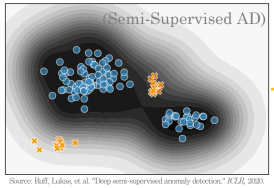
Understanding the Effect of Bias in Deep Anomaly Detection

Ziyu Ye Yuxin Chen Haitao Zheng

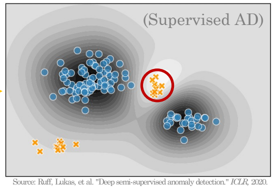
University of Chicago

1 *Motivation*: Bias from Additional Labeled Anomalies

Existing approaches for Anomaly Detection (AD)



Train with *additional labeled anomalies*. →



Semi-Supervised AD

Pro A compact enclosing of the normal.

Con Unable to use *additional labels*.
→ Underfitting bias.

Supervised AD

Pro A compact enclosing of the normal.
+
Discriminating on known anomalies.

A Counter-Intuitive Example

Training with *additional labeled anomalies* can induce disastrous harmful bias.

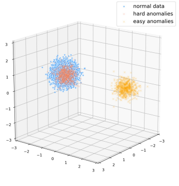


Fig 1. Original 3D Space

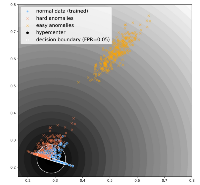


Fig 2. 2D Latent Space (*Semi-Supervised* AD)

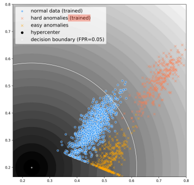


Fig 3. 2D Latent Space (*Supervised* AD)

Research Question

- Will unseen anomalies suffer from bias due to *additional labeled data in training*?
- If so, how can we *estimate* the bias? What is the *impact* of the bias?

[Clarification] Bias in AD ≠ Bias in Supervised Learning

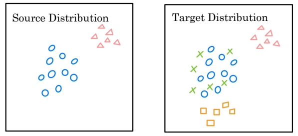


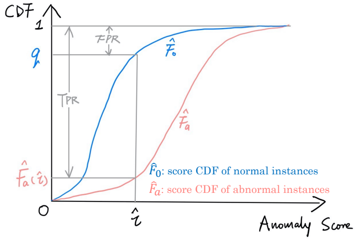
Fig 4. Data distribution of AD problem. The blue is the normal, others are different *subtypes* of anomalies.

Task Type	Distribution Shift	Known Target Distribution	Known Target Label Set
Imbalanced Classification [Johnson and Khoshgoftaar, 2019]	No	N/A	N/A
Closed Set Domain Adaptation [Saenko *et al.*, 2010]	Yes	Yes	Yes
Open Set Domain Adaptation [Panareda Busto and Gall, 2017]	Yes	Yes	No
Anomaly Detection [Chalapathy and Chawla, 2019]	Yes	No	No

Table 1: Comparison of anomaly detection tasks with other relevant classification tasks.

2 *Define Bias*: an ERM Framework

Scoring Bias $\mathrm{bias}(\hat{s}_\theta, \hat{\tau}_\theta) := \underset{(s_\theta, \tau_\theta):\theta\in\Theta}{\arg\max}\ \mathrm{TPR}(s_\theta, \tau_\theta) - \mathrm{TPR}(\hat{s}_\theta, \hat{\tau}_\theta)$



Relative **Scoring Bias**

$$\xi(s, s') := \mathrm{bias}(s, \tau) - \mathrm{bias}(s', \tau') = \mathrm{TPR}(s', \tau') - \mathrm{TPR}(s, \tau)$$

Empirical Relative **Scoring Bias**

$$\hat{\xi}(s, s') := \widehat{\mathrm{TPR}}(s', \tau') - \widehat{\mathrm{TPR}}(s, \tau)$$

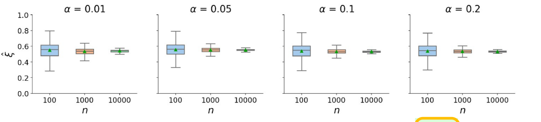
Proposition 1. *Given two scoring functions* s, s' *and a target FPR* q*, the relative scoring bias is* $\xi(s, s') = F_a(F_0^{-1}(q)) - F'_a(F_0'^{-1}(q))$.

3 *Estimate Bias*: a PAC Analysis

Theorem 3. *Assume that* $F_a, F'_a, F_0^-, F_0'^-$ *are Lipschitz continuous with Lipschitz constant* $\ell_a, \ell'_a, \ell_0^-, \ell_0'^-$*, respectively. Let α be the fraction of abnormal data from the mixture distribution. Then, w.p. at least* $1-\delta$*, with*

$$n = \mathcal{O}\left(\frac{1}{\alpha^2\epsilon^2}\log\frac{1}{\delta}\right)$$

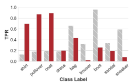
the empirical relative scoring bias satisfies $|\hat{\xi} - \xi| \le \epsilon$.



