Out-of-distribution Detection in High-dimensional Data Using Mahalanobis Distance - Critical Analysis

Henryk Maciejewski[0000−0002−8405−9987], Tomasz Walkowiak[0000−0002−7749−4251], and Kamil Szyc[0000−0001−6723−271X]

Wroclaw University of Science and Technology
{henryk.maciejewski,tomasz.walkowiak,kamil.szyc}@pwr.edu.pl

Abstract. Convolutional neural networks used in real-world recognition must be able to detect inputs that are Out-of-Distribution (OoD) with respect to the known or training data. A popular, simple method is to detect OoD inputs using confidence scores based on the Mahalanobis distance from known data. However, this procedure involves estimating the multivariate normal (MVN) density of high dimensional data using the insufficient number of observations (e.g., the dimensionality of features at the last two layers in the ResNet-101 model are 2048 and 1024, with ca. 1000-5000 examples per class for density estimation). In this work, we analyze the instability of parametric estimates of MVN density in high dimensionality and analyze the impact of this on the performance of Mahalanobis distance-based OoD detection. We show that this effect makes Mahalanobis distance-based methods ineffective for near OoD data. We show that the minimum distance from known data beyond which outliers are detectable depends on the dimensionality and number of training samples and decreases with the growing size of the training dataset. We also analyzed the performance of modifications of the Mahalanobis distance method used to minimize density fitting errors, such as using a common covariance matrix for all classes or diagonal covariance matrices. On OoD benchmarks (on CIFAR-10, CIFAR-100, SVHN, and Noise datasets), using representations from the DenseNet or ResNet models, we show that none of these methods should be considered universally superior.

Keywords: Out-of-Distribution Detection · Mahalanobis distance · Convolutional Neural Networks

1 Introduction

Machine learning systems used in real-world recognition tasks need to classify inputs far from the known or training data as unrecognized or Out-of-Distribution (OoD). This is important in image or text recognition, where it is infeasible to train models for all categories encountered in open-world recognition. Recognition of OoD samples is vital in safety-critical applications or incremental-learning

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19

systems [13], [5], [19], [4]. However, popular models for image or text classification, e.g., ResNet, DenseNet, EfficientNet, are still vulnerable to OoD or adversarial examples that are easily recognized by humans [3], [9], [20], [14]. This is despite high classification accuracy realized on the benchmark datasets (e.g., top-1 accuracy on the ImageNet is ca. 90 [15]).

Many current approaches recognize OoD inputs using confidence scores obtained from class-conditional posterior distributions. A popular method, due to its simplicity, is to use multivariate Gaussian distributions as models of class-conditional distributions [1], [11], [18], [16]. This approach leads to estimating the uncertainty of prediction using Mahalanobis distance.

However, these procedures rely on the estimation of probability density in high-dimensional data. The dimensionality of the representations generated by CNNs used for image classification is usually ca 10^3. E.g., the dimensionality of features at the last layer of the ResNet-101[7] is 2048, and of the EfficientNet-B3 is 1536. Class conditional distributions are estimated from training data, typically based on an insufficient number of examples, e.g., using 5000 observations per class in the CIFAR-10 dataset. The purpose of this work is the analysis of the quality of such parametric density estimates in high-dimensional data and the impact of errors in density estimation on the performance of the Mahalanobis distance-based OoD detection.

Our contributions are the following.

- We analyzed the instability of estimated densities in high-dimensional data. We showed, using simulated data, that the generative MVN models fitted to the training data are far from the testing samples from the same distribution. Hence, OoD detection based on this model will tend to reject testing samples as outliers. We analyzed this effect as a function of dimensionality and training sample size.
- We analyzed the limitations of Mahalanobis distance-based OoD detection: we showed that due to the model estimation error, near OoD samples are not distinguishable from known data. The minimum distance from known data beyond which outliers are detectable depends on the dimensionality of the features and the training sample size and decreases for larger training samples.
- We analyzed simple modifications of the method used to reduce the impact of model fitting errors: Mahalanobis distance using one covariance matrix shared by all classes of known data, or using diagonal covariance matrices. We illustrate the performance of these methods on OoD benchmarks, with the CIFAR-10 as in-distribution and the CIFAR-100, the SVHN, and the Noise datasets as OoD, and with features generated by different CNN models. We showed that none of these Mahalanobis distance-based methods should be declared universally best, as the performance depends on the characteristics of benchmark datasets. On some benchmarks, Mahalanobis distance-based OoD detectors are outperformed by simple methods, which use the Euclidean or standardized Euclidean distance.

