ANOMALY DETECTION BY DEEP DIRECT DENSITY RATIO ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the ratio of two probability densities without estimating each density separately has been shown to provide useful solutions to various machine learning tasks such as domain adaptation, anomaly detection, feature extraction, and conditional density estimation. However, density ratio estimation in the context of deep learning has not been extensively explored yet. In this paper, we apply a Bregman-divergence minimization method for density ratio estimation to deep neural networks and investigate its properties and practical performance in image anomaly detection. Our numerical experiments on the CIFAR-10, CIFAR-100 and Fashion-MNIST datasets demonstrate that deep direct density ratio estimation greatly improves the anomaly detection ability and reduces the computation time over state-of-the-art methods.

1 INTRODUCTION

Anomaly detection (also known as outlier detection) has received a lot of attention in diverse research areas such as monitoring (Lavin & Ahmad, 2015), credit card fraud detection (Phua et al., 2010), and medical diagnosis (Schlegl et al., 2017). The aim of anomaly detection is to identify outliers in a given dataset. A standard anomaly detection problem falls into the category of unsupervised learning, due to lack of labeled anomaly data. While (semi-)supervised anomaly detection methods perform better than unsupervised methods (Gao et al., 2006), they require anomalous data for training, which are not always available in practice. Furthermore, the anomalous properties may be diverse, and thus such (semi-)supervised methods are not necessarily useful in detecting an unknown type of anomaly.

Traditional approaches for unsupervised anomaly detection such the as one-class support vector machine (OC-SVM) (Schölkopf et al., 2001) and support vector data description (SVDD) (Tax & Duin, 2004) have been widely used, which relies on the the assumption that a sample located in a low-density region is regarded as an outlier. However, these approaches often face difficulties when they are applied to high-dimensional data such as images, due to the curse of dimensionality. Furthermore, these approaches depend heavily on the choice of tuning parameters (e.g., the Gaussian kernel width) and there seems to be no universal method to appropriately determine the values of such tuning parameters.

Recently, *convolutional neural networks* (CNNs) have significantly improved their performance in various computer vision tasks, e.g., image classification and object detection (Krizhevsky et al., 2012; He et al., 2016; Redmon et al., 2016). With the advent of deep learning, various methods have been developed for anomaly detection in the context images (Chalapathy & Chawla, 2019). For example, the generative model methods, which are based either on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) or Autoencoders (AEs) (Hinton & Salakhutdinov, 2006), have been applied in anomaly detection. While there are several other approaches, these are mostly based on the idea of obtaining a good representation, e.g., intermediate representations in AE, i.e., latent spaces in GANs, of normal data with a neural network. Then the obtained representation is used to define anomaly scores via reconstruction errors. However, since the representation learning and the anomaly score calculation are performed separately, methods based on deep generative models are suboptimal. To avoid such two-step optimization, different methods have been proposed based on SVDD (Ruff et al., 2018) , but they cannot utilize the superior representation power of neural networks. Also all these unsupervised approaches suffer the problem of hyperparameter optimization due to lack of supervision.

To overcome the weakness of unsupervised anomaly detection, weakly-supervised anomaly detection has been explored, where normal samples and unlabeled samples are utilized. More specifically, an approach based on density ratio estimation (Sugiyama et al., 2012a) has been investigated thoroughly. In this approach, the ratio of probability densities of normal and unlabeled samples are directly estimated without estimating each density separately, and it is used as an outlier score. A notable advantage of this direct density ratio estimation approach is that hyperparameter tuning can be performed objectively through cross-validation. So far, various direct density ratio estimation methods have been developed, e.g., unconstrained least-squares importance fitting (uLSIF) (Kanamori et al., 2009) and the Kullback-Leibler importance estimation procedure (KLIEP) (Sugiyama et al., 2008). In the context of anomaly detection, kernel-base KLIEP was demonstrated to be superior in accuracy and stability compared to OC-SVM, kernel mean matching (KMM) (Huang et al., 2006) and uLSIF (Hido et al., 2011).

As explained above, direct density ratio estimation is a promising approach to anomaly detection. However, direct density ratio estimation in the context of deep learning has not been extensively explored yet. In this paper, we apply the Bergman-divergence minimization method for density ratio estimation to deep neural networks and investigate its properties and practical performance in image anomaly detection. An interesting finding is that batch normalization (BatchNorm), which is an effective method in training deep neural networks (Ioffe & Szegedy, 2015; Bjorck et al., 2018), does not work well in our context. We explain the reason for this phenomenon and propose not to use BatchNorm in our proposed method. We perform numerical experiments on the CIFAR-10, CIFAR-100 and Fashion-MNIST datasets and demonstrate that deep direct density ratio estimation significantly improves the anomaly detection ability and reduces the computation time over state-of-the-art methods.

