
Self-Supervised Adversarial Example Detection by Disentangled Representation

Anonymous Author(s)

Affiliation

Address

email

1 This document contains the full experimental results that support the work titled “Self-Supervised
2 Adversarial Example Detection via Disentangled Representation”.

3 In order to verify the efficiency of DRR, we prepare adversarial examples generated by 5 kinds of
4 attack methods (FGSM, BIM, PGD, DeepFool, CW) with 3 different norms (L_1 , L_2 , L_{inf}) over the
5 datasets MNIST, Fashion-MNIST, CIFAR-10.

6 Table 1, Table 3, and Table 5 respectively report the accuracy of the victim models on benign test
7 examples and various adversarial examples. It is clear that DeepFool and CW attacks are stronger
8 than FGSM, BIM, and PGD, and CW is the strongest attack. It is also true that for a specific attack
9 and a given norm, the success rate of the attack is becoming higher as the value of the norm increases
10 (model accuracy becomes lower).

11 Figure 4 depicts some reconstruction results of all the detectors: DRR, HLR-P, HLR-L, HVR-P,
12 HVR-L. It is very clear that when reconstructing adversarial examples, HVR-P and HVR-L generally
13 produce a small Reconstruction Error (RE), which means that the manifold directly draw by the AE is
14 too large (we call this unnecessary generalization capability over adversarial examples). And HLR-P
15 and HLR-L generally produce a reconstructed image with large RE, but sometimes they still fail (for
16 example, rows 5 and 7 of the left part, rows 3 and 4 of the middle part of Fig. 4). These observations
17 indicate that logits (outputted by the victim model) is a good choice to further constrain the volume
18 of the manifold drawn by AE (so to reduce the unnecessary generalization capability over adversarial
19 examples), but it is not always enough. By disentangling the representation of image as semantic
20 and class features (we use logits as our class feature), DRR can mimic the behavior of adversarial
21 examples to even better reduce unnecessary generalization capability of the encoder-decoder network.

22 Figures 1, 3, 2 box-plot the values of REs of all the benign test and adversarial examples for CIFAR-
23 10, MNIST, Fashion-MNIST, respectively. From these figures, it is clear REs from HVR-P, HVR-L
24 are most of the time mixed. For HLR-P and HLR-L, if we look at specific attack on a fixed dataset,
25 it seems that it works as good as DRR. However, to put detector into real usage, it should have a
26 universal threshold value for all the possible attacks against a fixed dataset. From this view, HLR-P
27 and HLR-L perform worse than DRR, even if we can use another offline technique to learn this
28 threshold value (by assuming we know all the possible attacks).

29 For quantitative comparison, Tables 2, 4, 6 list all the AUC values of the ROC of the detectors with
30 respect to all the attacks over the three datasets. The general trend in these three tables is, the large
31 the value of the norm (used for the attack), the large the AUC value, i.e., attacks with larger norm
32 can be detected easier. And it is also clear that DRR has the highest value of AUC over all the three
33 datasets. Specifically, for CIFAR-10 (Table 2), DRR has an AUC value larger than 0.99, which is
34 close to the ideal value 1.

Table 1: Test accuracy of the victim models on CIFAR-10.

Attack	Norm	VGG
	benign	0.869
BIM	$L_1 \epsilon = 5$	0.588
	$L_1 \epsilon = 10$	0.298
	$L_2 \epsilon = 0.1$	0.702
	$L_2 \epsilon = 0.3$	0.252
	$L_{inf} \epsilon = 0.005$	0.500
	$L_{inf} \epsilon = 0.01$	0.195
PGD	$L_1 \epsilon = 5$	0.665
	$L_1 \epsilon = 10$	0.411
	$L_2 \epsilon = 0.1$	0.744
	$L_2 \epsilon = 0.3$	0.352
	$L_{inf} \epsilon = 0.005$	0.553
	$L_{inf} \epsilon = 0.01$	0.267
FGSM	$L_1 \epsilon = 5$	0.651
	$L_1 \epsilon = 10$	0.457
	$L_2 \epsilon = 0.3$	0.431
	$L_2 \epsilon = 1$	0.170
	$L_{inf} \epsilon = 0.01$	0.352
	$L_{inf} \epsilon = 0.05$	0.092
DeepFool	L_2	0.039
	L_{inf}	0.054
CW	L_2	0.001

Table 2: AUC of ROC of different detectors over CIFAR-10.