2 OoD Detection with Mahalanobis Distance - Simulation Study

2.1 Method - Using Mahalanobis Distance for OoD Detection

Using Mahalanobis distance as the score for OoD detection relies on the estimation of multivariate Gaussian (MVN) distribution as a model of class-conditional posterior distribution. Here we briefly summarize the method. Given the known (in-distribution) dataset $X_c \subset R^d$ for the class $c \in C = \{1, 2, \ldots, m\}$, with N_c examples, we estimate the model $\mathcal{N}(\mu_c, \Sigma_c)$ with the mean vector $\mu_c = \frac{1}{N_c} \sum_{x \in X_c} x$ and the covariance matrix $\Sigma_c = \frac{1}{N_c} \sum_{x \in X_c} (x - \mu_c)(x - \mu_c)^\top$.

Given a test sample u, Mahalanobis distance to the MVN model of class c is computed as

$$d_{Mah,c}(u) = \sqrt{(u - \mu_c)^\top \Sigma_c^{-1} (u - \mu_c)}. \tag{1}$$

The confidence score used to label the sample u as in-distribution or OoD is calculated as $s(u) = -\min_{c \in C} d_{Mah,c}(u)$.

To minimize errors due to unreliable estimation of Σ_c in high dimensional data, some works (e.g., [11], [16]) assume that all m classes share the common covariance matrix, estimated from the larger sample of size $N = \sum_c N_c$ as $\Sigma = \frac{1}{N} \sum_{c \in C} \sum_{x \in X_c} (x - \mu_c)(x - \mu_c)^\top$.

The Mahalanobis distance of a test sample u to class c is then computed as

$$d_{MahUF,c}(u) = \sqrt{(u - \mu_c)^\top \Sigma^{-1} (u - \mu_c)}. \tag{2}$$

Other modifications / simplifications of this procedure assume that the covariance matrix $\Sigma_c = V_c$ is the diagonal matrix with diagonal components calculated as variances of features computed over samples in X_c. Then the distance of a test sample u to the MVN model of class c is calculated as the standardized Euclidean distance:

$$d_{SEuc,c}(u) = \sqrt{(u - \mu_c)^\top V_c^{-1} (u - \mu_c)}. \tag{3}$$

Finally, the distance of a sample u to the model of class c can be computed as the Euclidean distance $d_{Euc,c}(u) = |u - \mu_c|_2$, (which implies that all variances in V_c in Equation 3 are equal 1).

2.2 Non-robust Estimation of MVN Model in High Dimensional Data

We performed a simulation study in which we analyzed the instability of the MVN model of in-distribution data as a function of dimensionality and sample size. We generated n training and n testing observations from the MVN distribution in d dimensions, with the mean at $[0]_d$ and with uncorrelated variables with variance 1. We estimated the MVN model from the training sample and compared the distances of the training and testing samples from the model.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19

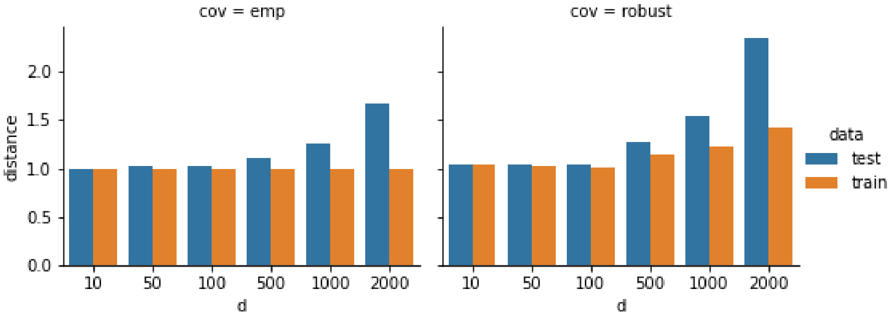

Fig. 1: Mean Mahalanobis distances of train and test samples to the MVN model fitted to the train samples. Train and test samples drawn from the same MVN. Number of samples $n = 5000$. Note: distance shown here is the squared Mahalanobis distance divided by d.

Results as a function of dimensionality d, for fixed sample size $n = 5000$ are shown in Figure 1. We used the scikit learn library MLE estimator of covariance (referred to as *empirical*), and the Minimum Covariance Determinant estimator (MCD) [17], referred to as *robust* due to its resistance to outliers. We observe that when the dimensionality of data grows, the test data tend to lie significantly further off the model than the train data. Since this effect is more prominent with the robust estimator, we conclude that the robust estimator is not appropriate for high-dimensional data. Note that this observation holds even if $n > 5d$, a condition deemed to guarantee a low error of the MCD estimator.