2 Related work

An extensive review of classical anomaly detection methods can be found in Chandola et al. (2009). In this section we focus on anomaly detection in the context of images and deep learning.

2.1 Density ratio estimation

Previous work (Nam & Sugiyama, 2015) has already applied deep density ratio estimation to anomaly detection. However, this study only reported that CNN-based uLSIF is superior to kernel-based uLSIF and kernel-based KLIEP for image datasets. However, from the kernel-based density ratio estimation studies (Sugiyama et al., 2012a), it is known that KLIEP is more sensitive to outliers than uLSIF and is more effective in detecting outlier samples. In addition, LeNet-5 (LeCun et al., 1998), which was used in Nam & Sugiyama (2015), is a network architecture originally designed for hand-written character recognition. So it has poor expressive ability compared to more recently proposed neural network architecture for complex image datasets. To the best of our knowledge, there are no studies investigating whether deep density ratio estimation under the KLIEP criterion with modern deep learning techniques is effective compared to recent deep anomaly detection methods. This is what we will investigate in this paper.

2.2 Deep generative model and Deep SVDD

A typical method of deep anomaly detection in the context of image data is based on deep generative models such as AEs or GANs. The main idea is based on the fact that it is difficult to generate outlier samples from a latent space obtained by learning with only normal samples. In the context of anomaly detection, Schlegl et al. (2017) first introduced an approach based on GANs, which is called AnoGAN. AnoGAN uses a convex combination of the ℓ_2 norm and a discrimination loss between an input image and generated image as an anomaly score. Similary to AnoGAN, Deecke et al. (2018) proposed ADGAN that improved the performance slightly. Since ADGAN never uses the discriminator loss to calculate an anomaly score, the discriminator can be discard after training the GAN.

As a different method, Ruff et al. (2018) recently proposed an approach to detect outliers using a deep neural network inspired by SVDD, which is a widely used one-class classification method for anomaly detection. The main idea of the method, named Deep SVDD, is using a deep neural network to minimize the volume of a hyper-sphere that encloses the network representations of normal samples. Anomaly scores in the Deep SVDD approach is the distance of a data point from the center of the hyper-sphere.

2.3 Geometric Transformations

The geometric transformations (GTs) method (Golan & El-Yaniv, 2018) first creates a self-labeled dataset by performing 72 distinct geometric transformations consisting of horizontal flips, translations, and rotations on normal data. Then a multi-class classifier is trained over the self-labeled dataset, where the labels are the types of transformations. An anomaly score is defined based on the Dirichlet distribution obtained by maximum likelihood estimation using the softmax output from the classification network for the labels. GTs greatly exceed the accuracy of Deep SVDD and ADGAN on benchmark datasets. Thus, in this paper, we will compare it with our method.

3 Anomaly Detection via Density Ratio Estimation

Here we briefly review the framework of density ratio estimation by density ratio fitting under the Bregman divergence for anomaly detection (Sugiyama et al., 2012b)[1].

3.1 Formulation

Let $\mathcal{X} \subset \mathbb{R}^d$ be the data domain for positive integer d. Suppose that we are given independent and identically distributed (i.i.d.) training samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ from a training distribution with density $p_{\mathrm{tr}}^*(\boldsymbol{x})$ on $\mathcal{X}$ and i.i.d. test samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ from a test distribution with density $p_{\mathrm{te}}^*(\boldsymbol{x})$ on $\mathcal{X}$ [2]. The training samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ are all inliers, while the test samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ do not only contain inliers but can also contain some outliers if any. The goal of anomaly detection based on density ratio estimation is to estimate the *density ratio*,

$$r^*(\boldsymbol{x}) \coloneqq \frac{p_{\mathrm{tr}}^*(\boldsymbol{x})}{p_{\mathrm{te}}^*(\boldsymbol{x})}, \tag{1}$$

from $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$. The density ratio is close to one when $\boldsymbol{x}$ is an inlier and it is close to zero when $\boldsymbol{x}$ is an outlier. Thus, the density ratio would be a suitable anomaly score.

A naive approach to estimate the density ratio Eq.(1) is to first estimate the numerator and denominator densities separately from their associated samples and then take their ratio. However, such a two-step approach is not reliable because the first step of density estimation is performed without regard to the second step of taking the ratio. Below, we review a direct density ratio estimation method that does not involved density estimation.

