the dynamic sparse coding approach, which could capture the possible concepts drift in video contents and update the event dictionary continuously based on the ability to reconstruct the query segment signals in an online fashion. Then the abnormality of the event is computed based on the sparse reconstruction cost proposed in [115]. However, this model does not include sparsity regularization into its training process. Therefore, the extracted representations may not be optimal to be applied with the dynamic sparse coding method.

However, different from the above two models, the model proposed in this section is end-to-end optimized. By incorporating the sparsity regularization into the optimization objective, our model can extract features more optimal for measuring sparsity. Also, our model does not require any pre-defined thresholds, which makes it easier to be applied to different datasets.

Proposed Solution

The abnormal images can not be well represented by the auto-encoder trained on normal images with the sparse regularization. Therefore, the abnormal images can not be well reconstructed from the latent representations with high fidelity. So the reconstruction loss

is used as the anomaly score for distinguishing the abnormal images.

The traditional criterion for training the auto-encoder is to minimize the reconstruction loss (mean square error), given by:

$$\underset{\Theta,\Phi}{L_{res}} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - g_\Phi(f_\Theta(x_i))||_2^2 \tag{3.1}$$

where N is the number of normal images for training, x_i is a normal image. Θ is the learnable parameter set of the encoder $f_\Theta(\cdot)$ and Φ is the learnable parameter set of the decoder $g_\Phi(\cdot)$.

The objective function contains two terms. The first term is the reconstruction loss. We choose Mean Square Error as the reconstruction loss because of its simplicity. The other term is a sparse regularization term. As in [37], the L1 norm is introduced as the sparse regularization to control the sparsity of the latent representations. The L1 norm penalizes the complexity of the latent representations. Therefore, we use L1 norm as the sparse regularization term. By combining the two terms, the proposed model are not only sensitive to the reconstruction quality but also takes into account the sparsity of the latent representations. The overall loss function is given by:

$$\underset{\Theta,\Phi}{L_{sparse}} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - g_\Phi(f_\Theta(x_i))||_2^2 + \lambda||f_\Theta(x_i)||_1 \tag{3.2}$$

where λ is a balancing parameter.

The pseudo-code of the training and testing process of the proposed model is presented in **Algorithm** 1.

Algorithm 1: Auto-Encoder based Sparse Representation

Input: Normal images X_n, Abnormal images X_{ab}, sparsity parameter λ
Random initialization of the model
Encoder parameter set Θ, Decoder parameter set Φ
repeat for each epoch:
  foreach normal image $x_n \in X_n$:
    Loss $= ||x_n - g_\Phi(f_\Theta(x_n))||_2^2 + \lambda||f_\Theta(x_n)||_1$
  end for
end TRAINING
repeat for each test image:
  foreach test image $x_{ab} \in X_{ab}$:
    Anomaly Score $= ||x_{ab} - g_\Phi(f_\Theta(x_{ab}))||_2^2$
  end for
return AUC
end

As for the network architecture of the proposed model, we adopt the same architecture as in [119]. Figure 3.1 shows the encoder part of the proposed model. The encoder contains four successive convolution blocks: the first block only contains one convolutional layer and one ReLU layer, the next three blocks contain an extra batch normalization layer between the convolutional layer and the ReLU layer. The last convolution layer in the encoder is followed by a Sigmoid function instead of ReLU. The decoder architecture is symmetric to the encoder. The last deconvolution layer in the decoder is followed by Tanh function instead of ReLU. The encoder architecture of the proposed model benefits from several improvements developed in recent research work. The first improvement mentioned in [60]

indicates that consequent convolutional layers will not compromise accuracy for feature selection when compared to a down-sampling layer. The second improvement is the batch normalization technique introduced in [95]. Batch normalization layers help stabilize the distribution of each hidden unit in convolution blocks and reduce the randomness caused by the random initialization of the model. As suggested in [10], by omitting the batch normalization layer in the first block of the encoder and the last block of the decoder, we alleviate the model oscillation problem caused by batch normalization. The third improvement [13] suggests that introducing a non-zero slope to the negative part of the rectifier linear unit could preserve the information when data is passing through a concatenation structure of deep layers. Therefore, the leaky ReLU activation layer is added after each batch normalization layer.

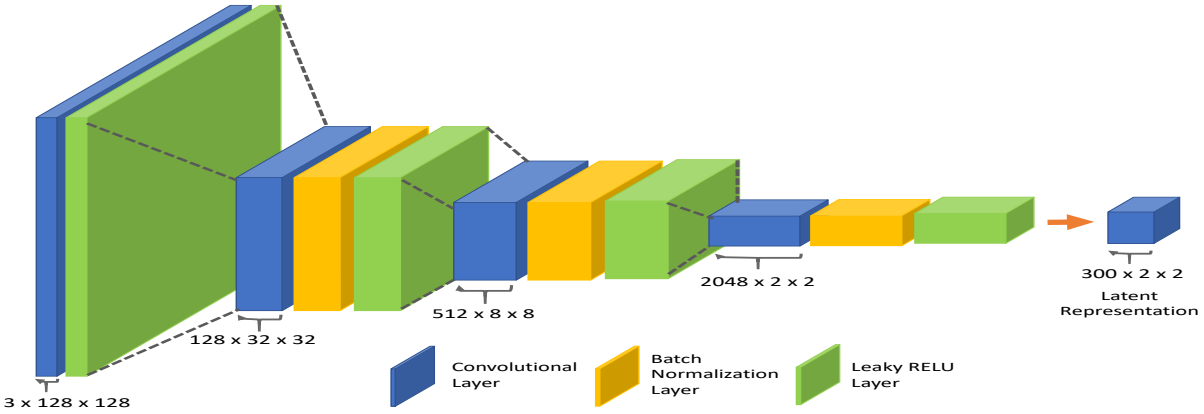


Figure 3.1: The architecture of the proposed convolutional encoder

Anomaly Score

The anomaly score is used to represent how well a given test image conforms to the normal patterns learned from the proposed model. Since the proposed model is not able to

reconstruct the anomaly images from the sparse representation with high fidelity, the abnormality of images is evaluated by the reconstruction loss, which is calculated by Equation 3.1. The AUC is then employed to demonstrate the effectiveness of the proposed method.

3.1.2 Auto-Encoder with ACB

Problem Formulation

The performance of CNN has been significantly enhanced in recent years. Recent researches are aiming at developing even more sophisticated architectures. Exploring a new customized architecture of the CNN requires a lot of work and numerous GPU hours [122]. There are three active lines of research about the architecture of the CNN. One pertains to building connections among layers and the other focuses on the combination of layer-wise outputs to improve the quality of learned representations. For the first kind, the related work in [48] enhances the feature propagation by concatenating the layer inputs from multiple directions. For the second kind, the works in [102] [103] [44] explicitly reformulate the identity mapping between layers.