In Figure 2 and 3 we analyze the effect of the growing sample size. This analysis can be used to determine, for a given dimensionality of features d, the required size of training data to guarantee a robust model of known data.

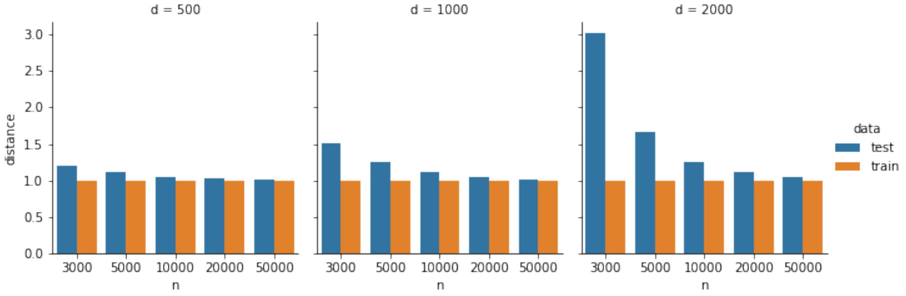

Fig. 2: Mean Mahalanobis distances of train and test samples to the MVN model fitted to the train samples, as a function of the size of samples n and dimensionality of data d. Note: distance is the squared Mahalanobis distance divided by d.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19

For instance, considering the case $d = 2000$, $n = 5000$ (Figure 3, left panel), which corresponds to the dimensionality of the representations from the ResNet-101 model, and the size of the CIFAR-10 train data, we conclude, that Mahalanobis-based OoD detection with per-class covariance matrices (Equation 1) will fail to recognize OoD samples as different from known data unless sufficiently far from the in-distribution data ($d_{Mah,c} > 62$). Increasing the sample size (Figure 3, right panel) allows to recognize nearer OoD samples (with $d_{Mah,c} > 50$). We further analyze this effect in Sections 2.3 and 3.4.

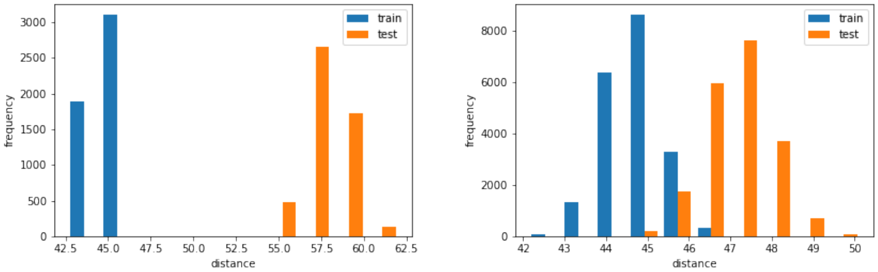

Fig. 3: Distribution of Mahalanobis distance of train and test samples to the MVN model fitted to the train data, dimensionality $d = 2000$, sample size $n = 5000$ (left), $n = 20000$ (right)

.

2.3 Non-robust MVN Model Used for OoD Detection

In the second experiment, we compare the Mahalanobis distance of in-distribution test data and OoD data to the MVN model fitted to the train data. We model known data as in Section 2.2, and OoD data as MVN with mean μ shifted from $[0]_d$ by r, ie. $|\mu - [0]_d| = r$, and with uncorrelated variables with variance 1. We realized three schemes of OoD data, denoted *ood1*: shift by r along only one axis; *ood3*: $\mu = [\frac{r}{\sqrt{d}}]_d$, ie. shift along all the axes; *ood2*: shift along $\frac{d}{2}$ axes. (As we later show, the scheme effect is visible if known and OoD data differ in terms of correlation structure).

Results as a function of the sample size n and OoD shift r are summarized in Figures 4 and 5. In the left panel of Figure 4 (with n and d corresponding the CIFAR-10 training data and the ResNet-101 features), we observe that Mahalanobis distance is unable to distinguish in-distribution test and OoD data. To quantify the dissimilarity between groups shown in Figure 4, we use the measure

$$\mathit{\Delta}(group1, group2) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2}}, \tag{4}$$

where $\bar{X}_i$ and s_i are the mean and its standard error in group i. Note that this measure is used as the test statistic in Welch's t-test. We observe that with growing n, $\Delta(test, train)$ decreases, and $\Delta(ood, test)$ increases. Hence, with increasing sample size, the model stabilizes and leads to better separation between in- and OoD data.

