Window-Based Distribution Shift Detection for Deep Neural Networks

Guy Bar-Shalom[1], Yonatan Geifman[2], Ran El-Yaniv[1,2]

Technion[1], Deci Ai[2]

Introduction

Scan for paper Scan for Repo.

The study introduces Coverage-Based Detection (CBD) for detecting distribution shifts in deep neural networks, focusing on continuous monitoring to identify deviations in input data during operational phases.

Problem formulation and goal

We are given a pretrained model f, trained on a labeled set, $S_n \triangleq \{(x_1, y_1), \ldots, (x_n, y_n)\} \sim P^n$. We are also given an **unlabeled** and typically large detection-training set (or calibration set), denoted as $S_m \sim (P_X)^m$. The goal can be formulated as follows,

1. Given an unlabeled test sample, $W_k \sim Q^k$, where Q may be a different distribution from P_X, the task is to determine whether $Q \neq P_X$.
2. Achieve (1) while ensuring that the time and space complexity of each detection decision over a test window remains within $o(m)$ – avoiding continuously referencing S_m.

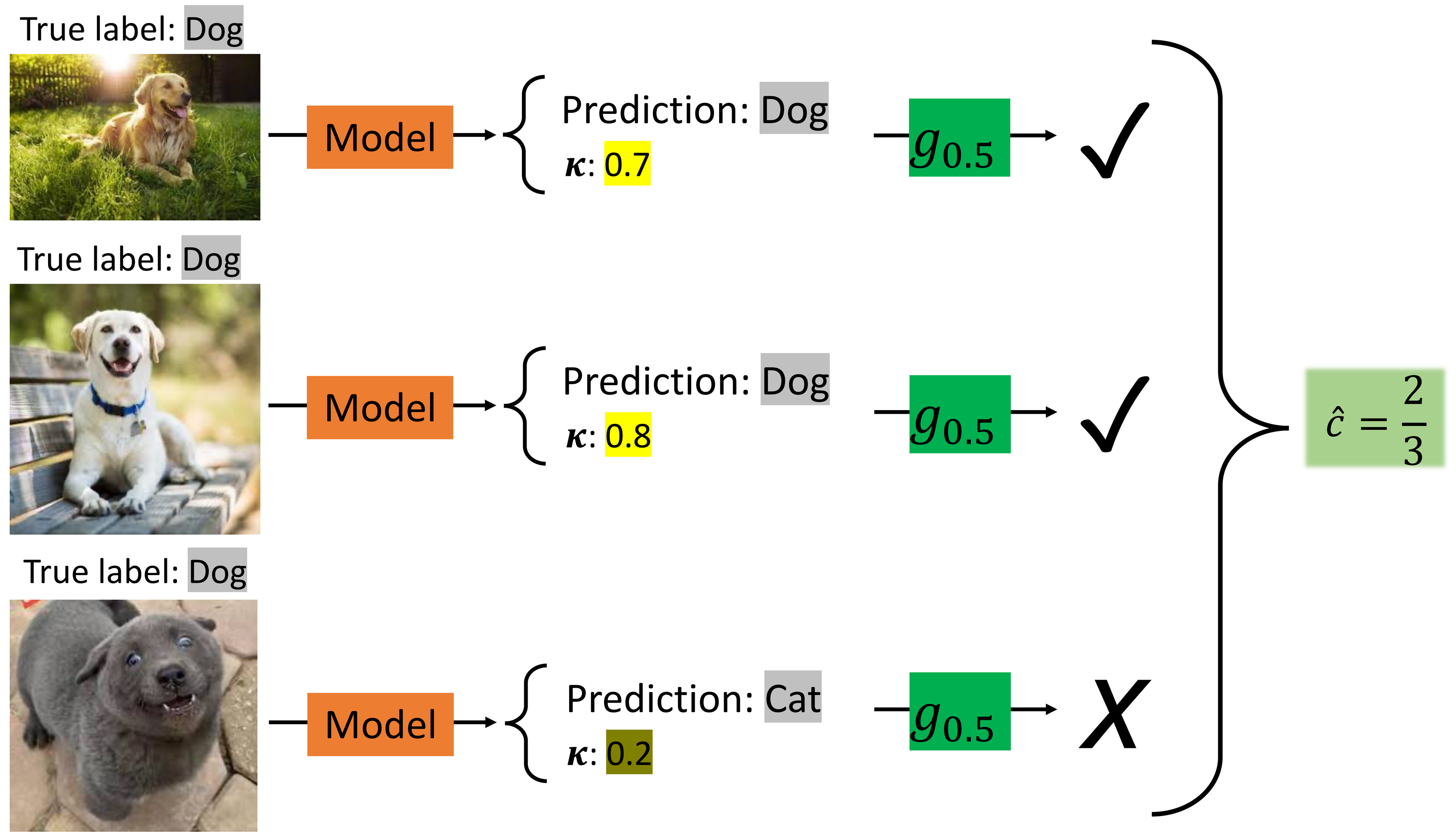
Contributions

1. A distribution shift detector, CBD, which can easily be integrated to any classification model, significantly outperforming earlier methods.
2. Given a test window of k samples, W_k, determine whether or not it is deviated from the original distribution, with $O(k)$ time and space complexities (independent of the size of S_m) – dramatic improvement to previous baselines.

