The Best of Both Worlds:
On the Dilemma of Out-of-Distribution Detection

Qingyang Zhang, Qiuxuan Feng, Joey Tianyi Zhou, Yatao Bian
Qinghua Hu and Changqing Zhang*

Tianjin University, A*Star, Tecent AI Lab

In a Nutshell

Goal: Out-of-distribution (OOD) detection and generalization.

Previous: OOD detection and generalization are conflict to each other.

Problem: Notable lack of **theorical justification**:

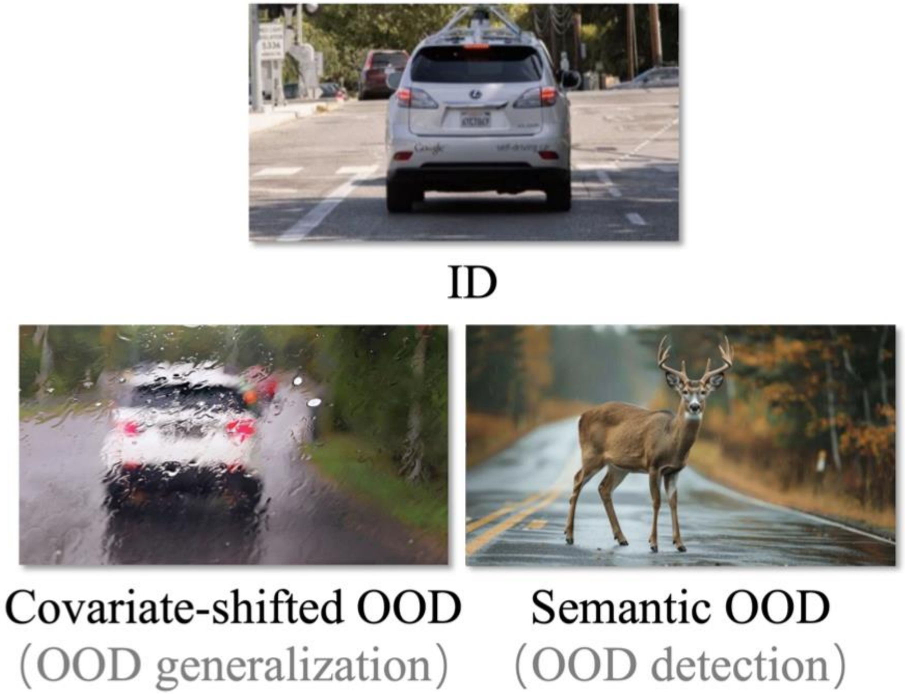
What is the relationship between these two tasks?

Our work: A joint framework for **dual optimal** OOD detection and generalization.

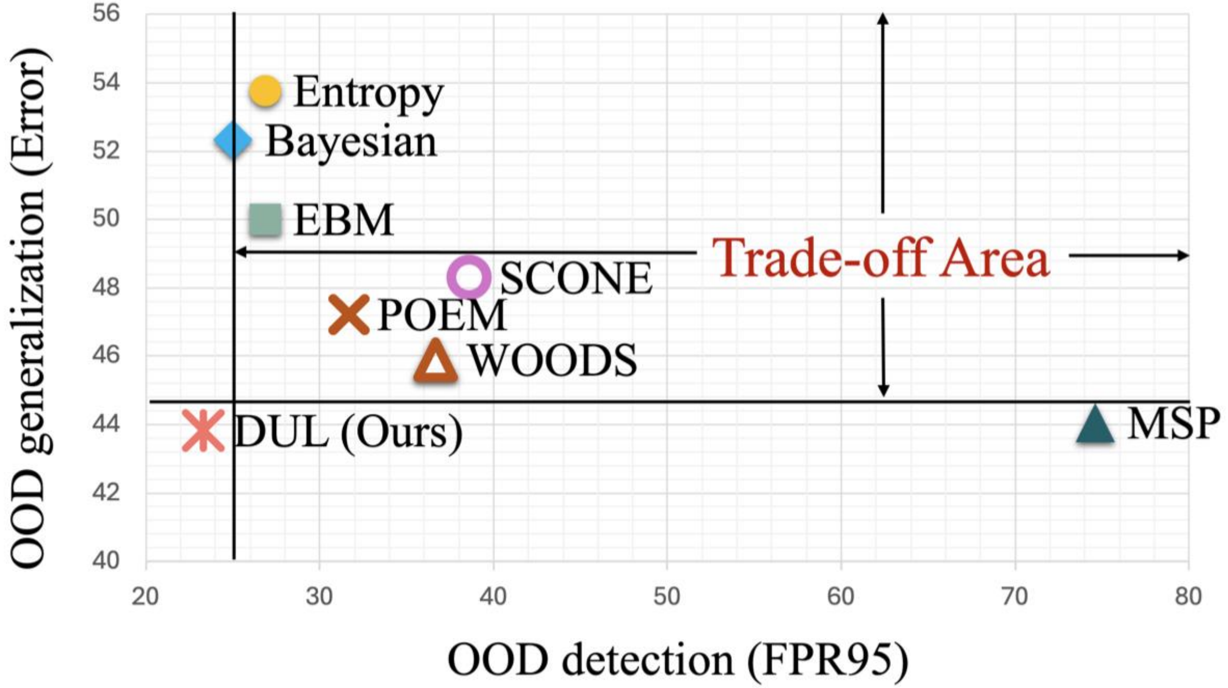
- ★ Theoretical analysis to understand the dilemma between OOD detection and generalization
- ★ A novel Bayesian optimization framework for dual optimal performance