In Figure 5 we analyze the effect of shift r on the separability of in- and OoD data. We observe that for sufficiently far OoD data (e.g., $r = 32$), the inaccuracy in MVN model estimation no longer matters: OoD samples are significantly more distant from the model than the test in-distribution samples. We argue that this effect accounts for the success or failure of the Mahalanobis distance-based OoD detection in CNN benchmarks, as further analyzed in Section 3.4.

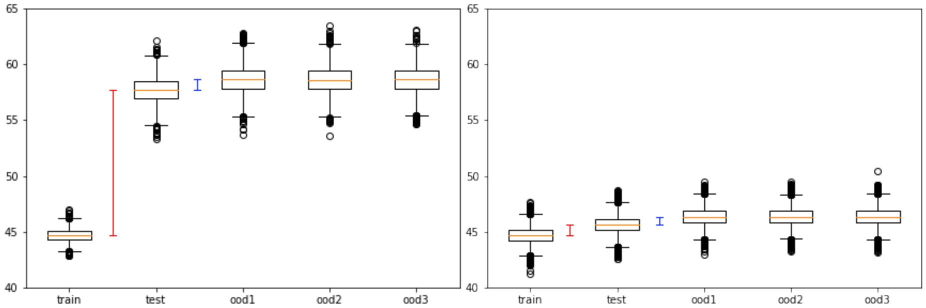

Fig. 4: Distribution of Mahalanobis distance of train, test and OoD samples to the MVN model fitted to the train data, dimensionality $d = 2000$, OoD shift $r = 8$, sample size $n = 5000$ (left), $n = 50000$ (right). Dissimilarity between distances: $\Delta(test, train)$= 708 (left), 204 (right); $\Delta(ood1, test)$= 39 (left), 154 (right). Large training samples lead to more robust models of in-distribution data (difference between train and test data decreases), and better separability of in- and OoD data (difference between test and ood increases).

Finally, in Figure 6, we signal the effect of feature correlation on OoD performance. We assume correlated in-distribution and uncorrelated OoD data. We observe that if the correlation schemes of in-distribution and OoD data differ, the Mahalanobis distance-based separability of OoD and known data improves, and distances to the model of $ood1, 2, 3$ schemes become significantly different, hence in this case, the distance from the model depends on the direction of OoD shift.

2.4 Mahalanobis Distance-based vs. Nonparametric Outlierness Factor-based OoD in High Dimensional Data

Since the estimation of density in high-dimensional data is generally considered unattractive [6], we want to empirically show that the confidence scores obtained from density estimates lead to the limited performance of OoD detection. On the

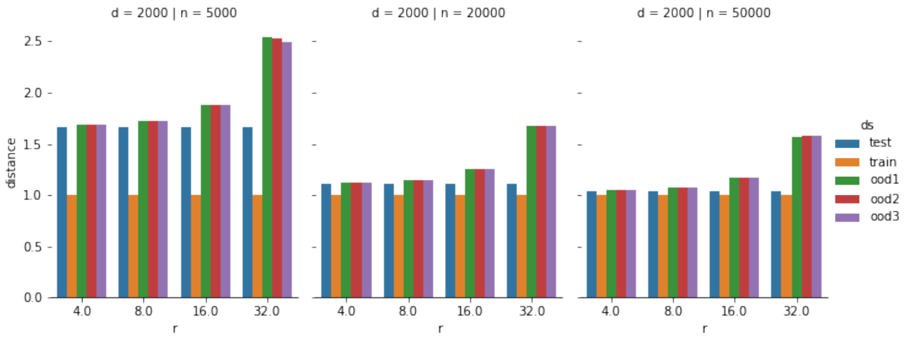

Fig. 5: Comparing Mahalanobis distances of train, test and OoD samples, as a function of sample size n, OoD shift r, for dimensionality $d = 2000$. Note: distance is the squared Mahalanobis distance divided by d.

