Adaptive Out-of-Distribution Detection with Human-in-the-Loop

Heguang Lin [*1] **Harit Vishwakarma** [*1] **Ramya Korlakai Vinayak** [1]

Abstract

Robustness to Out-of-Distribution (OOD) samples is essential for successful deployment of machine learning models in the open world. Many existing approaches focus on offline setting and maintaining a true positive rate (TPR) of 95% which is usually achieved by using an uncertainty score with a threshold based on the In-Distribution (ID) data available for training the models. In contrast, practical systems have to deal with OOD samples on the fly (online setting) and many critical applications, e.g., medical diagnosis, demand the system to meet quality constraints in terms of controlling FPR (false positive rate) at most 5%. This is challenging since having adequate access to the variety of OOD data, the system encounters after deployment is hard. To meet this challenge, we propose a human-in-the-loop system for OOD detection that can adapt to variations in the OOD data while adhering to the quality constraints. Our system is based on active learning approaches and is complementary to the current OOD-detection methods. We evaluate our system empirically on a mixture of benchmark OOD datasets in image classification task on CIFAR-10 and show that our method can maintain FPR at most 5% while maximizing TPR and making a limited number of human queries.

1. Introduction

Deploying machine learning (ML) models in the open world makes it subject to out-of-distributions (OOD) inputs — points dissimilar to the training data points e.g. in the classification setup one can think of OOD data points as those that do not belong to any of the classes in the training data. The modern ML models, in particular deep neural networks, can fail silently with high confidence on OOD points (NYC15; AOS+16) rather than flagging them as OOD and asking for human intervention. Such failures can have serious consequences in high risk applications e.g. medical diagnosis, autonomous driving etc. For a successful deployment of an ML model in the open world, we need mechanisms that ensure robustness to the OOD inputs.

A few recent works have addressed this problem (LLS17; LLLS18; LWOL20; MSDL22). Broadly, these works propose methods to quantify a *score* that can be used to decide OOD vs ID label for a given point. Many of these methods are based on distance between data points or a model's confidence score in prediction. For a detailed survey of literature in the area of generalized OOD detection, see (YZLL21). However, many of these works are largely limited to static in-distribution (ID) and OOD data (LLS17; LWOL20; MSDL22). However, in practice even if the ID data remains the same, the OOD data can vary, making it very difficult to foresee and prepare for all possible OOD data.

Moreover, in many applications the consequences of classifying an OOD point as ID (false positive) could be worse than classifying an ID point as OOD (false negative), e.g. in medical diagnosis it is better to classify a chest scan as OOD and defer the decision to humans rather than classifying it as ID and giving a disease label when the uncertainty is high. While ID data is usually available in plenty as a part of training set, having adequate access to OOD data that one can encounter during deployment is difficult. Most of the recent literature in OOD detection have focused on guaranteeing a certain true positive rate (TPR) e.g., TPR at 95% and set the threshold accordingly using the ID data. In applications such as medical diagnosis, it is crucial to guarantee that the false positive rate (FPR) is below certain acceptable rate, e.g., FPR below 5%. Since the availability of exact type of OOD data that the system can encounter during deployment is rare, it is crucial to design systems that can adapt to the OOD data while controlling the FPR during deployment.

To address this challenge, we propose an adaptive OOD detection system with human-in-the-loop (Figure 1) which can get human feedback and adapt itself after being deployed in the open world. The proposed system leverages existing methods for computing the uncertainty score (LLLS18; LWOL20; SCM21) and adaptively decides on the score

[*]Equal contribution [1]University of Wisconsin-Madison, WI, USA. Correspondence to: Heguang Lin <hlin324@wisc.edu>, Harit Vishwakarma <hvishwakarma@cs.wisc.edu>.

ICML 2022 Workshop on Human-Machine Collaboration and Teaming, Baltimore, Maryland, USA, 2022.

