The Supplementary Material for ML-LOO: Detecting Adversarial Examples with Feature Attribution

Puyudi Yang,[1] Jianbo Chen,[2] Cho-Jui Hsieh,[3] Jane-Ling Wang,[1] Michael I. Jordan,[2]
[1]University of California, Davis
[2]University of California, Berkeley
[3]University of California, Los Angeles

Abstract

We evaluate the performance of ML-LOO under another method for constructing feature attribution maps: Integrated Gradients. We carry out a comparison between three dispersion measures. Finally, we plot the ROC curves of detection methods on all attacks, data sets and models, both the augmented curves when FPR ranging from 0.0 to 0.2 and the full curves.

Performance of Integrated Gradients

In this section, we evaluate the detection of adversarial examples by thresholding the IQR of another popular feature attribution method Integrated Gradients (IG), and compare it with KD+BU, LID, MAHA, and ML-LOO. We consider three attacks FGSM, C&W and JSMA, which are optimized for L_∞, L_2 and L_0 distances respectively, on CIFAR-10 with ResNet. We can see that IQR of IG achieves competitive performance in detecting adversarial examples, but not as powerful as the detection methods which incorporated multi-layer information like LID, MAHA and our proposed method ML-LOO. The IG feature is also not as effective as the LOO feature (whose performance is shown in Figure 3).

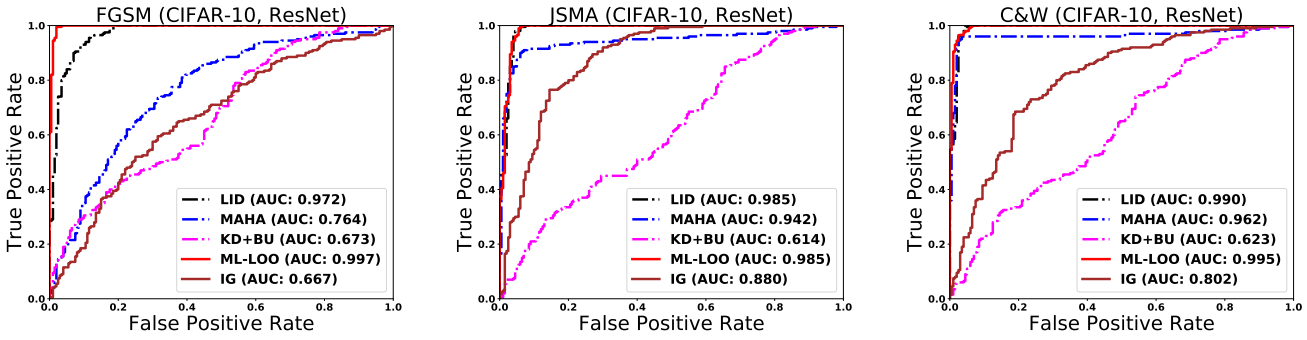

Figure 1: ROC curves of detection methods on CIFAR-10 with ResNet. We restrict FPR between 0 and 0.2, which is meaningful in practice. See Appendix for full plots.

Comparison Based on Dispersion Measures

In this section, we compare performance of detection using three different dispersion measures of feature attributions: IQR, STD and MAD.

Figure 2 shows the histograms of these three dispersion measures of feature attributions for ResNet on natural test images from CIFAR-10 with those on adversarially perturbed images, where the adversarial perturbation is carried out by C&W Attack. We can see there is a significant difference in the distributions of the dispersion measures between natural and adversarial images.

Figure 3 shows the ROC curves of the three dispersion statistics on CIFAR-10 with ResNet. We can see that all three dispersion measures achieve competitive performance in detecting adversarial examples generated by three attacks C&W, JSMA and L_∞-PGD, but IQR achieves the largest AUC values across all attacking methods.