3.2 Density Ratio Estimation under Bregman Divergence

The basic idea of direct density ratio estimation is to fit a density ratio model $r(\boldsymbol{x})$ to the true density ratio function $r^*(\boldsymbol{x})$ under some divergence. Here we employ the *Bregman* (BR) divergence (Bregman, 1967) for measuring the discrepancy between the true density ratio function and the density ratio model. This framework includes various existing approaches of density ratio estimation as special cases.

The BR divergence from t^* to t is defined as follows:

$$\mathrm{BR}'_f(t^*||t) \coloneqq f(t^*) - f(t) - \partial f(t)(t^* - t), \tag{2}$$

[1] See Sugiyama et al. (2012a) for a comprehensive review on the application of density ratio estimation to tasks other than anomaly detection.

[2] We assume that $p_{\mathrm{te}}^*(\boldsymbol{x})$ is strictly positive for all $x \in \mathcal{X}$.

where $f(t)$ is a strictly convex function and differentiable. Minimizing the BR divergence between the true density ratio $r^*(\boldsymbol{x})$ and a model of the density ratio $r(\boldsymbol{x})$, weighted by p^*_{te}, gives

$$\begin{aligned}\mathrm{BR}'_f(r^*\|r) &\coloneqq \int p^*_{\mathrm{te}}(\boldsymbol{x})\left[f(r^*(\boldsymbol{x})) - f(r(\boldsymbol{x})) - \partial f(r(\boldsymbol{x}))\left(r^*(\boldsymbol{x}) - r(\boldsymbol{x})\right)\right]\mathrm{d}\boldsymbol{x} \\ &= C + \mathrm{BR}_f(r), \qquad (3)\end{aligned}$$

where $C \coloneqq \int p^*_{\mathrm{te}}(\boldsymbol{x}) f(r^*(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$ is a constant independent of the density ratio model r and

$$\begin{aligned}\mathrm{BR}_f(r) &\coloneqq \int p^*_{\mathrm{te}}(\boldsymbol{x})\partial f(r(\boldsymbol{x}))r(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{te}}(\boldsymbol{x}) f(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ &\quad - \int p^*_{\mathrm{tr}}(\boldsymbol{x})\partial f(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \qquad (4)\end{aligned}$$

Then an empirical approximation $\widehat{\mathrm{BR}}_f(r)$ of $\mathrm{BR}_f(r)$ is given by

$$\begin{aligned}\widehat{\mathrm{BR}}_f(r) &\coloneqq \frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} \partial f(r(\boldsymbol{x}^{\mathrm{te}}_j))r(\boldsymbol{x}^{\mathrm{te}}_j) - \frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} f(r(\boldsymbol{x}^{\mathrm{te}}_j)) \\ &\quad - \frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} \partial f(r(\boldsymbol{x}^{\mathrm{tr}}_i)). \qquad (5)\end{aligned}$$

This immediately gives the following optimization criterion:

$$\min_r \widehat{\mathrm{BR}}_f(r). \qquad (6)$$

As a particular BR divergence Eq.(3), *Basu's Power* divergence (BA divergence) (Basu et al., 1998) can be induced by the function,

$$f(t) = \frac{t^{1+\alpha} - t}{\alpha}, \qquad (7)$$

where $\alpha > 0$. By substituting Eq.(7) into Eq.(3), an empirical approximation $\widehat{\mathrm{BA}}_f(r)$ of the BA divergence without an irrelevant constant term is given by

$$\widehat{\mathrm{BA}}_\alpha(r) \coloneqq \frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} r(\boldsymbol{x}^{\mathrm{te}}_j)^{\alpha+1} - \left(1 + \frac{1}{\alpha}\right)\frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}^{\mathrm{tr}}_i)^\alpha + \frac{1}{\alpha}. \qquad (8)$$

The BA divergence includes uLSIF ($\alpha = 1$) and KLIEP ($\alpha \to 0$) as special cases, and is more general.

