2021 IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) | 978-1-7281-7691-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/USBEREIT51232.2021.9455004

Anomaly Detection in Image Datasets Using Convolutional Neural Networks, Center Loss, and Mahalanobis Distance

Garnik Vareldzhan
R&D Department
OOO Code Laboratory
Perm, Russian Federation
gavid19912@gmail.com

Kirill Yurkov
Mechanics and Mathematics
Perm State University
Perm, Russian Federation
opcheese@gmail.com

Konstantin Ushenin
Mathematical Physiology
Institute of Immunology and Physiology
Ekaterinburg, Russian Federation
konstantin.ushenin@urfu.ru

***Abstract*—User activities generate a significant number of poor-quality or irrelevant images and data vectors that cannot be processed in the main data processing pipeline or included in the training dataset. Such samples can be found through manual analysis by an expert or with anomalous detection algorithms. There are several formal definitions for anomalous samples. For neural networks, anomalies are usually defined as out-of-distribution samples. This work proposes methods for supervised and semi-supervised detection of out-of-distribution samples in image datasets. Our approach extends a basic neural network that solves the image classification problem. Thus, after extension one neural network can solve image classification and anomalous detection problems simultaneously. The proposed methods are based on the center loss and its effect on deep feature distribution in a last hidden layer of the neural network. This paper provides an analysis of the proposed methods for the LeNet and EfficientNet-B0 on the MNIST and ImageNet-30 datasets.**

***Index Terms*—anomaly detection, novelty detection, outlier detection, deep feature space, EfficientNet**

I. Introduction

As an image-processing approach, convolutional neural networks demonstrate the best performance for the image classification problem. However, the usage of neural networks in real-life applications has a number of practical challenges. Primarily, these challenges are related to datasets: the small size of the training dataset, train-test leakage, sample imbalance, and others. Other problems are related to user behavior, such as irrelevant and poor-quality user-produced data.

Detection of unusual, irrelevant, or adversarial data is important both for the creation of a proper training dataset and for filtering irrelevant samples during the inference. Several formal definitions are proposed for these problems, such as anomaly detection, novelty detection, outlier detection, or out-of-distribution detection. In this work, we use the definitions from [1]. Anomalous detection is a general term for the detection of any unusual or unwanted data. Meanwhile, out-of-distribution (OOD) detection is a more specific problem. This refers to the detection of samples that are included in the distribution of test samples (inference samples) but are not included in the distribution of training samples. The OOD detection problem assumes that detected samples unintentionally appear in the test dataset and that their inclusion in the training dataset is not required after detection. This problem focus differentiates OOD detection from adversarial attack detection and novelty detection, respectively.

According to [1], OOD detection approaches may be trained with supervised, semi-supervised, or unsupervised learning. In the first case, the algorithm observes in-distribution and OOD samples with proper markup. Thus, ODD detection is equivalent to a binary classification problem. Semi-supervised learning uses only the in-distribution samples. Unsupervised learning does not use any data markup. Evaluation of OOD detection methods requires the main dataset with in-distribution samples and an anomalous dataset with proper out-of-distribution samples.

There are numerous methods for anomaly and OOD detection in vector datasets: the local outlier factor, Mahalanobis distance, isolation forest, one-class support vector machine, autoencoder, variational autoencoder [2], [3], the Gaussian mixture model for variational autoencoder, and others. Most of them are implemented in open-source libraries [4], [5]. Methods for anomaly detection in image datasets have also been proposed. For example, this is a convolutional autoencoder. Recent studies propose methods based on contrastive learning [6] and variations of a RotNet [7].

All listed approaches focus only on anomalous or ODD detection problems. However, these problems are rarely the main goals of data processing, and usually they are only small parts of a bigger data processing workflow. Neural networks are a computationally expensive approach. For this reason, an extension of neural networks is preferable to adding one more neural network to the workflow.

This study proposes two methods that extend convolutional neural networks for image classification problems. The extended networks solve classification and OOD detection problems simultaneously. Thus, they reduce time and computational costs for the training and inference of neural networks. The first proposed method is a semi-supervised approach that combines Mahalanobis distance and a center loss [8]. The second method is a supervised OOD detection approach based

on multi-layer perception as the second head of the neural network. Both approaches were evaluated with the MNIST [9], FashionMNIST [10], and ImageNet30 [7] datasets using LeNet [11] and EfficientNet-B0 [12] neural networks.

