Rethinking Assumptions in Deep Anomaly Detection

Radek Bartyzal

GLAMI AI

16. 6. 2020

Anomaly Detection (AD)

Unsupervised

- classic approach
- we usually have unlabeled corpus of mostly nominal data

Semi-supervised

- = AD with negative examples
- = like unsupervised but push negative samples away from positive
- OE = outlier exposure = enrich dataset with data known to be anomalous = download random images

Supervised

- usually small number of known anomaly samples
- tricky to construct dataset that covers "everything else" as anomaly

Decision boundaries

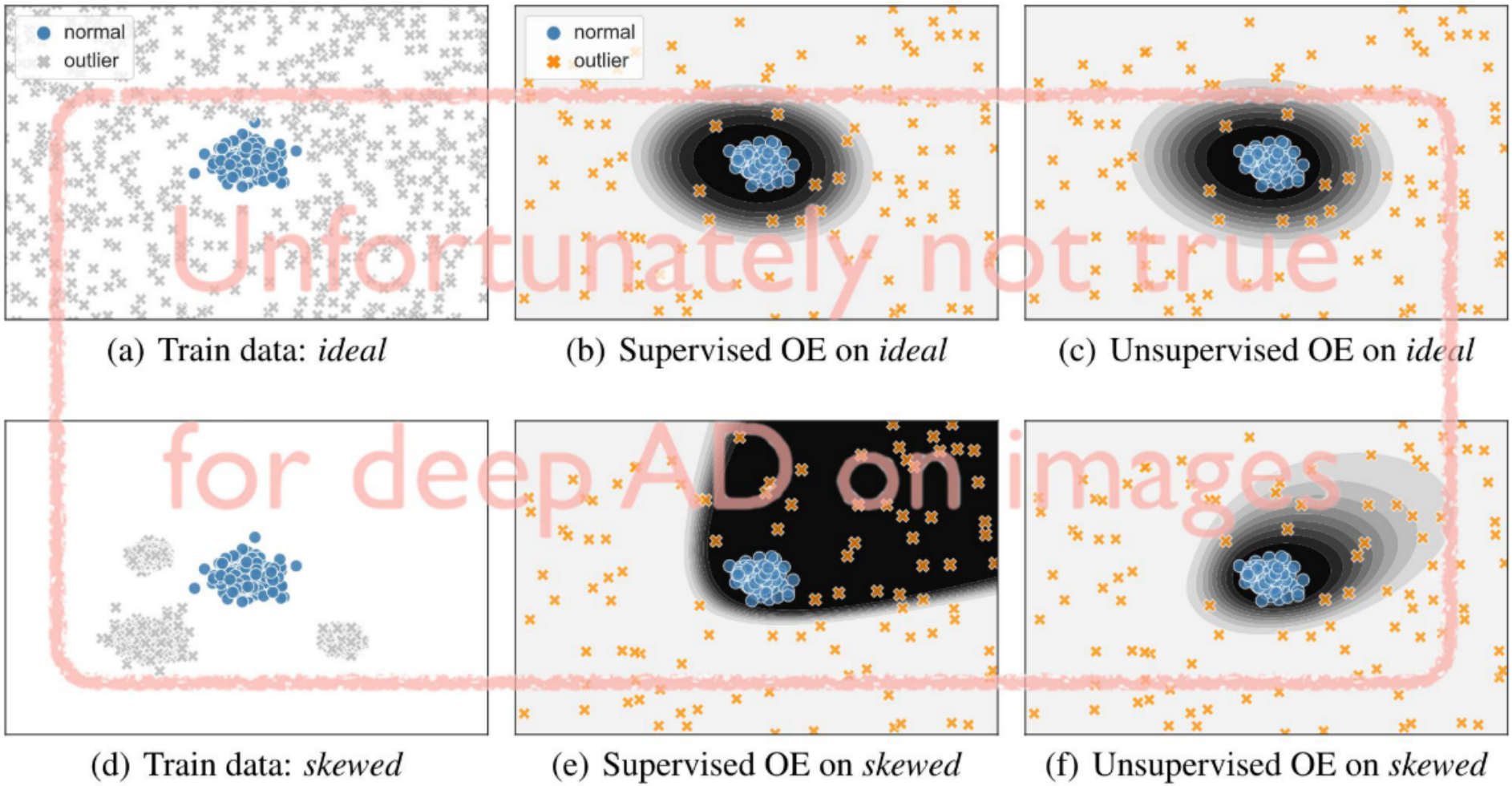


Figure: Supervised methods look for a clean decision boundary. Unsupervised / semi-sup. = clustering.

