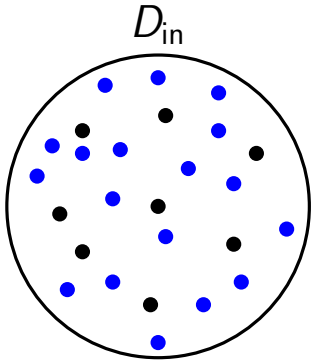
ICLR

Removing Undesirable Feature Contributions Using Out-of-Distribution Data

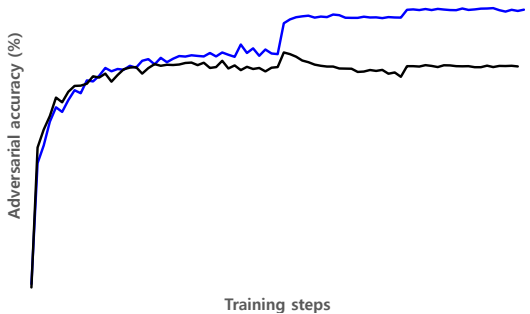
Saehyung Lee Changhwa Park Hyungyu Lee Jihun Yi Jonghyun Lee Sungroh Yoon

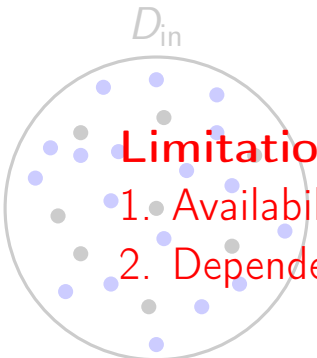
SEOUL
NATIONAL
UNIVERSITY



●: Training data

●: Unlabeled data



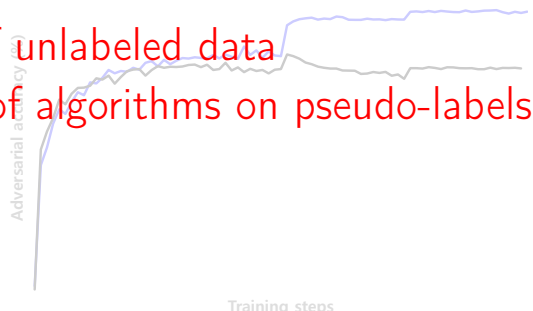


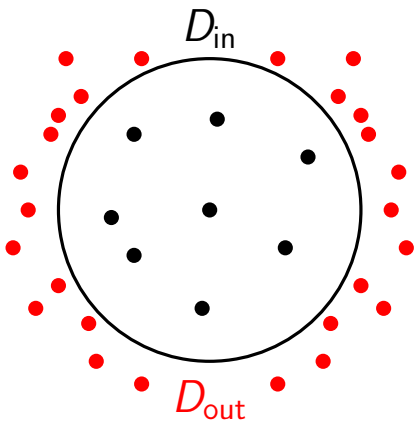
: Training data

: Unlabeled data

Limitations:

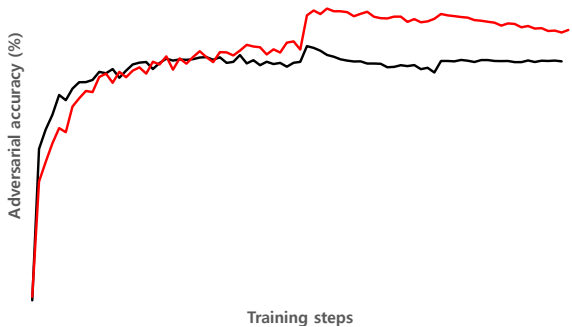
1. Availability of unlabeled data
2. Dependence of algorithms on pseudo-labels





●: Training data

●: OOD data



Assumption

Undesirable features are shared among diverse image datasets.

Non-robust classifier

$D_{\text{in}} \sim$ 

Robust feature extractor → $z_1 \underset{\sim \mathcal{N}(y, A)}{\times} w_1$

Non-robust feature extractor → $z_2 \underset{\sim \mathcal{N}(\eta y, B)}{\times} w_2$

$\Uparrow$ $Loss(\sigma(z_1 w_1 + z_2 w_2), y)$

where y = given label

Adversarial training

$D_{\text{in}} \sim$

Robust feature extractor

$z_1 \times w_1$
$\sim \mathcal{N}(y, A)$

Non-robust feature extractor

$z_2 \times w_2$
$\sim \mathcal{N}(\eta y, B)$

0

$Loss(\sigma(\tilde{z}_1 w_1 + \tilde{z}_2 w_2), y)$

where $\tilde{z}$ = perturbed feature

Proposed method

D_{out}~

Robust feature extractor

Non-robust feature extractor

$z_1 \times w_1$
$\sim\mathcal{N}(0, a)$

$z_2 \times w_2$
$\sim\mathcal{N}(\eta q, B)$

0

$Loss(\sigma(\tilde{z}_1 w_1 + \tilde{z}_2 w_2), y')$

where $y' =$ decision boundary

Proposed method

$$y' = \left[\frac{1}{c}, \frac{1}{c}, \ldots, \frac{1}{c}\right], \qquad \text{where c} = \#\text{ of classes}$$

