Deep One-Class Classification on MNIST using Deep SVDD

Omar Qawasmi
Artificial intelligence and data engineering department
Istanbul Technical University
Istanbul, Turkey

***Abstract*—This project addresses the problem of one-class classification (OCC), specifically focusing on identifying anomalies from a set norm in a visual dataset. This is crucial for applications such as fault detection and quality control, where anomaly data is scarce or undefined. We compare Deep Support Vector Data Description (Deep SVDD) with traditional anomaly detection methods like One-Class SVM (OC-SVM) and Isolation Forest (IF). Our hypothesis is that a Convolutional Neural Network (CNN) trained with Deep SVDD can learn a superior feature representation where normal instances cluster tightly around a center, thereby outperforming traditional methods in anomaly detection on the MNIST dataset. The digit '1' was used as the normal class, with all other digits considered anomalies.**

***Index Terms*—One-Class Classification, Deep SVDD, Anomaly Detection, MNIST, Convolutional Neural Networks, Deep Learning**

I. Introduction

The challenge in anomaly detection often lies in the availability of data: typically, only examples of normal behavior are available during training, with anomalies being rare or unknown. Unlike traditional supervised classification, where labels for all classes are present, one-class classification tackles this by building a model based solely on the 'normal' class. This project applies OCC to the MNIST dataset, treating one specific digit (digit '1') as the normal class and all other digits as anomalies. We hypothesize that Deep SVDD, which leverages a neural network for feature extraction, will be more effective than conventional methods like OC-SVM and Isolation Forest, which are applied after explicit feature engineering.

II. Methodology

A. Data Loading and Preprocessing

The MNIST dataset, consisting of 60,000 training and 10,000 testing images of handwritten digits (28x28 pixels, grayscale), was used. Pixel values were normalized to $[0, 1]$, and a channel dimension was added for CNN input.

- **Normal Class:** Digit '1' was designated as the normal class.
- **Training Data:** `x_train_normal` consists of 6742 samples of digit '1'.
- **Test Data:** `x_test_final` consists of 10000 samples, comprising 1135 normal samples (digit '1') and 8865 anomaly samples (other digits). The test set was shuffled to ensure a balanced evaluation.

B. Deep SVDD Model Development and Training

The Deep SVDD model employs a Convolutional Neural Network (CNN) as its feature extractor, $\phi(x; W)$.

1) CNN Architecture (Feature Extractor): The CNN architecture used is a sequential model:

- `Conv2D` (32 filters, 3x3 kernel, ReLU activation)
- `MaxPooling2D` (2x2 pool size)
- `Conv2D` (64 filters, 3x3 kernel, ReLU activation)
- `MaxPooling2D` (2x2 pool size)
- `Flatten`
- `Dense` layer with an `EMBEDDING_DIM` of 128 (output feature vector).

The total number of trainable parameters in this CNN is 223,744.

2) Deep SVDD Loss Function: The objective of Deep SVDD is to map normal data points close to a fixed center c in the output feature space. The loss function minimizes the volume of a data-enclosing hypersphere:

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} ||\phi(x_i; W) - c||^2 + 2\lambda \sum_{l=1}^{L} ||W_l||_F^2$$

where W represents the network weights, c is the hypersphere center, and λ is a regularization parameter (1×10^{-6}) for weight decay.

3) Center 'c' Initialization: The center c was initialized in a data-dependent manner by computing the mean of the feature embeddings of 10 batches of normal training data. This ensures the center is representative of the normal class's features.

4) Training Loop: The model was trained using the SGD optimizer with a learning rate of 1×10^{-2} for a maximum of 50 epochs. An early stopping mechanism was implemented, triggering if the average epoch loss fell below 0.0020. The training process converged in 28 epochs.

C. Deep SVDD Anomaly Scoring and Evaluation

After training, anomaly scores for the test dataset were calculated as the squared Euclidean distance between the embeddings of the test samples and the learned center c: $Score(x) = ||\phi(x; W*) - c||^2$.

A threshold for anomaly detection was determined from the normal training data by taking the 99th percentile of their anomaly scores. This threshold was then used to classify test samples as normal or anomalous.

D. Comparison with Traditional Methods

To provide a comprehensive evaluation, the performance of Deep SVDD was compared against two traditional anomaly detection algorithms:

1) **One-Class SVM (OC-SVM):** Configured with `nu=0.1` and an 'rbf' kernel, with `gamma='scale'`.
2) **Isolation Forest (IF):** Configured with `random_state=42` and `contamination='auto'`.

For both traditional methods, the input images were flattened into 784-dimensional vectors. Anomaly scores were derived from the `decision_function` outputs of these models.

III. Results

A. Deep SVDD Performance

- **Training Convergence:** Deep SVDD training completed in 28 epochs due to early stopping, with the average epoch loss dropping to 0.0019, below the 0.0020 threshold.
- **Anomaly Scores (Test Set):**
  - Min: 0.0006
  - Max: 0.0244
  - Mean: 0.0072
  - Median: 0.0072
- **ROC AUC Score:** Deep SVDD achieved a ROC AUC score of 0.9975.
- **Threshold Calculation (from Normal Training Data):**
  - Min anomaly score: 0.0005
  - Max anomaly score: 0.0130
  - **Calculated Threshold (99th percentile of NORMAL TRAINING scores):** 0.0053
- **Classification Metrics (using threshold 0.0053):**
  - Accuracy: 0.8715
  - Precision: 0.9993
  - Recall: 0.8556
  - F1-Score: 0.9219

B. Traditional Methods Performance

- **One-Class SVM ROC AUC Score:** 0.9971
- **Isolation Forest ROC AUC Score:** 0.9929

C. Performance Comparison (ROC AUC)

A summary of the ROC AUC scores for all methods:

- Deep SVDD (Custom CNN): 0.9975
- One-Class SVM: 0.9971
- Isolation Forest: 0.9929

IV. Conclusion

The experimental results demonstrate that the Deep SVDD model, utilizing a custom CNN for feature extraction, achieves exceptional performance in anomaly detection on the MNIST dataset, with a ROC AUC score of 0.9975. This performance is marginally superior to that of One-Class SVM (0.9971) and significantly better than Isolation Forest (0.9929), confirming our hypothesis that combining deep learning for feature extraction with the Deep SVDD objective is highly effective. The classification metrics (Accuracy: 0.8715, Precision: 0.9993, Recall: 0.8556, F1-Score: 0.9219) obtained using a robust threshold derived from normal training data further underscore its practical utility. This project successfully highlights the advantages of Deep SVDD for visual anomaly detection tasks where only normal data is available for training.

References

References

[1] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443-1471.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. https://arxiv.org/abs/2010.11929

[3] Liu, F. T., Ting, K. M., Zhou, Z. H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.