Selective prediction

Selective prediction techniques aim to create models that make reliable predictions but can abstain under high uncertainty. We introduce key definitions and concepts for their use in detecting distribution shifts,

- $\kappa_f(x)$ - a confidence-rate function.
- $g_\theta(x|\kappa) \triangleq \mathbf{1}[\kappa_f(x) \geq \theta]$ - a selection function.
- $\hat{c}(\theta, S_k) \triangleq \frac{1}{k}\sum_{i=1}^{k} g_\theta(x_i|\kappa)$ - the empirical coverage of S_k given θ.
- $c(\theta, P_X) \triangleq \mathbf{E}_{P_X}[g_\theta(x|\kappa)]$ - the coverage (or true coverage) of P_X given θ.



Selection with Guaranteed Coverage (SGC)

SGC relies on Lemma 4.1 (see paper), which gets as input $\hat{c}(\theta, S_m)$, and returns b^*, such that,

$$\mathrm{Pr}_{S_m}\{c(\theta, P_X) < b^*(m, m \cdot \hat{c}(\theta, S_m), \delta)\} < \delta,$$

i.e., returns a lower bound on the true coverage, $c(\theta, P_X)$.

SGC gets as input:

- A detection-training set, $S_m \sim (P_X)^m$.
- A desired coverage (lower bound), c^*.
- Confidence parameter, δ.

And outputs:

- The actual guaranteed coverage (true coverage lower bound), b^*.
- The corresponding threshold, θ, for constructing the appropriate g_θ.

Algorithm 1: *Selection with guaranteed coverage* (SGC)