II. Methods

Fig. 1 shows a basic convolutional network that solves a classification problem [13]. The neural network obtains a mini-batch of images $\{\mathbf{X}_i\}_{i=0}^{m}$ and produces a set of predicted classes $\{\hat{y}_i\}_{i=0}^{m}$. A set of true classes is $\{y_i\}_{i=0}^{m}$. The last hidden layer provides a set of deep feature vectors $\{\mathbf{x}_i\}_{i=0}^{m}$. Our approach to OOD detection extends such neural networks.

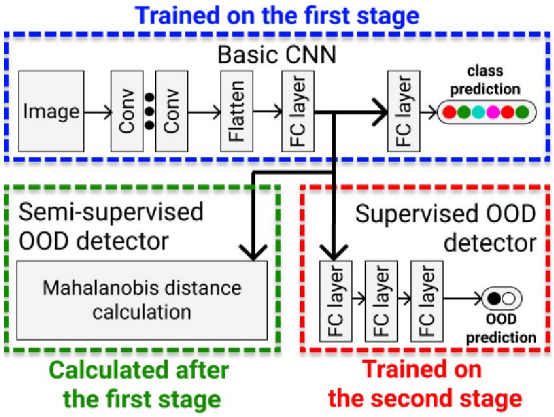

Fig. 1. The basic convolutional neural network (CNN) for image classification and two modifications for out-of-distribution (OOD) detection.

The semi-supervised OOD detection method complements cross-entropy loss with a center loss:

$$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_C = \tag{1}$$

$$= -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \mathbf{x}_i + b_j}} + \frac{\lambda}{2}\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{c}_{y_i}\|_2^2, \tag{2}$$

where L denotes a complex loss function; $\mathcal{L}_S$ denotes a softmax function with a cross entropy; $\mathcal{L}_C$ denotes the center loss function; λ is a balancing coefficient; n and m denote the class number and mini-batch size, respectively; $\mathbf{x}_i \in \mathbb{R}^d$ denotes the ith deep feature, beloning to the y_ith class; d is a feature dimension; $W_j \in \mathbb{R}^d$ denotes the jth column of the weights $W \in \mathbb{R}^{d\times n}$ in the last fully connected layer, and $\mathbf{b} \in \mathbb{R}^n$ is the bias term; and $\mathbf{c}_{y_i} \in \mathbb{R}^d$ denotes the y_ith class center of deep features.

Training of the neural network for the classification problem is performed with the main training dataset. After the training, all samples $\mathbf{X}_i,\ i \in [1, M],\ i \in \mathbb{N}$ are passed to the neural network, and this generates deep feature vectors $\mathbf{x}_i = T(\mathbf{X}_i)$. Each vector belongs to one given class $\{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$. Let us denote a subset of deep features that relates to the y_ith class as $\{\mathbf{x}\}^{(y_i)}$. Each subset of deep features determines a mean vector $\boldsymbol{\mu}^{(y_i)} = \mathrm{E}[\{\mathbf{x}\}^{(y_i)}]$ and covariance matrix $\mathbf{S}^{(y_i)}$. The obtained vectors determine the Mahalanobis distances for each class: $D_M^{(y_i)}(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu}^{(y_i)})^T(\mathbf{S}^{(y_i)})^{-1}(\mathbf{x}-\boldsymbol{\mu}^{(y_i)})}$. Distances of deep features obtained for y_ith determine $\boldsymbol{\theta}^{(y_i)}$ which is a threshold criterion. We set the threshold as the 97.5% percentile: $\boldsymbol{\theta}^{(y_i)} = \mathrm{percentile}(\{D_M^{(y_i)}(\mathbf{x})|\mathbf{x} \in \{\mathbf{x}\}^{(y_i)}\}, 0.975)$.