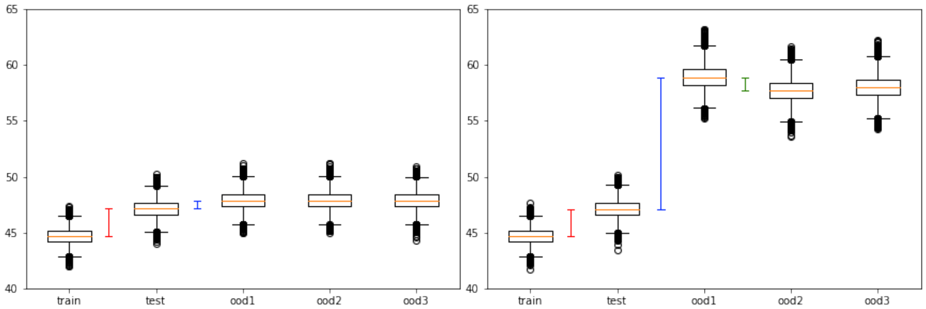

Fig. 6: Effect of correlated features: distribution of Mahalanobis distance of train, test and OoD samples to the MVN model, for dimensionality $d = 2000$, OoD shift $r = 8$, sample size $n = 20000$, uncorrelated features (left); 1000 features correlated with coefficient 0.5. (right). Dissimilarity between distances: $\Delta(test, train)$= 330 (left and right); $\Delta(ood1, test)$= 95 (left), 1278 (right). If in- and OoD data differ in correlation structure, separability of OoD and in-distribution data improves. Note that with correlated data, distance of different outlier groups $ood1, 2, 3$ to the model is significantly different, e.g., $\Delta(ood1, ood2)$= 122 (right), whereas in previous examples (left panel and Figure 5) differences between ood schemes were not significant.

other hand, outlier or out-of-distribution detection in high-dimensional data can be performed reasonably well using scores obtained from the Local Outlierness Factor (LOF) method [2] (see Section 3.3 for technical details of LOF).

We performed a simple simulation study in which we compared the performance of OoD detection (see Section 2.5 for technical details of the used metrics) based on confidence scores obtained using the Mahalanobis distance vs confidence scores obtained with the LOF algorithm. As in-distribution (known) data, we generated two clusters from the MVN distribution in d dimensions, with the mean at $[0]_d$ and $[-1]_d$ and uncorrelated variables with variance 1. As

OoD, we used a cluster with mean at $[\frac{r}{\sqrt{d}}]_d$, uncorrelated, with variance 1. Confidence scores were calculated as the $MahUF$ distance (see Equation 2) between a test sample and the closest class conditional Gaussian distribution, which can be interpreted as the log of the probability density of the test sample. In the alternative approach, confidence scores were obtained as local outlierness factors (LOF) calculated for test samples with respect to the closest cluster of known data.

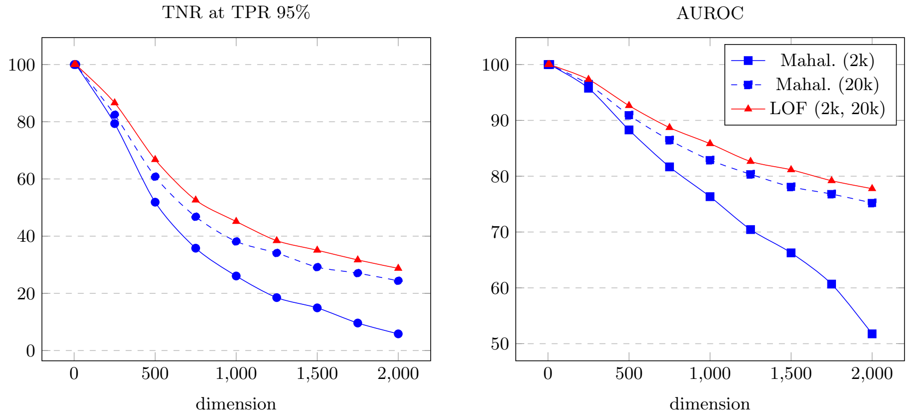

Fig. 7: Comparison of the averaged TNR at TPR 95 % and the AUROC for $MahUF$ and LOF on simulation data. The inlier set consists of two classes, simulated by MVN with variance 1 and distance 1. The outlier is also simulated by MVN moved from the closest inlier class by a given distance (r=8). The number of training (inlier) examples was set to 2k (solid) and 20k (dashed line), with 2k for test inliers and outliers. The results for 2k and 20k for LOF are undistinguished. By increasing the complexity of the problem, expanding the input data dimension, the LOF method is much more stable and achieves better results. The dimension above 2000 is common in the last layers of CNNs. Moreover, LOF is less prone to the changes of the training set size.

We observe that for $d = 2000$ with 1000 training points per class (there are two classes), the Mahalanobis procedure no longer detects outliers (AUROC $\approx$ 50%), while LOF is more reliable (AUROC $>$ 80%). Where for 10k training points per class, the Mahalanobis procedure gives results closer to the LOF ones. It shows how the number of examples is important for the Mahalanobis approach.

