Exploring Out-of-Distribution Detection

Matthew Bloom, Luis Cunha

Duke
PRATT SCHOOL of ENGINEERING

Introduction

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 DNN when faced with inputs that belong to these OOD classes - an issue that is common in real-world scenarios.

- We compare three OOD detection algorithms used today: Baseline, ODIN, and Energy, and evaluate them on a variety of OOD datasets.
- We examine the effect of network architecture (depth) on OOD detection performance.
- We explore the impact of outlier exposure (OE) during training on OOD performance and how it varies relative to outlier diversity.

Methodology

We utilize the following implementations of OOD Detectors:
Baseline: this technique works by observing that the softmax probabilities output as a result of an in-distribution (ID) input usually have a higher maximum probability. So, the algorithm establishes a threshold probability, and if none of the softmax probabilities are greater than that baseline, the input is classified as OOD.
ODIN: this technique relies on temperature scaling to distinguish examples between ID and OOD. By scaling the logits by the temperature when computing softmax probabilities, the difference between softmax probabilities for ID and OOD inputs increases, allowing the algorithm to more easily detect if an input is in the training classes or not using the same thresholding as the baseline method.
Energy: instead of utilizing the softmax scores, this technique implements a function that returns a scalar “free energy” score for each input. Energy scores are calculated for each input, with a higher energy corresponding to OOD inputs and a lower energy corresponding to ID inputs. The energy score is then similarly thresholded to classify inputs as OOD or ID.

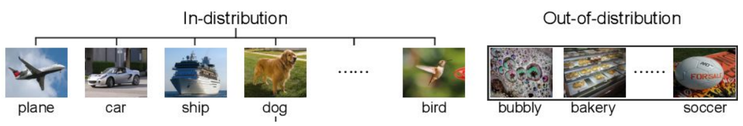

Figure 1. An illustration showing an example of the separation between ID and OOD classes.

Outlier Exposure (OE): this 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.

We consider the following evaluation metrics in our experiments: AUPR: the area under the precision-recall curve (which plots precision vs. recall, where high precision indicates low false positive rates and high recall indicates low false negative rates). FPR@95TPR: the false positive rate of OOD samples when the true positive rate of in-distribution (ID) samples is as high as 95%.

We train all models for 20 epochs using stochastic gradient descent (SGD) with the following hyperparameters: 0.1 learning rate, 0.875 momentum, and 0.0005 weight decay. The models all implement the residual network (ResNet) architecture, with depths of 18, 20, or 34 layers.

Experiments

Our initial exploration involves comparing the performance of the three OOD detection methods previously mentioned: Baseline, ODIN, and Energy. The ID classes are CIFAR-10, and we report and compare here the AUPR and FPR@95TPR metrics for each of these methods across six different OOD datasets (CIFAR-100, SVHN, LSUN, MNIST, Uniform-Random, and Gaussian-Random), and three different ResNet implementations (ResNet-18, ResNet-20, and ResNet-34).

OOD Dataset	Baseline	ODIN	Energy
CIFAR-100	0.8432/0.8493/0.8538 0.7294/0.7412/0.7431	0.8528/0.8609/0.8628 0.6610/0.6732/0.6675	0.8517/0.8599/0.8617 0.6538/0.6718/0.6674
SVHN	0.9085/0.9233/0.9152 0.7279/0.6369/0.7015	0.9390/0.9254/0.9301 0.5057/0.612/0.5859	0.9407/0.9184/0.9279 0.4904/0.6587/0.6065
LSUN-Crop	0.9074/0.9233/0.9132 0.6125/0.6369/0.6248	0.9505/0.9455/0.9472 0.3337/0.3606/0.3598	0.9516/0.9462/0.9510 0.3041/0.3581/0.3344
LSUN-Resize	0.8249/0.8432/0.8508 0.7598/0.739/0.7372	0.8682/0.8855/0.8832 0.6443/0.5860/0.5841	0.8697/0.8864/0.8845 0.6325/0.5692/0.5673
MNIST	0.9409/0.9128/0.8915 0.4551/0.5961/0.6902	0.9831/0.9640/0.9559 0.1294/0.2849/0.3896	0.9874/0.9676/0.9616 0.0763/0.2498/0.3322
Uniform-Random	0.9744/0.9441/0.9682 0.2396/0.8075/0.4338	0.9823/0.9588/0.9502 0.0888/0.8075/0.9991	0.9789/0.9568/0.9374 0.1566/0.9508/0.9999
Gaussian-Random	0.9310/0.9265/0.9714 0.7725/0.8905/0.2713	0.9245/0.9406/0.9416 0.9869/0.9836/0.9971	0.9188/0.9387/0.9253 0.9980/0.9982/0.9999

Figure 2. Summary table of results from experiments using Baseline, ODIN, and Energy OOD detection methods. The results are separated as follows: ResNet-18/20/34. The first line reports the AUPR, while the second reports the FPR@95TPR. LSUN-Crop is a random 32x32 cropped image, while LSUN-Resize is a downsampled 32x32 image. Uniform-Random images have random pixel values from 0 to 1, while Gaussian-Random samples the pixel values from a normal distribution with mean 0.5 and variance 1.

