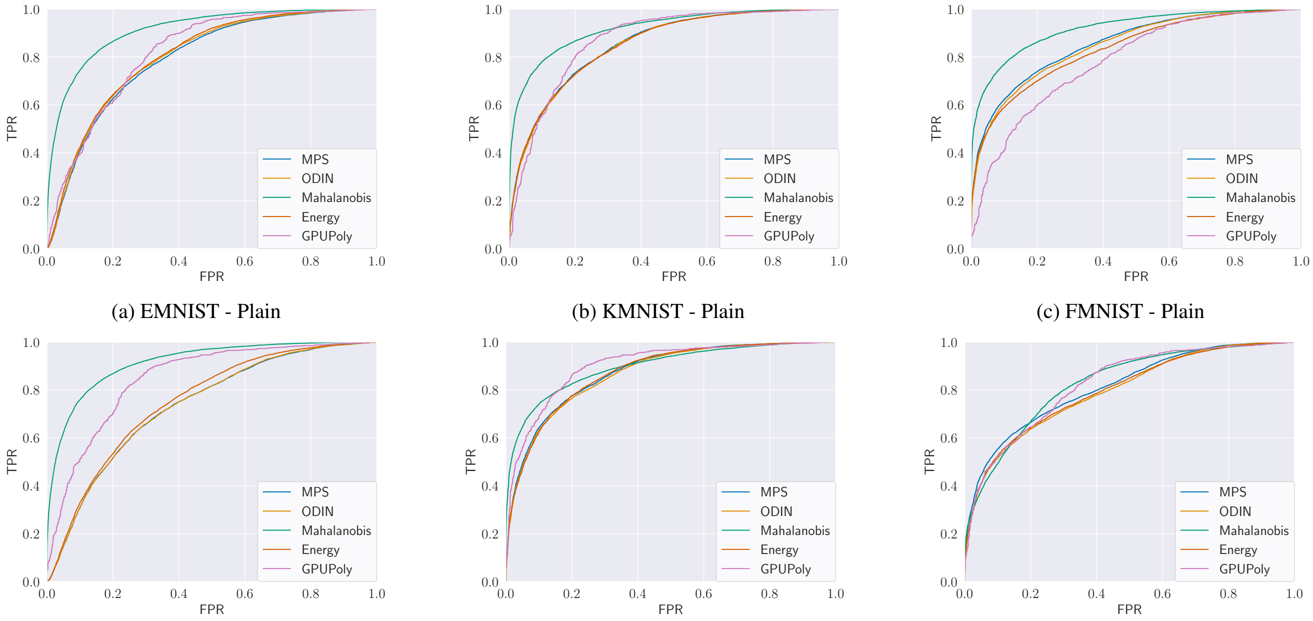
# Understanding ReLU Network Robustness Through Test Set Certification Performance

Supplementary Material

## 6. Extension to the out-of-distribution experimental analysis.

Here, we extend our evaluation of robustness certificates for OOD detection. Below, we show ROC curves and other results for different trainings, networks and datasets. As previously mentioned, GPUPoly is used with $\ell_\infty$-norm robustness certificates and a range of 4 000 values of $\epsilon$ between 0 and 0.2. Instead, all other methods uses a range of $10e5$ samples for the threshold.