Except for the above two kinds of models, another kind of approach is to incorporate the contextual information with multi-scale layer-wise representations, such as [18] and [82]. The work in [18] uses spatial recurrent neural networks to extract contextual information and skipping pooling layers to exploit multi-level abstraction representations. Moreover, the work in [82] developed hourglass modules to improve the model accurately. The superior performance is achieved by the intermediate supervision done by successive steps of down-sampling and up-sampling processes. Another work in [55] introduces the spatial attention to the network architecture and explicitly transforms feature maps within the network. The work in [47] explores the channel relationship among the convolution

operator and proposes a novel unit that can be stacked together to strengthen the representational power of the CNN.

However, due to the high computational cost, simply adding more adjustable parameters and building complicated connections among networks is not optimal for enhancing the model. The kernel method is becoming another active research topic in the CNN community. The work in [101] explores the relationship between the shape of convolution kernels and the learned representations. Another work in [64] explores how kernel size could influence the model performance and leads the trend of applying smaller kernels in deep CNN.

Proposed Solution

The redundancy of the weights of the convolutional kernels has been extensively studied through CNN literature. Previous works like [24] and [56] speed up the convolutional layers by equivalently transforming the standard two-dimensional $d \times d$ convolutional kernels to one-dimensional $d \times 1$ kernel and one-dimensional $1 \times d$ kernel. However, applying transformed low-rank kernels may lead to significant information loss as mentioned in [59]. The work in [24] tackled this problem by employing the matrix decomposition to find the low dimensional projections that could significantly decrease the number of fine-tuning parameters. The experiment results shown in [24] also suggest that the regularized low-rank approximations have improved the generalization ability of the model. The work in [56] tackled this problem using both filter reconstruction optimization technique and data reconstruction optimization technique to learn the horizontal and vertical kernels. The work proposed in [59] achieves two times speedup comparing to the baseline model with ten times fewer parameters by converting three-dimensional convolution to a sequence of one-dimensional convolution. Similarly, [72] employs the asymmetric convolution structure and

dense connectivity to increase the receptive field in a CNN. Moreover, the work proposed in [88] factorizes the square kernels to save computational cost. Most recent researches are working on developing information-preserving kernel transformations to enhance the model performance.

The proposed model is inspired by the work in [25] that explores the effectiveness of the kernel spatial locations for learning the feature representations. This work strengthens the performance of the CNN by replacing each standard convolution kernels with the ACB in every convolution layer of the model. The Figure 3.2 shows an example of the ACB architecture. One of the advantages of using ACB in the CAE is that it does not introduce extra parameters to the model, which makes it able to be simply incorporated into main-stream CNNs without extra computational cost. ACB also benefits from one of the important properties of the convolution operator shown in Equation 3.3. The equation indicates that if several two-dimensional kernels with consistent sizes filtering have the same feature mapping with the same stride, their respective resolution outputs could be added up to form the same outputs from an equivalent kernel which is formed by adding corresponding positions of these kernels.

$$I * (\mathbf{K^{(1)}} \oplus \mathbf{K^{(2)}}) = I * \mathbf{K^{(1)}} + I * \mathbf{K^{(2)}} \tag{3.3}$$

where I is the identity feature matrix, $\mathbf{K^{(1)}}$ and $\mathbf{K^{(2)}}$ are two-dimensional kernels, $\oplus$ is the element-wise addition operator for adding weights from two kernels on the corresponding spatial locations.

Based on Equation 3.3, the additivity may support two-dimensional convolutions with different kernel sizes, for example, a "smaller" kernel may be patched to a "bigger" one. As mentioned in [25], compared to the corner weights learned by convolution kernels, the

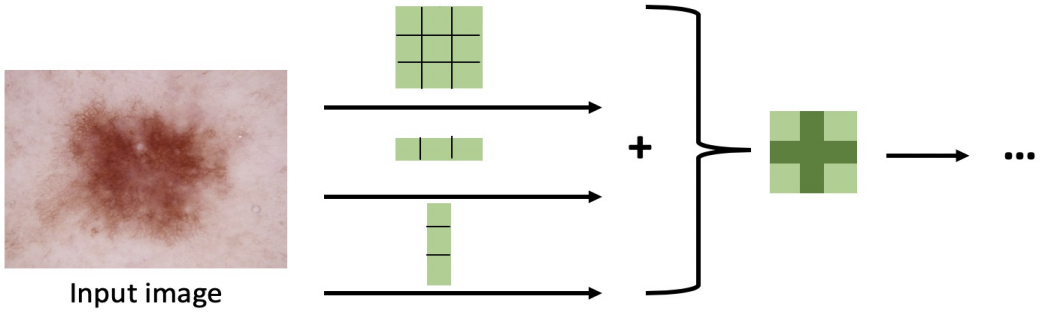


Figure 3.2: An example of the architecture of the ACB. For example, every standard 3 x 3 convolution layer is replaced by ACB that consists of three layers with 3 x 3, 1 x 3, and 3 x 1 kernel. The outputs are summed up to construct the feature map.

weights obtained by the central crisscross part of the kernel are the key point of improving the model accuracy. Therefore, adding one-dimensional horizontal and one-dimensional vertical kernels to the standard square kernels could improve the model performance.

Our proposed model replaces each convolutional kernel in 3.1 by ACB. By using ACB, the central part of weights is added twice in each convolution kernel. Unlike the sparse coding method proposed in Section 3.1.1, the ACB method is another way to enhance the robustness of the model. The sparse coding method proposed in Section 3.1.1 improves the model performance by making constraints on latent representations. The ACB improves the model performance by strengthening the central intersection parts of standard convolution kernels.

3.2 Summary

In this chapter, we proposed two CAE based methods for image anomaly detection. The two proposed models explore the potential of producing robust latent representations of normal images in different aspects. The first model described in Section 3.1.1 learns to extract distinct representations for abnormal patterns by imposing sparse regularizations on the latent space during the optimization process. The second model described in Section 3.1.2 detects abnormal patterns using ACB, which could extract features less sensitive to geometric transformations. These two proposed models provide examples of constraining intermediate representations and establishing connections among the layers in improving the visual recognition ability of models. Experimental results are tackled in Chapter 4 to illustrate the validity of our proposition.

Chapter 4

Experiments and Analysis

In this Chapter, the performance evaluations of the two proposed methods are presented. Also, two datasets are presented in Section 4.1 to validate the two approaches. Compared with several auto-encoder based models, the proposed methods achieve the best results in detecting abnormal images.

