Exploring Out-Of-Distribution Detection

Matthew Bloom
Duke University
Durham, NC 27708

Luis Cunha
Duke University
Durham, NC 27708

Abstract

It is often the case that during deployment, machine learning models encounter examples outside of the distribution of classes seen during training. Given this possibility, there is a concern for classifying these anomalous inputs when they should not be, and it becomes important to be able to detect them and not output an erroneous prediction. Our project examines the effectiveness of various out-of-distribution (OOD) detectors in order to improve the robustness of a deep neural network (DNN) when faced with image inputs that belong to these OOD classes - an issue that is common in real-world scenarios. We also explore the impact of outlier exposure during training in improving the OOD detection performance, which we find to be highly effective.

1 Introduction

The ability of neural networks to perform well when classifying samples from the set of classes that they were trained on has been long-established. However, what happens when a network is faced with an input from outside of this set? Without any modification, these out-of-distribution (OOD) samples will still result in the output of a confident prediction from one of the classes from the in-distribution (ID) training classes, even when it is not relevant. In real-world deployment, this is a common issue faced as the inputs of a neural network cannot be ensured to be in-distribution. In recent years, there have been a variety of methods proposed for the flagging of OOD samples, where a detector identifies an input as out-of-distribution and and the network does not output an erroneous prediction.

Our main contributions are as follows:

- We implement, examine, and compare the performance of three different OOD detection algorithms in use today: baseline, energy, and ODIN.
- The detectors' performance is tested across multiple OOD datasets, and the effect of varying the network capacity is evaluated.
- We also explore the impact of outlier exposure (OE), an alternative training method where OOD samples are included in order to improve detection performance. The advantages of varying the diversity of this OOD set are also examined.

2 Related Works

In this section, we acknowledge the previous research that was instrumental in our contributions.

2.1 Baseline Method for Detecting Out-of-Distribution Examples in Neural Networks

Hendrycks and Gimpel established the baseline for OOD detection, which simply relies on the observation that the softmax probabilities output following an ID input tend toward having a higher

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

maximum probability than that of OOD samples. By using a simple threshold, baseline level performance can be achieved.

2.2 ODIN for Enhancing OOD Detection

Liang et al. proposed ODIN (Out-of-Distribution detector for Neural networks) as an algorithm that focuses on the effect of temperature scaling and input perturbation in order to further separate the softmax scores from ID and OOD samples and ease identification of each.

2.3 OOD Detection using Energy

Liu et al. introduced the idea of calculating a scalar energy score for each input as an alternative to methods based on the softmax probabilities such as the baseline method and ODIN. The authors observe that using the softmax outputs results in overconfident predictions of the OOD inputs.

2.4 Training using Outlier Exposure to Enhance OOD Detection

As opposed to a direct detection method, Hendrycks et al. suggested an alternative training pipeline, where the neural network is exposed to out–of-distribution samples during training. By utilizing a training objective that results in lower softmax probabilities for OOD inputs, OOD detection becomes easier during testing.

3 Methodology

Here, we explain the technical motivation and mathematical formulation for everything we explored.

3.1 Out-of-Distribution Detectors

Three different methods of differentiation between out-of-distribution and in-distribution samples were explored.

3.1.1 Baseline

As mentioned earlier, the baseline method of detection is motivated by the lower maximum softmax probabilities for out-of-distribution samples. So, we implement the baseline detector using the standard softmax function that is normally applied to the logits (outputs) of the neural network's layers to produce a set of probabilities. The "softmax score" is the maximum value of these probabilities. Then, given an input $\boldsymbol{x}$ and a threshold probability δ, the baseline out-of-distribution detector is described as follows:

$$g(\mathbf{x};\delta) = \begin{cases} 1 & \text{if } \max_i p(\boldsymbol{x}) \leq \delta \\ 0 & \text{if } \max_i p(\boldsymbol{x}) > \delta \end{cases}$$

A 1 is considered a positive, in-distribution sample, while a 0 is a negative, out-of-distribution sample. The detector predicts that a sample is OOD for inputs with a softmax score below the threshold, and ID for a sample whose softmax score exceeds the threshold.

3.1.2 ODIN

Our implementation only includes the temperature scaling step of the original method outlined. This is implemented simply by applying a scaling factor T to the logits that are inputs to the softmax, formulated as follows, where the neural network $\boldsymbol{f}$ classifies N classes and outputs N logits f_i:

$$p_i(\mathbf{x};T) = \frac{\exp(f_i(\boldsymbol{x}/T))}{\sum_{j=1}^{N} \exp(f_j(\boldsymbol{x}/T))}$$

We utilize a factor of $T = 1000$ for all experiments outlined below as in the paper. This temperature scaling has the effect of creating a further distinction between the softmax scores for in-distribution and out-of-distribution inputs. Then, the same thresholding applied in the baseline method can be used to greater effect for out-of-distribution detection.

