One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy

Jiacheng Zhang, Benjamin I. P. Rubinstein, Jingfeng Zhang, Feng Liu* (fengliu.ml@gmail.com)

THE UNIVERSITY OF MELBOURNE
TMLR Group — TRUSTWORTHY MACHINE LEARNING AND REASONING
RIKEN
THE UNIVERSITY OF AUCKLAND, Te Whare Wānanga o Tāmaki Makaurau, NEW ZEALAND

ICML International Conference On Machine Learning

An Upper Bound Without Constant: Significance of Distributional Discrepancy to Adversarial Defense

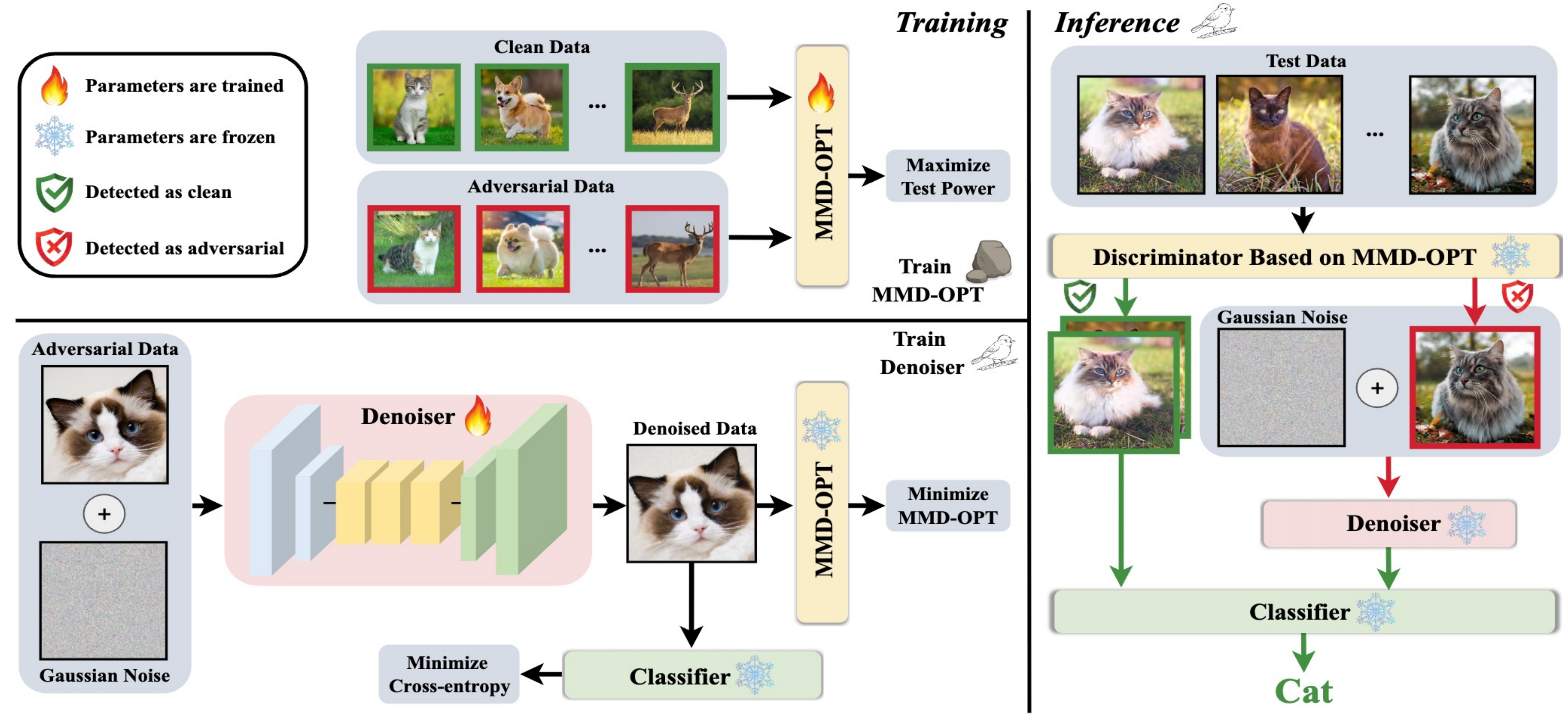
Theorem 1. *For a hypothesis $h \in \mathcal{H}$ and a distribution $\mathcal{D}_A \in \mathbb{D}$:*

$$R(h, f_A, \mathcal{D}_A) \leq R(h, f_C, \mathcal{D}_C) + d_1(\mathcal{D}_C, \mathcal{D}_A)$$