4.1 Datasets

Two datasets were selected for testing the proposed models described in Section 3.1.1 and Section 3.1.2. The two datasets are closely related to real-life challenges. One of the challenges is that the need for medical experts who can diagnose disease using images has increased tremendously, and the automation in diagnosis is required to support doctors. The first dataset contains healthy skin images and several kinds of unhealthy ones that we treated as anomalies. Also, the autonomous driving is becoming increasingly popular in recent years. It needs to deal with a number of unexpected anomalous situations to

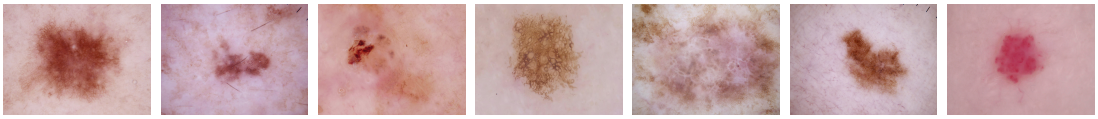
ensure the safety of passengers. The second dataset contains images of different types of distracted driving. The relevant information of the two datasets are provided in Section 4.1.1 and Section 4.1.2.

4.1.1 The HAM10000 dataset

The HAM10000 dataset is a pigmented skin disease image set that contains several dermatoscopic lesions [85]. The dataset includes seven types of skin images: melanocytic nevi (NV), actinic keratoses and intraepithelial carcinoma / Bowen's disease (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL) and vascular lesions (VASC). The number of images for each category is shown in Table 4.1. Figure 4.1 shows an example of each type of skin disease.

Table 4.1: The number of images for each category in the HAM10000 dataset

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



(a) NV (b) AKIEC (c) BCC (d) BKL (e) DF (f) MEL (g) VASC

Figure 4.1: The image of each type of skin diseases included in HAM10000 dataset

4.1.2 Daytime Driving Distraction dataset

The daytime driving distraction dataset [83] is collected from the integral upper body movement view of twenty-five drivers when they are driving on a simulated driving environment. The original dataset contains two versions of images: daytime and nighttime, the nighttime images are collected from an infrared camera. In this work, only daytime distraction driving images are used. The dataset contains abnormal behaviors such as talking, texting, and focusing on the GPS while driving, as well as normal driving behaviors. The number of images of each distraction driving class is given in Table 4.2.

Table 4.2: The number of images for each category in the Daytime Distraction Driving dataset

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

The images are collected from 25 participants. The participants include 16 males and 9 females of different ages coming from different countries. Before the experiment, each participant had enough time to get familiar with the simulator driving environment. Then, each participant performs distraction behaviors shown in Figure 4.2: talking on a cell phone with his/her right hand while driving, typing messages with a cell phone on his/her right hand while driving, operating a device near the gear stick of the simulator to mimic using GPS while driving, as well as the normal driving behavior. The camera is located on the right frontal side of each participant and the images are extracted from a recorded video. The images are captured in 5 frames per second.

(a) Normal (b) Talking (c) Texting (d) GPS

Figure 4.2: The image of each type of distraction driving class included in Daytime Distraction Driving dataset

4.2 Experimental Settings

To show the superiority of the two proposed methods, we compare several variants of auto-encoder based models against ours. The first model is the CAE trained with the binary cross-entropy loss function. This model has the same auto-encoder architecture as shown in Figure 3.1. The second model is the variational auto-encoder [27]. And the third model is the CAE trained with the mean square error loss function. All models included in the experiments have the same auto-encoder architecture and they only differ from each other in terms of loss functions. These model also share the same training dataset and pre-processing steps, which are listed below:

- Each model performs the same data pre-processing steps on both datasets: each image is resized to 128×128 and is normalized to the range of -1 to 1.
- The batch size is set to 64 and the Adam optimizer [26] is adopted for training all models. The number of epoch and learning rate are adjusted to make sure all models converge.
- For all auto-encoders in the models, the dimension of latent spaces is set to 300. And

for the proposed auto-encoder based sparse representation method, the parameter λ balancing the sparsity of the latent representations with the reconstruction quality is set to 0.002 for the HAM10000 dataset and 0.0004 for the daytime driving distraction dataset.

- The normal images $\{x_1, x_2, ..., x_n\}$ are used to train all models and are considered as independent and identically distributed samples from an unknown distribution. For both datasets, the prior distribution of the VAE is set to be Gaussian.

- During the testing stage, we randomly pick abnormal images from each abnormal category with the same number of normal images. Training and testing images are disjoint. We use Equation 3.1 to calculate the anomaly score of each testing image and the AUC score is computed based on the anomaly scores of the testing images.

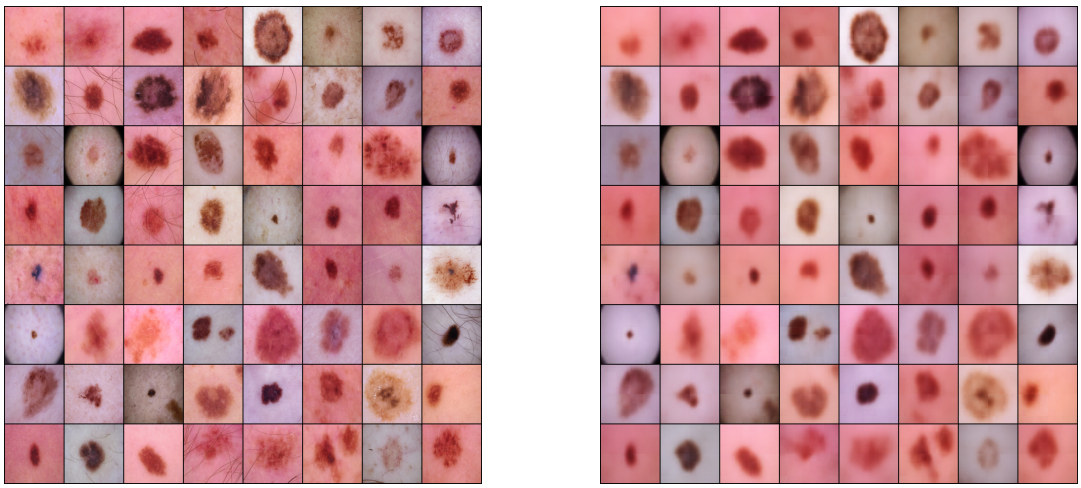
The experiment results on both datasets are presented in detail in the following subsection 4.3 and subsection 4.4.

4.3 Experiments on Sparse Representation Method

Symmetrical to the architecture of the encoder shown in Figure 3.1, the decoder of the sparse representation model has four deconvolution blocks that could reconstruct the images from sparse representations. The CAE with sparse representation is trained with normal images for 70 epochs in each run. By incorporating sparse regularization, the model is able to learn the sparse representations that could be used to reconstruct normal images. However, the trained model is not able to produce sparse representations for abnormal images. Therefore, the abnormal images reconstructed from those representations will produce a big loss since the decoder could not reconstruct them with high quality.