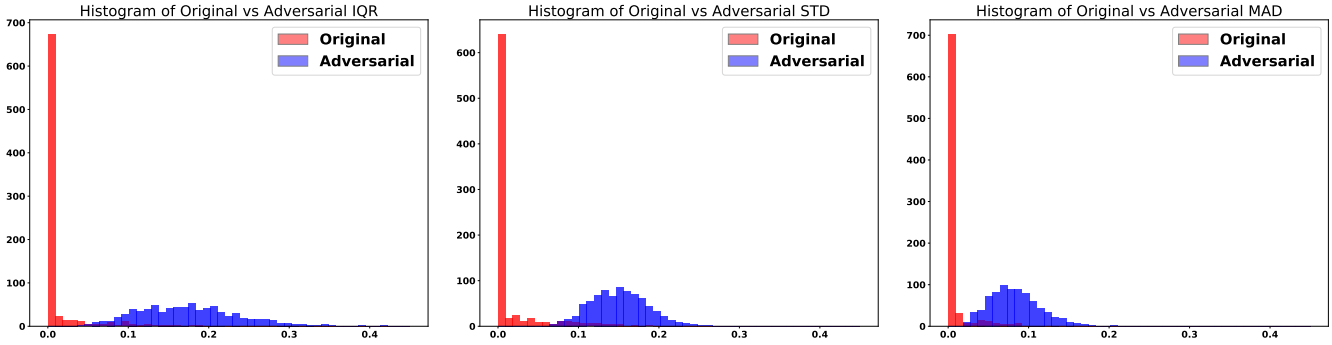

Figure 2: Histogram of Statistics

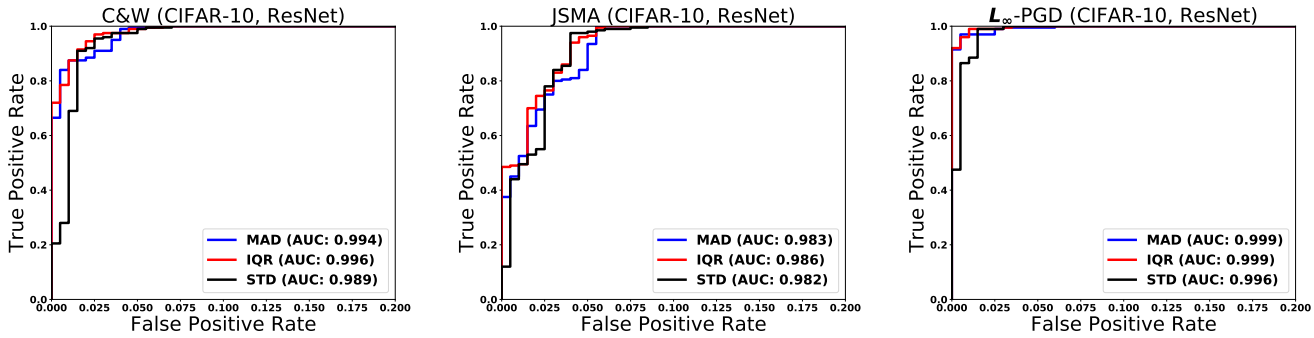

Figure 3: ROC curves of different statistics on CIFAR-10 with ResNet

ROC curves on CIFAR-10, MNIST and CIFAR-100 with FPR from 0.0 to 0.2

Figure 4, Figure 5, Figure 6, Figure 7, and Figure 8 show the ROC curves of four detection method (LID, MAHA, KD+BU, ML-LOO) on three data sets (CIFAR-10, MNIST, CIFAR-100) with three models (CNN, ResNet, DenseNet) under six attacks (FGSM, JSMA, C&W, DeepFool, Boundary, L_∞-PGD) where FPR is from 0.0 to 0.2, which is the setting of practical interest. The ROC curves where FPR is from 0.0 to 1.0 are shown in Appendix .

