Appendix to Combining Outlierness Scores and Feature Extraction Techniques for Improvement of OoD and Adversarial Attacks Detection in DNNs

Tomasz Walkowiak[0000−0002−7749−4251], Kamil Szyc[0000−0001−6723−271X], and Henryk Maciejewski[0000−0002−8405−9987]

Wroclaw University of Science and Technology
{tomasz.walkowiak,kamil.szyc,henryk.maciejewski}@pwr.edu.pl

Table 1. OoD detection results for CIFAR-10 (as inliers) versus SVHN/CIFAR-100 (as outliers). Mahalanobis, MSP and LOF are working on features extracted from the last layer of CNN using the classic GAP method. WNN (our method) is a combination of 15 OoD detectors (obtained by three OoD methods and five feature extractors: GAP, CroW, lcGAP, SCDA, and GMP). We can see that adding information from additional feature extractors increases the OoD detection. The results show the mean and std values of the metrics achieved. The main message is that the proposed WNN is better than other methods.

Model	Method	AUROC	DATACC	TNR at TPR 95%	AUROC	DATACC	TNR at TPR 95%
		SVHN			CIFAR-100		
ResNet	Mah	93.1±0.09	85.4±0.15	64.2±0.47	70.8±0.13	66.7±0.12	8.5±0.44
	MSP	78.5±0.08	72.5±0.10	18.6±0.33	77.3±0.10	71.2±0.17	19.7±0.32
	LOF	92.2±0.08	84.7±0.12	60.5±0.59	69.6±0.21	66.0±0.15	8.4±0.25
	WNN	**97.1±0.05**	**91.1±0.12**	**85.8±0.29**	**78.9±0.20**	**72.3±0.17**	**20.5±0.92**
	KNN	84.5±0.16	77.7±0.15	30.1±0.45	70.5±0.18	67.2±0.19	9.0±0.25
AlexNet	Mah	28.5±0.21	50.0±0.00	0.9±0.06	54.6±0.15	53.9±0.13	5.2±0.13
	MSP	73.4±0.20	67.0±0.16	16.5±0.37	62.7±0.13	59.5±0.09	7.6±0.24
	LOF	26.3±0.11	50.0±0.00	0.8±0.07	52.8±0.18	52.8±0.16	4.4±0.17
	WNN	**86.4±0.16**	**78.4±0.15**	**41.9±0.56**	**64.2±0.34**	**61.0±0.26**	**7.9±0.27**
	KNN	22.7±0.12	50.1±0.01	0.4±0.02	45.9±0.19	50.3±0.03	2.5±0.08
ShuffleNet	Mah	84.8±0.14	77.6±0.09	30.0±0.52	61.3±0.13	59.7±0.11	5.8±0.21
	MSP	79.0±0.15	72.7±0.18	20.0±0.31	76.8±0.14	70.7±0.14	18.0±0.42
	LOF	87.2±0.10	80.7±0.08	28.5±0.30	68.8±0.16	64.3±0.19	11.5±0.13
	WNN	**93.5±0.07**	**86.0±0.15**	**65.2±0.74**	**77.9±0.14**	**71.5±0.13**	**20.1±0.56**
	KNN	82.9±0.14	76.8±0.11	18.9±0.30	62.8±0.21	60.3±0.13	7.1±0.24
WideResNet	Mah	94.5±0.07	86.9±0.11	70.7±0.36	72.1±0.17	68.0±0.13	10.4±0.34
	MSP	83.3±0.18	76.6±0.20	25.6±0.39	80.3±0.11	73.6±0.20	24.1±0.45
	LOF	92.4±0.09	84.7±0.13	60.4±0.53	78.1±0.16	72.0±0.13	19.8±0.50
	WNN	**97.4±0.05**	**91.5±0.11**	**86.0±0.48**	**81.4±0.16**	**74.4±0.18**	**24.8±0.79**
	KNN	90.1±0.09	82.2±0.15	49.8±0.35	77.2±0.19	72.2±0.18	17.1±0.65
MobileNet	Mah	80.0±0.11	73.9±0.16	18.8±0.39	67.2±0.09	63.5±0.08	9.2±0.20
	MSP	80.3±0.14	73.1±0.14	25.3±0.58	76.3±0.14	70.3±0.06	17.4±0.54
	LOF	84.7±0.14	78.8±0.14	22.4±0.78	70.5±0.21	65.4±0.19	12.8±0.31
	WNN	**94.2±0.14**	**86.4±0.27**	**70.3±1.01**	**77.3±0.20**	**70.8±0.20**	**18.8±0.61**
	KNN	66.5±0.21	65.9±0.14	6.3±0.18	68.2±0.23	65.0±0.20	10.3±0.32
VGG16	Mah	78.8±0.17	73.7±0.07	14.7±0.49	75.5±0.12	70.5±0.12	14.1±0.25
	MSP	76.4±0.13	70.3±0.16	19.6±0.35	73.3±0.22	68.8±0.16	14.7±0.25
	LOF	85.5±0.14	78.8±0.13	31.3±0.38	74.2±0.14	69.7±0.14	14.1±0.21
	WNN	**92.6±0.19**	**85.5±0.24**	**58.4±1.38**	**77.3±0.20**	**71.7±0.24**	**17.3±0.43**
	KNN	80.3±0.09	73.4±0.08	23.1±0.42	75.3±0.16	70.2±0.14	16.0±0.26