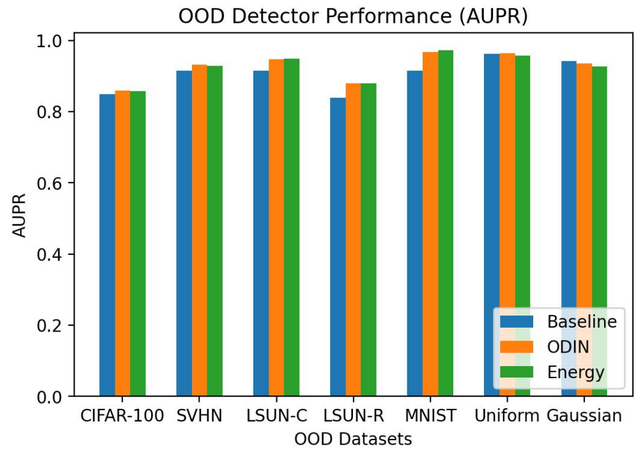

Figure 3. Bar graph representing the AUPR results for Baseline, ODIN, and Energy methods across the OOD datasets.

Following these experiments, we consider the implementation of OE in the training process of a network to analyze its impact on OOD detection. We report the AUPR and FPR@95TPR 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 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.

OOD Dataset	Outlier Diversity (CIFAR-100)					
	0 (non-OE)	1	5	20	50	100
CIFAR-10 (6-10)	0.5438 0.9722	0.6492 0.9517	0.6122 0.9042	0.7932 0.6785	0.8845 0.6435	0.8809 0.6480
SVHN	0.8457 0.7547	0.2971 0.9092	0.6281 0.9261	0.8368 0.3919	0.8855 0.4712	0.9668 0.0797
LSUN	0.8676 0.5801	0.3966 0.8629	0.5987 0.8170	0.6872 0.6618	0.8740 0.4938	0.9296 0.2112
MNIST	0.8634 0.6403	0.8015 0.8160	0.6308 0.9005	0.3763 0.9560	0.7661 0.8072	0.8332 0.6824
Uniform-Random	0.8036 0.9339	0.7120 1.0000	0.6285 0.9999	0.8152 0.9999	0.9743 0.0876	0.9992 0.0000
Gaussian-Random	0.7572 0.9370	0.3113 1.0000	0.6511 0.9985	0.8472 0.9984	0.9917 0.0143	0.9991 0.0000

Figure 4. Table of results from our experiments using OE. The first line reports AUPR, while the second reports FPR@95TPR. Outlier diversity reflects the number of classes of the OOD dataset (CIFAR-100) that were included in training. CIFAR-10 (6-10) are the last 5 classes of CIFAR-10.

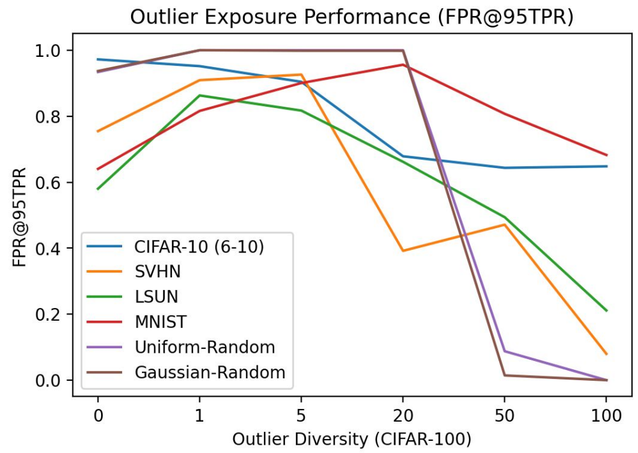

Figure 5. Plot of the trend in FPR@95TPR when testing on various OOD datasets as the diversity of the OOD training dataset increases.

Conclusions

OOD Detectors: Though the three detectors demonstrate comparable results, ODIN and Energy tend to outperform (i.e. produce higher AUPR and lower FPR@95TPR) Baseline, with Energy usually performing best, as is reflected in the literature. These methods were successful in detecting OOD samples in all of the models tested. We did not observe consistent trends in performance when changing the network architecture (number of ResNet layers).

Outlier Exposure: With low outlier diversity, the OE performance is poor, however as the diversity in the OOD training set increased, testing performance increased, with AUPR trending up and FPR@95TPR trending down. 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.

References

Hendrycks and Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” (https://arxiv.org/abs/1610.02136)
Hendrycks et al., “Deep Anomaly Detection with Outlier Exposure,” (https://openreview.net/forum?id=HyxCxhRcY7)
Liang et al., “ENHANCING THE RELIABILITY OF OUT-OF-DISTRIBUTION IMAGE DETECTION IN NEURAL NETWORKS”, (https://arxiv.org/pdf/1706.02690.pdf)
Liu et al., “Energy-based Out-of-distribution Detection,” (http://arxiv.org/abs/2010.03759)