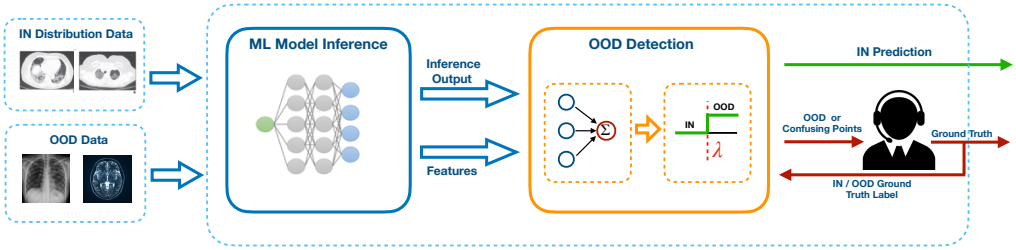

Figure 1. Illustration of Adaptive OOD System with Human-in-the-Loop

threshold for OOD classification. To minimize the number of human queries, we use a simple active querying mechanism – in which label for a point is queried only if the model is uncertain about its label. We evaluate our proposed system empirically on CIFAR-10 (ID) and mixture of 5 benchmark OOD datasets namely, SVHN (NWC+11), LSUN-Crop (YSZ+15), LSUN-Resize (YSZ+15), ImageNet-Resize (DDS+09), and iSUN (XEZ+15), with a variety of scoring functions (Mahalanobis distance (LLLS18), Energy score(LWOL20) and SSD score (SCM21) and show that our system can work well in the online setting.

Problem Setup: Consider a training dataset sampled from some distribution which we refer to as ID-dataset[1] and IN-distribution respectively. Consider a ML model (parameterized with w) trained for classification task on the ID-dataset. We assume access to scoring functions g_w such that the score $g_w(x)$ can be used to decide whether the given data point (x) is OOD or ID. Examples of g_w include Mahalanobis distance and energy based scores (LLLS18; LWOL20).

A threshold λ_0 on the *score* is usually calibrated using some fixed ID and OOD data to decide whether a point is OOD or not during deployment. This λ_0 may not work well due to variations in the OOD data and have to adapted as we see more data. To enable this, we assume that there is a human-in-the loop in the OOD detection system that can be used to inspect a point that it is uncertain on and confirm whether it is indeed OOD or is an ID.

We consider ID as positive class and OOD as negative class and we focus on applications that are more sensitive to false positives. Therefore, in the beginning, when the system has a lot of uncertainty, it is better off deferring the decision to humans. However, getting human labels, e.g., an expert radiologist to take a look at the image, is expensive. So, ideally over time, we want to achieve the objective of improving the system for identifying OOD points (which can get better with more labels) while minimizing the number of labels needed from the human experts.

Our research problem can be summed up as follows: *Given an OOD detection system bootstraped with some initial ID and OOD data, augment the system with a human-in-the-loop system so that it can adapt itself to better match the OOD data (i.e. have low false positive rate and high true positive rate) with minimal human supervision.*

[1]where 'ID' stands for "In-Distribution"

2. Methodology

Accurately detecting OOD points in the online setting needs a good scoring function g_w that separates the ID and OOD points at some threshold score λ, and a method to adjust the threshold score λ based on the human inputs. We leverage existing works on construction of g_w and propose simple methods based on online and active learning to adapt the system with minimal labeling queries while maintaining the performance. The two main components are:

1. **Computing OOD Score**: We use the following methods to compute the OOD score for a given data point,
   (a) **Mahalanobis Distance:** For a given point x, the Mahalanobis Distance (MD) based score is its MD from the closest class conditional distribution. We use MD based score as given in (LLLS18) for detecting OOD and adversarial samples. They compute the scores using representations from various layers of DNNs and combine them to get a better scoring function.
   (b) **Energy Score:** This score was proposed in (LWOL20) and it is well aligned with the probability density of the samples, with low energy implying ID and high energy implying OOD.
   (c) **SSD:** It is based on computing the Mahalanobis distance in the feature space of the model trained on the unlabeled in-distribution data using self-supervised learning. More details can be found in (SCM21).
   After computing the score the task is to make a prediction and adapt the system in case a mistake is made.
2. **Online Learning**: We assume that we have access to a small fraction of ID and OOD data to begin with. Using this we calculate an initial threshold (λ_0) that meets the given criterion e.g., FPR at most 5%. Then we simulate an online learning setting in which the samples (including both ID and OOD samples) arrive one at a time. When we see a new sample $\boldsymbol{x}_i$, we first compute the score t_i for this sample and then assign label $z(x_i) = \text{ID}$ if $t_i \geq \lambda_i$ and OOD otherwise. The system can get human feedback by querying for true labels

in various ways to adapt itself. We compare following methods for this:

(a) **Fixed-Threshold (No-Feedback)**: Keep a fixed threshold λ_0. This method serves as a baseline.

(b) **Always-Querying**: True label y_i is queried for each sample and we adapt the threshold as in (1).

(c) **Active-Querying**: Query labels only for confusing points – the points whose score t_i is close to current threshold λ_i. Formally, we query the label if $|t_i - \lambda_i| \leq \epsilon$, where ϵ can be interpreted as a margin parameter. Otherwise we do not query the label. The threshold is adapted if a mistake is made according to (1).

