Postprint of Anomalous Distribution Detection Method Based on Sparse Optimization

Authors: Chen Qichao, Li Kuan, Li Kuan

Date: 2023-02-15T00:00:00+00:00

Abstract

Modern neural networks may produce high-confidence predictions for inputs from outside the training distribution, posing a potential threat to machine learning models. Detecting out-of-distribution inputs is a core problem for the safe deployment of models in the real world. Energy-based detection methods directly utilize feature vectors extracted by the model to compute the energy score of samples, yet relying on unimportant features may affect detection performance. To address this issue, we propose a loss function based on sparse optimization. We fine-tune a pretrained classification model, maintaining its classification capability during the learning process while increasing the sparsity of features for normal samples, thereby reducing the energy scores of normal samples and enlarging the score discrepancy between normal and anomalous samples, thus improving detection effectiveness. This method does not introduce an auxiliary anomalous dataset, avoiding the influence of correlations between samples. Experimental results on the CIFAR-10 and CIFAR-100 datasets demonstrate that the proposed method reduces the average FPR@95 for detecting six anomalous datasets by 15.02% and 15.41%, respectively.

Full Text

Abstract

Modern neural networks may produce high-confidence predictions for inputs drawn from outside the training distribution, posing a potential threat to machine learning models. Detecting out-of-distribution (OOD) inputs is a core problem for the safe deployment of models in real-world scenarios. While energy-based detection methods compute sample energy scores using feature vectors extracted from models, reliance on non-essential features may degrade detection performance. To address this issue, we propose a sparsity-optimized loss function that fine-tunes a pre-trained classification model. During learning, this approach preserves the model' s classification capability while increasing

the sparsity of in-distribution sample features, thereby lowering energy scores for normal samples and enlarging the score discrepancy between normal and anomalous samples to improve detection effectiveness. Our method introduces no auxiliary anomalous data, avoiding issues related to inter-sample correlation. Experimental results on CIFAR-10 and CIFAR-100 demonstrate that the proposed method reduces the false positive rate at 95% true positive rate (FPR95) by 15.02% and 15.41%, respectively.

Keywords: neural networks; out-of-distribution detection; energy score; fine-tuning; sparsity regularization

1. Introduction

In recent years, deep neural network (DNN) based algorithms have achieved remarkable success in classification tasks within the machine learning domain [**?**]. These algorithms are typically designed and trained on data from specific distributions in static, closed, ideal environments, and are expected to make predictions on in-distribution input data. However, in real-world applications, models frequently encounter out-of-distribution inputs and will nevertheless attempt to make predictions. An increasing body of work demonstrates that models suffer from overfitting to training data, producing overconfident yet incorrect results for inputs from unknown distributions [**?**, **?**]. When deploying machine learning models in the real world, a reliable system must not only produce accurate predictions on known inputs from familiar distributions but also detect unknown inputs from unfamiliar distributions to ensure system reliability and safety. For instance, in autonomous vehicles [**?**], we hope the driving system can detect anomalous scenes or objects it has never encountered before and promptly alert human users to take control when it cannot make safe decisions. Consequently, the out-of-distribution detection problem has been proposed and has rapidly attracted widespread attention [**?**]. Its goal is to determine whether model inputs originate from the training data distribution, thereby preventing models from producing unreliable predictions.

2. Methodology

2.1 Problem Formulation

Out-of-distribution detection can be formalized as a binary classification problem. Let P_X denote the data distribution defined on any sample space $\mathcal{X}$, from which we independently and identically sample a batch of training data $\mathcal{D}_{\text{in}} = \{(x_i, y_i)\}_{i=1}^{n}$ to train the model. Samples from the same distribution are called normal samples, while those from unrelated distributions are considered anomalous samples. The label set for normal samples is $\mathcal{Y} = \{1, 2, \dots, K\}$, where K represents the number of categories. The goal of OOD detection is to design an evaluation function $G(x)$ to assess whether input data originates from distribution P_X:

$$G(x) = \begin{cases} 0, & S(x) \geq \delta \\ 1, & S(x) < \delta \end{cases}$$

where $S(x)$ is a scoring function that computes the anomaly degree of a test sample, and δ is a detector threshold used to adjust model confidence. The threshold δ is selected based on practical requirements, typically choosing the value that correctly identifies normal samples.