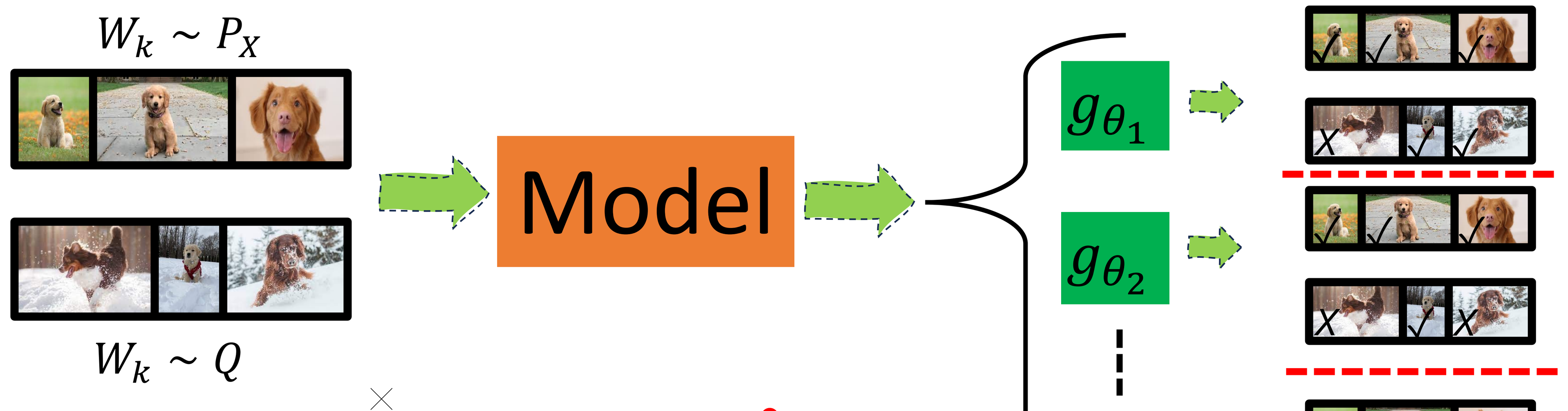
Input: detection-training set: S_m, confidence-rate function: κ_f, confidence parameter δ, target coverage: c^*.
Sort S_m according to $\kappa_f(x_i)$, $x_i \in S_m$ (and now assume w.l.o.g. that indices reflect this ordering).
$z_{\text{min}} = 1$, $z_{\text{max}} = m$
for $i = 1$ ***to*** $k = \lceil \log_2 m \rceil$ **do**
  $z = \lceil (z_{\text{min}} + z_{\text{max}})/2 \rceil$
  $\theta_i = \kappa_f(x_z)$
  Calculate $\hat{c}_i(\theta_i, S_m)$
  Solve for $b_i^*(m, m \cdot \hat{c}_i(\theta_i, S_m), \frac{\delta}{k})$ {see Lemma 4.1 in the paper}
  if $b_i^*(m, m \cdot \hat{c}_i(\theta_i, S_m), \frac{\delta}{k}) \leq c^*$ **then**
    $z_{\text{max}} = z$
  end
  $z_{\text{min}} = z$
end
Output: bound: $b_k^*(m, m \cdot \hat{c}_k(\theta_k, S_m), \frac{\delta}{k})$, threshold: θ_k.

Theorem. Assume S_m is sampled i.i.d. from P_X, and consider an application of Algorithm 1. For $k = \lceil \log_2 m \rceil$, let $b_i^*(m, m \cdot \hat{c}_i(\theta_i, S_m), \frac{\delta}{k})$ and θ_i be the values obtained in the i$^{\text{th}}$ iteration of Algorithm 1. Then,