Overview of methods

- autoencoders trained on nominal data, where samples not reconstructed well are deemed anomalous
- use of OE: train net to predict uniform distribution (=random) for all outliers (=random images/data) = we expect that the nominal data are not random while everything else is
- **Unsupervised** deep version of support vector data description (Deep SVDD) Network is trained to map nominal samples to a center c:

$$\min_{\theta} \quad \frac{1}{n} \sum_{i=1}^{n} \|\phi(\boldsymbol{x}_i; \theta) - \boldsymbol{c}\|^2 .$$

- Deep **Semi-supervised** Anomaly Detection (Deep SAD). Deep SAD trains a network to concentrate nominal data near a predetermined center and maps anomalous samples away from that center.

Proposed method

Change Deep SAD to **Hypersphere classifier (HSC)** = cross-entropy classification that concentrates nominal samples together and pushes anomalies away = still unsupervised.

Cross entropy:

$$-\frac{1}{n}\sum_{i=1}^{n} y_i \log l(\phi(\boldsymbol{x}_i;\theta)) + (1-y_i)\log\left(1-l(\phi(\boldsymbol{x}_i;\theta))\right).$$

Use $l(z) = \exp(-\|z\|^2)$:

$$\frac{1}{n}\sum_{i=1}^{n} y_i \left\|\phi(\boldsymbol{x}_i;\theta)\right\|^2 - (1-y_i)\log\left(1-\exp\left(-\left\|\phi(\boldsymbol{x}_i;\theta)\right\|^2\right)\right).$$

Proposed method

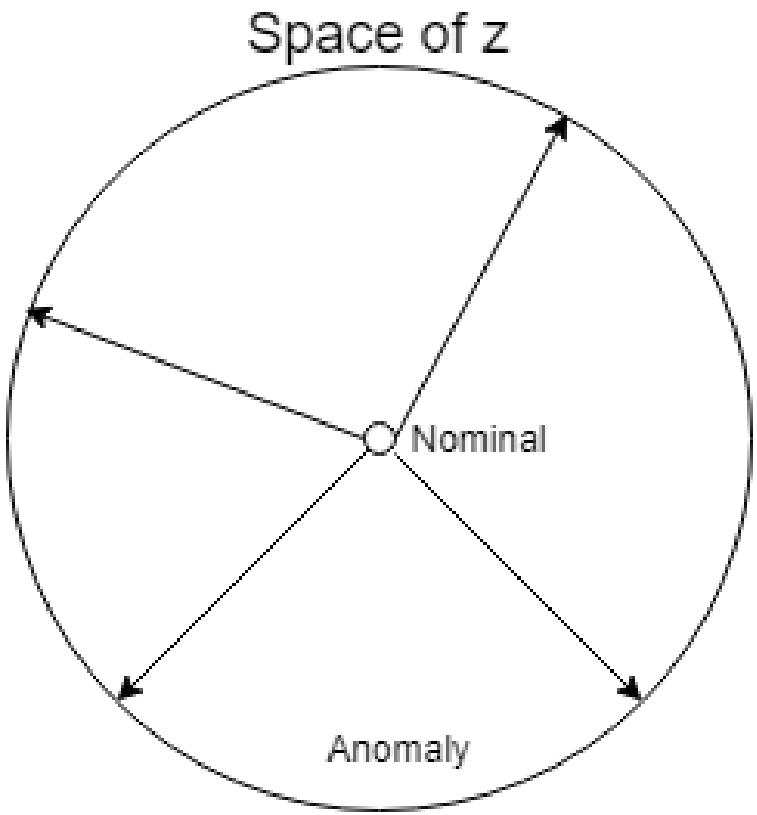


Figure: z is the output of the ANN.

Anomaly score $s(x) = \|ANN(x)\|^2$.

Goals

This paper challenges following common assumptions:

- Many (thousands) samples are needed for a deep method to understand a class of data.
- Anomalies are unconcentrated and thus inherently difficult to characterize with data.
- Therefore unsupervised approach is better than supervised.

Experiments

One vs. Rest Benchmark = "one" class (e.g, digit 0) as being nominal and the "rest" classes (e.g., digits 1–9) as being anomalous at test time.