2.2 Energy Models and Energy Score Functions

Energy models [**?**] express probability density through Gibbs distribution. Given a function mapping $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that maps inputs from the sample space to energy values, the collection of energy values can be expressed as:

$$p(x, y) = \frac{\exp(-F(x, y))}{Z}$$

where Z is the normalization constant, also known as the partition function. For any classification model $f : \mathcal{X} \to \mathbb{R}^K$ that maps inputs to K-dimensional feature vectors (commonly called Softmax logits), the probability of a sample being predicted as class y is:

$$p(y|x) = \frac{\exp(f_y(x))}{\sum_{i=1}^{K} \exp(f_i(x))}$$

From the energy model perspective, the denominator of the Softmax function represents the energy function. When model parameters are fixed, the energy function can be defined as:

$$E(x; f) = -\log \sum_{i=1}^{K} \exp(f_i(x))$$

Energy-based detection methods use this as the scoring function, considering samples with lower energy values as normal and those with higher energy values as anomalous.

2.3 Sparsity-Optimized Fine-Tuning Framework

Given a batch of labeled normal sample training data $\mathcal{D}_{\text{in}} = \{(x_i, y_i)\}_{i=1}^{n}$, we first train a convolutional neural network (CNN) using the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{D}_{\text{in}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{in}}} \log(p_{y_i}(y|x_i))$$

After training, we obtain a pre-trained model f_θ with the ability to classify normal samples. However, this model' s detection capability for anomalous samples is often insufficient because the training process focuses solely on the classification task. To further enhance the model' s ability to detect anomalous samples, we propose a sparsity regularization term:

$$\mathcal{L}_{\text{sparsity}} = \frac{1}{|\mathcal{D}_{\text{in}}|} \sum_{x_i \in \mathcal{D}_{\text{in}}} \|g(x_i)\|_1$$

where $g(x_i)$ represents the feature vector extracted by the model and $\|\cdot\|_1$ denotes the L1 norm. This regularization increases the sparsity of normal sample feature vectors. Since sparser features have lower energy values, this improves the performance of energy-based detection algorithms by enlarging the energy value discrepancy between normal and anomalous samples.

The total optimization objective function for fine-tuning is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{sparsity}}$$

where λ is a hyperparameter balancing the sparsity degree. Notably, this sparsity optimization loss is a nearly hyperparameter-free scheme, which significantly reduces objective function complexity compared to energy margin loss functions that require dataset-specific hyperparameters.

3. Experiments

3.1 Datasets

We select the widely-used CIFAR-10 and CIFAR-100 datasets [**?**] as normal sample training data for training deep neural network image classification models. CIFAR-10 consists of 60,000 color images ($3 \times 32 \times 32$) depicting common objects in daily life. CIFAR-100 has richer semantic information with more specific labels (e.g., not just "vehicle" but specific types of vehicles). To ensure experimental accuracy, we strictly control the evaluation setup to ensure that the training dataset, test dataset, and any anomalous data exposed to the model have minimal semantic overlap—meaning the model has not previously seen the test data.

We use four different test datasets to evaluate our method, covering most common real-world scenarios: - Texture dataset [**?**] - Places365 scene dataset [**?**] - LSUN-Crop and LSUN-Resize datasets [**?**]

All images are downsampled to 32×32 to maintain consistency with normal sample dimensions.

3.2 Evaluation Metrics

Following standard experimental settings in anomaly detection research, we use three metrics to measure method effectiveness: FPR95, AUROC, and AUPR.

Let TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives, respectively. Then: - True Positive Rate: TPR = TP/(TP+FN) - False Positive Rate: FPR = FP/(FP+TN) - Precision: Precision = TP/(TP+FP)

FPR95 represents the false positive rate when the true positive rate is 95%. **AUROC** is the area under the ROC curve, which plots TPR against FPR. **AUPR** is the area under the Precision-Recall curve. Lower FPR95 and higher AUROC/AUPR values indicate better performance.

3.3 Implementation Details

We use a WideResNet [**?**] with residual connections as the backbone network for training normal sample classification models. Data augmentation includes random flipping and cropping to increase dataset diversity. During training, each channel of the datasets is normalized using standard mean and variance. The algorithm employs stochastic gradient descent (SGD) with momentum. The momentum parameter is set to 0.01, and the L2 norm weight decay coefficient is also set to 0.01.

In the pre-training phase, the learning rate is relatively large (0.1) for 200 epochs to facilitate rapid exploration of the parameter space. In the fine-tuning phase, we start from a model that already possesses certain knowledge, requiring only a small learning rate (0.001) for local optimization. We set the balancing constant λ to 0.001 and train for 10 epochs, keeping all other configurations unchanged.

3.4 Experimental Results

[Figure 1: see original paper] shows the energy score distributions before and after fine-tuning. The green curve represents normal samples, while the red curve represents anomalous samples. The overlapping area between the curves indicates where normal and anomalous samples cannot be distinguished. After fine-tuning with our method, this overlapping region is significantly reduced, demonstrating substantially improved detection capability. The shift in absolute score values occurs because fine-tuning produces a new model with transformed feature representations.

