Learning Competitive and Discriminative Reconstructions for Anomaly Detection

Kai Tian,[1] **Shuigeng Zhou,**[1*] **Jianping Fan,**[2] **Jihong Guan**[3]
[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
[2]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223 USA
[3]Department of Computer Science & Technology, Tongji University, China
[1]{ktian14, sgzhou}@fudan.edu.cn; [2]jfan@uncc.edu; [3]jhguan@tongji.edu.cn

Abstract

Most of the existing methods for anomaly detection use only positive data to learn the data distribution, thus they usually need a pre-defined threshold at the detection stage to determine whether a test instance is an outlier. Unfortunately, a good threshold is vital for the performance and it is really hard to find an optimal one. In this paper, we take the discriminative information implied in unlabeled data into consideration and propose a new method for anomaly detection that can learn the labels of unlabelled data directly. Our proposed method has an end-to-end architecture with one encoder and two decoders that are trained to model inliers and outliers' data distributions in a competitive way. This architecture works in a discriminative manner without suffering from overfitting, and the training algorithm of our model is adopted from SGD, thus it is efficient and scalable even for large-scale datasets. Empirical studies on 7 datasets including KDD99, MNIST, Caltech-256, and ImageNet etc. show that our model outperforms the state-of-the-art methods.

Introduction

Anomaly detection is to identify the data that do not conform to the expected normal patterns. These data may come from a new class or some noisy data that has no meaning. Usually, we call these abnormal data *outliers*, and *inliers* for the normal data. Anomaly detection is closely related with many real-world applications such as outlier detection, novelty detection in computer vision area (Khan and Madden 2009; Chandola, Banerjee, and Kumar 2009; Khan and Madden 2014) and medical diagnoses, drug discovery in bioinformatics (Wei et al. 2018). It can be categorized into one-class learning, where the profile of negative class is not well defined. According to the real application contexts, the negative data could be hard to collect or verify. Besides, there could be any kind of abnormal data that are unpredictable. Thus, those data are considered as novelties (or outliers), while the positive data (or inliers) are well characterized by the training data. It is hard to use traditional multiple class classification methods to learn from only positive labeled data due to the inertness of classifiers.

Over the past decades, researchers have proposed lots of methods to deal with anomaly detection problems (Eskin 2000; Chandola, Banerjee, and Kumar 2009; Liu, Hua, and Smith 2014; Malhotra et al. 2016). Generally, these methods either build a model configuration for normal data examples and identify the examples that disobey the normal profiles as outliers, or explicitly isolate outliers based on statistical or geometric measures of abnormality. Usually, different models have different capacities to characterize the data distribution. Most of the traditional methods are linear models which have limited model capacity. Although kernel function can be used to improve their capacities, it is not suitable for the context of high-dimensional and large-scale data.

Recently, deep learning methods have shown their powerful representation ability and gained immense success in many applications (Vincent et al. 2010; Krizhevsky, Sutskever, and Hinton 2012; Bengio, Courville, and Vincent 2013). However, due to the unavailability of negative data, it is hard to train a supervised deep neural network for one-class classification directly. Although some efforts have been made to learn one-class classifier, most of them could not establish a discriminative model for anomaly detection. The detection is done by choosing a pre-defined threshold. From the probability perspective, it can be explained as that outliers should lie on the low-density regions of the model's distribution. However, since the outliers are not predictable, it is hard to determine a threshold that works for all cases. Meanwhile, as the model is only trained on the positive data, overfitting is another key factor that may destroy the model's generalization performance. That is the reason why we cannot simply train a DNN classifier based on positive data. Although one can use some strategies such as early stopping to avoid overfitting, it is very tricky and one can not decide when to stop is best for the test data.

To address those issues, we propose a new model named *Co*mpetitive *R*econstruction *A*utoencoder (*CoRA*). Our model has the benefits of both supervised and unsupervised approaches. We formulate a transductive semi-supervised method for anomaly detection, which uses positive training data and unlabeled test data for learning.

Fig. 1 demonstrates the differences of the learning procedure between our method and most existing approaches. The proposed architecture, different from classical auto-encoder, comprises one encoder and two decoders. These two decoders are designed to compete during the reconstruction procedure, one of them is learned to reconstruct positive

[*]Correspondence author.