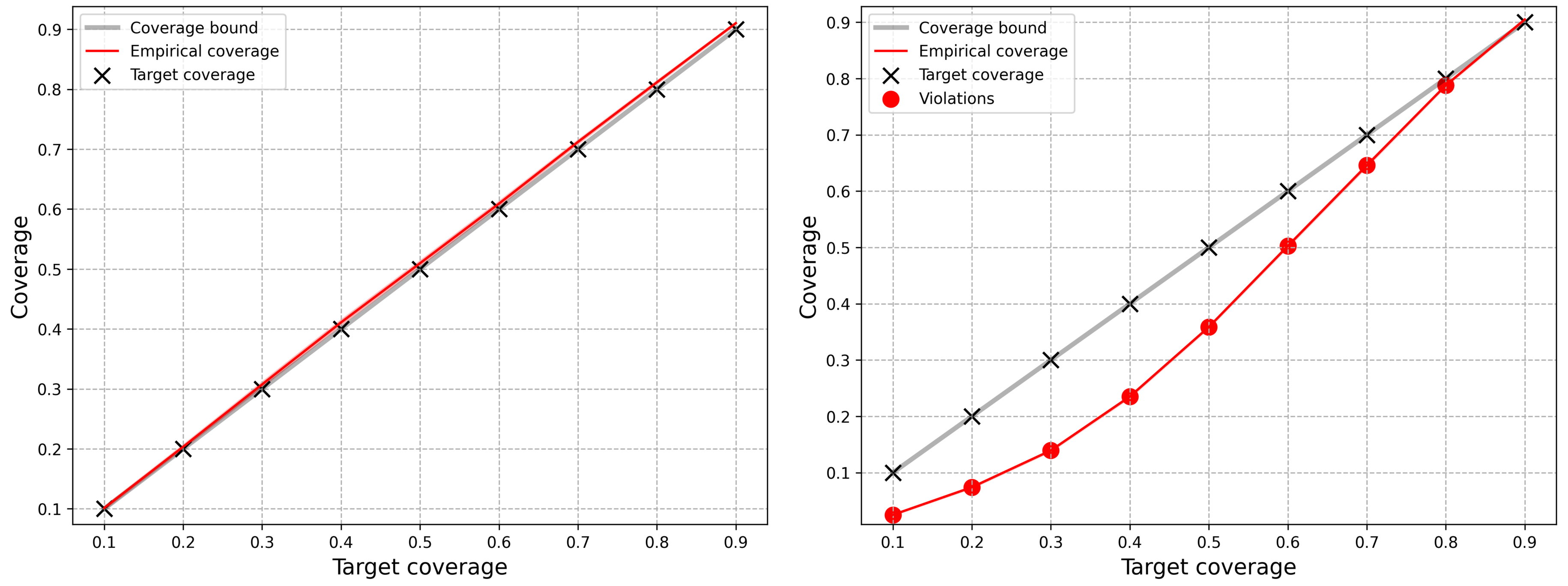
$$\mathrm{Pr}_{S_m}\{\exists i : c(\theta_i, P_X) < b_i^*(m, m \cdot \hat{c}_i(\theta_i, S_m), \tfrac{\delta}{k})\} < \delta.$$

Coverage-Based Detection (CBD)

Our CBD technique employs SGC across multiple target coverages (C_{target}) to identify the corresponding lower bounds and thresholds, $\{b_j^*, \theta_j\}_{j=1}^{C_{\text{target}}}$, for the true coverage of the underlying distribution.



$$V = \frac{1}{C_{\text{target}}} \sum_{j=1}^{C_{\text{target}}} \overbrace{(b_j^* - \hat{c}(\theta_j, W_k))}^{\text{Magnitude}} \cdot \overbrace{\delta(\hat{c}(\theta_j, W_k)|b_j^*)}^{\mathbf{1}[\hat{c} < b^*]}$$