Out-of-distribution Detection

天津大学 Tianjin University



(a) Three types of data arise in the open world



(b) Dilemma of current SOTA methods

Why do we care: one might want the model to be aware of outliers for safety, but certainly does not want to sacrifice the classification accuracy.

Theorical Analysis

Theorem 1 (Sensitive-robust dilemma)

[…] it holds that

$$\underbrace{\mathrm{GError}_{\mathcal{P}^{\mathrm{COV}}}(f)}_{\text{OOD generalization error}} \geq -\sqrt{\frac{1}{8\kappa^2}}\,\mathbb{E}_{\mathcal{P}^{\mathrm{SEM}}_{\mathrm{test}}}[\underbrace{\mathcal{L}_{\mathrm{reg}}(f)}_{\text{OOD detection loss}} - \log K]^{\frac{1}{2}} - \frac{1}{2\kappa} d_{\mathcal{F}}(\mathcal{P}^{\mathrm{COV}}, \mathcal{P}^{\mathrm{SEM}}_{\mathrm{test}}) + C$$

Under mild conditions, the error lower bound of MPS-based detectors are negatively correlated with OOD detection loss.

ID

Semantic OOD

Covariate-shifted OOD

OOD detection $\xrightarrow[\text{①}]{\textbf{encourages}}$ High uncertainty on auxiliary outliers $\xrightarrow[\text{②}]{\textbf{results in}}$ Uncertain prediction on target distribution $\xrightarrow[\text{③}]{\textbf{results in}}$ Degraded generalization

Decoupled Uncertainty Learning

天津大學 Tianjin University

Different types of uncertainty in Bayesian framework:

$$\underbrace{p(\hat{y}|x,\theta)}_{\text{overall uncertainty}} = \int \overbrace{p(\hat{y}|\mu)}^{\text{data uncertainty}} \underbrace{p(\mu|x,\theta)}_{\text{distributional uncertainty}} \, d\mu.$$

Related to OOD generalization (theorem 1).

Related to OOD detection by the definition.

One essential property is that high distributional uncertainty does not ensure high overall uncertainty.

The proposed decouple uncertainty learning:

$$\underbrace{\mathbb{E}_{P_{\mathrm{ID}}}[\mathcal{L}_{\mathrm{CE}}(f(x),y)]}_{\text{ID classification}} + \mathbb{E}_{P^{\mathrm{SEM}}_{\mathrm{train}}}\{\lambda \underbrace{||\max(0,(h_0+m_{\mathrm{OUT}})-h)||_{\tau}}_{\text{high distributional uncertainty}} + \underbrace{\gamma \mathrm{KL}(p(\hat{y}|\tilde{x})||p_0(\hat{y}|\tilde{x}))}_{\text{unchanged overall uncertainty}} \}$$

Decoupled Uncertainty Learning

Different types of uncertainty in Bayesian framework:

$$\underbrace{p(\hat{y}|x,\theta)}_{\text{overall uncertainty}} = \int \overbrace{p(\hat{y}|\mu)}^{\text{data uncertainty}} \underbrace{p(\mu|x,\theta)}_{\text{distributional uncertainty}} d\mu.$$

Related to OOD generalization (theorem 1).

Related to OOD detection by the definition.

One essential property is that high distributional uncertainty does not ensure high overall uncertainty.

Now the dilemma does not hold anymore.

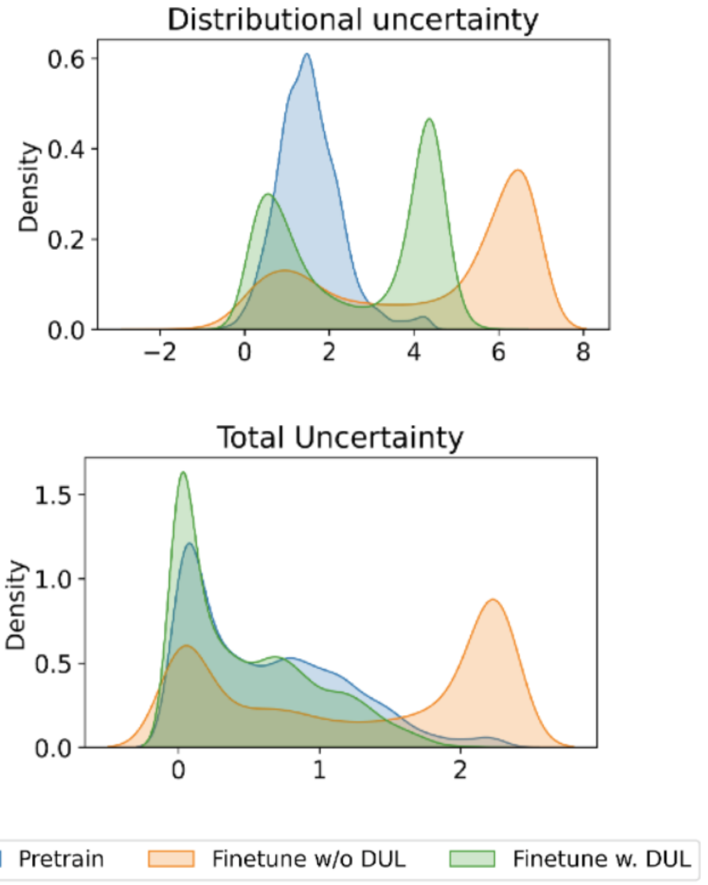
The logic chain has been broken down from here

