GAN Ensemble for Anomaly Detection

Xu Han[*], **Xiaohui Chen**[*], **Li-Ping Liu**

Department of Computer Science, Tufts University
{Xu.Han, Xiaohui.Chen, Liping.Liu}@tufts.edu

Abstract

When formulated as an unsupervised learning problem, anomaly detection often requires a model to learn the distribution of normal data. Previous works modify Generative Adversarial Networks (GANs) by using encoder-decoders as generators and then apply them to anomaly detection tasks. Related studies also indicate that GAN ensembles are often more stable than single GANs in image generation tasks. In this work, we propose to construct GAN ensembles for anomaly detection. In this new method, a group of generators interact with a group of discriminators, so every generator gets feedback from every discriminator, and vice versa. Compared to a single GAN, an ensemble of GANs can better model the distribution of normal data and thus better detect anomalies. We also make a theoretical analysis of GANs and GAN ensembles in the context of anomaly detection. The empirical study constructs ensembles based on four different types of detecting models, and the results show that the ensemble outperforms the single model for all four model types.

Introduction

Anomaly detection is an important problem in machine learning and has a wide range of applications such as fraud detection (Abdallah, Maarof, and Zainal 2016; Kou et al. 2004), intrusion detection (Garcia-Teodoro et al. 2009), and event detection (Atefeh and Khreich 2015). In most anomaly detection problems, only samples of normal data are given, and the task is to detect *anomalies* that deviates from normal data. A large class of anomaly detection algorithms (Hodge and Austin 2004; Gupta et al. 2013; Chalapathy and Chawla 2019) directly or indirectly model the data distribution and then report samples atypical in the distribution as anomalies.

Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), which are flexible models for learning data distributions, provide a new set of tools for anomaly detection. A GAN consists of a generator network and a discriminator network, which learn a data distribution through a two-player game. Several studies apply GANs to anomaly detection, such as AnoGAN (Schlegl et al. 2017), f-AnoGAN (Schlegl et al. 2019), EGBAD (Zenati et al. 2018), GANomaly (Akcay, Atapour-Abarghouei, and Breckon 2018), and Skip-GANomaly (Akcay, Atapour-Abarghouei, and Breckon 2019). All these models use an encoder-decoder as the generator since the generator of a vanilla GAN cannot check a sample. A synthetic sample from the decoder is either a reconstruction of a real sample or a new sample. The discriminator needs to distinguish synthetic samples and normal samples, so the discriminator acquires the ability of differentiate anomalies from normal data. For the discriminator, the training is the same as binary classification but avoids manual labeling of training data (Liu and Fern 2012). The anomaly score is usually computed by checking the reconstruction of a sample and the internal representation of a sample in the discriminator. These models perform well in a series of detection tasks.

There still lacks a thorough understanding of the role of adversarial training in anomaly detection. Theoretical analysis (Goodfellow et al. 2014; Arora et al. 2017) indicates that the discriminator should do no better than random guessing at the end of an ideal training procedure. However, the internal representation of a discriminator is very effective in differentiating normal samples and anomalies in practice. Then there is a gap between theory and practice: how does the discriminator characterize normal samples' distribution?

Training a GAN is challenging because it needs to optimize multiple deep neural networks in a min-max problem. The optimization often has stability issues. If neural networks are not carefully regularized, there are also problems of mode collapse. Recent works show multiple generators or/and discriminators help to overcome those problems. Several studies (Durugkar, Gemp, and Mahadevan 2016; Neyshabur, Bhojanapalli, and Chakrabarti 2017) use multiple discriminators to provide stable gradients to the generator, making the training process more smooth. Multiple generators, which essentially defines a distribution mixture, can capture multiple data modes (Hoang et al. 2018). Arora et al. (2017) analyzes the equilibrium of GAN models and show that a mixture of both generators and discriminators guarantees an approximate equilibrium in a GAN's min-max game. These ensemble methods have improved the performance in various generation tasks.

In this work, we propose to use GAN ensembles for anomaly detection. A GAN ensemble consists multiple

[*] The first two authors contribute equally.

encoder-decoders and discriminators, which are randomly paired and trained via adversarial training. In this procedure, an encoder-decoder gets feedback from multiple discriminators while a discriminator sees "training samples" from multiple generators. The anomaly score is the average of anomaly scores computed from all encoder-decoder-discriminator pairs. This ensemble method works for most existing GAN-based anomaly detection models.

We further analyze GAN models in the context of anomaly detection. The theoretical analysis refines the function form of the optimal discriminator in a WGAN (Arjovsky, Chintala, and Bottou 2017). The analysis further explains why the discriminator helps to identify anomalies and how GAN ensemble improves the performance.

In the empirical study, we test the proposed ensemble method on both synthetic and real datasets. The results indicate that ensembles significantly improve the performance over single detection models. The empirical analysis of vector representations verifies our theoretical analysis.

Background

The anomaly detection problem is formally defined as follows. Let $\mathbf{X}$ be a training set of N normal data samples, $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^d : n = 1, \ldots, N\}$ from some unknown distribution $\mathcal{D}$. There is also a test sample $\mathbf{x}' \in \mathbb{R}^d$, which may or may not be from the distribution $\mathcal{D}$. The problem of anomaly detection is to train a model from $\mathbf{X}$ such that the model can classify $\mathbf{x}'$ as *normal* if $\mathbf{x}'$ is from the distribution $\mathcal{D}$ or *abnormal* if $\mathbf{x}'$ is from a different distribution. The model often computes an anomaly score $y' \in \mathbb{R}$ for $\mathbf{x}'$ and decide the label of $\mathbf{x}'$ by thresholding y'.