Training mini-batch

Experimental details

- **80M-TI**: An OOD dataset created from the 80 Million Tiny Images dataset (Torralba et al.,2008) using confidence-based OOD detection algorithm (Carmon et al., 2019).
- Mainly compare the following setting:
  - **PGD**: The model trained using PGD-based adversarial training (Madry et al., 2017)on the target dataset.
  - **TRADES**: The model trained using TRADES (Zhang et al., 2019) on the target dataset.
  - $\textbf{OAT}_{\textbf{PGD}}$: The model which is adversarially trained with OAT based on a PGD approach.
  - $\textbf{OAT}_{\textbf{TRADES}}$: The model which is adversarially trained with OAT based on TRADES.

Adversarial training results

Model	Target	OOD	Clean	PGD100	CW100	AA
Standard	CIFAR10	-	95.48	0.00	0.00	0.00
PGD		-	87.48	49.92	50.80	48.29
PGD+CutMix		-	89.35	53.39	52.35	49.05
TRADES		-	85.24	55.69	54.04	52.83
OAT_{PGD}		80M-TI	86.63	**56.77**	**52.38**	**49.98**
OAT_{TRADES}		80M-TI	**86.76**	**59.66**	**55.71**	**54.63**
Standard	CIFAR100	-	78.57	0.02	0.00	0.00
PGD		-	61.37	24.66	24.68	22.76
TRADES		-	58.84	30.24	27.97	26.91
OAT_{PGD}		80M-TI	**61.54**	**30.02**	**27.85**	**25.36**
OAT_{TRADES}		80M-TI	**63.07**	**34.23**	**29.02**	**27.83**
Standard	ImgNet10 (64 x 64)	-	86.03	0.11	0.06	0.00
PGD		-	82.80	48.77	48.86	48.34
OAT_{PGD}		ImgNet990	81.91	**59.03**	**54.69**	**53.83**

Adversarial training results

	CIFAR10				ImgNet10 (64 x 64)		
OOD	None	SVHN	Simpson	Fashion	None	Places365	VisDA17
Clean	87.48	86.16	86.79	85.84	82.80	82.37	82.46
PGD20	50.41	**53.70**	**53.88**	**53.27**	49.00	**59.86**	**55.34**
CW20	51.11	**52.21**	**52.15**	**51.70**	48.91	**56.23**	**53.80**

Standard training results

Dataset N	CIFAR10 2,500 / Full	CIFAR100 2,500 / Full
Standard	65.44 / 94.46	24.41 / 74.87
OAT_{SVHN}	**68.56** / 94.45	**24.82 / 75.65**
$\text{OAT}_{\text{Simpson}}$	**70.08** / 94.43	**27.04 / 76.03**
$\text{OAT}_{\text{80M-TI}}$	**72.49 / 95.20**	**26.13 / 76.30**
Pseudo-label	- / 95.28	- / 77.24
Fusion	- / **95.53**	- / **77.36**

Dataset N	ImgNet10 (64 x 64) 100 / Full	ImgNet10 (160 x 160) 100 / Full
Standard	37.90 / 86.93	33.36 / 90.91
$\text{OAT}_{\text{VisDA17}}$	36.21 / 86.71	**35.93 / 91.23**
$\text{OAT}_{\text{Places365}}$	**41.84 / 88.37**	**40.11 / 91.42**
$\text{OAT}_{\text{ImgNet990}}$	**42.18 / 87.88**	**40.41 / 91.87**

Conclusion

- Propose out-of-distribution data augmented training (OAT) to leverage OOD data for adversarial and standard learning.
- Our theoretical analyses demonstrate how our proposed method can improve robust and standard generalization.
- The experimental results on CIFAR-10, CIFAR-100, and a subset of ImageNet suggest that OAT can help reduce the generalization gap in adversarial and standard learning.
- By applying OAT using various OOD datasets, it is shown that undesirable features are shared among diverse image datasets.

Thank you!

https://github.com/Saehyung-Lee/OAT

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) [2018R1A2B3001628], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2020, and AIR Lab (AI Research Lab) in Hyundai & Kia Motor Company through HKMC-SNU AI Consortium Fund.