### 6.1. Grayscale Category Extension



(a) EMNIST - Plain

(b) KMNIST - Plain

(c) FMNIST - Plain

(d) EMNIST - Randomized

(e) KMNIST - Randomized

(f) FMNIST - Randomized

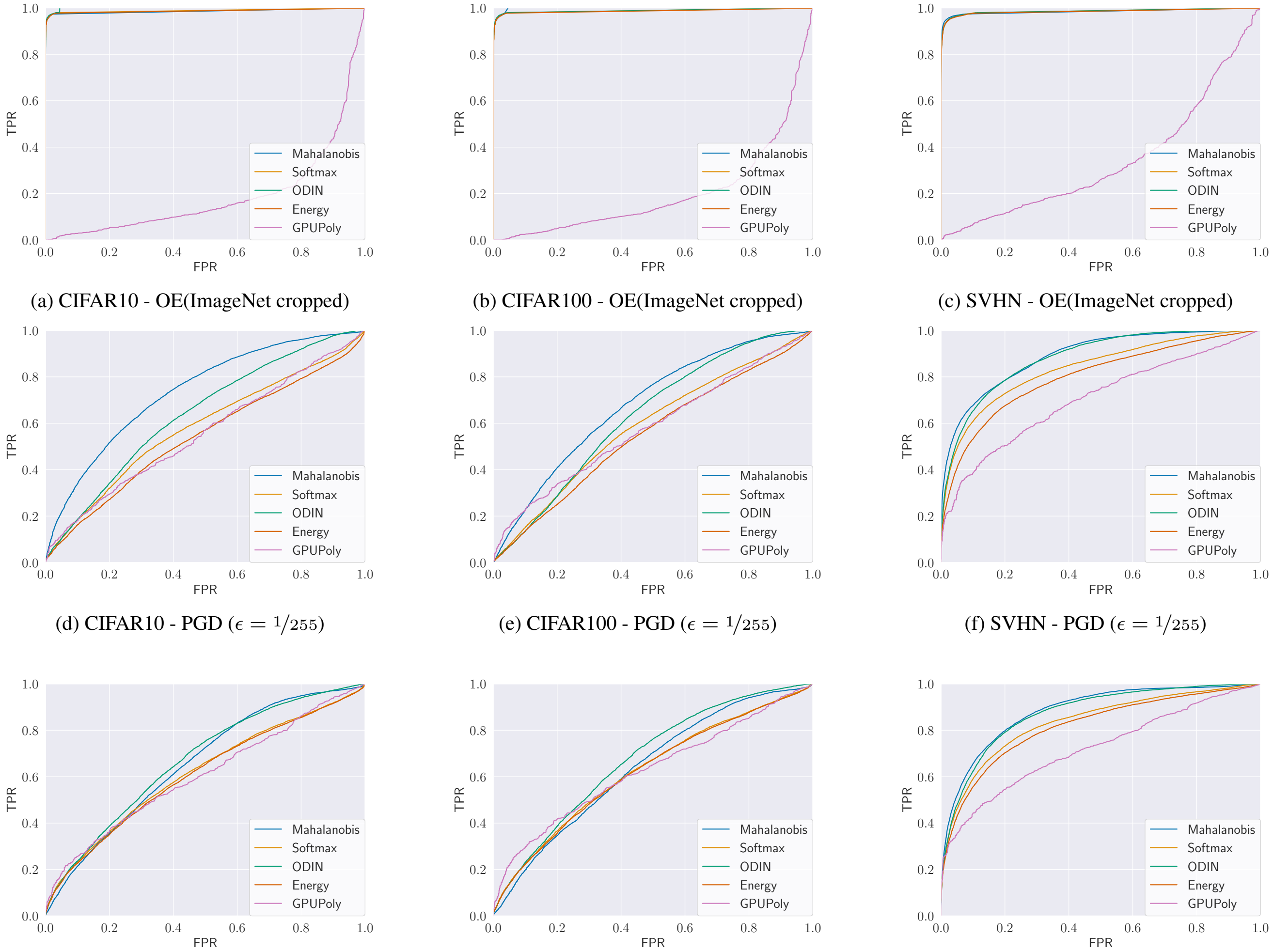
Figure 6. **ConvSmall - ID:MNIST** Comparison of ROC curves for OOD detection methods on KMNIST, EMNIST datasets. GPUPoly is used with $\ell_\infty$ robustness certificates and a range of 4000 values of $\epsilon$ between 0 and 0.2. All other methods uses a range of $10e5$ for the threshold.

In Fig. 6, we report the ROC curves for convolutional networks trained normally and with randomized smoothing on the MNIST dataset. We observe that GPUPoly($\ell_\infty$) perform relatively better than standard OOD detection methods on EMNIST and KMNIST.

### 6.2. Colored Category Extension

Here, we test two convolutional networks of different sizes: ConvSmall and ConvMed, trained on CIFAR10. In Fig. 7, we show the ROC curves for the ConvMed model trained with OE (ImageNet cropped), PGD and randomized smoothing. In the context of OE, we clearly see that the ROC curves of GPUPoly are below the average value of 0.5. Besides his average performance on PGD and randomized trained networks, the results graphically demonstrate that OOD samples are more likely to be certified than ID samples for OOD aware networks.