Training with center loss provides a deep feature space, where the vector sets $\mathbf{x}^{(y_i)}$ group around proper centroids. We assume that OOD samples $\mathbf{X}^{(-1)}$ are transformed to deep feature vector $\mathbf{x}^{(-1)}$ that are far from any centroids. The following criteria are used to differentiate OOD and normal samples:

$$P(\mathbf{X}) = Q(T(\mathbf{X})), \tag{3}$$

$$Q(\mathbf{x}) = \bigvee_{i=1}^{n} (D_M^{(y_i)}(\mathbf{x}) \le \boldsymbol{\theta}^{(y_i)}), \tag{4}$$

$$P, Q \in \{\mathrm{False}, \mathrm{True}\}, \tag{5}$$

where $P(\mathbf{X}) \equiv \mathrm{False}$ for OOD samples, and $P(\mathbf{X}) \equiv \mathrm{True}$ for normal samples.

The supervised OOD detection method also uses the center loss, but analyses of the deep feature vectors are performed with the multi-layer perceptron instead of the Mahalanobis distance criteria. The multi-layer perceptron includes three fully connected layers with 256, 256, and 1 neuron, respectively. Two layers use ReLU as the activation function, and the last layer uses the sigmoid function. The second head is trained with the binary cross-entropy loss function.

Training of the neural network is performed in two stages. The first stage involves training for the classification problem with the main training dataset. The backpropagation algorithm uses complex loss ($\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_C$) and propagates a prediction error from the main head to the general part of the neural network. The second stage consists of a training the second head for OOD prediction, using samples from the anomalous dataset as the zeros class and samples from the main dataset as the first class. This problem is equivalent to binary classification, where the zeroth class is OOD samples, and the first class is normal samples.

Analyses of the proposed method were performed using modified LeNet [11] and EfficientNet-B0 [12] neural networks. The modified LeNet has the same structure as the network from the original work [11], but we replace average pooling layers with max-pooling layers and replace hyperbolic tangent activation functions with rectifier linear unit functions [13]. These changes aims to increase the training and inference performance. The EfficientNet-B0 [12] was adopted from [12] without change.

This study uses datasets of the handwritten digests (MNIST), images of clothing (FashionMNIST), and images of real-world objects (ImageNet30, [7]). The MNIST and FashionMNIST consist of grayscale images with a size of 28x28 px. The ImageNet30 dataset consists of RGB images with a size of 256x256 px. Datasets of main data and anomaly data were obtained by splitting the MNIST dataset into MNIST-0 and MNIST-1..9 datasets according to the zero and other digits. MNIST and MNIST-1..9 were used as the main datasets, and the FashionMNIST and MNIST-0 as the anomaly datasets. We experimented with three different splits

978-1-7281-7691-8/21/$31.00 ©2021 IEEE

of ImageNet-30: ImageNet-30a consists of images from 0 to 9 classes, ImageNet-30b consists of images from 10 to 19, and ImageNet-30c consists of images from 10 to 19. In computational experiments, one of three datasets becomes the main dataset, and another one becomes the anomalous dataset.

III. Results

Fig. 2 shows F1-scores for the classification problem. The LeNet network on MNIST and MNIST-1..9 yielded F1-score ranging from 0.9852 to 0.9915. The EfficientNet-B0 on the ImageNet-30a, ImageNet-30b, and ImageNet-30c showed F1-scores ranging from 0.9897 to 0.9989. The performance of EfficientNet in the last case was better because this neural network includes more hidden layers. The value of the balancing coefficient λ significantly affected the classification accuracy and F1-score. Usage of the center loss improved the classification accuracy of the LeNet in all experiments. The center loss also improved the results for ImageNet-30 in one of the three experiments. The worst decrease in classification accuracy and F1-score caused by center loss did not exceed 0.0052 and 0.005295, respectively. Thus, the usage of the center loss is appropriate in real-life applications. According to the ranges of the F1-scores, the optimal value of the balancing coefficient λ was 0.1 or 1. Fig. 3 shows the anomaly detection

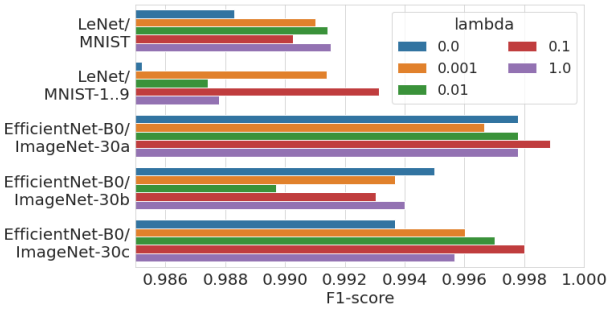

Fig. 2. Performance of neural networks on the classification problem for some values of the balancing coefficient λ.