To investigate robustness, let us take the derivative Eq.(8) with respect to parameters in the density-ratio model r and equate it to zero. Then we have the following estimation equation:

$$\frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} r(\boldsymbol{x}^{\mathrm{te}}_j)^\alpha \nabla r(\boldsymbol{x}^{\mathrm{te}}_j) - \frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}^{\mathrm{tr}}_i)^{\alpha-1}\nabla r(\boldsymbol{x}^{\mathrm{tr}}_i) = \mathbf{0}_b, \qquad (9)$$

where ∇ is the differential operator with respect to the parameters in the density-ratio model r, b denotes the number of parameters, and $\mathbf{0}_b$ denotes the b-dimensional vector with all zeros. In the case of $\alpha \to 0$ which corresponds to KLIEP, the estimation equation is given as follows:

$$\frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} \nabla r(\boldsymbol{x}^{\mathrm{te}}_j) - \frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}^{\mathrm{tr}}_i)^{-1}\nabla r(\boldsymbol{x}^{\mathrm{tr}}_i) = \mathbf{0}_b. \qquad (10)$$

Comparing this with Eq.(9), the BA method can be regarded as a weighted version of KLIEP according to $r(\boldsymbol{x}^{\mathrm{te}}_j)^\alpha$ and $r(\boldsymbol{x}^{\mathrm{tr}}_i)^\alpha$. As mentioned above, since outliers tend to take smaller ratio values, the BA method down-weights the effect of those samples. Thus, the KLIEP (i.e., $\alpha \to 0$) can provide a more sensitive anomaly score than uLSIF, which corresponds to $\alpha = 1$, in the above sense.

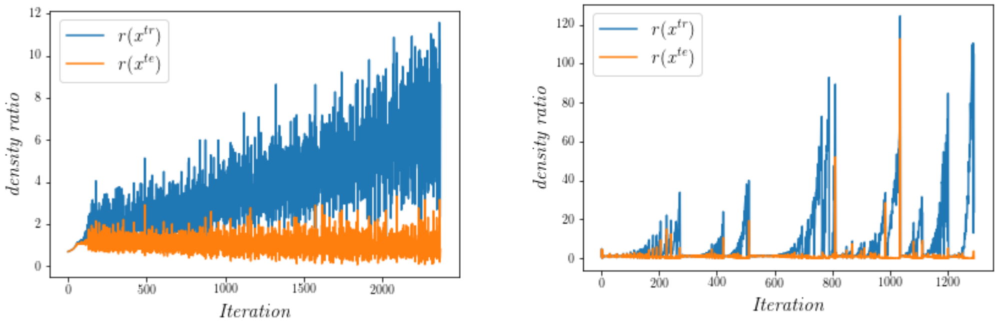

Figure 1: Evolution of the averages of density ratio values for $\frac{1}{n_{\mathrm{tr}}}\sum_{i=1} r(\boldsymbol{x}_i^{\mathrm{tr}})$ and $\frac{1}{n_{\mathrm{te}}}\sum_{j=1} r(\boldsymbol{x}_j^{\mathrm{te}})$ during training with the KLIEP criterion. The left graph contains results on without BatchNorm, while the right graph contains the results with BatchNorm.

4 DENSITY RATIO ESTIMATION AND BATCH NORMALIZATION

BatchNorm (Ioffe & Szegedy, 2015) has become a de facto standard for training deep neural networks with various architectures. Its effectiveness is still being investigated from various angles. Bjorck et al. (2018) argued that its effect may be smoothing the loss surface. This enables training with larger learning rates, which results in faster convergence and better generalization. Despite its empirical success on many tasks and recent theoretical progress, we argue that BatchNorm is incompatible with density ratio estimation using deep neural networks.

To explain the reason, let us consider using CNN to estimates the density ratio function under the KLIEP criterion:

$$\lim_{\alpha\to 0}\widehat{\mathrm{BA}}_\alpha(r) = \frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} r(\boldsymbol{x}_j^{\mathrm{te}}) - \frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} \ln(r(\boldsymbol{x}_i^{\mathrm{tr}})). \tag{11}$$

In the direct density ratio estimation problem, we use not only training data but also test data that include both inliers and outliers during density ratio fitting. Thus, outliers are heterogeneously distributed in a mini-batch.

Figure 1 plots the transition of the averages of density ratio values $\frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}_i^{\mathrm{tr}})$ and $\frac{1}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} r(\boldsymbol{x}_j^{\mathrm{te}})$ during training with and without BatchNorm under the KLIEP criterion in Eq.(11). Minimizing this objective function, the model is optimized to increase the second term in Eq.(11). However, when BatchNorm is used, density ratio estimation becomes unstable and $\frac{1}{n_{\mathrm{tr}}}\sum_{i=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}_i^{\mathrm{tr}})$ takes a large value suddenly compared to the case where BatchNorm is not used. In this figure, the density ratio obtained with BatchNorm diverges after the 1300th iteration, and consequently no outliers can be detected. Therefore, we decided not to use BatchNorm in this paper, which resulted in good empirical performance.