Anomaly detection essentially depends on the characterization of the distribution of normal samples. Adversarial training, which is designed to learn data distributions, suits the task well. Models based on adversarial training usually have an encoder-decoder as the generator and a classifier as the discriminator. Previous models such as f-AnoGAN, EGBAD, GANomaly, and Skip-GANomaly all share this architecture. It is important to have an encoder-decoder as the generator because a detection model often needs the low-dimensional encoding of a new sample. Below is a review of these models.

The generator consists of an encoder $G_e(\cdot;\phi) : \mathbb{R}^d \to \mathbb{R}^{d'}$, which is parameterized by ϕ, and a decoder $G_d(\cdot;\psi) : \mathbb{R}^{d'} \to \mathbb{R}^d$, which is parameterized by ψ. The encoder maps a sample $\mathbf{x}$ to an *encoding vector* $\mathbf{z}$ while the decoder computes a reconstruction $\tilde{\mathbf{x}}$ of the sample from $\mathbf{z}$.

$$\mathbf{z} = G_e(\mathbf{x};\phi), \;\; \tilde{\mathbf{x}} = G_d(\mathbf{z};\psi). \tag{1}$$

Skip-GANomaly uses U-Net with skip connections as the encoder-decoder, which does not compute $\mathbf{z}$ explicitly.

The discriminator $D(\cdot;\gamma)$, which is parameterized by γ, takes a sample and predicts the probability of the sample being from the data $\mathbf{X}$ instead of the generator. In the context of WGAN, the critic function plays a similar role as a discriminator, so it is also denoted by $D(\cdot;\gamma)$ and called as "discriminator" for easy discussion. The discriminator $D(\cdot;\gamma)$ should give higher values to normal data samples than reconstructions. The discriminator from a vanilla GAN and a WGAN has the form of $u = D(\mathbf{x};\gamma)$. In a EGBAD model, which is based on a BiGAN, the discriminator takes both a sample and its encoding as the input, so it has the from of $u = D((\mathbf{x}, G_e(\mathbf{z}));\gamma)$.

Since a model consists of an encoder-decoder and a discriminator, the training often considers losses inherited from both models. The *adversarial loss* is from GAN training. The losses are defined as follows when the discriminator works in three GAN types (vanilla GAN, WGAN, and BiGAN).

$$L_{a\text{-}g}(\mathbf{x}) = \log D(\mathbf{x}) + \log\big(1 - D(G_d(G_e(\mathbf{x})))\big) \tag{2}$$

$$L_{a\text{-}wg}(\mathbf{x}) = D(\mathbf{x}) - D\big(G_d(\tilde{\mathbf{z}})\big) \tag{3}$$

$$L_{a\text{-}bg}(\mathbf{x}) = \log D(\mathbf{x}, G_e(\mathbf{x})) + \log\big(1 - D(G_d(\tilde{\mathbf{z}}), \tilde{\mathbf{z}})\big) \tag{4}$$

In $L_{a\text{-}g}(\mathbf{x})$, the sample from the generator is the reconstruction of the original sample. The WGAN objective will be used by f-AnoGAN, which does not train the generator $G_e(\cdot;\phi)$ in the objective, so $L_{a\text{-}wg}(\mathbf{x})$ has no relation with ϕ. An encoding $\tilde{\mathbf{z}}$ in (3) or (4) is sampled from a prior distribution $p(\mathbf{z})$, e.g. multivariate Gaussian distribution, so $L_{a\text{-}wg}(\mathbf{x})$ and $L_{a\text{-}bg}(\mathbf{x})$ are considered as stochastic functions of $\mathbf{x}$.

The second type of loss is the *reconstruction loss*, which is inherited from encoder-decoders. The loss is the difference between a reconstruction and the original sample. The difference is measured by ℓ-norm with either $\ell = 1$ or $\ell = 2$.

$$L_r(\mathbf{x}) = \|\mathbf{x} - G_d\big(G_e(\mathbf{x})\big)\|_\ell^\ell \tag{5}$$

Previous research also indicates that the *hidden vector* $\mathbf{h}$ of a sample in the last hidden layer of the discriminator $D(\cdot;\gamma)$ is often useful to differentiate normal and abnormal samples. Denote $\mathbf{h} = f_D(\mathbf{x};\gamma)$ as the hidden vector of $\mathbf{x}$ in $D(\mathbf{x};\gamma)$, then the *discriminative loss* based on $\mathbf{h}$ is

$$L_d(\mathbf{x}) = \|f_D(\mathbf{x}) - f_D\big(G_d(G_e(\mathbf{x}))\big)\|_\ell^\ell \tag{6}$$

The GANomaly model also considers the difference between the encoding vectors of a normal sample $\mathbf{x}$ and its reconstruction $\tilde{\mathbf{x}}$. Particularly, it uses a separate encoder $G_e(\cdot;\tilde{\phi})$ to encode the recovery $\tilde{\mathbf{x}}$. Then the *encoding loss* is

$$L_e(\mathbf{x}) = \|G_e(\mathbf{x};\phi) - G_e\big(G_d(G_e(\mathbf{x};\phi));\tilde{\phi}\big)\|_\ell^\ell \tag{7}$$

Equation (7) explicitly indicates that the parameters ϕ and $\tilde{\phi}$ of the two encoders are different.

To train their discriminators, these GAN models need to maximize the adversarial loss.

$$\max_{\gamma} \quad \sum_{i=1}^{N} L_a(\mathbf{x}_i;\phi,\psi,\gamma) \tag{8}$$

Here L_a can be any of the three losses in (2), (3), and (4). This equation explicitly shows all trainable parameters in the loss function.

