NeuralFP: Out-of-distribution Detection using Fingerprints of Neural Networks

Wei-Han Lee, Steve Millman, Nirmit Desai, Mudhakar Srivatsa, Changchang Liu

IBM T. J. Watson Research Center

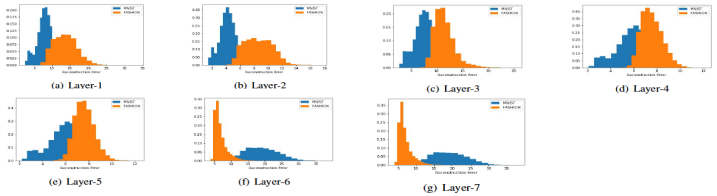
Background:

Edge devices use neural network models learnt on cloud to predict labels of its data records

- may lead to incorrect predictions especially for OOD data involved in the training process

However, recent efforts in OOD detection

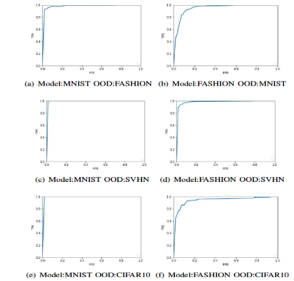
- require the retraining of the model
- assume the existence of a certain amount of OOD records



1) reconstruction errors are strong signals for distinguishing OOD records from in-distribution data
2) reconstruction errors of OOD records are not always larger than in-distribution data (Layer-6 vs Layer-7)
3) different layers have various capability in distinguishing OOD records (Layer-1 vs Layer-3)

Experimental Results

1. Effectiveness in Detecting OOD Data



AUCs of detecting various OOD records

Training \ OOD	MNIST	FASHION	CIFAR10	SVHN
MNIST	N.A.	0.9826	0.9921	0.9921
FASHION	0.9498	N.A.	0.9442	0.9809

2. Usefulness of One-Out Integration Strategy

AUCs for different layer under MNIST model

	FASHION	CIFAR10	SVHN
Layer-1	0.7298	0.8817	0.8821
Layer-2	0.8260	0.8848	0.8850
Layer-3	0.6893	0.8394	0.8634
Layer-4	0.4190	0.5340	0.7661
Layer-5	0.4298	0.4821	0.6927
Layer-6	0.8678	0.8902	0.8899
Layer-7	0.8741	0.8955	0.8944

AUCs for different layer under Fashion-MNIST model

	MNIST	CIFAR10	SVHN
Layer-1	0.5280	0.6983	0.7499
Layer-2	0.5995	0.7178	0.7606
Layer-3	0.6013	0.7146	0.7546
Layer-4	0.8163	0.7851	0.7801
Layer-5	0.8236	0.8062	0.8447
Layer-6	0.7980	0.7868	0.8217
Layer-7	0.8009	0.7523	0.8064
Layer-8	0.6668	0.3737	0.3797
Layer-9	0.6446	0.2530	0.2418
Layer-10	0.1961	0.5216	0.5369
Layer-11	0.3826	0.5893	0.4637

3. Advantages over Previous Methods

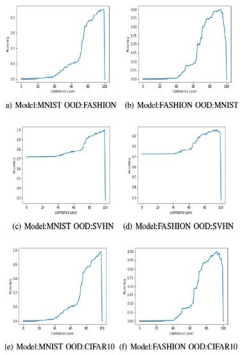
AUCs by using Mahalanobis [4] under MNIST model

OOD \ ϵ	FASHION	CIFAR10	SVHN
0.0005	0.4566	0.6472	0.5088
0.001	0.5046	0.9145	0.9621
0.014	**0.5139**	**0.9846**	**0.9788**
0.002	0.5048	0.974	0.9684
0.005	0.4659	0.9295	0.9304
0.01	0.4437	0.9165	0.9006

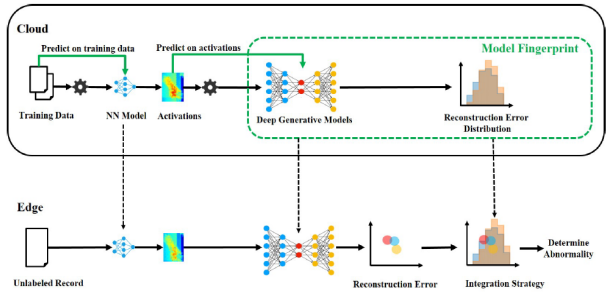
AUCs by using ODIN [1] under MNIST model

OOD \ ϵ	FASHION	CIFAR10	SVHN
0.0005	**0.9289**	0.9650	0.9858
0.001	0.9286	0.9649	0.9858
0.014	0.9285	0.9650	0.9859
0.002	0.9284	0.9650	0.9860
0.005	0.9276	0.9657	0.9870
0.01	0.9270	**0.9677**	**0.9891**

4. Practical Guidelines for Selecting Parameters



Design Details



Fingerprinting on the Cloud

1) obtain activations for each layer by passing training data through neural network model
2) construct deep generative models of each layer based on activations
3) compute the reconstruction errors of these autoencoders applied on training data

OOD Detection on the Edge

pass a given data through the neural network model to compare with the stored model fingerprints for determining abnormality.

Specifically, a data record $\boldsymbol{x}^*$ would be classified as Outlier if its reconstruction error $e_l^* = \mathcal{L}(a_l(x^*), \hat{a}_l(x^*))$ satisfies

$$\exists\, l, \quad e_l^* > \tau_l \quad \text{or} \quad e_l^* < \mu_l$$ **One-out integration strategy**

Discussion and Future Directions

NeuralFP can serve as a key technique of detecting outliers in practical edge computing scenarios.

1) NeuralFP extracts representative information of the training set by constructing non-linear fingerprints of neural network models.
2) NeuralFP successfully distinguishes the differences between in-distribution data and OOD data, through carefully integrate the fingerprints across multiple layers
3) We have verified the effectiveness of NeuralFP on multiple real-world datasets, showed its advantages over existing detection methods, and provided useful guidelines for parameter selection in practice

Future Directions

1) Investigate the performance of NeuralFP in detecting various types of adversarial outliers
2) Integrate other deep generative models such as Generative Adversarial Networks (GAN)