4.3.1 The HAM10000 data

All models are tested to detect skin diseases using the HAM10000 dataset. The experiment results show the effectiveness of using a CNN in the image feature learning task. And the superiority of the proposed model indicates the success of combining sparse coding with the auto-encoder in learning sparse representations of images.



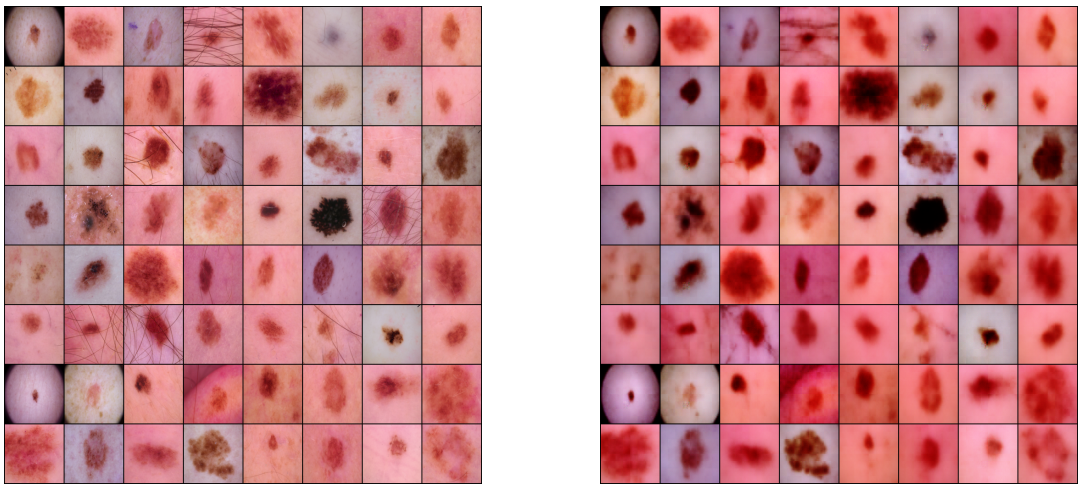
(a) the original images in test set

(b) the reconstructed images

Figure 4.3: The test images and corresponding reconstructed images on the HAM10000 dataset without sparse coding

Some examples of the original testing images and the images reconstructed from the latent representations without sparse coding are shown in Figure 4.3. Compared to the images reconstructed without sparse representations, the images reconstructed from the sparse representations shown in Figure 4.4 are less blurry and still maintain plenty of textures such as the contours of the objects.

To alleviate the randomness caused by the random initialization of the model, the experiment results presented in Table 4.3.1 are calculated based on the average of 10 runs



(a) the original images in test set

(b) the reconstructed images

Figure 4.4: The test images and corresponding reconstructed images on the HAM10000 dataset with sparse coding

of the experiments. As shown in Table 4.3.1, the proposed sparse representation model achieves overall highest AUC scores which indicate the best anomaly detection performance among other listed models.

The different types of pigmented skin are hard to distinguish by human eyes since many skin diseases have similar shape and texture patterns. As shown in Figure 4.1, the sample images of the NV (the normal type) and MEL (one of the abnormal types) type are similar to each other. Even for an expert, it may be challenging to identify the normal one. The experiment results in Table 4.3.1 show that the proposed sparse representation model could improve the performance of detecting anomalous skin images, especially for the anomaly types that performed poorly in the baseline model. For example, the proposed model improves the detection performance of AKIEC disease from 0.59 AUC in the baseline model to 0.78 AUC in the proposed model. For MEL, BCC, and VASC diseases, the

Table 4.3: The AUC results of the HAM10000 dataset

	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

* Baseline is the CAE trained by binary cross entropy loss.

proposed sparse representation model also greatly improves the AUC scores.

4.3.2 The Driving Distraction data

In recent years, autonomous driving is becoming an increasingly popular industrial field. There is a high demand for anomaly detection in distraction driving scenarios. To improve the safety of autonomous driving, anomaly detection techniques have been widely deployed in the industry. The experiments run on the driving distraction data to test how the proposed model performs on the distraction driving images. And same to skin disease images, the experiment results of driving distraction images also show the superiority of the proposed sparse representation model over several other variants of the auto-encoders.

The original images and the images reconstructed from corresponding latent representations without sparse coding are shown in Figure 4.5. Unlike the HAM10000 dataset, the reconstructed images look quite similar to the original ones, indicating that the model without sparse representation can learn the normal features. The images reconstructed from their corresponding sparse representations are shown in Figure 4.6. The reconstructed

(a) the original images in test set

(b) the reconstructed images

Figure 4.5: The test images and corresponding reconstructed images on the Driving Distraction dataset without sparse coding

(a) the original images in test set

(b) the reconstructed images

Figure 4.6: The test images and corresponding reconstructed images on the Driving Distraction dataset with sparse coding

images have a darker background and have fewer details in the background. The reconstructed images lose detailed features such as the color of the participants' clothes and the shape of the driving seat. But it can also be regarded as the sparse representation model ignored unimportant features.

To minimize the impact of the random initialization of the model, the experiment results presented in Table 4.3.2 are calculated based on the average of 10 runs of the experiments. As shown in Table 4.3.2, the proposed sparse representation model achieves the best overall AUC scores. It significantly improves the detection performance of the using GPS while driving anomaly type from 0.63 AUC to 0.83 AUC and the talking on the phone anomaly type from 0.56 AUC to 0.64 AUC. This surprising fact also reveals that the deep convolutional encoder trained by normal images is able to extract useful normal information when combined with sparse coding. Among those three different driving distraction abnormal behaviors, using GPS while driving is the most similar to the normal driving behavior. But sparse representations bring it the biggest improvement. The huge improvement also indicates that the proposed sparse representation model could successfully detect the anomaly images that look similar to the normal images.

4.4 Experiments on ACB Method

The proposed model has the same auto-encoder architecture as the model previously proposed in 3.1, except that all convolutional layers are replaced with ACB. As shown in Figure 3.2, the ACB combines every two-dimensional convolution kernel with two parallel one-dimensional convolution kernels.