To train their generators, different models minimize some of the four losses mentioned above. We write their objective in a general form.

$$\min_{\phi,\psi,\tilde{\phi}} \sum_{i=1}^{N} \alpha_1 L_a(\mathbf{x}_i;\phi,\psi,\gamma) + \alpha_2 L_r(\mathbf{x}_i;\phi,\psi) + \alpha_3 L_e(\mathbf{x}_i;\phi,\psi,\tilde{\phi}) + \alpha_4 L_d(\mathbf{x}_i;\phi,\psi,\gamma) \quad (9)$$

F-AnoGAN (Schlegl et al. 2019) trains the decoder and encoder separately: it first trains the decoder $G_d(\cdot;\psi)$ and the discriminator $D(\cdot;\gamma)$ by setting $\alpha_2 = \alpha_3 = \alpha_4 = 0$, then it trains the encoder with the decoder and the discriminator fixed and $\alpha_1 = \alpha_3 = 0$.

After training a model, we need to compute an anomaly score $A(\mathbf{x}')$ for a test instance $\mathbf{x}'$. The anomaly score is usually a weighted sum of the reconstruction loss and the discriminative loss.

$$A(\mathbf{x}') = L_r(\mathbf{x}') + \beta L_d(\mathbf{x}') \quad (10)$$

The relative weight β is usually empirically selected. The exception is GANomaly, which uses the encoding loss $A(\mathbf{x}') = L_e(\mathbf{x}')$ as the anomaly score. The higher an anomaly score is, the more likely $\mathbf{x}'$ is an anomaly.

Proposed Approach

In this work, we propose an ensemble learning framework for anomaly detection based on GAN models. We will have multiple generators and discriminators, which are parameterized differently. We define I generators, $\mathcal{G} = \{(G_e(\cdot;\phi_i), G_d(\cdot;\psi_i)) : i = 1,\ldots,I\}$, and J discriminators $\mathcal{D} = \{D(\cdot;\gamma_j) : j = 1,\ldots,J\}$. A single generator or discriminator is the same as that of a base model. GANomaly has two encoders, so its generators are $\mathcal{G} = \left\{\left(G_e(\cdot;\phi_i), G_d(\cdot;\psi_i); G_e(\cdot;\tilde{\phi}_i)\right) : i = 1,\ldots,I\right\}$ in this framework. We will omit $G_e(\cdot;\tilde{\phi}_i)$ in the following discussion, but the optimization with GANomaly should be clear from the context.

In the adversarial training process, we pair up every generator with every discriminator. Then a generator is critiqued by every discriminator, and a discriminator receives synthetic samples from every generator. In our method, we do not use weights for generators or discriminators as (Arora et al. 2017) since we want to sufficiently train every generator and every discriminator.

With multiple pairs of generators and discriminators, the adversarial loss and the discriminative loss are both computed from all generator-discriminator pairs. Denote the losses between each generator-discriminator pair (i, j) by

$$L_a^{ij} = L_a(\mathbf{x};\phi_i,\psi_i,\gamma_j), \quad L_d^{ij} = L_d(\mathbf{x};\phi_i,\psi_i,\gamma_j). \quad (11)$$

Here $L_a(\mathbf{x};\phi_i,\psi_i,\gamma_j)$ can be any of the three adversarial loss functions defined in (2), (3), and (4). We explicitly write out parameters of the pair of generator and discriminator to show the actual calculation.

Similarly denote the recovery loss and the encoding loss of a single generator i by

$$L_r^i = L_r(\mathbf{x};\phi_i,\psi_i), \quad L_e^i = L_e(\mathbf{x};\phi_i,\psi_i). \quad (12)$$

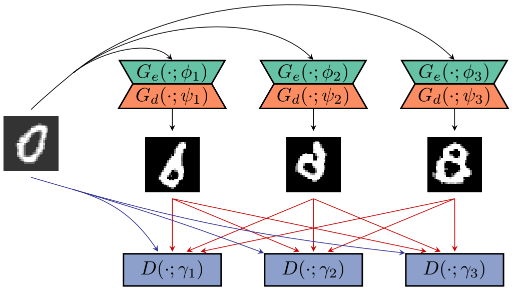

Figure 1. GAN ensemble with multiple generators and discriminators. The "0" image on the left is an anomaly, and the three images from the three encoder-decoders are reconstructions of the "0" image.

Then we maximize the sum of adversarial losses to train discriminators while minimize the sum of all losses to train generators. The training objectives are as follows.

$$\max_{(\gamma_j)_{j=1}^{J}} \sum_{i=1}^{I}\sum_{j=1}^{J} L_a^{ij} \quad (13)$$

$$\min_{(\phi_i,\psi_i)_{i=1}^{I}} \sum_{i=1}^{I}\sum_{j=1}^{J} \alpha_1 L_a^{(ij)} + \alpha_2 L_r^{(i)} + \alpha_3 L_d^{(ij)} + \alpha_4 L_e^{(i)} \quad (14)$$

This ensemble method works with all base models reviewed in the last section. The training of the ensemble is shown in Figure 1.

Batch training In one training iteration we update only one generator-discriminator pair instead of all generators and discriminators. Particularly, we randomly pick a generator and a discriminator and compute the loss with a random batch of training data. With the spirit of stochastic optimization, we still optimize the objectives in (13) and (14). The training algorithm is shown in Algorithm 1. Note that the training does not take as much as (IJ) times of the training time of a base model – it is much faster than that. This is because a generator is updated once in I iterations on average, and a discriminator is similar. In actual implementations, small I and J values (e.g. $I = J = 3$) often provide significant performance improvement. The training of the model is shown in Algorithm 1.