In Tab. 7, we report the results for the RGB category with CIFAR10 as ID dataset. In this evaluation, we compare PGD, FGSM and normally trained convolutional networks. Similarly to the networks trained on the GTSRB dataset, we observe



(a) CIFAR10 - OE(ImageNet cropped)

(b) CIFAR100 - OE(ImageNet cropped)

(c) SVHN - OE(ImageNet cropped)

(d) CIFAR10 - PGD ($\epsilon = {}^{1}/_{255}$)

(e) CIFAR100 - PGD ($\epsilon = {}^{1}/_{255}$)

(f) SVHN - PGD ($\epsilon = {}^{1}/_{255}$)

(g) CIFAR10 - Randomized ($\sigma = 0.1$)

(h) CIFAR100 - Randomized ($\sigma = 0.1$)

(i) SVHN - Randomized ($\sigma = 0.1$)

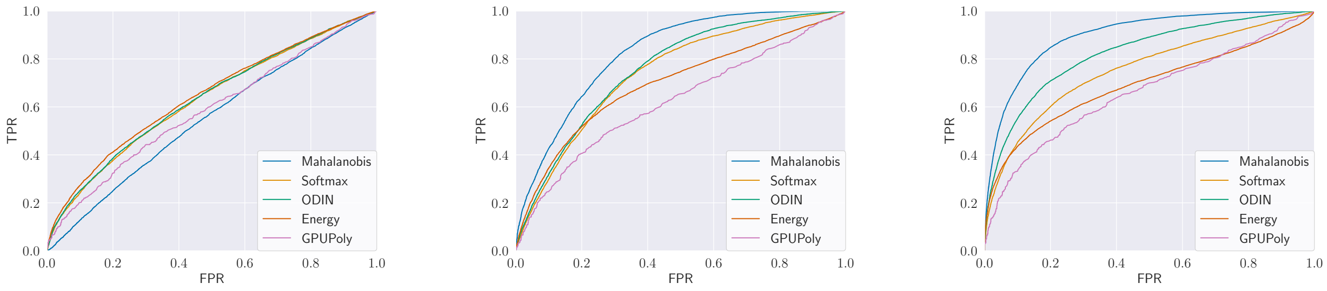
Figure 7. **ConvMed - ID: GTSRB.** Comparison of ROC curves for OOD detection methods on CIFAR10/100 and SVHN datasets. GPUPoly is used with $\ell_\infty$ robustness certificates and a range of 4 000 values of $\epsilon$ between zero and 0.2. All other methods uses a range of $10e5$ for the threshold.

similar performances as standard OOD detection methods. As previously mentioned, the accuracy is generally low compared to state-of-the-art networks, but is aligned with related work on verification methods [32, 34].

In Fig. 8 and Fig. 9, we report the ROC curves for ConvSmall and ConvMed, respectively. All networks are trained with CIFAR10 as ID dataset.

Table 7. **ID: CIFAR10.** Comparison between standard OOD detection methods and robustness certificates of $\ell_\infty$-norm: GPUPoly($\ell_\infty$) [32]. We report the clean accuracy (ACC) on the in-distribution (ID) dataset: CIFAR10. In the context of GPUPoly, the AUC and FPR95 are computed by varying the adversarial power $\epsilon$.

| **Network/ Training** | **ACC** | **Method** | **CIFAR100** | | **GTSRB** | | **SVHN** | |
|---|---|---|---|---|---|---|---|---|
| | | | **AUC↑** | **FPR95↓** | **AUC↑** | **FPR95↓** | **AUC↑** | **FPR95↓** |
| ConvSmall/ Plain | 58.0 | Mahalanobis | 54.9 | 93.2 | **82.4** | **51.4** | **89.8** | **41.9** |
| | | Softmax | 63.0 | **90.3** | 74.1 | 75.6 | 75.4 | 85.8 |
| | | ODIN | 63.3 | **90.3** | 75.6 | 68.6 | 82.4 | 70.0 |
| | | Energy | **64.5** | 90.4 | 69.6 | 91.4 | 69.2 | 96.1 |
| | | GPUPoly($\ell_\infty$) | 58.3 | 91.3 | 62.1 | 91.4 | 65.8 | 91.2 |
| ConvMed/ FGSM ($\epsilon = 1/255$) | 57.3 | Mahalanobis | 57.7 | 92.4 | **77.7** | **65.5** | **88.0** | **55.2** |
| | | Softmax | 64.8 | 89.7 | 70.1 | 80.9 | 69.9 | 90.2 |
| | | ODIN | **65.3** | **89.1** | 68.0 | 86.2 | 77.2 | 78.5 |
| | | Energy | 64.3 | 90.5 | 70.3 | 85.6 | 64.5 | 96.2 |
| | | GPUPoly($\ell_\infty$) | 58.8 | 95.1 | 64.4 | 92.2 | 66.0 | 93.3 |
| ConvMed/ PGD ($\epsilon = 1/255$) | 56.1 | Mahalanobis | 57.6 | 92.4 | **80.4** | **57.4** | **91.5** | **41.3** |
| | | Softmax | 64.7 | **89.8** | 75.6 | 77.5 | 66.3 | 92.1 |
| | | ODIN | **64.9** | 89.9 | 75.2 | 69.1 | 79.2 | 73.9 |
| | | Energy | 64.1 | 90.8 | 70.8 | 89.9 | 59.0 | 98.1 |
| | | GPUPoly($\ell_\infty$) | 58.4 | 92.1 | 66.2 | 89.7 | 68.0 | 90.3 |