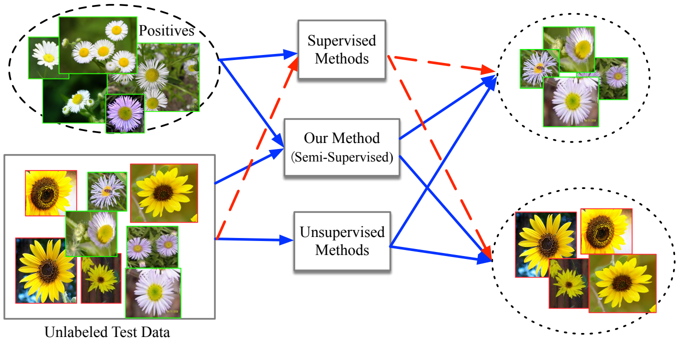


Figure 1: Comparison of the training/testing process among supervised methods, unsupervised methods and our method. The blue solid arrows indicate the training data flow. The red dashed arrows indicate the data flow of testing stage. Unsupervised methods and our method do not have a separate testing stage, they discriminate the outliers during the training stage.

data which is referred to as *inlier decoder*, while the other is learned to reconstruct outliers which is referred to as *outlier decoder*. With the guidance of the positive training data, *inlier decoder* can build a proper profile for positive class, while most of the anomaly data will be assigned to the *outlier decoder*. Discriminative labeling is done by comparing the reconstruction errors of these two decoders. After training, the final assignments constitute the predictions for the unlabelled (or test) data.

It is known that auto-encoder will learn different feature subspaces for different data spaces. As our two competitive decoders share one encoder, it is reasonable to add a regularization term on the subspace to maximize the separability of the positive data and outliers, which will further improve *CoRA*. Thus, in this paper we propose a manifold regularizer to preserve the data structures of positive data and outliers in subspace.

In this paper, we propose a transductive semi-supervised anomaly detection approach that can be trained in an end-to-end manner. Our work is featured by the following merits:

- **Novel architecture.** A new transductive semi-supervised deep neural network framework implemented by an auto-encoder with one encoder and two competitive decoders is proposed for anomaly detection problems.
- **New criterion.** While the previous reconstruction-based algorithms use thresholds as the classification criterion, we propose a new scheme to make the decision, which can help us get rid of this hyper-parameter selection.
- **Robustness.** Extensive experiments show that the proposed model is more robust to the outlier ratio than many state-of-the-art methods.
- **Efficient and scalable optimization procedure.** We adopt the stochastic gradient descent (SGD) to our model, making it very efficient to train and can be used for large-scale data.
- **Comprehensive experiments.** We comprehensively evaluate the proposed model and compare it with a number of state-of-the-art methods on seven datasets.

Related Work

Anomaly detection belongs to one-class classification which is also closely related to outlier detection or novelty detection. The common objective of these applications is to discover novel concept, which has rarely or never shown up and is substantially different from the known concept.

Conventional researches often model the positive/target class and reject the samples that do not following them. Methods in this category usually estimate probability density function from positive data. By building up a parametric or non-parametric probability estimator from the positive samples, including kernel density estimator (KDE) (Parzen 1962) and more recent robust kernel density estimator (RKDE) (Kim and Scott 2012), one can identify an outlier sample when it has low probability density. These methods take the assumption that positive data are more densely distributed. Other statistical models such as PCA (De La Torre and Black 2003; Candès et al. 2011; Xu, Caramanis, and Mannor 2013) assume that positive data are more correlated with each other than outliers and they get better reconstruction from the low-dimensional subspace.

Most deep learning approaches for anomaly detection are built upon auto-encoders, where reconstruction error is used to discriminate outliers from inliers. Those methods learn from only positive data which induces some problems. On the one hand, these methods need to pre-define a threshold to determine the unlabelled samples, while it is hard to find an optimal threshold. On the other hand, training auto-encoders on positive data will suffer from the overfitting problem, especially when the number positive samples is small.

Recently, (Zhai et al. 2016) proposed a new energy-based deep neural network to detect outliers. Instead of directly using the reconstruction error as decision criterion, they showed that energy could be another criterion for identifying outliers. However, this method still need a pre-defined threshold to discriminate outliers. Instead of autoenocders, some researchers tried to solve one-class classification problem with generative adversarial networks (GANs). One of them is to combine a denoising auto-encoder with a discriminator and train them in an adversarial way (Sabokrou et al. 2018). The key idea is to enhance the decoder to reconstruct inliers perfectly while distort the outliers. Although they used discriminator of GAN to classify the samples, classification is actually based on a task-dependent threshold, i.e. x is a outlier if $f(x) < \sigma$.

Our method is closely related to some unsupervised approaches for anomaly detection or outlier removal. Different from supervised methods that just learn from positive data, unsupervised methods take the discriminative information from the unlabeled data by introducing auxiliary labels, and maximize the discrimination between inliers and outliers, they iteratively refine the labels and finally output the predictions. An unsupervised one-class learning (UOCL) method was proposed in (Liu, Hua, and Smith 2014) by utilizing an auxiliary label variable, and jointly optimizing the kernel-based max-margin classifier and the soft-label assignment.

Table 1: Comparison of state-of-the-art DNN-based models with our method from three perspectives.