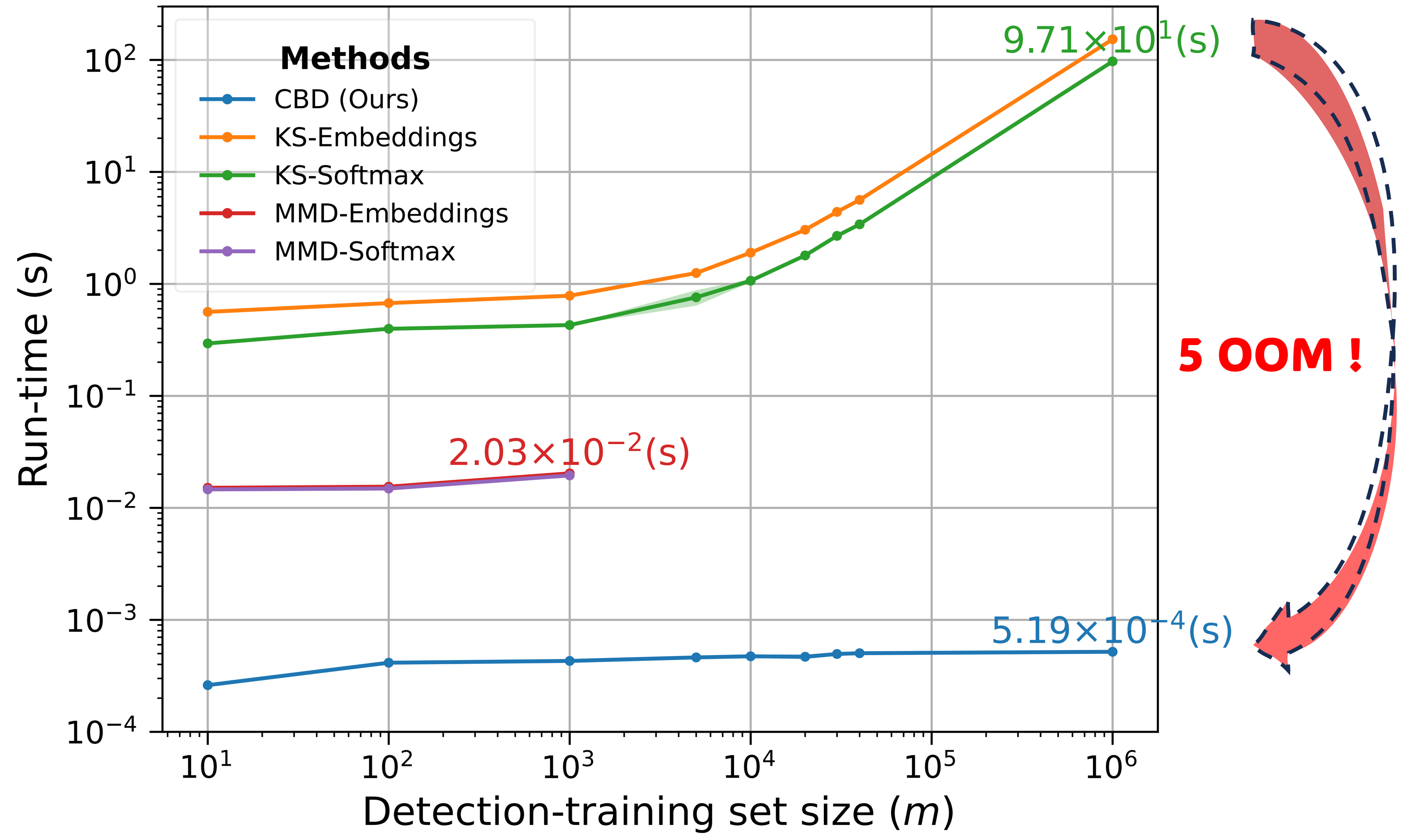
(a) No-shift case. (b) Shift case (ImageNet-O).

Complexity analysis

A theoretical complexity analysis is given →.

An empirical complexity analysis is given ↓ – showing our method's significant run-time advantage, 5 OOM reduction over the best baseline, with $m = 1,000,000$ detection-training samples.

Method	Space	Time
MMD	$O(m^2 + k^2 + mk)$	$O(d(m^2 + k^2 + mk))$
KS	$O(d(m + k))$	$O(d(m \log m + k \log k))$
Single-instance	$\boldsymbol{O(k)}$	$\boldsymbol{O(k)}$
CBD (Ours)	$\boldsymbol{O(k)}$	$\boldsymbol{O(k)}$



Experiments

We benchmark our method against a range of established benchmarks, including both population-based and single-instance detection techniques; CBD excels in most combinations of architecture, window size, and evaluation metrics. Notably, when applied over the ViT-T architecture, CBD achieves remarkable results, registering over 86% in all threshold-independent metrics (such as AUROC, AUPR-In, AUPR-Out), particularly across a window of 10 samples. This performance is significantly superior to its closest competitors, with CBD maintaining a substantial lead of approximately 20%.