Attack	Norm	VGG				
		HVR-P	HLR-P	HVR-L	HLR-L	DRR (ours)
BIM	$L_1 \epsilon = 5$	0.4380	0.5996	0.4364	0.8515	0.9975
	$L_1 \epsilon = 10$	0.4533	0.6173	0.4556	0.8806	0.9985
	$L_2 \epsilon = 0.1$	0.4312	0.5947	0.4358	0.8516	0.9981
	$L_2 \epsilon = 0.3$	0.4650	0.6231	0.4670	0.8892	0.9990
	$L_{inf} \epsilon = 0.005$	0.4549	0.5973	0.4554	0.8610	0.9981
	$L_{inf} \epsilon = 0.01$	0.4755	0.6340	0.4766	0.9039	0.9992
PGD	$L_1 \epsilon = 5$	0.4313	0.5981	0.4286	0.8516	0.9982
	$L_1 \epsilon = 10$	0.4436	0.6020	0.4446	0.8644	0.9985
	$L_2 \epsilon = 0.1$	0.4411	0.5992	0.4469	0.8568	0.9988
	$L_2 \epsilon = 0.3$	0.4594	0.6105	0.4613	0.8742	0.9983
	$L_{inf} \epsilon = 0.005$	0.4530	0.5978	0.4546	0.8606	0.9982
	$L_{inf} \epsilon = 0.01$	0.4704	0.6199	0.4723	0.8856	0.9990
FGSM	$L_1 \epsilon = 5$	0.4347	0.5948	0.4350	0.8560	0.9981
	$L_1 \epsilon = 10$	0.4535	0.6134	0.4554	0.8550	0.9982
	$L_2 \epsilon = 0.3$	0.4622	0.6139	0.4639	0.8571	0.9977
	$L_2 \epsilon = 1$	0.5146	0.6544	0.5157	0.8709	0.9981
	$L_{inf} \epsilon = 0.01$	0.4841	0.6205	0.4854	0.8620	0.9980
	$L_{inf} \epsilon = 0.05$	0.6620	0.7000	0.6430	0.8207	0.9977
DeepFool	L_2	0.5083	0.6621	0.5098	0.8251	0.9953
	L_{inf}	0.5051	0.6518	0.5038	0.8167	0.9954
CW	L_2	0.5020	0.6562	0.5019	0.8298	0.9982

Table 3: Test accuracy of the victim models over Fashion-MNIST.

Attack	Norm	CNN
	benign	0.926
BIM	$L_1 \epsilon = 20$	0.300
	$L_1 \epsilon = 50$	0.290
	$L_2 \epsilon = 2$	0.322
	$L_2 \epsilon = 5$	0.313
	$L_{inf} \epsilon = 0.01$	0.587
	$L_{inf} \epsilon = 0.05$	0.021
PGD	$L_1 \epsilon = 10$	0.294
	$L_1 \epsilon = 20$	0.207
	$L_2 \epsilon = 1$	0.240
	$L_2 \epsilon = 2$	0.095
	$L_{inf} \epsilon = 0.01$	0.633
	$L_{inf} \epsilon = 0.05$	0.008
FGSM	$L_1 \epsilon = 20$	0.509
	$L_1 \epsilon = 50$	0.425
	$L_2 \epsilon = 5$	0.437
	$L_{inf} \epsilon = 0.01$	0.731
	$L_{inf} \epsilon = 0.05$	0.327
DeepFool	L_2	0.233
	L_{inf}	0.052
CW	L_2	0.003

Table 4: AUC of ROC of different detectors over Fashion-MNIST.

Attack	Norm	CNN				
		HVR-P	HLR-P	HVR-L	HLR-L	DRR (ours)
BIM	$L_1 \epsilon = 20$	0.5420	0.8311	0.5481	0.8419	0.8944
	$L_1 \epsilon = 50$	0.6018	0.8484	0.6253	0.8591	0.9269
	$L_2 \epsilon = 2$	0.5716	0.8004	0.5800	0.8244	0.9050
	$L_2 \epsilon = 5$	0.7015	0.8391	0.7254	0.8563	0.9248
	$L_{inf} \epsilon = 0.01$	0.4920	0.9210	0.4917	0.9323	0.9596
	$L_{inf} \epsilon = 0.05$	0.5255	0.8295	0.5389	0.8419	0.8880
PGD	$L_1 \epsilon = 10$	0.5177	0.9532	0.5183	0.9592	0.9830
	$L_1 \epsilon = 20$	0.5658	0.9654	0.5797	0.9679	0.9852
	$L_2 \epsilon = 1$	0.5614	0.9610	0.5728	0.9662	0.9858
	$L_2 \epsilon = 2$	0.6972	0.9654	0.7259	0.9709	0.9914
	$L_{inf} \epsilon = 0.01$	0.4981	0.9259	0.4997	0.9347	0.9589
	$L_{inf} \epsilon = 0.05$	0.5267	0.8439	0.5415	0.8555	0.9000
FGSM	$L_1 \epsilon = 20$	0.6091	0.9362	0.6147	0.9455	0.9687
	$L_1 \epsilon = 50$	0.8390	0.9469	0.8489	0.9574	0.9825
	$L_2 \epsilon = 5$	0.9628	0.9068	0.9673	0.9008	0.9010
	$L_{inf} \epsilon = 0.01$	0.5281	0.9183	0.5253	0.9270	0.9536
	$L_{inf} \epsilon = 0.05$	0.6113	0.9201	0.6298	0.9193	0.9192
DeepFool	L_2	0.6204	0.9333	0.6267	0.9516	0.9696
	L_{inf}	0.6137	0.9088	0.6270	0.9273	0.9378
CW	L_2	0.5068	0.8981	0.5013	0.9058	0.9125

Table 5: Test accuracy of the victim models over MNIST.