3.1.3 Energy

As mentioned earlier, the energy-based method of OOD detection proposes that softmax scores lead to overconfidence in the probability distribution of OOD inputs. Therefore, we implement a scalar energy score based on the free energy of an input $\boldsymbol{x}$, where the free energy of ID inputs is lower (more negative) than that of OOD inputs:

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

In practice, the negative of the free energy is used in order to maintain the convention from the baseline method of in-distribution samples having larger scores. Then, the energy scores are similarly thresholded as they are in baseline and ODIN to detect OOD samples.

3.2 Outlier Exposure

The outlier exposure (OE) technique involves exposing a model to OOD examples during training and using a learning objective that yields lower softmax probabilities for these examples. As a result, it becomes easier for OOD detectors, e.g. baseline, to discriminate between ID and OOD samples. This is equivalent to minimizing the following learning objective, $\mathcal{L}$, over the parameters of a given model f, where $\mathcal{D}_{\text{in}}$ is the in-distribution dataset, and $\mathcal{D}_{\text{out}}^{\text{OE}}$ is the exposed outlier dataset.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}}[\mathcal{L}(f(x), y) + \lambda\mathbb{E}_{x'\sim\mathcal{D}_{\text{out}}^{\text{OE}}}[\mathcal{L}_{\text{OE}}(f(x'), f(x), y)]]$$

Due to the applicability of outlier exposure in a wide variety of tasks and datasets, the specific formulation of the OE learning objective, $\mathcal{L}_{\text{OE}}$, is a design choice. In our case, having chosen the baseline detector, we opt for implementing the cross-entropy loss.

4 Experiments

In this section, we provide details of the setup and results from the experiments we carried out. All experiments were run on Google Colab and PyTorch was used for implementation.

4.1 Training Setup

All (non-outlier exposure) models implemented the residual network (ResNet) architecture and were trained using the CIFAR-10 training set, which contains 50000 32x32 RGB images from 10 classes, 5000 each. Networks with depths of 18, 20, and 34 residual layers were trained using stochastic gradient descent for 20 epochs with the following hyperparameters: 0.1 learning rate, 0.875 momentum, and 0.0005 weight decay.

4.1.1 Outlier Exposure Training and Testing

We consider the implementation of OE in the training process of a network to analyze its impact on OOD detection. We report evaluation metrics for the baseline detection method when testing with six different OOD datasets (CIFAR-10 (6-10), SVHN, LSUN, MNIST, Uniform-Random, and Gaussian-Random) using the ResNet-20 architecture. The models are trained with the same hyperparameters as above, using the first 5 classes of CIFAR-10 as the ID dataset with the first 1, 5, 20, 50, and 100 classes of CIFAR-100 as the OOD datasets.

4.2 Out-of-Distribution Datasets

The following outlines all OOD datasets that were used for testing. Datasets were selected based on the ODIN paper's choices. The ID samples used for testing come from the CIFAR-10 testing set.

4.2.1 CIFAR-100

The CIFAR-100 dataset contains 100 classes of 32x32 RGB images with categories such as "lawn-mower," "dinosaur," or "mushrooms," with the test set used containing 100 images per class for 10000 total images.

4.2.2 SVHN

The Street View House Numbers dataset contains 32x32 RGB images of house numbers from Google Street View, with its 10 classes being the digits 0-9. For our out-of-distribution testing, a sample of 10000 images from the test set was used.

4.2.3 LSUN

The Large-scale Scene UNderstanding dataset contains a large variety of large images of scenes such as "bedroom" or "bridge," and the OOD samples used were the 10000 images in its test set. Given that the images are larger than 32x32, we separated the LSUN test set into random 32x32 cropped images (LSUN-crop) and downsampled 32x32 images (LSUN-resize).

4.2.4 MNIST

MNIST is a widely used database containing 28x28 single-channel images of handwritten digits. The 10000 images in its test set were used in our experiments, with the images padded to 32x32 and their pixel values mirrored across 3 channels.

4.2.5 Uniform-Random

We generated 10000 RGB images with pixel values randomly sampled from a uniform distribution with limits from 0 to 1 in order to test the OOD detectors on noise.

4.2.6 Gaussian-Random

The Gaussian images are 10000 RGB images, with each pixel's value drawn from a normal distribution with a mean of 0.5 and a variance of 1, then clamped between 0 and 1.