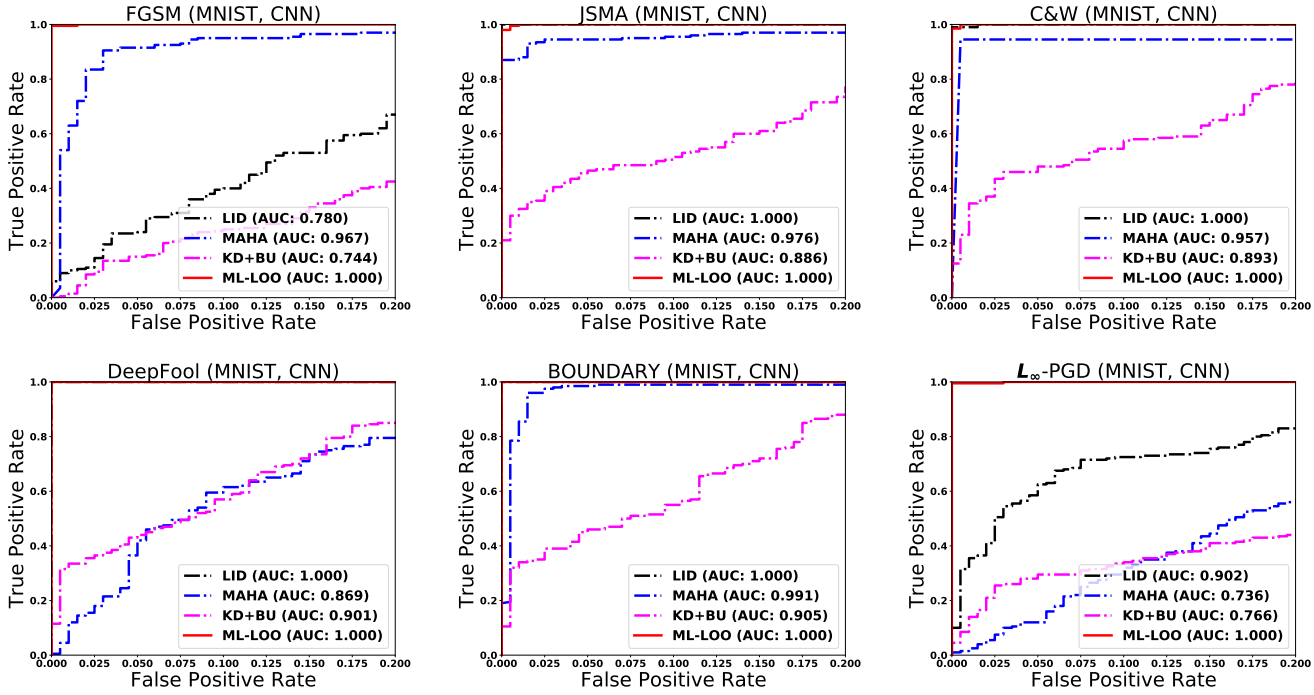

Figure 4: ROC curves of detection methods on MNIST with CNN

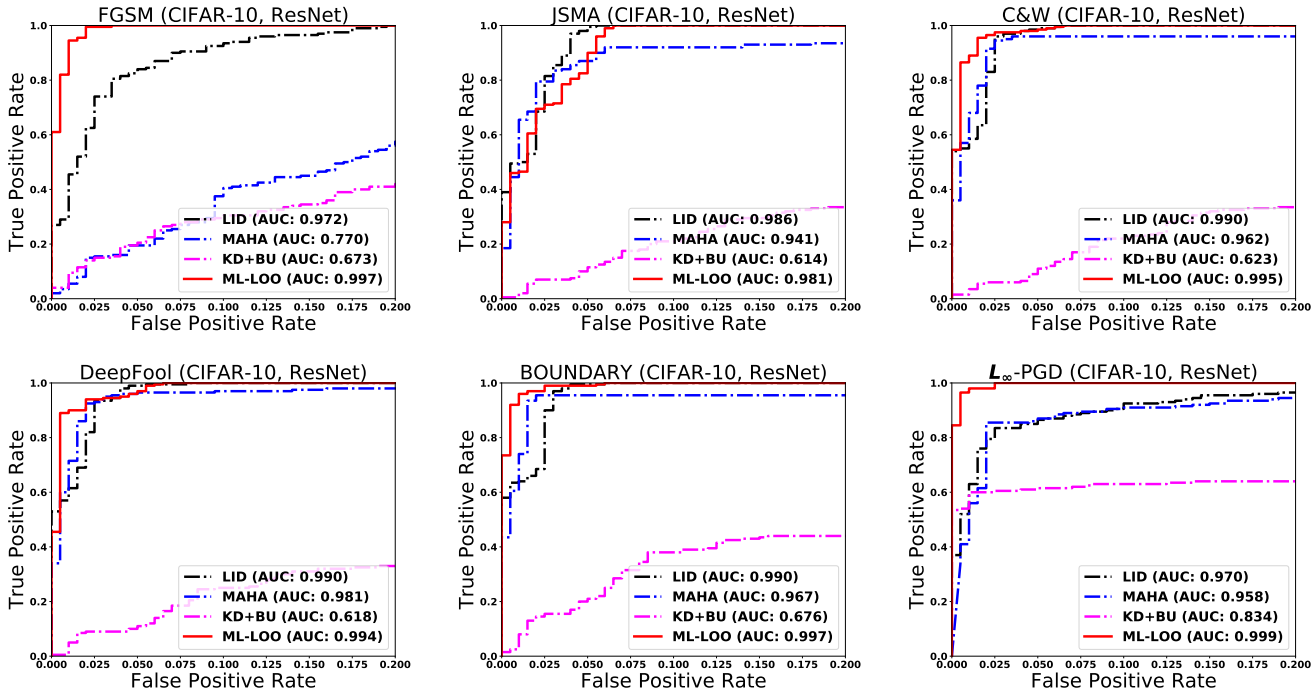

Figure 5: ROC curves of detection methods on CIFAR-10 with ResNet

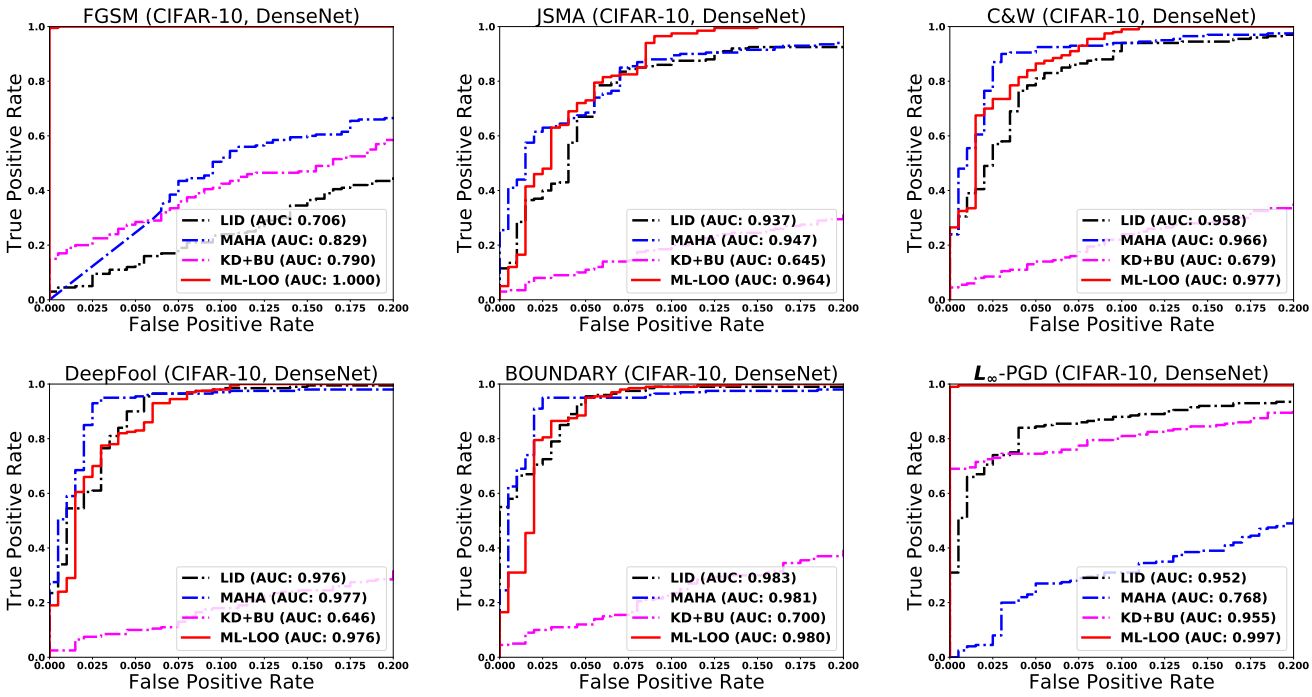

Figure 6: ROC curves of detection methods on CIFAR-10 with DenseNet

