One-Class Classification with Deep Adversarial Learning

Liane Xu
Winter Park High School
Orlando FL, USA
1-407-478-8688
lianex85@gmail.com

Zhiguang Xu
Department of Computer Science
Valdosta State University
Valdosta GA, USA
1-229-333-5783
zxu@valdosta.edu

ABSTRACT
One-class classification (OCC) seeks to build a machine-learning model when the negative class is either absent, poorly sampled, or not well defined. In this paper, we present a deep adversarial learning based architecture for one-class classification. Our architecture is composed of two deep neural networks, a generator and a discriminator, that are competing while collaborating with each other since it is inspired by the success of Generative Adversarial Networks (GANs). The generator network contains a Deconvolutional Neural Networks (aka. decoder), which is trained using a zero-centered Gaussian noise as the feature representation of the pseudo-negative class to learn a good representation as well as the boundary for the classifiable distortion of the target (or positive) class with the assistance from the target class. The outputs produced by the generator network are aggregated with the real positive class data samples, which are then used to train the discriminator network, whose goal is to understand the underlying concept in the positive class, and then classify the negative testing samples. The proposed architecture applies to a variety of OCC problems such as novelty detection, anomaly detection, and mobile user authentication. The experiments on MNIST and Caltech-256 images demonstrate that our architecture achieves superior results over the recent state-of-the-art approaches.

CCS Concepts
• Computing Methodologies → Artificial Intelligence → Knowledge Representation and Reasoning

Keywords
One-Class Classification; Generative Adversarial Networks; Convolutional Neural Networks.

1. INTRODUCTION
Traditional multi-class classification aims to build a machine-learning model that classifies an unknown object sample into one of the several pre-defined object classes. In contrast, in one-class classification (OCC), the objective of the model is to identify an object sample that does not belong to any of those classes. Despite that the absence of data from the negative class during the training of the OCC model makes it difficult, the OCC problem is closely related to a wide range of real-world applications including novelty detection, anomaly detection, and mobile user authentication, thus has gained great interest of the research community in recent years.

Sabokrou et al. [1] proposed a general framework for one-class classification and novelty detection in images and videos, trained in an adversarial and unsupervised manner. Inspired by Generative Adversarial Networks (GAN), their architecture consists of two modules, Reconstructor (aka. Generator) and Discriminator. The former is a pair of encoder-decoder Convolutional Neural Networks (CNNs) that learn the concept of a target class to reconstruct images such that the latter, which is another CNN, is fooled to consider those reconstructed images as real target class images. After training the model, R can reconstruct target class samples correctly, while it distorts and decimates samples that do not have the concept shared among the target class samples. This eventually helps D discriminate the testing samples even better.

In this paper, we focus on the Reconstructor module in the GAN-based architecture as described above and improve it in two ways. First, we remove the encoder CNN entirely from R such that not only do we cut the number of parameters in half, but also achieve a more efficient training procedure of the module. Second, output from the decoder (a Deconvolutional Neural Network that remains its presence in R) is then aggregated with the positive class before flowing to the Discriminator network. Specifically, assuming the features eventually extracted from the target class samples by D are D-dimensional, the decoder is trained using a zero-centered D-dimensional Gaussian noise as the feature representation of the pseudo-negative class to learn a good representation as well as the boundary for the classifiable distortion of the target class. That way, the goal of the decoder network is simplified because it only needs to learn to map a random vector to a distortion of the real data sample, rather than the sample itself following the real data distribution. The experiments on MNIST and Caltech-256 images demonstrate that our proposed architecture achieves superior results over the recent state-of-the-art approaches.

The outline of the rest of this paper is as follows. Section 2 provides a review of related works of OCC and Deep Adversarial Learning. In Section 3, we propose our approach to the OCC problem. Experimental results are presented in Section 4, which is followed by Section 5 that concludes this paper.

2. RELATED WORKS
2.1 One-Class Classification
The task in OCC is to define a classification boundary around the positive class, such that it accepts as many objects as possible

CSAI2019, December 6–8, 2019, Beijing, China

ACM ISBN 978-1-4503-7627-3/19/12…$15.00
DOI: https://doi.org/10.1145/3374587.3374609

from the positive class, while it minimizes the chance of accepting the outliner objects from the negative class. One-class classification is closely related to rare event detection, outlier detection/removal, and anomaly detection. All these applications share the search procedure for a novel concept, which is scarcely seen in the training data and hence can all be encompassed by the umbrella term novelty detection.