[Figure 2: see original paper] illustrates the ROC curve changes, where the blue curve represents the pre-trained model and the orange curve represents our fine-tuned model. The curve shifts toward the upper-left corner after fine-tuning, indicating improved detection performance.

presents quantitative comparison results. Each experiment is run 5 times to avoid random errors, with average values reported. Our method achieves com-

prehensive improvements across all datasets. On CIFAR-10, FPR95 is reduced by an average of 15.02%; on CIFAR-100, FPR95 is reduced by an average of 15.41%. Notably, our method requires no auxiliary anomalous datasets—only normal sample training data for fine-tuning—making it more practical and easier to implement.

3.5 Comparison with State-of-the-Art Methods

To demonstrate our method' s advantages in the current research landscape, we compare it with popular detection methods including ODIN [**?**], Mahalanobis [**?**], and Energy [**?**]. For fair comparison, all methods use the same WideResNet backbone trained only on normal samples without auxiliary datasets. shows that our method leads all evaluation metrics across test datasets, significantly improving upon current methods and demonstrating strong competitiveness.

4. Conclusion

We propose a simple yet effective loss function for fine-tuning pre-trained classification models, which significantly enhances OOD detection capability. By increasing the sparsity of normal sample features, normal samples obtain lower energy scores, thereby enlarging the energy discrepancy between normal and anomalous samples. Experimental results show that even without exposing the model to anomalous data, our method achieves comparable performance to approaches that do, while consistently improving detection across different anomaly types. Future work will explore performance on higher-dimensional and larger-scale datasets.

References

[1] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 1026-1034.

[2] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 427-436.

[3] HEIN M, ANDRIUSHCHENKO M, BITTERWOLF J. Why relu networks yield high-confidence predictions far away from the training data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 41-50.

[4] FILOS A. Can autonomous vehicles identify, recover from, and adapt to distribution shifts?[C]//Proceedings of the International Conference on Machine Learning. New York, NY: ACM Press, 2020: 3145-3153.

[5] HENDRYCKS D, GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[EB/OL]. (2016-10-07)[2021-10-03]. https://arxiv.org/abs/1610.02136.

[6] LIU Weitang, WANG Xiaoyun, OWENS J, et al. Energy-based out-of-distribution detection[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2020: 21464-21475.

[7] LIANG S, LI Yixuan, SRIKANT R. Enhancing the reliability of out-of-distribution image detection in neural networks[EB/OL]. (2017-07-08)[2020-08-30]. https://arxiv.org/abs/1706.02690.

[8] HSU Yilin, SHEN Yujie, JIN Hongxia, et al. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 10951-10960.

[9] LEE K, LEE H, LEE K, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 7167-7177.

[10] HENDRYCKS D, MANTAS M, KADAVATH S, et al. Deep anomaly detection with outlier exposure[EB/OL]. (2018-12-04)[2022-01-28]. https://arxiv.org/abs/1812.04606.

[11] TORRALBA A, FERGUS R, FREEMAN W T. 80 million tiny images: a large database for nonparametric object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958-1970.

[12] PAPADOPOULOS A, RAJATI M, SHAIKH N. Outlier exposure with confidence control for out-of-distribution detection[J]. Neurocomputing, 2021, 441: 138-150.

[13] YU Q, AIZAWA K. Unsupervised out-of-distribution detection by maximum classifier discrepancy[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 9518-9526.

[14] LEE K, LEE H, LEE K, et al. Training confidence-calibrated classifiers for detecting out-of-distribution samples[EB/OL]. (2017-11-26)[2022-02-23]. https://arxiv.org/abs/1711.09325.

[15] CHEN Y, WANG T, COATES A, et al. Informative outlier matters: robustifying out-of-distribution detection using outlier mining[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, German: Springer, 2021: 8-26.

[16] LI Y, VASCONCELOS N. Background resampling for outlier-aware classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 13218-13227.

[17] MOHSENI S, PITALE M, YADAWA J, et al. Self-supervised learning for out-of-distribution detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAAI Press, 2020, 34: 5216-5223.

[18] YANG Jingkang, WANG Haoqi, FENG Litong, et al. Semantically coherent out-of-distribution detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 8301-8309.

[19] LECUN Y, CHOPRA S, HADSELL R, et al. A tutorial on energy-based learning[M]//Predicting Structured Data. Cambridge, MA: MIT Press, 2006.

[20] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[D]. Toronto: University of Toronto, 2009: 48-60.

[21] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1452-1464.

[22] CIMPOI M, MAJI S, KOKKINOS I, et al. Describing textures in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 3606-3613.

[23] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]//NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Cambridge, MA: MIT Press, 2011.

[24] YU F, ZHANG Y, SONG S, et al. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop[EB/OL]. (2015-06-10)[2021-06-04]. https://arxiv.org/abs/1506.03365.

[25] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 580-587.

[26] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

Note: Figure translations are in progress. See original paper for figures.