Similar to the settings in Section 4.3, the proposed auto-encoder with ACB is trained with normal images. Without introducing extra computations, the ACB strengthens the

Table 4.4: The AUC results of the Driving Distraction dataset

	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	0.69	0.83	0.72

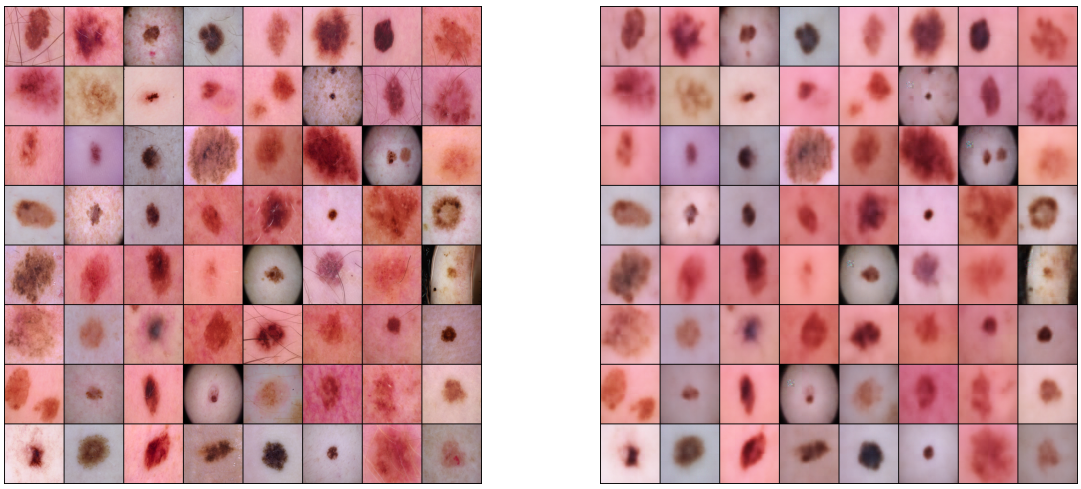
* Baseline is the CAE trained by binary cross entropy loss.

central intersection parts of standard convolution kernels. In consequence, the feature mappings in the down-sampling layers are enhanced. Since the corner weights have less impact in ACB, the proposed model is less sensitive to geometric transformations such as rotation.

4.4.1 The HAM10000 data

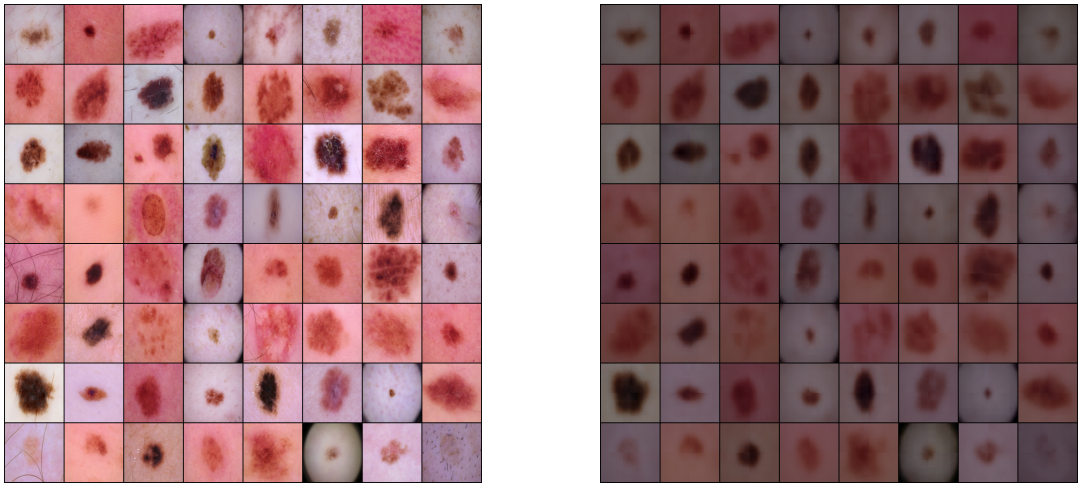
The original testing images and their corresponding reconstructed images of the VAE model are shown in Figure 4.7. As shown in Figure 4.1, the features of the HAM10000 skin images are mainly located in the center of the images. The experiment results presented in Table 4.3.1 are calculated based on the average of 10 runs of the experiments. As shown in Table 4.3.1, the proposed model achieves the best AUC scores of each anomaly category.

The skin disease images of different types have many similar local characteristics and are invariant to image rotation. Therefore, we need a robust representation sensitive to different types of diseases and invariant to image rotation. The proposed model meets the



(a) the original images in test set (b) the reconstructed images

Figure 4.7: The HAM10000 images and their corresponding reconstructed images of VAE



(a) the original images in test set (b) the reconstructed images

Figure 4.8: The HAM10000 images and their corresponding images reconstructed from the proposed ACB model

Table 4.5: The AUC results of the HAM10000 dataset

	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 + ACB	0.84	0.71	0.81	0.74	0.65	0.65	0.73

* Baseline is the CAE trained by binary cross entropy loss.

above two requirements by using ACB. Compared to the results of the sparse model shown in Table 4.3.1, the proposed model further improves the detection performance by a large margin. For example, the proposed model improves the detection of MEL disease from 0.60 AUC in the baseline model to 0.84 AUC, and the AKIEC disease from 0.59 AUC to 0.81 AUC. However, compared to the proposed sparse representation model, the model with ACB cannot further improve the detection performance of the BCC and VASC diseases. As shown in Figure 4.1, the BCC and VASC images are quite different from the normal NV image. Based on the observations, the proposed model with ACB is more suitable for detecting anomalies similar to the normal ones.

4.4.2 The Driving Distraction data

In the recorded driving scenarios, the images may suffer from different degrees of rotation due to imperfect camera setups. Anomaly detection models should be able to deal with images taken from different angles. ACB is a good solution to this problem. The proposed model with ACB improves the detection performance by adding a one-dimensional hori-

zontal kernel and one-dimensional vertical kernel to the standard two-dimensional kernel.

(a) the original images in test set

(b) the reconstructed images

Figure 4.9: The driving distraction images and their corresponding reconstructed images of VAE

The reconstructed images of the proposed model are shown in Figure 4.10. The experiment results presented in Table 4.4.2 are calculated based on the average of 10 runs of the experiments. The detection performance of the proposed model achieves the best AUC scores over the listing models. The proposed model with ACB improves the detection of talking while driving anomaly type from 0.56 AUC to 0.65 AUC and texting while driving anomaly type from 0.67 AUC to 0.72 AUC. The experiment results also reveal that the proposed model is invariant to the image rotation and can extract rotation invariant features that could represent the normal images.

(a) the original images in test set

(b) the reconstructed images

Figure 4.10: The driving distraction images and their corresponding images reconstructed from the proposed ACB model

4.5 Summary

In this chapter, two datasets are presented in Section 4.1 to validate the two approaches. And experiment results of the proposed sparse representation model are given in Section 4.3 and the experiment results of the proposed model with ACB are given in Section 4.4. The two proposed models enhance the anomaly detection performance by a large margin compared to the baseline models. This demonstrates the effectiveness of sparse representations and the ACB in detecting anomalies.