5 EXPERIMENTS

In this section, we use benchmark datasets to demonstrate the effectiveness of our method in anomaly detection. All experiments were performed using the PyTorch (Paszke et al., 2017) library. We used the AWS `p3.2xlarge` instance which has a single NVIDIA V100 GPU.

5.1 DATASET

Our method was evaluated on three publicly available benchmark image datasets: CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and Fashion-MNIST (Xiao et al., 2017). (i.) CIFAR-10 consists of various color images, which has 50000 32×32×3 training images in ten classes. (ii.) CIFAR-100 is similar to CIFAR-10, but with 100 classes containing 500 images per class. These classes are grouped into 20 superclasses each containing five classes. We used 20 superclasses in our experiments. (iii.) Fashion-MNIST which consists of 70000 $28\times28\times1$ grayscale images depicting fashion items in ten classes. To compatible with CIFAR-10 and CIFAR-100 classification architectures, we resize the images to 32×32.

5.2 Evaluation strategy

Our experimental settings are the same as the previous work (Golan & El-Yaniv, 2018). For all datasets, the inlier and outlier classes were defined as follows. One particular class was considered as the inlier class and all other classes were regarded as the outlier classes. For example, in the case of CIFAR-10, there are 5000 training data per class, so $n_{\mathrm{tr}} = 5000$. On the other hand, since there are 1000 test data for each class, the number of test samples is $n_{\mathrm{te}} = 10000$, which consists of 1000 inlier samples and 9000 outlier samples. The area under the receiver operating characteristic curve (AUROC) is used as a metric to evaluate whether an outlier class can be detected in the test data.

5.3 Experimental setup

We used the VGG11 (Simonyan & Zisserman, 2014) model as the backbone architecture without BatchNorm for density ratio estimation. Multiple convolutional layers in VGG11 are followed by three fully-connected (FC) layers. The first and second layers have 4096 channels, and the third layer has 1 channels. We used dropout regularization where the dropout rate was set to 0.5 in the convolution and FC layers. Taking into account the non-negativity of the density ratio, the output layer was set to the softplus function $\log(1 + e^{x})$. On the other hand, the Wide Residual Network (WRN) model (Zagoruyko & Komodakis, 2016) was only used as the backbone in Golan & El-Yaniv (2018). Thus, we conducted numerical experiments not only with WRN but also with VGG11 as backbone models in GTs for comparison. In WRN, we set the depth and width of the model to 10 and 4, respectively.

For all dataset, CNN-based KLIEP was trained by stochastic gradient descent (SGD) (Bottou et al., 2018) with batch size of 128. We set the learning rate to 0.02 and the number of epochs to 30. We used weight decay of 10^{-4}. Experiments were repeated over five trials. We converted the value of each pixel into the interval [0, 1] without other preprocesses and data augmentation. For fair comparison, GTs implemented by ourselves used the same settings such the batch size, optimizer and learning rate[3].

5.4 Results

The experimental results are shown in Table 1. The proposed CNN-based KLIEP clearly outperforms GTs on the benchmark datasets. In CIFAR-100, we omitted the name of the superclasses due to lack of space. The correspondence between indices and superclasses is listed in Appendix A. The inlier class consists of multiple classes in CIFAR-100. Experimental results show that KLIEP can stably achieve higher accuracy than the existing methods even in the multiclass setting.

Table 2 shows the computation time of each method for each dataset. Since the GTs method needs to perform geometric transformations to create the self-labeled dataset and training using that dataset, the computation time is long compared to our method. From the above results, it can be said that our proposed method is superior not only in terms of accuracy but also in terms of computational efficiency.

In addition, Appendix B shows that transfer learning is also effective in density ratio estimation. In this work, we used weight decay of 0.1 for fine-tuning the ImageNet-pretrained network from the PyTorch class `torchvision.models` [4]. We have also shown in Appendix B the result of changing the parameter α of the BA divergence in Eq.(8). Overall, KLIEP ($\alpha \to 0$) was found to be optimal for anomaly detection. This result is consistent with the theoretical analysis shown in Sec.3.2.

[3] For the details of the original implementation, refer to `https://github.com/izikgo/AnomalyDetectionTransformations`.

[4] `https://pytorch.org/docs/stable/torchvision/models.html`

Table 1: Average AUROC in % with standard deviation (over 5 trials with different seeds) per method. The best performing method in terms of the mean AUC is specified by bold face.