Datasets:

- MNIST - with EMNIST-Letters as OE data
- CIFAR10 - with Tiny Images (80MTI) without CIFAR as OE
- ImageNet 1K - with ImageNet-22K without 1K as OE

Models:

- Unsupervised: SVDD, state-of-the-art methods = GEO, IT
- Semi-Supervised: HSC
- Supervised: standard binary cross-entropy classifier: BCE

Experiments: CIFAR10 one vs rest

	Unsupervised					Unsupervised OE			Supervised OE		
Class	SVDD*	Deep SVDD*	Geo*	IT*	Geo+*	Geo+*	Deep SAD	HSC	Focal*	Focal	BCE
Airplane	65.6	61.7	74.7	78.5	77.5	90.4	94.2	96.3	87.6	95.9	**96.4**
Automobile	40.9	65.9	95.7	89.8	96.9	**99.3**	98.1	98.7	93.9	98.7	98.8
Bird	65.3	50.8	78.1	86.1	87.3	**93.7**	89.8	92.7	78.6	92.3	93.0
Cat	50.1	59.1	72.4	77.4	80.9	88.1	87.4	89.8	79.9	88.8	**90.0**
Deer	75.2	60.9	87.8	90.5	92.7	**97.4**	95.0	96.6	81.7	96.6	97.1
Dog	51.2	65.7	87.8	84.5	90.2	**94.3**	93.0	94.2	85.6	94.1	94.2
Frog	71.8	67.7	83.4	89.2	90.9	97.1	96.9	97.9	93.3	97.8	**98.0**
Horse	51.2	67.3	95.5	92.9	96.5	**98.8**	96.8	97.6	87.9	97.6	97.6
Ship	67.9	75.9	93.3	92.0	95.2	**98.7**	97.1	98.2	92.6	98.0	98.1
Truck	48.5	73.1	91.3	85.5	93.3	**98.5**	96.2	97.4	92.1	97.5	97.7
Mean AUC	58.8	64.8	86.0	86.6	90.1	95.6	94.5	95.9	87.3	95.8	**96.1**

Figure: BCE performs better than SOTA unsupervised OE.