$$\begin{array}{c}\text{OOD}\\\text{detection}\end{array} \xrightarrow[\text{①}]{\textbf{encourages}} \begin{array}{c}\text{High uncertainty}\\\text{on auxiliary outliers}\end{array} \xcancel{\xrightarrow[\text{②}]{\textbf{results in}}} \begin{array}{c}\text{Uncertain prediction}\\\text{on target distribution}\end{array} \xrightarrow[\text{③}]{\textbf{results in}} \begin{array}{c}\text{Degraded}\\\text{generalization}\end{array}$$

Experimental Results

$\mathcal{P}^{\text{ID}} / \mathcal{P}^{\text{SEM}}_{\text{train}}$	Method	Model generalization		OOD detection		
		ID-Acc ↑	OOD-Acc ↑	FPR ↓	AUROC ↑	AUPR ↑
CIFAR-100 / ImageNet-RC	Entropy	80.21	45.48	22.29	95.33	82.34
	EBM (finetune)	80.53	48.14	13.47	96.78	87.84
	POEM	78.15	42.18	**9.89**	**97.79**	**98.40**
	DPN	78.90	50.14	18.36	95.42	74.45
	WOODS	80.69	54.38	38.15	92.01	71.79
	SCONE	80.80	**56.73**	47.60	89.61	65.29
	DUL (ours)	**81.30**$^{\pm 0.04}$	56.27$^{\pm 3.29}$	12.49$^{\pm 0.05}$	95.24$^{\pm 0.01}$	86.72$^{\pm 0.58}$
	DUL† (ours)	81.23$^{\pm 0.05}$	55.41$^{\pm 0.54}$	11.12$^{\pm 0.62}$	95.46$^{\pm 0.36}$	96.49$^{\pm 0.13}$
CIFAR-100 / TIN-597	Entropy	80.15	46.25	26.88	93.50	79.81
	EBM (finetune)	79.94	50.00	26.87	91.68	80.08
	POEM	78.68	52.53	32.71	91.30	94.65
	DPN	78.44	47.67	24.99	93.55	81.63
	WOODS	79.26	53.13	36.71	92.15	73.42
	SCONE	79.53	52.70	35.60	92.47	73.58
	DUL (ours)	**80.85**$^{\pm 0.06}$	56.19$^{\pm 2.33}$	23.32$^{\pm 1.22}$	**94.48**$^{\pm 0.12}$	80.82$^{\pm 2.63}$
	DUL† (ours)	80.50$^{\pm 0.06}$	**56.22**$^{\pm 1.66}$	**22.75**$^{\pm 0.78}$	90.88$^{\pm 0.08}$	**96.33**
ImageNet-200 / None	MSP	85.15	74.84	58.23	86.98	82.27
	EBM (pretrain)	85.15	74.84	51.94	88.18	84.75
	Maxlogits	85.15	74.84	51.62	88.30	84.71
ImageNet-200 / ImageNet-800	Entropy	84.92	74.75	53.62	89.05	85.02
	EBM (finetune)	84.14	73.31	59.73	87.54	82.81
	DPN	84.87	74.40	63.84	87.18	80.69
	WOODS	84.99	74.98	51.71	88.30	84.80
	SCONE	84.93	74.91	52.52	88.19	84.50
	DUL (ours)	**85.65**$^{\pm 0.07}$	**75.59**$^{\pm 0.12}$	**49.14**$^{\pm 0.13}$	**89.27**$^{\pm 0.03}$	**85.62**$^{\pm 0.03}$

Dual optimal performance



Decoupled uncertainty

Summary

The Best of Both Worlds: On the Dilemma of Out-of-Distribution Detection

Qingyang Zhang, Qiuxuan Feng, Joey Tianyi Zhou, Yatao Bian
Qinghua Hu and Changqing Zhang*

Takeaway conclusion:

Dual optimal OOD detection and generalization can be achieved without trade-off.

Paper: https://arxiv.org/abs/2410.11576 **Code:** https://github.com/QingyangZhang/DUL