Anomaly scores The anomaly score of the ensemble $\mathcal{A}(\mathbf{x}')$ for a new instance $\mathbf{x}'$ is the average of anomaly scores from multiple generators and discriminators.

$$\mathcal{A}(\mathbf{x}') = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} A(\mathbf{x}';\phi_i,\psi_i,\gamma_j). \quad (15)$$

The average of anomaly scores helps to smooth out spurious scores if a model is not well trained at a specific test instance.

Analysis

Previous analysis of the GAN equilibrium (Goodfellow et al. 2014; Mescheder, Geiger, and Nowozin 2018) indicates that

Algorithm 1 GAN ensemble for anomaly detection

```
Input: Training set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$
Output: Trained generators $\{(G_e(\cdot;\phi_i), G_d(\cdot;\psi_i)\}_{i=1}^I$ and discriminators $\{D(\cdot;\gamma_j)\}_{j=1}^J$
Initialize parameters for$(\phi_i, \psi_i)_{i=1}^I$ an $(\gamma_j)_{j=1}^J$
$t \leftarrow 0$
while the objective not converge and $t <$ max_iter do
    Sample $i$ from $\{1, \ldots, \mathrm{I}\}$ and $j$ from $\{1, \ldots, J\}$
    Sample a minibatch $\mathbf{X}^t$ from $\mathbf{X}$
    Compute the adversearial loss $L_a^{(ij)}$
    Update $D(\cdot;\gamma_j)$: $\gamma_j \leftarrow \gamma_j + \nabla_{\gamma_j} L_a^{(ij)}$
    $\mathcal{L}^{(ij)} = \alpha_1 L_a^{(ij)} + \alpha_2 L_r^{(i)} + \alpha_3 L_d^{(ij)} + \alpha_4 L_e^{(i)}$
    Update $G_e(\cdot;\phi_i)$: $\phi_i \leftarrow \phi_i - \nabla_{\phi_i} \mathcal{L}^{(ij)}$
    Update $G_d(\cdot;\psi_i)$: $\psi_i \leftarrow \psi_i - \nabla_{\psi_i} \mathcal{L}^{(ij)}$
    $t \leftarrow t + 1$
end while
```

the discriminator does no better than random guesses at the equilibrium in the ideal case. In practice, however, the discriminator at convergence actually gives large values to training samples and small values to samples that are far away. The gap might be attributed to the assumption of real data distribution in previous analysis. Here we remove the assumption on real data distributions and analyze the discriminator in the context of anomaly detection.

We focus on the Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017; Zhou et al. 2019), and the analysis directly applies to f-AnoGAN. Suppose the critic function $D(\cdot)$ (also called as discriminator in the discussion above) is only restricted by the 1-Lipschitz condition. Given a generator $G_d(\cdot)$, the maximization objective of $D(\cdot)$ is (Arjovsky, Chintala, and Bottou 2017)

$$\max_{\|D(\cdot)\|_{L \le 1}} \quad \mathbb{E}_{\mathbf{x} \sim \mathrm{unif}(\mathbf{X})}\left[D(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x} \sim G_d(\cdot)}\left[D(\mathbf{x})\right] \quad (16)$$

Here we consider training samples only because the training use these samples instead of an imaginary real distribution.

To define the Lipschitz continuity, we assume there is a norm $\|\cdot\|$ defined for the sample space $\mathbb{R}^d$. $D(\cdot)$ is restricted to be 1-Lipschitz with respect to this norm. The following theorem shows that the function $D(\cdot)$ is decided by it function values $\{D(\mathbf{x}_i) : i = 1, \ldots, N\}$ on real samples $\mathbf{X}$ and the support $\mathcal{S}$ of the generator $G_d(\cdot)$.

Theorem 1. *Assume the generator $G_d(\cdot)$ defines a continuous distribution that has positive density on its support $\mathcal{S}$. Suppose the optimizer $D^*(\cdot)$ of* (16) *takes values $(D^*(\mathbf{x}_i) : i = 1, \ldots, N)$ on training samples $\mathbf{X}$, then it must have the following form,*

$$D^*(\mathbf{x}) = \max_i D^*(\mathbf{x}_i) - \|\mathbf{x} - \mathbf{x}_i\|, \qquad \forall \mathbf{x} \in \mathcal{S}. \quad (17)$$

Furthermore, the value of $D^*(\mathbf{x})$ for $\mathbf{x} \notin (\mathcal{S} \cup \mathbf{X})$ does not affect the objective (16).

Proof. We first show that $D^*(\mathbf{x})$ achieves the smallest value on any sample in $\mathcal{S} - \mathbf{X}$ under the 1-Lipschitz constraint. Let $\mathbf{x}' \in \mathcal{S} - \mathbf{X}$ and $i' = \arg\max_i D^*(\mathbf{x}_i) - \|\mathbf{x}' - \mathbf{x}_i\|$, then $D^*(\mathbf{x}') = D^*(\mathbf{x}_{i'}) - \|\mathbf{x}' - \mathbf{x}_{i'}\|$. $D^*(\mathbf{x}')$ cannot take any value smaller than that; otherwise it violates the 1-Lipschitz constraint.