performance of the proposed semi-supervised OOD detection method. As can be seen, the center loss improved the F1-score in all five experiments. The F1-score increased more than 0.03 points in three cases. According to the F1-score ranges, the optimal value of the balancing coefficient λ is 0.1 or 1. Analysis of the results for supervised classification provided

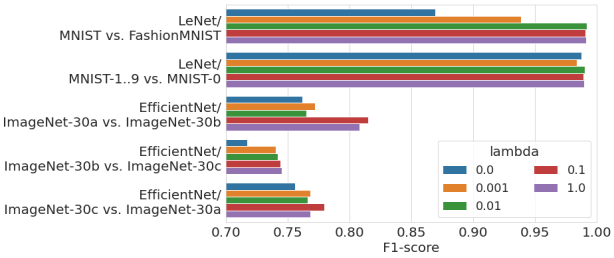

Fig. 3. Performance of neural networks in OOD sample detection using the semi-supervised approach for some values of the balancing coefficient λ.

opposite results. As shown in Fig. 4, the best F1-score for OOD detection was reached without the center loss ($\lambda = 0.0$). The OOD detector performance was decreased by 0.19 points in the worst case. This effect is more significant for datasets with real-world objects and for deeper neural networks.

OOD detection for the supervised method was better than for the unsupervised one. F1-scores for the semi-supervised approach were in range $[0.8695, 0.9913]$ for the LeNet and $[0.7172, 0.8149]$ for the EfficientNet. The supervised approach yielded F1-score in the ranges $[0.9881, 0.9990]$ and $[0.8083, 0.9727]$, respectively. Thus, we conclude that the supervised method is preferable in real applications. According

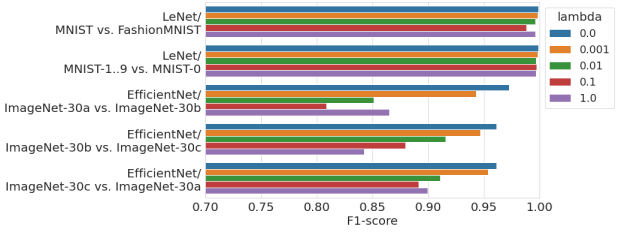

Fig. 4. Performance of neural networks in ODD sample detection using the supervised approach for some values of the balancing coefficient λ.

to the obtained results, the center loss is suitable for the semi-supervised method, but it does not benefit supervised OOD detection. This was be studied more precisely using ROC curve analysis. Fig. 5 shows the ROC curves for both methods under $\lambda = 0$ or $\lambda = 1$. The ROC curves for the semi-supervised method were built by simultaneous variation of all thresholds $\boldsymbol{\theta}^{(y_i)}$ in range $[0, 6]$. ROC curves for the supervised method were obtained using a variety of thresholds in the sigmoid function. The plots show that the increase of λ affected the two methods in opposite directions. High λ improved the semi-supervised approach, but the center loss decreased the area under the curve for the supervised one.

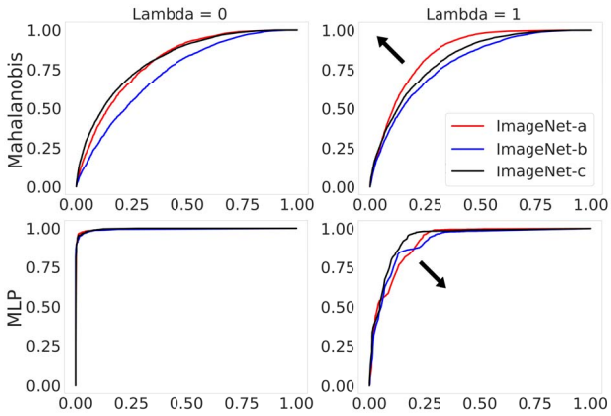

Fig. 5. Effect of the balancing coefficient λ on the ROC curves for the semi-supervised approach (Mahalanobis) and supervised approach (MLP). Black arrows show the direction of changes observed with increasing λ.