Method	Architecture	Criterion	Learning paradigm
DRAE (Xia et al. 2015)	auto-encoder	Discriminative	Unsupervised
DSEBM (Zhai et al. 2016)	DNN	Threshold (energy)	Supervised
DAOC (Sabokrou et al. 2018)	GAN	Threshold (score)	Supervised
SSGAN (Kimura and Yanagihara 2018)	GAN	Threshold (score)	Inductive semi-supervised
CoRA	auto-encoder	Discriminative	Transductive semi-supervised

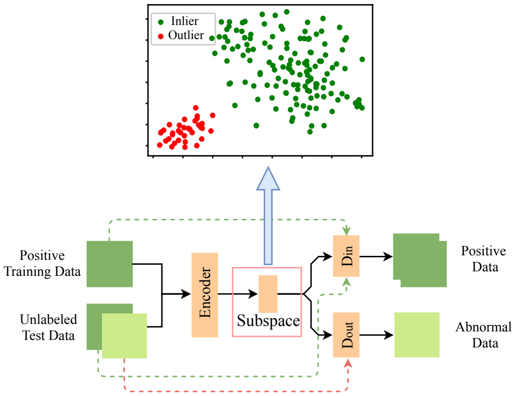


Figure 2: The architecture of the *CoRA* model. Both training data and test data are used for learning. All data share the same encoder but have to choose one of the two different decoders. The encoder aims to map positive data and unlabelled data in the subspace separately. Green dotted lines indicates that positive data in both training and test datasets would be assigned to the *inlier decoder*, while the red dotted line represents the assignment of abnormal data (or outliers).

Another deep learning based unsupervised approach is discriminative reconstruction auto-encoder (DRAE) (Xia et al. 2015), they also introduced discriminative labels for the unlabeled data and optimized the within-class variance of reconstruction error for both inliers and outliers in the learning process of auto-encoder.

Note that our formulation for anomaly detection is different from positive and unlabeled (PU) learning (Elkan and Noto 2008) or *inductive* semi-supervised anomaly detection (Kimura and Yanagihara 2018). In both cases, it is assumed that the unlabeled data that used for training share the same distribution with the test data. In our case, similar to unsupervised approaches, only positive data and unlabelled (or test) data are used for training. Our method belongs to *transductive* semi-supervised learning (Chapelle, Scholkopf, and Zien 2009) where learning is to infer the correct labels for the given unlabeled data only.

For better understanding the differences between most existing deep learning models and our method, we present a comparison of them from three perspectives in Table 1.

Competitive Reconstruction Auto-encoder

In this section, we present detailed description about our competitive reconstruction auto-encoder (*CoRA*). We introduce the architecture and the objective function of *CoRA* first and then we show how to optimize this model.

Model Architecture

The proposed competitive reconstruction auto-encoder is composed of three main modules: (1) encoder network E, (2) inlier decoder network $\mathcal{D}_{in}$, and (3) outlier decoder network D_{out}. The encoder maps the data into a common subspace shared by inliers and outliers. D_{in} performs the reconstruction for inlier samples while D_{out} acts as a reconstructor for outliers. D_{in} and D_{out} work in a competitive way as they both try to give low reconstruction error for the input samples. However, each sample could only be interpreted by one decoder. The architecture of the proposed model is shown in Fig.2. It can be seen that both positive training data and unlabeled test data are feeded into E. As the training data are given as positive, we directly select D_{in} to reconstruct them. For unlabeled test data, the label is determined by the reconstruction errors of two decoders.

Formally, let X^p be the positive training data where X_i^p is the i-th sample of X^p, $i = 1, ...m$. Similarly, let X^u be the unlabeled data and X_j^u is the j-th sample of X^u, where $j = 1, ..., n$. In order to assign labels for each X_j^u, we propose the following loss function.

Competitive Reconstruction Loss

Previous work (Xia et al. 2015) showed that the reconstruction error of an auto-encoder, trained on samples from the target class, is a useful measure for novelty sample detection. Concretely, if an auto-encoder is trained to reconstruct target class samples (inliers), the reconstruction error for novelty samples would be high. In order to model the distributions of positive data and outliers separately. We design a competitive mechanism for our model. By the guidance of positive data, D_{in} is trained to learn the distribution of target class. For unlabeled data which may be inliers or outliers, if the reconstruction error of D_{in} is small than that of D_{out}, it could be an inlier with high probability. In this paper, we use mean square error as the reconstruction error for all our experiments. For simplicity, we define the reconstructions for sample x as $\mathcal{R}_{in}(x) = D_{in}\Big(E(x)\Big)$ and $\mathcal{R}_{out}(x) = D_{out}\Big(E(x)\Big)$.