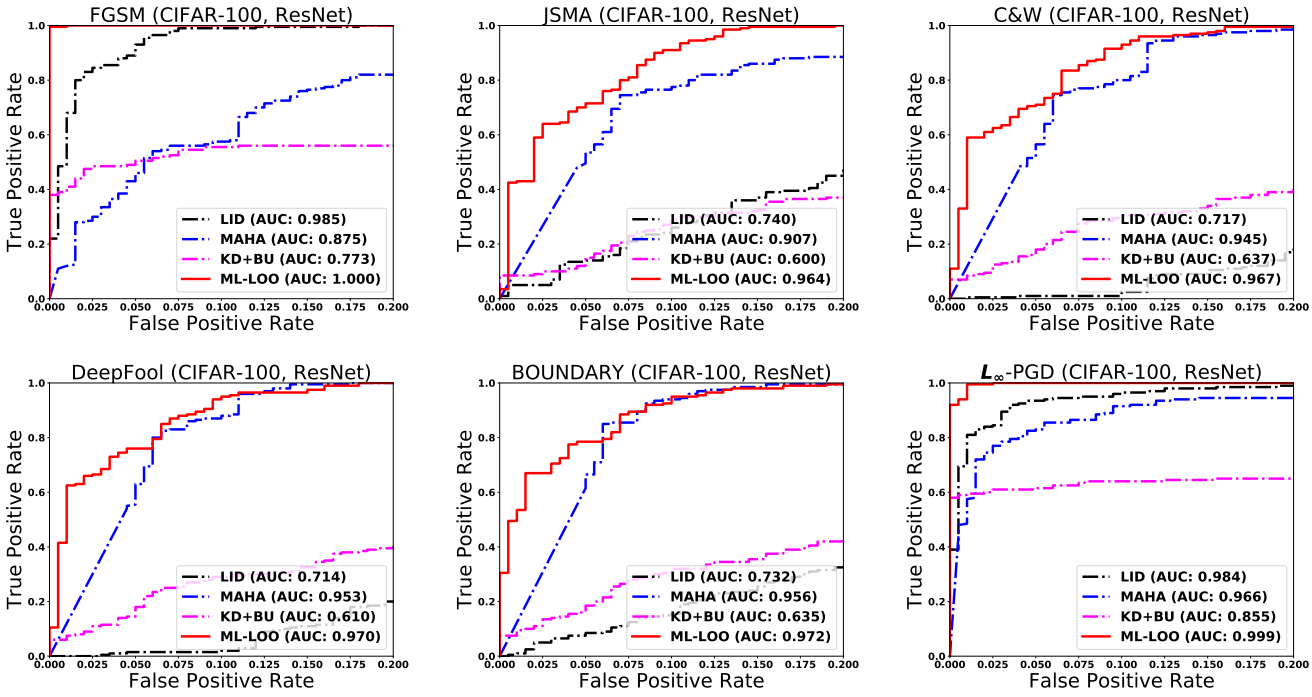

Figure 7: ROC curves of detection methods on CIFAR-100 with ResNet

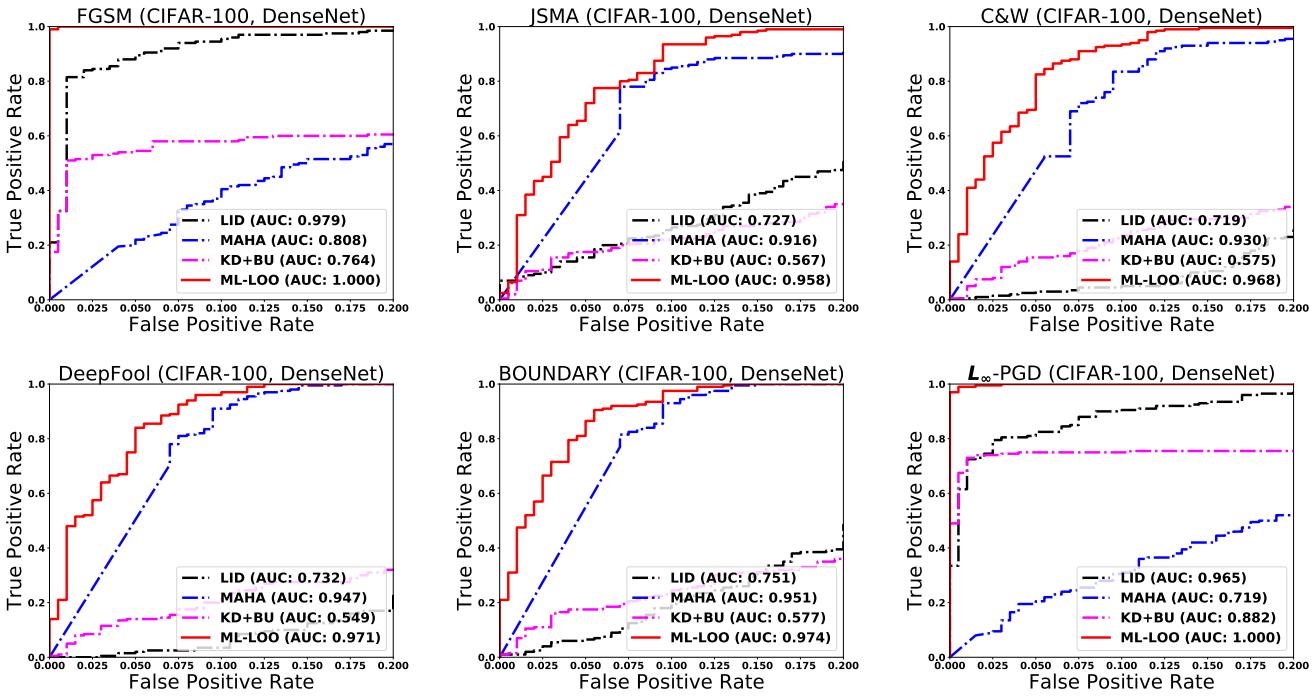

Figure 8: ROC curves of detection methods on CIFAR-100 with DenseNet

ROC curves on CIFAR-10, MNIST and CIFAR-100 with FPR from 0.0 to 1.0

In this section, we show the ROC curves of four detection method (LID, MAHA, KD+BU, ML-LOO) on three data sets (CIFAR-10, MNIST, CIFAR-100) with three models (CNN, ResNet, DenseNet) under six attacks (FGSM, JSMA, C&W, DeepFool, Boundary, L_∞-PGD) where FPR is from 0.0 to 1.0.

