Deep One-Class Classification

Parag Goyal (IIT2018164), Bhavya Girotra (IIT2018167) , Yash Katiyar (IIT2018170)

VI Semester B.tech IT Section B

Indian Institute of Information Technology, Allahabad, UP, India

Abstract*—Despite the great advances made by deep learning in many machine learning problems, there is a relative dearth of deep learning approaches for anomaly detection. Those approaches which do exist involve networks trained to perform a task other than anomaly detection, namely generative models or compression, which are in turn adapted for use in anomaly detection; they are not trained on an anomaly detection based objective. In this paper we introduce a new anomaly detection method—Deep Support Vector Data Description—, which is trained on an anomaly detection based objective. The adaptation to the deep regime necessitates that our neural network and training procedure satisfy certain properties, which we demonstrate theoretically. We show the effectiveness of our method on MNIST and CIFAR-10 image benchmark datasets as well as on the detection of adversarial examples of GTSRB stop signs.

I. INTRODUCTION

Anomaly detection (AD) (Chandola et al., 2009; Aggarwal, 2016) is the task of discerning unusual samples in data. Typically, this is treated as an unsupervised learning problem where the anomalous samples are not known a priori and it is assumed that the majority of the training dataset consists of “normal” data (here and elsewhere the term “normal” means not anomalous and is unrelated to the Gaussian distribution). The aim then is to learn a model that accurately describes “normality.” Deviations from this description are then deemed to be anomalies. This is also known as one class classification (Moya et al., 1993). AD algorithms are often trained on data collected during the normal operating state of a machine or system for monitoring (Lavin & Ahmad, 2015). Other domains include intrusion detection for cybersecurity (Garcia-Teodoro et al., 2009), fraud detection (Phua et al., 2005), and medical diagnosis (Salem et al., 2013; Schlegl et al., 2017). As with many fields, the data in these domains is growing rapidly in size and dimensionality and thus we require effective and efficient ways to detect anomalies in large quantities of high-dimensional data.

III. SVDD OPTIMIZATION

We use stochastic gradient descent (SGD) and its variants to optimize the parameters W of the neural network in both Deep SVDD objectives using backpropagation. Training is carried out until convergence to a local minimum. Using SGD allows Deep SVDD to scale well with large datasets as its computational complexity scales linearly in the number of training batches and each batch can be processed in parallel (e.g. by processing on multiple GPUs). SGD optimization also enables iterative or online learning.

IV. ALGORITHM

Step 1: We build on the kernel-based SVDD and minimum volume estimation by finding a data-enclosing hypersphere of smallest size.

Step 2: we employ a neural network that is jointly trained to map the data into a hypersphere of minimum volume

Step 3: Deep SVDD optimization and selection of the hypersphere is done.

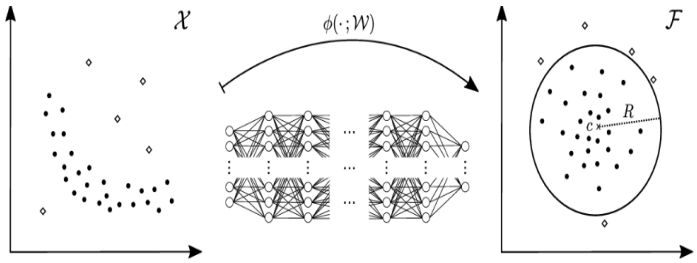


Fig : Deep SVDD learns Neural Network Transformation

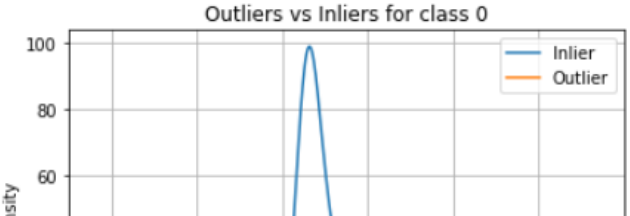
V. RESULT

We evaluate Deep SVDD on the well-known MNIST and CIFAR-10 datasets. Adversarial attacks have seen a lot of attention recently and here we examine the possibility of using anomaly detection to detect such attacks.To do this we apply Boundary Attack to the GTSRB stop signs dataset.

ROC SCORES FOR EACH CLASSES WERE:

```
ROC scores for class  0  is:  98.79004706095299
ROC scores for class  1  is:  99.60754439450295
ROC scores for class  2  is:  91.3406906727797
ROC scores for class  3  is:  90.24187491051664
ROC scores for class  4  is:  92.38666816626986
ROC scores for class  5  is:  85.27766453775423
ROC scores for class  6  is:  97.23715100812308
ROC scores for class  7  is:  95.31597221619877
ROC scores for class  8  is:  94.77695282303326
ROC scores for class  9  is:  95.8508558112126
```

We compare our method against a diverse collection of state of-the-art methods from different paradigms. We use image data since they are usually high-dimensional and moreover allow for a qualitative visual assessment of detected anomalies by human observers. Using classification datasets to create one-class classification setups allows us to evaluate the results quantitatively via AUC measure by using the ground truth labels in testing.



Normal Class	OC-SVM/ SVDD	KDE	IF	DCAE	AnoGAN	Soft-bound. Deep SVDD	One-Class Deep SVDD
0	**98.6**±0.0	97.1±0.0	98.0±0.3	97.6±0.7	96.6±1.3	97.8±0.7	98.0±0.7
1	99.5±0.0	98.9±0.0	97.3±0.4	98.3±0.6	99.2±0.6	99.6±0.1	**99.7**±0.1
2	82.5±0.1	79.0±0.0	88.6±0.5	85.4±2.4	85.0±2.9	89.5±1.2	**91.7**±0.8
3	88.1±0.0	86.2±0.0	89.9±0.4	86.7±0.9	88.7±2.1	90.3±2.1	**91.9**±1.5
4	**94.9**±0.0	87.9±0.0	92.7±0.6	86.5±2.0	89.4±1.3	93.8±1.5	**94.9**±0.8
5	77.1±0.0	73.8±0.0	85.5±0.8	78.2±2.7	88.3±2.9	85.8±2.5	**88.5**±0.9
6	96.5±0.0	87.6±0.0	95.6±0.3	94.6±0.5	94.7±2.7	98.0±0.4	**98.3**±0.5
7	93.7±0.0	91.4±0.0	92.0±0.4	92.3±1.0	93.5±1.8	92.7±1.4	**94.6**±0.9
8	88.9±0.0	79.2±0.0	89.9±0.4	86.5±1.6	84.9±2.1	92.9±1.4	**93.9**±1.6
9	93.1±0.0	88.2±0.0	93.5±0.3	90.4±1.8	92.4±1.1	94.9±0.6	**96.5**±0.3

VI. CONCLUSION

We introduced the first fully deep one-class classification objective for unsupervised AD in this work. Our method, Deep SVDD, jointly trains a deep neural network while optimizing a data-enclosing hypersphere in output space. Through this Deep SVDD extracts common factors of variation from the data. We have demonstrated theoretical properties of our method such as the v-property that allows to incorporate a prior assumption on the number of outliers being present in the data. Our experiments demonstrate quantitatively as well as qualitatively the sound performance of Deep SVDD.

VII. REFERENCES

[1] An, J. and Cho, S. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. SNU DataMining Center, Tech. Rep., 2015.

[2] Hinton, G. E. and Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. Science, 313(5786):504–507, 2006.

[3] Andrews, J. T. A., Morton, E. J., and Griffin, L. D. Detecting Anomalous Data Using Auto-Encoders. IJMLC, 6(1):21,2016.