Dataset	inlier class	**GTs**	**GTs**	**KLIEP**
		(VGG11)	(WRN)	
CIFAR-10	plane	69.0 ± 1.0	76.3 ± 0.6	**93.6 ± 0.3**
	car	94.3 ± 0.3	**95.0 ± 0.1**	94.8 ± 0.7
	bird	76.2 ± 2.0	84.9 ± 1.0	**86.7 ± 0.3**
	cat	64.1 ± 0.8	77.1 ± 0.4	**85.8 ± 0.6**
	deer	83.4 ± 1.0	88.5 ± 0.2	**89.1 ± 0.5**
	dog	83.7 ± 0.8	86.7 ± 0.3	**87.4 ± 1.0**
	frog	89.3 ± 1.0	88.4 ± 0.1	**93.2 ± 0.4**
	horse	94.5 ± 0.2	**95.8 ± 0.0**	88.5 ± 0.5
	ship	92.2 ± 0.2	94.3 ± 0.1	**95.6 ± 0.3**
	truck	90.0 ± 0.2	90.9 ± 0.1	**92.6 ± 1.0**
	avg	83.7	87.8	**90.7**
CIFAR-100	0	72.9 ± 1.4	76.8 ± 1.1	**84.5 ± 0.5**
	1	66.0 ± 1.9	66.2 ± 2.0	**81.9 ± 2.6**
	2	74.3 ± 1.4	78.8 ± 1.9	**96.0 ± 0.3**
	3	76.3 ± 0.7	73.3 ± 3.1	**86.7 ± 1.0**
	4	76.2 ± 1.5	78.2 ± 1.4	**90.8 ± 1.6**
	5	59.8 ± 2.7	54.9 ± 2.7	**81.9 ± 1.2**
	6	69.2 ± 1.8	72.5 ± 2.6	**86.7 ± 1.1**
	7	65.2 ± 2.1	63.5 ± 1.4	**88.2 ± 0.5**
	8	75.3 ± 2.0	**86.6 ± 0.7**	82.7 ± 0.5
	9	87.3 ± 0.4	89.1 ± 0.3	**92.0 ± 0.5**
	10	78.9 ± 1.7	85.4 ± 2.1	**94.1 ± 0.3**
	11	83.1 ± 0.3	**85.7 ± 0.4**	85.4 ± 0.7
	12	78.3 ± 0.5	**84.1 ± 0.8**	84.0 ± 0.5
	13	59.5 ± 1.2	57.3 ± 0.7	**74.8 ± 1.6**
	14	82.5 ± 0.6	**90.7 ± 0.9**	90.1 ± 1.6
	15	66.1 ± 0.7	70.5 ± 0.8	**78.1 ± 1.1**
	16	64.1 ± 1.5	73.0 ± 1.7	**82.0 ± 0.5**
	17	92.5 ± 0.2	93.9 ± 0.3	**96.0 ± 0.2**
	18	89.0 ± 0.2	**90.2 ± 0.5**	90.1 ± 0.9
	19	82.6 ± 0.7	82.8 ± 1.7	**87.2 ± 0.7**
	avg	75.0	77.7	**86.7**
Fashion-MNIST	T-shirt/top	88.2 ± 0.3	94.1 ± 0.3	**98.4 ± 0.1**
	Trouser	98.9 ± 0.3	99.0 ± 0.5	**99.9 ± 0.0**
	Pullover	86.9 ± 0.6	92.2 ± 0.2	**98.5 ± 0.1**
	Dress	92.7 ± 0.3	89.3 ± 1.1	**99.2 ± 0.0**
	Coat	91.1 ± 0.1	91.7 ± 0.7	**98.3 ± 0.1**
	Sandal	95.7 ± 0.4	92.8 ± 0.5	**99.8 ± 0.1**
	Shirt	83.6 ± 0.4	85.2 ± 0.2	**95.7 ± 0.2**
	Sneaker	95.8 ± 0.3	97.9 ± 0.1	**99.8 ± 0.0**
	Bag	98.0 ± 0.1	96.7 ± 0.2	**99.8 ± 0.1**
	Ankle boot	99.4 ± 0.0	99.4 ± 0.4	**99.8 ± 0.0**
	avg	93.0	93.8	**98.9**

6 Conclusion and Future Work

In this paper, density ratio estimation under the KLIEP criterion was performed with CNN, and its effectiveness for anomaly detection was investigated. The method of deep anomaly detection has been actively discussed in recent years, but its main approach is to use a deep generative model (Schlegl et al., 2017; Deecke et al., 2018), Deep SVDD (Ruff et al., 2018), and Geometric Transformations (Golan & El-Yaniv, 2018). Our numerical experiments on the CIFAR-10, CIFAR-100 and Fashion-MNIST datasets demonstrated that deep direct density ratio estimation greatly improves the

Table 2: Average computation time in seconds with standard deviation for training on each datasets per methods.