(d) **Random-Querying**: Randomly query labels for a fixed fraction of testing samples, and adapt the threshold if a mistake is made according to (1).

Adaptation Rule and Querying Strategies: Upon receiving the true label (y_i), if there was a mistake, i.e., ($z(\boldsymbol{x}_i) \neq y_i$), then the threshold is updated as:

$$\lambda_{i+1} = \lambda_i + \eta|t_i|\mathbb{1}_{y_i \text{ is OOD}} - \eta|t_i|\mathbb{1}_{y_i \text{ is ID}}, \quad (1)$$

where η is the learning rate that represents how aggressively the threshold is adapted in response to a mistake.

3. Experiments

Data Stream: We combine the test set of CIFAR-10 (ID) with five OOD datasets (SVHN (NWC+11), LSUN-Crop (YSZ+15), LSUN-Resize (YSZ+15), ImageNet-Resize (DDS+09), and iSUN (XEZ+15)) and create a stream of samples by randomly shuffling them.

Score Computation: To compute the scores we use the pretrained models on CIFAR-10 and methods as given in ((LLLS18; LWOL20; SCM21)). Distributions of the computed scores are shown in Fig. 3 in the appendix.

Evaluation: We evaluate the performance on the data stream discussed above. We run 25 trials on each method by shuffling the testing stream. We keep track of the following three metrics:

1. **Average Mistakes**: A mistake is that when one ID sample is classified as OOD, or an OOD sample is classified as ID. Then the mistakes are averaged by the total number of the samples we have seen so far.
2. **Adapted Threshold**: The adapted threshold when more samples are revealed.
3. **False Positive Rate**: The false positive rate is the fraction of OOD samples classified as ID, divided by total number of the OOD samples.
4. **True Positive Rate**: The true positive rate is the fraction of ID samples classified as OOD, divided by total number of the ID samples.

To compare the result of Active-Querying and Random-Querying, we set the number of queries in both methods approximately the same in long run. To better compare the adapted thresholds used by different methods, we do a grid search over the all the scores and compute the threshold that gives the least error rate, and the threshold that gives 5% FPR. We denote this threshold as the least average mistakes threshold, and 5% FPR threshold, respectively. The results are shown in Fig. 3, in which we use initial threshold λ_0 for 0% FPR: we have no access to any ID or OOD samples initially. We consider it to be very conservative that all the samples are considered as OOD (Thus a 0% FPR) to start with. Therefore, we set the initial threshold to be the largest score among all the scores of samples.

Results and Discussion: For Mahalanobis Score, Always-, Active-, and Random-Querying achieves the same average mistakes and TPR in the long run. Whereas Active-Querying achieves better FPR. Active- and Random-Querying only query 21.1% of the samples.

For Energy Score, the ID and OOD samples are well separated. Always-, Active-, and Random-Querying are all able to converge to the least average mistakes threshold. Always- and Active-Querying converge at approximately the same rate, which is faster than that of Random-Querying. In the process of convergence, Always- and Active-Querying presents better average mistakes and TPR. Random-Querying exhibits a larger variance across 25 trails, comparing to Always- and Active-Querying. Active- and Random-Querying query only 13.4% of the samples.

For SSD score, Always-, Active-, and Random-Querying exhibit different behaviors. Always-Querying does better in TPR, Active-Querying does better in FPR, and Random-Querying is in the middle. In terms of average mistakes, three methods converge in the long run. Active- and Random-Querying query 49.9% of the samples.

Overall, Active-Querying achieves the same average error rate as Always-Querying, while making much fewer queries. In Energy Score, Active-Querying outperforms Random-Querying and shows less variance. Notice that Always-Querying is not practical in the real-world applications as it defeats the purpose of having the ML model for decision making. We list Always-Querying as a bench mark in comparison to Active- and Random-Querying. Our experiments show that Active-Querying can do as well as Always-Querying, and can outperform Random-Querying in Energy Score and has lower variance.

4. Related Work

Out-Of-Distribution Detection: The problem of OOD detection has been addressed in many recent works where the main contributions have been methods to quantify a score (uncertainty) which gives a better separation of OOD and ID data points. Liang et al. (LLS17) proposed ODIN, which uses temperature scaling to separate the softmax score distributions between ID and OOD images. Liu et