- $R(h, f_A, \mathcal{D}_A)$: expected loss on adversarial data
- $R(h, f_C, \mathcal{D}_C)$: expected loss on clean data
- $d_1(\mathcal{D}_C, \mathcal{D}_A)$: distributional discrepancy

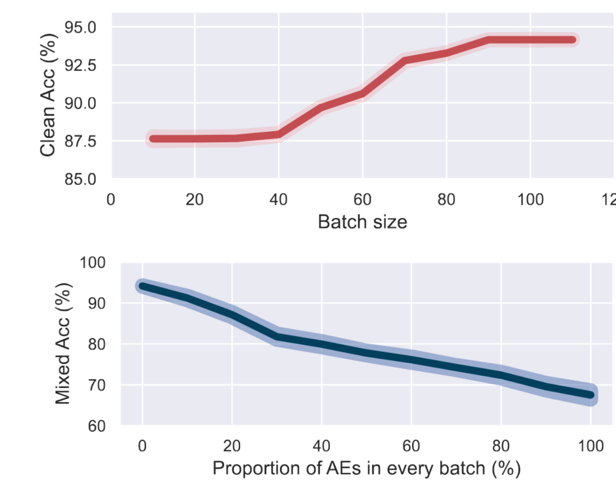
Distributional Discrepancy Minimization **reduces the expected loss on adversarial data**

A New Framework: Distributional-discrepancy-based Adversarial Defense



Experiment Results

Type	Method	ℓ_∞ ($\epsilon = 8/255$) Clean	ℓ_∞ ($\epsilon = 8/255$) Robust	Type	Method	ℓ_2 ($\epsilon = 0.5$) Clean	ℓ_2 ($\epsilon = 0.5$) Robust
WRN-28-10				**WRN-28-10**			
AT	Gowal et al. (2021)	87.51	63.38	AT	Rebuffi et al. (2021)*	91.79	78.80
	Gowal et al. (2020)*	88.54	62.76		Augustin et al. (2020)†	93.96	78.79
	Pang et al. (2022a)	88.62	61.04		Sehwag et al. (2022)†	90.93	77.24
AP	Yoon et al. (2021)	85.66	33.48	AP	Yoon et al. (2021)	85.66	73.32
	Nie et al. (2022)	90.07	46.84		Nie et al. (2022)	91.41	79.45
	Lee & Kim (2023)	90.16	55.82		Lee & Kim (2023)	90.16	83.59
Ours	DAD	**94.16 ± 0.08**	**67.53 ± 1.07**	Ours	DAD	**94.16 ± 0.08**	**84.38 ± 0.81**
WRN-70-16				**WRN-70-16**			
AT	Rebuffi et al. (2021)*	92.22	66.56	AT	Rebuffi et al. (2021)*	**95.74**	82.32
	Gowal et al. (2021)	88.75	66.10		Gowal et al. (2020)*	94.74	80.53
	Gowal et al. (2020)*	91.10	65.87		Rebuffi et al. (2021)	92.41	80.42
AP	Yoon et al. (2021)	86.76	37.11	AP	Yoon et al. (2021)	86.76	75.66
	Nie et al. (2022)	90.43	51.13		Nie et al. (2022)	92.15	82.97
	Lee & Kim (2023)	90.53	56.88		Lee & Kim (2023)	90.53	83.57
Ours	DAD	**93.91 ± 0.11**	**67.68 ± 0.87**	Ours	DAD	93.91 ± 0.11	**84.03 ± 0.75**



Type	Method	ℓ_∞ ($\epsilon = 4/255$) Clean	ℓ_∞ ($\epsilon = 4/255$) Robust
RN-50			
AT	Salman et al. (2020a)	64.02	34.96
	Engstrom et al. (2019)	62.56	29.22
	Wong et al. (2020)	55.62	26.24
AP	Nie et al. (2022)	71.48	38.71
	Lee & Kim (2023)	70.74	42.15
Ours	DAD	**78.61 ± 0.04**	**53.85 ± 0.23**

Trained on WRN-28-10

Unseen Transfer Attack		WRN-70-16	RN-18	RN-50	Swin-T
PGD+EOT (ℓ_∞)	$\epsilon = 8/255$	80.84 ± 0.46	80.78 ± 0.60	81.47 ± 0.30	81.46 ± 0.29
	$\epsilon = 12/255$	80.26 ± 0.60	80.54 ± 0.45	80.98 ± 0.36	80.40 ± 0.41
C&W (ℓ_2)	$\epsilon = 0.5$	82.45 ± 0.19	91.30 ± 0.20	89.26 ± 0.11	93.45 ± 0.17
	$\epsilon = 1.0$	81.20 ± 0.39	90.37 ± 0.17	88.65 ± 0.22	93.41 ± 0.18

Paper | Code | Contact