Table 4.6: The AUC results of the Driving Distraction dataset

	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 + ACB	0.65	0.72	0.83	0.73

* Baseline is the CAE trained by binary cross entropy loss.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this research work, two proposed methods aim at obtaining the feature presentation for images in two different aspects. From the perspective of constraining the latent representation in a sparse way, the proposed model combined the sparse coding to produce the sparse representation in a low-dimensional feature space in which the abnormal images are expected to present in a different way in this latent space. The proposed sparse representation model is able to produce robust representations for normal images and improve the anomaly detection performance in both datasets. From the perspective of the impact of the convolution kernel spatial locations in extracting the feature representations, the proposed model with ACB is another alternative at producing the robust features that invariant to rotation images and distortion images among normal samples.

Both proposed methods could be easily combined with other anomaly detection techniques. Especially for the proposed ACB, it could be adapted to most of the main-stream

CNN. While the approaches used have provided advantages in terms of flexibility, there are still challenges such as being unable at handling normal images with strong diversity and thus deal with complicated anomaly detection scenarios. Further improvements are required by adjoining other modules or by proposing new architectural designs of the deep neural networks. Future work is summarized next.

5.2 Future Work

Although the proposed methods explore the potential of using the encoding and decoding mechanism of the auto-encoder to identify abnormal images, some other relevant topics still need to be tackled by researchers in future work:

1. The combination of the sparse representation with manifold learning techniques in producing latent representations with geometrically sparse structures is one of the interesting topics of anomaly detection. This idea makes an assumption that the normal patterns follow a unified sparse structure and it can be applied to high dimensional data that contains various structural anomalies.

2. Except for looking into the central crisscross part of the convolution kernels, an alternative way to explore how spatial locations of the convolution kernels will affect the feature representation performance is to see the problem from the opposite perspective. Therefore, exploring the effectiveness of the corner weights of the learned feature maps for detecting the anomalies can be regarded as another interesting topic in future work.

References

[1] Bagnell J. A. and Bradley D. M. *Differentiable sparse coding.* in NIPS, pp. 113–120, 2009.

[2] Coates A. and Ng A. Y. *The importance of encoding versus training with sparse coding and vector quantization.* in ICML, 2011.

[3] Coates A. and Ng A. Y. *The importance of encoding versus training with sparse coding and vector quantization.* in ICML, 2011.

[4] Hyvärinen A. *Estimation of non-normalized statistical models using score matching.* in Journal of Machine Learning Res, vol. 6, 2005.

[5] Hyvärinen A. *Some extensions of score matching.* in Computational Statistics and Data Analysis, vol. 51, pp. 2499–2512, 2007.

[6] Hyvärinen A. *Optimal approximation of signal priors.* in Neural Computation, 20(12), pp. 3087–3110, 2008.

[7] Krizhevsky A. and Hinton G. *Learning multiple layers of features from tiny images.* in Technical report, University of Toronto, 2009.

[8] Olshausen B. A. and Field D. J. *Emergence of simplecell receptive field properties by learning a sparse code for natural images.* in Nature, vol. 381, pp. 607–609, 1996.

[9] Olshausen B. A. and Field D. J. *Emergence of simplecell receptive field properties by learning a sparse code for natural images.* in Nature, vol. 381, pp. 607–609, 1996.

[10] Radford A., Metz L., and Chintala S. *Unsupervised representation learning with deep convolutional generative adversarial networks.* in the International Conference on Learning Representations (ICLR), 2015.

[11] Charu C Aggarwal. *An introduction to outlier analysis.* in Outlier analysis, Springer, pages 1–40, 2013.

[12] Daneshpazhouh Armin and Sami Ashkan. *Entropy-based outlier detection using semi-supervised approach with few positive examples.* in Pattern Recognition Letters, vol. 49, pp. 747–752, 2014.

[13] Xu B., Wang N., Chen T., and Li M. *Empirical Evaluation of Rectified Activations in Convolution Network.* arXiv preprint arXiv:1505.00853, 2015.

[14] Zhao B., Fei-Fei L., and 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.

[15] G. Baudat and F. Anouar. *Generalized discriminant analysis using a kernel approach.* in Neural Computation, vol. 12, pp. 2385–2404, 1936.

[16] P. Belhumeur, J. Hespanha, and D. Kriegman. *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection.* in IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, pp. 711–720, 1997.

[17] M. Belkin and P. Niyogi. *Laplacian eigenmaps and spectral techniques for embedding and clustering.* in NIPS, pp. 585–591, 2001.

[18] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. *Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks.* in the IEEE conference on computer vision and pattern recognition, 2016.

[19] Y. Bengio, A. Courville, and P. Vincent. *Representation learning: A review and new perspectives.* in IEEE Trans, Pattern Anal. Mach. Intell., vol. 35, pp. 1798–1828, 2013.

[20] Zhou C. and Paffenroth 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.

[21] M. Cheriet, R. Moghaddam, E. Arabnejad, and G. Zhong. *Manifold Learning for the Shape-Based Recognition of Historical Arabic Documents.* in Elsevier,, pp. 471–491, 2013.

[22] Tsallis Constantino. *Possible generalization of Boltzmann-Gibbs statistics.* in Journal of Statistical Physics, Springer, vol. 52, pp. 479-487, 1988.

[23] Erhan D., Bengio Y., Courville A., Manzagol P.-A., Vincent P., and Bengio S. *Why does unsupervised pre-training help deep learning?* in Journal of Machine Learning Research, vol. 11, pp. 625–660, 2010.

[24] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. *Exploiting linear structure within convolutional networks for efficient evaluation.* in Advances in neural information processing systems, pages 1269–1277, 2014.

[25] X. Ding, Y. Guo, G. Ding, and J. Han. *Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks.* In Proceedings of the IEEE International Conference on Computer Vision, pages 1911–1920, 2019.

[26] Kingma D.P. and Ba J.L. *Adam: a method for stochastic optimization.* in International Conference on Learning Representations (ICLR), 2015.

[27] Kingma D.P. and Welling M. *Auto-Encoding Variational Bayes.* The 2nd International Conference on Learning Representations (ICLR), 2014.

[28] S. Dumais. *Latent semantic analysis.* in ARIST, vol. 38, pp. 188–230, 2004.

[29] Hinton G. E. *Products of experts.* in ICANN, 1999.

[30] Hinton G. E., Osindero S., and Teh Y. *A fast learning algorithm for deep belief nets.* in Neural Computation, vol. 18, pp. 1527–1554, 2006.

[31] Tipping M. E. and Bishop C. M. *Probabilistic principal components analysis.* in J. Roy. Stat. Soc. B, (3), 1999.

[32] Otey Matthew Eric, Ghoting Amol, and Parthasarathy Srinivasan. *Fast distributed outlier detection in mixed-attribute data sets.* in Data mining and knowledge discovery, Springer, vol 12, pp. 203–228, 2006.

[33] R. Fisher. *The use of multiple measurements in taxonomic problems.* in Annals of Eugenics, vol. 7, pp. 179–188, 1936.

[34] R. W. Floyd. *Algorithm 97: Shortest path.* in Communications of the Acm, 1962.