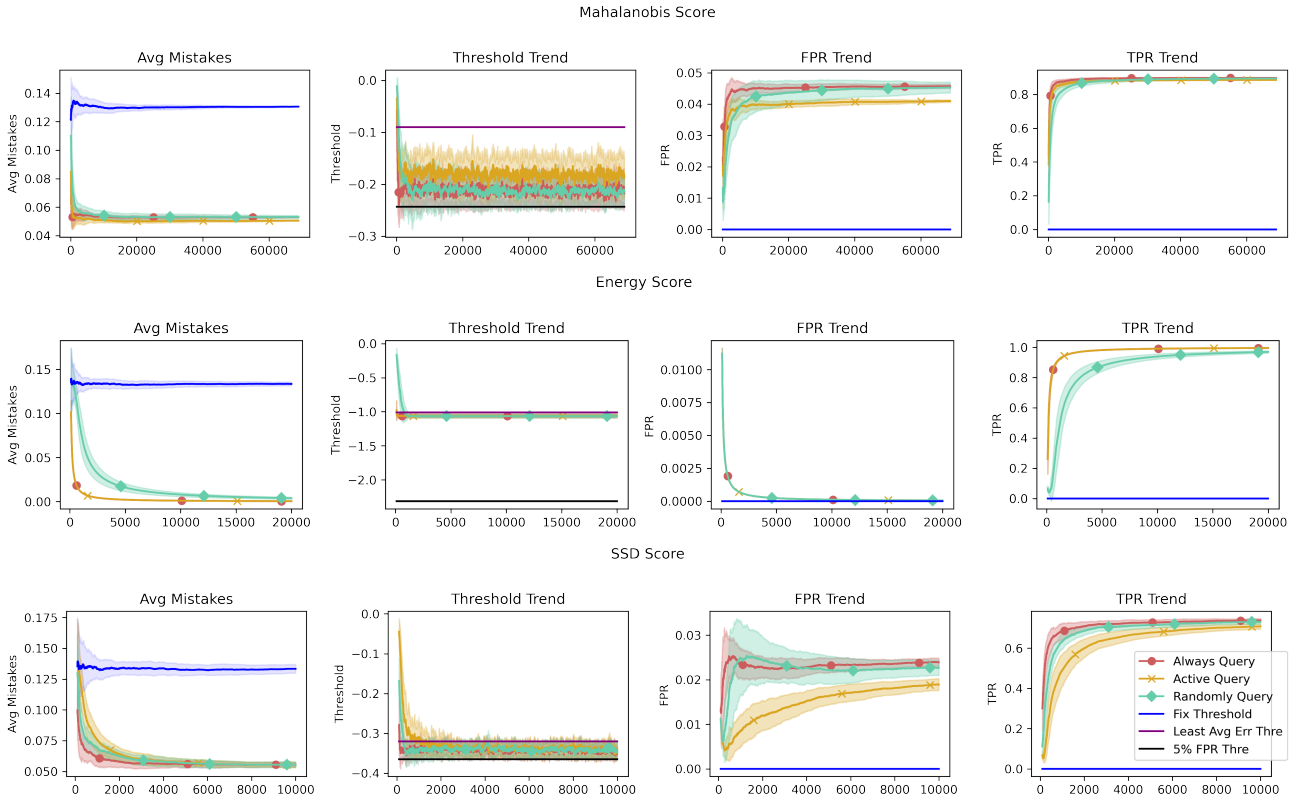

Figure 2. Threshold adaptation using CIFAR10 as ID dataset and mixture of five OOD datasets. Three scoring methods are used (from top to bottom). The x-axis is the number of testing samples revealed so far. The y-axis indicates: **1)** the average mistakes on the revealed test samples. **2)** The threshold adaption. **3)** FPR the the revealed testing samples. **4)** TPR on the revealed testing samples. The result are averaged across the 25 trails. The shading around the line denotes the variance across 25 trails.

al. (LWOL20) proposed a framework using energy score to perform OOD detection on pre-trained neural classifiers. Lee et al. (LLLS18), Sehwag et al. (SCM21), and Ming et al. (MSDL22) proposed mahalanobis distance-based scores to detect OOD samples. While these methods perform well, the evaluation setup is rather static and does not reflect the real-world deployment scenario, wherein the system has to adapt to new and evolving OOD data. In our work we are proposing a simple and extensible system for online OOD detection. Moreover the system can also adapt by getting ground truth labels from humans on selected points.

Online Anomaly Detection: There is a rich literature on anomaly (or outlier) detection in offline settings (CBK09; CZS+16; CC19). However, our setting is akin to online anomaly (outlier) detection – wherein the system receives samples one at a time and it has to figure out the outliers or anomalous behaviour within a given window of time. Some of the notable works along this line are (SPP+06; AF07; ZMH13). The methods proposed are unsupervised and perform density or distance based detection.