The objective of one-class feature learning could be formulated as an optimization problem, where any feature g extracted from the training samples t must simultaneously satisfy both of the following characteristics: Compactness C and Descriptiveness D.

$$\hat{g} = max_g\ D(g(t)) + \lambda C(g(t)) \#(1)$$

where λ is a positive constant, D is the descriptiveness or inter-class distance of samples from different classes in the feature space, C is the compactness or intra-class distance of samples from the same class in the feature space.

In the literature for one-class classification where information regarding the negative class data is unavailable, various methods have been proposed and they can be classified under two broad categories: those that are Support Vector Machine (SVM) based and those that are Deep Neural Network based. Scholkopf et al. [2] proposed One-Class Support Vector Machine (OC-SVM), which maximizes the boundary between data classes with respect to the origin. Another popular SVM based approach is Support Vector Data Description (SVDD) introduced by Tax et al. [3], in which a hypersphere that encloses the target (or positive) class data is sought. Over the last several years, methods based on Convolutional Neural Networks (CNN) have shown impressive performance improvements for image object detection and recognition. Although CNNs alone could not be directly used for the OCC problems, especially in an end-to-end and straightforward manner, they are commonly seen as building blocks in quite a few models [1, 4, 5, 6] inspired by Generative Adversarial Networks (GAN), which are also what our research as described in this paper is based on.

2.2 Deep Adversarial Learning

In the recent years, GANs [7, 8] have shown outstanding success in generating data for learning models. They have also been extended to classification models even in the presence of not enough labeled training data (e.g., in [9, 10, 11]). They are based on a two-player game between two different networks, both concurrently trained in an unsupervised fashion. One network is the generator (G), which aims at generating realistic data (e.g., images), while the second network poses as the discriminator (D), and tries to discriminate real data from the data generated by G. One of the different types of GANs, closely related to our work, is the conditional GANs, in which G takes an image X as the input and generates a new image X'. Whereas, D tries to distinguish X from X', while G tries to fool D producing more and more realistic images. Very recently, Isola et al. [12] proposed an "Image-to-image translation" framework based on conditional GANs, where both G and D are conditioned on the real data. They showed that a U-Net encoder-decoder [13] with skip connections could be used as the generator coupled with a patch-based discriminator to transform images with respect to different representations. In a concurrent work, [14] proposed to learn the generator as a reconstructor of normal events, and hence if it cannot properly reconstruct a chunk of the input frames, that chunk is considered an anomaly.

3. PROPOSED APPROACH

Figure 1 above gives an overview of the proposed Deep Adversarial Learning for One-Class Classification (DAL-OCC) architecture. Like in Sabokrou's work [1], the overall architecture is composed of two modules – a Reconstructor (aka. Generator) and a Discriminator. These two modules are trained in an adversarial and unsupervised fashion within an end-to-end setting.

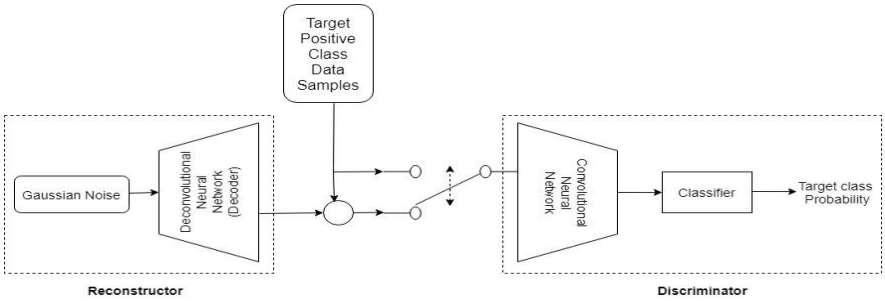

Figure 1. Overview of the proposed architecture.

3.1 Reconstructor – A Deconvolutional Neural Network

Assuming that the extracted features from the target class are D-dimensional, the input to the Reconstructor module is a Gaussian noise, $\mathcal{N}(\bar{\mu}, \sigma^2 \cdot I)$, where $\bar{\mu}$ and σ are the parameters of Gaussian and I is a $D \times D$ identity matrix. Hence, $\mathcal{N}(\bar{\mu}, \sigma^2 \cdot I)$ can be seen as generating D independent one dimensional Gaussian with σ standard deviation. Sitting in the center of R is a Deconvolutional Neural Network of three layers. To improve the stability of the network, we do not use any pooling layers, instead, a batch normalization operation is exploited after each deconvolutional layer.