Attack	Norm	CNN
	benign	0.993
BIM	$L_1 \epsilon = 20$	0.277
	$L_1 \epsilon = 50$	0.012
	$L_2 \epsilon = 2$	0.061
	$L_2 \epsilon = 5$	0.002
	$L_{inf} \epsilon = 0.1$	0.407
	$L_{inf} \epsilon = 0.3$	0.001
PGD	$L_1 \epsilon = 10$	0.882
	$L_1 \epsilon = 20$	0.503
	$L_2 \epsilon = 1$	0.724
	$L_2 \epsilon = 2$	0.086
	$L_{inf} \epsilon = 0.1$	0.443
	$L_{inf} \epsilon = 0.3$	0.000
FGSM	$L_1 \epsilon = 20$	0.800
	$L_1 \epsilon = 50$	0.340
	$L_2 \epsilon = 2$	0.596
	$L_2 \epsilon = 5$	0.173
	$L_{inf} \epsilon = 0.1$	0.762
DeepFool	L_2	0.080
	L_{inf}	0.078
CW	L_2	0.000

Table 6: AUC of ROC of different detectors over MNIST.

Attack	Norm	CNN				
		HVR-P	HLR-P	HVR-L	HLR-L	DRR (ours)
BIM	$L_1 \epsilon = 20$	0.3823	0.9960	0.4183	0.9931	0.9955
	$L_1 \epsilon = 50$	0.7572	0.8876	0.7391	0.8968	0.9714
	$L_2 \epsilon = 2$	0.4037	0.9978	0.4449	0.9925	0.9981
	$L_2 \epsilon = 5$	0.9119	0.9156	0.8904	0.9124	0.9790
	$L_{inf} \epsilon = 0.1$	0.0946	0.9780	0.1057	0.9806	0.9864
	$L_{inf} \epsilon = 0.3$	0.4711	0.8509	0.3972	0.8672	0.9939
PGD	$L_1 \epsilon = 10$	0.3006	0.9959	0.3363	0.9939	0.9741
	$L_1 \epsilon = 20$	0.2285	0.9926	0.2772	0.9894	0.9895
	$L_2 \epsilon = 1$	0.2313	0.9897	0.2633	0.9886	0.9691
	$L_2 \epsilon = 2$	0.1645	0.9948	0.2149	0.9915	0.9977
	$L_{inf} \epsilon = 0.1$	0.0796	0.9752	0.0946	0.9795	0.9854
	$L_{inf} \epsilon = 0.3$	0.5226	0.8291	0.4350	0.8551	0.9970
FGSM	$L_1 \epsilon = 20$	0.3609	0.9856	0.3771	0.9849	0.9557
	$L_1 \epsilon = 50$	0.5697	0.9765	0.5585	0.9753	0.9819
	$L_2 \epsilon = 2$	0.3968	0.9766	0.4059	0.9767	0.9647
	$L_2 \epsilon = 5$	0.8534	0.9810	0.8134	0.9810	0.9916
	$L_{inf} \epsilon = 0.1$	0.0724	0.9352	0.0811	0.9569	0.9367
DeepFool	L_2	0.6801	0.9921	0.6879	0.9855	0.9694
	L_{inf}	0.3824	0.9653	0.4301	0.9748	0.9815
CW	L_2	0.7211	0.9871	0.6947	0.9828	0.9750

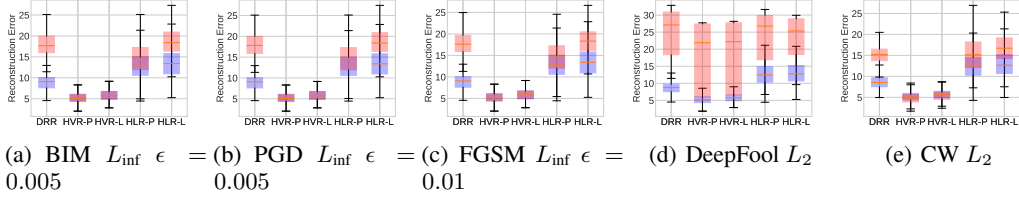


Figure 1: Box-plot of the reconstruction errors for CIFAR-10 (red box is for adversarial and blue box is for benign).