The competitive reconstruction loss is defined as

$$\begin{aligned}\mathcal{L}_{\mathcal{CR}} &= \left\|X^p - \mathcal{R}_{in}\Big(X^p\Big)\right\|_2^2 \\ &+ \sum_{j=1}^{n}\Bigg(y_j \left\|X_j^u - \mathcal{R}_{in}\Big(X_j^u\Big)\right\|_2^2 \\ &+ (1-y_j)\left\|X_j^u - \mathcal{R}_{out}\Big((X_j^u\Big)\right\|_2^2\Bigg)\end{aligned} \tag{1}$$

where $y_j, j = 1, ..., n$ is the label assignment for the j-th sample, and it is evaluated by

$$y_j = \begin{cases} 1, & \left\|X_j^u - \mathcal{R}_{in}\Big(X_j^u\Big)\right\|_2^2 < \left\|X_j^u - \mathcal{R}_{out}\Big((X_j^u\Big)\right\|_2^2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Note that y_j is updated for each iteration. We will give detailed description on how to optimize the model later. We omit the average of the sum notation for clarity.

As mentioned before, *CoRA* maps different data to different regions in the subspace. Thus, an intuitive idea is to preserve the data structures of positive data and abnormal data in the subspace, so that similar examples stay close in subspace. We propose a structure-preserving regularization term to improve the performance of our model.

Subspace Structure-preserving Regularization

In addition to competitive loss, we also propose a regularization term to keep the data structure in subspace. Subspace regularization has been explored in previous works (Agarwal, Gerber, and Daume 2010; Liu, Hua, and Smith 2014; Yao et al. 2015). Here we employ a block symmetric affinity matrix, as we use both positive data and unlabeled test data for training. Formally, we build the affinity matrix of the training samples as W:

$$W = \begin{cases} \exp\left(\frac{-D(x_i,x_j)}{\epsilon^2}\right) & i \in \mathcal{N}_j \text{ or } j \in \mathcal{N}_i \\ 0 & \text{else} \end{cases} \tag{3}$$

where $D(x_i, x_j)$ is the distance measure of the data. $\mathcal{N}_i$ is the neighborhood of the i-th data point. $\epsilon > 0$ is the bandwidth parameter. In our method, W can be formulated as

$$W = \begin{bmatrix} W_{pp} & W_{pu} \\ W_{up} & W_{uu} \end{bmatrix} \tag{4}$$

where W_{pp} and W_{uu} indicate the affinity matrix among positive data and unlabeled data respectively, W_{pu} and W_{up} all represent the affinity matrix between positive data and unlabeled data. As for unlabeled data, we set W_{uu} to a zero matrix in order to avoid subspace structure being destroyed by the uncertain structure of unlabeled data. In other words, we do not consider the manifold structure among unlabeled data due to the existence of uncertainty.

The manifold regularizer can be defined as

$$\mathcal{L}_{\mathcal{R}} = \sum_{i,j=1}^{n+m} W_{ij}\Big(E(x_i) - E(x_j)\Big)^2 \tag{5}$$

where $E(x_i)$ denotes the encoder feature of sample x_i. This regularization term can force the encoder to map similar examples into a neighborhood. Note that there may be a confusion: the sizes of positive data and unlabeled data are different, how could we formalize W in Eq. (4). It is a problem related to the optimization algorithm, which will be explained in the next subsection.

```
Algorithm 1 Training Algorithm of CoRA
  Input: positive data $X^p$, unlabeled test data $X^u$, encoder parameter $\theta_E$, decoder parameters $\theta_{\mathcal{D}_{in}}, \theta_{\mathcal{D}_{out}}$
  Initialize parameters.
  repeat
    Sample an unlabeled batch data $B^u$ and a positive batch data $B^p$, $|B^p| = |B^u|$.
    /* Forward propagation */
    $\hat{B}^p = \mathcal{D}_{in}\Big(E(B^p)\Big)$
    $\hat{B}^{iu} = \mathcal{D}_{in}\Big(E(B^u)\Big)$ /* Reconstruction on $\mathcal{D}_{in}$*/
    $\hat{B}^{ou} = \mathcal{D}_{out}\Big(E(B^u)\Big)$ /* Reconstruction on $\mathcal{D}_{out}$*/
    for each unlabeled sample $i = 1$ to $|B^u|$ do
      if $\left\|\hat{B}_i^{iu} - B_i^u\right\|_2^2 < \left\|\hat{B}_i^{ou} - B_i^u\right\|_2^2$ then
        Assign $B_i^u$ inlier decoder $\mathcal{D}_{in}$.
      else
        Assign $B_i^u$ outlier decoder $\mathcal{D}_{out}$.
      end if
      Compute $\mathcal{L}_{\mathcal{CR}}$ according to Eq. 1
    end for
    Compute loss $\mathcal{L}_{\mathcal{R}}$ with Eq. 3, 4, 5.
    Optimize $\mathcal{L}$ with SGD, and back propagate.
  until $convergence$
```