Experiments: CIFAR10 one vs rest

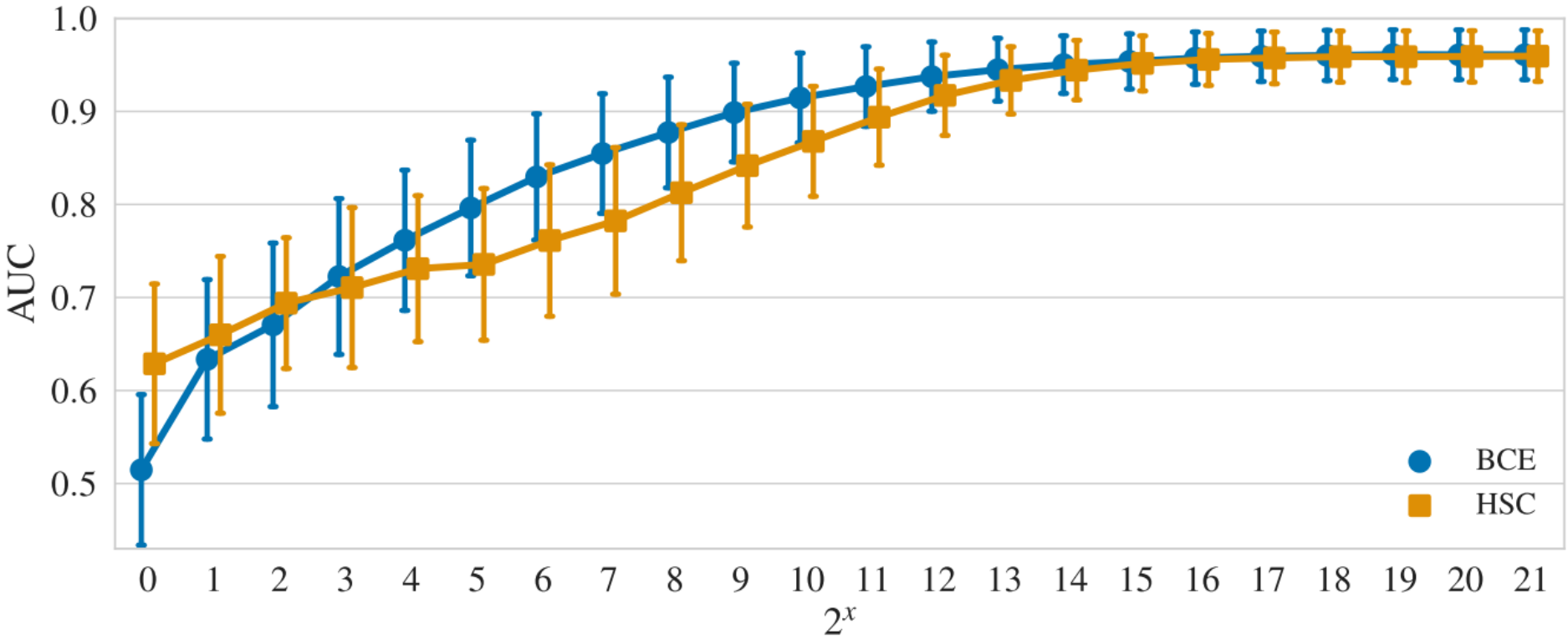


Figure: Detection performance in mean AUC in % (over 10 classes with 10 seeds per class) on the CIFAR-10 one vs. rest benchmark when varying the number of 80MTI OE samples. 32 OE samples is enough.

Experiments: ImageNet one vs rest

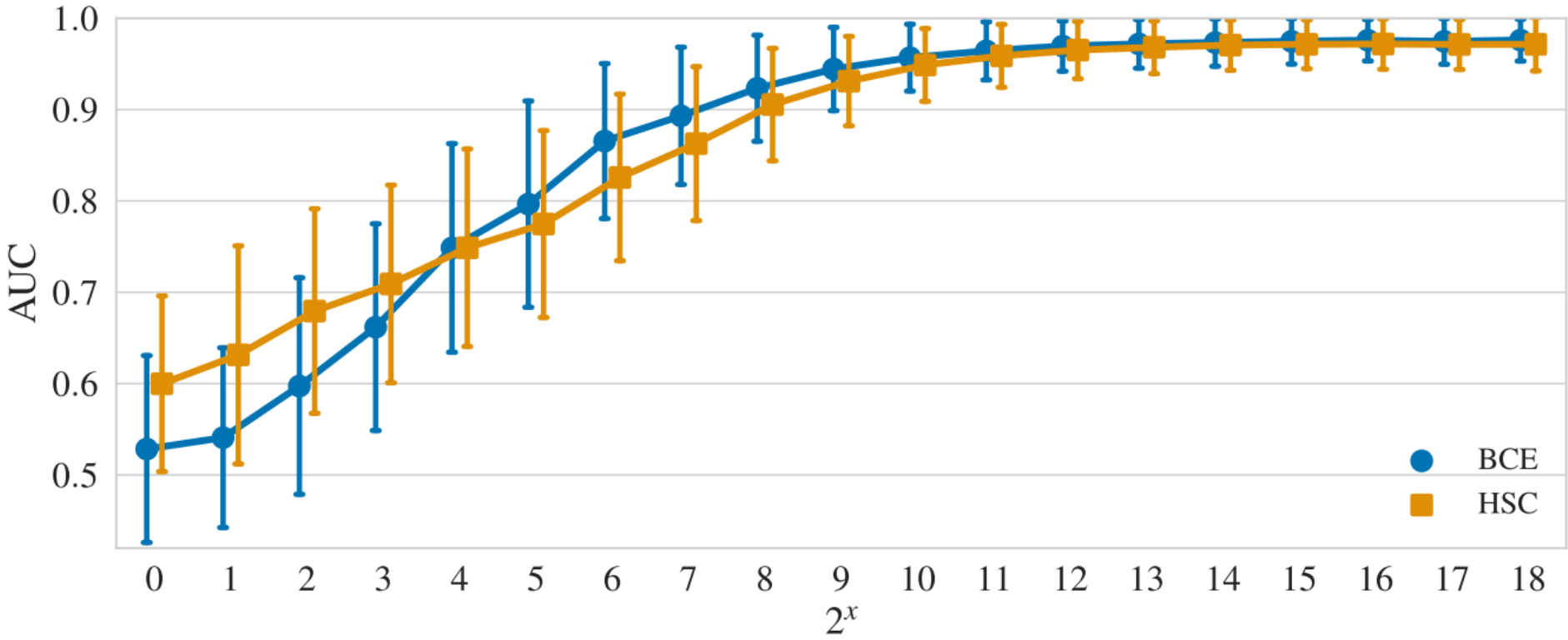
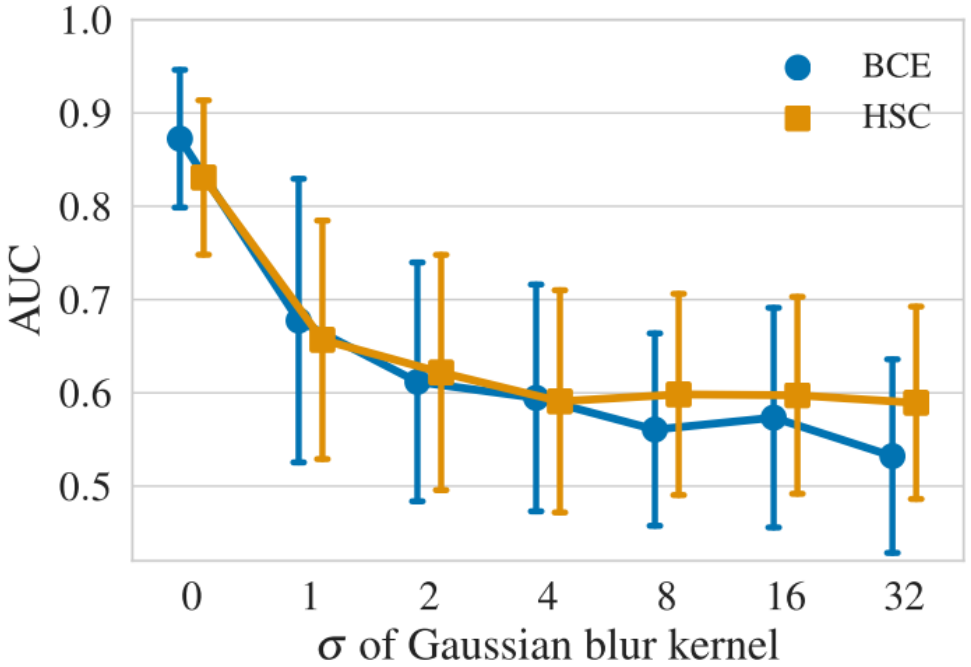


