Breaking Down Out-of-Distribution Detection

Many Methods Based on OOD Training Data Estimate a Combination of the Same Core Quantities

Julian Bitterwolf
Alexander Meinke
Maximilian Augustin
Matthias Hein

University of Tübingen

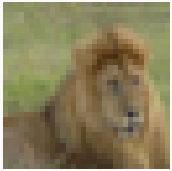
The OOD detection problem

Standard CIFAR-10 Classifier:


98.75%: bird


92.36%: bird


98.82 %: ship


99.47%: dog

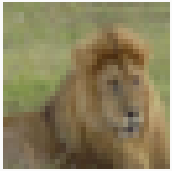
The OOD detection problem

Standard CIFAR-10 Classifier:


98.75%: bird


92.36%: bird


98.82 %: ship


99.47%: dog

Classify?

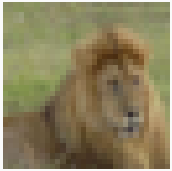
The OOD detection problem

Standard CIFAR-10 Classifier:


98.75%: bird


92.36%: bird


98.82 %: ship


99.47%: dog

~~Classify?~~ **Detect!** ➔ Discriminate In-distribution vs. OOD.

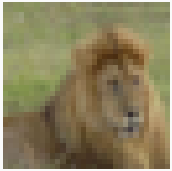
The OOD detection problem

Standard CIFAR-10 Classifier:


98.75%: bird


92.36%: bird


98.82 %: ship


99.47%: dog

~~Classify?~~ **Detect!** ➔ Discriminate In-distribution vs. OOD.

A **binary decision task!**

But can we train it like one?

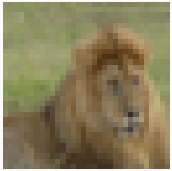
The OOD detection problem

Standard CIFAR-10 Classifier:


98.75%: bird


92.36%: bird


98.82 %: ship


99.47%: dog

~~Classify?~~ **Detect!** ➔ Discriminate In-distribution vs. OOD.

A **binary decision task!**

But can we train it like one?

Note: No training access to OOD samples from the same **distribution**.

Scoring functions and AUC

- OOD detector $\equiv$ **scoring function** $f : [0, 1]^d \to \mathbb{R}$

Scoring functions and AUC

- OOD detector $\equiv$ **scoring function** $f : [0, 1]^d \to \mathbb{R}$
- $\mathrm{AUC}_f\big(p(x|i), p(z|o)\big) = \mathop{\mathbb{E}}_{\substack{x\sim p(x|i)\\ z\sim p(z|o)}} \mathbb{1}_{f(x)>f(z)}$

Scoring functions and AUC

- OOD detector $\equiv$ **scoring function** $f : [0, 1]^d \to \mathbb{R}$
- $\mathrm{AUC}_f\big(p(x|i), p(z|o)\big) = \underset{\substack{x\sim p(x|i)\\ z\sim p(z|o)}}{\mathbb{E}} \mathbb{1}_{f(x)>f(z)}$
- Theorem 1:

  $$f \cong g \;\Leftrightarrow\; \forall p(x|i), p(z|o) : \mathrm{AUC}_f\big(p(x|i), p(z|o)\big) = \mathrm{AUC}_g\big(p(x|i), p(z|o)\big)$$

  is an equivalence relation.

  E.g.: $g(x) = 2 \cdot f(x)$ or $g(x) = \mathrm{sigmoid}(f(x))$

Bayes optimal predictions for given distributions

- Standard classifier: $\min_{\theta} \mathbb{E}_{(x,y)\sim p(x,y|i)} \mathcal{L}_{\mathrm{CE}}(\hat{p}_\theta(y|x), y)$

Bayes optimal predictions for given distributions

- Standard classifier: $\min_{\theta} \mathbb{E}_{(x,y)\sim p(x,y|i)} \mathcal{L}_{\mathrm{CE}}(\hat{p}_\theta(y|x), y)$
  - Bayes optimal solution: $\hat{p}^*_\theta(y|x) = p(y|x, i)$

Bayes optimal predictions for given distributions

- Standard classifier: $\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim p(x,y|i)} \mathcal{L}_{\text{CE}}(\hat{p}_\theta(y|x), y)$
  - Bayes optimal solution: $\hat{p}^*_\theta(y|x) = p(y|x, i)$