Outlier Detection with Human in the Loop: The notion of outlier may not always be based on statistical rarity and might need input from humans to learn the notion of outlier in the application of interest. Some of the recent works (CCL+20; IDDN18) have given methods for outlier detection in offline setting leveraging human inputs. The focus has been on minimizing the human effort by figuring out some candidate outliers and designing good questions and context for getting human inputs.

While there are a number of works on outlier or OOD detection, the main focus has been on designing methods (scoring functions) to distinguish inliers vs outliers mostly in the offline setting. Our work is rather complementary – we consider a deployed OOD system that takes into account an ensemble of various scoring functions and propose ways for online adaptation of this system based on human inputs.

5. Conclusion

We studied the problem of online OOD detection and proposed a simple system that can take human inputs and adapt itself to better match with a variety of OOD data. With an initial empirical evaluation on `CIFAR-10` and a mixture of 5 benchmark OOD datasets we show that the Active-Querying method makes less mistakes and has less FPR in comparison to other methods. While the initial results are promising, a more comprehensive study with new methods for adaptation and label noise is left as future work.

References

[AF07] Fabrizio Angiulli and Fabio Fassetti. Detecting distance-based outliers in streams of data. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 811–820, 2007.

[AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), jul 2009.

[CC19] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[CCL+20] Chengliang Chai, Lei Cao, Guoliang Li, Jian Li, Yuyu Luo, and Samuel Madden. Human-in-the-loop outlier detection. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 19–33, New York, NY, USA, 2020. Association for Computing Machinery.

[CZS+16] Guilherme O. Campos, Arthur Zimek, Jörg Sander, Ricardo J. Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30(4):891–927, 2016.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[IDDN18] Md Rakibul Islam, Shubhomoy Das, Janardhan Rao Doppa, and Sriraam Natarajan. Glad: Glocalized anomaly detection via human-in-the-loop learning. *arXiv preprint arXiv:1810.01403*, 2018.

[LLLS18] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[LLS17] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[LWOL20] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

[MSDL22] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. *arXiv preprint arXiv:2203.04450*, 2022.

[NWC+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[NYC15] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436. IEEE Computer Society, 2015.

[SCM21] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[SPP+06] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, page 187–198. VLDB Endowment, 2006.

[XEZ+15] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[YSZ+15] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[YZLL21] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[ZMH13] Yang Zhang, Nirvana Meratnia, and Paul J.M. Havinga. Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. *Ad Hoc Networks*, 11(3):1062–1074, 2013.

6. Appendix

Scores Distribution: We show the score distribution of the CIFAR10 and five OOD datasets in Fig. 6. Normalization is used to re-scale the scores so that the higher the score is, the more likely it is a ID sample.

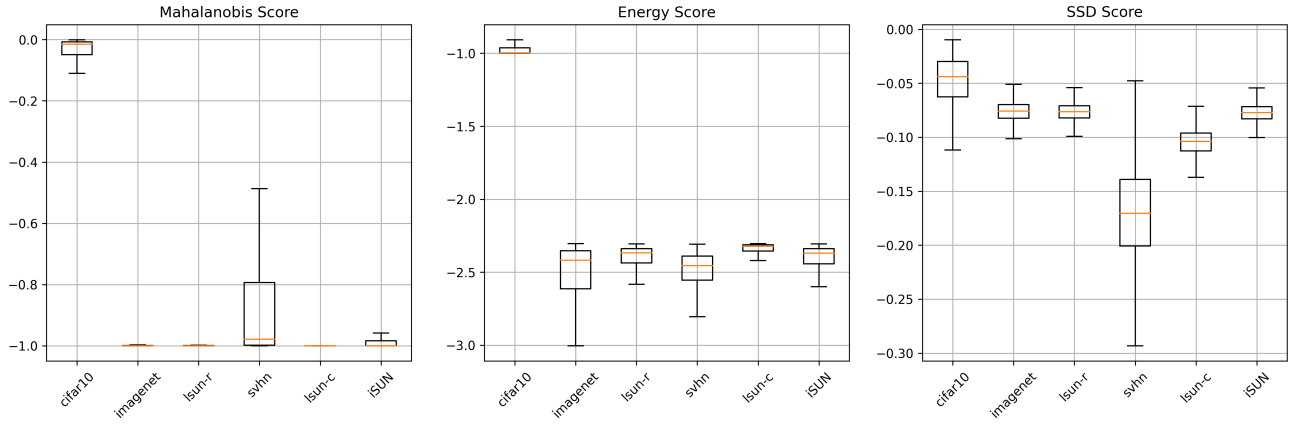

Figure 3. The score distribution of CIFAR10 and five OOD datasets with three scoring methods.