Optimization

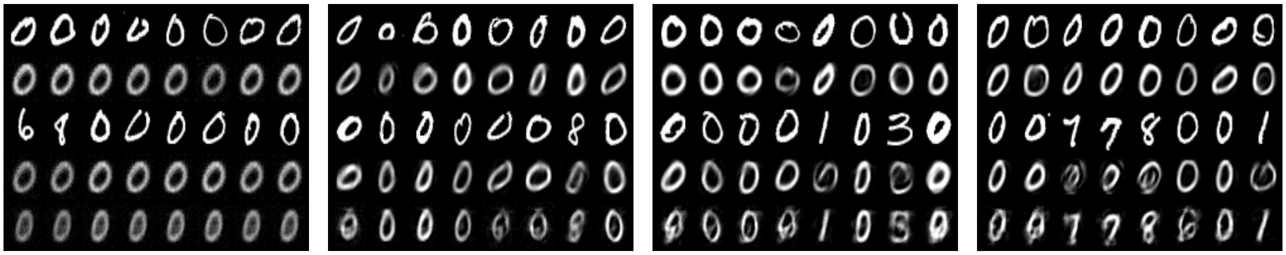
Competitive reconstruction auto-encoder is optimized to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{\mathcal{CR}} + \lambda\mathcal{L}_{\mathcal{R}} \tag{6}$$

where $\lambda > 0$ is a hyperparameter that controls the relative importance of the regularization terms. We tuned λ for different tasks and observed that $\lambda = 0.1$ produces promising results for all our experiments. To optimize this loss function, stochastic gradient descent (SGD) method is adopted for training our model. As we use mini-batch SGD, for each iteration we sample the same number of positive data such that W can be calculated as in Eq. (4). The detailed optimization algorithm is described in Alg. 1.

Discussions

The learning procedure. To understand the learning procedure of *CoRA*, we explore the reconstruction results of inlier decoder and outlier decoder for different learning iterations. As shown in Fig. 3, in the very beginning of training, both decoders give similar reconstruction for different samples. However, after a few epochs, the inlier (positive) decoder begins to generate better reconstruction on target class



(a) Initialization. (b) Epoch = 20. (c) Epoch = 50. (d) Epoch = 100.

Figure 3: Reconstruction results of *CoRA* after different training iterations. In each figure, samples in the 1st line are positive data, the 2nd line gives the reconstruction of the inlier decoder for the positive samples. Samples in the 3rd line are unlabeled data, the 4th line gives the reconstruction of the inlier decoder and the 5th line gives the reconstruction of the outlier decoder, all for the unlabeled samples.

than the outlier decoder. Finally, with the convergence of the learning procedure, the inlier decoder generates bad reconstruction for outliers which like distorted positive samples. Although the outlier decoder can output understandable reconstructions for positive samples, they all have some distortion. Thus, the positive samples are assigned to the inlier decoder, and outlier samples are assigned to outlier decoder.

Furthermore, we investigate the distributions of reconstruction residual between the inlier decoder and the outlier decoder on positive data/inliers and outliers during training. The residual is defined as $R_{residual}(x) = \|x - \mathcal{R}_{out}(x)\|_2^2 - \|x - \mathcal{R}_{in}(x)\|_2^2$. The results are shown in Fig. 4. We can see that both decoders have similar reconstruction error for inliers and outliers at the initialization stage. However, as the training proceeds, the residuals are different for inliers and outliers. After about 50 epochs, the learning process begins to converge and the discrimination between inliers and outliers becomes more and more clear. However, as there always exist some outliers that are hard to tell from inliers, thus there is a small overlap area.

The merits of CoRA. As mentioned before, our model does not require a pre-defined threshold to discriminate outliers. This is quite important for many real-world applications. Another merit of our algorithm is that it does not suffer from the overfitting problem while most reconstruction-based models do. When the learning process converges, the assignment of each unlabeled sample rarely changes, which means CoRA is robust. Moreover, we do not impose any restriction on the design of the auto-encoder. By exploiting the representation learning ability of deep autoecoder, our model is flexible for various data. Meanwhile, mini-batch gradient descent optimization makes our approach efficient in handling large-scale datasets, which is a basic requirement in the big data era.

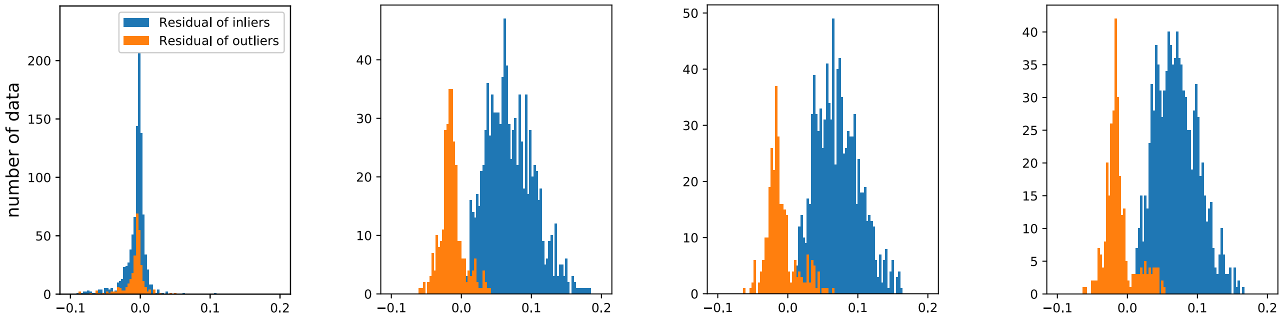
Performance Evaluation

In this section, we evaluate CoRA for anomaly detection task, and compare it with existing methods including the state of the art ones. We use seven datasets including low-dimensional data and image data.

