Supplementary Material

for Metric Learning for Novelty and Anomaly Detection

Marc Masana
mmasana@cvc.uab.cat
Idoia Ruiz
iruiz@cvc.uab.cat
Joan Serrat
joans@cvc.uab.cat
Joost van de Weijer
joost@cvc.uab.cat
Antonio M. Lopez
antonio@cvc.uab.cat

Computer Vision Center
Universitat Autònoma de Barcelona
Bellaterra, Spain

A Out-of-Distribution detection metrics

In out-of-distribution detection, comparing different detector approaches cannot be done by measuring only accuracy. The question we want to answer is if a given test sample is from a different distribution than that of the training data. The detector will be using some information from the classifier or embedding space, but the prediction is whether that processed sample is part of the in-distribution or the out-distribution. To measure that, we adopt the metrics proposed in [4]:

- **FPR at 95% TPR** is the corresponding False Positive Rate (FPR=FP/(FP+TN)) when the True Positive Rate (TPR=TP/(TP+FN)) is at 95%. It can be interpreted as the misclassification probability of a negative (out-distribution) sample to be predicted as a positive (in-distribution) sample.
- **Detection Error** measures the probability of misclassifying a sample when the TPR is at 95%. Assuming that a sample has equal probability of being positive or negative in the test, it is defined as $0.5(1-\text{TPR})+0.5\text{FPR}$.

where TP, FP, TN, FN correspond to true positives, false positives, true negatives and false negatives respectively. Those two metrics were also changed to **TNR at 95% TPR** and **Detection Accuracy** in [3], which can be calculated by doing $1-x$ from the two metrics above explained respectively. We use the latter metrics only when comparing to other state-of-the-art methods. This is also done because the implementation in both [3, 4] allows for using a TPR which is not at 95% in some cases, meaning that the Detection Error can go below 2.5 since TPR is not fixed to 0.95.

In order to avoid the biases between the likelihood of an in-distribution sample to being more frequent than an out-distribution one, we need threshold independent metrics that

Table 1: Quantitative comparison between cross-entropy and metric learning based methods training on LeNet for MNIST – 2, 6, 7 (In-dist), 0, 3, 4 and 8 (Seen Out-dist) and 5, 9, 1 (Unseen Out-dist Novelty).

Method	In-dist accuracy	Out-dist	FPR at 95% TPR	Detection Error	AUROC	AUPR-in	AUPR-out
CE	99.70	Novelty	33.76	19.38	92.33	92.73	92.29
		Gaussian noise	0.70	2.85	98.85	99.21	98.14
		SVHN	0.23	2.60	99.48	98.64	99.91
		CIFAR-10	2.86	3.93	98.96	98.02	99.57
Ours - ML	99.54	Novelty	21.05	13.03	94.48	94.02	94.46
		Gaussian noise	0.00	1.95	98.54	99.21	95.15
		SVHN	0.00	1.74	98.88	98.76	99.61
		CIFAR-10	0.01	2.36	98.87	98.93	99.12
Ours - ODM	99.64	Novelty	0.16	1.67	99.95	99.94	99.96
		Gaussian noise	0.00	1.76	99.14	99.46	97.66
		SVHN	0.00	0.96	99.65	99.41	99.89
		CIFAR-10	0.00	1.31	99.54	99.45	99.68

measure the trade-off between false negatives and false positives. We adopt the following performance metrics proposed in [2]:

- **AUROC** is the Area Under the Receiver Operating Characteristic proposed in [1]. It measures the relation between between TPR and FPR interpreted as the probability of a positive sample being assigned a higher score than a negative sample.
- **AUPR** is the Area Under the Precision-Recall curve proposed in [5]. It measures the relationship between precision (TP/(TP+FP)) and recall (TP/(TP+FN)) and is more robust when positive and negative classes have different base rates. For this metric we provide both AUPR-in and AUPR-out when treating in-distribution and out-distribution samples as positive, respectively.

B Quantitative results of the MNIST experiment

In this section we present the quantitative results of the comparison on the MNIST dataset. In this case we allowed a 5-dimensional embedding space for ML so the representation is rich enough to make the discrimination between in-dist and out-dist. For CE, as it is fixed to the number of classes, the embedding space is 3-dimensional. In Table 1 we see that ML performs a better than CE on all cases. ODM almost solves the novelty problem while keeping a similar performance on anomalies as ML. It is noticeable that CE struggles a bit more with Gaussian noise than the other anomalies. In this case, CE still produces highly confident predictions for some of the noise images.