Figure: Detection performance in mean AUC in % (over 30 classes with 5 seeds per class) on the ImageNet-1K one vs. rest benchmark when varying the number of ImageNet-22K OE samples. 64 OE samples is enough.

Experiments: ImageNet blurring



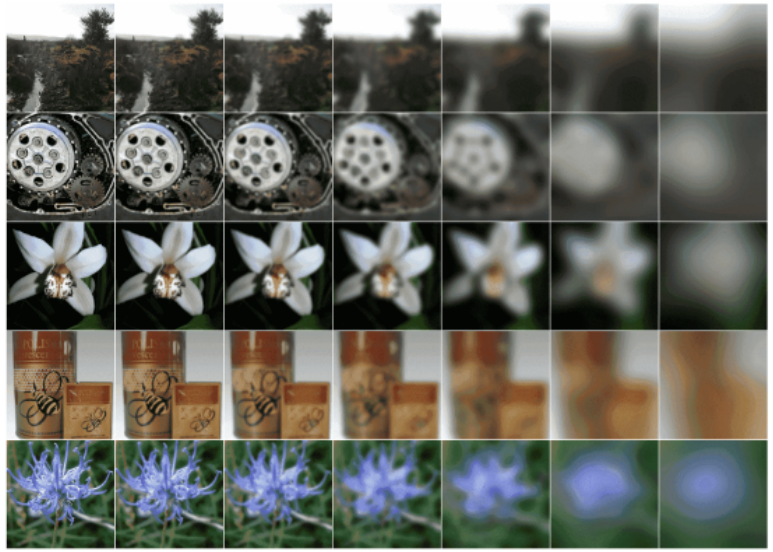


Figure: Detection performance in mean AUC on ImageNet when the OE data samples become increasingly blurred with a Gaussian kernel for having $2^6 = 64$ OE samples (left). An example of the various degrees of blurring is shown on the right. The abrupt decrease in AUC suggests the exceptional informativeness of OE on images is due to the multiscale structure of images.

Results

Key difference between classic AD and deep image AD is the presence of information at multiple spatial scales in images = small number of images contains a lot of information.

Supporting evidence:

- advantage of supervised OE over unsupervised OE is most evident on ImageNet, a high-resolution dataset. Smaller on CIFAR10 and nonexistent at MNIST.
- blurring ImageNet images = corrupting small scale features = drastic reduction in performance

Other results:

- very large amount of OE examples = unsupervised OE and supervised OE are equal

Sources

1. Ruff, Lukas, et al. "Rethinking Assumptions in Deep Anomaly Detection." arXiv preprint arXiv:2006.00339 (2020).
https://arxiv.org/abs/2006.00339