2.5 Evaluation Metrics

In the evaluation of OoD performance, we follow the approach used in [8] where the outlier detection is considered a binary classification. The outliers are defined as the positive class and the closed set examples (test set) as the negative class. The confidence score allows the binary classification. In the result presentation,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19

we used the standard metrics: TNR at TPR 95%, AUROC, DTACC, and AUPR – the higher the values of all metrics, the better the OoD detection is. The True Negative Rate at 95% True Positive Rate (TNR at TPR 95%) can be interpreted as the probability of correctly classifying the Out-of-Distribution examples when the In-Distribution (test) samples are classified as high as 95%. The Area Under Receiver Operating Characteristic curve (AUROC) defines the OoD method's ability to discriminate between cases (test examples) and non-cases (OoD examples). It can be calculated by the area under the false positive rate against the true positive rate curve. The detection accuracy (DTACC) defines the ratio of correct classification of the test and OoD examples to all examples. The AUPR is calculated by the Area Under the Precision and Recall curve, where test (AUPR In) or OoD (AUPR Out) images are specified as positive. We denote AUPR as the mean of both due to equal numbers of examples in both sets.

3 Using Mahalanobis Distance for OoD Detection in CNNs

In this section, we illustrate the efficiency of the Mahalanobis-based method for OoD in CNN models and show the characteristics of the representations generated by CNNs which make it feasible to use the Mahalanobis method.

3.1 Datasets and CNN Models

We used popular benchmark datasets successfully used in OoD detection in computer vision: the CIFAR-10, the CIFAR-100, the SVHN, and the Noise. The CIFAR-10[1] dataset contains $60,000$ 32x32 color images divided into 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. There are $5,000$ images per class in the training set and $1,000$ in the test set. CIFAR-100 is similar – there are 100 classes (disjoint from the CIFAR-10 classes) with 500 and 100 images per class, respectively, train and test subsets. SVHN contains real-world images of Street View House Numbers from Google Street View - they are easily distinguishable for humans compared to CIFARs ones. The Noise dataset consists of randomly generated images - a theoretically straightforward recognition task. To evaluate OoD methods, we used the testing partitions of the in-distribution datasets and the given Out-of-Distribution dataset, with a 1:1 proportion of known and unknown samples.

We trained two models, ResNet-101 [7] trained on CIFAR-10 (achieving 94.75 % accuracy) and DenseNet-169 [10] trained on CIFAR-100 (achieving 74.04 % accuracy). These model architectures were chosen due to their high popularity in commercial applications and OoD detection problems. We used the classic method of feature extraction from deep models, which uses vectors after applying the Global Average Pooling [12] on the last convolutional layer. Dimensionality of feature vectors is $2,048$ (for ResNet-101), and $1,664$ (for DenseNet-169). The

[1] https://www.cs.toronto.edu/~kriz/cifar.html

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19

procedure in our experiments is as follows: (1) train the model using training subset, (2) extract features from the model for images in the training subset, (3) fit the in-distribution model for OoD detection, (4) evaluate confidence scores (or distances) for in- and OoD test data.

3.2 Analysis of Mahalonobis Distances for the CIFAR-10

First, we analyze the characteristics of features generated by ResNet-101 for the CIFAR-10 train and test data and OoD data. We estimated the distribution of distances from the class centers (calculated on the train data sets) to different groups of data. Results for one of the CIFAR-10 classes are shown in Figure 8. We can notice the large difference between train and test data in the case of Mah distance. It is the same phenomenon as discussed in Section 2.5, i.e., insufficient number of data. Moreover, one can see the relative shift of the noise set (black curves) position with respect to the SVHN and the CIFAR-100 (green and red) for distances with the full covariance matrix (Mah and $MahUF$) and with limited ($SEuc$) or non-existing (Euc) one. This suggests that the type of distance (still from the same family) may have a big influence on the OoD detection.