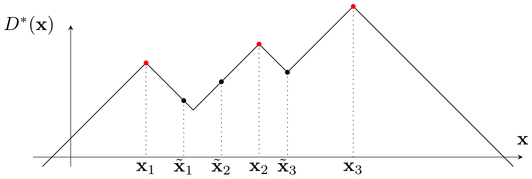

Figure 2. An illustration of $D^*(\mathbf{x})$ in Theorem 1. $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ are training samples, and $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \tilde{\mathbf{x}}_3\}$ are synthetic samples

Then we show that $D^*(\cdot)$ itself is 1-Lipschitz. Suppose we have another sample $\mathbf{x}'' \in \mathcal{S} - \mathbf{X}$ and $i'' = \arg\max_i D^*(\mathbf{x}_i) - \|\mathbf{x}'' - \mathbf{x}_i\|$, then $D^*(\mathbf{x}'') = D^*(\mathbf{x}_{i''}) - \|\mathbf{x}'' - \mathbf{x}_{i''}\| \ge D^*(\mathbf{x}_{i'}) - \|\mathbf{x}_{i'} - \mathbf{x}''\|$. We have $D^*(\mathbf{x}') - D^*(\mathbf{x}'') \le -\|\mathbf{x}_{i'} - \mathbf{x}'\| + \|\mathbf{x}_{i'} - \mathbf{x}''\| \le \|\mathbf{x}' - \mathbf{x}''\|$ by the triangle inequality. We also have $D^*(\mathbf{x}') - D^*(\mathbf{x}'') \ge -\|\mathbf{x}' - \mathbf{x}''\|$ by switching $\mathbf{x}'$ and $\mathbf{x}''$ and using the same argument. Therefore, $D^*(\mathbf{x})$ is 1-Lipschitz.

For $\mathbf{x} \notin (\mathcal{S} \cup \mathbf{X})$, $D^*(\mathbf{x})$ is not considered in the calculation in the objective, so it does not affect the objective.

An illustration of $D^*(\mathbf{x})$ is shown in Figure 2. □

From the theorem, we form two views about why a discriminator helps to identify anomalies.

Anomalies at the discriminator The theorem indicates that the discriminator should give small values to samples that are generated by the generator, particularly those that are far from training samples. On purpose the generator should not generate only training samples. The generated samples that are very different from training samples become "positive samples" (e.g. shirts with asymmetrical sleeves) of anomalies for the discriminator, then the discriminator is trained to give small values to these samples and also anomalies. A similar principle should exist in other types of GANs. This explains why we can include discriminator outputs in anomaly scores (Zenati et al. 2018).

The hidden vector in the last layer of the discriminator is often very different for an anomaly and its reconstruction. The difference is also used to compute an anomaly score in (6) in f-AnoGAN, EGBAD, and Skip-GANomaly. The reconstruction of an anomaly is often like a normal example because the unique property of an anomaly often cannot go through the encoder. In Figure 1, the sample "0" is the abnormal class, and its reconstructions are more like normal samples. Then the discriminator gives very different values to an anomaly and its reconstruction. These values are just a linear transformations of hidden vectors in the last layer, so the hidden vector of an anomaly is different from that of its reconstruction. A normal sample and its reconstruction are often very similar, and their hidden vectors are also similar, so a normal sample has small discriminative loss.

	0	1	2	3	4	5	6	7	8	9	average
f-AnoGAN	0.895	0.699	0.863	0.785	0.781	0.761	0.896	0.702	**0.889**	0.630	0.799
EGBAD	0.782	0.298	0.663	0.511	0.458	0.429	0.571	0.385	0.521	0.342	0.496
GANomaly	0.881	0.661	**0.951**	0.796	0.809	0.868	0.859	0.671	0.653	0.533	0.673
f-AnoGAN$^{\mathbf{en}}$	**0.961**	**0.943**	0.914	**0.913**	0.817	0.767	**0.957**	0.782	0.830	0.681	**0.857**
EGBAD$^{\mathbf{en}}$	0.804	0.202	0.671	0.577	0.438	0.480	0.595	0.425	0.595	0.458	0.525
GANomaly$^{\mathbf{en}}$	0.901	0.598	0.931	0.883	**0.838**	**0.875**	0.892	**0.801**	0.847	**0.733**	0.830

Table 1. AUROC results on the MNIST dataset. Ensemble methods generally outperform base models.

	bird	car	cat	deer	dog	frog	horse	plane	ship	truck	average
f-AnoGAN	0.427	0.778	0.541	0.442	0.601	0.582	0.628	0.688	0.637	0.781	0.611
EGBAD	0.383	0.514	0.448	0.374	0.481	0.353	0.526	0.577	0.413	0.555	0.462
GANomaly	0.510	0.631	0.587	0.593	0.628	0.683	0.605	0.633	0.710	0.617	0.620
Skip-GANomaly	0.448	**0.953**	0.607	0.602	0.615	**0.931**	0.788	0.797	0.659	0.907	0.731
f-AnoGAN$^{\mathbf{en}}$	0.531	0.804	0.581	0.584	0.616	0.642	0.653	0.704	0.830	0.907	0.685
EGBAD$^{\mathbf{en}}$	0.573	0.620	0.451	0.563	0.388	0.554	0.429	0.522	0.612	0.668	0.538
GANomaly$^{\mathbf{en}}$	0.533	0.669	0.599	0.719	0.667	0.856	0.614	**0.948**	0.854	0.682	0.714
Skip-GANomaly$^{\mathbf{en}}$	**0.998**	0.917	**0.691**	**0.766**	**0.937**	0.764	**0.992**	0.703	**0.991**	**0.917**	**0.868**

Table 2. AUROC results on the CIFAR-10 dataset. Ensemble methods generally outperform base models.