Datasets

- **KDD99**. A classical anomaly detection dataset that records the network attacks and normal connections. We use a subset of it which is referred as *KDD99 10-percent*. There are 494,021 instances, where 30% are outliers.
- **MNIST**. It has 70,000 training and test samples from 10 digit classes. For each class, we select the training positive data and simulate outliers from the test data with different ratios, from 10% to 50%.
- **Fashion MNIST**. The dataset composes of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28×28 grayscale image, associated with a label from 10 classes, each of which is a cloth category. We preprocess it as for MNIST.
- **ImageNet-20**, This dataset consists of images in 20 semantic concepts from ImageNet dataset, and each concept contains about 2,800 images on average. For each concept, outliers are simulated from the other concepts with a ratio of 30%.
- **Caltech-101**. This dataset consists of 101 classes of images. Following previous work (Zhai et al. 2016), we choose the 11 object categories that contain at least 100 images. For each category, outliers are sampled from other categories with a ratio 30%.
- **CIFAR-10**. There are 50,000 training and 10,000 test images from 10 categories. Similar to Caltech-101, outliers are sampled from other categories at the ratio of 30%.
- **Caltech-256**. This dataset contains 256 object classes with a total of 30,607 images. Each category has at least 80 images. Following previous works (You, Robinson, and Vidal 2017; Sabokrou et al. 2018), we randomly select images from $n \in \{1, 3, 5\}$ categories as inliers, and for those categories that have more than 150 images, only the first 150 images are used. A certain number of outliers are randomly selected from the "clutter" category, such that each experiment has exactly 50% outliers. We repeat each experiment three times and report the average performance.

In most of our experiments, images are represented by features extracted from a pre-trained Vgg-16 network (Si-



(a) Initialization.

(b) Epoch = 20.

(c) Epoch = 50.

(d) Epoch = 100.

Figure 4: The evolution of reconstruction residual in the learning process of CoRA.

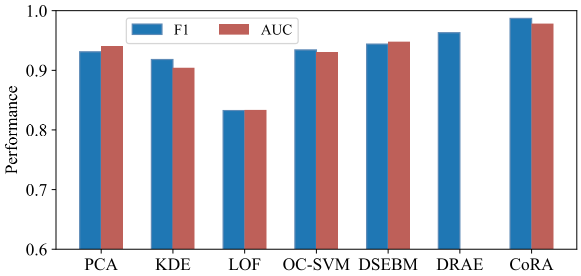


Figure 5: Performance comparison on KDD99 dataset.

monyan and Zisserman 2014) on ImageNet. We use the features after the first fully connected layer and the feature dimension is 4096. On MNIST and Fashion-MNIST, we directly use raw pixels as inputs for all methods. As some approaches are restricted to simple multiple-layer neural networks, we use the same architecture for most of the networks for fair comparison.

Evaluation Metrics Similar to previous works (Sabokrou et al. 2018), we utilize *F1* and *AUC* scores to measure performance. For most image datasets, we report the mean performance of all classes. Note that for discriminative algorithms such as DRAE, AUC may not suitable for evaluation. For our model, we calculate AUC with the transformed reconstruction residual in order to compare it with other supervised approaches.

Experimental Results

Results on KDD99 We first evaluate our *CoRA* method on low-dimensional dataset. Since feature dimension of KDD99 is 41 which is quite small, we compare *CoRA* with several traditional methods that are effectiveness on low-dimension datasets, including PCA, KDE, LOF and OC-SVM. For deep learning models, we select deep energy-based model (DSEBM) (Zhai et al. 2016) and discriminative auto-encoder (DRAE) (Xia et al. 2015) for comparison. The results are shown in Fig. 5. We can see that *CoRA* achieves the best performance in terms of both *F1* and *AUC* scores. Although conventional methods such as PCA and OC-SVM perform comparably to deep neural network models including DSEBM and DRAE, the training of deep learning models is about 10 times faster than the conventional methods.