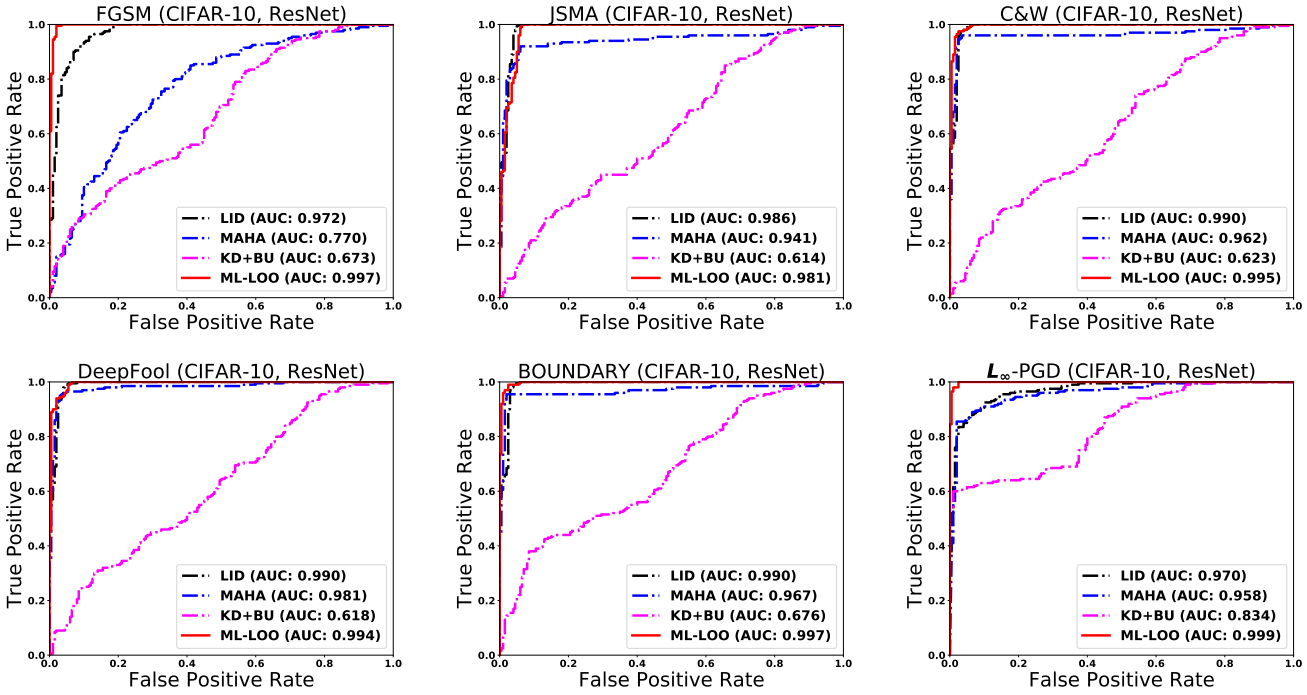

Figure 9: ROC curves of detection methods on CIFAR-10 with ResNet

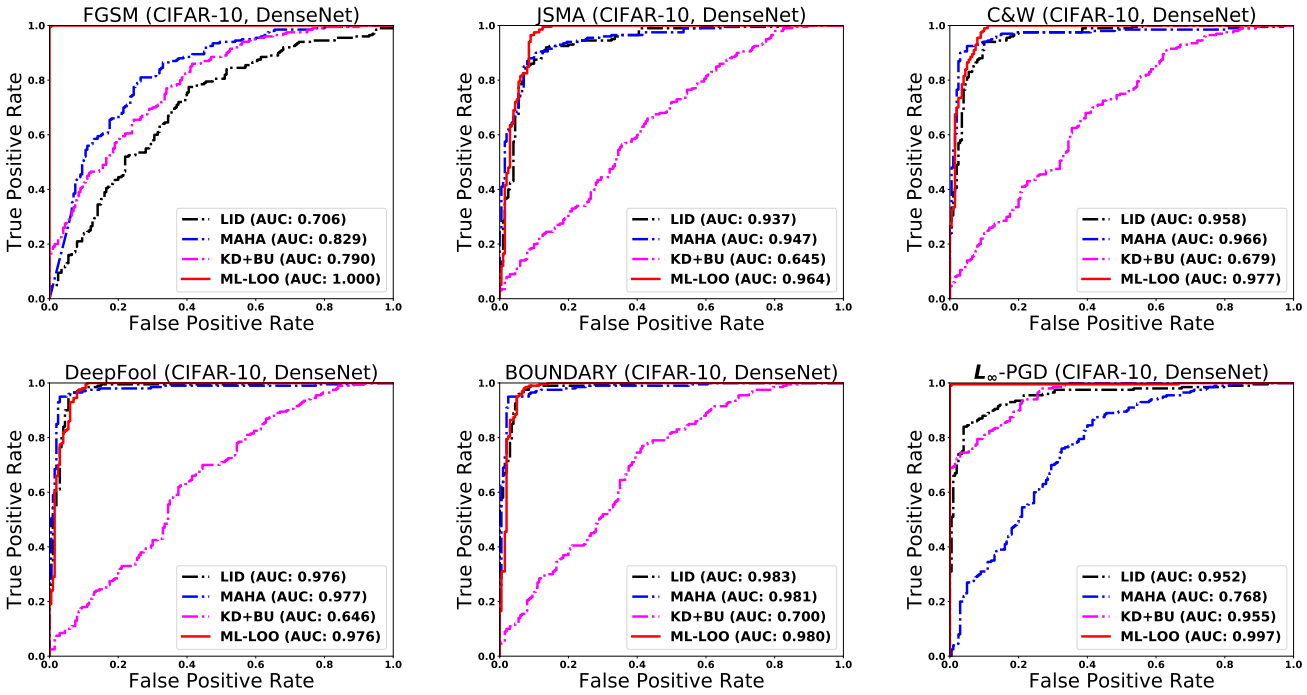

Figure 10: ROC curves of detection