Reconstruction guided by the discriminator We have the form of the ideal discriminator $D^*(\cdot)$ given the norm $\|\cdot\|$, but it is hard to find such a norm used by the discriminator. We can get some information about the norm by considering two similar samples, e.g. $\mathbf{x}_{i'}$ and $\mathbf{x}'$ when $\mathbf{x}'$ is a good reconstruction of a normal sample $\mathbf{x}_{i'}$. The training sample $\mathbf{x}_{i'}$ is likely to be nearest to $\mathbf{x}'$ among all training samples, and then $\|\mathbf{x}' - \mathbf{x}_{i'}\| = (D^*(\mathbf{x}_{i'}) - D^*(\mathbf{x}'))$. The minimization of (16) with respect to the generator is essentially the minimization of $\|\mathbf{x}' - \mathbf{x}_{i'}\|$, which corresponds to the minimization of $\|\mathbf{x}' - \mathbf{x}_{i'}\|_2^2$ in an encoder-decoder. Therefore, a discriminator implicitly defines a norm over samples. The adversarial loss, which is like a reconstruction loss using this norm, guides the training of the encoder-decoder.

The benefit of GAN ensembles Previous research (Arora et al. 2017) shows that multiple generators helps to capture data modes and provide synthetic samples with varieties. Multiple generators are also likely to have a larger joint support $\mathcal{S}$ than a single one. Therefore, the generator ensemble trains a discriminator better than a single model. Our experiments later show that the training of discriminators is very important to anomaly detection.

Experiments

In this section, we evaluate the proposed method in several anomaly detection tasks. We also design experiments to analyze the method.

We evaluate our method against baseline methods on four datasets. KDD99 (Dua and Graff 2019) is a dataset for anomaly detection. OCT (Kermany et al. 2018) has three classes with small number of samples, and these three classes are treated as abnormal classes. MNIST (LeCun and Cortes 1998) and CIFAR-10 (Krizhevsky, Hinton et al. 2009) are datasets for multiclass classification. We leave a class as the abnormal class and use other classes as normal data. Among these four datasets, MNIST and CIFAR-10 contain low-resolution images, OCT has high-resolution images from clinical applications, and KDD99 is a non-image dataset.

We consider four base models in our proposed framework: f-AnoGAN (Schlegl et al. 2019), EGBAD (Zenati et al. 2018), GANomaly (Akcay, Atapour-Abarghouei, and Breckon 2018) and Skip-GANomaly (Akcay, Atapour-Abarghouei, and Breckon 2019). Then we have four types of ensembles, which are compared against their respective base models. All the experiments are conducted using three $I = 3$ generators and three $J = 3$ discriminators. I and J are chosen for a good tradeoff between running time and performance. The implementation is available at https://github.com/tufts-ml/GAN-Ensemble-for-Anomaly-Detection.

Comparisons of Detection Performances

MNIST dataset MNIST has 10 classes of hand-written digits from 0 to 9. From it we construct 10 anomaly detection tasks, treating each class as abnormal and using other classes as normal. We compare three ensemble methods and three base models using the AUROC score as the evaluation measure. Table 1 reports performances of different methods. Ensemble methods are indicated by the superscript **en**. In the comparison of an ensemble and its corresponding base model, the better result is underscored. The best result across all models is emboldened. The performance values of EGBAD and GANomaly are taken from their original papers.

In general, ensemble methods have superior performance over their corresponding base models on most of the 10 classes. For classes 1, 3, 7, and 9, the best ensemble method improves more than 0.1 in AUROC over baseline methods. On average, the ensembles of f-AnoGAN and GANomaly improves the performances of their respective base models by 0.07 and 0.23. We do not include Skip-GANomaly because it performs poorly on the MNIST dataset.

Method	CNV	DME	DRUSEN	overall
f-AnoGAN	0.863	0.754	0.610	0.666
EGBAD	0.793	0.782	0.601	0.610
GANomaly	0.852	0.796	0.543	0.637
Skip-GANomaly	0.915	**0.929**	0.609	0.766
f-AnoGAN$^{\mathbf{en}}$	0.949	0.866	0.705	0.800
EGBAD$^{\mathbf{en}}$	0.889	0.821	0.611	0.705
GANomaly$^{\mathbf{en}}$	0.911	0.845	0.626	0.741
Skip-GANomaly$^{\mathbf{en}}$	**0.985**	0.889	**0.952**	**0.869**

Table 3. AUROC results on the OCT dataset. Ensemble methods generally outperform base models.

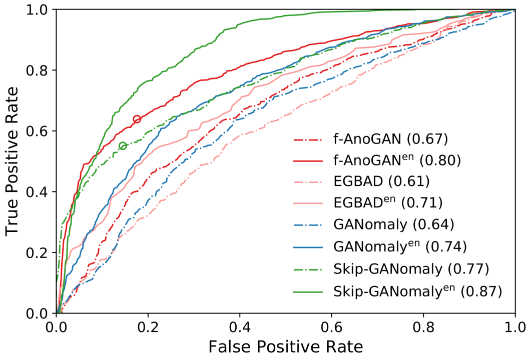

Figure 3. ROC curves of different models on the OCT dataset with "overall" being the anomlous class.

CIFAR-10 dataset CIFAR-10 has 10 classes of images of size 32 × 32. Every time one class is treated as anomalous, and other classes are used as normal data. All four base models and their ensembles are compared and evaluated by AUROC scores. The performance values of EGBAD, GANomaly, and Skip-GANomaly are taken from the original papers.

We report AUROC scores of all models in Table 2. Ensembles constructed from f-AnoGAN and GANomaly outperform their corresponding base models on all 10 classes. The improvements from our ensemble method for EGBAD and Skip-GANomaly are also apparent in the results. The ensemble of Skip-GANomaly models, which performs the best on 7 out of 10 classes, is generally the best model for this dataset.