Table 2: Performance on ImageNet-20, Caltech-101 and CIFAR-10. The best performance in each column is in **bold.**

Dataset	ImageNet-20		Caltech-101		CIFAR-10	
Method	F1	AUC	F1	AUC	F1	AUC
AE	0.911	0.907	0.919	0.923	0.828	0.822
DSEBM	0.926	0.923	0.923	0.892	0.867	0.872
DRAE	0.930	-	**0.968**	-	0.862	-
DAOC	0.945	0.939	0.956	0.944	0.876	0.869
CoRA	**0.960**	**0.964**	0.954	**0.985**	**0.910**	**0.897**

Results on MNIST/Fashion-MNIST Raw pixels are used as input features for these two datasets, and the feature dimension is 784. In our experiments, the architecture of the encoder is $[784, 64, 32]$, the decoders use a symmetric architecture. As PCA is a special case of auto-encoder with a linear hidden layer, we omit PCA-based models in these experiments. We compare *CoRA* with several state-of-the-art methods including DSEBM, DRAE and deep adversarial one-class learning (DAOC) method (Sabokrou et al. 2018). auto-encoder (AE) is used as the baseline algorithm. The results are shown in Fig. 6.

When the outlier ratio is small (*e.g.* 0.1), DAOC slightly outperforms CoRA on both datasets. The reason is that when the number of outliers is small, the outlier decoder of our model is not able to learn the anomaly data distribution well. However, CoRA still achieves comparable performance to unsupervised method DRAE and the other state-of-the-art algorithms, and it also outperforms the other traditional methods including AE.

With the increase of outlier ratio, most methods perform worse. However, CoRA shows its robustness to outlier ratio. As we can see, CoRA significantly outperforms all the other methods on MNIST dataset when outlier ratio is larger than 0.2. Although similar trend can be observed on Fashion-MNIST, most state-of-the-art methods perform worse than on MNIST, since images in Fashion-MNIST are much more blurred than those in MNIST.

Table 3: Performance comparison on Caltech-256. Inliers are images taken from one, three, or five randomly chosen categories, and outliers are randomly selected from category 257 — clutter with a ratio of 50%. In each row the best result is in **bold** and the second best in *italic* typeface.

# categories	Measure	CoP	REAPER	OutlierPursuit	LRR	DPCP	R-graph	SSGAN	DAOC	*CoRA*
1	AUC	0.905	0.816	0.837	0.907	0.783	*0.948*	-	0.942	**0.968**
1	$F1$	0.880	0.808	0.823	0.893	0.785	0.914	**0.977**	0.928	*0.967*
3	AUC	0.676	0.796	0.788	0.479	0.798	0.929	-	*0.938*	**0.962**
3	$F1$	0.718	0.784	0.779	0.671	0.777	0.880	**0.963**	0.913	**0.963**
5	AUC	0.487	0.657	0.629	0.337	0.676	0.913	-	*0.923*	**0.952**
5	$F1$	0.672	0.716	0.711	0.667	0.715	0.858	*0.945*	0.905	**0.950**

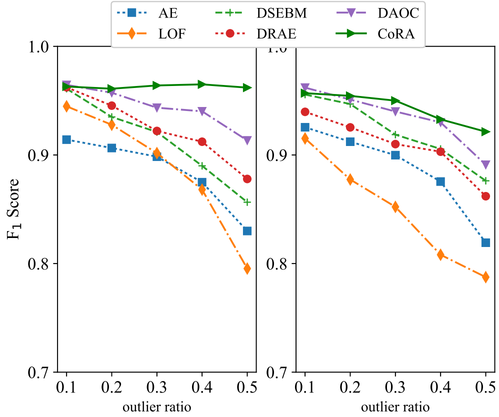


Figure 6: Performance comparison on MNIST (left) and Fashion-MNIST (right).

Results on ImageNet-20/Caltech-101/CIFAR-10 Except for DAOC (Sabokrou et al. 2018), most algorithms represent images by deep learning features. As DAOC uses auto-encoder as the generator and is trained in adversarial manner, we use the original pixel features as inputs to DAOC. The architecture of auto-encoder for ImageNet-20 and CIFAR10 has two hidden layers with 256 and 128 units respectively. As there are too few samples in Caltech-101, we use a smaller architecture with two hidden layers. The encoder architecture is $[4096, 64, 32]$. We use *ReLU* as the activation function of the hidden layers. Table. 2 presents the results of state-of-the-art methods and AE. On ImageNet-20, DSEBM and DRAE achieve similar results while DAOC outperforms both of them.

CoRA performs best on ImageNet-20. For Caltech-101, DRAE achieves the best result in terms of *F1* while CoRA achieves the best result in terms of *AUC*. Meanwhile, CoRA and DAOC have close performance in terms of *F1* measure. As mentioned before, the number of images in each concept of Caltech-101 is less than that in ImageNet-20 and CIFAR-10, it may not be sufficient for DAOC and CoRA to learn the inlier concept. As for CIFAR-10, DRAE slightly underperforms DSEBM, while DAOC outperforms the rest of approaches but CoRA. CoRA significantly outperforms the other methods in terms of both *F1* and *AUC*.