Dataset	**GTs**	**KLIEP**
CIFAR-10	450.1 ± 1.8	85.8 ± 0.1
CIFAR-100	380.6 ± 1.0	47.0 ± 0.1
Fashion-MNIST	436.3 ± 2.0	100.1 ± 0.2

anomaly detection ability and reduces the computation time over state-of-the-art methods. We also showed that BatchNorm is not compatible with density ratio estimation using deep neural networks.

The objective function Eq.(8) continues to decrease regardless of whether or not BatchNorm is adopted. A similar phenomenon has been investigated in the context of learning from positive and unlabeled data (Kiryo et al., 2017) , which caused overfitting when a model with high expressive ability such as a deep neural network is used. More specifically, the empirical risk tends to be negative during training, and they proposed to cleverly impose a non-negativity constraint to avoid overfitting. On the other hand, when density ratio estimation is performed under the BA divergence, there is a constant term $C = \int p^*_{\mathrm{te}}(\boldsymbol{x}) f(r^*(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$ that is dropped during training. Since the value of C is unknown, we cannot directly impose a suitable non-negativity constraint in the current case. Thus, it is an important future work to explore a more stable learning algorithm for deep density ratio estimation.

REFERENCES

Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. In *Advances in Neural Information Processing Systems*, pp. 7694–7705. 2018.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *CoRR*, abs/1901.03407, 2019. URL `http://arxiv.org/abs/1901.03407`.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 3–17. 2018.

Jing Gao, Haibin Cheng, and Pang-Ning Tan. A novel framework for incorporating labeled examples into anomaly detection. In *Proceedings of the Sixth SIAM International Conference on Data Mining, April 20-22, 2006, Bethesda, MD, USA*, pp. 594–598, 2006.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pp. 9758–9769. 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680. 2014.

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*, pp. 770–778. 2016.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26(2):309–336, 2011.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Scholkopf. Correcting sample selection bias by unlabeled data. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pp. 601–608, 2006.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 448–456. 2015.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems 30*, pp. 1675–1685. 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105. 2012.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 38–44. 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

H. Nam and M. Sugiyama. Direct density ratio estimation with convolutional neural networks with application in outlier detection. *IEICE Transactions on Information and Systems*, E98-D(5): 1073–1079, 2015.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788. 2016.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4393–4402. 2018.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

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

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

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pp. 1433–1440, 2008.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York, NY, USA, 1st edition, 2012a. ISBN 0521190177, 9780521190176.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012b.

David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1): 45–66, 2004.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. BMVA Press, 2016.

A SUPERCLASS NAMES

Here is the list of superclass and classes in the CIFAR-100.

index	superclass	classes
0	Aquatic mammals	beaver, dolphin, otter, seal, whale
1	Fish	aquarium fish, flatfish, ray, shark, trout
2	Flowers	orchids, poppies, roses, sunflowers, tulips
3	Food containers	bottles, bowls, cans, cups, plates
4	Fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
5	Household electrical devices	clock, computer keyboard, lamp, telephone, television
6	Household furniture	bed, chair, couch, table, wardrobe
7	Insects	bee, beetle, butterfly, caterpillar, cockroach
8	Large carnivores	bear, leopard, lion, tiger, wolf
9	Large man-made outdoor things	bridge, castle, house, road, skyscraper
10	Large natural outdoor scenes	cloud, forest, mountain, plain, sea
11	Large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
12	Medium-sized mammals	fox, porcupine, possum, raccoon, skunk
13	Non-insect invertebrates	crab, lobster, snail, spider, worm
14	People	baby, boy, girl, man, woman
15	Reptiles	crocodile, dinosaur, lizard, snake, turtle
16	Small mammals	hamster, mouse, rabbit, shrew, squirrel
17	Trees	maple, oak, palm, pine, willow
18	Vehicles 1	bicycle, bus, motorcycle, pickup truck, train
19	Vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

B ROBUSTNESS AND TRANSFER LEARNING

Table 3: Average AUROC in % with standard deviation (over 5 trials with different seeds) per method.