- Binary discriminator In vs. $\text{OOD}_{\text{train}}$: $\min_\theta \mathop{\mathbb{E}}_{x\sim p(x|i)} -\log \hat{q}_\theta(i|x) + \mathop{\mathbb{E}}_{x\sim p(x|o)} -\log(1 - \hat{q}_\theta(i|x))$

Bayes optimal predictions for given distributions

- Standard classifier: $\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim p(x,y|i)} \mathcal{L}_{\text{CE}}(\hat{p}_\theta(y|x), y)$
  - Bayes optimal solution: $\hat{p}^*_\theta(y|x) = p(y|x, i)$

- Binary discriminator In vs. $\text{OOD}_{\text{train}}$: $\min_\theta \mathop{\mathbb{E}}_{x\sim p(x|i)} -\log \hat{q}_\theta(i|x) + \mathop{\mathbb{E}}_{x\sim p(x|o)} -\log(1 - \hat{q}_\theta(i|x))$
  - Bayes optimal solution: $\hat{q}^*_\theta(i|x) = \dfrac{p(x|i)}{p(x|i) + p(x|o)}$

Bayes optimal predictions for given distributions

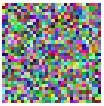
- Standard classifier: $\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim p(x,y|i)} \mathcal{L}_{\text{CE}}(\hat{p}_\theta(y|x), y)$
  - Bayes optimal solution: $\hat{p}^*_\theta(y|x) = p(y|x, i)$
- Binary discriminator In vs. $\text{OOD}_{\text{train}}$: $\min_\theta \mathop{\mathbb{E}}_{x\sim p(x|i)} -\log \hat{q}_\theta(i|x) + \mathop{\mathbb{E}}_{x\sim p(x|o)} -\log(1 - \hat{q}_\theta(i|x))$
  - Bayes optimal solution: $\hat{q}^*_\theta(i|x) = \dfrac{p(x|i)}{p(x|i) + p(x|o)}$

Lemma 3: If $p(x|o) = 1$, then $\hat{q}^*_\theta(i|x)$ is equivalent to $p(x|i)$.

Bayes optimal predictions for given distributions

- Standard classifier: $\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim p(x,y|i)} \mathcal{L}_{\text{CE}}(\hat{p}_\theta(y|x), y)$
  - Bayes optimal solution: $\hat{p}^*_\theta(y|x) = p(y|x, i)$
- Binary discriminator In vs. $\text{OOD}_{\text{train}}$: $\min_\theta \mathop{\mathbb{E}}_{x\sim p(x|i)} -\log \hat{q}_\theta(i|x) + \mathop{\mathbb{E}}_{x\sim p(x|o)} -\log(1-\hat{q}_\theta(i|x))$
  - Bayes optimal solution: $\hat{q}^*_\theta(i|x) = \dfrac{p(x|i)}{p(x|i)+p(x|o)}$

Lemma 3: If $p(x|o) = 1$, then $\hat{q}^*_\theta(i|x)$ is equivalent to $p(x|i)$.



Uniform Noise

BinDisc In vs. Uni

Density

Equivalences at their Bayes optima

- OOD detection with **background class** [1] $\cong$ BinDisc In vs. $\text{OOD}_{\text{train}}$

[1] S. Thulasidasan et al. 2021: An effective baseline for robustness to distributional shift.

Equivalences at their Bayes optima

- OOD detection with **background class** [1] $\cong$ BinDisc In vs. $\text{OOD}_{\text{train}}$
- **Energy**-based OOD detection [2] $\cong$ BinDisc In vs. $\text{OOD}_{\text{train}}$

[1] S. Thulasidasan et al. 2021: An effective baseline for robustness to distributional shift.
[2] W. Liu et al. 2020: Energy-based out- of-distribution detection.