methods on CIFAR-10 with DenseNet

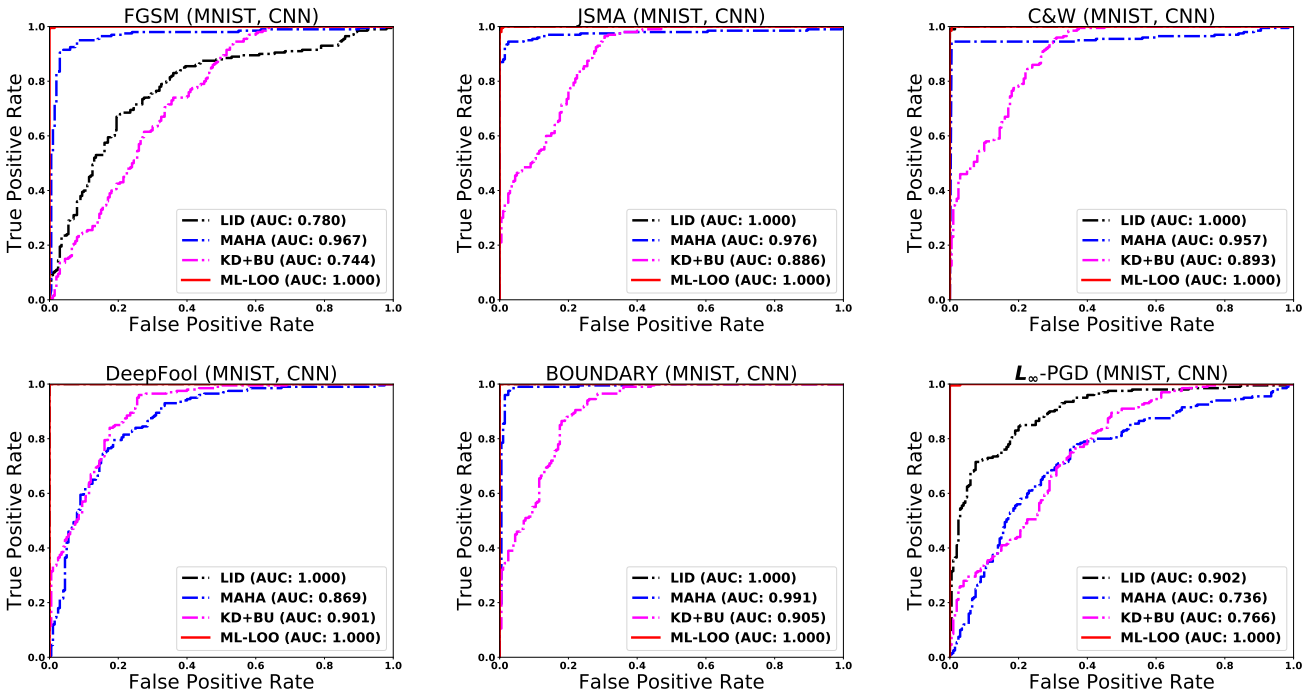

Figure 11: ROC curves of detection methods on MNIST with CNN

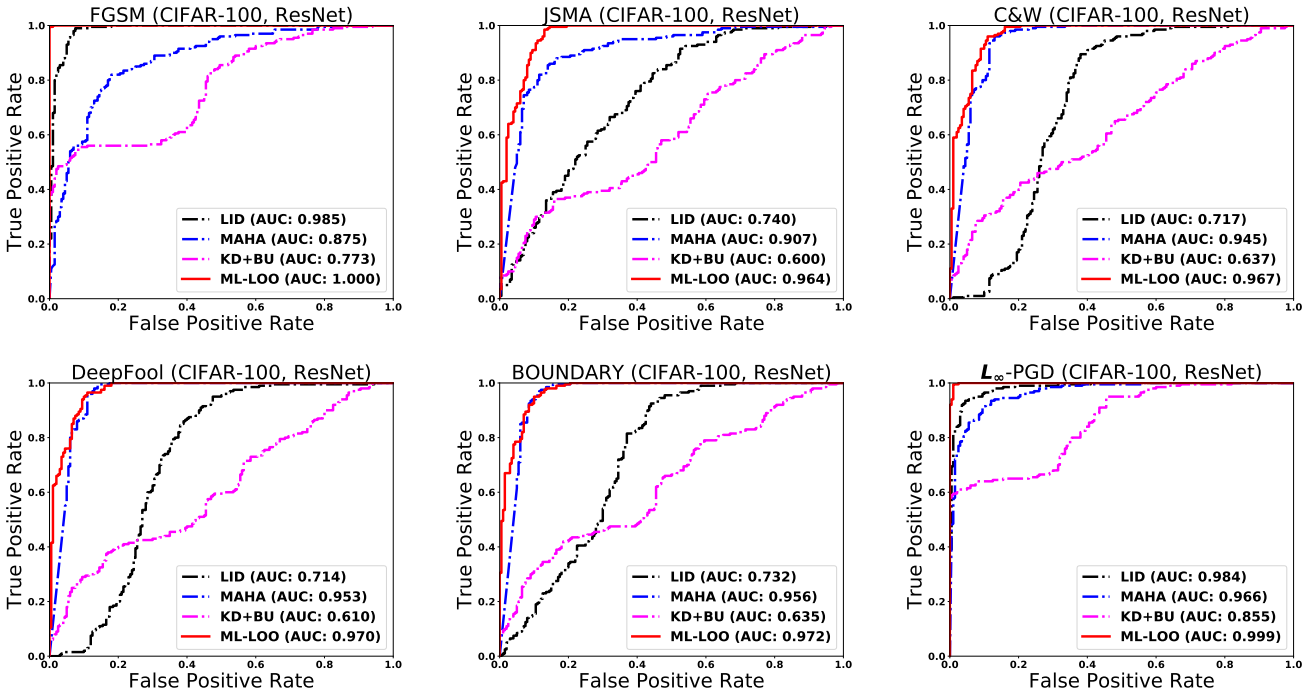