OCT dataset OCT contains high-resolution clinical images. It has three small classes (CNV, DME, DRUSEN) and a large class of images. We use the large class as normal data; then, we use three small classes separately as anomalous classes and use the three classes together as a single anomalous class (denoted as "overall").

Table 3 shows AUROC performances of different methods over four types of anomaly categories (CNV, DME, DRUSEN, and overall). The ensemble method's benefit is more obvious in this dataset: except Skip-GANomaly on the DME class, all ensemble models show significant performance improvements on all anomaly classes. Particularly on the DRUSEN class, the ensemble of Skip-GANomaly has a much better performance than all base models. Figure 3 shows ROC Curves for all single models as well as ensemble models. This result indicates that an ensemble model can make reliable decisions in medical applications.

Method	Precision	Recall	F1
OC-SVM	0.746	0.8526	0.795
DSEBM-r	0.852	0.647	0.733
DSEBM-e	0.862	0.645	0.740
DAGMM	0.929	0.944	0.937
EGBAD	0.920	0.958	0.939
f-AnoGAN	0.935	0.986	0.960
EGBAD$^{\mathbf{en}}$	**0.972**	0.960	0.966
f-AnoGAN$^{\mathbf{en}}$	0.967	**0.990**	**0.979**

Table 4. Performance comparison on the KDD99 dataset. Ensembles outperform base models and other baselines.

KDD99 dataset KDD99 (Dua and Graff 2019) is a benchmark dataset for anomaly detection. Each sample in it is a 122-dimensional vector. We compare ensembles of EGBAD and f-AnoGAN against two base models and four other baselines (OC-SVM (Schölkopf et al. 2001) , DSEBM (Zhai et al. 2016), DAGMM (Zong et al. 2018)). We do not include GANomaly and Skip-GANomaly because they are designed for image data. Following prior work, we use precision, recall, and F1-score to evaluate methods in comparison. The performance values of EGBAD, f-AnoGAN, and other baselines are from their original papers, except that the performance values of OC-SVM are from (Zhai et al. 2016).

The results in Table 4 show that the two ensemble models constructed from f-AnoGAN and EGBAD outperform base models by all evaluation measures. The ensemble of f-AnoGAN outperforms all baseline methods.

Framework Analysis

Analysis of ensemble sizes In this experiment, we check how ensemble sizes affect performances. We vary the ensemble sizes in $\{1, 3, 5, 7\}$ and keep the same number of generators and discriminators. We test these configurations on the CIFAR-10 dataset with *ship* as the anomalous class. Figure 5 shows that there is a significant improvement from size 1 (a single model) to size 3. The average increase of AUROC is 35.9% across all models. However, the performance gain from size 3 to size 7 is often marginal. In our experiment with $I = J = 3$, the ensemble takes about 3 times of a base model's running time.. Note that every generator or discriminator in an ensemble is updated once in every 3 iterations. As a rule of thumb, ensembles of size 3 works the best for all tasks in our experiments considering both performance and running time.

Analysis of encoding vectors and hidden vectors We analyze encoding vectors and hidden vectors of normal and abnormal samples to understand why an ensemble significantly improves the performance of a base model. Note that when a sample is present to a GAN, we get an encoding vector from the encoder and a hidden vector from the discriminator's last hidden layer. For an ensemble method, we take the average

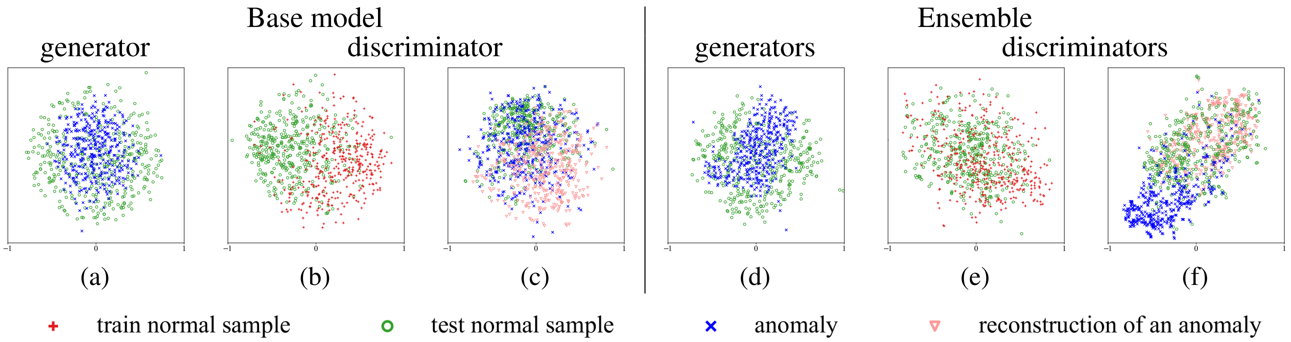

Figure 4. Latent analysis for discriminator and generator

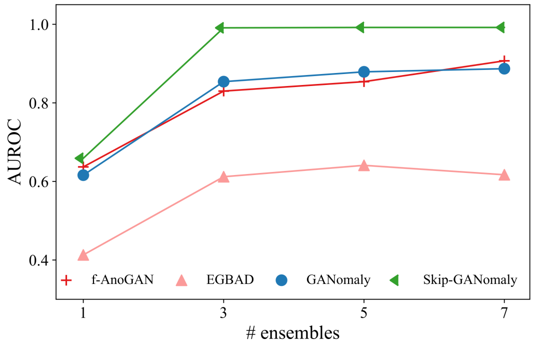

Figure 5. Difference detection performances with different ensemble sizes: $I = J \in \{1,3,5,7\}$.

of encoding vectors and hidden vectors from multiple models.