Architecture	Method		Window size AUROC ↑ / AUPR-In ↑ / AUPR-Out ↑ / DetectionError ↓ / FPR@95TPR ↓						
			10	20	50	100	200	500	1000
ResNet50	KS	Softmax	61/67/62/34/67	73/74/74/31/64	87/90/85/13/27	89/89/89/15/29	94/95/92/7/14	99/99/99/2/4*	100/100/100/0.4/0.9
		Embeddings	72/74/73/28*/56*	68/73/74/24/48	81/84/79/18/37	75/76/79/22/44	76/79/79/20/40	84/87/84/13/26	86/88/84/13/26
	MMD	Softmax	54/61/56/36/72	62/65/62/37/72	73/76/72/29/56	73/73/78/33/59	79/79/79/35/54	83/85/83/15/30	85/85/85/22/37
		Embeddings	75/72/77/38/70	79/78/79/29/57	87/87/86/18/37	83/86/81/15/30	83/85/82/14/29	83/85/83/17/32	83/86/82/13/26
	Single-instance	SR	56/65/55/34/68	72/73/72/32/63	71/75/72/28/56	77/78/79/25/50	84/85/83/19/40	87/88/87/14/28	88/88/89/15/30
		Entropy	64/69/63/32/64	73/73/73/32/63	74/78/73/26/52	80/80/81/23/47	84/85/84/17/35	87/87/87/15/31	90/90/91/13/26
	CBD (Ours)		78/70/82*/42/84	88*/91*/87*/15*/30*	95*/95*/93*/9*/17*	93*/93*/92*/10*/20*	97*/97*/97*/5/10	98/98/98/4/7	100/100/100/0.4/0.9
MobileNetV3-S	KS	Softmax	71/72/75/32*/63*	84*/84/83/21/43	89/91/88/13/27	92/93/91/10/20	95/97/94/5/11	96/96/97/6/11	100/100/100/1/2
		Embeddings	63/67/63/37/75	65/66/67/37/75	77/78/76/27/54	72/73/76/27/53	84/83/86/22/43	86/87/86/15/30	79/81/81/18/36
	MMD	Softmax	75/73/75/38/72	78/78/78/30/59	86/89/82/17/32	86/89/84/14/26	87/88/86/14/28	89/90/88/12/24	90/91/88/11/22
		Embeddings	67/67/68/39/75	66/67/68/37/74	72/77/71/23/47	75/75/79/28/53	89/87/87/20/39	81/82/81/21/40	82/86/80/15/30
	Single-instance	SR	58/62/60/37/74	65/70/66/30/60	86/87/86/19/39	86/88/84/15/30	93/93/93/11/22	96/97/95/5/10	98/98/97/3/7
		Entropy	52/61/57/36/72	64/70/65/29/57	85/86/86/17/33	87/88/85/15/29	93/93/94/11/22	96/96/96/6/12	98/99/98/3/7
	CBD (Ours)		80*/73/82*/40/80	80/85/76/20/40	94*/95*/93*/8*/15*	94/94/94*/11/21	95/96/95/6/13	97/98/96/4/8	99/99/99/1/2
ViT-T	KS	Softmax	62/68/66/30/59	85/86/83/19/41	82/82/83/21/42	90/91/91/11/22	88/89/90/11/22	95/96/95/5/11	98/98/98/3/6
		Embeddings	68/66/71/38/76	76/83/73/19/38	82/86/80/17/34	81/82/81/20/39	81/84/80/17/34	76/75/82/22/44	84/83/86/19/38
	MMD	Softmax	58/59/64/44/82	69/70/73/38/69	77/80/77/22/44	75/80/76/20/40	80/86/78/15/29	89/91/90/12/23	93/94/92/7/15
		Embeddings	61/59/68/46/85	74/77/74/26/53	82/84/80/20/40	80/82/83/21/39	82/81/82/20/40	78/76/81/22/44	77/78/79/24/45
	Single-instance	SR	67/70/70/31/61	78/76/77/29/58	76/79/76/23/45	89/90/86/14/29	91/93/91/9/20	97/97/97/5/11	99/99/98/2/5
		Entropy	69/74/69/27/53	79/78/78/27/55	75/78/74/22/44	89/89/90/15/31	86/87/87/14/28	93/94/93/7/14	97/97/97/4/9
	CBD (Ours)		89*/86*/90*/28/56	91*/87/92*/24/47	94*/93*/95*/13*/25*	95*/96*/96*/8*/15*	97*/96*/97*/8/16	98*/98/98/4/9	99/99/99/3/6

My Email: guy.b@campus.technion.ac.il