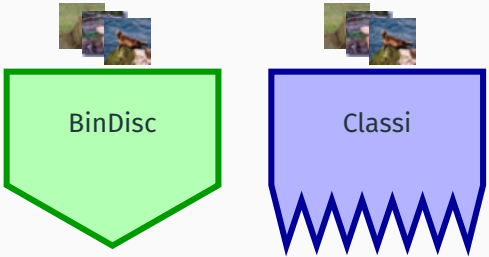
Equivalences at their Bayes optima

- OOD detection with **background class** [1] $\cong$ BinDisc In vs. $\text{OOD}_{\text{train}}$
- **Energy**-based OOD detection [2] $\cong$ BinDisc In vs. $\text{OOD}_{\text{train}}$
- **Outlier Exposure** [3] with confidence loss [4] $\cong \hat{q}^*_\theta(i|x) \cdot \left[\max_{y=1,\dots,K} \hat{p}^*_\theta(y|x,i) - \frac{1}{K} \right] + \frac{1}{K}$

[1] S. Thulasidasan et al. 2021: An effective baseline for robustness to distributional shift.
[2] W. Liu et al. 2020: Energy-based out- of-distribution detection.
[3] D. Hendrycks et al. 2019: Deep anomaly detection with outlier exposure.
[4] K. Lee et al. 2017: Training confidence-calibrated classifiers for detecting out-of-distribution samples.

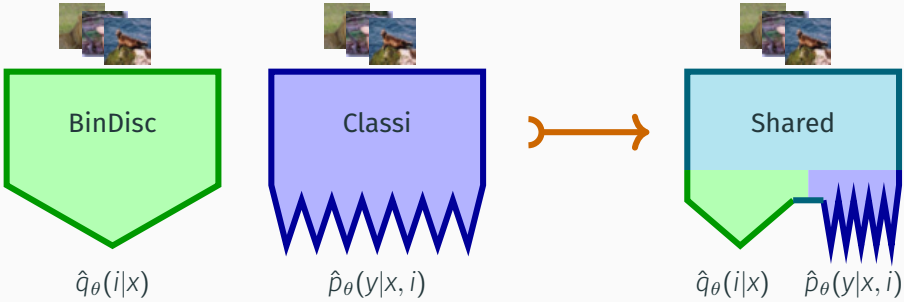
Shared training

- Classification (In) and Discrimination (In vs. $\text{OOD}_{\text{train}}$) can be treated by individual models.



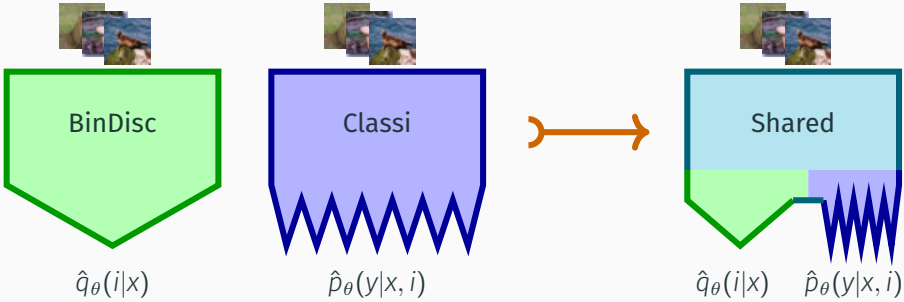
Shared training

- Classification (In) and Discrimination (In vs. $\text{OOD}_{\text{train}}$) can be treated by individual models.
- Shared training benefits both!



Shared training

- Classification (In) and Discrimination (In vs. $\text{OOD}_{\text{train}}$) can be treated by individual models.
- Shared training benefits both!
- Using shared training, the different equivalent methods behave similarly in practice.



Experimental results

	In: CIFAR-10 OOD_{train}: OpenIm		In: CIFAR-100 OOD_{train}: OpenIm		In: Restr. ImgNet OOD_{train}: Other classes	
Model	Acc.↑	$\overline{\text{FPR}}$↓	Acc.↑	$\overline{\text{FPR}}$↓	Acc.↑	$\overline{\text{FPR}}$↓
Plain Classi	95.16	53.01	77.16	67.57	96.34	36.71
Energy	94.13	18.59	73.47	35.91		
OE ($\cong s_3$)	95.06	**15.20**	77.19	35.03	97.10	4.26
BCG	95.21	18.83	**77.61**	**31.14**	97.50	4.22
BinDisc (Shared)		19.56		31.86		2.73
Classi (Shared)	**95.28**	29.34	77.35	67.23	**97.59**	17.51
Combi s_3	**95.28**	16.06	77.35	33.06	**97.59**	**2.62**

FPR at 95% TPR mean over 5-7 OOD datasets.

Combi s_3: Combination of BinDisc and Classi, equivalent to OE.