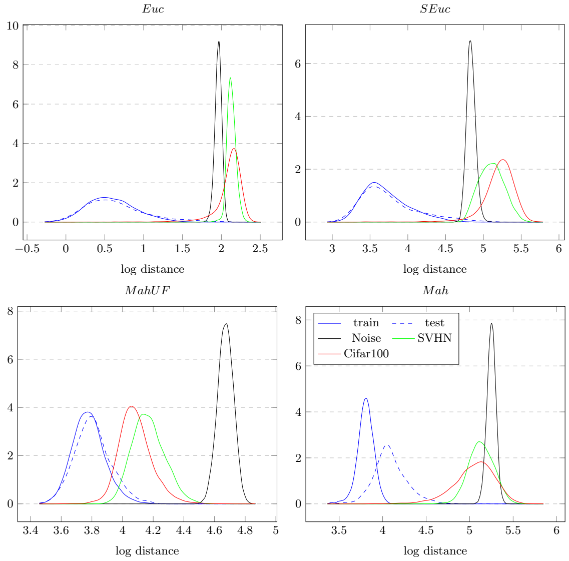

Fig. 8: Log distance distribution for different distance metrics (Euc, $SEuc$, $MahUF$, and Mah). Distances are from the center of one of the CIFAR-10 labels to train and test data for the same label, and also to outlier data sets (Noise, SVHN, and CIFAR-100). Plots represent the probability density function of log distances obtained by the Gaussian kernel-density estimator with Scott's bandwidths.

Table 1: Difference between train and test data distances to the model of one the CIFAR-10 class, and the difference averaged over all the CIFAR-10 classes. Differences between train and test distances are measured using Welch's t-test statistic - Eq. 4.

	CIFAR-10 (ResNet-101)		CIFAR-100 (DenseNet-169)	
distance	class 1	avg over all classes	class 1	avg over all classes
Euc	5.11	8.86	6.02	5.66
$SEuc$	**4.95**	**8.72**	**5.77**	5.17
$MahU$F	6.63	8.81	6.41	**4.5**
Mah	39.36	40.12.	14.41	15.03

Table 1 shows the Welch's t-test statistic comparing the distances of the train and test data for a selected CIFAR-10 class (the same class as used in Figure 8), and averaged over all classes for each of the analyzed Mahalanobis distances. It can be seen that Mah gives the largest values of statistic t suggesting that the train and test population means are the most distant (the same conclusion as from Figure 8). The train and test populations are closest for $SEucl$ but the differences to Euc and $SEuc$ are small.

3.3 LOF-based OoD

The Local Outlier Factor [2] (LOF) is based on an analysis of the local density of points. It works by calculating the so-called local reachability density $LRD_k(x, X)$ of input x with regard to the known dataset X. LRD is defined as an inverse of an average reachability distance between a given point, its k-neighbors, and their neighbors (for details refer to [2]). K-neighbors ($N_k(x, X)$) includes a set of points that lie in the circle of radius k-distance, where k-distance is the distance between the point, and it's the farthest k^{th} nearest neighbor ($||N_k(x, X)|| >= k$). The local outlier factor (LOF) is formally defined as the ratio of the average LRD of the k-neighbors of the point x to the LRD of the point.

$$d_{LOF}(u) = \frac{\sum_{x \in N_k(u,X)} LRD_k(x, X)}{||N_k(u, X)|| LRD_k(u, X)} \tag{5}$$

Intuitively, if the point is an inlier, the ratio of the average LRD of neighbors is similar to the LRD of the point. Therefore, the LOF is around 1. For outliers, it should be above 1 since the density of an outlier is smaller than its neighbor density.

3.4 OoD Experiments

In Table 2, we demonstrate the performance of the Mahalanobis based OoD detection using popular CNN architectures: ResNet-101 (trained on CIFAR-10)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19

Table 2: The comparison of analysed OoD methods for CIFAR-10 and CIFAR-100. Note that there is no best or worst method.

In-dist (Model)	OOD	Method	TNR at TPR 95%	AUROC	DTACC	AUPR
CIFAR-10 (ResNet-101)	Noise	Euc	97.12	98.81	96.21	98.56
		$SEuc$	45.92	94.57	94.71	91.24
		$MahUF$	100.00	100.00	100.00	100.00
		Mah	100.00	100.00	100.00	100.00
		LOF	100.00	99.90	99.30	99.89
	SVHN	Euc	55.03	93.00	86.75	92.52
		$SEuc$	13.55	86.59	82.57	83.85
		$MahUF$	53.84	91.13	83.26	90.96
		Mah	41.34	89.33	82.25	88.77
		LOF	57.80	93.00	86.43	92.73
	CIFAR-100	Euc	41.64	87.38	80.69	86.49
		$SEuc$	35.75	86.30	80.19	84.84
		$MahUF$	24.07	78.18	71.38	77.40
		Mah	37.59	85.91	78.55	85.16
		LOF	47.84	87.22	79.49	86.74
CIFAR-100 (DenseNet-169)	Noise	Euc	99.98	98.99	98.02	98.32
		$SEuc$	100.00	100.00	100.00	100.00
		$MahUF$	100.00	100.00	100.00	100.00
		Mah	100.00	100.00	100.00	100.00
		LOF	81.05	95.91	95.81	92.99
	SVHN	Euc	12.24	75.21	70.65	73.95
		$SEuc$	37.30	85.85	78.41	85.07
		$MahUF$	29.11	81.82	74.30	80.93
		Mah	19.37	81.31	75.73	79.74
		LOF	24.48	83.46	76.86	81.96
	CIFAR-10	Euc	15.48	73.29	68.39	71.19
		$SEuc$	9.30	69.77	66.40	68.07
		$MahUF$	8.05	65.65	62.22	63.97
		Mah	10.08	69.42	65.98	68.33
		LOF	13.51	73.36	68.28	71.71