C Experimental results on additional Tsinghua splits

Alternatively to the Tsinghua split generated with the restrictions introduced in Section 4.2, we also perform the comparison in a set of 10 random splits without applying any restriction to the partition classes. We still discard the classes with less than 10 images per class. Table 2 shows the average performance for this set of splits with their respective standard

Table 2: Comparison between ODIN and our proposed learning strategies on a WRN-28-10 architecture, when using novelty, anomaly (background patches and Gaussian noise) as seen out-of-distribution data as well as not seen out-of-distribution. The experiments are performed on a set of 10 random splits and the metrics provided are the mean of the metrics on the individual splits ± its standard deviation.

Method	In-dist accuracy	Out-dist	FPR at 95% TPR	Detection error	AUROC	AUPR-in	AUPR-out
ODIN	99.29±0.05	Tsinghua (unseen)	20.85±2.28	12.92±1.14	93.50±1.05	93.78±1.93	92.41±0.73
		Background (unseen)	8.39±6.34	6.70±3.17	98.06±1.26	97.02±3.15	98.79±0.60
		Noise (unseen)	0.03±0.43	2.53±0.85	99.67±0.34	99.60±0.39	99.74±0.41
Ours - ML	99.16±0.16	Tsinghua (unseen)	21.05±3.25	13.03±1.62	94.18±0.92	94.42±1.12	92.75±1.08
		Background (unseen)	1.91±1.02	3.45±0.51	99.14±0.32	98.79±0.35	99.40±0.22
		Noise (unseen)	0.30±0.96	2.65±0.48	99.27±0.36	99.09±0.40	99.43±0.35
Ours - ODM	99.13±0.22	**Tsinghua** (seen)	16.29±4.53	10.65±2.26	96.27±0.86	96.78±0.93	95.11±1.15
		Background (unseen)	0.39±1.63	2.71±0.31	99.50±0.27	99.30±0.31	99.66±0.20
		Noise (unseen)	0.01±1.39	2.51±0.70	99.59±0.54	99.51±0.60	99.69±0.43
Ours - ODM	99.09±0.18	Tsinghua (unseen)	20.36±3.63	12.68±1.81	93.47±1.55	93.58±2.10	92.00±1.74
		Background (seen)	0.01±0.03	2.51±0.01	99.97±0.02	99.92±0.03	99.98±0.01
		Noise (unseen)	0.00±0.00	2.50±0.01	99.99±0.03	99.97±0.05	99.99±0.01
Ours - ODM	99.02±2.42	Tsinghua (unseen)	20.87±1.63	12.93±0.81	93.65±1.05	94.01±1.48	92.33±0.89
		Background (unseen)	0.97±1.19	2.99±0.60	99.14±0.19	98.90±0.23	99.39±0.19
		Noise (seen)	0.00±0.00	2.50±0.01	100.00±0.00	99.98±0.01	99.99±1.85

deviation. Since the split of the classes is random, this leads to highly similar or mirrored classes to be separated into in-distribution and out-distribution, creating situations that are very difficult to predict correctly. For instance, detecting that a turn-left traffic sign is part of the in-distribution while the turn-right traffic sign is part of the out-distribution, is very difficult in many cases. Therefore, the results from the random splits have a much lower performance, specially for the novelty case.

When comparing the metric learning based methods, ODM improves over ML for the test set that has been seen as out-distribution during training. In general, using novelty data as out-distribution makes an improvement over said test set, as well as for background and noise. However, when using background images to push the out-of-distribution further from the in-distribution class clusters in the embedding space, novelty is almost unaffected. The same happens when noise is used as out-distribution during training. This could be explained by those cases improving the embedding space for data that is initially not so far away from the in-distribution class clusters. This would change the embedding space to push further the anomalies, but would leave the novelty classes, originally much closer to the clusters, almost at the same location.

When introducing out-of-distribution samples, the behaviour on the random splits is the same as for the restricted splits: while introducing novelty helps the detection on all cases, introducing anomaly helps the detection of the same kind of anomaly.

D Embeddings on Tsinghua

Figure 1 shows the embeddings for ODM (with novelty as seen out-od-distribution) and ML after applying PCA. When using ML, the novelties are not forced to be pushed away from the in-distribution clusters so they share the embedding space in between those same in-distribution clusters. In the case of ODM, the out-of-distribution clusters are more clearly separated from the in-distribution ones.