3.2 Discriminator – A Convolutional Neural Network

The Discriminator module consists of a Convolutional Neural Network (CNN) of three layers, which is trained to eventually distinguish the novel or outlier sample in an unsupervised fashion. This CNN is followed by a simple classifier (i.e. a fully-connected layer plus a softmax regression layer). The following binary cross-entropy loss function is used to train the entire DAL-OCC network.

$$L_c = -\frac{1}{K}\sum_{j=1}^{K}(y \ln(p) + (1-y)\ln(1-p)) \#(2)$$

where $y \in \{0,1\}$ indicates whether the classifier input corresponds to the target positive class (i.e. $y = 0$), or it is generated by R from $\mathcal{N}(\bar{\mu}, \sigma^2 \cdot I)$, (i.e. $y = 1$). Additionally, K represents the batch size and p indicates the softmax probability of $y = 0$. Both D and R are optimized using the Adam optimizer with the learning rate of 10^{-4}.

3.3 Adversarial Learning Orchestrated by the Target Class

The DAL-OCC approach that we propose is based on the Generative Adversarial Networks (GANs) introduced by Goodfellow et al. [14]. Like in GANs, the Reconstructor (aka. Generator) learns to map any noise Z from a latent space following a specific distribution $\boldsymbol{p_z}$ (Gaussian in this case) to a data sample that follows the target data distribution $\boldsymbol{p_t}$ whereas the Discriminator tries to discriminate between actual data and the fake data generated by R. R and D are trained in a two-player min-max game:

$$\min_R \max_D \left(\mathbb{E}_{X \sim p_t}[\ln D(X)] + \mathbb{E}_{Z \sim p_z}[\ln(1 - D(R(Z)))] \right) \#(3)$$

As shown in Figure 1, the adversarial training of R and D is orchestrated and assisted by the target positive class. We train both of them in alternating steps and lock them into a competition to improve themselves. We fix the models' parameters in R and perform a single iteration of gradient descent on D using the data samples from target positive class. Then we switch sides. We fix D parameters and train R for another single iteration. Notice that in order to accelerate the training of R, its output is aggregated with the positive class before flowing to D. That way, during gradient descent training, D only needs to learn to map a random vector to a distortion of the real data sample, rather than the sample itself following the real data distribution.

Eventually, the Discriminator identifies the tiny difference between the real and the generated, and the Reconstructor creates results that the Discriminator cannot tell the difference. The DAL-OCC architecture eventually converges and can be used for solving the One-Class Classification problems.

4. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed DAL-OCC architecture for outlier images detection on two different image datasets. All the results reported in this section are from our implementation using the TensorFlow/Keras framework and Python 3 running on an Nvidia Tesla C2070 GPU. They are analyzed in details and compared with state-of-the-art approaches.

4.1 Outlier Images Detection

Machine learning methods often experience considerable performance degradation in the presence of gross outliers when they fail to deal with data contaminated by noise and outliers in realistic vison-based datasets. Our DAL-OCC architecture is capable of learning the shared concepts among all inlier images, and hence classify the outliers. We evaluate the performance of our proposed method using the MNIST[1] and Caltech-256 datasets[2].

MNIST: This dataset includes 60,000 handwritten digits from "0" to "9". Each of the ten categories of digits is taken as the target positive class (i.e., inliers), and we simulate outliers by randomly sampling images from other categories with a proportion of 10% to 50%. This experiment is repeated for all of the ten digit categories.

Caltech-256: This dataset contains 256 object categories with a total of 30,607 images. Each category has at least 80 images. Similar to previous works [15], we repeat the procedure three times and use images from $n \in \{1,3,5\}$ randomly chosen categories as inliers (i.e., target). The first 150 images of each category are used, if that category has more than 150 images. A certain number of outliers are randomly selected from the "clutter" category, such that each experiment has exactly 50% outliers.

4.2 MNIST Results

The DAL-OCC architecture is trained on the MNIST images of the target classes, in the absence of any outliers. Then it is tested with the presence of outliers. In Figure 2 below, we report the F_1-score of our approach and compare it with two state-of-the-art methods (LOF [16] and DRAE [17]) as a function of the portion of the outliers in the test dataset.

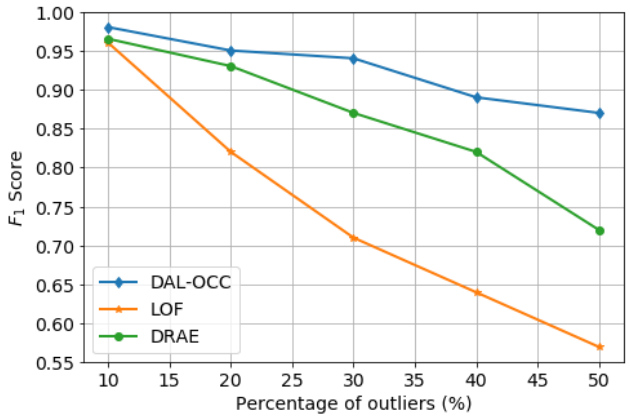

Figure 2. F_1 scores on MNIST dataset.