and DenseNet-169 (trained on the CIFAR-100). We used three outlier data sets: Noise, SVHN, and CIFAR-100 (for ResNet-101) or CIFAR-10 (for DenseNet-169). We evalauetd of the four versions of Mahalonobis distances (Euc,$SEuc$, $MahUF$, Mah) presented in Section 2.1. LOF is shown as an alternative, non-parametric approach.

The Noise dataset is very well detected as OoD. However, there are some problems (TNR worse then 50%) in case of $SEuc$ for ResNet-101. The SVHN and CIFARs datasets are harder to be detected as OoD for DenseNet-169 than in the case of ResNet-101.

Comparison of OoD methods gives no straightforward conclusions. There is no best or worst method. We can find a failure scenario (when a method is much worse than the best one) for each of the analyzed methods and a situation when a given method significantly outperforms others. For example, Euc fails for DenseNet-169 and SVHN, and outperforms others for the CIFAR-10 and the same model, $SEuc$ fails for ResNet-101 and Noise, and outperforms for SVHN and DenseNet-169, $MahUF$ fails in CIFAR-10/DenseNet-169, and Mah in SVHN/DenseNet-169. Our results suggest that we should carefully state that the given method is the best since the results (OoD metric) strongly depend on data (CNN features), so not only on image data sets but also network architecture and the process of model training.

4 Conclusion

In this paper, we analyzed the performance of the Mahalanobis distance-based OoD detection method in high-dimensional data. This method is popular due to its simplicity, but it relies on parametric density estimates in high-dimensional data. We analyzed the instability of MVN estimates of density and showed that this issue leads to the intrinsic limitation of this method: near OoD samples are not distinguishable from known data. For fixed dimensionality of features and the size of training data, we can estimate the minimum distance from known data beyond which outliers are detectable. We showed that this distance decreases with the growing number of training samples.

We also analyzed common modifications of the method used to mitigate the density estimation errors: Mahalanobis distance with single covariance matrix shared by all classes in known data, or standardized Euclidean distance with diagonal covariance matrices. We compared the performance of these methods using OoD benchmarks with CIFAR-10 as in-distribution vs. CIFAR-100, SVHN, Noise as OoD, and CIFAR-100 vs. CIFAR-10, SVHN, and Noise datasets. We showed that none of these methods should be seen as universally superior, as the performance of OoD detectors depends on the benchmark dataset and the CNN model used to generate representations.

References

1. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1893–1902 (2015)
2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: Identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (May 2000). https://doi.org/10.1145/335191.335388
3. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology **6**(1), 25–45 (2021), https://doi.org/10.1049/cit2.12028
4. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual

classification. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1625–1634. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00175

5. Feng, D., Rosenbaum, L., Dietmayer, K.: Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). pp. 3266–3273. IEEE (2018)
6. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media (2009)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. Proceedings of the International Conference on Learning Representations (2019)
9. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. arXiv preprint arXiv:1907.07174 (2019)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 7167–7177. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
12. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
13. McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., Weller, A.: Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. p. 4745–4753. IJCAI'17, AAAI Press (2017)
14. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
15. Pham, H., Dai, Z., Xie, Q., Luong, M.T., Le, Q.V.: Meta pseudo labels. In: IEEE Conference on Computer Vision and Pattern Recognition (2021), https://arxiv.org/abs/2003.10580
16. Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A simple fix to mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022 (2021)
17. Rousseeuw, P.J.: Least median of squares regression. Journal of the American statistical association **79**(388), 871–880 (1984)
18. Sehwag, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=v5gjXpmR8J
19. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 acm sigsac conference on computer and communications security. pp. 1528–1540 (2016)
20. Zhou, Z., Firestone, C.: Humans can decipher adversarial images. Nature communications **10**(1), 1–9 (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_19