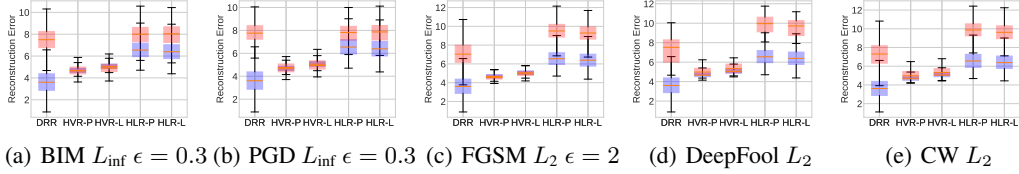


Figure 2: Box-plot of the reconstruction errors for MNIST (red box is for adversarial and blue box is for benign).

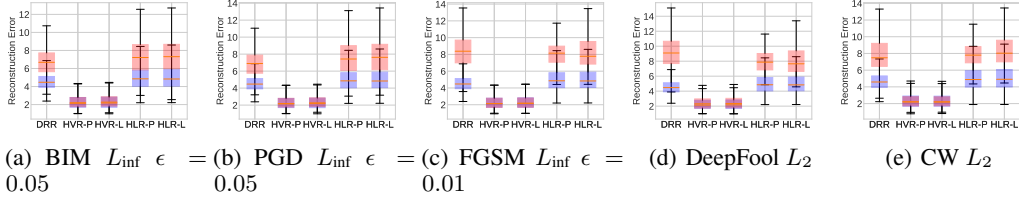


Figure 3: Box-plot of the reconstruction errors for Fashion-MNIST (red box is for adversarial and blue box is for benign).

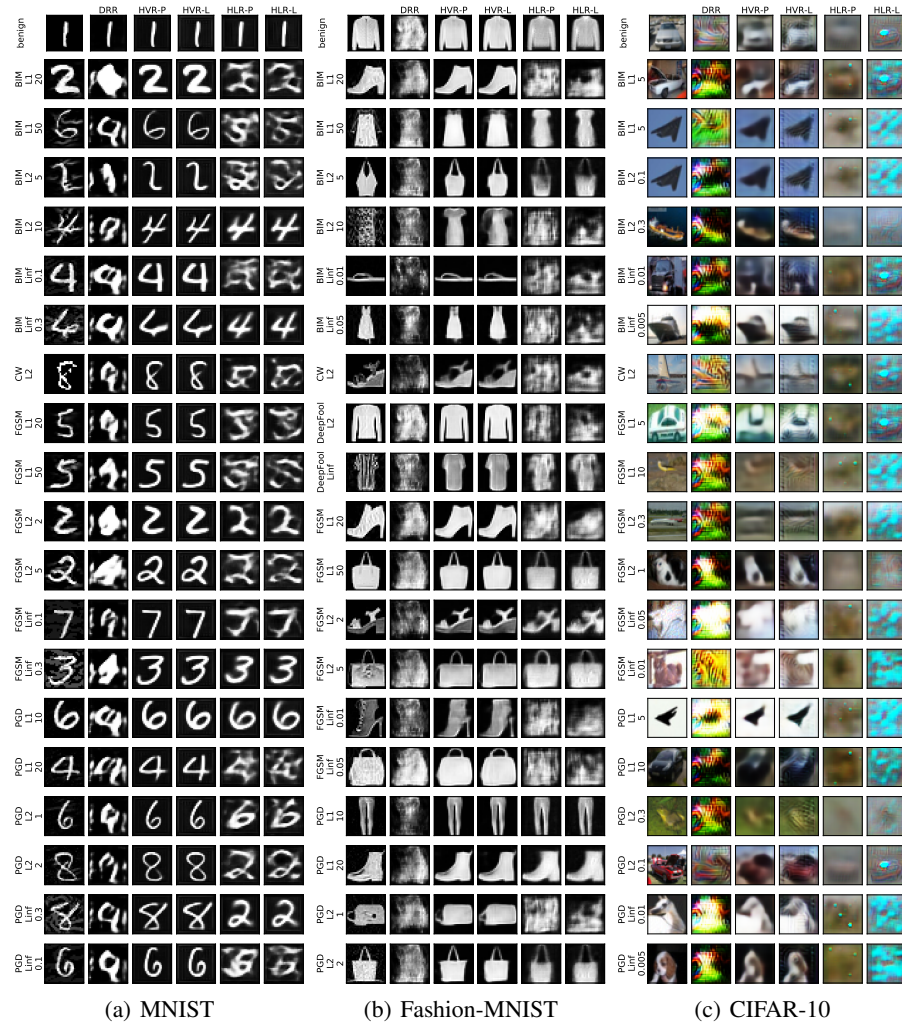


Figure 4: Reconstructed images of each detection methods over different datasets.