As shown above, our proposed approach not only outperforms both LOF and DRAE, but also displays consistent robustness while the two baseline methods fail to detect the outliers as their portion increases.

4.3 Caltech-256 Results

In this experiment, we compare our DAL-OCC approach with six state-of-the-art methods specifically designed for outlier detection – Coherence Pursuit (CoP), REAPER, OutlierPursuit (OP), Low-Rank Representation (LRR), Dual Principal Component Pursuit (DPCP), and OutRank. The measurements of these methods are borrowed from [15]. The comparison results of F_1-score and Area Under the ROC Curve (AUC) are summarized in Table 1 below.

Table 1. F_1 and AUC scores on the Caltech-256 dataset

	Inliers from one Category		Inliers from three Categories		Inliers from five Categories	
	AUC	F_1	AUC	F_1	AUC	F_1
CoP	0.905	0.880	0.676	0.718	0.487	0.672
REAPER	0.816	0.808	0.796	0.784	0.657	0.716
OP	0.837	0.823	0.788	0.779	0.629	0.711
LRR	0.907	0.893	0.479	0.671	0.337	0.667
DPCP	0.783	0.785	0.798	0.777	0.676	0.715
OutRank	0.948	**0.914**	0.929	0.880	0.913	0.858
DAL-OCC	**0.949**	0.913	**0.937**	**0.908**	**0.921**	**0.909**

[1] Available at http://yann.lecun.com/exdb/mnist

[2] Available at http://www.vision.caltech.edu/Image_Datasets/Caltech256

As shown above, in all cases except $\mathbf{F_1}$ score for the case of inliers from one category, our proposed architecture is superior to the baseline methods. Also, similar to what was seen in the MNIST experiment, our approach demonstrates great consistency as the number of inlier classes increase.

5. CONCLUSIONS

In One-Class Classification (OCC), it is normally assumed that one does not have a priori knowledge of the novel class data. Hence, the learning process involves only the target positive class data. In this paper, we propose a Deep Adversarial Learning based solution to the OCC problem. Inspired by the immense success of GANs, our method has two modules. The first module contains a Deconvolutional Neural Network, which not only reconstructs the target class, but also helps to improve the performance of the second module (a Convolutional Neural Network) on any given testing image, and decimating/distorting the anomaly or outlier samples. The training of these two modules is conducted in alternating steps and orchestrated by the target class data samples. We have used our approach in a couple of outlier detection applications. The experimental results demonstrate consistent performance improvements over the start-of-the-art approaches.

6. REFERENCES

[1] Sabokrou, M., Khalooei, M., and Fathy, M. 2018. Adversarially Learned One-Class Classifier for Novelty Detection. *IEEE/CVF Conf. Computer Vision and Pattern Recognition* 2018. DOI=10.1109/cvpr.2018.00356.

[2] B. Sch ̈olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, Estimating the support of a high-dimensional distribution, *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[3] D. M. Tax and R. P. Duin, Support vector data description, *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[4] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, Abnormal event detection in videos using generative adversarial nets, *in Image Processing (ICIP), 2017 IEEE International Conference on. IEEE, 2017*, pp. 1577–1581.

[5] C. Zhou and R. C. Paffenroth, Anomaly detection with robust deep autoencoders, *in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, 2017, pp. 665–674.

[6] R. Chalapathy, A. K. Menon, and S. Chawla, Robust, deep and inductive anomaly detection, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 36–51.

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

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

[9] W. Lawson, E. Bekele, and K. Sullivan. Finding anomalies with generative adversarial networks for a patrolbot. In *CVPR*, pages 12–13, 2017.

[10] T. Schlegl, P. Seeb ̈ock, S. M. Waldstein, U. Schmidt-Erfurth, and G. 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.

[11] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. *arXiv preprint arXiv:1706.07680*, 2017.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

[14] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In *IEEE International Conference on Image Processing (ICIP)*, 2017.

[15] C. You, D. P. Robinson, and R. Vidal. Provable self-representation based outlier detection in a union of subspaces. In *CVPR*, 2017.

[16] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[17] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *ICCV*, pp. 1511–1519, 2015.