[35] Alain G. and Bengio Y. *What regularized auto-encoders learn from the data generating distribution.* in Technical Report Arxiv report 1211.4246, Université de Montréal, 2012.

[36] Mesnil G., Dauphin Y., Glorot X., Rifai S., Bengio Y., Goodfellow I., Lavoie E., Muller X., Desjardins G., Warde-Farley D., Vincent P., Courville A., and Bergstra J. *Unsupervised and transfer learning challenge: a deep learning approach.* in JMLR WCP: Proc. Unsupervised and Transfer Learning, vol. 7, 2011.

[37] A. C. GILBERT, M. J. STRAUSS, J. A. TROPP, and R. VERSHYNIN. *Algorithmic linear dimension reduction in the l1 norm for sparse vectors.* the 44th Annual Allerton Conference on Communication, Control, and Computing,, 2006.

[38] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, Massachusetts, 1994.

[39] Bourlard H. and Kamp Y. *Auto-association by multilayer perceptrons and singular value decomposition.* in Biological Cybernetics, vol. 59, pp. 291–294, 1988.

[40] Larochelle H. and Bengio Y. *Classification using discriminative restricted Boltzmann machines.* in ICML, 2008.

[41] Lee H., Battle A., Raina R., and Ng A.Y. *Efficient sparse coding algorithms.* Advances in neural information processing systems (NIPS), pp. 801–808, 2007.

[42] Lee H., Ekanadham C., and Ng A. *Sparse deep belief net model for visual area V2.* in NIPS, 2008.

[43] D. Hawkins. *Identification of Outliers.* Chapman and Hall, London, 1980.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep residual learning for image recognition.* in Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[45] X. He and P. Niyogi. *Locality preserving projections.* in NIPS, 2003.

[46] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. *Face Recognition Using Laplacianfaces.* in IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, pp. 328–340, 2005.

[47] Jie Hu, Li Shen, and Gang Sun. *Squeeze-and-excitation networks.* in Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

[48] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. *Densely connected convolutional networks.* in Proceedings of the IEEE conference on computer vision and pattern recognition, volume 1, page 3, 2017.

[49] Goodfellow I., Courville A., and Bengio Y. *Spike-and-slab sparse coding for unsupervised feature discovery.* in NIPS Workshop on Challenges in Learning Hierarchical Models, 2011.

[50] Goodfellow I., Le Q., Saxe A., and Ng A. *Measuring invariances in deep networks.* in NIPS, pp. 646–654, 2009.

[51] Goodfellow I.J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., and Bengio Y. *Generative Adversarial Nets.* in Conference on Neural Information Processing Systems (NIPS), 2014.

[52] Hastad J. and Goldmann M. *On the power of small-depth threshold circuits.* in Computational Complexity, vol. 1, pp. 113–129, 1991.

[53] Kivinen J. J. and Williams C. K. I. *Multiple texture Boltzmann machines.* in AISTATS, 2012.

[54] Ngiam J., Chen Z., Koh P., and Ng A. *Learning deep energy models.* in ICML, 2011.

[55] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. *Spatial transformer networks.* in Conference on Neural Information Processing Systems, 2015.

[56] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. *Speeding up convolutional neural networks with low rank expansions.* arXiv preprint arXiv:1405.3866, 2014.

[57] Kingdon Jason. *AI fights money laundering.* in IEEE Intelligent Systems, IEEE, vol 19, pp. 87–89, 2004.

[58] Y. Jia, F. Nie, and C. Zhang. *Trace Ratio Problem Revisited.* in IEEE Trans. Neural Networks, vol. 20, pp. 729–735, 2009.

[59] Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. *Flattened convolutional neural networks for feedforward acceleration.* arXiv preprint arXiv:1412.5474, 2014.

[60] Springenberg J.T., Dosovitskiy A., Brox T., and Riedmiller M. *Striving for simplicity: The all convolutional net.* CoRR, abs/1412.6806, 2014.

[61] Sun J.Y., Wang X.Z., Xiong N.X., and Shao J. *Learning Sparse Representation With Variational Auto-Encoder for Anomaly Detection.* IEEE Access, 33353-33361, 2018.

[62] Fukushima K. *Neocognitron: A hierarchical neural network capable of visual pattern recognition.* in Neural networks, vol. 1, pp. 119-130, 1988.

[63] Kavukcuoglu K., Ranzato M., and LeCun Y. *Fast inference in sparse coding algorithms with applications to object recognition.* in CBLL-TR-2008-12-01, NYU, 2008.

[64] Simonyan K. and Zisserman A. *Very deep convolutional networks for large-scale image recognition.* in Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[65] Swersky K., Ranzato M., Buchman D., Marlin B., and de Freitas N. *On score matching for energy based models: Generalizing autoencoders and simplifying deep learning.* in ICML, 2011.

[66] Yu K., Lin Y., and Lafferty J. *Learning image representations from the pixel level via hierarchical sparse coding.* in CVPR, 2011.

[67] Donald Knuth. *The TEXbook.* Addison-Wesley, Reading, Massachusetts, 1986.

[68] Leslie Lamport. *LATEX — A Document Preparation System.* Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[69] N. Lawrence. *Probabilistic non-linear principal component analysis with Gaussian process latent variable models.* in Journal of Machine Learning Research, vol. 6, pp. 1783–1816, 2005.

[70] Y. LeCun, Y. Bengio, and G. Hinton. *Deep learning.* in Nature, vol. 521, pp. 436–444, 2015.

[71] W.-J. Li, D.-Y. Yeung, and Z. Zhang. *Probabilistic relational PCA.* in NIPS, pp. 1123–1131, 2009.

[72] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, and Jing-Jhih Lin. *Efficient dense modules of asymmetric convolution for real-time semantic segmentation.* arXiv preprint arXiv:1809.06323, 2018.

[73] Chen M., Xu Z., Winberger K. Q., and Sha F. *Marginalized denoising autoencoders for domain adaptation.* in ICML, 2012.

[74] Gutmann M. and Hyvarinen A. *Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.* in AISTATS, 2010.

[75] Neal R. M. *Connectionist learning of belief networks.* in Artificial Intelligence, vol. 56, pp. 71–113, 1992.

[76] Neal R. M. *Probabilistic inference using Markov chain Monte-Carlo methods.* in Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

[77] Ranzato M., Poultney C., Chopra S., and LeCun Y. *Efficient learning of sparse representations with an energy-based model.* in NIPS, 2007.

[78] Ranzato M. and Hinton G. H. *Modeling pixel means and covariances using factorized third-order Boltzmann machines.* in CVPR, pp. 2551–2558, 2010.

[79] Ranzato M., Mnih V., and Hinton G. *Generating more realistic images using gated MRF's.* in NIPS, 2010.

[80] Ranzato M., Boureau Y., and LeCun Y. *Sparse feature learning for deep belief networks.* in NIPS, 2008.