Results on Caltech-256

As there are less images in each category, detecting outliers from Caltech-256 is much challenging. Considering that Caltech-256 is a benchmark dataset for outlier detection task, in addition to DAOC, we compare our method with 7 other methods therein designed specifically for detecting outliers. Those methods include Coheerence Pursuit (CoP) (Rahmani and Atia 2017), OutlierPursuit (Xu, Caramanis, and Sanghavi 2010), REAPER(Lerman et al. 2015), Dual Principal Component Pursuit (DPCP) (Tsakiris and Vidal 2015), Low-Rank Representation (LRR) (Liu, Lin, and Yu 2010), OutRank (Moonesignhe and Tan 2006), and inductive semi-supervised GAN (SSGAN) (Kimura and Yanagihara 2018).

We use similar setup as in (You, Robinson, and Vidal 2017), the experimental results are presented in Table 3. The results of the other methods are from (Sabokrou et al. 2018) and (Kimura and Yanagihara 2018). These results show that even for a small number of training samples, our method performs at least as well as the state-of-the-art algorithms, and in many cases it superior to them. Furthermore, when the inliers come from only one categories, CoRA slightly underperforms SSGAN. When the number of inlier classes increases to 3 or 5, CoRA outperforms the other 8 methods in terms of both $F1$ and AUC.

Conclusion

In this paper, we propose a competitive reconstruction auto-encoder model for transductive semi-supervised anomaly detection task. Our model learns from positive data and unlabeled test data, and predicts the labels of unlabelled data directly after learning. The two decoders are designed to compete with each other to achieve lower reconstruction error. With the guidance of positive data, inlier decoder is more likely to build the configuration of positive class. The new discriminative criterion does not need a pre-defined threshold, which is different from most existing methods. Moreover, we adopt SGD to optimize this model, which enables it to be efficient and scalable for large-scale datasets. Experimental results on seven benchmark datasets shown that this model can beat many state-of-the-art methods.

Acknowledgement: This work was supported by NSFC under gant No. U1636205. Jihong Guan was supported by NSFC under grant No. 61772367.

References

Agarwal, A.; Gerber, S.; and Daume, H. 2010. Learning multiple tasks using manifold regularization. In *Advances in Neural Information Processing Systems*, 46–54.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1798–1828.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58(3):11.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41(3):15.

Chapelle, O.; Scholkopf, B.; and Zien, A. 2009. *Semi-Supervised learning*. MITPress.

De La Torre, F., and Black, M. J. 2003. A framework for robust subspace learning. *International Journal of Computer Vision* 54(1-3):117–142.

Elkan, C., and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220. ACM.

Eskin, E. 2000. Anomaly detection over noisy data using learned probability distributions. In *In Proceedings of the International Conference on Machine Learning*. Citeseer.

Khan, S. S., and Madden, M. G. 2009. A survey of recent trends in one class classification. In *Irish Conference on Artificial Intelligence and Cognitive Science*, 188–197. Springer.

Khan, S. S., and Madden, M. G. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review* 29(3):345–374.

Kim, J., and Scott, C. D. 2012. Robust kernel density estimation. *Journal of Machine Learning Research* 13(Sep):2529–2565.

Kimura, M., and Yanagihara, T. 2018. Semi-supervised anomaly detection using gans for visual inspection in noisy training data. *arXiv preprint arXiv:1807.01136*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

Lerman, G.; McCoy, M. B.; Tropp, J. A.; and Zhang, T. 2015. Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics* 15(2):363–410.

Liu, W.; Hua, G.; and Smith, J. R. 2014. Unsupervised one-class learning for automatic outlier removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3826–3833.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning*, 663–670.

Malhotra, P.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; and Shroff, G. 2016. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*.

Moonesignhe, H., and Tan, P.-N. 2006. Outlier detection using random walks. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, 532–539. IEEE.

Parzen, E. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* 33(3):1065–1076.

Rahmani, M., and Atia, G. K. 2017. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Transactions on Signal Processing* 65(23):6260–6275.

Sabokrou, M.; Khalooei, M.; Fathy, M.; and Adeli, E. 2018. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3379–3388.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tsakiris, M. C., and Vidal, R. 2015. Dual principal component pursuit. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 10–18.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11(Dec):3371–3408.

Wei, Q.; Ren, Y.; Hou, R.; Shi, B.; Lo, J. Y.; and Carin, L. 2018. Anomaly detection for medical images based on a one-class classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, 105751M. International Society for Optics and Photonics.

Xia, Y.; Cao, X.; Wen, F.; Hua, G.; and Sun, J. 2015. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, 1511–1519.

Xu, H.; Caramanis, C.; and Mannor, S. 2013. Outlier-robust pca: the high-dimensional case. *IEEE Transactions on Information Theory* 59(1):546–572.

Xu, H.; Caramanis, C.; and Sanghavi, S. 2010. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, 2496–2504.

Yao, T.; Pan, Y.; Ngo, C.-W.; Li, H.; and Mei, T. 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2142–2150.

You, C.; Robinson, D. P.; and Vidal, R. 2017. Provable self-representation based outlier detection in a union of subspaces. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 1–10.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, 1100–1109.