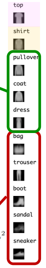
The estimation error ϵ decreases at the rate of $\frac{1}{\sqrt{n}}$.

4 *Characterize Bias*: Empirical Experiments

Scenario 1 Training w/ *hard* anomalies.



training normal = top, training abnormal = shirt



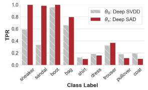
Positive bias!

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to shirt
shirt	0.09 ± 0.01	0.71 ± 0.01 ↑	0.70 ± 0.01 ↑	0.12 ± 0.01	0.72 ± 0.01 ↑	0.72 ± 0.01 ↑	0
pullover	0.13 ± 0.02	0.90 ± 0.01 ↑	0.89 ± 0.01 ↑	0.19 ± 0.02	0.84 ± 0.02 ↑	0.85 ± 0.01 ↑	0.01
coat	0.14 ± 0.03	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.15 ± 0.02	0.92 ± 0.02 ↑	0.92 ± 0.01 ↑	0.01
dress	0.17 ± 0.03	0.24 ± 0.03 ↑	0.24 ± 0.03 ↑	0.11 ± 0.01	0.20 ± 0.03 ↑	0.21 ± 0.03 ↑	0.04
bag	0.49 ± 0.07	0.38 ± 0.08 ↓	0.36 ± 0.07 ↓	0.70 ± 0.03	0.52 ± 0.09 ↓	0.53 ± 0.07 ↓	0.04
trouser	0.32 ± 0.10	0.07 ± 0.04 ↓	0.06 ± 0.03 ↓	0.59 ± 0.04	0.07 ± 0.04 ↓	0.16 ± 0.07 ↓	0.06
boot	0.92 ± 0.03	0.29 ± 0.15 ↓	0.27 ± 0.16 ↓	0.98 ± 0.02	0.90 ± 0.09 ↓	0.90 ± 0.08 ↓	0.08
sandal	0.30 ± 0.04	0.26 ± 0.08 ↓	0.26 ± 0.12 ↓	0.82 ± 0.02	0.46 ± 0.10 ↓	0.56 ± 0.09 ↓	0.09
sneaker	0.55 ± 0.09	0.12 ± 0.10 ↓	0.14 ± 0.12 ↓	0.74 ± 0.09	0.47 ± 0.19 ↓	0.46 ± 0.18 ↓	0.10

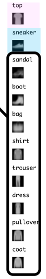
Table 3: The model TPR under scenario 1, Fashion-MNIST. The normal class top is similar to the abnormal training class shirt. Their L^2 distance = 0.02.

Negative bias!

Scenario 2 Training w/ *easy* anomalies.



training normal = top, training abnormal = sneaker



Mostly harmless bias!

Test data	Deep SVDD	Deep SAD	HSC	AE	SAE	ABC	L^2 to sneaker
sneaker	0.55 ± 0.09	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.74 ± 0.09	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0
sandal	0.30 ± 0.04	0.99 ± 0.01 ↑	0.98 ± 0.02 ↑	0.82 ± 0.02	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.02
boot	0.92 ± 0.03	1.00 ± 0.00 ↑	0.97 ± 0.02 ↑	0.98 ± 0.02	1.00 ± 0.00 ↑	1.00 ± 0.00 ↑	0.07
bag	0.49 ± 0.07	0.80 ± 0.05 ↑	0.81 ± 0.11 ↑	0.70 ± 0.03	0.84 ± 0.03 ↑	0.82 ± 0.03 ↑	0.07
shirt	0.09 ± 0.01	0.11 ± 0.02 ↑	0.12 ± 0.01 ↑	0.12 ± 0.01	0.13 ± 0.01 ↑	0.15 ± 0.01 ↑	0.10
trouser	0.32 ± 0.09	0.31 ± 0.10	0.11 ± 0.12 ↓	0.58 ± 0.04	0.58 ± 0.03	0.58 ± 0.05	0.12
dress	0.16 ± 0.03	0.16 ± 0.04	0.11 ± 0.01 ↓	0.11 ± 0.01	0.11 ± 0.01	0.12 ± 0.01	0.13
pullover	0.13 ± 0.02	0.13 ± 0.03	0.14 ± 0.05	0.19 ± 0.02	0.21 ± 0.03	0.19 ± 0.02	0.13
coat	0.14 ± 0.03	0.13 ± 0.03	0.13 ± 0.06	0.15 ± 0.02	0.16 ± 0.02	0.15 ± 0.02	0.14

Table 6: The model TPR under scenario 2, Fashion-MNIST. The normal class top is dissimilar to the abnormal training class sneaker, and the L^2 distance between the two is 0.13.

5 Takeaways and Future Directions

Additional labeled data in AD poses a *hidden threat* for model practitioners.

Data-Based **Debiasing Strategy**

- Using *active learning* to obtain representative anomaly labels.
- Leveraging *synthetic examples*.

Model-Based **Debiasing Strategy**

- Using *robust model design* (e.g., ensembles of semi-supervised and supervised models).