Above, we introduced criteria for OOD detection, assuming that normal samples are localized near the centroids. Fig. 6

978-1-7281-7691-8/21/$31.00 ©2021 IEEE

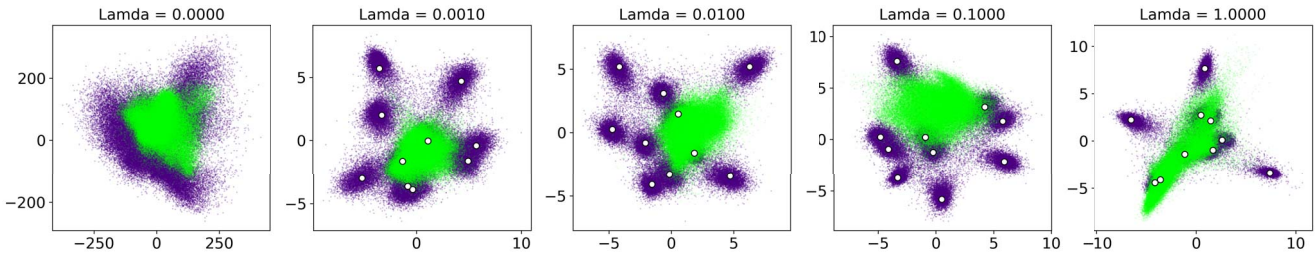

Fig. 6. Multidimensional deep feature vectors that are represented on a 2D plane. This projection was obtained using principal component analysis. Green dots are deep feature vectors transformed from OOD samples. Blue dots are deep feature vectors transformed from normal samples. White circles are centroid locations.

proves this assumption. Without centroids, OOD and normal samples form an unstructured distribution with a mean near to the zero vector (see Fig. 6 for $\lambda = 0$). Usage of the center loss modified the distributions of the deep feature vectors. OOD samples transformed to deep vectors $\mathbf{x}^{(-1)}$ surrounding the zero vector, but normal samples are distributed around centroids. The results of all experiments are presented in our GitHub repository [14].

IV. Discussion

In this study, we propose supervised and semi-supervised methods for anomaly detection in an image dataset (out-of-distribution samples detection). The proposed approaches extend a basic convolutional neural network that solves the problem of image classification. Thus, the modified network performs classification and OOD detection simultaneously. The semi-supervised approach is based on the use of the center loss to build the suitable distribution of deep vectors and use of Mahalanobis distance to analyze this distribution. The supervised approach is based on the second head that analyzes deep features.

The semi-supervised method is more agile because it does not require a dataset with anomalous samples for training. Usage of the center loss has an insignificant effect on image classification accuracy, but it strongly affects the deep feature distributions. Normal samples are grouped near the centroids, while out-of-distribution samples are grouped near the zero vector. Then, criterion (4) separates OOD and normal samples from image datasets. Increasing the balancing coefficient λ improves the accuracy and area under the ROC curve for OOD detection.

The supervised approach shows better performance than the semi-supervised one, making it favorable for real applications. However, this approach is strongly dependent on the choice of a proper anomalous dataset. The center loss should not be used with the supervised approach.

References

[1] S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney, and D. Song, "Anomalous instance detection in deep learning: A survey," *arXiv preprint arXiv:2003.06979*, 2020.

[2] A. Masaki, K. Nagumo, B. Lamsal, K. Oiwa, and A. Nozawa, "Anomaly detection in facial skin temperature using variational autoencoder," *Artificial Life and Robotics*, pp. 1–7, 2020.

[3] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini, "Anomaly detection with conditional variational autoencoders," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 1651–1657.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] A. Van Looveren, G. Vacanti, J. Klaise, and A. Coca, "Alibi-Detect: Algorithms for outlier and adversarial instance detection, concept drift and metrics." 2019.

[6] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *arXiv preprint arXiv:2007.08176*, 2020.

[7] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 663–15 674.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[9] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[10] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[11] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.

[12] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[13] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*, 2020, https://d2l.ai.

[14] G. Vareldzhan, K. Yurkov, and K. Ushenin, "Anomaly detection in image datasets," May 2021. [Online]. Available: https://github.com/GarnikOriginal/Anomaly-Detection-in-Image-Datasets