Dataset	inlier class		**BA divergence**		**KLIEP**
		($\alpha = 1$)	($\alpha = 0.5$)	($\alpha \to 0$)	(**ImageNet**)
CIFAR-10	plane	73.8 ± 6.7	82.2 ± 5.2	93.6 ± 0.3	96.9 ± 0.4
	car	80.2 ± 6.7	94.4 ± 2.0	94.8 ± 0.7	99.0 ± 0.1
	bird	82.4 ± 2.3	82.5 ± 1.3	86.7 ± 0.3	94.3 ± 0.3
	cat	80.2 ± 1.2	81.7 ± 2.0	85.8 ± 0.6	91.4 ± 0.8
	deer	80.2 ± 1.1	85.6 ± 0.8	89.1 ± 0.5	96.1 ± 0.2
	dog	79.7 ± 3.5	86.2 ± 0.7	87.4 ± 1.0	94.5 ± 0.6
	frog	83.3 ± 3.4	90.8 ± 0.7	93.2 ± 0.4	97.5 ± 0.4
	horse	73.8 ± 6.7	88.9 ± 3.4	88.5 ± 0.5	97.3 ± 0.2
	ship	89.4 ± 2.5	93.1 ± 2.6	95.6 ± 0.3	98.6 ± 0.3
	truck	83.4 ± 1.1	92.2 ± 0.7	92.6 ± 1.0	98.5 ± 0.1
	avg	80.6	87.8	90.7	96.4
CIFAR-100	0	76.3 ± 0.5	91.4 ± 2.5	84.5 ± 0.5	90.4 ± 1.4
	1	67.7 ± 2.3	87.8 ± 1.3	81.9 ± 2.6	89.9 ± 0.9
	2	85.3 ± 3.6	77.3 ± 7.5	96.0 ± 0.3	96.4 ± 1.3
	3	66.3 ± 4.3	81.6 ± 1.6	86.7 ± 1.0	94.6 ± 0.4
	4	70.8 ± 1.9	83.5 ± 1.7	90.8 ± 1.6	93.1 ± 2.2
	5	62.5 ± 1.5	74.1 ± 2.0	81.9 ± 1.2	92.2 ± 0.7
	6	67.9 ± 4.5	83.9 ± 4.8	86.7 ± 1.1	94.0 ± 0.7
	7	74.6 ± 4.3	73.7 ± 6.9	88.2 ± 0.5	92.5 ± 0.5
	8	76.2 ± 3.4	79.1 ± 2.7	82.7 ± 0.5	89.3 ± 4.7
	9	79.8 ± 6.4	79.4 ± 2.9	92.0 ± 0.5	97.1 ± 0.3
	10	86.5 ± 0.5	69.7 ± 4.5	94.1 ± 0.3	95.9 ± 0.5
	11	72.5 ± 3.6	75.2 ± 4.1	85.4 ± 0.7	88.0 ± 4.4
	12	75.0 ± 1.6	87.3 ± 1.1	84.0 ± 0.5	89.2 ± 0.6
	13	71.9 ± 1.5	78.9 ± 2.9	74.8 ± 1.6	88.0 ± 1.2
	14	71.3 ± 9.2	77.2 ± 5.9	90.1 ± 1.6	93.7 ± 1.4
	15	73.7 ± 1.7	74.6 ± 5.8	78.1 ± 1.1	86.7 ± 1.4
	16	76.0 ± 2.4	72.4 ± 1.3	82.0 ± 0.5	89.2 ± 0.2
	17	79.3 ± 6.6	92.4 ± 1.0	96.0 ± 0.2	97.4 ± 1.8
	18	71.7 ± 8.8	83.0 ± 1.2	90.1 ± 0.9	95.1 ± 0.8
	19	73.2 ± 3.6	79.9 ± 3.3	87.2 ± 0.7	93.4 ± 0.7
	avg	73.9	80.1	86.2	92.3
Fashion-MNIST	T-shirt/top	93.4 ± 3.1	97.8 ± 0.2	98.4 ± 0.1	-
	Trouser	86.1 ± 2.7	99.8 ± 0.2	99.9 ± 0.0	-
	Pullover	94.3 ± 1.8	97.1 ± 1.3	98.5 ± 0.1	-
	Dress	95.1 ± 0.8	98.6 ± 0.1	99.2 ± 0.0	-
	Coat	89.2 ± 4.9	97.3 ± 0.6	98.3 ± 0.1	-
	Sandal	99.6 ± 0.2	92.8 ± 0.5	99.8 ± 0.1	-
	Shirt	89.8 ± 2.4	92.7 ± 2.3	95.7 ± 0.2	-
	Sneaker	95.4 ± 4.6	99.7 ± 0.1	99.8 ± 0.0	-
	Bag	87.3 ± 1.5	99.5 ± 0.2	99.8 ± 0.1	-
	Ankle boot	94.3 ± 5.8	99.8 ± 0.0	99.8 ± 0.0	-
	avg	92.5	97.5	98.9	-