4.3 Evaluation Metrics

The evaluation metrics used for all our experiments were the area under the precision-recall curve (AUPR) and the false positive rate at 95% true positive rate (FPR@95TPR) in order to maintain consistency with the literature. The precision-recall curve plots precision (true positives/(true positives + false positives) versus recall (true positives/(true positives + false negatives)). We define positives as ID samples and negatives as OOD samples. Therefore, a high AUPR indicates both a low false positive rate, where OOD samples are classified as ID, and a low false negative rate, which indicates the reverse. FPR@95TPR indicates the proportion of out-of-distribution samples classified as ID at the threshold where 95% of in-distribution samples are classified correctly.

4.4 Comparing OOD Detector Experimental Results

Here, we discuss the main results from the comparison of the three OOD detectors across datasets, seen in Table 1.

Performance for each detector and OOD dataset. Though the three detectors demonstrate comparable results, ODIN and Energy tend to outperform the baseline method (i.e. produce higher AUPR and lower FPR@95TPR), with Energy usually performing best, as is reflected in the literature. This performance is mostly comparable across each of the OOD datasets, with minor fluctuations in a few cases, such as MNIST being easier to detect as OOD while CIFAR-100 is more difficult, as indicated by the AUPR bar heights in Figure 1.

Influence of model capacity (ResNet depth). In some cases, as the number of layers in the ResNet increases, AUPR increases, but FPR@95TPR also did. The trend is not consistent; for some OOD datasets, increasing the model capacity improved performance for both metrics (increase in AUPR and decrease in FPR@95TPR) and in some subcases both metrics became worse (a decrease in AUPR and an increase in FPR@95TPR). In some other cases, performance did not change much at all. All this is to say: results were inconsistent when changing base model depth. Nonetheless, the three detectors were successful to some extent in detecting OOD samples in all of the models tested.

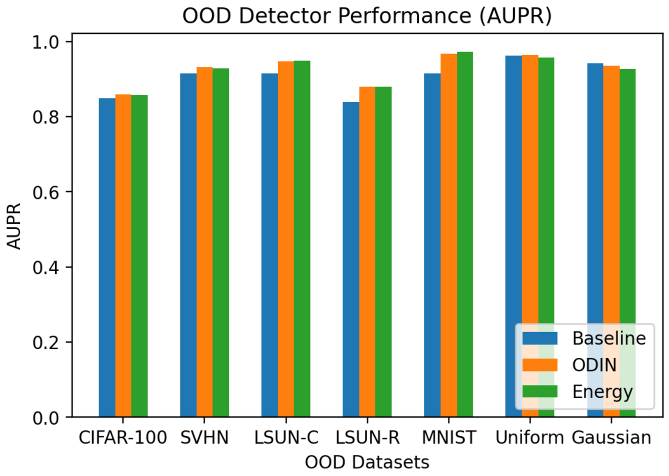

Figure 1: Bar graph representing the AUPR results for Baseline, ODIN, and Energy methods across the OOD datasets.

Table 1: Results for the three out-of-distribution detection methods across datasets and model capacities. All metrics are reported in percentages, and ↑ signifies that a higher value indicates better performance, while ↓ signifies the opposite.

Model	**Out-of-distribution dataset**	**Baseline**	**ODIN**	**Energy**
		AUPR (↑) / FPR@95TPR (↓)		
ResNet-18	CIFAR-100	84.3/72.9	85.3/66.1	85.2/65.3
	SVHN	90.9/72.8	93.9/50.6	94.1/49.0
	LSUN (crop)	90.7/61.3	95.1/33.4	95.2/30.4
	LSUN (resize)	82.5/76.0	86.8/64.4	86.9/63.3
	MNIST	94.1/45.5	98.3/12.9	98.7/7.63
	Uniform	97.4/24.0	98.2/8.88	97.9/15.7
	Gaussian	93.1/77.3	92.5/98.7	91.9/99.8
ResNet-20	CIFAR-100	84.9/74.1	86.1/67.3	86.0/67.2
	SVHN	92.3/63.7	92.5/61.2	91.8/65.9
	LSUN (crop)	92.3/63.7	94.6/36.1	94.6/35.8
	LSUN (resize)	84.3/73.9	88.6/58.6	88.6/56.9
	MNIST	91.3/59.6	96.4/28.5	96.8/25.0
	Uniform	94.4/80.8	95.9/80.8	95.7/95.1
	Gaussian	92.7/89.1	94.1/98.4	93.9/99.8
ResNet-34	CIFAR-100	85.4/74.3	86.3/66.8	86.2/66.7
	SVHN	91.5/70.2	93.0/58.6	92.8/60.7
	LSUN (crop)	91.3/62.5	94.7/36.0	92.8/60.7
	LSUN (resize)	85.1/73.7	88.3/58.4	88.5/56.7
	MNIST	89.2/69.0	95.6/39.0	96.2/33.2
	Uniform	96.8/43.4	95.0/99.9	93.7/99.9
	Gaussian	97.1/27.1	94.2/99.7	92.5/99.9