Figure 12: ROC curves of detection methods on CIFAR-100 with ResNet

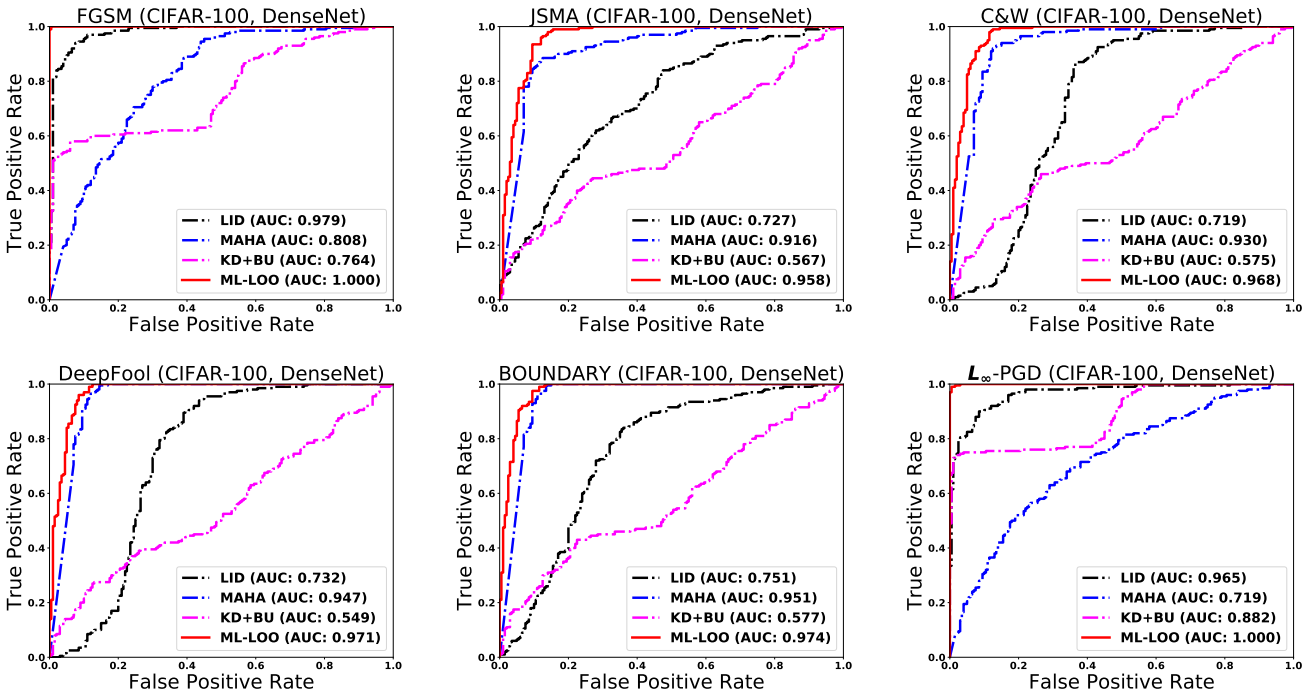

Figure 13: ROC curves of detection methods on CIFAR-100 with DenseNet