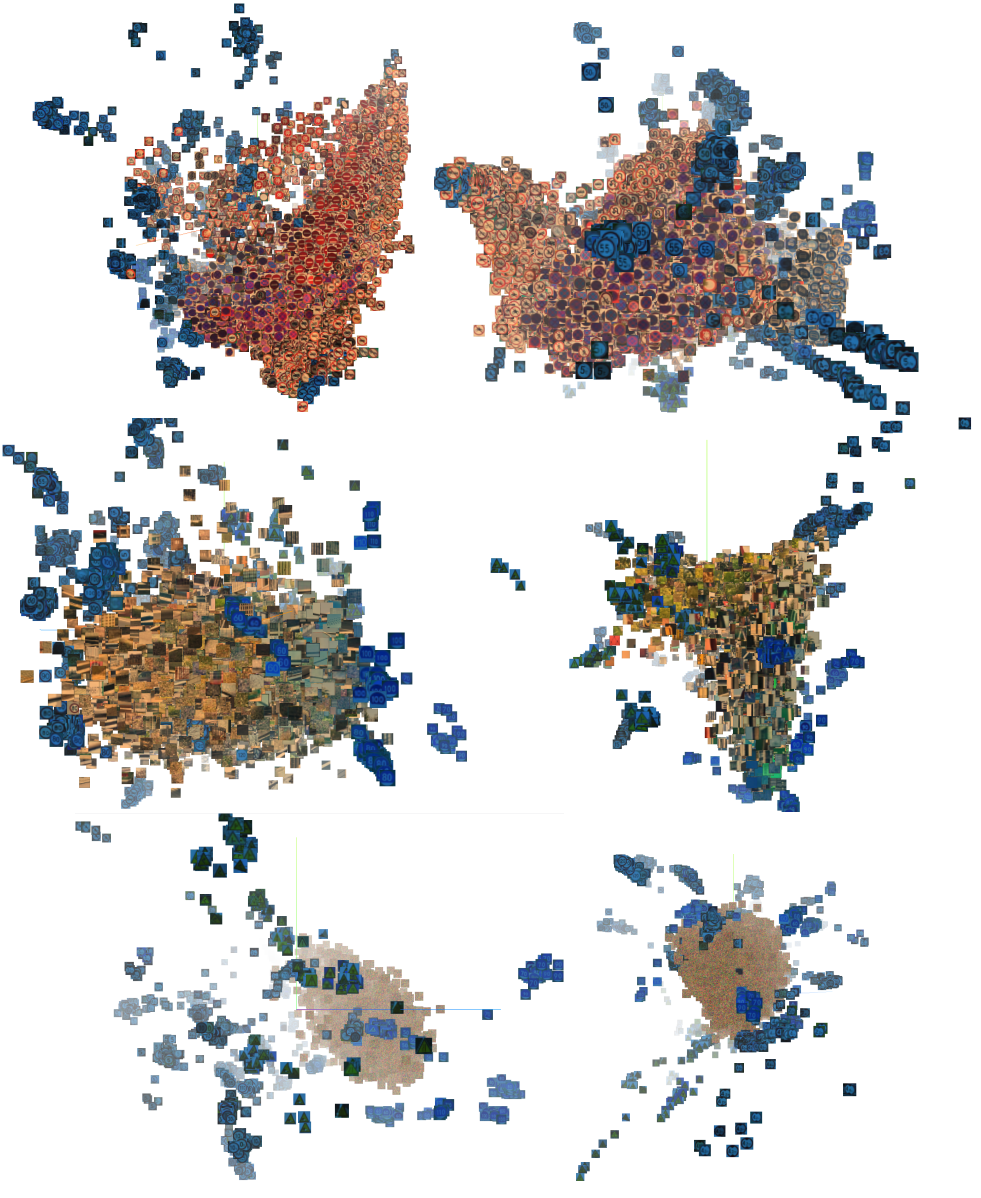

Figure 1: Embedding spaces after PCA for ODM (left) and ML (right) tested for in-dist (blue shaded) and out-dist (yellow shaded). Results are for TSinghua (first row), background patches (second row) and Gaussian noise (third row). Best viewed in color.

References

[1] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[2] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Int. Conference on Learning Representations (ICLR)*, 2017.

[3] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Int. Conference on Learning Representations (ICLR)*, 2018.

[4] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Int. Conference on Learning Representations (ICLR)*, 2018.

[5] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.