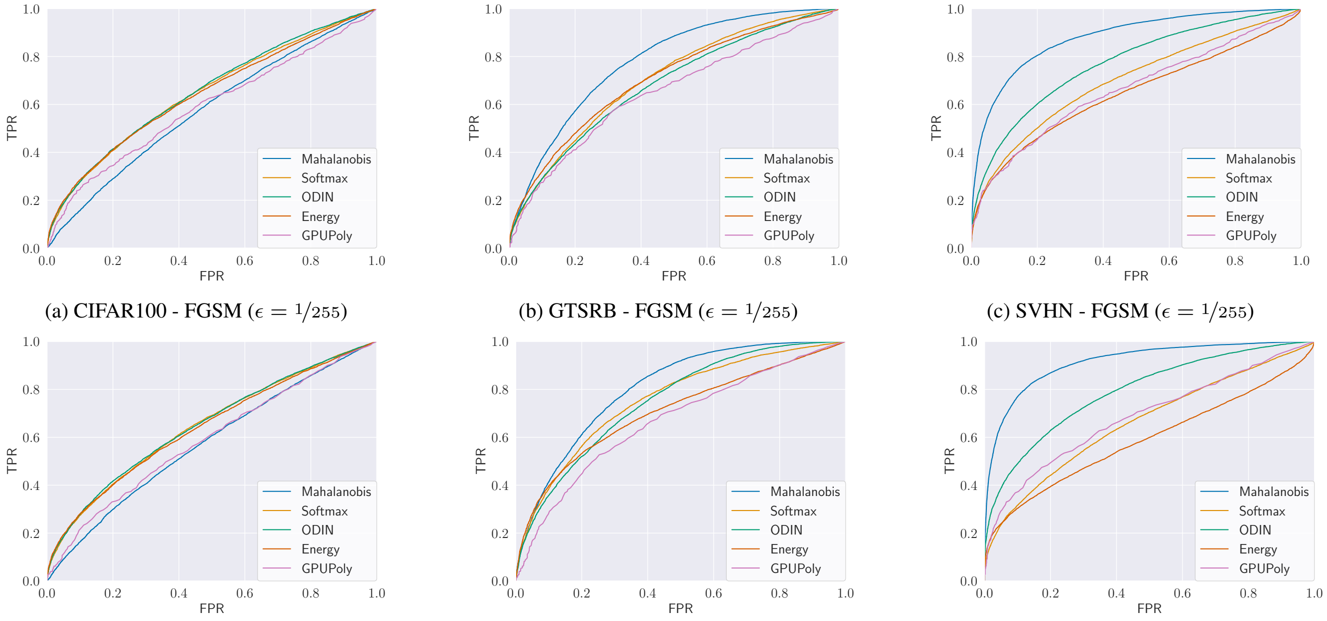


(a) CIFAR100 - Plain

(b) GTSRB - Plain

(c) SVHN - Plain

Figure 8. **ConvSmall - ID: CIFAR10.** Comparison of ROC curves for standard OOD detection methods and GPUPoly on CIFAR100, GTSRB and SVHN datasets.



(a) CIFAR100 - FGSM ($\epsilon = 1/255$)

(b) GTSRB - FGSM ($\epsilon = 1/255$)

(c) SVHN - FGSM ($\epsilon = 1/255$)

(d) CIFAR100 - PGD ($\epsilon = 1/255$)

(e) GTSRB - PGD ($\epsilon = 1/255$)

(f) SVHN - PGD ($\epsilon = 1/255$)

Figure 9. **ConvMed - ID: CIFAR10.** Comparison of ROC curves for standard OOD detection methods and GPUPoly on CIFAR100, GTSRB and SVHN datasets.