We first check encoding vectors. From (a) and (d) in Figure 4, we see that encoding vectors of normal samples and abnormal samples are mixed together. This is because the encoder does not encode the unique properties of abnormal samples because the encoder is trained to compress common patterns. Then it means that the reconstruction of abnormal samples will be like normal samples.

We then check hidden vectors of normal samples from the discriminator. Figure 4 (b) shows that training and testing samples get different representations in the discriminator in a base model. It means that the discriminator of a single model might overfit the training data. The ensemble seems to train better discriminators by checking (e), which shows that representations of training and test samples are very similar.

We also check hidden vectors of abnormal samples and their reconstructions. From Figure 4 (c), we see that the discriminator of a single model cannot distinguish test samples and abnormal samples, then the detection performance is not optimal. Figure 4 (f) shows that hidden vectors of abnormal samples are far from test normal samples. At the same time, the reconstructions of abnormal samples are similar to normal samples, which is consistent with our observation in (d). Then the difference between an anomaly and its reconstruction is likely to be large in (6), then the anomaly will have a large score.

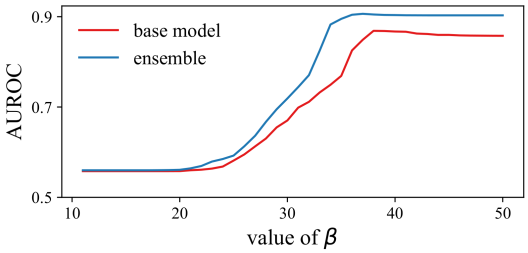

Figure 6. Different detection performances with different relative weight β of discriminative loss in (10).

In summary, the experimental analysis verifies our theoretical analysis in . It shows how the ensemble improves the training of discriminators and also the computation of anomaly scores.

We further verify that the ensemble mainly improves the second term of the calculating of anomaly scores in (10). We vary the value of the relative weight β of the reconstruction loss and the discriminative loss and check detection performances. Figure 6 shows the results. As β increases, the discriminative loss contributes larger fractions to anomaly scores, and the detection performance improves. It means the discriminative loss is important for anomaly detection. Compared with a single model, the ensemble benefits more from large β values. It indicates that an ensemble mainly improves the discriminative loss over a base model, as shown in our analysis above.

Conclusion

This work introduces ensemble learning to GAN-based anomaly models for anomaly detection. Extensive experiments show that the ensemble method achieves superior results across various applications over single models. We conduct theoretical analysis to rigorously explain why the discriminator of a GAN is very useful for anomaly detection and why an ensemble further improves the training of discriminators. Our experimental analysis of encoding vectors and hidden vectors verifies our theoretical analysis and further shows the benefit of ensemble learning.

Acknowledgments

Thank all reviewers for constructive comments. Li-Ping Liu were supported by DARPA under the SAIL-ON program (Contract number: W911NF-20-2-0006). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Abdallah, A.; Maarof, M. A.; and Zainal, A. 2016. Fraud detection system: A survey. *Journal of Network and Computer Applications* 68: 90–113.

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, 622–637. Springer.

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2019. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *arXiv preprint arXiv:1701.07875* .

Arora, S.; Ge, R.; Liang, Y.; Ma, T.; and Zhang, Y. 2017. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573* .

Atefeh, F.; and Khreich, W. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence* 31(1): 132–164.

Chalapathy, R.; and Chawla, S. 2019. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407* .

Dua, D.; and Graff, C. 2019. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml. Last accessed in July 2020.

Durugkar, I.; Gemp, I.; and Mahadevan, S. 2016. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673* .

Garcia-Teodoro, P.; Diaz-Verdejo, J.; Maciá-Fernández, G.; and Vázquez, E. 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security* 28(1-2): 18–28.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gupta, M.; Gao, J.; Aggarwal, C. C.; and Han, J. 2013. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and data Engineering* 26(9): 2250–2267.

Hoang, Q.; Nguyen, T. D.; Le, T.; and Phung, D. 2018. MGAN: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations*.

Hodge, V.; and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22(2): 85–126.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172(5): 1122–1131.

Kou, Y.; Lu, C.-T.; Sirwongwattana, S.; and Huang, Y.-P. 2004. Survey of fraud detection techniques. In *IEEE International Conference on Networking, Sensing and Control, 2004*, volume 2, 749–754. IEEE.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical report, Citeseer* .

LeCun, Y.; and Cortes, C. 1998. MNIST handwritten digit database URL http://yann.lecun.com/exdb/mnist/. Last accessed in July 2020.

Liu, L.-P.; and Fern, X. Z. 2012. Constructing training sets for outlier detection. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 919–929. SIAM.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which training methods for GANs do actually converge? In *International conference on machine learning*, 3481–3490. PMLR.

Neyshabur, B.; Bhojanapalli, S.; and Chakrabarti, A. 2017. Stabilizing GAN training with multiple random projections. *arXiv preprint arXiv:1705.07831* .

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Langs, G.; and Schmidt-Erfurth, U. 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis* 54: 30–44.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, 146–157. Springer.

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13(7): 1443–1471.

Zenati, H.; Foo, C. S.; Lecouat, B.; Manek, G.; and Chandrasekhar, V. R. 2018. Efficient GAN-based anomaly detection. *arXiv preprint arXiv:1802.06222* .

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717* .

Zhou, Z.; Liang, J.; Song, Y.; Yu, L.; Wang, H.; Zhang, W.; Yu, Y.; and Zhang, Z. 2019. Lipschitz Generative Adversarial Nets. In *International Conference on Machine Learning*, 7584–7593.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*.