4.5 Outlier Exposure Experimental Results

As seen in Table 2, by changing the the outlier diversity during training in the CIFAR-100 dataset, we notice performance impacts in the ability to detect out-of-distribution samples from the model's output using the baseline method. With low outlier diversity, i.e. by including the first 1, or 5 classes from CIFAR-100, the OOD detection performance is poor, with AUPR trending down and FPR@95TPR trending up. However as the diversity in the OOD training set increased, i.e. by including the first 20, 50, or 100 classes of CIFAR-100, the OOD detection performance improved, with AUPR trending up

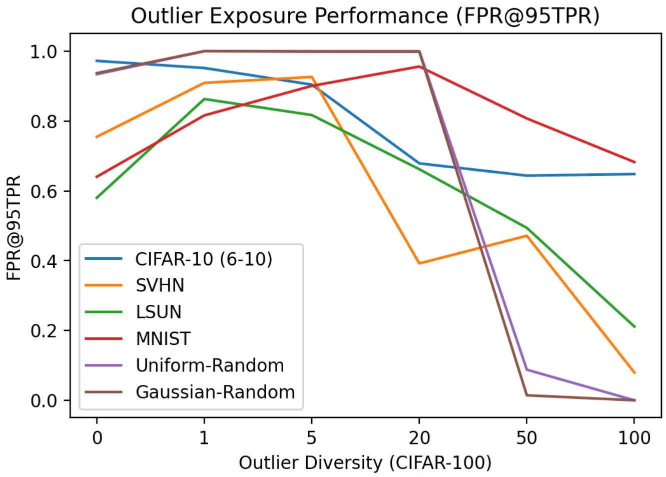

Figure 2: Plot of the trend in FPR@95TPR when testing on various OOD datasets as the diversity of the OOD training dataset increases.

Table 2: Results of baseline out-of-distribution detection after training with outlier exposure. Outlier diversity indicates the number of classes of the out-of-distribution dataset (CIFAR-100) that were included when training the model. CIFAR-10 (6-10) contains the last 5 classes of CIFAR-10. All trained models were ResNets with 20 layers.

Out-of-distribution dataset	**Outlier Diversity (CIFAR-100)**					
	0 (no OE)	**1**	**5**	**20**	**50**	**100**
	AUPR (↑) / FPR@95TPR (↓)					
CIFAR-10 (6-10)	54.4/97.2	64.9/95.2	61.2/90.4	79.3/67.9	88.5/64.4	88.1/64.8
SVHN	84.6/75.5	29.7/90.9	62.8/92.6	83.7/39.2	88.6/47.1	96.7/7.97
LSUN	86.8/58.0	39.7/86.3	59.9/81.2	68.7/66.2	87.4/49.4	93.0/21.1
MNIST	86.3/64.0	80.2/81.6	63.1/90.1	37.6/95.6	76.6/80.7	83.3/68.2
Uniform	80.4/93.4	71.2/100.	62.9/99.9	81.5/99.9	97.4/8.76	99.9/0.00
Gaussian	75.7/93.7	31.1/100.	65.1/99.9	84.7/99.8	99.2/1.43	99.9/0.00

and FPR@95TPR trending down in Figure 2. OE brought the least improvement when testing on MNIST and CIFAR-10 (6-10), but saw comparably large improvements for all other datasets. These results are consistent with our expectations from the literature.

5 Discussions and Conclusions

In our experiments we were able to not only reproduce some of the results outlined in the original research papers proposing these techniques, but also offered new performance indicators for potential applications of the same. This exploration provides valuable insights about the effectiveness of the baseline, ODIN, and energy OOD detectors across a variety of image datasets from different domains, and potential network architecture implementations. It also covers the benefits offered by outlier exposure during training across multiple image domains relative to outlier diversity. OE appears to be more powerful than the detectors used by themselves with a sufficiently diverse OOD training set, and a future opportunity for exploration could be examining the performance of various detectors when combined with OE.

References

[1] Hendrycks and Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” (https://arxiv.org/abs/1610.02136)

[2] Liang et al., “Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks”, (https://arxiv.org/pdf/1706.02690.pdf)

[3] Liu et al., “Energy-based Out-of-distribution Detection,” (http://arxiv.org/abs/2010.03759)

[4] Hendrycks et al., “Deep Anomaly Detection with Outlier Exposure,” (https://openreview.net/forum?id=HyxCxhRcY7)

[5] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a largescale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.