[81] Zhang Zhongfei Mark, Salerno John J, and Yu Philip S. *Applying data mining in investigating money laundering crimes.* in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 747–752, 2003.

[82] A. Newell, K. Yang, and J. Deng. *Stacked hourglass networks for human pose estimation.* in ECCV, 2016.

[83] C. Ou, Q. Zhao, F. Karray, and A. E. Khatib. *Design of an End-to-End Dual Mode Driver Distraction Detection System.* in the 16th International Conference on Image Analysis and Recognition, 2019.

[84] Smolensky P. *Information processing in dynamical systems: Foundations of harmony theory.* in D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing, vol. 1, chapter 6, pp. 194–281, MIT Press, Cambridge, 1986.

[85] Tschandl P., Rosendahl C., and Kittler H. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.* Online, 2018.

[86] Vincent P., Larochelle H., Lajoie I., Bengio Y., and Manzagol P.-A. *Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.* Journal of Machine Learning Research, pp. 3371–3408, 2010.

[87] Vincent P., Larochelle H., Bengio Y., and Manzagol P.-A. *Extracting and Composing Robust Features with Denoising Autoencoders.* The 25th International Conference on Machine Learning, pp. 1096–1103, 2008.

[88] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. *Enet: A deep neural network architecture for real-time semantic segmentation.* arXiv preprint arXiv:1606.02147, 2016.

[89] K. Pearson. *On lines and planes of closest fit to systems of points in space.* in Philosophical Magazine, vol. 2, pp. 559–572, 1901.

[90] Le Q., Ngiam J., Coates A., Lahiri A., Prochnow B., and Ng A. *On optimization methods for deep learning.* in ICML, 2011.

[91] Chalapathy R. and Chawla S. *Deep Learning for Anomaly Detection: A Survey.* arXiv:1901.03407, 2019.

[92] Grosse R., Raina R., Kwong H., and Ng A. Y. *Shift-invariant sparse coding for audio classification.* in UAI, 2007.

[93] Salakhutdinov R. and Hinton G. E. *Deep Boltzmann machines.* in AISTATS, pp. 448–455, 2009.

[94] S. Roweis and L. Saul. *Nonlinear dimensionality reduction by locally linear embedding.* in Science, vol. 290, pp. 2323–2326, 2000.

[95] Ioffe s. and Szegedy C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* the International Conference on Machine Learning, 2015.

[96] Lazebnik S., Schmid C., and Ponce J. *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.* in CVPR, 2006.

[97] Rifai S., Vincent P., Muller X., Glorot X., and Bengio Y. *Contractive auto-encoders: Explicit invariance during feature extraction.* in ICML, 2011.

[98] B. Schölkopf, A. Smola, and K.-R.Müller. *Nonlinear component analysis as a kernel eigenvalue problem.* in Neural Computation, vol. 10, pp. 1299–1319, 1998.

[99] J. Schmidhuber. *Deep learning in neural networks: An overview.* in Neural Networks, vol. 61, pp. 85–117, 2015.

[100] L. Sirovich and M. Kirby. *Low-dimensional procedure for the characterization of human faces.* in Journal of the Optical Society of America, pp. 519–524, 1987.

[101] Zhun Sun, Mete Ozay, and Takayuki Okatani. *Design of kernels in convolutional neural networks for image classification.* in European Conference on Computer Vision, pages 51–66. Springer, 2016.

[102] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. *Inception-v4, inception-resnet and the impact of residual connections on learning.* in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[103] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. *Rethinking the inception architecture for computer vision.* in Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[104] J. Tenenbaum, V. Silva, and J. Langford. *A global geometric framework for nonlinear dimensionality reduction.* in Science, vol. 290, pp. 2319–2323, 2000.

[105] R. Urtasun and T. Darrell. *Discriminative Gaussian process latent variable models for classification.* in Proceedings of the International Conference on Machine Learning, pp. 927–934, 2007.

[106] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. *Ratio Trace for Dimensionality Reduction.* in CVPR, 2007.

[107] Bengio Y. *Learning deep architectures for AI.* in Foundations and Trends in Machine Learning, 2(1), pp. 1–127, 2009.

[108] Bengio Y. and Monperrus M. *Non-local manifold tangent learning.* in In NIPS, MIT Press, pp. 129–136, 2005.

[109] Bengio Y. and Delalleau O. *Justifying and generalizing contrastive divergence.* in Neural Computation, 21(6), pp. 1601–1621, 2009.

[110] Bengio Y. and Delalleau O. *On the expressive power of deep architectures.* in ALT, 2011.

[111] Bengio Y., Delalleau O., and Simard C. *Decision trees do not generalize to new variations.* in Computational Intelligence, 26(4), pp. 449–467, 2010.

[112] Bengio Y., Delalleau O., and Le Roux N. *The curse of highly variable functions for local kernel machines.* in NIPS, 2006.

[113] Bengio Y., Lamblin P., Popovici D., and Larochelle H. *Greedy layer-wise training of deep networks.* in NIPS, 2007.

[114] Bengio Y. and LeCun Y. *Scaling learning algorithms towards AI.* in L. Bottou, O. Chapelle, D. DeCoste and J. Weston, editors, Large Scale Kernel Machines, MIT Press, 2007.

[115] Cong Y., Yuan J., and Liu J. *Sparse Reconstruction Cost for Abnormal Event Detection.* In Computer vision and pattern recognition (CVPR), pp. 3449-3456, 2011.

[116] LeCun Y., Bottou L., Bengio Y., and Haffner P. *Gradient based learning applied to document recognition.* in IEEE, 1998.

[117] Zou W. Y., Ng A. Y., and Yu K. *Unsupervised learning of visual invariance with temporal coherence.* in NIPS, Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[118] Z. Zhang and H. Zha. *Principal manifolds and nonlinear dimensionality reduction via tangent space alignment.* in SIAM J. Scientific Computing, vol. 26, pp. 313–338, 2004.

[119] Q. Zhao and F. Karray. *Anomaly Detection for Images using Auto-Encoder based Sparse Representation.* in the 17th International Conference on Image Analysis and Recognition, 2020.

[120] G. Zhong, K. Huang, X. Hou, and S. Xiang. *Local Tangent Space Laplacian Eigenmaps.* in SNOVA Science Publishers, pp. 17–34, 2012.

[121] G. Zhong, Y. Shi, and M. Cherietg. *Relational fisher analysis: A general framework for dimensionality reduction.* in IJCNN, pp. 2244–2251, 2016.

[122] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc VLe. *Learning transferable architectures for scalable image recognition.* in Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8697–8710, 2018.

[123] H. Zou, T. Hastie, and R. Tibshirani. *Sparse principal component analysis.* in Journal of Computational and Graphical Statistics, vol. 15, pp. 265–286, 2006.